UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA
INTERNATIONAL DOCTORATE IN STRUCTURAL BIOLOGY

CICLO XXIX

COORDINATORE Prof. Claudio Luchinat

Conformational heterogeneity in multidomain biological systems studied through averaged NMR restraints

Settore Scientifico Disciplinare CHIM/03

Dottorando
Dott. Witold Andrałojć

Tutore
Prof. Giacomo Parigi

Coordinatore
Prof. Claudio Luchinat

Novembre 2013 - 2016

This thesis has been approved by the University of Florence, the University of Frankfurt and the Utrecht University

Universiteit Utrecht

CERM Firenze

FRANKFURT AM MAIN

This page intentionally left blank

Table of Contents

This page intentionally left blank

1 Introduction

1.1 Studying mobility in multidomain proteins – why and how?

Internal mobility is an intrinsic property of biological macromolecules[1–7]. It is present throughout an exceptional range of both spatial and temporal regimes, extending from local, sub-nanosecond backbone vibrations[8], through concerted motions on a variety of intermediate timescales[9,10], up to rearrangements of entire sub-domains or sets of secondary structure elements occurring up to the range of seconds – during, for example, folding/unfolding events[11,12]. Many of these internal motions turn out to be critical for the biological function of the molecule in question[13,14] and thus their localization and extent are under tight evolutionary control[15]. The importance of the function-mobility relationship is well-reflected in the attention attributed, especially in the NMR community, to the development of experimental methods suitable for characterizing the internal motions at a variety of timescales[16–18]. Protein systems composed of multiple quasi-rigid domains connected through flexible linkers constitute a peculiar and interesting type of systems in which the local, fast dynamics around the few degrees of freedom of the linker are coupled to the global and much slower rearrangements of the reciprocal positions of the domains. Such large scale inter-domain motions are often of key importance to the mechanism of action of the given protein, for example through providing it with the ability to adapt its shape to the interaction with a multitude of different substrates[19], regulating the possibility of cooperative binding of a substrate by multiple domains[20], or even allowing the assembly of the binding site itself, if its constituting residues are spread between domains[21].

While being critically important, interdomain motions are also rather difficult to access experimentally.

2 Introduction

The most widespread technique in structural biology, X-ray crystallography, provides information on the state of the macromolecule, when it is introduced into a rigid lattice with translational symmetry. Forcing a system capable of exploring many different shapes into such an arrangement can be difficult (i.e. multidomain proteins usually do not crystallize easily) and even if this is achieved the obtained conformation can be quite different from the one encountered in solution, due to the introduction of additional interactions – the crystal packing forces. A classic example of such a behavior are the first crystal structures of calmodulin[22,23] in which its two domains, now known from solution studies to be highly mobile with respect to one another[24], are present in a single dumbbell-like conformation. To emphasize how misleading such solid state data can be, one can mention that even the four residues now identified to form the disordered linker make part of an α-helical element in the crystalline state. The fact that relatively weak crystal contacts can severely reshape the interdomain free energy landscape clearly illustrates the fragile nature of the latter, whose energetics are mostly governed by weak unspecific interactions (electrostatic, hydrophobic) and conformational entropy[18,25]. This feature is also the reason why multidomain systems are difficult to study theoretically through, for example, molecular dynamics (MD) simulations, as it makes the observed conformational sampling strongly dependent on the fine details of the applied force field. Another factor undermining such studies is the relatively long (likely up to microsecond) timescale of large amplitude domain motions, which require very lengthy runs to achieve convergence (compare for example [26] and [27]). Thus, experimental solution state techniques remain the best tools to glance into interdomain motions in proteins. Among them NMR spectroscopy stands forward for its ability to obtain atomic level information on the studied system. However, in the current context of correlating spin positions between protein domains, classical NOE-based NMR methods would be of very limited use, as in a general conformation there may be no spins close enough to give rise to a interdomain NOE, and even more importantly the system is unlikely to reside in any given conformation long enough for any such possible NOE to build-up. The answer to this shortcoming can be found in other, less commonly used, types NMR observables, which can probe interactions over significantly longer distances or even global features of a spin distribution (such as directions of spin-spin vectors in some fixed frame of reference). Such observables include pseudocontact shifts (PCS) and paramagnetic relaxation enhancements (PRE) which can measure long inter-spin distances (in the case of PCS the distance information is convoluted with orientation dependence), as well as residual dipolar couplings (RDC) which report, among other[28], on bond orientations in the frame of the molecular alignment tensor. One can gain access to all of these highly complementary observables through the introduction of a paramagnetic center into the protein of interest[29,30]. This makes paramagnetic NMR a

wealthy source of information for the study of conformational variability in multidomain biomolecules. Other solution methods like small angle scattering techniques (SAXS, SANS), EPR based methodologies (DEER) or FRET can be, and actually were[31,32], used to complement paramagnetic NMR with orthogonal information and thus yield integrated approaches to interdomain mobility in proteins.

1.2 Data averaging

As the interdomain motions are usually fast on the time scales of the chemical shift differences between conformers, NMR experiments provide a single set of signals resulting from time and ensemble averaging. Depending on the type of observable, the ensemble averaging may simply correspond to taking population weighted means of the properties expected for single sampled conformers - as is the case for PCS and RDC - or it can be more involved, like in the case of PRE where the interdomain motions can influence not only the distance but also the correlation times for the dipolar interaction between the unpaired electron and the reporter nuclear spin.

Whatever the actual rules underlying the motional averaging, one is left with the task of performing an inverse transformation – use the measured mean values to conclude which exact combination of conformers gave rise to these specific means. In any practical case however, the number of independent experimental observations one can collect is significantly smaller than the number of possible distinct conformations that the studied system can assume. The available experimental data is thus clearly insufficient to uniquely determine the set of conformers which generated it. In other words, an infinite number of conformational ensembles, often not having a single structure in common, can reproduce the gathered ensemble averaged data with similar quality and there is no way to distinguish the actually sampled ensemble from the other ones. The spectrum of possible solutions is broadened even further by the presence of experimental uncertainty. Mathematically the problem can be described as ill-posed and ill-determined[33]. Even though, as just described, the information contained in the averaged data is not sufficient to pinpoint the unique, true solution of the ensemble recovery problem, these data still set many constraints on the properties of the plausible conformational ensembles and on the share that specific conformations can have in them.

These observations, together with the biological importance of interdomain motions, as outlined in the previous section, urged multiple groups to seek for methodologies that allow to extract as much information as possible from ensemble averaged observables[27,34–36]. To give a brief overview of the very many methods that arose in the field over the last decade, reference can be made to the classification

proposed in a recent review[37], in which the authors broadly divide the ensemble recovery methods into those based on the Maximum Entropy Principle (MEP) and those applying the Occam's Razor or sparsity assumption (the large weight – LW – methods). The MEP methods search for the ensembles which can reproduce the experimental data while deviating as little as possible from a uniform population of all the sterically available conformations. The LW methods on the other hand aim at reproducing the data with the fewest states possible, taken in high populations.

1.3 The Maximum Occurrence (MaxOcc) method

The specific methodology developed at CERM is called the Maximum Occurrence approach and can formally be classified into the LW group[37]. It however differs from the majority of other methods in the fact that instead of seeking for tentative conformational ensembles, which both fit the experimental data and have some desired property (that is obey an imposed regularizing assumption), it turns its attention towards particular conformations and scores them according to their ability to participate in plausible conformational ensembles. The scoring consists of assigning to each conformation a quantity called its Maximum Occurrence. It is defined as the highest population that the given conformation can have in any conformational ensemble that is in agreement with the experimental data[31]. It therefore constitutes an upper bound that the experimental data impose on the weight of that structure in the real conformational ensemble. The Maximum Occurrence values would converge to the actual weights in the limiting case of sufficiently many independent observables (and no experimental noise). Through a number of numerical simulations it was demonstrated that already the MaxOcc values calculated using the amount of data typically available in real studies, are in most cases able to correctly identify the highly sampled conformations[31,38]. My PhD project mainly consisted in further developing the MaxOcc method and applying it for the study of biological systems.

The idea of Maximum Occurrence (MaxOcc) can actually be generalized from single structures to entire sections of the conformational space, yielding a new quantity: the Maximum Occurrence of conformational Regions (MaxOR). Through direct analogy to MaxOcc, the MaxOR of a given conformational region is defined as the highest *total* population that the structures located within that region can have in any ensemble capable of reproducing the experimental observables. No assumptions are made regarding how the population is partitioned between different conformers within the region. The introduction of MaxOR is not only a theoretical exercise, but indeed it significantly broadens the spectrum of problems that the Maximum Occurrence approach can address. Through the application of

MaxOR one can, for example, quantify the smallest amount of mobility that has to be present in the system in order to explain the experimental observations. Another crucial advantage emerging when moving on from single conformers to entire sections of the conformational space is the possibility of pinpointing regions whose presence is *necessary* for the fitting of the experimental data. For such a region the experiment imposes not only an upper bound on its population (equal to its MaxOR value), but also a lower one. This lower bound can be dubbed the minimum Occurrence of that Region (minOR) and it constitutes an invaluable piece of information for the characterization of the conformational ensemble of a biomolecular system.

1.4 Contributions of the current thesis

The contributions that my thesis brought to the field can be divided into two main branches: 1) the methodological developments related to the Maximum Occurrence approach and ensemble recovery methods in general and 2) the applications of the MaxOcc method to the study of mobility in a variety of biologically important systems.

The first branch included complete rewriting of the original MaxOcc software in a way which exploits the linearity[34] of the ensemble recovery problem when only PCS and RDC data are used. Such reimplementation allowed for an over hundredfold decrease in the required computational time. For the first time a software capable of estimating the MaxOR and minOR values for conformational regions was also developed using an algorithm similar to the new implementation of MaxOcc. Both pieces of software were thoroughly tested on different sets of simulated PCS and RDC data. The detailed description of the new software and the performed simulations can be found in Article 1: 'Exploring regions of conformational space occupied by two-domain proteins'. Still along the lines of the methodological developments, I was involved in a project aiming at comparing the relative usefulness of paramagnetic and diamagnetic NMR observables for ensembles recovery applications. This work gave rise to Article 2: 'Information content of long-range NMR data for the characterization of conformational heterogeneity'. My contribution to this work consisted of performing the computational analysis with both the MaxOcc and SES (Sparse Ensemble Selection)[34] approaches on simulated paramagnetic and diamagnetic data, including the preparation of the simulated observables and performing all calculations needed for the recovery of the main sampled states.

The second branch of my activities, consisting of the study of biological questions through the MaxOcc approach, included the following: 1) the MaxOcc analysis of calmodulin interacting with the DAKp

peptide, included in Article 1; 2) the study of the interhelical mobility in the HIV1TAR using diamagnetic RDC, as described in Article 3: ‘Inter-helical conformational preferences of HIV-1 TAR-RNA from maximum occurrence analysis of NMR data and molecular dynamics simulations’; and 3) the MaxOcc study of the encounter complex between cytochrome P450cam and putidaredoxin presented in Article 4: ‘Identification of productive and futile encounters in an electron transfer protein complex’. In all of these works my contribution concentrated on the data analysis with the MaxOcc approach. This included the design of the exact procedure taken during the analysis which had to be adjusted each time based on the properties of the available data and the information sought. Each of these projects also required specific additions and modifications to the MaxOcc software itself. Developing them was also part of my contributions to the listed papers.

Finally I am also a co-author of the review Article 5: ‘How to tackle protein structural data from solution and solid state: An integrated approach’. Besides contributing to the writing of the review, and of the paragraphs related to the data analysis with the MaxOcc approach in particular, I performed the simulations of the averaging of RDC data, presented in Fig 4 and Table 1 of this paper.

2 Methodological aspects

The following sections are intended to give a unified description of the procedure followed during the MaxOcc (and MaxOR/minOR) analysis. The preliminary steps of the process (generation of the pool of conformers and back-prediction of the experimental observables) are usually mentioned just briefly in the respective papers and thus they are discussed more in-depth here, aiming at highlighting all the decisions that have to be made at each point. In the later sections (the algorithm used, determination of MaxOcc/MaxOR) a more general description of the other steps is provided (for details the reader can refer to the specific sections in my published manuscripts), and focus is made on a systematic overview of the differences between the exact implementations of MaxOcc used in the different publications. The final section intends to draft a general flowchart of the decisions made during the later steps of the MaxOcc analysis and to show how the partial results obtained in the course of the study are interactively used to shape the subsequent rounds of calculations.

2.1 Generation of a broad pool of sterically allowed conformers

Even though the exact course of the MaxOcc (and/or MaxOR) analysis depends on the specific system under study and on the type of information one seeks to extract, there exist two invariant preliminary steps common to every application of MaxOcc. These are: 1) the generation of a wide pool of rigid conformers of the studied system, in which all the sterically allowed conformational space is sampled with some desired resolution (understood as the distance between neighboring structures) and 2) the

prediction of the values of the experimental observables that would have been measured for each rigid structure in the generated pool. An exhaustive sampling of the entire conformational space is of primary importance for the correct determination of MaxOcc values as the absence of specific type of conformations in the pool can in principle lead to an underestimation of the MaxOcc of all the other structures. This section aims at presenting the available methods for the generation of a broad, un-biased pool of conformers. The most straightforward approach would be of course to construct the pool through a systematical variation of all the degrees of freedom present in the system (e. g. the dihedral angles in the linker residues of a two domain protein) with some constant step size. Such an approach was indeed used in the MaxOcc study of the HIV1TAR element (Article 3) in which the reciprocal position of the two rigid elements can be described by just three variables, for example the inter-helical Euler angles. This kind of method however quickly becomes impractical with the increasing number of degrees of freedom, as the amount of conformers to be generated becomes prohibitively high. One thus have to resort to a more statistical approach. For two domain proteins one can perform a random sampling of the combinations of the mobile dihedral angles using for example the program RanCh[39]. This software generates random conformations of the mobile linker, places the two domains at its ends and then evaluates whether the obtained conformation is sterically acceptable (no overlap between the domains, no knots on the linker etc.). It can also restrict the sampling to specific, predefined combinations of angles, allowing for example to sample only the (generously) allowed regions of the Ramachandran plot. Already in the case of calmodulin, with just four mobile residues, such an approach has proven indispensable and thus both my works involving calmodulin (Articles 1 and 2) made use of a pool generated using RanCh. One can note that the values of the experimental observables expected for the moving domain actually depend exclusively on the reciprocal positions and orientations of the two domains and thus only indirectly on the exact values of the dihedral angles in the linker. It may thus seem beneficial to construct the pool by sampling (systematically or randomly, depending on the size of the system) directly the different positions and orientations of one domain with respect to the other, with the added advantage that in such case the sampled space is always just six dimensional (with three coordinates describing the position of the center of mass of the domain and other three its orientation through for example Euler angles). The disadvantage of such a method for practical applications arises from a much more difficult procedure for evaluating whether a specific conformation is sterically allowed. In fact, such task would require for example the usage of a loop reconstruction software (like [40]) in order to assess for each conformation whether a physically correct conformation of the linker can be found to connect the two domains. Still the coordinate systems based on the six-dimensional rotation-

translation space remain the most convenient for the later stages of the calculations (verification of the actual resolution and completeness of the pool, definition of conformational regions etc.) and thus even the pools generated in the dihedral angle space are subsequently mapped into such coordinates.

Systems without physical linkers between the domains, like protein-protein complexes, are on the other hand perfectly suited for the pool generation directly in the inter-protein rotation-translation space. Here the only steric restriction – apart from the lack of overlap – that one can impose is the requirement for the two proteins to be in contact with each other. This is however easily satisfied by translating, for each generated conformer, one of the partners along the line connecting the centers of masses of the two proteins until a contact is achieved[41]. On the other hand, the virtual lack of restrictions on the reciprocal positions and orientations of the two proteins immensely broadens the dimensions of the conformational space to be sampled. Already for two proteins of moderate sizes, sampling of the entire conformational space with an average resolution (distance between neighboring structures) of the order of a few Angstrom of translation and a dozen degrees of rotation would require the generation of not just thousands of conformations (as is usually the case for two domain proteins) but rather of over a million of them. The usage of pools of such sizes for MaxOcc analysis is impossible with the computational resources available nowadays. This problem was encountered during the MaxOcc analysis of the cytP450cam complex with Pdx (Article 4). Luckily the specific properties of the available experimental data – PCS and RDC which are almost not averaged and thus compatible only with a very limited set of structures; minor conformations sensed only by the PRE which decay to 0 for all conformers far enough from the paramagnetic center – allowed us to disregard considerable parts of the conformational space as not even possibly capable of contributing to the fit of the experimental data. The exact procedure devised for the identification of the possibly important regions of the conformational space and for the generation of the pool is described in detail in the SI of Article 4.

2.2 Experimental restraints and their back-prediction

The Maximum Occurrence approach relies on comparing the values of experimental observables expected for specific rigid structures with the actually measured ensemble averages. A reliable prediction of the former is thus of uttermost importance for the success of the method. The following paragraphs introduce the main types of experimental observables used so far in the MaxOcc analysis and outline the quantities that have to be known and/or assumptions that have to be made in order to predict their values for arbitrary rigid structures. Throughout this discussion the structures of single protein domains are

considered known and rigid, unless stated otherwise.

2.2.1 Pseudocontact shifts (PCS)

General properties

The pseudocontact shift is the contribution to the chemical shift of a given nuclear spin originating from its dipolar interaction with the time average of the unpaired electron's magnetic moment $\langle\mu\rangle$ (the so called the Curie spin)[42]. Under isotropic reorientation the dipolar interaction between two magnetic moments of constant magnitude is known to average to 0. Thus PCS can be observed in an isotropic solution only in the cases in which the magnitude of the Curie spin becomes orientation dependent due to a non-zero orbital angular momentum of the unpaired electron and its interactions with the spin angular momentum (spin-orbit coupling). This orientation dependence of $\langle\mu\rangle$ manifests itself experimentally in the form of the anisotropy of the molecular magnetic susceptibility $\boldsymbol{\chi}$ of the system, which becomes a tensorial quantity[43]. The magnitude of the anisotropy and the orientation of the PAF (principle axis frame) of the magnetic susceptibility tensor with respect to the molecular frame of reference are the main quantities which have to be known in order to predict the PCS experienced by an arbitrary nuclear spin. The PCS (in ppm) is then given by[43]:

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

where θ and ϕ are the polar angles of the metal-proton vector in the PAF of $\boldsymbol{\chi}$, r the metal-proton distance and $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic components of the anisotropy of $\boldsymbol{\chi}$ defined as:

$$\Delta\chi_{ax} = \chi_{zz} - \frac{\chi_{xx}+\chi_{yy}}{2} \quad (2)$$

$$\Delta\chi_{rh} = \chi_{xx} - \chi_{yy} \quad (3)$$

where χ_{xx} , χ_{yy} and χ_{zz} are the principal components of $\boldsymbol{\chi}$.

Equation 1 assumes that the electron is fully localized on the metal ion (point dipole approximation) and neglects the influence of molecular tumbling anisotropy induced by $\boldsymbol{\chi}$ (see below). For lanthanides the first approximation should have a negligible effect for all nuclei far enough from the metal not to be rendered unobservable by an excessive paramagnetic relaxation enhancement. The second effect can be evaluated to modify the predicted PCS value by at most a fraction of a percent even for the most anisotropic metals at the highest magnetic fields accessible today.

Application to MaxOcc

In order to apply the PCS to the study of conformational heterogeneity in a multidomain protein,

obtaining a reliable estimate of χ is the prime requirement. Despite the recent advances in the *ab initio* prediction of the magnetic susceptibility anisotropy parameters[44], an experimental approach developed in CERM over a decade ago[45] remains the best established method for the determination of the χ tensor for the current type of applications. The methodology requires a rigid attachment of the paramagnetic center to one of the domains of the system under study. The nuclei of this domain then provide a firm network of spins with fixed (and known) positions with respect to the paramagnetic metal. The values of PCS experienced by these nuclei allow one to determine the parameters defining the χ tensor as well as to refine the metal position through a fitting procedure with eight degrees of freedom (axial and rhombic anisotropy, three variables defining the PAF of the tensor – e. g. Euler angles – and the three Cartesian coordinates of the metal). The rigidity of the paramagnetic metal position with respect to the protein framework is of uttermost importance for a robust determination of the χ tensor parameters. As a consequence, metal substitution into existing binding sites or the use paramagnetic tags with two attachment sites (for example tags from the CLaNP series[46,47] or LBPs fused into internal loops[48,49]) is strongly preferable over their single attachment counterparts for our kind of applications. The lack of rigidity of the metal attachment can usually be spotted rather easily as it becomes evident in a variety of ways, including: a poor quality of the PCS fit to a tensor (no 'average tensor' exists for motionally averaged PCS), large discrepancies between the tensor parameters determined using PCS and RDC (the same motions average PCS and RDC in different manners[50]), unrealistic metal position and/or anisotropy values much smaller than expected for the given metal.

Once the χ tensor has been determined it can be used to predict the PCS values experienced by the nuclei of the domain not containing the paramagnetic metal ion (called the 'moving' domain) in any single rigid conformation of the system.

The PCS were among the observables most frequently used during my PhD work and they were used in Articles 1, 2 and 4.

2.2.2 Residual dipolar couplings (RDC)

General properties

As already noted in the previous paragraph the dipolar interaction between two spins of constant magnitudes (like the nuclear spins) averages to 0 under isotropic tumbling in solution. If however, for whatever reason, the chemical entity harboring the two spins does not assume all of its orientations with equal probabilities a non-zero dipolar coupling can be measured. Its sign and magnitude bear information

on the (average) orientation of the spin pair in the frame of the molecular alignment tensor $\boldsymbol{A}$. The alignment tensor, which describes the anisotropy of the orientational sampling, is a real, symmetric tensor with trace equal to 0. Its principal components indicate whether the ensemble of orientations present in solution is enriched in (positive principal component) or depleted of (negative principal component) molecules in which the given axis of $\boldsymbol{A}$ points towards the direction of the magnetic field. The elements of $\boldsymbol{A}$ (expressed in an arbitrary reference frame) are defined as[43,51]:

$$\boldsymbol{A}_{pq} = \langle \cos\theta_p \cos\theta_q - \frac{1}{3}\delta_{pq} \rangle \quad (4)$$

where θ_i is the angle which the i[th] axis of the arbitrary frame makes with the external magnetic field and $\delta_{pq} = 1$ when $p = q$ and $\delta_{pq} = 0$ otherwise.

For a rigid specie the expected residual dipolar coupling between a pair of spins A-B can be calculated according to the following relationship (written in the PAF of $\boldsymbol{A}$):

$$RDC\ [Hz] = -\frac{\mu_0 \gamma_A \gamma_B \hbar}{8\pi^2 r_{AB}^3} \times \left[\Delta A_{ax} (3\cos^2\theta - 1) + \frac{3}{2} \Delta A_{rh} \sin^2\theta \cos 2\phi \right] \quad (5)$$

where μ_0 is the magnetic permeability of the vacuum, γ_A and γ_B are the magnetogyric ratios of the spins A and B, $\hbar$ is the reduced Planck constant, r_{AB} the interspin distance in the A-B pair, ΔA_{ax} and ΔA_{rh} are the axial and rhombic anisotropies of $\boldsymbol{A}$ defined by relations analogical to eq. (2) and eq. (3), while θ and ϕ are the polar angles defining the orientation of the A-B vector in the PAF of $\boldsymbol{A}$.

Equation 5 is general and valid for any source of molecular alignment. In the specific case in which the entire anisotropy of the molecular tumbling can be attributed to the presence of a metal ion with an anisotropic magnetic susceptibility $\boldsymbol{\chi}$ the alignment tensor can be directly derived from the latter through the relationship[51]:

$$\boldsymbol{A} = \frac{B_0^2}{15\mu_0 kT} \boldsymbol{\chi} \quad (6)$$

in which k is the Boltzmann constant, T the absolute temperature and $B_0$the magnetic field strength.

Application to MaxOcc

Quite interestingly the amount of information regarding the inter-domain rearrangements carried by the measured RDC, the methods used to back-predict their values and even the solidity of the theoretical framework underlying their analysis differ rather substantially, depending on the source of the molecular alignment used. The RDC will be thus discussed in separate paragraphs depending on their origin (paramagnetic or diamagnetic) with the scope of highlighting their similarities and differences.

Paramagnetic RDC (pRDC)

When a paramagnetic metal ion (with an anisotropic $\boldsymbol{\chi}$), is rigidly attached to a multidomain protein it causes the domain which bares it to partially align in the magnetic field, inducing measurable RDC between its nuclei, according to eq. (5) and (6). The alignment of one part of the protein is naturally passed over to its integrity, yet the presence of internal degrees of freedom between the metal attachment site and a given spin pair attenuates the degree of alignment of the latter. Thus, the set of RDC measured in any domain other than the paramagnetically tagged one can still be described by eq. (5), but with an alignment tensor $\boldsymbol{A}_{mobile}$ different (and smaller) than the one dictated by eq. (6) (call it $\boldsymbol{A}_{rigid}$)[52]. The elements of $\boldsymbol{A}_{mobile}$ correspond to (population weighted) averages of the elements of $\boldsymbol{A}_{rigid}$ expressed in the frame of the moving domain over the different sampled conformations. It is actually in the values of the attenuated RDC from the untagged domain (and thus in the properties of $\boldsymbol{A}_{mobile}$) that the information about the inter-domain mobility is encoded. This by no means signifies the lack of use for the RDC measured in the rigid domain. Indeed, through a procedure similar to the one outlined for the PCS these RDC can be used to obtain a direct experimental measure of $\boldsymbol{\chi}$ (or equivalently $\boldsymbol{A}_{rigid}$), which then serves to predict the RDC expected for the moving domain in any possible conformation of the macromolecule. The direct experimental access to $\boldsymbol{A}_{rigid}$ relieves the analysis of paramagnetic RDC from the need of any theoretical modelling of the molecular alignment, thus making the procedure fully self-contained and highly robust. Moreover, PCS and pRDC are usually measured together and both arise from the presence of the same $\boldsymbol{\chi}$ anisotropy. They can be thus used together to estimate the elements of $\boldsymbol{\chi}$ with an even better robustness (when the two types of observables are combined a Lipari-Szabo order parameter[53] is usually introduced into eq. (5) to account for some averaging of the RDC by the local intra-domain mobility).

The information that the paramagnetic RDC provide on the respective positions of the 'rigid' and 'moving' domains is only limited to their relative orientation. Thus all the conformers in which the inter-domain orientation is conserved give rise to degenerate pRDC values regardless the respective Cartesian position of the two domains. This is because the RDCs arising between the nuclei of the 'moving' domain in any rigid conformer of the protein depend solely on the orientation of the respective spin-spin vector in the reference frame of $\boldsymbol{A}_{rigid}$ (eq. (5)), which coincides with the reference frame of the 'rigid' domain. Paramagnetic RDC were widely used across most of my research projects, in each case in conjunction with PCS measures for the same sample. Their usage is reported in Articles 1, 2 and 4.

14 Methodological aspects

External alignment diamagnetic RDC (dRDC)

When a macromolecule is introduced into a dilute liquid crystalline phase (an alignment medium) it begins to tumble partially anisotropically, through the mechanisms of steric exclusion and unspecific electrostatic interactions (if the alignment medium is charged)[54]. The alignment properties of a given molecule are thus dictated by its shape and charge distribution. Both these properties change in the course of intramolecular interdomain rearrangements and so does the magnitude and direction of the alignment tensor. Consequently, in the case of external alignment RDC none of the domains is fixed in the frame of the alignment tensor and all the dRDC are subject to motional averaging. It the absence of experimental information about the parameters of the alignment tensor (which for pRDC was provided by the 'rigid' domain), the prediction of dRDC for any specific rigid conformation have to rely on theoretical modeling of the molecular alignment. This is usually achieved through the usage of the PALES[55] or PATI[56] software. Even though the dRDC values predicted using these programs for rigid structures are generally quite reliable and accurate, one can express doubts whether the average dRDC measured for a flexible entity can actually be viewed as simple averages over the dRDC calculated this way for specific sampled conformers. When PALES predicts the steric alignment of a protein conformation it generates a set of orientations of that structure and excludes those that show steric overlap with a model of the aligning entity (e. g. Pf1 phagi are modelled as rods, bicelles as discs). This process is repeated while moving the two objects closer and closer together, increasing the number of orientations that become prohibited. When the process is finished the alignment tensor can be calculated by applying ex. (4) to the set of retained conformations. The applicability of this simple treatment to a macromolecule exhibiting large scale flexibility may be questionable. When a flexible macromolecule (in some conformation) approaches the excluded volume of the aligning entity it can avoid the clash not only by the change of orientation but also by the change of shape. This possibility rises doubts not only regarding the usage of the PALES predicted dRDC for such systems, but also regarding a more fundamental question of the influence of the aligning medium on the interdomain conformational sampling itself. The mechanism of avoiding collisions through the change of shape may result in the enrichment of the immediate vicinities of the excluded volume in the most compact (closest to spherical shape) conformations of the studied macromolecule. The practical significance of the mentioned effects is difficult to evaluate and thus their existence is usually neglected. One however have to remember that the usage of dRDC for the study of interdomain mobility in macromolecules thus relies on several assumptions whose correctness was not so far objectively verified.

Assuming the correctness of the theoretical framework usually employed to analyze dRDC in

multidomain systems, these observables are richer in information then their paramagnetic counterparts (for which in turn the analysis is founded on a more solid basis of the experimentally determined alignment tensor). The dRDC are sensitive to both the position and orientation of the two domains. Even those rotations of single domains that do not noticeably modify the shape of the macromolecule (and thus its alignment properties) can have a pronounced effect of the dRDC as they still change the orientations of the spin-spin vectors in the frame of the alignment tensor. Moreover in case of dRDC the observations made in both the domains are reporters of the inter-domain motions and thus a set of external alignment dRDC provides twice as many restraints as a set of pRDC.

One more practical consideration to keep in mind when using dRDC for MacOcc analysis is that the experimental parameters dictating the magnitude of the measured dRDC (the exact concentration of the alignment medium and the degree of its own alignment in the magnetic field) are often difficult to determine precisely enough[34]. The values of these parameters assumed during the PALES prediction will affect the magnitude of the back-calculated alignment tensor (but not its orientation or rhombicity). Thus to correct for the possible inaccuracies in their determination (leading to an incorrect magnitude of the back-predicted dRDC) one has to introduce an additional degree of freedom during the ensemble optimization in MaxOcc: an optimizable scaling factor between the experimental dRDC values on one side and the population weighted averages of the back-predicted data on the other.

In the course of my PhD the diamagnetic RDCs were applied in two distinct projects which led to the publication of Articles 2 and 3.

Special case: alignment properties dictated by a single domain

When one of the domains of a biological macromolecule is significantly larger, more elongated and/or much more charged than the other one, it can, to a good approximation, fully dictate the shape and/or charge distribution of the molecule as a whole, regardless the actual inter-domain conformation[57]. If the external alignment dRDC are measured for such a system their properties become very similar to those discussed for pRDC[58]. The alignment tensor practically does not change between conformers and it is anchored to the large domain. As a consequence, the dRDC measured in that domain are not averaged by inter-domain motions and can be used to experimentally determine the properties of the alignment tensor. Only the dRDC induced in the smaller domain report on the inter-domain rearrangements and the information is once again of purely orientational character, as was the case for pRDC.

Diamagnetic RDCs measured for such systems were used in Article 3. However, the work also made use of additional sets of dRDC originating from constructs in which the two domains were actually of similar sizes. Thus, for the sake of consistency, all the data were treated using a single approach, involving the

usage of PALES for the back-prediction of the dRDC. For the constructs with large disproportions in the domain sizes, the alignment tensors predicted with both formalisms (PALES and a pRDC-like approach) were actually very similar.

2.2.3 Paramagnetic relaxation enhancements (PRE)

General properties

The interactions of the nuclear spin with the unpaired electron manifest themsevles also in the form of contributions to the longitudinal and transverse nuclear relaxation rates – the paramagnetic relaxation enhancements[42]. For nuclei located in a different protein domain than the paramagnetic center, the PRE is caused practically exclusively by the dipolar interactions with the unpaired electron (Solomon term) and its time average (Curie term). The magnitude of these two contributions to the nuclear transverse relaxation rate in a rigid system can be calculated according to the following equations[42,59]:

$$R_{2,para}^{Solomon} = \frac{\gamma_A^2 g_e^2 \mu_B^2 \mu_0^2 S(S+1)}{240\pi^2 r^6}\left(4\tau_{Solo} + \frac{3\tau_{Solo}}{1+\omega_A^2\tau_{Solo}^2} + \frac{13\tau_{Solo}}{1+\omega_e^2\tau_{Solo}^2}\right) (7)$$

$$R_{2,para}^{Curie} = \frac{\gamma_A^2 B_0^2 g_e^4 \mu_B^4 \mu_0^2 S^2(S+1)^2}{720\pi^2 k^2 T^2 r^6}\left(4\tau_{Curie} + \frac{3\tau_{Curie}}{1+\omega_A^2\tau_{Curie}^2}\right) (8)$$

where, $\omega_A$and $\omega_e$are the Larmor frequencies of the nucleus and the unpaired electron respectively, $g_e$is the free electron g factor, S the electron spin quantum number, while τ_{Solo} and τ_{Curie} are the correlation times governing the Solomon and Curie interactions respectively. All the other symbols have the same meaning as in the previous equations. In the case of lanthanide ions in which the spin orbit coupling effect is very strong, S should be substituted with a different quantum number J, which is a vector combination of S with the electron orbital quantum number L[42]. The Solomon and Curie contributions to PRE are in general governed by two distinct correlation times because the sets of processes modulating the magnetic field at the position of the nucleus related to these two terms are different. The direct dipolar interaction (giving rise to the Solomon term) is modulated by transitions between the spin levels of the electron (occurring at the timescales of τ_e – the correlation time of electronic relaxation), the rotational reorientation of the molecular entity in question (described by a correlation time τ_r) and by any possible exchange processes (with a correlation time τ_M). Thus, the correlation time τ_{Solo} will be dictated by the fastest of the abovementioned processes or when several of them occur at similar timescales it will be given by the following expression:

$$\frac{1}{\tau_{Solo}} = \frac{1}{\tau_e} + \frac{1}{\tau_r} + \frac{1}{\tau_M} (9)$$

On the other hand the dipolar interaction with the time average of the electron spin (the Curie term)

clearly cannot depend on the rate of electronic transitions and thus τ_e must not be considered when evaluating τ_{Curie}, which thus is given by:

$$\frac{1}{\tau_{Curie}} = \frac{1}{\tau_r} + \frac{1}{\tau_M} \quad (10)$$

Application to MaxOcc

In a system with internal mobility, the latter can influence the observed paramagnetic relaxation enhancements (PRE) in several ways. First, the electron-nucleus distance will in general vary between the different sampled conformers and thus the single value of r^{-6} in equations (7) and (8) should be substituted with its population weighted average. Second, the internal dynamics constitute an additional process modulating the strength of the electron magnetic field sensed by the nucleus and as such, if fast enough, they may influence the values of the correlation times τ_{Solo} and τ_{Curie}. Finally, in the presence of any internal dynamics slow with respect to the nuclear transverse relaxation rate, the different conformations may provide magnetization decays with different time constants, which sum up to a decay that may be different from that corresponding to the average time constant. Thus, only in the cases in which all the inter-domain dynamics occur on the time scales intermediate between the two extremities just mentioned (that is they are slow with respect to molecular rotation and fast with respect to the nuclear relaxation), the measured PRE correspond to population weighted averages of the values calculated using eq. (7) and (8) for single sampled conformers.

For the two domain proteins whose sizes make them amenable to paramagnetic NMR studies the integrity of inter-domain dynamics is expected to occur at microsecond or faster timescales and consequently even the states with PREs different by several thousands of s^{-1} can be considered to be in fast exchange. The effects that slow motions can have on the PRE are therefore not important for the current studies and will not be discussed here any further.

On the other hand, the impact of the fast dynamics on the autocorrelation function for paramagnetic relaxation deserves more attention here, as at least some part of the interdomain motions can occur on timescales faster than those of molecular tumbling or electron relaxation. Under the assumption that the overall tumbling (or electron relaxation) and the interdomain motions occur on sufficiently different timescales and that they are independent of each other one can treat the paramagnetic relaxation in the moving domain according to the extended model free formalism[60]. The effect of the fast local motions occurring within the domain itself can safely be disregarded as the electron-nucleus vectors are usually long and so restricted dynamics within a given interdomain conformation cannot alter them significantly

(the Lipari-Szabo order parameters for such type of motions S_l are close to 1). What thus remains to be considered are the large scale interdomain motions occurring on a timescale τ_d, intermediate between the fast local dynamics and the overall tumbling of the protein (or the electron relaxation time). Such motions are in general restricted in space and cannot lead to the total decay of the autocorrelation function. We thus define for them an order parameter S_d. With the interdomain motions described in terms of τ_d and S_d the expected ensemble averaged PRE can be calculated as:

$$R_{2,para}^{Solomon} = \langle r^{-6} \rangle [S_d^2 f(\tau_{Solo}) + (1 - S_d^2) f(\tau_d)] \quad (11)$$

with:

$$f(\tau) = \frac{\gamma_A^2 g_e^2 \mu_B^2 \mu_0^2 S(S+1)}{240\pi^2} \left(4\tau + \frac{3\tau}{1+\omega_A^2 \tau^2} + \frac{13\tau}{1+\omega_e^2 \tau^2} \right) \quad (12)$$

The presented equation is written for the Solomon contribution. The corresponding relationship for the Curie term can be straightforwardly obtained by comparison with eq. (7) and (8). It can be noted that when the current formalism is used the PRE values assigned to each single conformer during the MaxOcc analysis are calculated according to eq. (11) taken with a single value of r and thus no longer correspond to the PRE values that would be measured for rigid structures (eq. (8)).

In order to use eq. (11) one needs a way to determine both τ_d and S_d. The correlation time τ_d can be estimated through the analysis of amide relaxation data[61]. The determination of S_d on the other hand is far less obvious. Given the set of sampled conformers S_d for each electron-nucleus vector could be calculated using for example the N-state discrete jump model[62]. However, the set of conformers sampled by the system is exactly the unknown parameter that we would like to study through MaxOcc! Moreover, it may happen that only a fraction of interdomain rearrangements occur of timescales faster than τ_{Solo} (or τ_{Curie}) and thus only some unknown subset of the sampled conformations would have to be considered for the calculation of S_d. Given these difficulties we chose to treat S_d as an optimizable parameter (taken equal for all the electron-nucleus vectors for simplicity) and repeat the MaxOcc calculations for different values of S_d. The S_d value giving the highest MaxOcc value for a given conformer would be then considered optimal as the true MaxOcc value has to be the highest one that can be obtained irrespective on the specific parameters one uses to model the data.

During my PhD research the PREs were used in the work described in Article 4. In this specific case, the simple approach using a correlation function based entirely on the overall rotational reorientation was applied, as – for the reasons discussed in the publication itself – it was expected to give rise to the most conservative MaxOcc (and MaxOR/minOR) values.

2.3 The ensemble optimization algorithm

As already mentioned in the Introduction, the MaxOcc of a conformation is defined as the highest weight (population) that this structure can assume in a conformational ensemble without unavoidably leading to a misfit of the experimental observations. For the practical determination of MaxOcc values one thus needs a method of identifying, among all the ensembles containing the evaluated conformer at some fixed population x_{eval}, the ones most capable of reproducing the experimental data. That means finding the combination of other structures (at arbitrary populations) best suited to complement the selected conformer when it is present for the x_{eval} fraction of time. The particular method used for such ensemble optimization can vary depending on the implementation of the MaxOcc procedure. The specific implementation developed as part of my PhD project and used in all of my published work (with some variations; see below) exploits the fact that, when only the PCS and paramagnetic RDC data are used, the ensemble recovery problem becomes linear (ensemble averaged data are linear combinations of the observables predicted for rigid conformers)[34]. This new implementation, using a Non-Negative Least Squares algorithm, is described in detail in Article 1 main text and SI. Its introduction led to an over hundredfold decrease of the computational time, as compared to the previous version of the MaxOcc software, making possible the evaluation of the maximum occurrence values for entire pools of conformations. Along the course of my PhD work, new features were progressively added to the program in order to adapt it to the requirements of the subsequent projects. As an example, the MaxOcc study of HIV1TAR (Article 3), which employed four sets of diamagnetic RDC, required upgrading the program to be able to deal with cases in which the absolute magnitude of the back-calculated data is uncertain (for the origin of this uncertainty when using dRDC see section 2.2). Allowing for a changeable magnitude of the back-predicted data during the ensemble optimization would break the linearity of problem and thus an iterative procedure was implemented in which the composition of the ensemble and the magnitude of the data are optimized in separate steps which are repeated until both parameters converge. This procedure is described in detail in Article 3 and its SI. This generalization of the algorithm allows the new linear formulation of MaxOcc to be applied also to other types of experimental observables for which the scaling between the observed and back-calculated data is unknown, like for example the SAXS profiles. Yet another variation of the ensemble optimization procedure was devised and applied in Article 4 (see the Materials and methods section of that work for a detailed description), allowing to determine the MaxOcc values faster in cases when they are expected to span a very broad dynamic range across the pool (from a small fraction of a percent to almost 100%).

The algorithms used in the different projects to calculate the Maximum and minimum Occurrences of

Regions (MaxOR and minOR values) are always straightforward extensions of the corresponding procedures devised to calculate MaxOcc, described in the previous paragraph. The main conceptual difference with respect to MaxOcc – in which during each ensemble optimization the weight of a single conformer was kept fixed ($x_{eval} = const$) – is that in MaxOR it is only the total population of the evaluated region that is restrained ($\sum_{i \in eval} x_i = const$). The repartition of this fixed total population between different structures within the region is not subjected to any constraint.

2.4 Determination of the MaxOcc of a conformer (and MaxOR of a conformational region)

In order to calculate the maximum occurrence of a specific conformation a series of ensemble optimizations are performed (as described in the previous section) in which the fixed weight of the evaluated conformer (x_{eval}) is gradually increased by a constant increment (say 5%) starting from 0. The quality of the reproduction of the experimental data achieved with $x_{eval} = 0$ should be as good as the one obtained during an unrestrained fit. This is because effects of the exclusion of a single conformer (a point in the conformational space) can always be compensated for by populating its closest neighbors (in the ideal case the structures infinitesimally close to it). Obtaining a different result would actually point towards an under-sampling of the part of the conformational space containing the evaluated conformer. As higher and higher contributions of the evaluated conformer to the sampled ensemble are imposed (increasing x_{eval}), the quality of the fit of the experimental data begins to deteriorate. When the value of x_{eval} is reached for which the experimental observations can no longer be reproduced (the errors in their fit surpass a certain threshold) it is a sign that the imposed x_{eval} have surpassed the maximum occurrence of the evaluated conformer. The exact MaxOcc can be calculated through linear interpolation between the last two values of x_{eval} (just before and just after passing the threshold).

For NMR observables like PCS and RDC the actual values of the experimental errors may be difficult to estimate and moreover apart from the random noise they may also contain systematical contributions originating from slight differences in sample conditions between the diamagnetic and paramagnetic samples. For this reason, the position of the threshold defining a good quality of the fit becomes somewhat arbitrary. In all my work the following approach was used to determine a reasonable 'threshold of misfitting' of the experimental data. The quality of the fit of the each type of experimental observables is followed by the use of Q-factors[63]:

$$Q = \sqrt{\frac{\sum_i^N (\delta_{e,i} - \delta_{p,i})^2}{\sum_i^N \delta_{e,i}^2}} \quad (13)$$

where, N is the number of measured restraints of the given type and $\boldsymbol{\delta_e}$ and $\boldsymbol{\delta_p}$ are the sets of experimental and back-predicted observables respectively. The Q-factors describing the quality of the fit of different types of data are added together to yield a single Target Function TF (weighting factors can later be introduced to ensure comparable contributions of the different typess of data to TF). An unrestrained fit (without imposing any value to x_{eval}) of all the experimental data is performed to determine the lowest value of the TF achievable for the current set of experimental data (TF_{min}). This lowest value of TF is then used as a reference when setting the threshold above which the experimental data is considered misfitted (the threshold TF_{thres} is usually set equal to 120%-130% of TF_{min}). It is thus the quality of the best obtainable fit that is used to judge the actual extent of uncertainty present in the measured data. Figure 1a presents the concepts just introduced in a pictorial way.

The determination of the MaxOR and minOR values of a conformational region proceeds along very similar lines. The ensemble optimization is this time performed for different fixed total weights of the structures in the region ($\sum_{i \in eval} x_i = const$). The calculation uses the same target function TF and threshold of misfitting TF_{thres} as were described above for MaxOcc. An example of theis kind of calculations is depicted in Figure 1b. An important novelty when passing from MaxOcc of single conformations to MaxOR of conformational regions is the fact that the quality of the fitting may decrease (the TF may increase) also when $\sum_{i \in eval} x_i$ is taken too low. If this increase of the TF for the lowest weights of the region is significant enough to surpass TF_{thres} then it becomes clear that sampling the considered region is actually indispensable and the $\sum_{i \in eval} x_i$ value at which the threshold is reached defines the minimum occurrence of this region (minOR).

This kind of behavior will occur when the examined region contains all the conformers having some specific characteristic in the data space, indispensable to obtain the correct averaged data. One obvious type of such regions could consist of all the conformers located in the vicinity of a paramagnetic probe for which high PRE values were measured – excluding all these structures (setting $\sum_{i \in eval} x_i = 0$) would make reproducing the related set of PRE impossible. Predicting which conformational regions may have a minOR > 0 when only PCS and RDC data are used is less straightforward, yet in some cases compact regions with this property can be found (see Article 1, the calmodulin bound to DAPk peptide study).

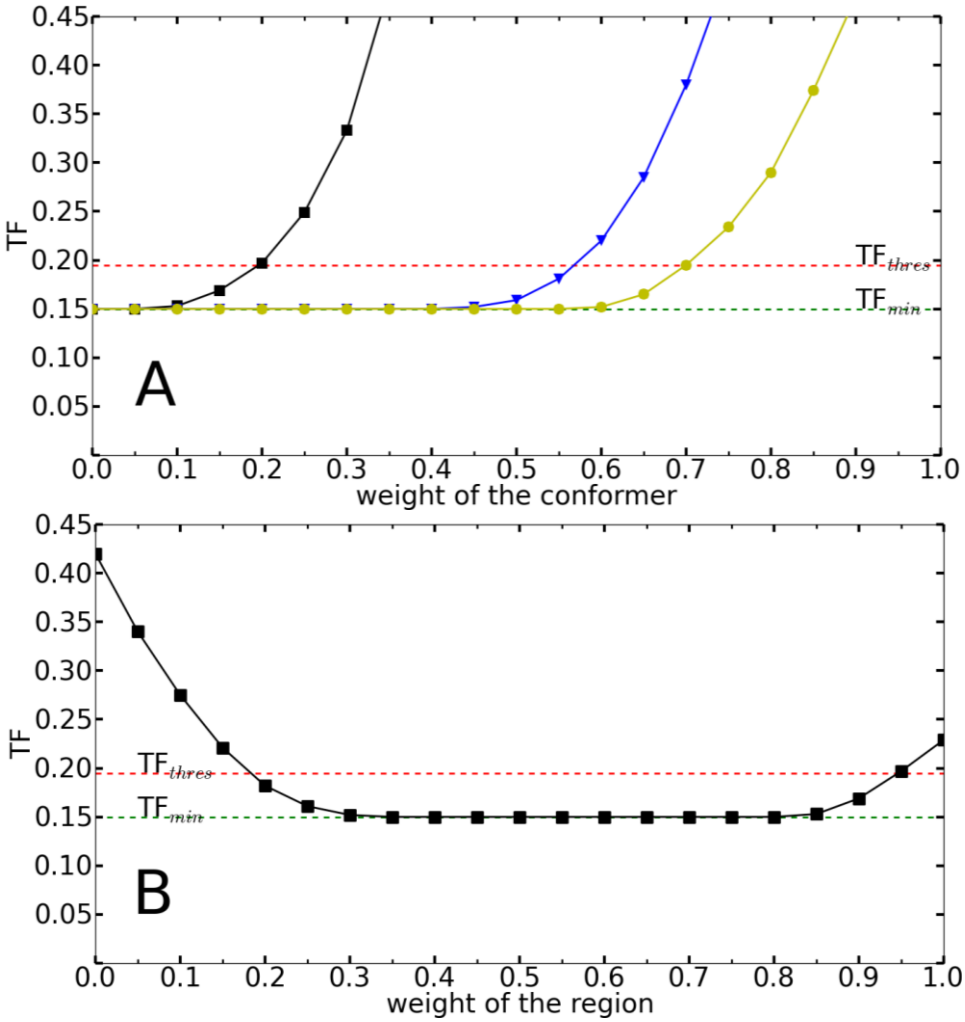


Figure 1. A graphical representation of the calculation of MaxOcc (panel A) and MaxOR/minOR (panel B). Both panels show the TF_{min} and TF_{thres}values as green and red dashed lines respectively. In panel A the *TF* versus x_{eval} curves are traced for three different conformers (black, blue and yellow curves). The MaxOcc value of each conformer corresponds to the x_{eval} value at which the respective curve crosses the red dashed line of TF_{thres}(it is thus around 20% for the black conformation, 57% for the blue and 70% for the yellow one). In panel B the *TF* versus $\sum_{i \in eval} x_i$ curve is presented for one conformational region. The MaxOR of this region corresponds to the $\sum_{i \in eval} x_i$ value at which TF_{thres} is surpassed on the high weight side (that is to around 95%) while the minOR to the $\sum_{i \in eval} x_i$ value for which the same occurs on the low weight side (that is to around 18%).

2.5 General course of the Maximum Occurrence analysis

Having discussed the details of the specific steps that are required to set up a MaxOcc analysis (pool generation. restraint back-calculation, possible adjustment of the optimization algorithm) the ground is now prepared for a presentation of the most common strategies adapted during a MaxOcc study. The choice of a specific approach is aimed at extracting as much information as possible with the available

experimental restraints and thus depends on the properties of both the system under investigation and of the collected experimental data. Moreover the MaxOcc procedure have an interactive character, in the sense that the exact course of the analysis can be adjusted ‘on the fly’ based on the obtained partial results.

As long as the MaxOcc of single conformations is considered the goal of the analysis is unique and straightforward, that is to determine to entire set of structures with the highest MaxOcc values (the peaks of the MaxOcc profile). With the new linear implementation of MaxOcc the calculations are in many cases fast enough to achieve this goal in the most direct way through simply evaluating the MaxOcc value for every single conformer in the generated pool. Such an approach was pursued in Articles 3 and 4. If on the other hand one disposes of more restricted computational resources or if the calculations become particularly demanding (due to for example especially high number of experimental restraints or especially numerous pool) one can adapt a different strategy which exploits the fact that MaxOcc is a well-behaved, continuous function. The continuous character of the MaxOcc function implies that the structures with the highest MaxOcc will be surrounded by a neighborhood of other conformers with still reasonably high MaxOcc values. On can thus expect to find the peaks of the MaxOcc profile through the following iterative procedure. First the MaxOcc values for some random, broad selection of conformers (usually a few percent of the entire pool) are evaluated. Among these structures the ones with the highest MaxOcc are identified and then the MaxOcc of all their neighbors in the six dimensional rotation-translation space are calculated as the next iteration of the search. If among these newly evaluated conformers there are some with even higher MaxOcc values then in turn the MaxOcc of all of their neighbors is calculated until a (local) maximum of the MaxOcc profile is found (that is: a conformer with a MaxOcc value higher than that of any of its neighbors). The success of such an approach depends on evaluating a large enough set of structures in the initial step to cover (coarsely) the entire conformational space and on choosing a large enough radius of the neighborhood searched in each iteration. These two parameters must thus be set with care, considering the actual vastness of the conformational space of the system under study and the resolution of the conformational pool at hand. This procedure was applied in Articles 1 and 2. With either approach one arrives at the complete set of structures having the highest MaxOcc values among all the possible conformations of the system. It is the distribution of such structures that dictates the further steps taken during the subsequent analysis with MaxOR/minOR. The following paragraphs will present the preferable courses of the further analysis to be performed for the different distributions of the high MaxOcc structures.

In the simplest case all the conformers with the highest MaxOcc values form a single and relatively

narrow peak in the six dimensional rotation-translation space. In such situation the most reasonable next step is to check whether allowing for local mobility around the MaxOcc peak is enough to fit all the experimental observables. In practice such a test involves constructing a series of conformational regions centered at the MaxOcc peak (the structure with the highest MaxOcc) and containing all conformers which can be superimposed on it through performing a translation and rotation smaller than some fixed threshold (different for each region in the series). The MaxOR value of each such region is then calculated and if a MaxOR value of 100% is found for any of such regions (with still reasonable small translations and rotations involved) then it can be concluded that local mobility around a single structure is enough of explain all the experimental data. The size of the smallest region with a MaxOR = 100% quantifies the lowest amount of mobility that have to present in the studied system according to the experimental data available and determining it concludes the analysis. Such a case was found in practice in the study of calmodulin bound to the DAPk peptide (Article 1).

If on the other hand no compact region constructed around the MaxOcc peak can explain all the experimental observables (MaxOR < 100 % for all the considered regions) is a clear indication that also structures clearly distinct from the ones encountered at the MaxOcc peak have to be present in the conformational ensemble of the system. One thus needs a methods of identifying the most probable candidates for such additional states. Assuming that we indeed deal with well-defined minor states (and not for examples a broad distribution of additional structures) and that their population is high enough, one can try to locate them by an additional series of MaxOR calculations performed on regions composed of the MaxOcc peak and different possible canditates for the minor states. Such an approach was pursued in Article 3 and thus the reader can refer there for more details.

If instead we arrive at a situation in which local mobility doesn't explain all the data (MaxOR < 100%) yet the minor states are necessarily lowly populated (the MaxOR near to the MaxOcc peak is already very high) one can still try locate them by exploiting the notion of minOR. This is possible however only if the structures composing the minor state(s) have some distinguishing characteristics in the experimental data space (see section 2.4). Still a system for which such an approach proves very useful was actually already encountered in practice and its MaxOcc study is described in detail in Article 4.

Moving further, if the set of structures with the highest MaxOcc values forms not one but instead a series of well-defined peaks, generalized versions of the approaches described in the previous paragraph can still be applied. One would thus check whether all the different MaxOcc peaks taken together can explain the experimental data (MaxOR > 100%). Then, if it is not the case, proceed with allowing local mobility around each center and finally, if it is required, search for additional minor states with the methods similar

to those briefly mentioned in the previous paragraph. A case in which the structures with the highest MaxOcc form two distinct well-defined peaks was encountered for one of the simulations described in Article 2.

Finally, in cases in which instead of well-defined peaks the structures with high MaxOcc values adapt a broad continuous distribution, one can concentrate on assessing whether such a broad distribution is indeed a reflection of the real conformational sampling of the studied system or whether maybe some of the encountered high MaxOcc values are actually just artifacts of the data averaging. The second situation may occur in the case of systems exchanging between several different states present at similar populations. For such systems one sometimes observe high MaxOcc values for structures laying in between the actually sampled states. This is because of the similarity of the experimental restraints expected for such an 'average structure' to the population weighted averages of the restraints back-predicted for the actually sampled conformers. A series of MaxOR calculations using regions composed of different subsets of the structures with high MaxOcc can in such cases distinguish between real states and such 'average structures'. Such cases were encountered in several simulations described in Articles 1 and 2 and the respective publications provide detailed information about how exactly the MaxOR calculations are able to identify the real states.

The conceivable distributions of the structures with the highest MaxOcc are of course infinitely many and it is thus impossible to formulate guidelines for every case one may encounter in practice. The examples described above should however be enough to provide insights into the way of thinking usually adapted during the subsequent MaxOR analysis.

26 Methodological aspects

This page intentionally left blank

3 Results

3.1 Exploring regions of conformational space occupied by two-domain proteins

Witold Andrałojć[a,b], Claudio Luchinat[a,b], Giacomo Parigi[a,b], Enrico Ravera[a,b]

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

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

*J. Phys. Chem. B, **2014**, 118 (36), pp 10576–10587*

28 Results

This page intentionally left blank

THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Exploring Regions of Conformational Space Occupied by Two-Domain Proteins

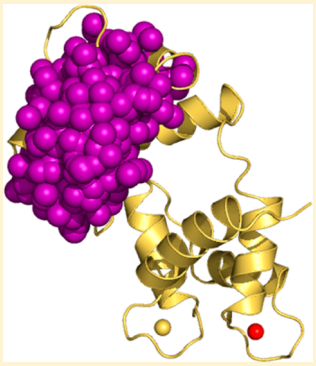
Witold Andrałojć,[†,‡] Claudio Luchinat,*[,†,‡] Giacomo Parigi,[†,‡] and Enrico Ravera[†,‡]

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

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

Supporting Information

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



1. INTRODUCTION

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

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

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

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

2. METHODS

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

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

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

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

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

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

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

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

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

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

The maxOR and minOR values were determined by solving

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

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

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

3. RESULTS AND DISCUSSION

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

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

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

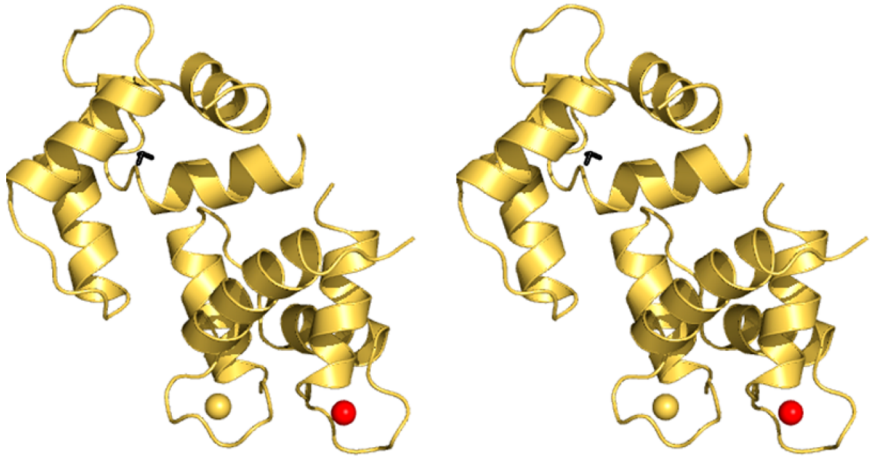


Figure 1. Stereoview (cross-eye) of the conformation selected for the generation of synthetic data. The red and yellow spheres represent the calcium-substituted lanthanide and the calcium ions, respectively. This conformation can be represented by a triad of Cartesian axes (in black), centered at the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

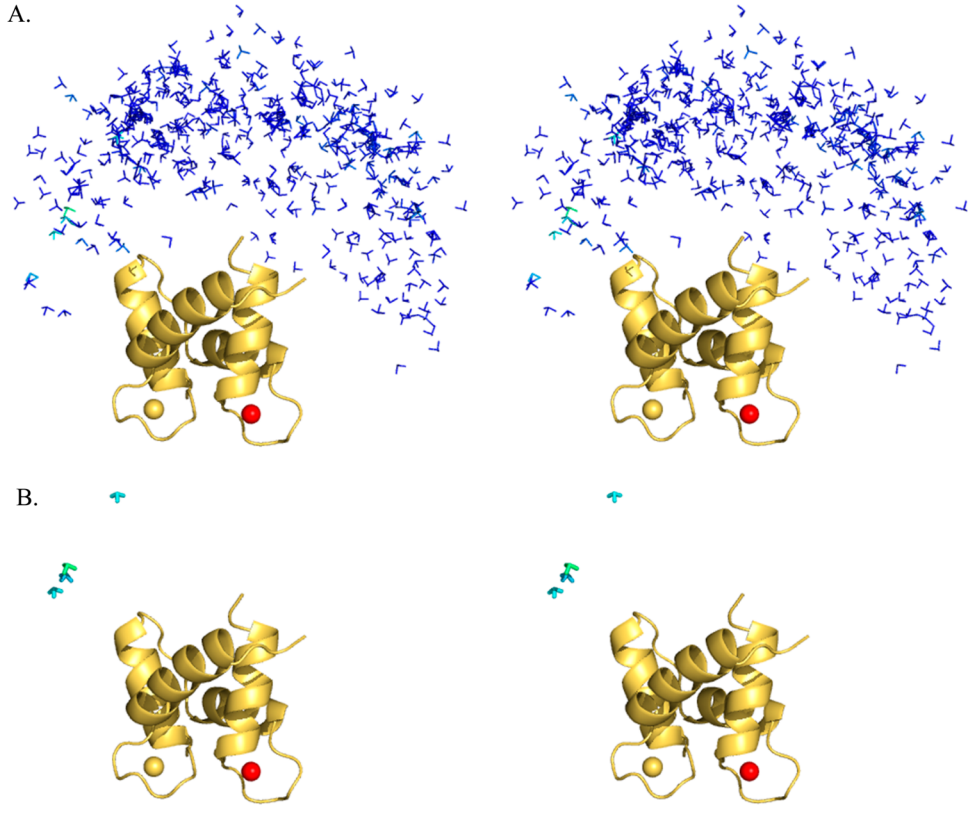


Figure 2. Stereoview (cross-eye) of (A) 400 sterically allowed conformations, color-coded according to their MaxOcc values and (B) conformations with the highest MaxOcc (ranging from 0.17 to 0.29) among these 400. Each conformation is represented as a triad of axes, centered at the center of mass of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

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

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

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

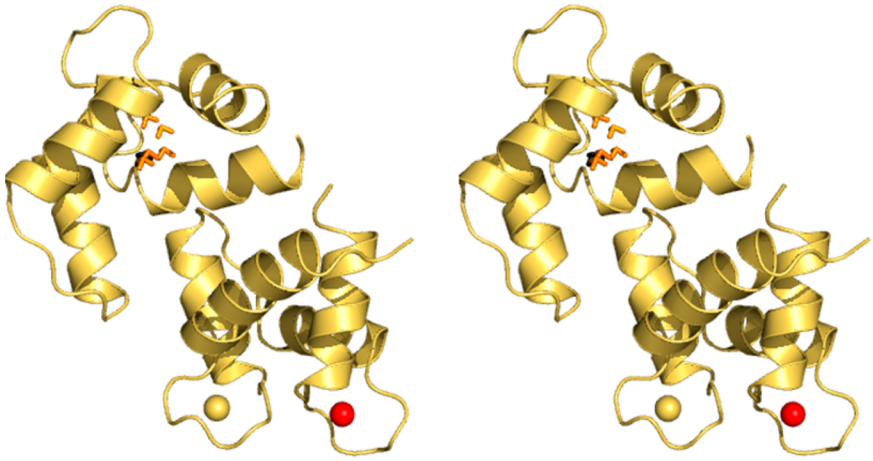


Figure 3. Stereoview (cross-eye) of conformations with the highest MaxOcc (0.66−0.71) in comparison to the reference conformation (in black).

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

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

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

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

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

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

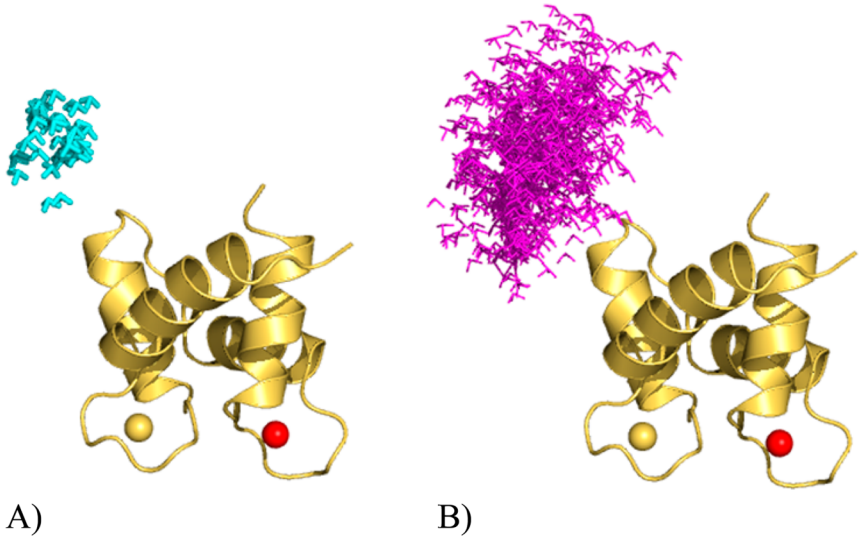


Figure 4. (A) Smallest region with maxOR = 1 and (B) smallest region with minOR = 0.54.

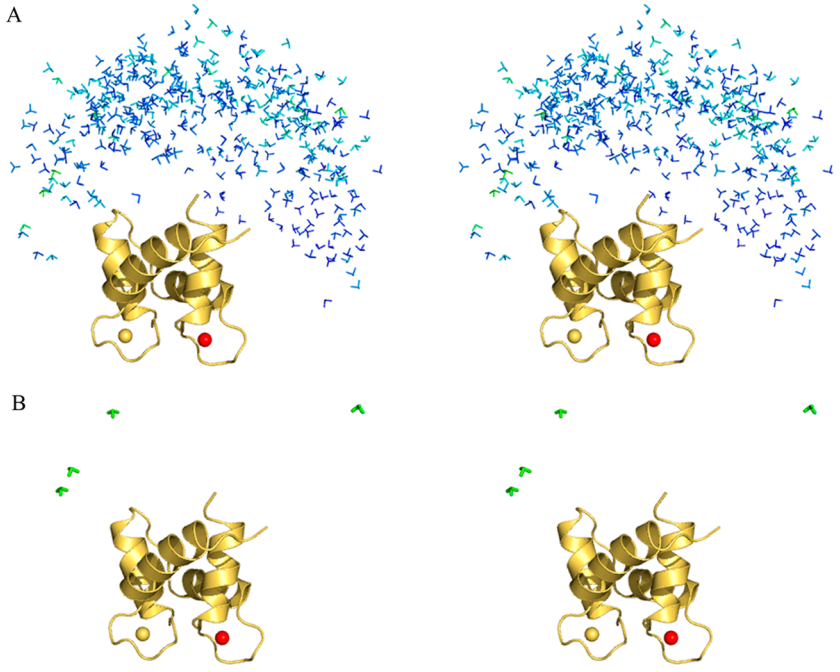


Figure 5. Stereoview (cross-eye) of (A) 400 sterically allowed conformations of CaM when bound to the DAPk1 peptide, color-coded according to their MaxOcc values, and (B) conformations with the highest MaxOcc (ranging from 0.36 to 0.43) among these 400. Color code is from <0.0 (blue) to >0.9 (red).

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

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

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

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

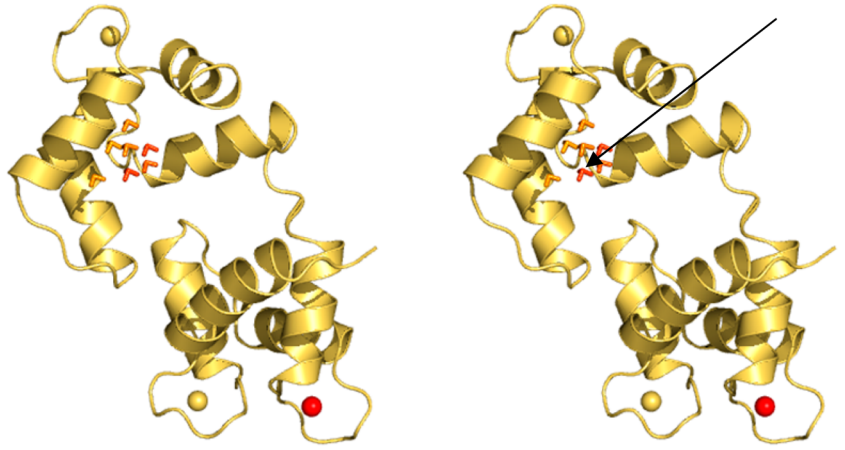


Figure 6. (A) Stereoview (cross-eye) of the conformations of CaM, when bound to the DAPk1 peptide, with the highest MaxOcc (0.65−0.81) in comparison to the experimentally determined solution structure (indicated by the arrow).

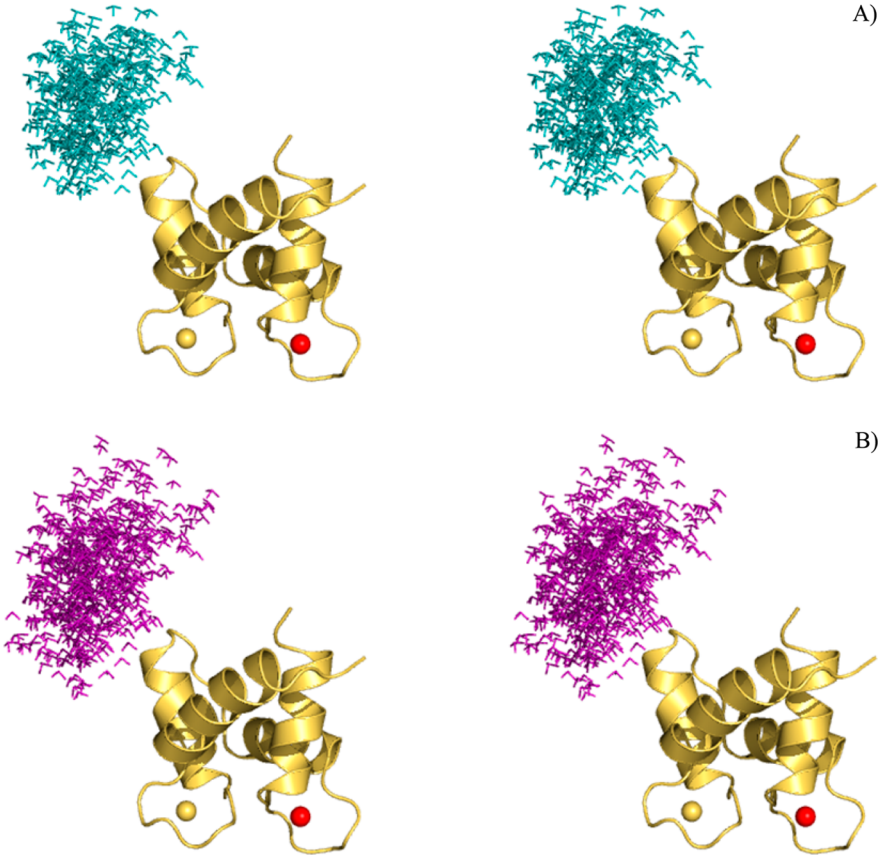


Figure 7. (A) Smallest region with maxOR = 1 and (B) smallest region with minOR = 0.10 for CaM, when bound to the DAPk1 peptide.

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

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

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

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

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

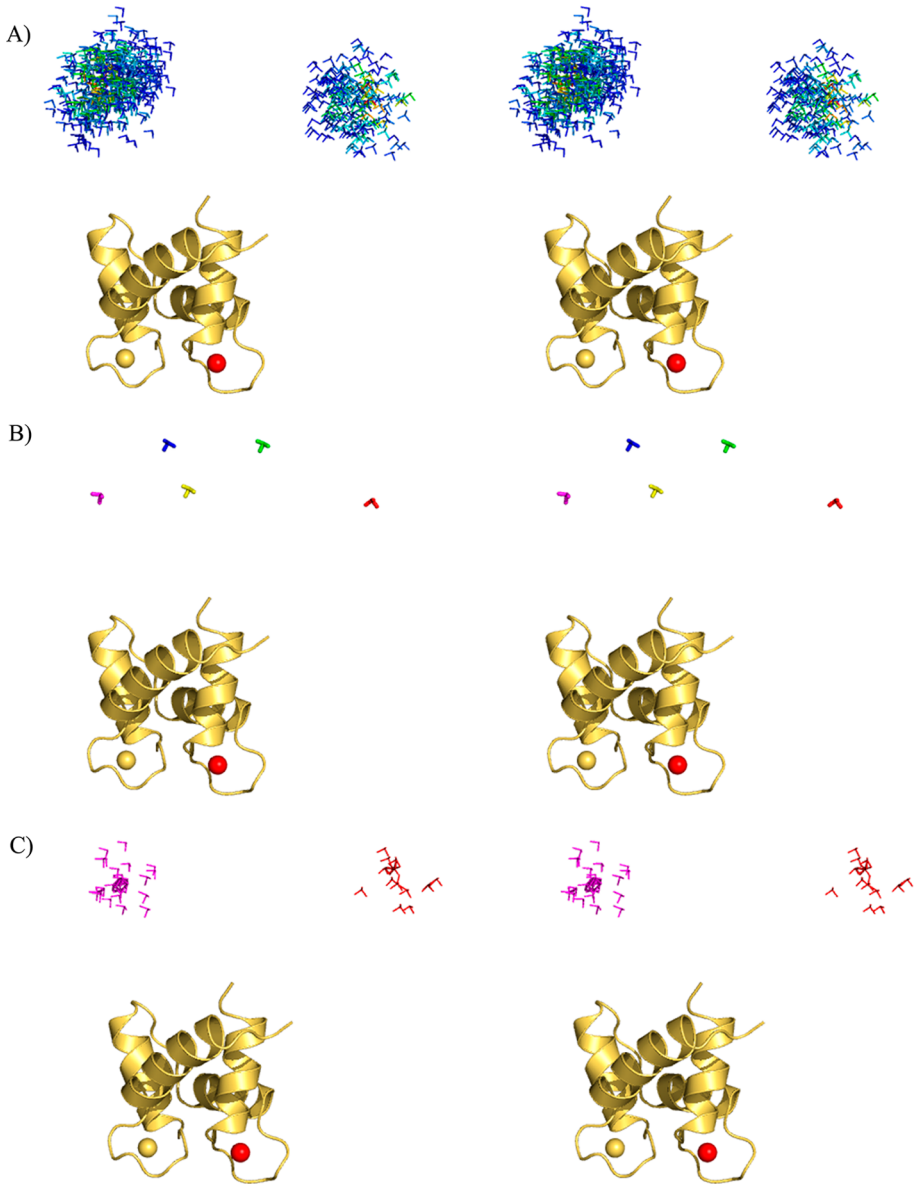


Figure 8. Stereoview (cross-eye) of (A) the regions of structures used for generating the second synthetic data set (the color indicates the relative weight, from 0.0 (blue) to >0.9 (red); the structure in the center of each region was considered with a relative weight of 1), (B) the structures having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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

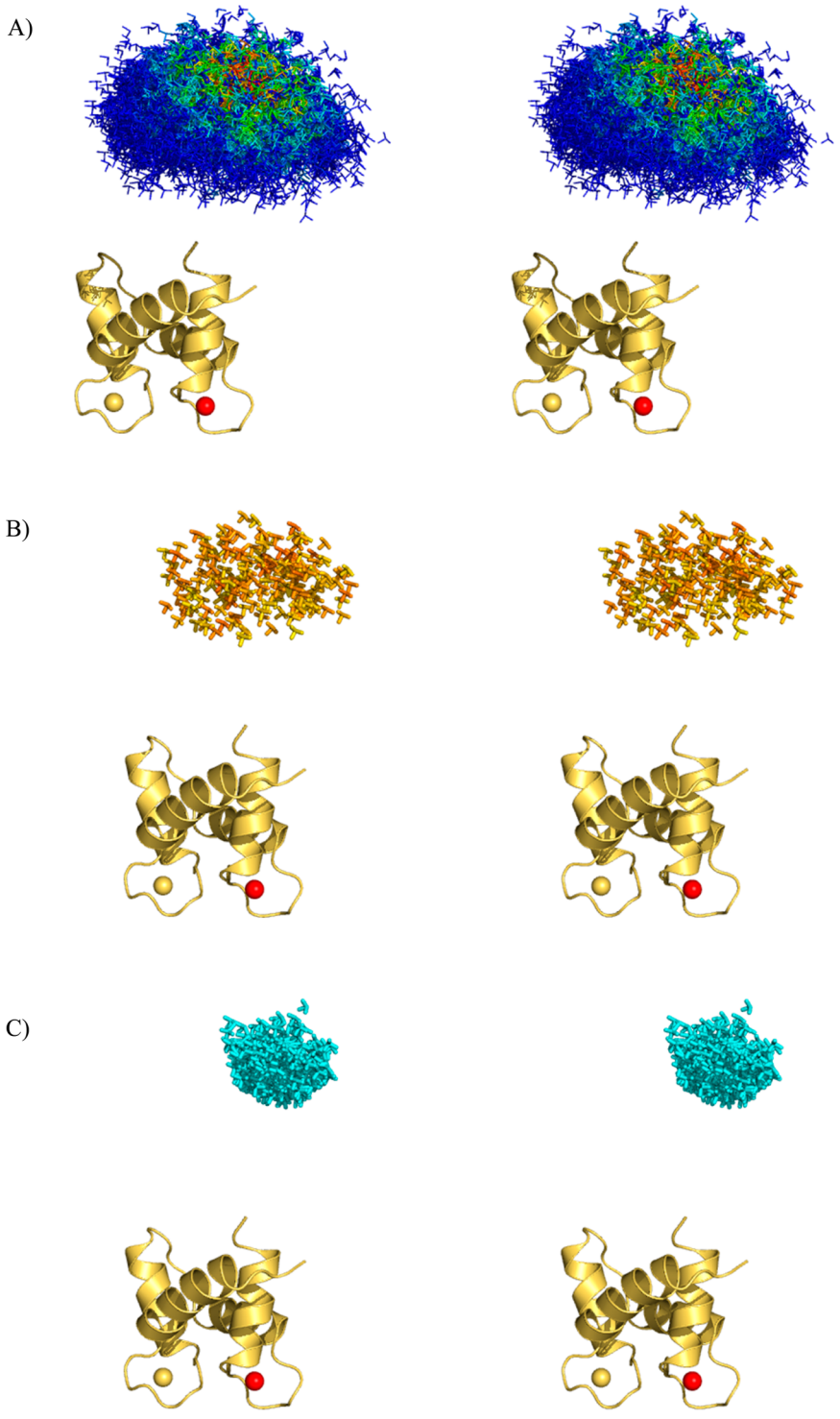


Figure 9. Stereoview (cross-eye) of (A) the region of conformations used for generating the third synthetic data set (the color indicates the relative weight, from 0.0 (blue) to >0.9 (red); the structure in the center of each region was considered with a relative weight of 1), (B) the conformations having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

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

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

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

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

4. CONCLUSIONS

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

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

■ ASSOCIATED CONTENT

Supporting Information

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

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

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

■ REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Exploring regions of conformational space occupied by two-domain proteins

Witold Andralojc, Claudio Luchinat*, Giacomo Parigi, Enrico Ravera

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

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

Supporting information

Synthetic test 1

One ensemble of conformations was generated around one selected closed CaM structure. The conformations to be included in the ensemble were chosen based on the proximity in the Cartesian space (measured by the distance *d* between the centers of the mass of the C-terminal domain of each structure and the selected one) and in orientation (measured by α, which is the angle between the quaternions describing the orientation of the C-terminal domain of each structure and the selected one). The two parameters were merged into a single equation as described in [1;2]:

$$\Delta = d + f(1-\cos\alpha) \qquad \text{(S1)}$$

All the structures with $\Delta<7$, as calculated with f=117, were included in the simulated conformational ensemble. The conformations were given a Gaussian weight depending on their distance from the selected structure. The simulated PCS and RDC data were then generated from this ensemble, and a Gaussian error was added (with standard deviation of 1.5 Hz for RDCs and 0.02 ppm for PCSs).

The search for the structures displaying the highest MaxOcc values was performed according to the following methodology. As the first step, the MaxOcc values for 400 randomly chosen conformations were calculated, as described in references [1;3]. The values of the calculated MaxOcc ranged from 0.01 to 0.29. Out of these structures, those having MaxOcc higher than 80 % of the maximum occupancy of the best scoring structure found, were selected (with a minimum number of 4 selected structures). Around each of them, a region in the conformational space was built, consisting of the structures most similar to it. For the present case, the regions were defined by setting Δ to 5 and f to 37.3, corresponding to a maximum angular displacement of 30°.

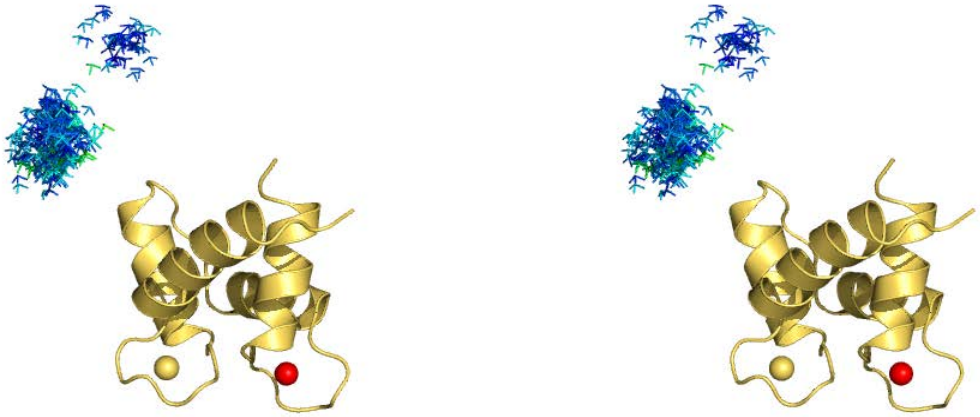
The MaxOcc values were then calculated for all the structures of the pool belonging to these regions (i.e. with a distance $\Delta<5$ from the best scoring structures), and again the best scoring structures were selected and new regions were built around them as done previously. This procedure was iteratively repeated until no better scoring structures were found in the explored regions. The search is summarized by the steps given in table S1, where the number of structures evaluated in all iterations, the range of their MaxOcc values and the number of the selected highest scoring structures is reported. A graphical representation of the best scoring structures is given in Figures 2, S1, and 3.

Table S1. Number of evaluated structures, range of their MaxOcc values and number of the selected highest scoring structures relative for each step of the procedure used to determine the highest MaxOcc structures.

Step	Number of structures	Figure	MaxOcc range	Number of selected structures	Figure
1	400	2a	0.01-0.29	4	2b
2	219	S1a	0.04-0.46	6	S1b
3	178	S1c	0.03-0.68	4	S1d
4	164	S1e	0.05-0.71	4	S1f
5	110	S1g	0.05-0.71	4	S1h
6	77	S1i	0.06-0.71	5	S1j
7	87	S1k	0.06-0.71	5	3

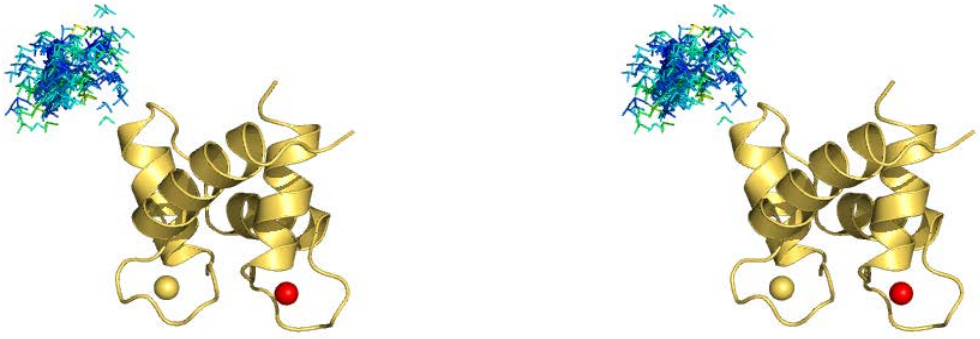
Figure S1: Stereoviews (cross-eye) representing the intermediate steps of the search described in table S1.

a)



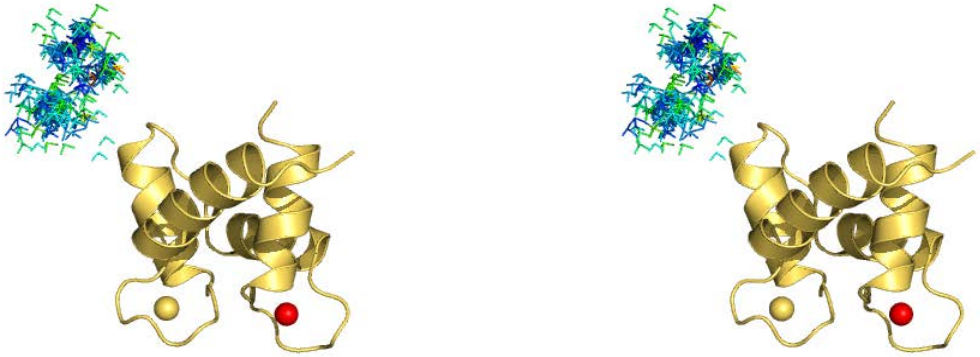
b)

c)



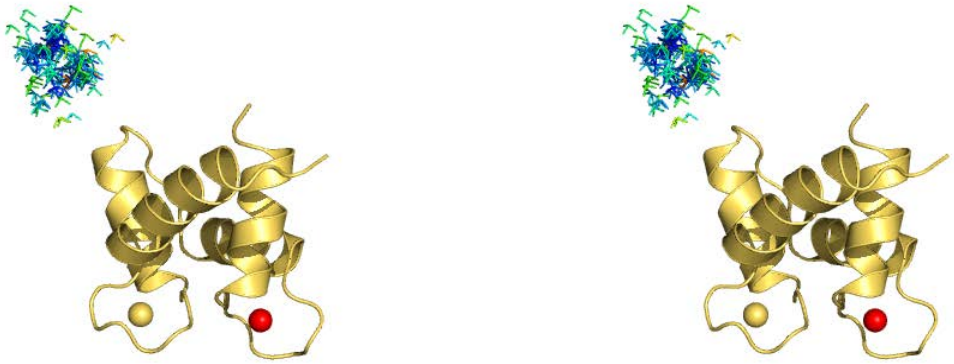
d)

e)



f)

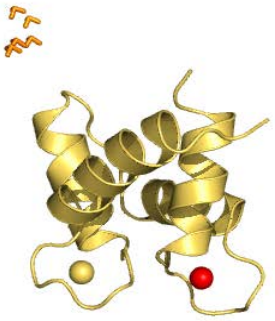
g)

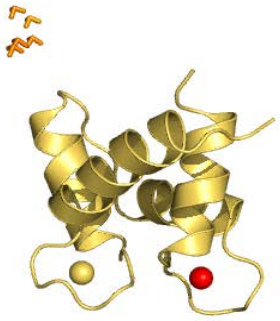


h)

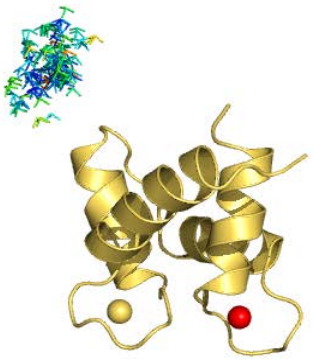
i)

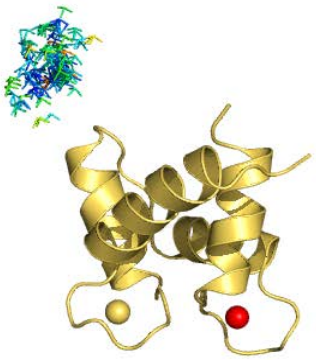
j)





k)



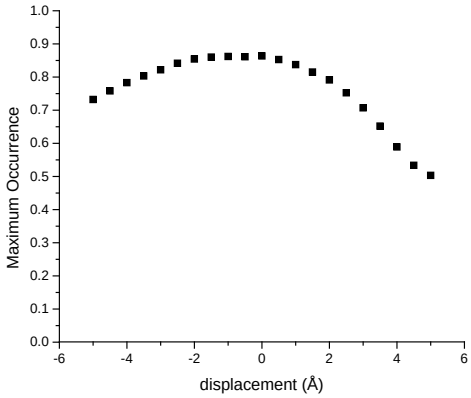


Behavior of MaxOcc as a function of displacement and validity of the sampling

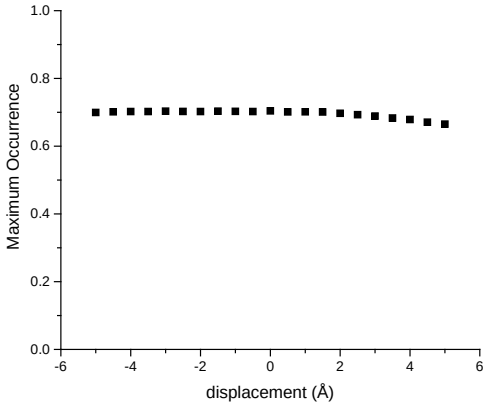
To test the validity of the sampling scheme used for the present analysis, we have checked the behavior of the MaxOcc values as a function of the distance from the reference conformation. We have thus calculated the MaxOcc values of the conformations that are generated from the reference conformation changing either one of the coordinates of the center of mass (along the x axis of the 2K61 pdb file) or rotating the C-terminal domain around its center of mass along the same axis. This was done for both simulation 1 (Figure S2a,c) and simulation 3 (Figure S2b,d). To endow these values with significance on the overall calculation performance, we have calculated the average minimum distance in terms of translation and rotation of the nearest neighbor of each conformation in the pool. The average minimum distance is 3 Å and 7°. Therefore, we conclude that the sampling is sufficient also in the case of the extremely limited mobility of simulation 1.

Figure S2: MaxOcc values as a function of distance in terms of translation (a,b) or rotation (c,d) from the reference conformations of simulation 1 (a,c) and simulation 3 (b,d).

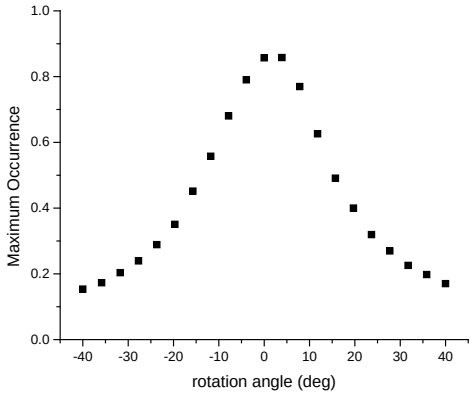
a)



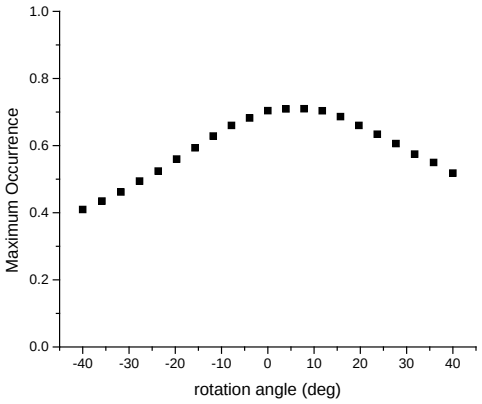
b)



c)



d)



Determination of regions

The structures within the pool were scored via equation S1 in terms of distance from a selected structure. The search for regions with maxOR=1 and with minOR> 0 was accomplished by calculating the maxOR and minOR values for regions of different size (in terms of rotational or spatial difference) through a grid analysis.

The results of the grid search is reported in Table S2 (maxOR) and S3 (minOR). Table S4 reports the dimensions of the regions, as the percentage of the structures considered with respect to the entire pool. The regions discussed in the main text are highlighted in the tables.

Table S2. MaxOR values for different sizes of the region defined around a selected structure

		distance [Å]							
		5	7	9	11	13	15	17	19
angle [deg]	10	0.93	0.96	0.96	0.96	0.96	0.97	0.97	0.96
	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	35	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	40	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table S3. MinOR values for different sizes of the region defined around a selected structure

		distance [Å]							
		5	7	9	11	13	15	17	19
angle [deg]	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	15	0.00	0.04	0.17	0.27	0.38	0.37	0.39	0.37
	20	0.00	0.09	0.22	0.39	0.54	0.56	0.59	0.59
	25	0.00	0.08	0.23	0.37	0.60	0.65	0.68	0.70
	30	0.00	0.09	0.20	0.37	0.62	0.68	0.71	0.75
	35	0.00	0.08	0.21	0.38	0.62	0.67	0.72	0.75
	40	0.00	0.09	0.23	0.38	0.60	0.68	0.71	0.74
	45	0.00	0.08	0.22	0.38	0.63	0.67	0.72	0.75

Table S4. Percentage of the structures in the analyzed regions (with respect to the whole pool) on varying their size, defined by the maximum rotation and translation from a selected structure

		distance [Å]							
		5	7	9	11	13	15	17	19
angle [deg]	10	0.03	0.07	0.13	0.18	0.20	0.22	0.23	0.23
	15	0.07	0.17	0.36	0.51	0.61	0.68	0.72	0.74
	20	0.15	0.34	0.72	1.02	1.29	1.50	1.63	1.71
	25	0.23	0.57	1.13	1.64	2.11	2.52	2.81	2.99
	30	0.32	0.80	1.58	2.34	3.08	3.74	4.31	4.71
	35	0.39	1.03	2.03	3.04	4.06	5.05	5.94	6.63
	40	0.45	1.20	2.42	3.70	5.01	6.33	7.63	8.73
	45	0.49	1.33	2.73	4.26	5.85	7.53	9.20	10.73

Synthetic test 2

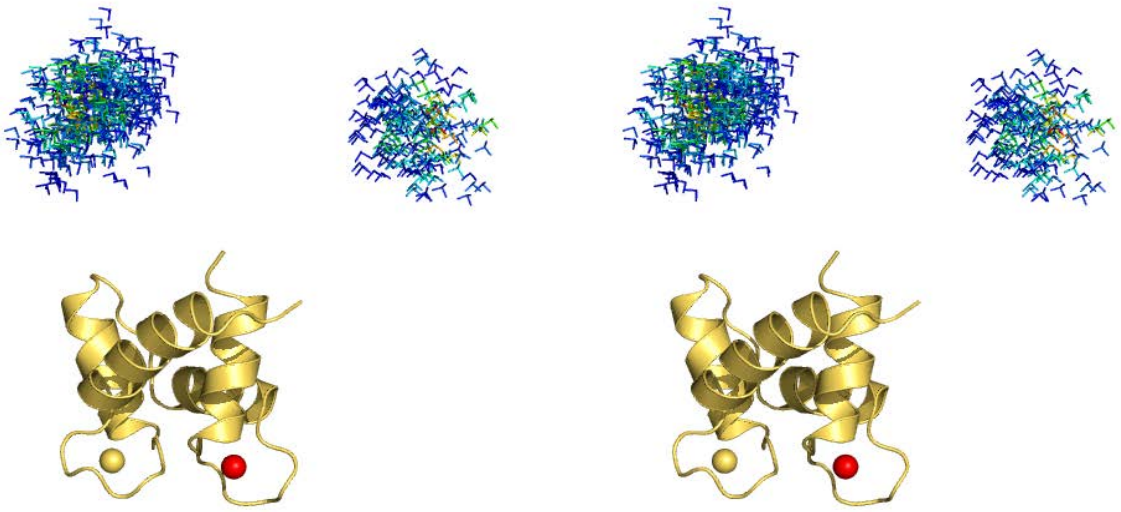
Two ensembles of conformations were generated around two randomly selected structures. All the structures fulfilling equation S1 with Δ<15 and f=64.1 were included in the simulated ensembles. The conformational space sampled by the protein in this case was thus composed of the two regions shown in Figure S4a.

As described for the synthetic test 1, the search for structures having the highest MaxOcc was first accomplished. Since the MaxOcc distribution is broad and with rather low values as compared to the previous case, moving with the same criteria would end up in evaluating a prohibitively high number of structures, so the pursuit was dropped at the second iteration, when 5 regions of high MaxOcc structures were visible (Figure S4b-f).

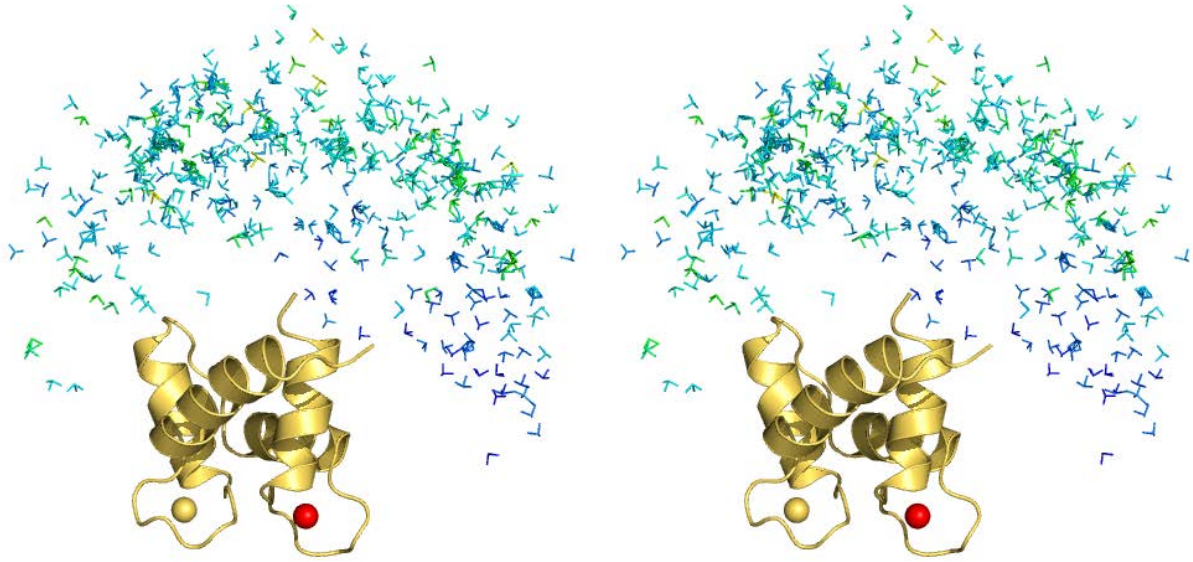
For each of those regions, the structure having highest MaxOcc was selected (Figure S4g). Regions were then built around each of the five selected highest MaxOcc structures (Figure S4h) by taking all conformations from the calculated pool with maximum translation of 5 Å and rotation of 20° with respect to the selected structures.

Figure S3: Stereoview (cross-eye) of a) the regions of structures used for generating the synthetic data; b-f) the steps of the search for high MaxOcc structures, g) the structures having highest MaxOcc from each region and h) the regions built around them; i) pair of regions with the highest MaxOcc.

a)

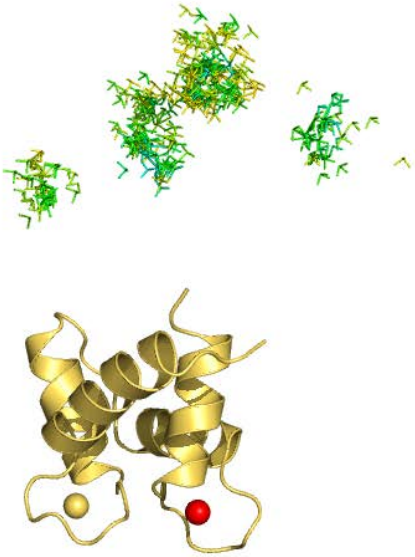


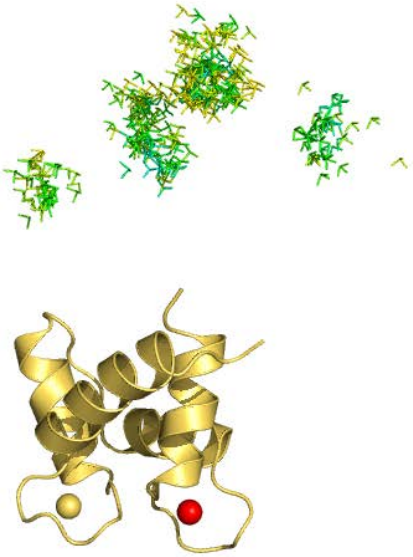
b)



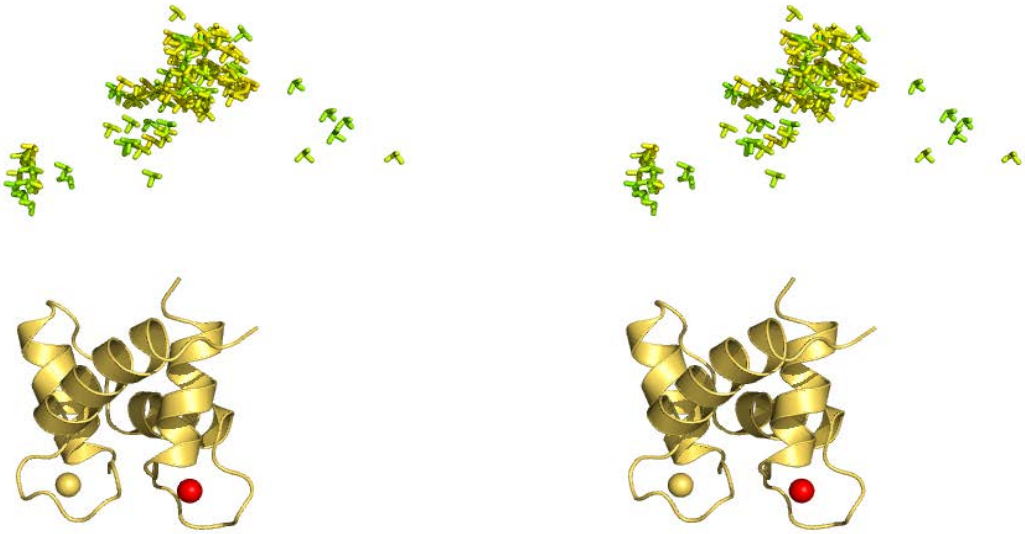
c)

d)

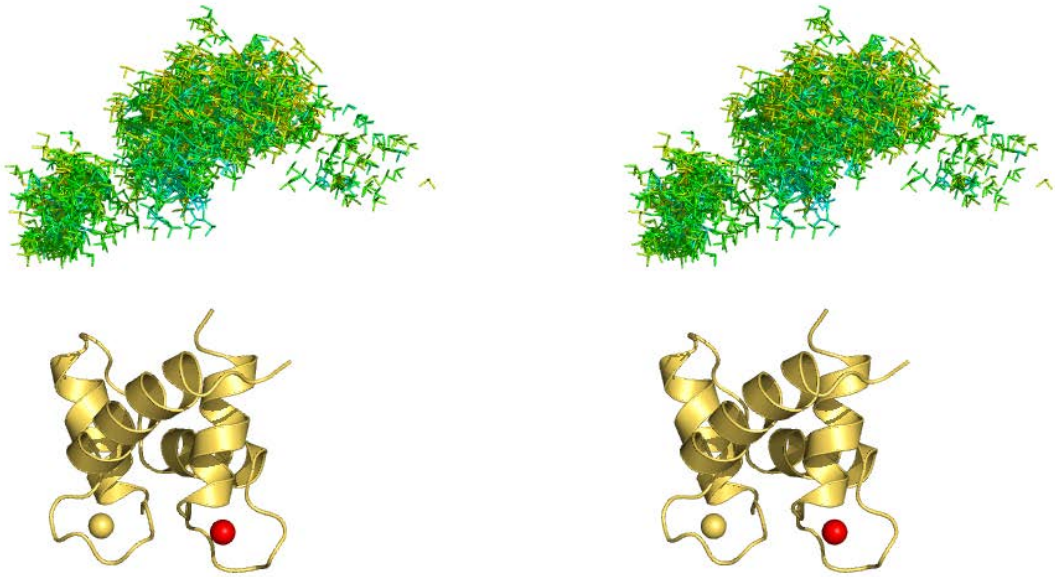




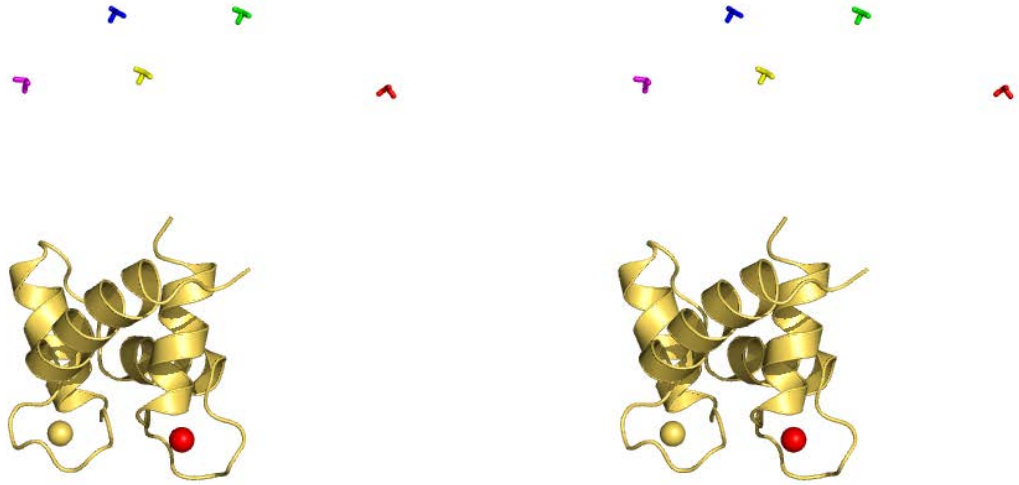
e)



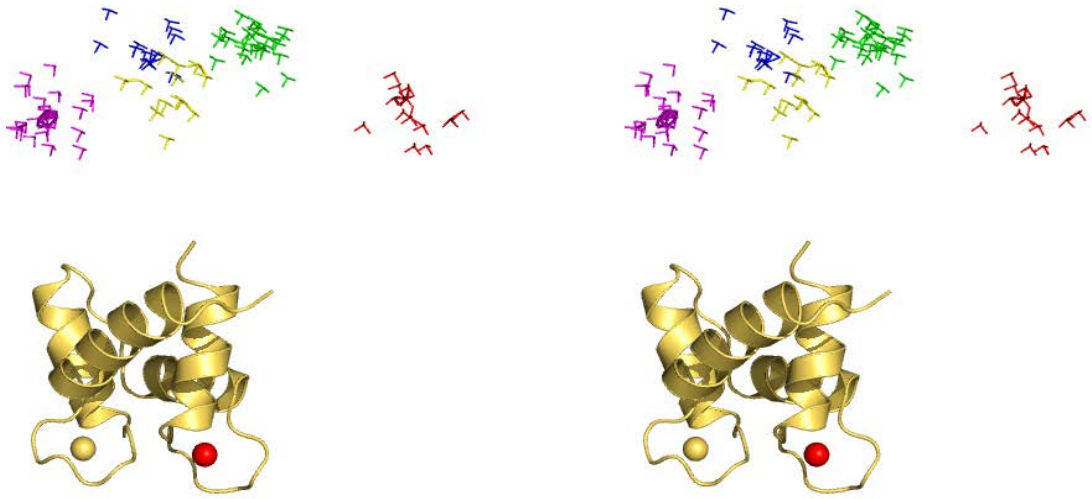
f)



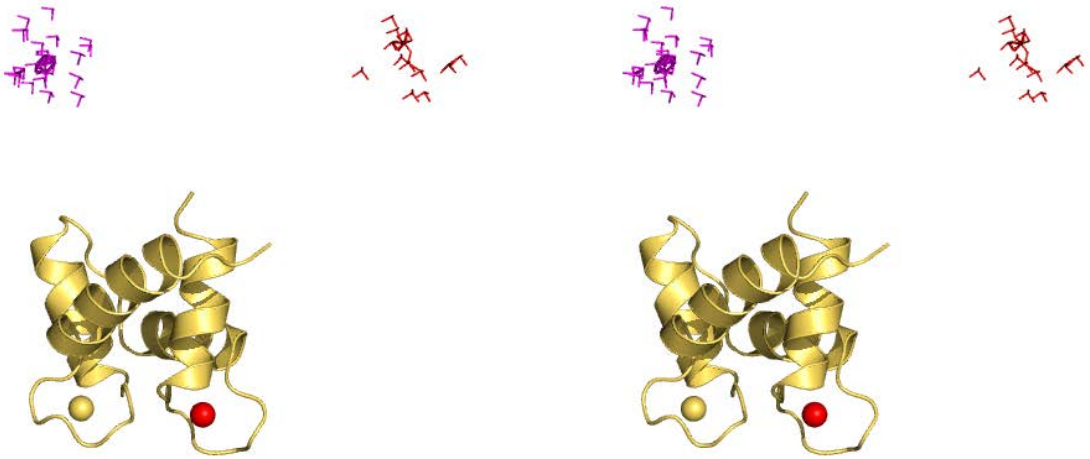
g)



h)



i)



Synthetic test 3

A wide pool of conformations was generated around a structure positioned between the two central structures used in the synthetic test 2 (structure B in Table 1). Different values of Δ and f to be plugged into eq. S1 were tested in order to obtain a reduction of the C-terminal RDC-derived tensor similar to the previous simulation (similar amount of mobility). In the end, all the structures of the pool fulfilling equation S1 with a maximum Δ of 30 and with f=36.3 were included in the simulated conformational ensemble. This selected ensemble is shown in Figure 7a. The conformations were given a Gaussian weight as described for Synthetic test 1.

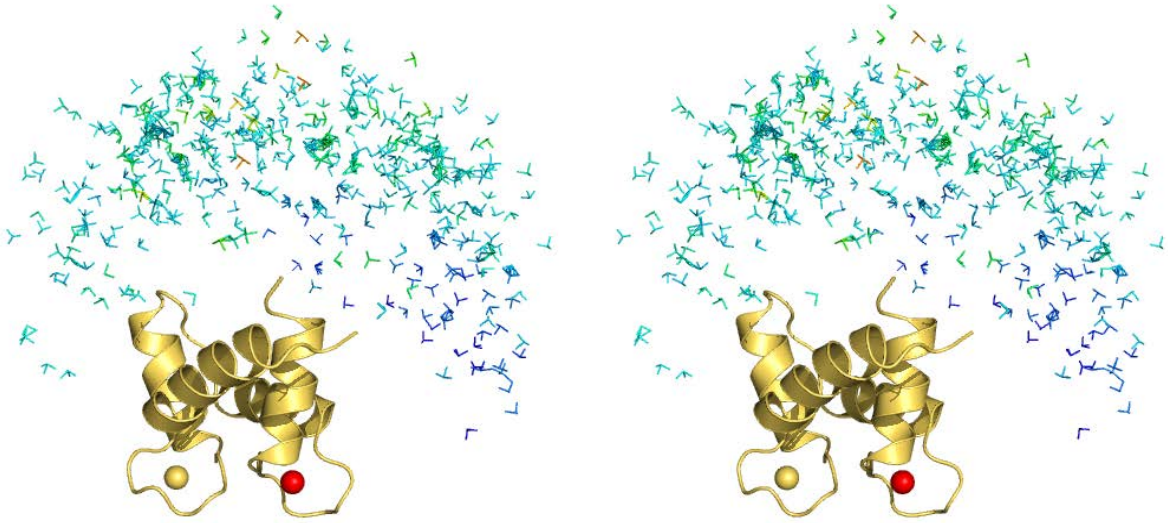
As in the previous cases, the search for structures having the highest MaxOcc was accomplished (Figure S5a-g). The number of the highest scoring conformations increased largely at each iteration, so the pursuit was dropped on the 3rd iteration due to prohibitively high number of structures that would need to be evaluated (Figure 7b). The MaxOcc values of the best scoring structures are noticeably higher that in the previous simulation (up to 0.71) and in opposition to it all those structures form only one, relatively big, but well defined region in space.

Out of them the structure with the highest MaxOcc was selected (Figure S5i) and used as the center of regions of different sizes (a grid search was performed as described for Simulation 1), the maxOR of which were calculated.

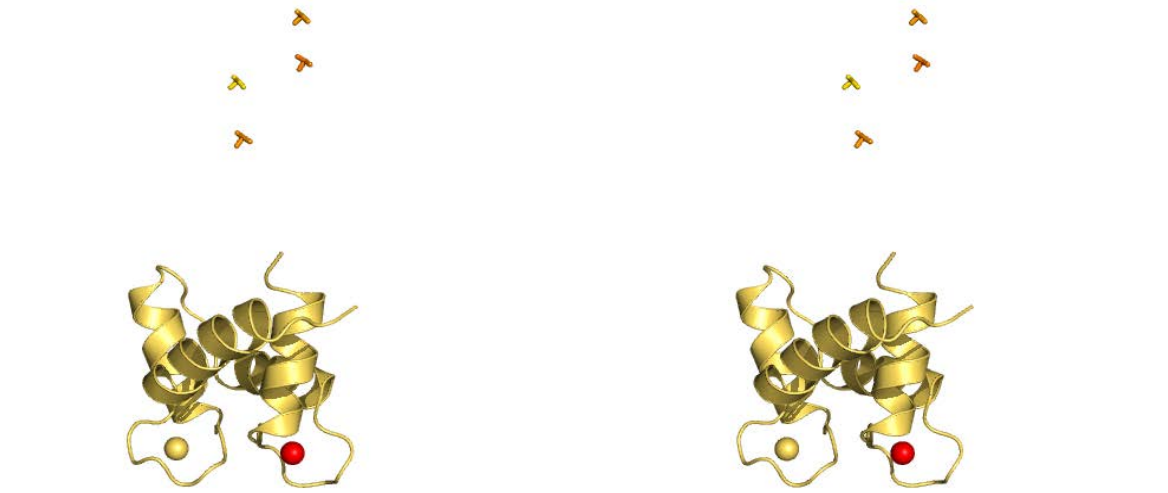
The results of the performed grid search are reported in Table S5 and the smallest region with maxOR=1 is depicted in the Figure 7c.

Figure S4: Stereoview (cross-eye) of a-g) the steps of the search for high MaxOcc structures, i) the conformation with highest MaxOcc.

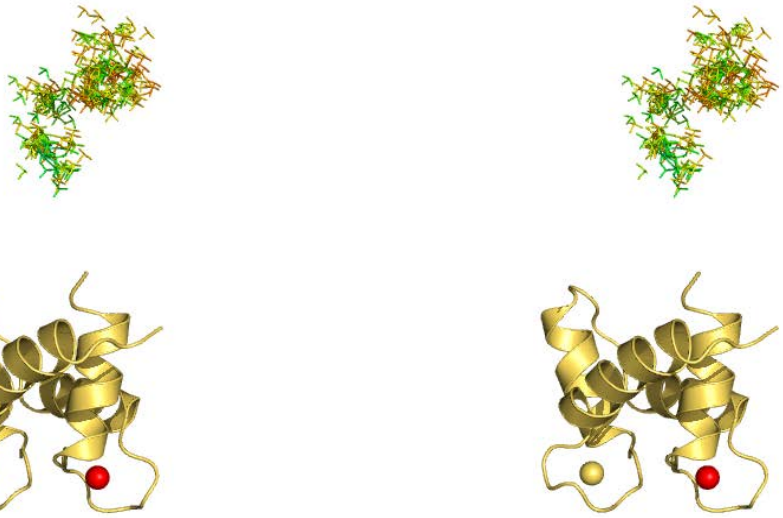
a)



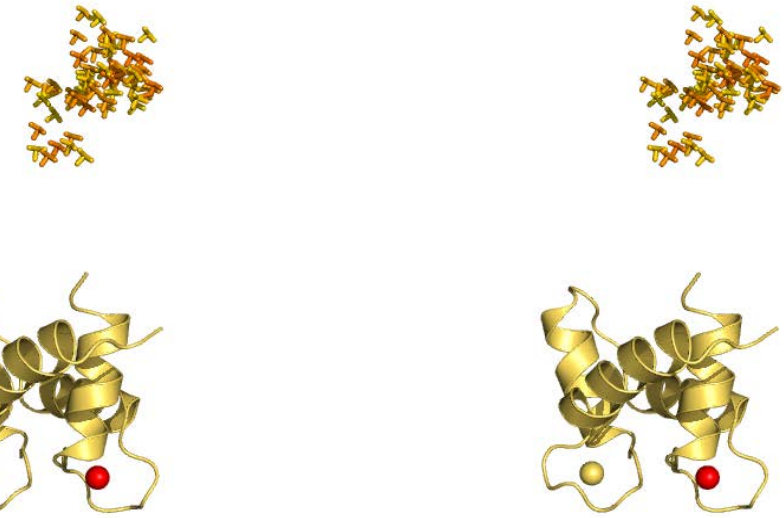
b)



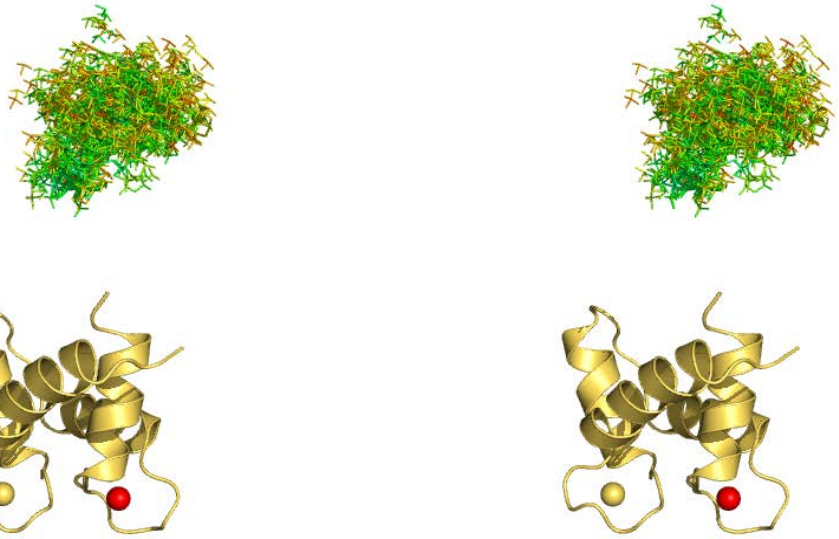
c)



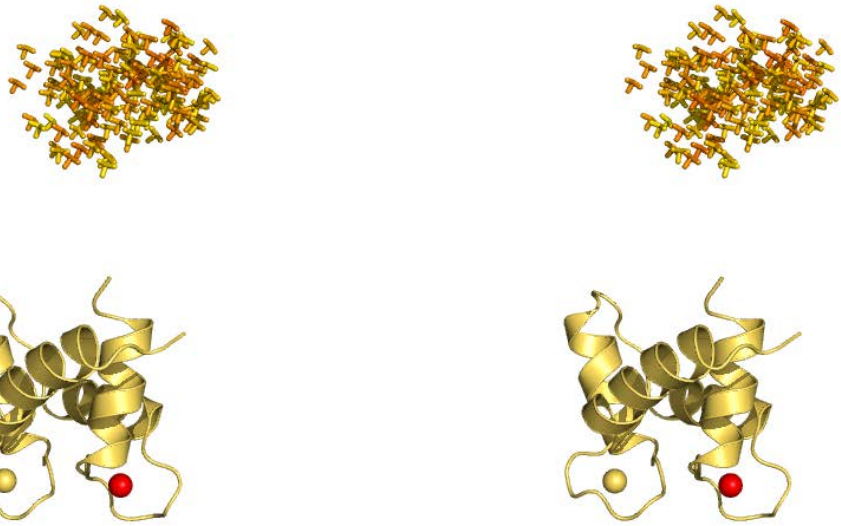
d)



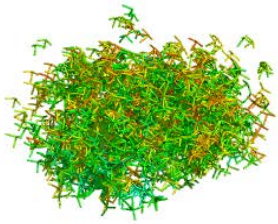
e)

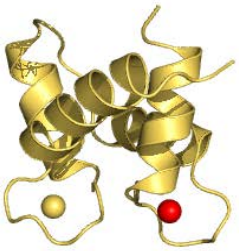
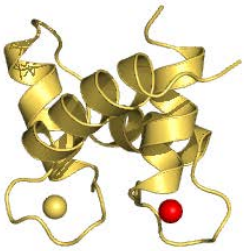


f)



g)



h)



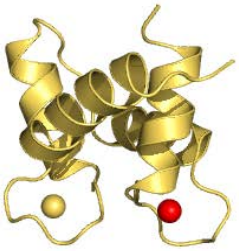
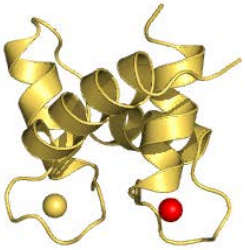



Table S5

		distance [Å]							
		5	7	9	11	13	15	17	19
angle [deg]	10	0.76	0.78	0.78	0.78	0.79	0.79	0.79	0.79
	15	0.85	0.87	0.89	0.90	0.90	0.91	0.91	0.91
	20	0.95	0.99	1.00	1.00	1.00	1.00	1.00	1.00
	25	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	30	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	35	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	40	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	45	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Minimum Occupancy for synthetic tests 2 and 3

To order to restrict the number of conformations to be included in regions with minOR>0, an approach different from that one used in the synthetic test 1 was chosen. An ensemble of structures able to fit the data (one of the many) was first found. The weight of the structures belonging to this ensemble (excluding those with a negligibly small weight) was then constrained to zero, and a completely different ensemble of structures was thus found when a further fit of the data was performed. A new ensemble comprising the previous structures and the new found structure (with non-negligibly small weight) was then defined, its weight was constrained to zero, and a new fit was performed. The procedure is repeated a number of times, until it is no longer possible to fit the data with the weight of the conformations of the constructed ensemble constrained to zero: an ensemble with a non-zero minimum occupancy is thus found. The procedure can still be iterated to find larger and larger ensembles of conformations with increased minOR values. In the present cases we stopped the iterations when a minOR=0.2 was calculated; the corresponding regions comprised <14% on the conformational space.

This procedure may lead to the inclusion of redundant structures (structures which could be excluded from the region without lowering its minOR) into this generated first ensemble. The number of redundant structures can be largely reduced by repeating the whole procedure and defining a second ensemble in which the structures determined after all fits are included only if present in the first ensemble. After a number of iterations this second ensemble has the same minOR as the previous one, but with a reduced number of structures.

Because in this procedure the choice of the structures was not based on a spatial proximity criterion, but rather on the representation of the structures in the data space, the distribution of the conformations belonging to the identified regions over the conformational space can be very broad. In order to describe them and endow them with some physical significance, we examined their position and size in two-dimensional projections of the six-dimensional conformational space. Two types of projections were used. In the first the two considered variables were the distance between the centers of mass of the structures belonging to these regions and of a reference one, as well as the angle between the quaternions describing the orientation of the structures and of a reference one (thus this projection contains information on both the position and orientation of the domains). In the second, the variables were the angles between the x-axes and the z-axes of the frames of reference of the structures and of a reference one (thus this representation contains exclusively information about the orientation of the domains).

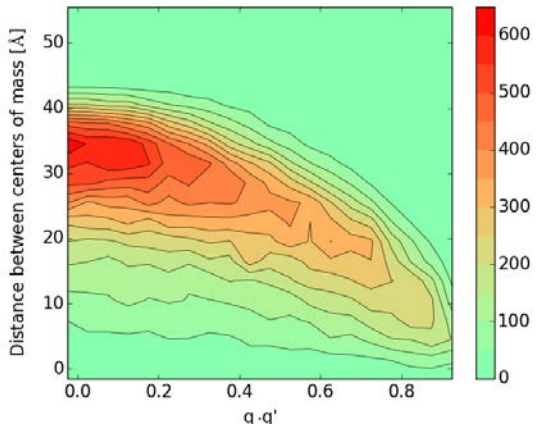
Figures S6 and S7 show the two-dimensional histograms of the number of structures for the first and the second projection, respectively, in the two site exchange simulation (synthetic test 2). These histograms are reported for all the structures present in the pool and representing the accessible conformational space (Fig. S6A,S7A), for the ensembles of structures really sampled by the protein in the simulations (Fig. S6B, S7B), for the smallest region with maxOR=1 (Fig. S6C,S7C) and for the just described ensembles with non-zero minOR (Fig. S6D,S7D). As clear from Figures S6A and S7A, the entire pool of structures is not evenly distributed along the considered variables, and this fact also affects the distributions in Fig. S6B-D and S7B-D. In order to monitor which parts of the space are actually over/under-represented in the considered sub-ensembles, we have shown in Figures S8 and S9 the histograms normalized by the total

amount of structures present in the pool with the same coordinate variables. Figures S10-13 show the histograms corresponding to the synthetic test 3.

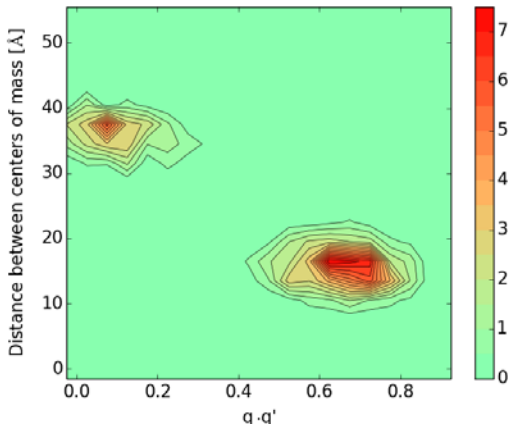
It is important to observe the peculiar appearance of the normalized histograms of the minimum occupancy calculations in the distance-angle representation (Fig. S8d and S12d). In this case regions with large usage-over-population ratios are obtained at the edge of the conformational space. On the one hand, this indicates that these regions contain conformations in fair agreement with the experimental observables and thus cannot be safely excluded; on the other hand these regions only contain few tens of structures, with the effect that their relevance appears overestimated.

Figure S5: The two site exchange simulation: (a-d) 2D histograms of the number of structures as a function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and of a reference a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR=1 d) for the ensemble with minOR =17% (e) highest MaxOcc values among all structures sharing the same distance and rotation.

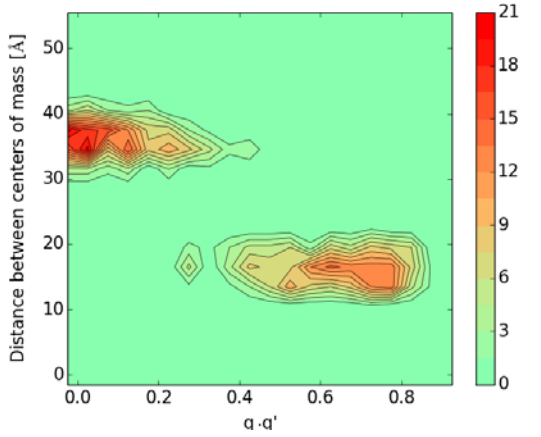
a.



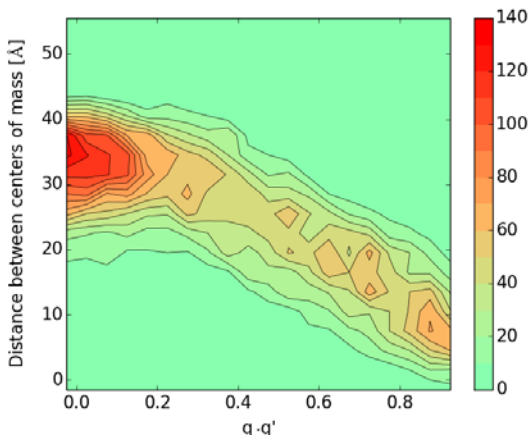
b.



c.



d.



e.

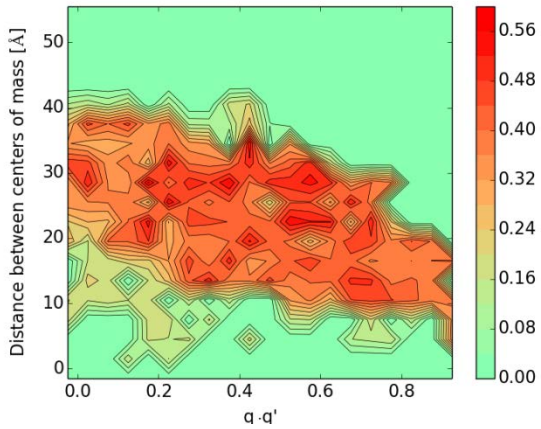
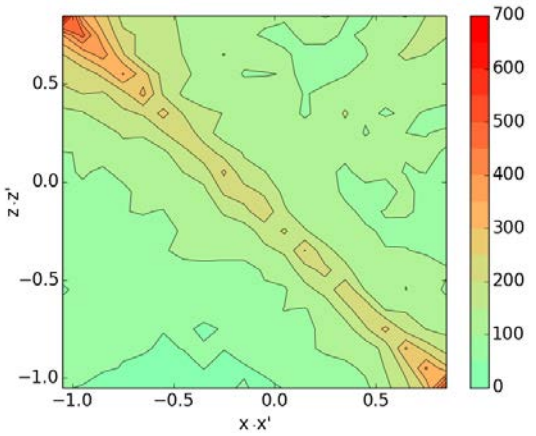
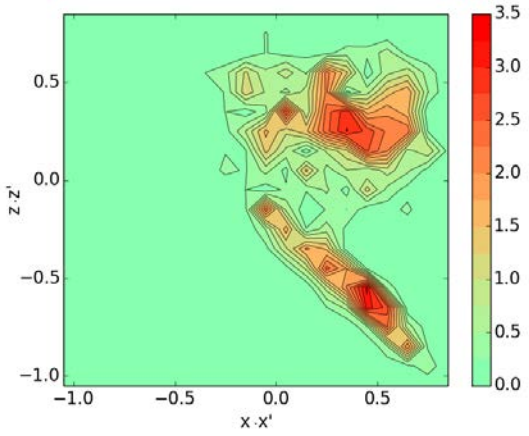


Figure S6: The two site exchange simulation: (a-d) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR=1 d) for the ensemble with minOR =17% (e) highest MaxOcc values among all structures sharing the same angles between x-axes and z-axes.

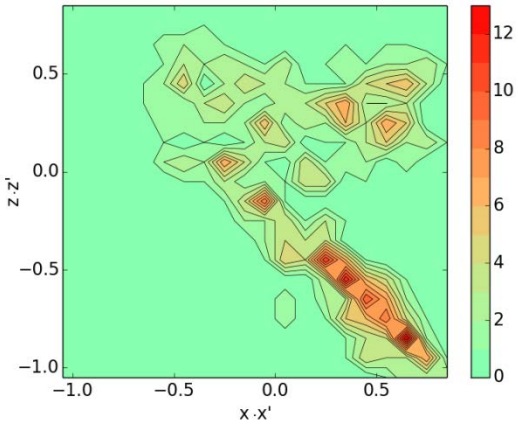
a.



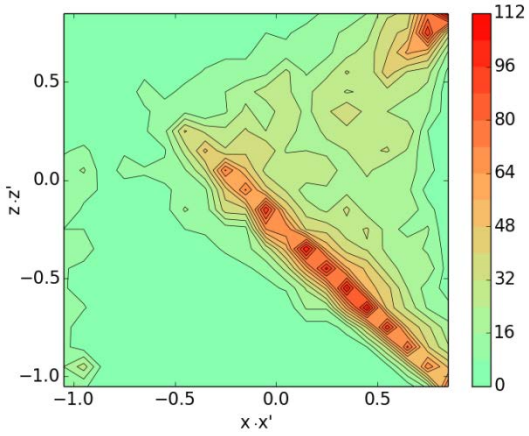
b.



c.



d.



e.

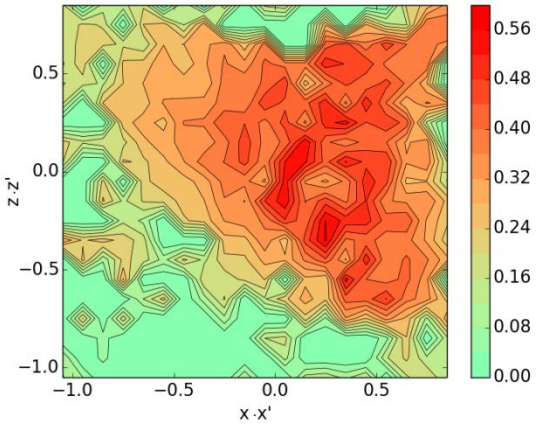
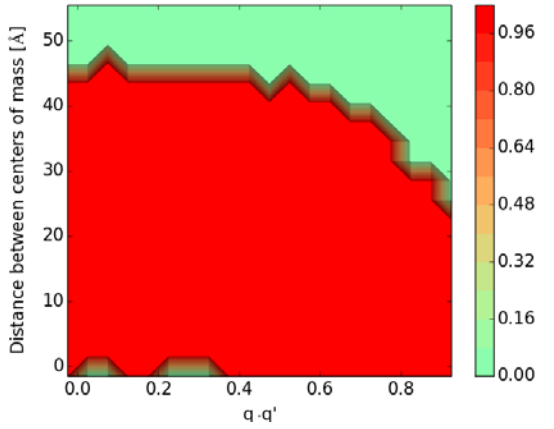
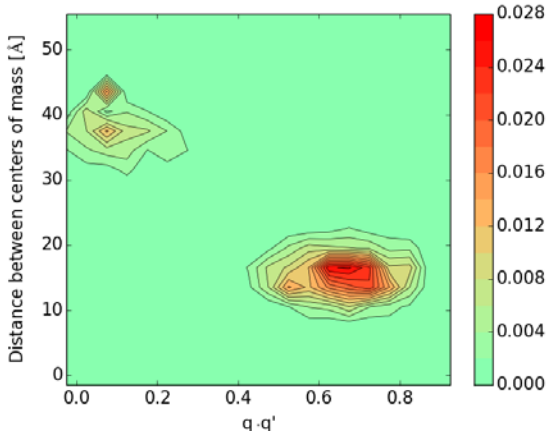


Figure S7: The two site exchange simulation: normalized (see text) 2D histograms of the number of structures as function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and a reference a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =17%.

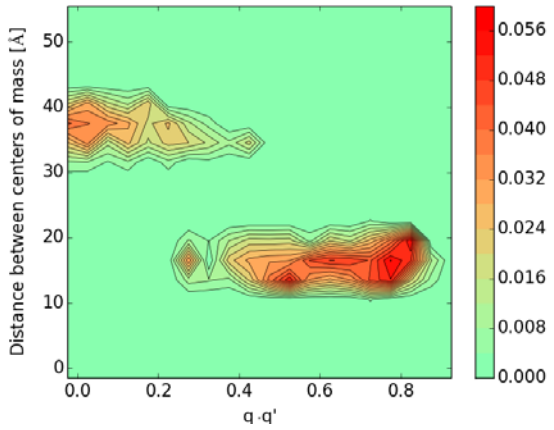
a.



b.



c.



d.

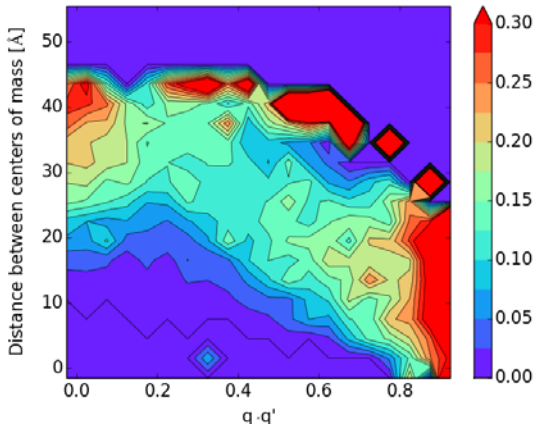
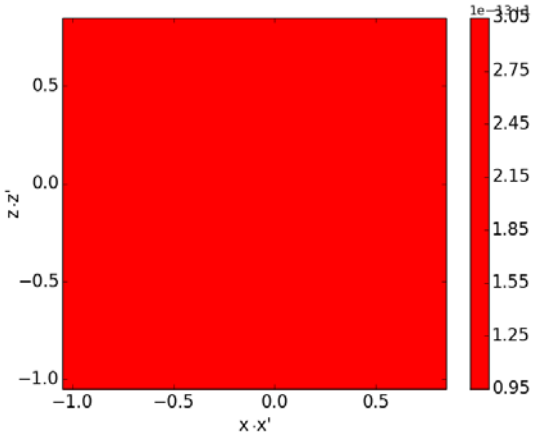
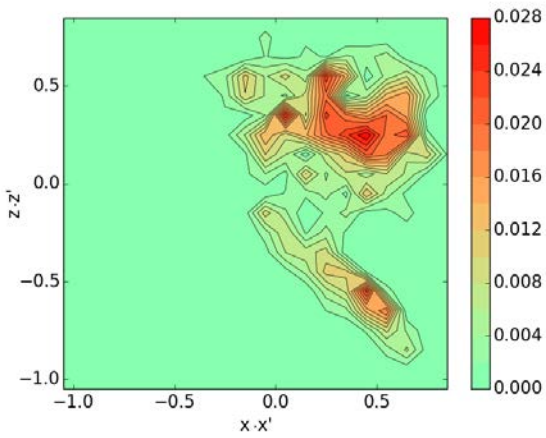


Figure S8: The two site exchange simulation: normalized (see text) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =17%.

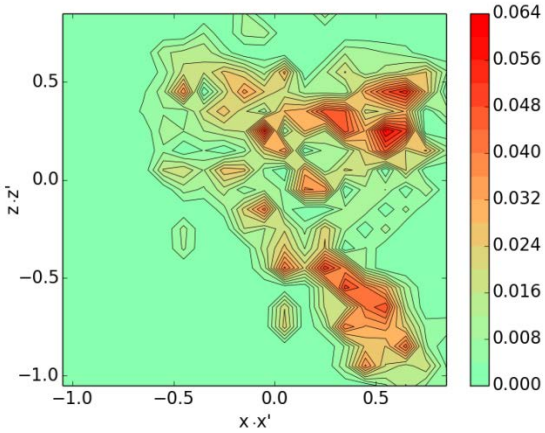
a.



b.



c.



d.

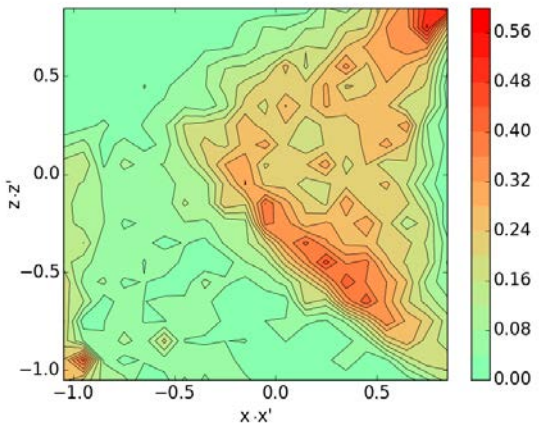
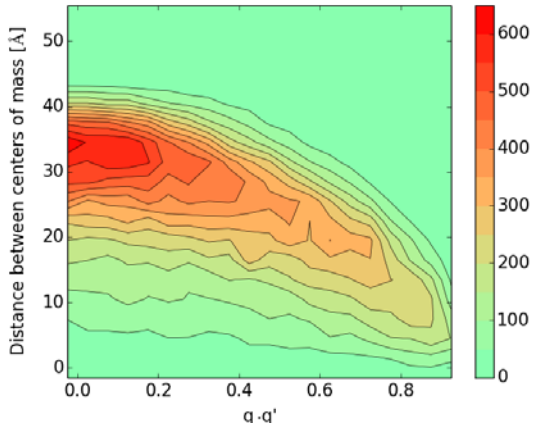
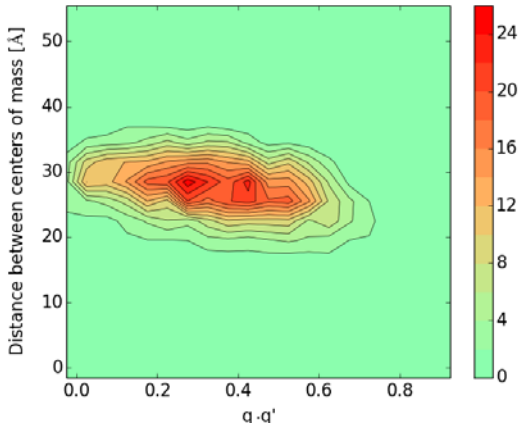


Figure S9: The single center mobility simulation: (a-d) 2D histograms of the number of structures as a function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and a reference structure a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minxOR =21% (e) highest MaxOcc values among all structures sharing the same distance and rotation.

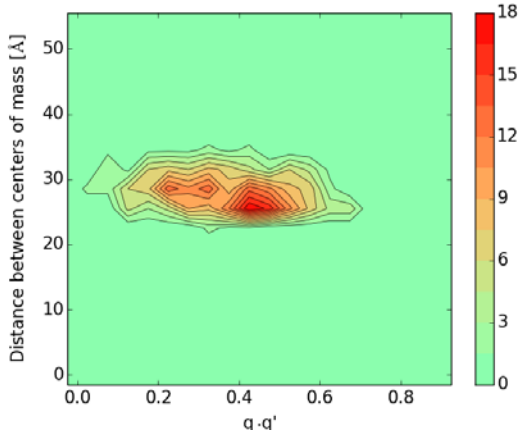
a.



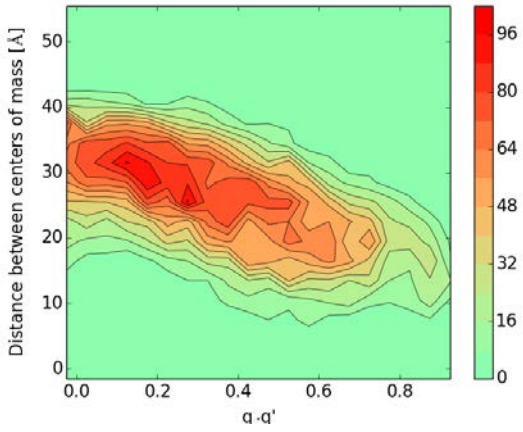
b.



c.



d.



e.

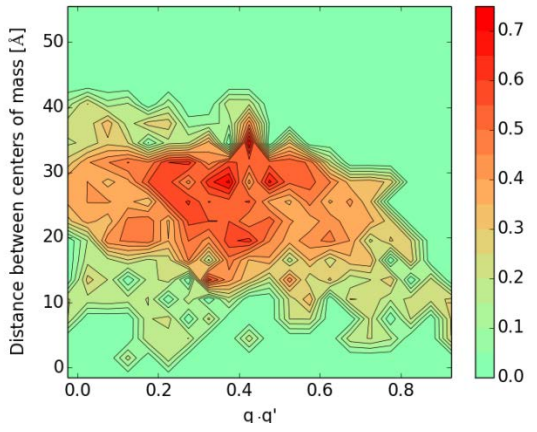
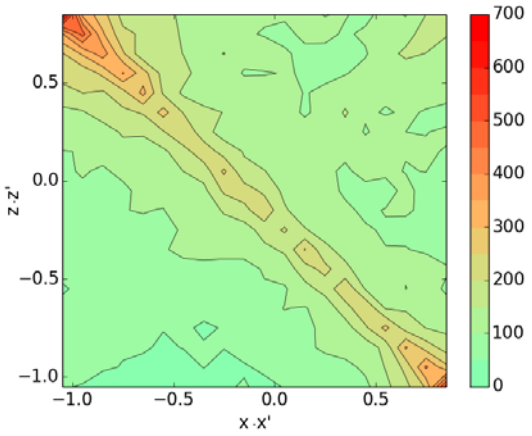
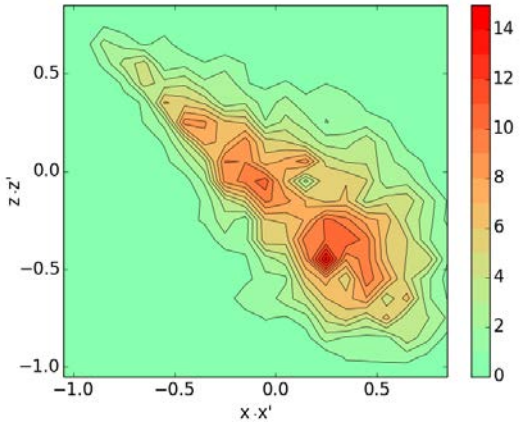


Figure S10: The single center mobility simulation: (a-d) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =21% (e) highest MaxOcc values among all structures sharing the same angles between x-axes and z-axes.

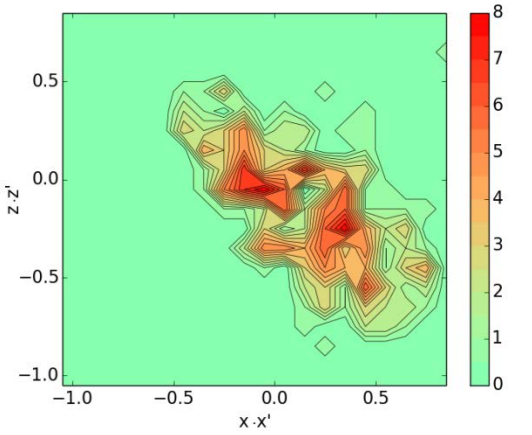
a.



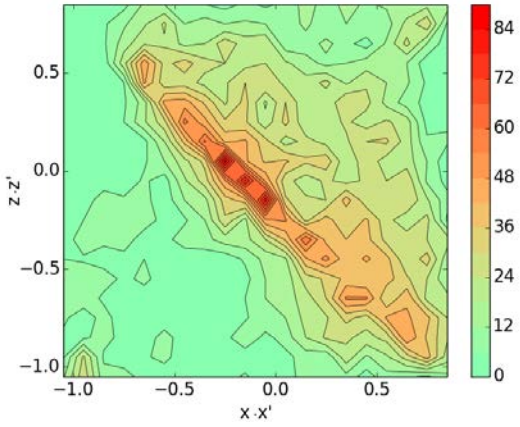
b.



c.



d.



e.

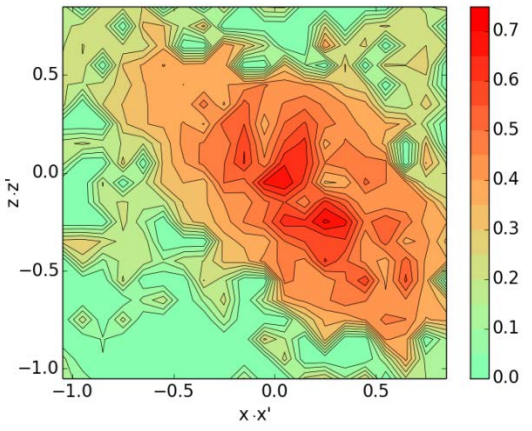
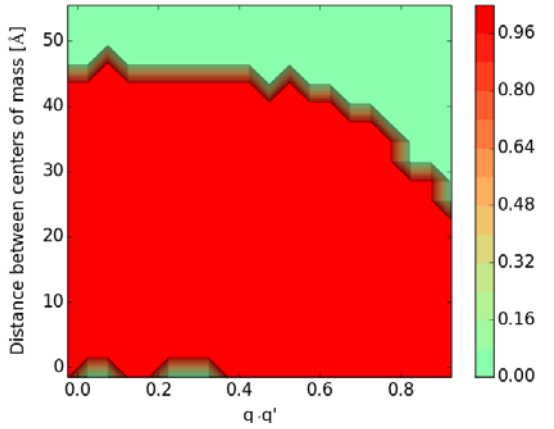
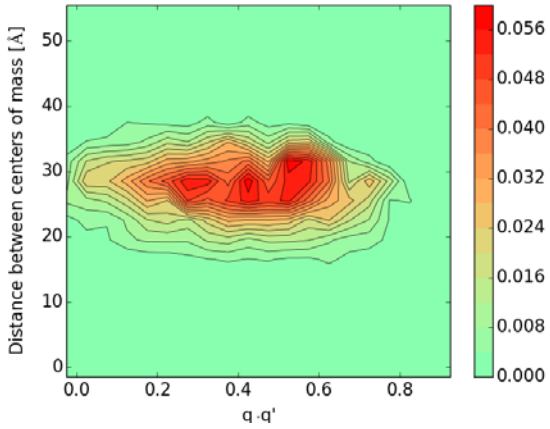


Figure S11: The single center mobility simulation: normalized (see text) 2D histograms of the number of structures as a function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and a reference structure a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =21%.

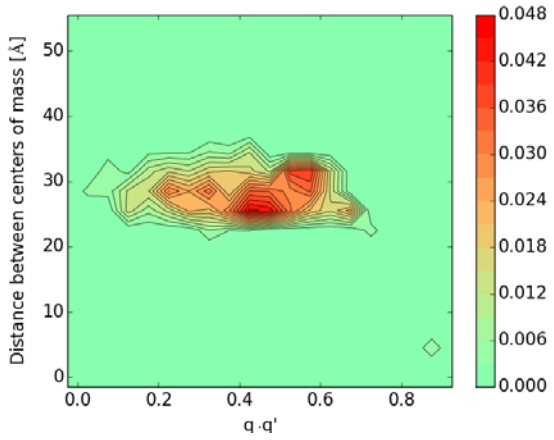
a.



b.



c.



d.

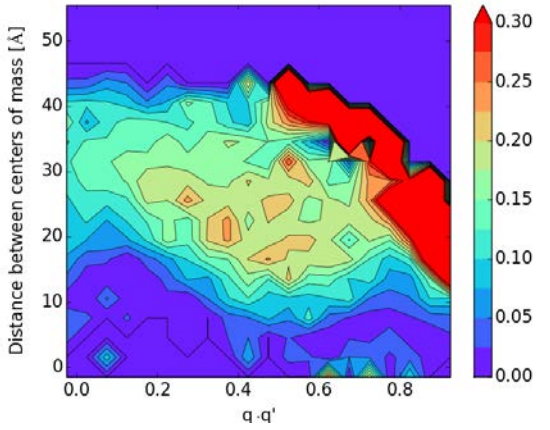
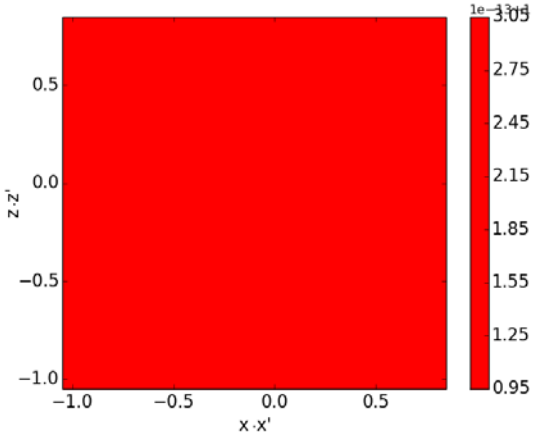
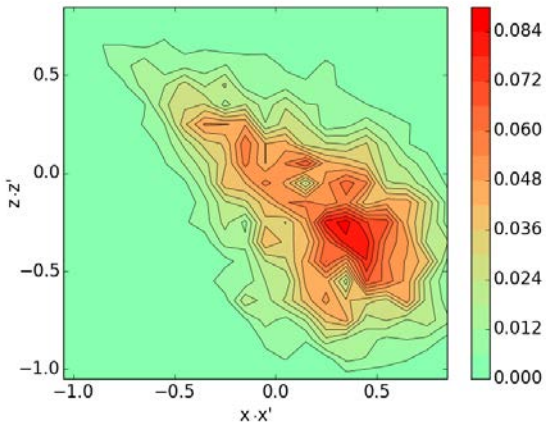


Figure S12: The single center mobility simulation: normalized (see text) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =21%.

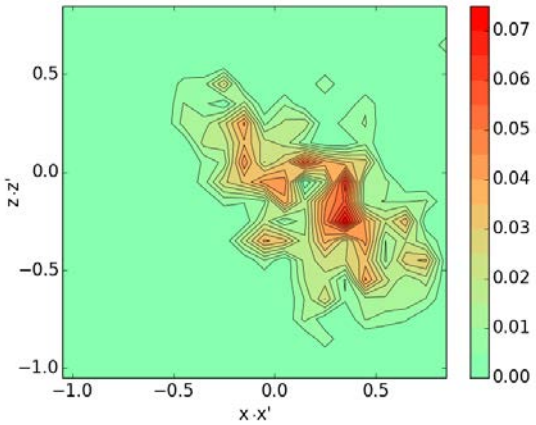
a.



b.



c.



d.

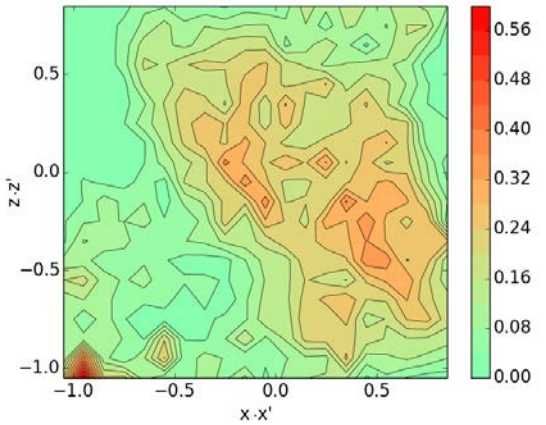
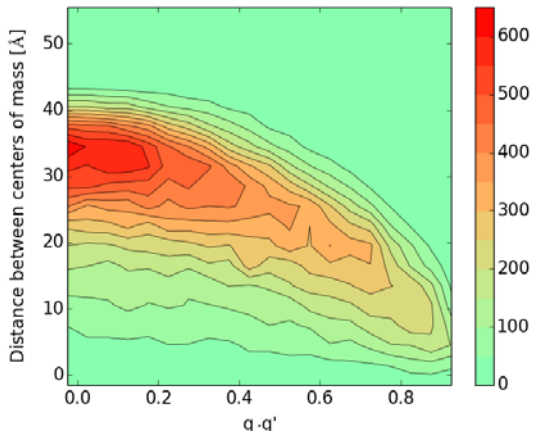
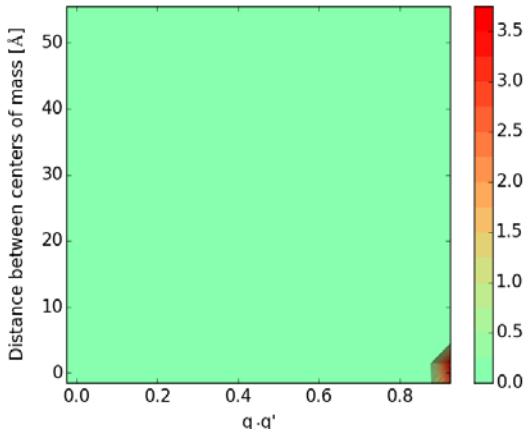


Figure S13: The very limited mobility simulation: (a-d) 2D histograms of the number of structures as a function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and a reference structure a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =54% (e) highest MaxOcc values among all structures sharing the same distance and rotation.

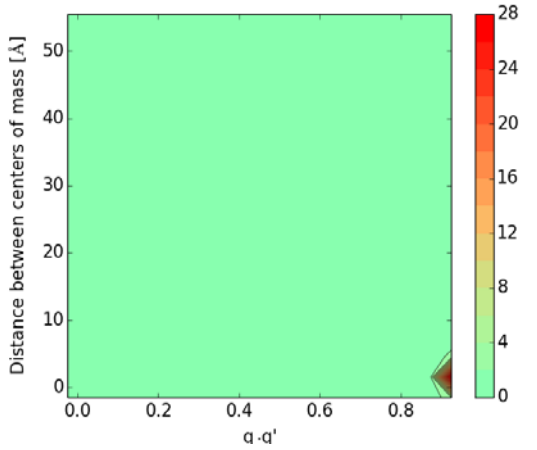
a.



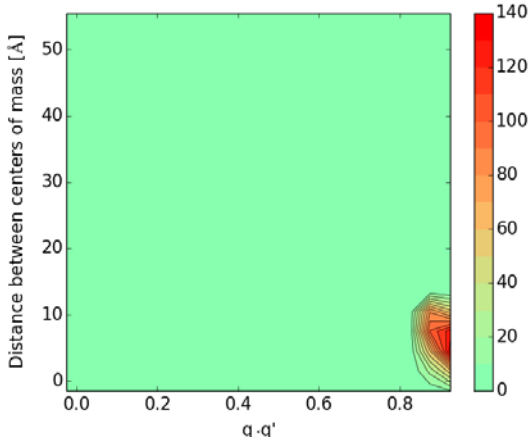
b.



c.



d.



e.

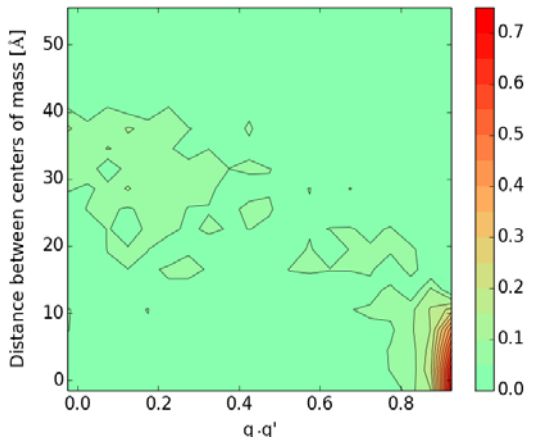


Figure S14: The very limited mobility simulation: (a-d) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =54% (e) highest MaxOcc values among all structures sharing the same angles between x-axes and z-axes.

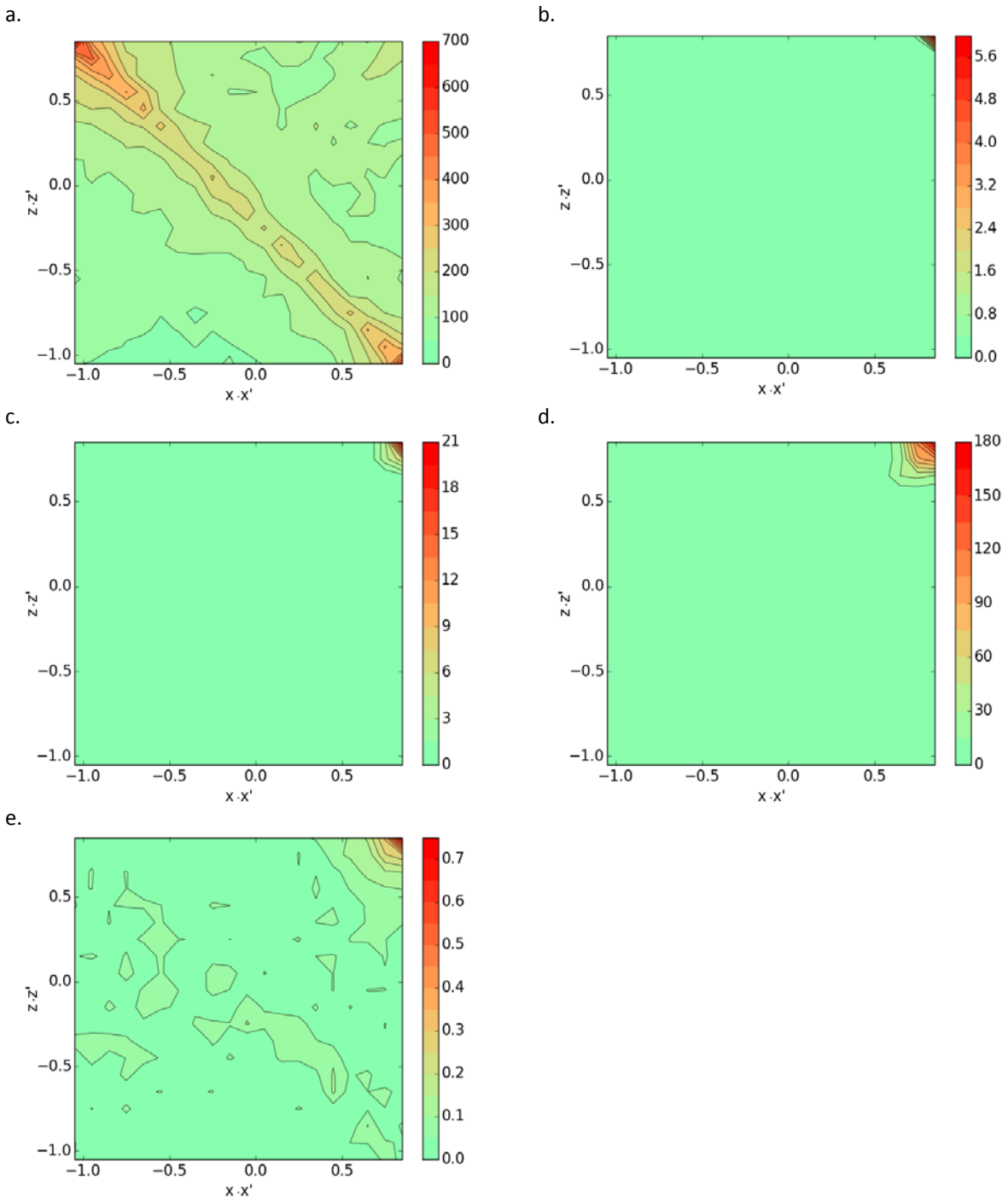
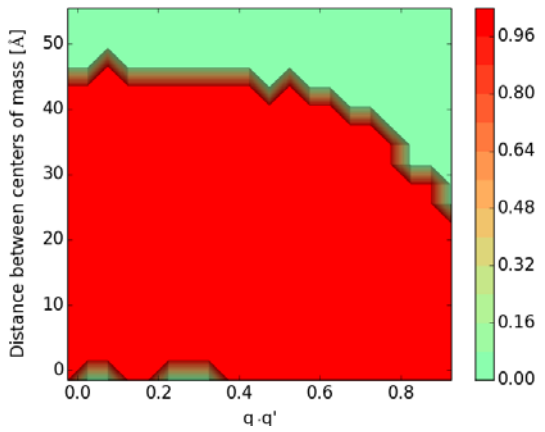
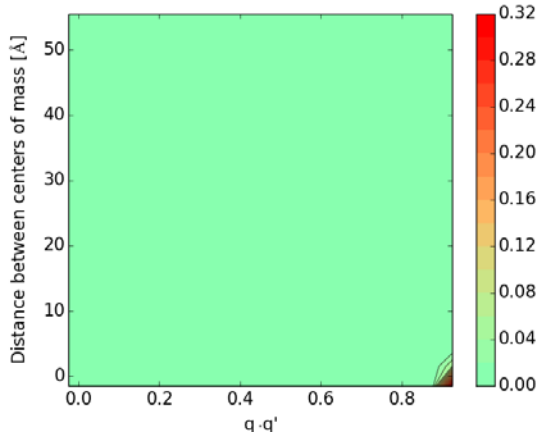


Figure S15: The very limited mobility simulation: normalized (see text) 2D histograms of the number of structures as a function of the distance from a reference structure and the angle between the quaternions describing the orientations of the structure and a reference structure a) for all the structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =54%.

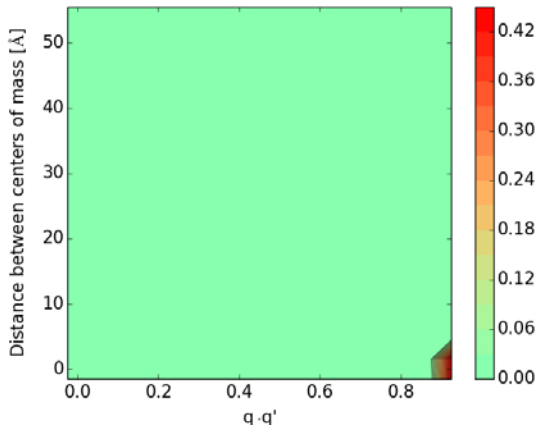
a.



b.



c.



d.

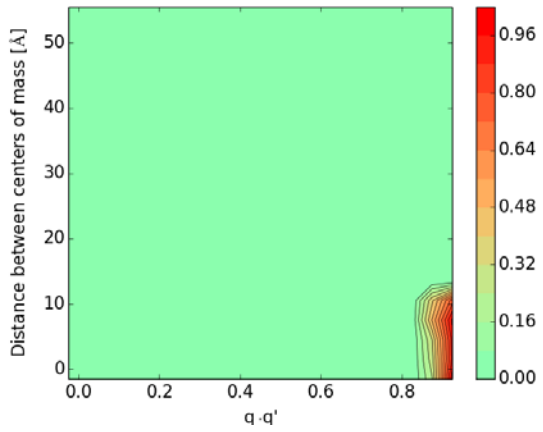
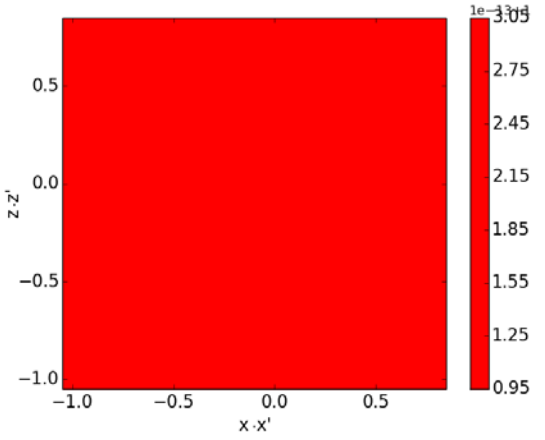
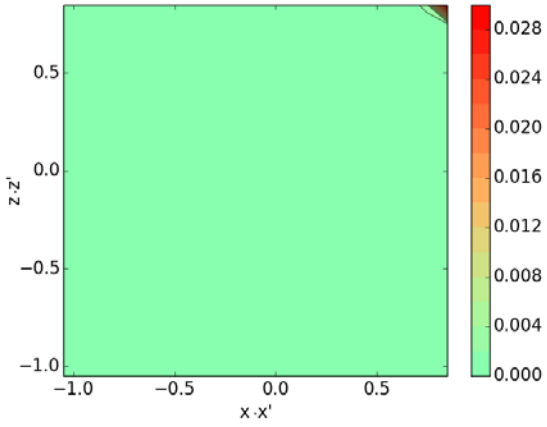


Figure S16: The very limited mobility simulation: normalized (see text) 2D histograms of the number of structures as a function of the angles between the x-axes and z-axes of the principal frames of the structure and of a reference a) for all structures b) for the simulated ensemble c) for the smallest ensemble with maxOR =1 d) for the ensemble with minOR =54%.

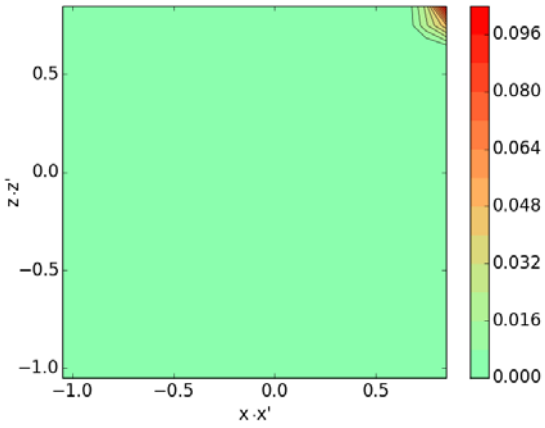
a.



b.



c.



d.

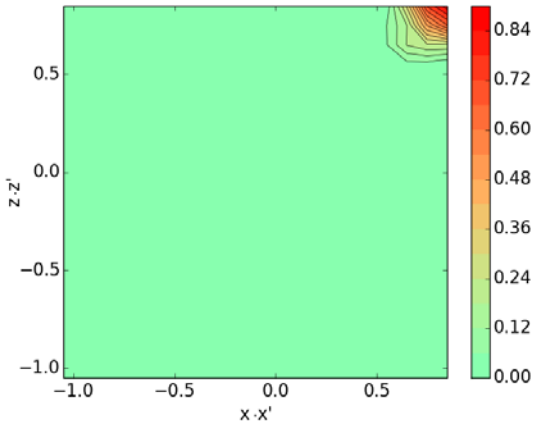


Figure S17: The regions with 17% of minOR in the two center exchange simulation, visualized in the Cartesian space.

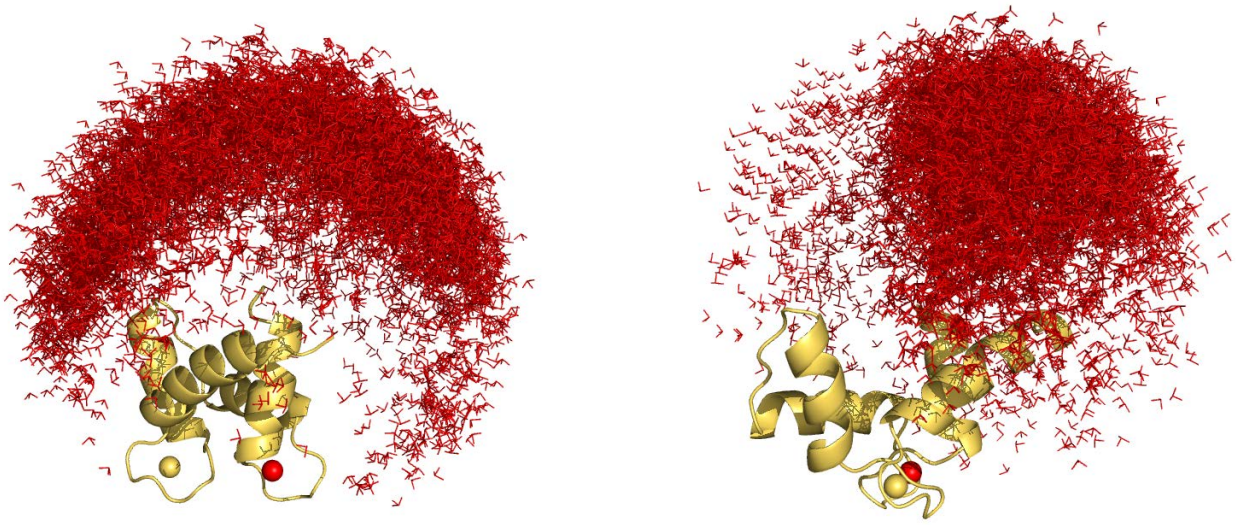
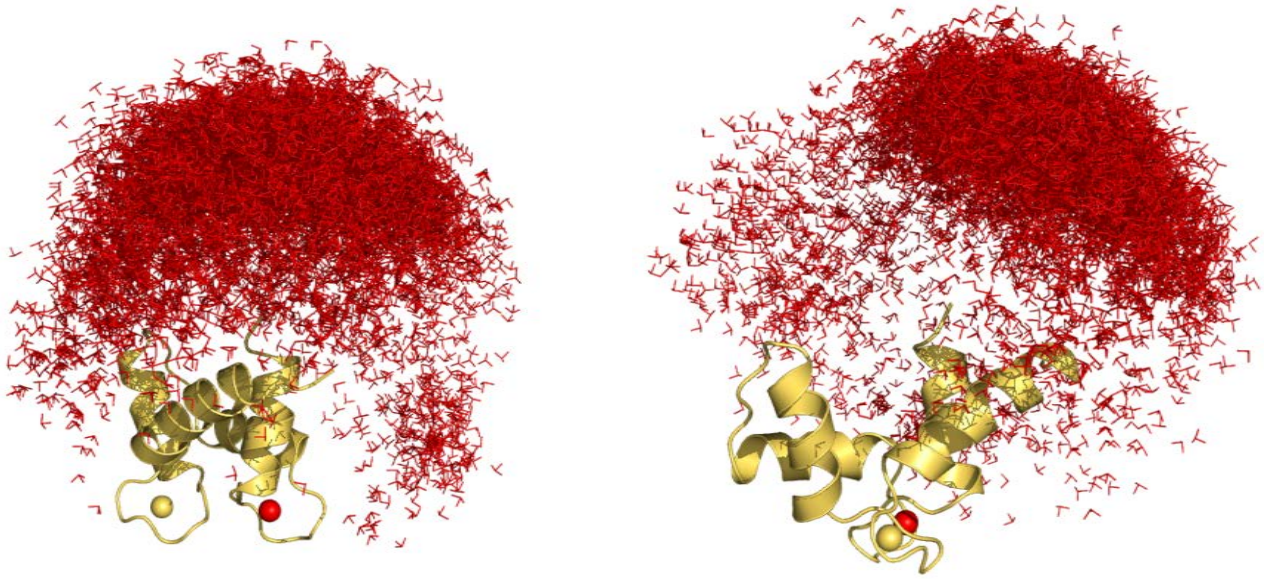


Figure S18: The regions with 21% of minOR in the two center mobility simulation, visualized in the Cartesian space.



Reference List

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

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

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

This page intentionally left blank

3.2 Information content of long-range NMR data for the characterization of conformational heterogeneity

Witold Andrałojć,[a] Konstantin Berlin,[b] David Fushman,[b] Claudio Luchinat,[a,c] Giacomo Parigi,[a,c] Enrico Ravera,[a,c] Luca Sgheri[d]

[a]Center for Magnetic Resonance (CERM), University of Florence, Via L. Sacconi 6, 50019, Sesto Fiorentino, Italy

[b]Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization, University of Maryland, College Park, MD 20742, USA

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

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

*J Biomol NMR, **2015**, 62(3), pp 353-371*

This page intentionally left blank

J Biomol NMR (2015) 62:353–371
DOI 10.1007/s10858-015-9951-6

ARTICLE

Information content of long-range NMR data for the characterization of conformational heterogeneity

Witold Andrałojć[1] · Konstantin Berlin[2] · David Fushman[2] · Claudio Luchinat[1,3] · Giacomo Parigi[1,3] · Enrico Ravera[1,3] · Luca Sgheri[4]

Received: 3 March 2015 / Accepted: 25 May 2015 / Published online: 5 June 2015

Abstract Long-range NMR data, namely residual dipolar couplings (RDCs) from external alignment and paramagnetic data, are becoming increasingly popular for the characterization of conformational heterogeneity of multidomain biomacromolecules and protein complexes. The question addressed here is how much information is contained in these averaged data. We have analyzed and compared the information content of conformationally averaged RDCs caused by steric alignment and of both RDCs and pseudocontact shifts caused by paramagnetic alignment, and found that, despite the substantial differences, they contain a similar amount of information. Furthermore, using several synthetic tests we find that both sets of data are equally good towards recovering the major state(s) in conformational distributions.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9951-6) contains supplementary material, which is available to authorized users.

✉ Claudio Luchinat
luchinat@cerm.unifi.it

David Fushman
fushman@umd.edu

[1] Center for Magnetic Resonance (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[2] Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization, University of Maryland, College Park, MD 20742, USA

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

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

Keywords Paramagnetic NMR · Residual dipolar couplings · Two-domain proteins · Protein mobility · Conformational variability

Introduction

Biological macromolecules are inherently flexible objects and often accomplish their task through extensive conformational rearrangement (Sicheri and Kuriyan 1997; Pickford and Campbell 2004; Zhang and Zuiderweg 2004; Tonks 2006; Chuang et al. 2010). Characterization of such rearrangements and the relevant conformational states can provide important clues about the mechanisms underlying biological function. This however is a challenging task because the system is underdetermined, implying a large degeneracy in the reconstructed solutions, and requires extensive experimental work often involving multiple techniques (Bonvin and Brunger 1996; Choy and Forman-Kay 2001; Svergun et al. 2001; Burgi et al. 2001; Clore and Schwieters 2004; Schroeder et al. 2004; Iwahara et al. 2004; Bertini et al. 2004a; Blackledge 2005; Lindorff-Larsen et al. 2005; Fragai et al. 2006; Tolman and Ruan 2006; Boehr et al. 2006; Ryabov and Fushman 2006; Chen et al. 2007; Bernadò et al. 2007; Bertini et al. 2007; Ryabov and Fushman 2007; Lange et al. 2008; Hulsker et al. 2008; Korzhnev and Kay 2008; Nodet et al. 2009; Boehr et al. 2009; Stelzer et al. 2009; Huang and Grzesiek 2010; Fisher et al. 2010; Bashir et al. 2010; Rinnenthal et al. 2011; Bothe et al. 2011; Fisher and Stultz 2011; Berlin et al. 2013; Russo et al. 2013; Guerry et al. 2013; Kukic et al. 2014; Ravera et al. 2014; Torchia 2015). Therefore, it is important to know the information content provided by various experimental methods in order to decide on an optimal set of experiments a priori.

Residual dipolar couplings (RDC; Lohman and Maclean 1978) are widely used as a source of information on biomolecular structure and dynamics (Tolman 2001; Tolman and Ruan 2006; Berlin et al. 2013; Ravera et al. 2014). They arise in the presence of partial molecular orientation, which can be achieved by interactions with alignment media surrounding the molecule (Tolman et al. 1995; Tjandra and Bax 1997; Hansen et al. 1998; Losonczi and Prestegard 1998; Ramirez and Bax 1998; Wang et al. 1998; Al-Hashimi et al. 2000; Prestegard et al. 2000; Zweckstetter and Bax 2001; Lakomek et al. 2008) and/or by the preferential orientation of the molecule itself in a magnetic field due to its magnetic susceptibility anisotropy (Lohman and Maclean 1978; Tolman et al. 1995; Zhang et al. 2003; Latham et al. 2008; Ravera et al. 2014; Musiani et al. 2014). RDCs obtained by alignment induced by an external orienting medium, herein referred to as diamagnetic RDCs (dRDC), depend on the nature of the interactions of the biomolecule with the medium. These interactions can be steric and/or electrostatic and, because of this, dRDC are reporters also on the overall shape of the macromolecule and/or its charge distribution (Zweckstetter and Bax 2000; Zweckstetter 2008; Berlin et al. 2009; Maltsev et al. 2014). On the other hand, RDCs caused by molecular self-alignment, often induced by the presence of a paramagnetic center with an anisotropic magnetic susceptibility, herein termed paramagnetic RDCs (pRDC), only depend on the orientation of the internuclear vectors in the reference frame of the magnetic susceptibility tensor and are generally independent of the shape of the molecule. However, the presence of an anisotropic magnetic susceptibility also gives rise to pseudocontact shifts (PCS; Kurland and McGarvey 1970), which are reporters on the positions of the nuclei in the principal axis frame of the magnetic susceptibility tensor centered on the paramagnetic site, and therefore contain information about the structure/shape of a molecule. The use of paramagnetism-induced restraints (Gochin and Roder 1995a, b; Banci et al. 1996, 1998; Bertini et al. 2001a; Gaponenko et al. 2004; Bertini et al. 2005; Diaz-Moreno et al. 2005; Jensen et al. 2006; Bertini et al. 2008; Schmitz et al. 2012; Yagi et al. 2013b) is becoming increasingly popular because of the introduction of lanthanide binding tags (Barthelmes et al. 2011; Wöhnert et al. 2003; Rodriguez-Castañeda et al. 2006; Su et al. 2006; John and Otting 2007; Pintacuda et al. 2007; Zhuang et al. 2008; Su et al. 2008a, b; Keizers et al. 2008; Häussinger et al. 2009; Su and Otting 2010; Hass et al. 2010; Man et al. 2010; Das Gupta et al. 2011; Saio et al. 2011; Swarbrick et al. 2011a, b; Bertini et al. 2012a; Liu et al. 2012; Kobashigawa et al. 2012; Cerofolini et al. 2013; Yagi et al. 2013a; Gempf et al. 2013; Loh et al. 2013), that extend the range of applications from paramagnetic metalloproteins (Banci et al. 1996, 1997; or proteins in which the naturally occurring metal can be replaced by a paramagnetic one; Allegrozzi et al. 2000; Bertini et al. 2001a, b, c; Bertini et al. 2003, 2004b; Balayssac et al. 2008; Bertini et al. 2010a; Luchinat et al. 2012b) to, in principle, any protein.

Given the various possibilities and limited resources, choosing the optimal set of observables for the characterization of protein conformational heterogeneity is important. In this work we analyze the information content associated with the two commonly used types of experimental data (dRDC and paramagnetic data) and discuss their features and advantages and pitfalls. Specifically, we want to understand what information can be recovered and to what extent. Importantly, the methodology that we develop below is not limited to dRDC or paramagnetic data, and can be applied to any set of experimental observables.

Theory

Formulation of the ensemble problem

We focus on analyzing the ensemble information content of three specific types of NMR restraints, dRDC, pRDC, and PCS, in the case of proteins composed of two domains connected by a flexible linker. We have used the two-domain protein calmodulin (CaM) as a test case.

As done previously (Bertini et al. 2007; Berlin et al. 2013), we assume that all three types of NMR restraints considered here represent a population-weighted average of the corresponding values for the individual conformers, and therefore have a linear dependence on the ensemble populations, such that

$$\mathbf{y} = \mathbf{a}_1 x_1 + \cdots + \mathbf{a}_N x_N + \varepsilon = \mathbf{A}\mathbf{x} + \varepsilon \quad (1)$$

where $\mathbf{y}$ is a length-L column vector representing the experimental data (dRDC, pRDC, PCS, or some combination thereof), $\mathbf{A}$ is an $L \times N$ prediction matrix consisting of N column-vectors $\mathbf{a}_j$ ($j = 1,\ldots,N$) representing the predicted data for each of the N conformers, x_j is the population weight for the jth conformer, and ε is the difference between $\mathbf{y}$ and $\mathbf{Ax}$ due to the presence of experimental error. This assumption seems reasonable for pRDC and PCS (Bertini et al. 2012c), whereas for dRDC the interconversion between conformers can occur on a timescale that could be comparable to the one of the interaction with the alignment medium; additionally, the latter may perturb the system.

Since in general recovering $\mathbf{x}$ from Eq. 1 is an ill-posed problem, having an infinite number of solutions, we seek to recover the minimum ensemble (sparsest solution) satisfying the experimental observables, which we express as a

constrained linear least-squares problem (Berlin et al. 2013),

$$\mathbf{x}^* = \arg\min_{\mathrm{x}} \|\mathbf{W}(\mathbf{A}\mathbf{x} - \mathbf{y})\|_2 \text{ s.t. } \mathbf{x} \geq 0,\ \|\mathbf{x}\|_0 = M \quad (2)$$

where $\mathbf{W}$ is the weight matrix that non-uniformly weighs the residuals between $\mathbf{y}$ and $\mathbf{Ax}$, M is the desired ensemble size, $\|\ldots\|_2$ is the Euclidian norm, and $\|\mathbf{x}\|_0$ is the l_0 quasi-norm of $\mathbf{x}$, i.e. the number of nonzero elements in $\mathbf{x}$. Typically the experimental errors are assumed to be uncorrelated, in which case $\mathbf{W}$ is simply a diagonal matrix with $\mathbf{W}_{\mathrm{ii}} = 1/\sigma_{\mathrm{i}}$, where σ_{i} is the estimated experimental error of the ith observation y_i. For simplicity, for the rest of the manuscript we will drop $\mathbf{W}$ from our equations by assuming that $\mathbf{A}$ and $\mathbf{y}$ are already multiplied by $\mathbf{W}$. In the sparse ensemble selection (SES) method the ensemble size is chosen by solving the problem for reasonable values of M and using the L-curve to select the appropriate M value (Berlin et al. 2013). A different approach was also applied, based on the calculation of the maximum occurrence allowed for each conformer (MaxOcc, see below; Bertini et al. 2002a; Gardner et al. 2005; Longinetti et al. 2006; Bertini et al. 2007, 2010b, 2012b, c; Luchinat et al. 2012a; Andralojc et al. 2014).

Predicting RDC and PCS data

For steric dRDC data, we generate the prediction matrix $\mathbf{A}$ using program PATI (Berlin et al. 2009, 2013), which assumes the presence of a steric planar alignment medium (Fig. 1a). Electrostatically induced RDCs were similarly simulated using PALES (Zweckstetter 2008). The absolute scaling in the predicted dRDC values is regulated by changing the value of the parameter "liquid crystal concentration" (Zweckstetter and Bax 2000) that controls the distance between the planar steric barriers. In the SES model the absolute scaling of the predicted dRDC is treated as an implicit parameter since the sum of all weights ($\sum_j x_j$) is not constrained (Berlin et al. 2013).

For pRDC and PCS, without loss of generality, we can assume that a metal ion tag is located on the first (rigid) domain of the protein (Bertini et al. 2003). Therefore, the position of the metal ion relative to that domain is the same for all conformers. So, instead of performing the prediction of pRDC and PCS values for both domains, we obtain the prediction matrix $\mathbf{A}$ for a two-domain rigid system by first deriving the magnetic susceptibility anisotropy tensor (and metal ion's position) from the experimental data for the first domain, and then use these tensors to predict the matrix $\mathbf{A}$ values for the second domain based on its position relative to the first domain (Fig. 1b). This formulation assumes that the distribution of the relative positions of the two domains is independent of the orientation of the magnetic susceptibility anisotropy tensor in the magnetic field (Bertini et al. 2002a).

Given a specific conformer, the pRDC values in the $\mathbf{A}$ matrix are thus predicted by first deriving the vector containing the 5 independent components of the alignment tensor, $\mathbf{S}^*$, directly from the experimental data for the first domain:

$$\mathbf{S}^* = \arg\min_{\mathrm{S}} \|\mathbf{V}_1\mathbf{S} - \mathbf{y}_1\|_2 \quad (3)$$

where $\mathbf{V}_1$ is a 5-column matrix, the elements of which depend on the orientations of the normalized bond vectors in the fixed frame (Losonczi et al. 1999; Valafar and Prestegard 2004; Berlin et al. 2009; Simin et al. 2014) and $\mathbf{y}_1$ are the observed experimental pRDC values for the first domain. Then, using the derived $\mathbf{S}^*$, we predict the pRDC for the second domain of the jth conformer ($A_{pRDC,j}$) as

$$\mathbf{A}_{pRDC,j} = \mathbf{V}_{2j}\mathbf{S}^* \quad (4)$$

where $\mathbf{V}_{2j}$ is the 5-column matrix of the bond vectors for the second domain in the jth conformer.

Similarly, the PCS values for the first domain can be used to derive the magnetic susceptibility anisotropy tensor $\mathbf{T}^*$, represented by a 3×3 traceless symmetric matrix, and the metal ion's position $\mathbf{p}^*$ (computed by alternating between solving a non-linear least-squares problem for $\mathbf{p}^*$, and a linear problem for $\mathbf{T}^*$). These values are then used to predict the PCS for the second domain of the jth conformer ($\mathbf{A}_{PCS,j}$). The elements of the $\mathbf{A}_{PCS,j}$ vector are the PCSs predicted for each nucleus i of the second domain, according to the relationship

$$\mathbf{A}_{PCS,j,i} \approx \frac{1}{12\pi\|\mathbf{r}_{ij}\|_2^5}\mathrm{tr}\left(\begin{bmatrix} 3r_{ij,1}^2 - \|\mathbf{r}_{ij}\|_2^2 & 3r_{ij,1}r_{ij,2} & 3r_{ij,1}r_{ij,3} \\ 3r_{ij,1}r_{ij,2} & 3r_{ij,2}^2 - \|\mathbf{r}_{ij}\|_2^2 & 3r_{ij,2}r_{ij,3} \\ 3r_{ij,1}r_{ij,3} & 3r_{ij,2}r_{ij,3} & 3r_{ij,3}^2 - \|\mathbf{r}_{ij}\|_2^2 \end{bmatrix}\mathbf{T}^*\right) \quad (5)$$

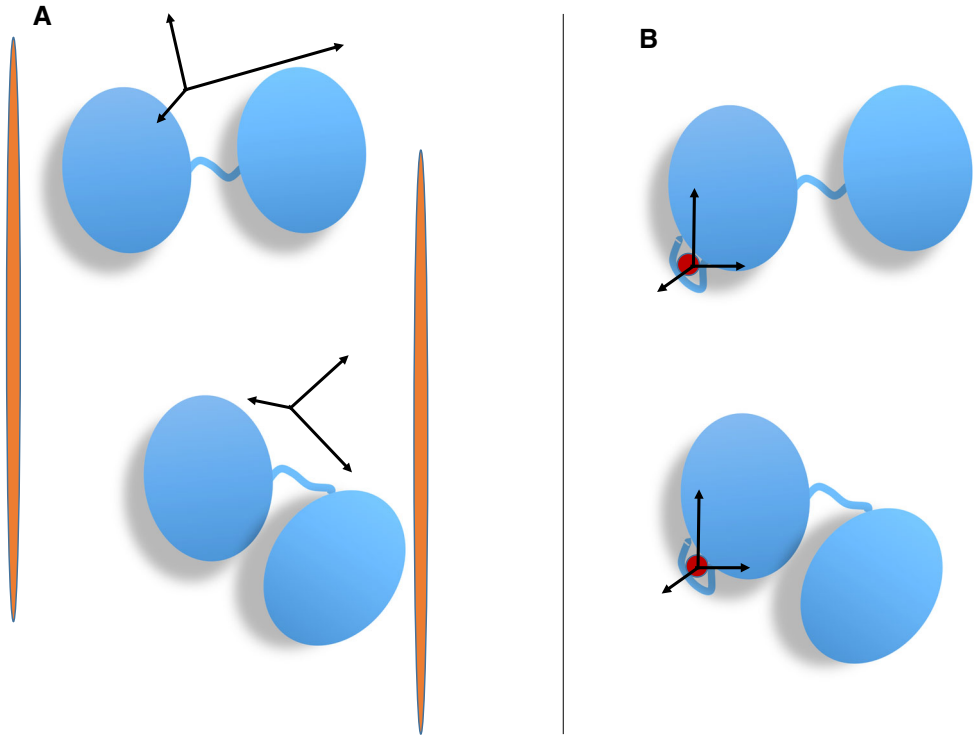


Fig. 1 Schematic illustration of the relationship between the conformation of a multidomain protein and the alignment tensor for the two experimental methods considered her (or: alignment tensor caused by external and internal alignment). In the case of partial orientation induced by external orienting media the alignment tensor changes for different conformations of a two-domain protein (**a**) whereas in the case of partial orientation induced by a paramagnetic metal ion attached to the protein the alignment tensor is invariant with respect to the orientation of the domain where it is attached (**b**)

where $\mathbf{r}_{ij} = [r_{ij,1}, r_{ij,2}, r_{ij,3}]$ is the vector connecting the metal ion (located at $\mathbf{p}^*$) and the ith atom in the jth conformer, and tr(…) designates the trace of a matrix. The elements of the tensor $\mathbf{T}^*$ and the components of the alignment tensor $\mathbf{S}^*$ are related to one another by a proportionality constant (Bertini et al. 2002b), so that each of the two can be easily calculated from the other.

Similarly to dRDC for multiple alignment media, pRDC and PCS from multiple metal ion derivatives (determined from the $\mathbf{S}^*$ and $\mathbf{T}^*$ tensors, respectively, of the corresponding metals) can be combined together in a single $\mathbf{A}$ matrix of predicted data.

Methods

Constraining SES ensemble populations

Since the scaling of the predicted dRDC values has an uncertainty (Berlin et al. 2013), when recovering SES ensembles using dRDC, we allow the total sum of $\mathbf{x}$, $\sum_j x_j$, to float, and only use the restraint $\mathbf{x} \geq 0$ (see Eq. 2).

By contrast, the values of pRDC and PCS are determined without any adjustable scaling factor, and thus the two datasets can be directly combined into a single population-constrained pRDC + PCS SES problem,

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \left\| \begin{bmatrix} \mathbf{y}_{pRDC} \\ \mathbf{y}_{PCS} \end{bmatrix} - \begin{bmatrix} \mathbf{A}_{pRDC} \\ \mathbf{A}_{PCS} \end{bmatrix} \mathbf{x} \right\|_2 \text{ s.t. } \mathbf{x} \geq 0, \quad \sum_j x_j \leq c, \|\mathbf{x}\|_0 = M \tag{6}$$

where c is the upper bound on the total population weight. Since $\sum_j x_j$ represents the total population weight $\sum_j x_j$ should be 1. However, we allow for the sum of the weights to be <1, since we aim at recovering the sparsest ensemble representing the major states (potentially there could be a very large set of transient minor states). The validity of the recovered solution can be evaluated from the geometrical interpretation of pRDC: a solution is a convex combination of a set of conformers such that the averaged pRDC belong to the polyhedron with vertices in the conformers (see Figure S5; Gardner et al. 2005; Longinetti et al. 2006). Since the problem is underdetermined, there will be many solutions, and the SES method chooses to limit the number of vertices to M. In order to find a solution with this constraint, we need to use a $c < 1$ in Eq. (6). This is equivalent to shrinking the vertices of the polyhedron towards the origin by a factor c and renormalizing the weighting factors

to 1. However, since the origin is an acceptable point (Sgheri 2010a) and the set is convex, the shrunk vertices will be anyway acceptable points. In other words, if c is relatively close to 1, the conformers representing the vertices are anyway good representatives of the conformational freedom of the system. Finally, the $\sum_j x_j \leq 1$ restraint prevents from finding unphysical solutions.

SES algorithm implementation

SES ensemble recovery was implemented using the multi-orthogonal matching pursuit (MOMP) algorithm (Berlin et al. 2013). We modified the MOMP method to handle the $\sum_j x_j \leq c$ requirement using the active set method (O'Leary 2009) to restrain our solution for each iteration of MOMP. Given that there are two restraints on $\mathbf{x}$: $\mathbf{x} \geq 0$ and $\sum_j x_j \leq c$, during each iteration of the MOMP algorithm there are four possible sets of active restraints: (1) no restraints are active; (2) $\sum_j x_j \leq c$ restraint is active; (3) the $\mathbf{x} \geq 0$ restraint is active; or (4) both $\mathbf{x} \geq 0$ and $\sum_j x_j - \leq c$ are active. To summarize, the constrained least-squares problem is solved as follows: update the solution using conjugate gradient (CG) method; if the solution violates $\mathbf{x} \geq 0$ or $\sum_j x_j \leq c$, solve the linearly constrained linear least-squares problem by using a "feasible direction" method (O'Leary 2009); if the solution still violates $\mathbf{x} \geq 0$, drop this solution from a list of possible solutions stored in a priority queue. This procedure is repeated for all propagated solutions from the previous iteration.

The time versus accuracy tradeoff in the MOMP algorithm is controlled by how many top solutions, K, from the current iteration are propagated to the next iteration of MOMP (Berlin et al. 2013). In order to improve the memory requirement for running SES using very large K values ($>10^6$), we modified the algorithm used to solve the overdetermined linear least-squares problem for each iteration of SES, when a new solution must be computed right after one new column is added to the list of active columns [see Supporting Information in Berlin et al. (2013)]. In the previous implementation (Berlin et al. 2013), the least-squares solution was efficiently updated by doing a rank-1 update of the QR decomposition. However, this approach requires us to store K QR decompositions during each iteration. In our current updated version, we switched to an iterative CG least-squares solver, which requires that we only store the previous-iteration solution, rather than the QR decomposition. This significantly reduced the SES memory footprint for large K. The full $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ matrix required for the CG algorithm is never explicitly formed, and instead the multiplication step in the CG algorithm is computed as $\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{x})$. With the CG implementation we are able to run SES on a 10 GB RAM desktop for $K = 10^6$, without any sacrifice in computational time or accuracy, as compared to the previous implementation.

MaxOcc calculations

The maximum occurrence (MaxOcc) of each and every conformer is defined as the maximum weight that it can obtain when part of a conformational ensemble without violating the constraints of the experimental data. No restriction is posed on the number of conformations to be included in the ensemble. Maximum occurrence (MaxOcc) can be interpreted as the maximum fraction of time that a conformation can exist, when taken together with any ensemble of conformations with optimized weights (Longinetti et al. 2006; Bertini et al. 2007; Sgheri 2010b; Bertini et al. 2010b; Das Gupta et al. 2011; Luchinat et al. 2012a; Bertini et al. 2012b, c; Cerofolini et al. 2013).

We formulate MaxOcc as a convex regularization problem, where for each conformer j we find the weight vector $\mathbf{x}$ which minimizes

$$\arg\min_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda (x_j - x_{MO})^2 + \lambda \left(1 - x_{MO} - \sum_{i=1, i\neq j}^{N} x_i \right)^2 \right\} \text{ s.t. } \mathbf{x} \geq 0 \quad (7)$$

where x_{MO} is the desired weight of the conformation j, and λ is a weighting factor. The calculations are repeated for increasing values of x_{MO}; the MaxOcc of conformation j is defined as the highest x_{MO} providing a value of the expression in Eq. 7 not exceeding the minimum value by more than a prefixed threshold, for example 20 %. The value of λ was fixed to 15, as found with the L-curve method, as a compromise between a good fit of the experimental observables and the proximity of the sum of the weights to 1. A frugal coordinate descent algorithm, combined with random coordinate search (Nesterov 2012), is used to solve Eq. 7.

Calculations are also performed to determine the maximum occurrence of a region (MaxOR) defined in the conformational space of the protein (Andralojc et al. 2014). The MaxOR, similar to MaxOcc, is defined as the maximum weight that a region in conformational space (composed of multiple structures) can have in an ensemble without causing a violation of the experimental restraints. First, the highest-MaxOcc structures are clustered according to their positions using a k-means algorithm as implemented in the Python library SciPy (Jones et al. 2001). The number of clusters is set to the highest value yielding reproducible clustering by the algorithm. Once the clusters are built, small regions are defined around the centers of the clusters, which include all conformations

within a given distance Δ from the center of the cluster. The MaxORs of these regions are determined by solving

$$\arg\min_{\mathbf{x}} \left\{ \|\mathbf{Ax}-\mathbf{y}\|_2^2+\lambda \left[\left(x_{MO}-\sum_{i\in C} x_i \right)^2 - \left(1-x_{MO}-\sum_{i\in D} x_i \right)^2 \right] \right\} \text{ s.t. } \mathbf{x}\geq 0 \quad (8)$$

where x_{MO} is the fixed value that must correspond to the sum of the weights of all conformations within the region, and C and D indicate the structures within and outside that region, respectively. Again, the largest x_{MO} providing a good fit of the experimental data defines the maxOR of the region.

Results and discussion

An important theoretical question that we would like to answer a priori, before performing any time-consuming simulation or experiment, is how much information for ensemble recovery is contained in dRDC versus pRDC versus PCS and in dRDC versus pRDC + PCS combined. For example, intuitively, dRDC should contain more information than pRDC, since dRDC contain shape/size-related information, while the relative informational content of PCS is harder to intuitively quantify. To what extent combining pRDC with PCS yields better results than each of these data separately? Is the information provided by pRDC + PCS similar to that provided by dRDC? Would using several different metal ions be needed to obtain results comparable to those obtained with multiple sets of dRDC, or do they produce a better set of experimental data for the characterization of the conformational heterogeneity?

In order to answer these questions, we analyzed several algebraic properties of eight experimentally feasible datasets: (1) single-alignment medium dRDC; (2) single-metal ion pRDC; (3) single-metal ion PCS; (4) single-metal ion pRDC + PCS combined; and (5–8) datasets analogous to (1–4) but with three alignment media or thee metal ions. We will refer to the one and three media/metal ions datasets as the one- and three-restraint datasets, respectively.

The datasets were generated for a pool of 32723 conformers of calmodulin (CaM), a protein composed of two rigid domains connected by a 4-residue flexible linker (Barbato et al. 1992; Tjandra et al. 1995; Chou et al. 2001; Kukic et al. 2014). This large pool of sterically allowed conformations of the protein was taken from reference (Bertini et al. 2010b), where it was generated using the program RanCh (Bernadò et al. 2007), For each conformer and for each aligning medium or metal ion, a set of dRDCs, pRDCs, and PCSs was generated, as described in the "Theory" section.

Simulated PCS and pRDC data

The paramagnetic restraints consisted of PCS of the amide H atoms and pRDC of amide N–H pairs of the C-terminal domain of CaM induced by the presence of a paramagnetic center in its N-terminal domain. Three metals with non-coinciding magnetic susceptibility tensors (corresponding to the experimental ones obtained for Tb(III), Tm(III), and Dy(III) CaM) were used to generate three sets of PCSs (132 observations in total) and pRDCs (112 observations in total). The magnetic susceptibility anisotropy tensors were taken from reference (Bertini et al. 2009).

Simulated dRDC data

The simulated diamagnetic restraints were amide ^{15}N-^{1}H dRDCs (219 in total) induced in both CaM domains by 3 independent external alignment media: flat uncharged discs and either positively or negatively charged rods. In the first case, dRDCs were generated using PATI (Berlin et al. 2009), in the other cases using PALES (Zweckstetter and Bax 2000; Zweckstetter 2008). In both cases, the calculation of the alignment tensors, and of the corresponding dRDC, are performed under the assumption that the protein's conformations are rigid during the time course of its interaction with the alignment medium. As a word of caution we note that every interaction of a protein with the alignment medium might actually perturb its conformation, and these interactions can occur on a timescale that is slower than the conformational averaging itself. The assumption that the averaged dRDCs correspond to a weighted average of the RDCs calculated for the individual conformations, although universally used, might fall short in representing the real physical picture.

SVD of prediction matrices

The first and simplest analysis we performed was aimed at evaluating the theoretical information content of the eight different datasets described above. This was done through the spectral analysis of the prediction matrix **A** for each dataset. The spectral analysis measures the number of significant linearly independent components present in the data, by counting the eigenvalues corresponding to linearly independent eigenvectors. This directly provides an upper bound on the number of independent conformers we can hope to extract. Trying to recover a larger number of

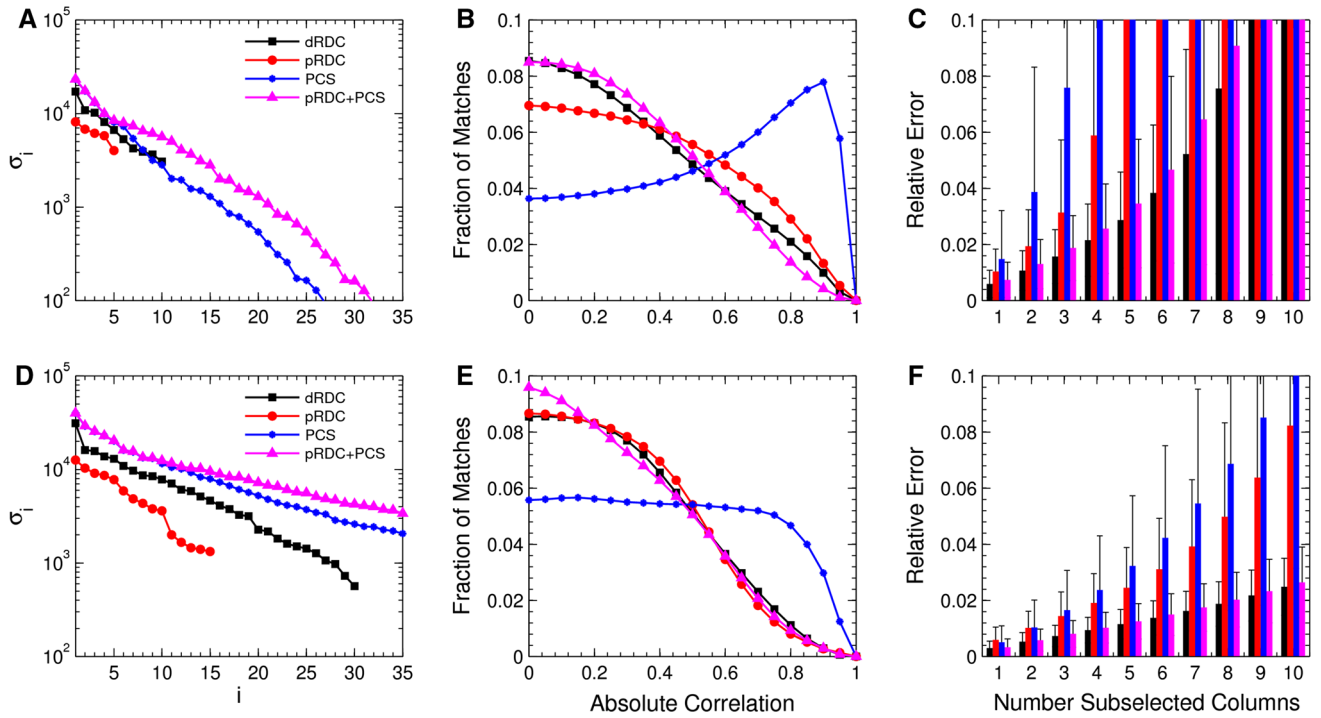


Fig. 2 SVD decomposition (*left panels*), histogram of column correlations (*center panels*), and condition number of randomly subselected set of columns (*right panels*), for the eight described datasets. The results for a single medium/metal ion are shown on the *top*, and the results for the 3 media/metal ions are shown on the *bottom*. **a–d** The 35 largest singular values of the associated **A** matrices. **b–e** The distribution of the uncentered correlations between all pairs of columns in the **A** matrix, estimated by performing 20,000 random samples. **c–f** The expected mean and SD of the relative error for recovering population weights from an arbitrary M = 1,…,10 subset of columns

independent conformers would result in overfitting. The results are shown in Fig. 2a, d.

As shown in Eq. 3, any vector of RDC values (either pRDC or dRDC) from a rigid domain can be expressed as a matrix **V**, which can be determined from the orientations of the bond vectors of that domain, multiplied by the 5 independent components of the alignment tensor matrix. Since there is a linear dependence of the observed data on the 5 components of the alignment tensor, we expect the **A** matrix for dRDC to have rank 10 (5 independent parameters for each of the two domains), and for pRDC to have rank 5, since only the second domain data are used for ensemble recovery. The number of unknowns in the paramagnetic case is also smaller because the alignment tensor for the first domain (5 parameters) can be easily determined from PCS and pRDC measured for this domain, as they are not averaged by conformational variability.

Numerical spectral analyses of the generated prediction matrices for dRDC and pRDC (Fig. 2a, d) support our theoretical analysis, and show that the number of singular values of matrix **A** for one-restraint dRDC and pRDC data is 10 and 5, respectively. Going from 1 to 3 alignment tensors triples the number of non-zero singular values for dRDC and pRDC, as would be expected for linearly independent alignments. The large decrease in the magnitude of singular values for the last 10 dRDC and 5 pRDC non-zero singular values in the three-restraint datasets likely reflects the difficulty in experimentally obtaining three fully independent alignment tensors. The larger magnitudes of dRDC singular values compared to the singular values for pRDC are not related to their information content, but merely reflect the relative strength of diamagnetic versus paramagnetic alignment in the simulated data. On the contrary, it is the decrease in the relative magnitude of the singular values with respect to the largest value, calculated from a set of data, that reflects the difficulty in exploiting the associated restraints, and is hence ultimately related to the information content.

Similarly, the observed PCS data for a rigid domain which is not containing the paramagnetic ion (i.e. for the second domain) can be expressed using 8 parameters: the 5 independent components/parameters defining the **T** tensor, and the 3 parameters describing the metal-ion's position **p** with respect to this domain. However, since the observed PCS vector **y** is not linearly related to **p**, the rank of $\mathbf{A}_{\mathrm{PCS}}$ (calculated from the PCSs in the second domain) is much higher than 8, and greater than that for dRDC or pRDC datasets. The rank of $\mathbf{A}_{\mathrm{PCS}}$ is actually close to (up to) the

number of observations; however as Fig. 2a, d show, the magnitude of the singular values decreases very rapidly. This decrease reflects the strong difference in the PCS values between conformers where the C-terminal (second) domain is close to the metal ion (paramagnetic center) and those where it is far away. After the first ≈ 15 entries, in the one-restraint case the singular values are very small because similar PCS values are calculated for conformers not very far from one another and for nuclei which are spatially close to several other nuclei. When using three sets of metal ions, the number of conformers with large and different PCS values increases. Thus, the decrease in the magnitude of the singular values is significantly slower than in the case of a single metal ion (Fig. 2a–d).

One major advantage of using metal ions instead of steric alignment is that both pRDC and PCS are collected from the same biochemical construct. Thus, two independent datasets can be directly combined, as described in Eq. 6. When combining these datasets, a significantly slower decay in singular values of **A** is obtained compared to the pRDC and PCS datasets analyzed independently. This supports the accepted intuition that pRDC and PCS provide orthogonal structural restraints (pRDCs are very sensitive to orientation, PCSs mostly provide distance restraints).

Histograms of prediction matrices

The spectral analysis of the **A** matrices suggests that pRDC + PCS and even PCS alone provide better restraints for ensemble selection than dRDC. However, singular values are not an exhaustive description of the overall vector distribution. Therefore, we directly analyzed the distribution of correlations between all columns of the matrix **A** calculated for dRDC, pRDC, and PCS. The uncentered correlation distributions between all pairs of columns are shown in Fig. 2b, e. The more uncorrelated the columns of each specific **A** ($\mathbf{A}_{\mathrm{dRDC}}$, $\mathbf{A}_{\mathrm{pRDC}}$, $\mathbf{A}_{\mathrm{PCS}}$) the smaller the chance that an alternative conformer can explain the same subset of experimental data, thus decreasing the number of viable alternative ensembles. In the optimal case, all columns would have zero correlation, and the ensemble solution would be unique.

Figure 2b, e clearly demonstrate that even though the number of singular values of PCS is larger than that of dRDC and pRDC, the correlation distribution is actually significantly worse than for any other dataset, so that their information content could not be larger. The higher correlation for large fraction of the conformers reflects a distribution of PCS where very large changes occur in proximity of the metal ion only, whereas almost no change occurs far away from the metal ion. Additional metal ions can significantly improve the distribution of correlations, although it remains poor with respect to that of the other restraints.

Since pRDCs are distance-independent, they provide a more uniform distribution of values, so that their correlation distribution is much better than for PCS. The pRDC distribution is anyway worse than that of dRDC in the one-restraint case; it significantly improves, essentially to the level of dRDC, in the three-restraint case. Interestingly, the dRDC distribution changes only slightly between one and three restraints, which suggests that the information contained in the additional dRDC datasets is more redundant than in the pRDC case.

Combining pRDC with PCS results in a better correlation distribution than for pRDC and PCS individually. In turn, the correlation distribution of pRDC + PCS is very similar to that of dRDC in the one-restraint case and actually somewhat better in the three-restraint case.

Expected relative error

While the correlation plots in Fig. 2b, e provide an estimate of the **A** matrix column vector distribution, they do not directly tell how well ensembles greater than two can be recovered, nor do they take signal-to-noise ratio into account. To assess how well larger ensembles can be recovered, we computed the mean and standard deviation (SD) of the relative error from a synthetically generated **y** data (with added Gaussian error) for $M = 1,\ldots,10$ columns. The mean and SD were computed by randomly sampling, for each M value, M columns and uniformly at random generating the associated population weights **x**. The synthetic **y** was generated as $\mathrm{y} = \mathbf{Ax} + N(0,1)$, where $N(0,1)$ is the zero-mean Gaussian distribution with $\sigma = 1$. The vector $\mathbf{x}^*$ and the associated relative error, $\|\mathbf{x} - \mathbf{x}^*\|_2/\|\mathbf{x}\|_2$, were recovered by solving Eqs. 2 and 6. In order to guarantee a <0.1 % relative error with >99.999 % confidence using Chernoff bound, the process was repeated 40,000 times for each M. The results for all datasets are shown in Fig. 2c, f.

For the one-restraint datasets, dRDC has lower relative error than pRDC, PCS, or pRDC + PCS. As expected, there is a rapid growth in pRDC errors due to the low matrix rank, and high errors overall in PCS due to the high correlation between columns. In the case of the three-restraint datasets, dRDC has significantly lower relative error than pRDC, even though on the correlation plot the two distributions are very similar. Interestingly, combining pRDC + PCS yields only slightly higher error rate than for dRDC.

Recovering the conformational variability from synthetic datasets

In the previous sections we theoretically analyzed the information content of 8 datasets of synthetic dRDC, pRDC and/or PCS data. Here we perform a direct comparison of the performance of the different restraints in recovering information on the structural variability of the system. To achieve this, we determined (1) the minimum-size sparsest ensemble solution using the SES method (Berlin et al. 2013) and (2) the conformations (as well as the regions in the conformational space) with the highest MaxOcc values. In this way it becomes possible to analyze the accuracy of the recovered solutions from the different sets of synthetic averaged data.

For this purpose, we devised three simulations modeling (1) extensive mobility around a single conformation, (2) two-site exchange with limited mobility around each center, and iii) two-site exchange with a reduced difference in the orientations of the two centers. In each of the simulations, the two-domain protein CaM was allowed to sample different, well defined, parts of its sterically allowed conformational space. Synthetic restraints were calculated as weighted averages over the values of dRDC, pRDC, and PCS of the individual conformations belonging to the sampled regions. These average data were perturbed with a Gaussian error with a SD of 1, 2, or 3 Hz for pRDC and dRDC and of 0.01, 0.02, or 0.03 ppm for PCS.

In the following descriptions of the simulated conformational ensembles, the N-terminal domain of CaM is taken as the frame of reference, and each conformation is described by the different position and orientation of the C-terminal domain with respect to the N-terminal domain. The exact details of each simulation, although described accurately for completeness, are not crucial for the success of the ensemble recovery attempts.

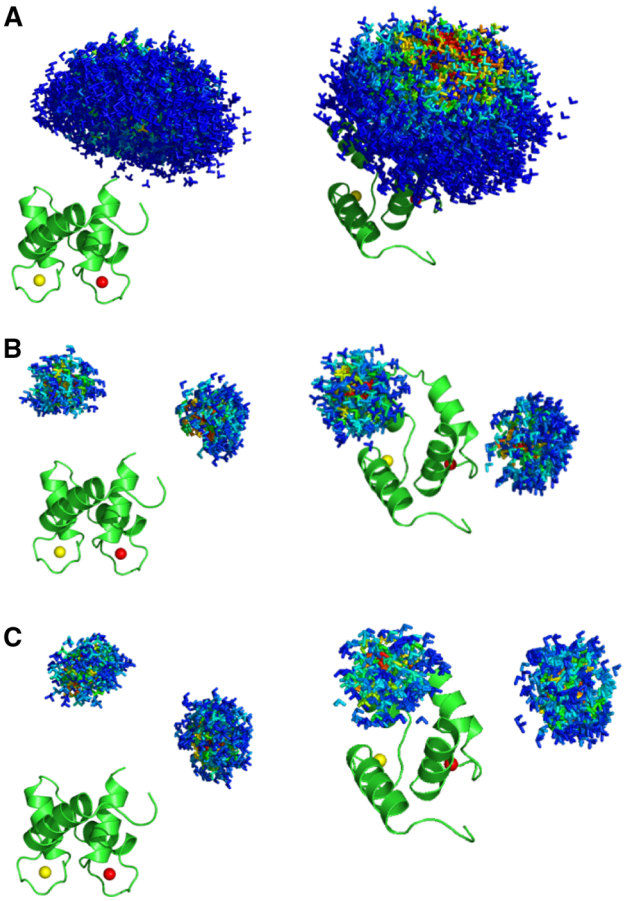


Fig. 3 The simulated ensembles. Different positions of the C-terminal domain of CaM are represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain. The conformers are color coded according to their relative weights (form *red* = high weight to *blue* = low weight). **a** Simulation 1, **b** Simulation 2, **c** Simulation 3. The ensembles are shown from two different points of view in the *left* and *right panels*. All the conformers are superimposed by the N-terminal domain, which is shown in cartoon representation

Simulation 1

In this first simulation we consider the case of conformational variability centered at a single extended conformation of CaM. The sampled ensemble consists of all the conformers, present in the pool of the 32723 sterically allowed conformers, within a distance Δ (measured as a combination of translation and rotation) from the central extended structure (Fig. 3a; Bertini et al. 2012b). Specifically, this distance is defined as:

$$\Delta = d + f(1 - \cos\alpha) \tag{9}$$

where d is the translation of the center of mass of the C-terminal domain from the central structure, and α is the angle of rotation from the central structure, calculated as $\alpha = \arccos(|q_c \cdot q|)$, where q_c and q are the unitary quaternions describing the central structure and the other structure. Note that the two structures are actually 2α apart in Cartesian space (Kuffner 2004). Δ defines the largest allowed spatial displacement (when α is 0) and the largest allowed rotation (when d is 0; it also depends on the factor f) from the position of the central conformer. In the present simulation, conformations with Δ up to 30 Å (a reasonable estimate for this system) were accepted and the value of f was set to 84 Å. In this way, the conformers in the constructed ensemble can have the center of mass of the C-terminal domain at a maximum distance of 30 Å with respect to the conformer at the center of the distribution, if they have the same orientation (the distance decreases with increasing the difference in the orientation). Their C-terminal domain can be rotated up to 100° ($\alpha = 50°$) with respect to the central conformer, if there is no translation of

the center of mass (and gradually less and less as the translational component increases). The weight of each conformation in the ensemble depends on its Δ, and is fixed according to a Gaussian distribution centered at $\Delta = 0$, with SD chosen to provide weights close to zero when Δ is close to 30 Å.

Simulation 2

This simulation models the case of a two-site exchange, with limited mobility allowed around each of the two main conformers (Fig. 3b). The two centers were separated by approximately 30 Å and their C-terminal domains were rotated by ca. 140° with respect to each other. The mobility around each center was simulated as in the previous case with the threshold on Δ set to 10 Å and f equal to 42.7 Å, which corresponds to a maximum allowed angular displacement with respect to the central conformer of 80° ($\alpha = 40°$).

Simulation 3

This simulation is similar to Simulation 2, with the difference that the angular distance between the two sites was decreased almost twofold (Fig. 3c). Sites with more similar orientations are likely to present a bigger challenge in ensemble recovery using restraints which depend on the domain orientations. The distance between the centers (both distinct from those used in Simulation 2) is 30 Å while the difference in orientation of the C-terminal domains is now 80°. The threshold of Δ and the value of f used to simulate the residual mobility around each center were the same as in Simulation 2, hence the same upper limit on the angle α.

SES ensembles

We applied the SES method to these simulated datasets and analyzed how the various restraints affect the recovery of the main conformations contained in the synthetic ensembles used to generate the data. The recovered ensembles were evaluated in terms of their sizes (number of major states) and of the proximity of recovered structures to the centers of the synthetic ensembles (in terms of spatial and angular displacement). As already mentioned in the "Theory" section, the ensemble size was chosen using the L-curve method (Berlin et al. 2013; see Figure S6).

The results are presented in Tables 1, 2 and 3. In general, dRDCs allowed a reasonably accurate recovery of the major states that were used to generate the synthetic datasets (see, for instance, Fig. 4a). However, in all three simulations, in some solutions one additional conformer was recovered, albeit with a relatively low weight. This additional conformer either belongs to the distribution of conformers around one of the main centers (as in Simulation 1 with error of 1 and 2 Hz, and in Simulation 2 with error of 2 Hz, Fig. 4b) or is positioned in-between the two major states (as in Simulation 2 with error of 1 Hz, Fig. 4c). In the first case its presence may reflect conformational heterogeneity; in the second case it is likely related to artifacts. The latter may arise because, 'average conformers' can be more compatible with the averaged experimental observables than any of the actually sampled conformations taken individually.

Table 1 Results of Simulation 1

Simulation 1		Ensemble				Conformer 1		Conformer 2	
Restraint	Error	Ensemble size	Total weight	Weight conformer 1	Weight conformer 2	Translation (Å)	Rotation (°)	Translation (Å)	Rotation (°)
PCS	0.01 ppm	1	0.792	0.792	–	3.85	15.6	–	–
	0.02 ppm	1	0.799	0.799	–	3.85	15.6	–	–
	0.03 ppm	1	0.771	0.771	–	3.85	15.6	–	–
dRDC	1 Hz	2	0.866	0.663	0.203	3.85	15.6	12.78	107.3
	2 Hz	2	0.839	0.482	0.356	8.68	22.4	6.34	39.5
	3 Hz	1	0.755	0.755	–	3.85	15.6	–	–
pRDC	1 Hz	1	0.598	0.598	–	10.48	11.8	–	–
	2 Hz	1	0.597	0.597	–	10.48	11.8	–	–
	3 Hz	1	0.608	0.608	–	14.01	15.6	–	–
pRDC + PCS	1 Hz/0.01 ppm	1	0.755	0.755	–	3.85	15.6	–	–
	2 Hz/0.02 ppm	1	0.761	0.761	–	3.85	15.6	–	–
	3 Hz/0.03 ppm	1	0.742	0.742	–	3.85	15.6	–	–

The table reports the sizes of the recovered ensembles, the specific weights ascribed to their constituent conformers, and for each of these conformers their spatial and angular displacement from the center of the original ensemble

Table 2 Results of Simulation 2

Simulation 2		Ensemble					Conformer 1			Conformer 2			Conformer 3		
Restraint	Error	Ensemble size	Total weight	Weight conformer 1	Weight conformer 2	Weight conformer 3	Closest center	Translation (Å)	Rotation (°)	Closest center	Translation (Å)	Rotation (°)	Closest center	Translation (Å)	Rotation (°)
PCS	0.01 ppm	2	1.000	0.617	0.383	–	B	3.65	26.2	A	0.00	0.0	–	–	–
	0.02 ppm	2	1.000	0.594	0.406	–	B	3.65	26.2	A	0.00	0.0	–	–	–
	0.03 ppm	2	1.000	0.557	0.443	–	A	5.44	16.8	B	3.67	42.6	–	–	–
dRDC	1 Hz	3	0.754	0.345	0.272	0.137	A	0.00	0.0	B	3.99	31.7	B	14.99	63.4
	2 Hz	3	0.881	0.423	0.328	0.130	A	3.12	8.6	B	3.67	42.6	B	4.17	37.5
	3 Hz	2	0.737	0.384	0.353	–	A	6.49	14.4	B	5.16	19.7	–	–	–
pRDC	1 Hz	2	0.674	0.338	0.336	–	A	8.45	14.1	B	14.52	21.6	–	–	–
	2 Hz	2	0.636	0.360	0.275	–	A	12.22	32.8	B	14.86	13.2	–	–	–
	3 Hz	2	0.692	0.365	0.327	–	B	9.66	22.0	A	15.56	14.6	–	–	–
pRDC + PCS	1 Hz/0.01 ppm	2	0.741	0.384	0.357	–	B	7.72	34.1	A	0.00	0.0	–	–	–
	2 Hz/0.02 ppm	2	0.926	0.481	0.445	–	B	7.85	24.3	A	0.00	0.0	–	–	–
	3 Hz/0.03 ppm	2	1.000	0.546	0.454	–	B	7.85	24.3	A	7.63	12.0	–	–	–

The table reports the sizes of the recovered ensembles, the specific weights ascribed to their constituent conformers and for each of these conformers indicates the closest of the two original sites together their spatial and angular displacement from it

Table 3 Results of Simulation 3

Simulation 3		Ensemble					Conformer 1			Conformer 2			Conformer 3		
Restraint	Error	Ensemble size	Total weight	Weight conformer 1	Weight conformer 2	Weight conformer 3	Closest center	Translation (Å)	Rotation (°)	Closest center	Translation (Å)	Rotation (°)	Closest center	Translation (Å)	Rotation (°)
PCS	0.01 ppm	1	1.000	1.000	–	–	A	10.92	24.8	–	–	–	–	–	–
	0.02 ppm	1	1.000	1.000	–	–	A	10.92	24.8	–	–	–	–	–	–
	0.03 ppm	1	1.000	1.000	–	–	A	10.92	24.8	–	–	–	–	–	–
dRDC	1 Hz	3	0.965	0.451	0.378	0.137	A	2.15	19.6	B	4.05	19.6	B	9.31	86.9
	2 Hz	2	0.820	0.431	0.389	–	A	7.09	21.6	B	4.25	24.3	–	–	–
	3 Hz	2	0.876	0.440	0.436	–	A	3.29	18.1	B	4.05	19.6	–	–	–
pRDC	1 Hz	2	0.739	0.374	0.364	–	B	9.60	20.5	A	12.36	11.0	–	–	–
	2 Hz	2	0.697	0.361	0.336	–	B	10.62	11.1	A	9.38	14.2	–	–	–
	3 Hz	2	0.720	0.362	0.358	–	B	4.25	24.3	A	5.20	13.9	–	–	–
pRDC + PCS	1 Hz/0.01 ppm	2	0.772	0.391	0.380	–	A	3.29	18.1	B	6.70	8.9	–	–	–
	2 Hz/0.02 ppm	2	0.799	0.408	0.392	–	B	17.82	19.9	A	8.03	20.0	–	–	–
	3 Hz/0.03 ppm	2	0.823	0.445	0.377	–	A	3.05	11.6	B	6.70	8.9	–	–	–

The table reports the sizes of the recovered ensembles, the specific weights ascribed to their constituent conformers and for each of these conformers indicates the closest of the two original sites together their spatial and angular displacement from it

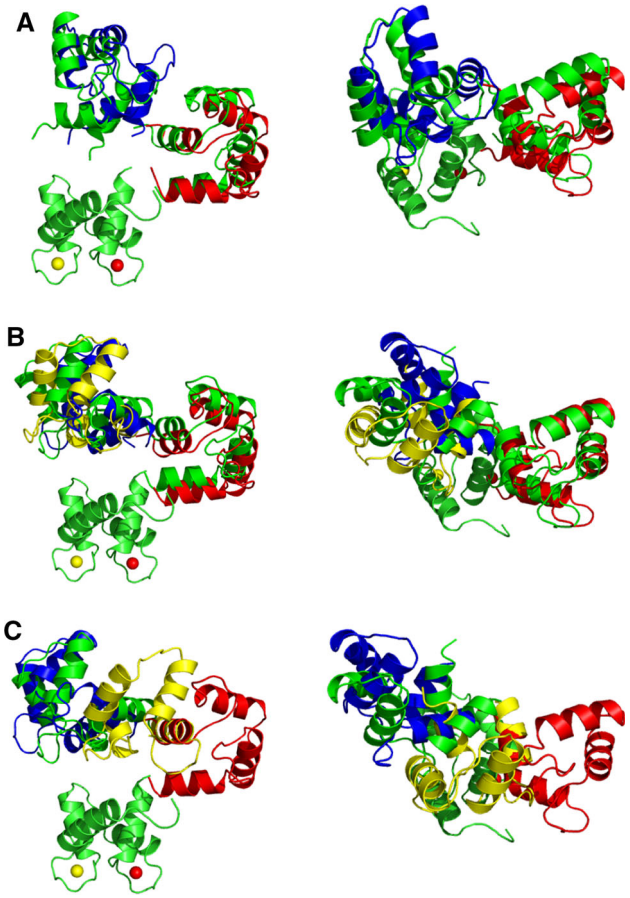


Fig. 4 SES recovery using dRDC data. Color code for the C-terminal domain: *green*—simulated conformers in the centers of the regions, *red*, *blue*, and *yellow*—reconstructed conformers with highest, intermediate, and lowest weight, respectively. **a** An ensemble with correctly recovered major states (Simulation 3, 3 Hz error), **b** An ensemble with an additional state present next to one of the centers (Simulation 2, 2 Hz error), **c** An ensemble with an additional state (*yellow*) recovered (Simulation 2, 1 Hz error). The ensembles are shown from two different points of view in the *left* and *right panels*. All conformers are superimposed by the N-terminal domain

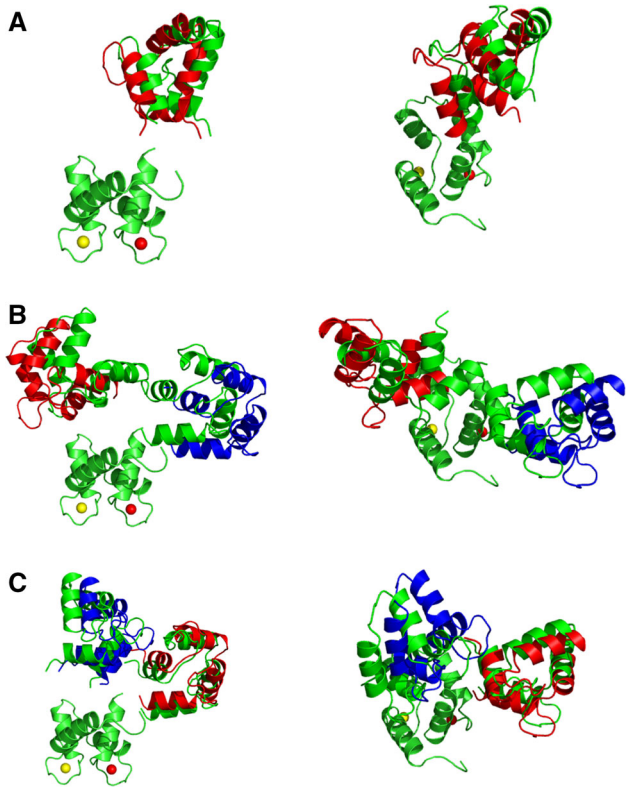


Fig. 5 SES recovery with all the paramagnetic data. Color code for the C-terminal domain: *green*—simulated conformers in the centers of the regions, *red*, *blue*—reconstructed conformers with higher and lower weight, respectively. **a** Simulation 1, **b** Simulation 2, **c** Simulation 3, with 0.03 ppm and 3 Hz errors. The ensembles are shown from two different points of view in the *left* and *right panels*. All conformers are superimposed by the N-terminal domain

In the case of pRDCs, the right number of major states was always recovered (Fig. S1), and in the corresponding conformers the domains were oriented with an accuracy comparable to that achieved with dRDC. It should be recalled that pRDCs contain no information whatsoever on the relative positions of the domains, which therefore results in inaccuracy of their positioning.

PCS data alone in two out of three simulations were sufficient to recover the correct solutions (Fig. S2) in terms of ensemble sizes and locations of the major states (with the accuracy similar to dRDC). However in Simulation 3, where the two states are more alike to one another, the calculations provide only a single state (Fig. S2B) situated in-between the two actual centers (in terms of both translation and orientation). The recovery of such an incorrect state is most likely, as already mentioned for dRDC, the outcome of the averaging of the experimental observables. Using all the paramagnetic data together (i.e. pRDC and PCS) improved the robustness of the recovery: both translations and orientations were satisfactory accurate in all cases (Fig. 5). The translation and rotation with respect to the conformers at the center of the distributions were within 4 Å and 16° for Simulation 1, 8 Å and 34° for Simulation 2, and 3 Å and 18° or 7 Å and 9° for Simulation 3 (1 Hz and 0.01 ppm error case). The ensemble recovery is robust, as increased errors did not noticeably affect the accuracy of solutions.

In conclusion, diamagnetic RDC, as well as the combination of paramagnetic RDC and PCS, are both equally suitable restraints for the recovery of the major states present in conformational ensembles. Special attention should be paid to the fact that, occasionally, 'average conformers' may be recovered.

MaxOcc analysis

Similar to the SES analysis, we performed MaxOcc analysis on the same datasets. From the MaxOcc values, it is

possible to determine which conformers can be sampled with the largest weights. In order to speed the computational analysis up, we used random sampling to detect regions of with potentially high MaxOcc conformers, and then expanded those regions, to find the globally best solution. To do this, we first computed MaxOcc for 400 conformers, randomly chosen from the generated pool (Bertini et al. 2010b, 2012b; Cerofolini et al. 2013). Then the conformers with the highest MaxOcc (up to 0.8 of the MaxOcc of the highest scoring conformer) were selected and the MaxOcc of their neighboring conformers (in the conformational space) were calculated. The procedure was repeated until no more neighbors with high MaxOcc were found. The neighboring conformers scored at each iteration were chosen using Eq. 9 with the threshold on Δ of 5 Å and $f = 40$ Å. If the final distribution of the highest MaxOcc conformers was broad, the analysis was supplemented by the maximum occurrence of regions (MaxOR) approach, which permitted to discriminate between the cases of high MaxOcc conformers corresponding to conformers actually sampled by the protein and the cases of high MaxOcc conformers corresponding to conformers arising from data averaging (Andralojc et al. 2014).

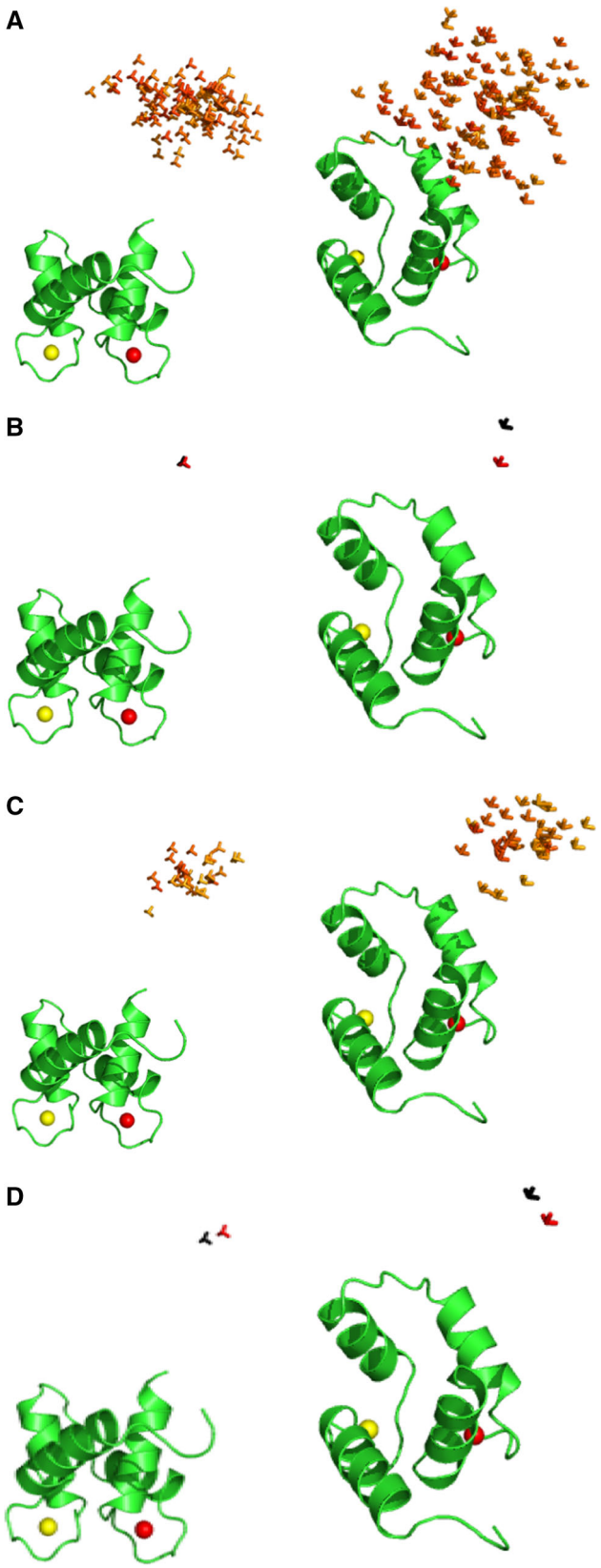
The results of the MaxOcc analysis are reported in Table 4 for all three simulations. In Simulation 1, for both the paramagnetic and diamagnetic data, the analysis revealed that all the conformers with the highest MaxOcc (from 0.8 to 1 of the highest MaxOcc, corresponding to 0.58–0.73 for the paramagnetic data and 0.57–0.71 for the dRDC) form a single, relatively compact, region in the conformational space (Fig. 6a, c). In order to quantify its agreement with the original distribution, the center of the region was calculated by averaging the translational and orientation parameters of the highest MaxOcc conformers. The conformation so obtained was then compared with the conformation at the center of the original distribution. As shown in Table 4 and Fig. 6b, d, the agreement was very good in terms of spatial and angular displacement for both the diamagnetic and the paramagnetic data, either for 1 Hz/0.01 ppm or for 3 Hz/0.03 ppm errors.

In simulation 2, i.e. the case of two well separated conformational regions, when dRDC are used, the highest MaxOcc conformers are positioned in two distinct, clearly separated regions (Fig. 7a), the centers of which are positioned very close to the centers of the actually sampled distribution (Table 4; Fig. 7b). When paramagnetic data (PCS + pRDC) are used, the highest MaxOcc (0.41–0.51) conformers are positioned in one elongated, banana-shape region in the conformational space (Fig. 8a), which includes the two actually sampled centers, but also many conformers situated between them (their high score is an outcome of conformational averaging as described in the SES results paragraph). From these results, one cannot conclude whether the studied conformational ensemble mainly reflects a two-site exchange case or the sampling of all the conformations within the determined region. In order to distinguish between these two cases, MaxOR calculations were performed. The highest MaxOcc conformers were clustered in 5 regions, shown in Fig. 8b, which include all conformations with distance $\Delta \leq 5$ Å from the central conformation (calculated using eq. 9, with $f = 147$ Å). The MaxOR values for these regions are

Table 4 The MaxOcc/MaxOR analysis

Simulation	Restraint	Error	Recovered center 1			Recovered center 2		
			Closest center	Translation (Å)	Rotation (°)	Closest center	Translation (Å)	Rotation (°)
Simulation 1	dRDC	1 Hz	–	3.42	5.7	–	–	–
		3 Hz	–	2.65	10.3			
	pRDC + PCS	1 Hz/0.01 ppm	–	4.72	9.1	–	–	–
		3 Hz/0.03 ppm	–	4.16	9.5	–	–	–
Simulation 2	dRDC	1 Hz	A	3.44	4.4	B	1.71	25.5
		3 Hz	A	3.27	9.7	B	3.26	29.8
	pRDC + PCS	1 Hz/0.01 ppm*	A	4.42	14.4	B	1.02	28.1
		3 Hz/0.03 ppm	A	7.15	25.1	B	6.60	24.7
Simulation 3	dRDC	1 Hz*	A	0.99	14.3	B	8.17	16.6
		3 Hz	A	5.32	28.9	B	5.35	16.0
	pRDC + PCS	1 Hz/0.01 ppm*	A	4.39	10.8	B	3.57	31.4
		3 Hz/0.03 ppm	A	5.45	9.9	B	5.80	40.3

For each simulation the spatial and angular displacement of the center of the ensemble of the best scoring conformers from the center of the actually sampled distribution is reported, together with the indication of the closest site, if applicable. The cases where MaxOR was used are indicated with an asterisk next to the error level



◀**Fig. 6** MaxOcc results for Simulation 1. Each conformation is represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain. Color code for **a**, **c**—according to the MaxOcc value (0.0—*blue*, 0.8—*red*), **a** The conformers with the highest MaxOcc recovered with the paramagnetic data (with error of 1 Hz for pRDC and 0.01 ppm for PCS), **b** The center of the distribution shown in *panel a* (*red*) versus the center of the simulated region (*black*), **c** The conformers with the highest MaxOcc recovered with dRDC (with error of 1 Hz), **d** The center of the distribution shown in *panel c* (*red*) versus the center of the simulated region (*black*). The results are shown from two different points of view in the *left* and *right panels*. All conformers are superimposed by the N-terminal domain, shown as a ribbon

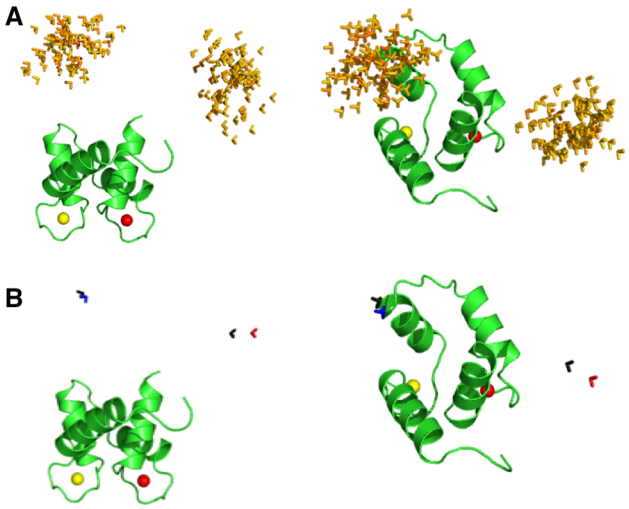


Fig. 7 MaxOcc results for Simulation 2 with dRDC with error of 1 Hz. Each conformation is represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain. **a** The conformers with the highest MaxOcc, color code: according to the MaxOcc value (0.0—*blue*, 0.6—*red*). **b** The center of the distribution shown in *panel a* (*red*) versus the center of the simulated region (*black*). All conformers are superimposed by the N-terminal domain, shown as a ribbon

reported in Table S1 (diagonal entries). All regions have similar MaxOR values (up to 0.60), not much higher than the largest MaxOcc values for the individual conformations. If however MaxOR values are calculated for pairs of regions (off-diagonal entries of Table S1), strong differences arise. All pairs yielding the highest MaxOR (0.90–1.00) are composed of regions at the opposite sides of the distribution of the highest MaxOcc conformers, whereas all pairs composed of the regions located on the same side of the distribution or more importantly containing a region in the middle, have significantly lower MaxOR (up to 0.63 and 0.78, respectively). This strongly suggests the occurrence of a two-site exchange model. The

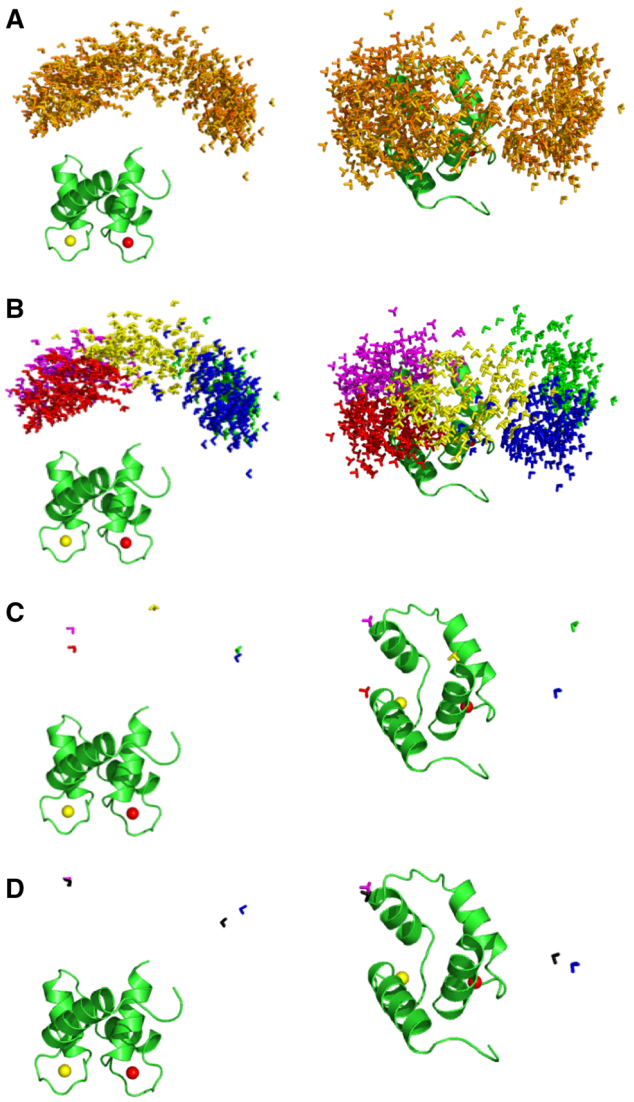


Fig. 8 MaxOcc/MaxOR results for Simulation 2 with paramagnetic data (1 Hz error for pRDC and 0.01 ppm error for PCS). Each conformation is represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain. **a** The conformers with the highest MaxOcc, color code—according to the MaxOcc value (0.0—*blue*, 0.6—*red*). **b** The five clusters formed by the conformers in *panel a*. **c** The centers of the clusters. **d** The centers of clusters with the largest MaxOR versus the centers of the simulated regions (*black*). All conformers are superimposed by the N-terminal domain, shown as a ribbon

pair of regions with the highest MaxOR has their central conformations in nice agreement with the conformations in the center of the distributions in the synthetic ensemble, with an accuracy comparable to that obtained by SES (Table 4; Fig. 8d).

In simulation 3, for both the paramagnetic and diamagnetic data, the conformers recovered by MaxOcc form elongated regions comprising both the two centers and conformers situated between them (Figs. S3A and S4A). MaxOR was thus applied in both cases. As in the previous simulation, no single region has MaxOR significantly higher than the others, but the analysis of pairs of regions indicated again the occurrence of a two-site exchange (Tables S2 and S3). The two central conformations of the synthetic ensemble were identified with good accuracy (Table 4; Figs. S3D and S4D) using both kinds of experimental restraints. Again, the results are robust, as increased errors did not largely affect the accuracy of the solutions.

The performed MaxOcc/MaxOR analysis, as it appears from Table 4 as a whole, confirms the conclusion from the SES results that paramagnetic and diamagnetic restraints are equally useful for the recovery of conformational ensembles.

Conclusions

In many experimental studies RDCs have been shown to be precious restraints for analyzing molecular conformational freedom (Montalvao et al. 2014; Ravera et al. 2014; Camilloni and Vendruscolo 2015; Torchia 2015). Here we compared paramagnetic and diamagnetic RDCs and found substantial differences in their information content in the case of multidomain proteins. We found that the information content of dRDC is larger than that of pRDC in terms of number of singular values, and this reflects the shape dependence of dRDC. However, since the internal alignment due to paramagnetism also gives rise to PCSs, the total informational content recovered in a paramagnetic experiment is at least on par with dRDCs.

We have performed several simulations to evaluate the capability of recovering the conformational variability of two-domain proteins by the use of two different approaches, SES and MaxOcc/MaxOR. The main states of the protein were recovered reasonably well for both paramagnetic and diamagnetic datasets, with both approaches (see Tables 1, 2, 3, 4 and also Table S4). Even for rather large experimental errors, we have found that both datasets still retain the ability of recovering the main conformational states, thus resulting appealing for the analysis of averaged experimental data possibly also in the case of large systems, where RDCs are affected by large errors. Of course, since the problem is underdetermined, a correct reconstruction of the main states may be unsuccessful for different rather unpredictable conformational distributions.

Such analysis suggests that pRDC + PCS provide a very promising alternative to dRDC data. It is important to note that this analysis does not include modeling error, which is harder to quantify. Therefore, our analysis does not capture the principal advantages of pRDC + PCS over

dRDC, in that it does not require assumption of a barrier model in order to predict the alignment. In addition, one has to consider that the interactions of the protein with the alignment medium might actually perturb the system, and that these interactions can occur on a timescale that is slower than the conformational averaging itself, so that the assumption that the measured dRDCs can be represented as a population-weighted average of the RDCs for the individual (rigid) conformers may fall short in representing the real physical picture.

Finally, the availability of a number of rigid lanthanide-binding tags nowadays may make the acquisition of three independent metal ion datasets more practical and safer than the acquisition and prediction of three independent alignment media. One current limitation of using metal ions is the low signal-to-noise ratio in pRDC and PCS data, which could potentially be improved with better technology and methodology.

Acknowledgments This work has been supported by Ente Cassa di Risparmio di Firenze, MIUR PRIN 2012SK7ASN, NIH Grant GM065334, European Commission projects BioMedBridges No. 284209, pNMR No. 317127, and Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions. Specifically, we thank the EU ESFRI Instruct Core Centre CERM, Italy.

Conflict of interest The authors declare that they have no conflict of interest.

Compliance with ethical standard This article does not contain any studies with human participants or animals performed by any of the authors.

References

Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH (2000) Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. J Magn Reson 143:402–406

Allegrozzi M, Bertini I, Janik MBL, Lee Y-M, Liu G, Luchinat C (2000) Lanthanide induced pseudocontact shifts for solution structure refinements of macromolecules in shells up to 40 Å from the metal ion. J Am Chem Soc 122:4154–4161

Andralojc W, Luchinat C, Parigi G, Ravera E (2014) Exploring regions of conformational space occupied by two-domain proteins. J Phys Chem B 118:10576–10587

Balayssac S, Bertini I, Bhaumik A, Lelli M, Luchinat C (2008) Paramagnetic shifts in solid-state NMR of proteins to elicit strucutral information. Proc Natl Acad Sci USA 105:17284–17289

Banci L, Bertini I, Bren KL, Cremonini MA, Gray HB, Luchinat C, Turano P (1996) The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyano-cytochrome *c* as an example. J Biol Inorg Chem 1:117–126

Banci L, Bertini I, Gori Savellini G, Romagnoli A, Turano P, Cremonini MA, Luchinat C, Gray HB (1997) Pseudocontact shifts as constraints for energy minimization and molecular dynamic calculations on solution structures of paramagnetic metalloproteins. Proteins Struct Funct Genet 29:68–76

Banci L, Bertini I, Huber JG, Luchinat C, Rosato A (1998) Partial orientation of oxidized and reduced cytochrome b_5 at high magnetic fields: magnetic susceptibility anisotropy contributions and consequences for protein solution structure determination. J Am Chem Soc 120:12903–12909

Barbato G, Ikura M, Kay LE, Pastor RW, Bax A (1992) Backbone dynamics of calmodulin studied by ^{15}N relaxation using inverse detected two-dimensional NMR spectroscopy; the central helix is flexible. Biochemistry 31:5269–5278

Barthelmes K, Reynolds AM, Peisach E, Jonker HRA, DeNunzio NJ, Allen KN, Imperiali B, Schwalbe H (2011) Engineering encodable lanthanide-binding tags into loop regions of proteins. J Am Chem Soc 133:808–819

Bashir Q, Volkov AN, Ullmann GM, Ubbink M (2010) Visualization of the encounter ensemble of the transient electron transfer complex of cytochrome *c* and cytochrome *c* peroxidase. J Am Chem Soc 132:241–247

Berlin K, O'Leary DP, Fushman D (2009) Improvement and analysis of computational methods for prediction of residual dipolar couplings. J Magn Reson 201:25–33

Berlin K, Castañeda CA, Schneidman-Dohovny D, Sali A, Nava-Tudela A, Fushman D (2013) Recovering a representative conformational ensemble from underdetermined macromolecular structural data. J Am Chem Soc 135:16595–16609

Bernadò 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

Bertini I, Donaire A, Jiménez B, Luchinat C, Parigi G, Piccioli M, Poggi L (2001a) Paramagnetism-based versus classical constraints: an analysis of the solution structure of Ca Ln calbindin D_{9k}. J Biomol NMR 21:85–98

Bertini I, Janik MBL, Lee Y-M, Luchinat C, Rosato A (2001b) Magnetic susceptibility tensor anisotropies for a lanthanide Ion series in a fixed protein matrix. J Am Chem Soc 123:4181–4188

Bertini I, Janik MBL, Liu G, Luchinat C, Rosato A (2001c) Solution structure calculations through self-orientation in a magnetic field of cerium (III) substituted calcium-binding protein. J Magn Reson 148:23–30

Bertini I, Longinetti M, Luchinat C, Parigi G, Sgheri L (2002a) Efficiency of paramagnetism-based constraints to determine the spatial arrangement of α-helical secondary structure elements. J Biomol NMR 22:123–136

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

Bertini I, Gelis I, Katsaros N, Luchinat C, Provenzani A (2003) Tuning the affinity for lanthanides of calcium binding proteins. Biochemistry 42:8011–8021

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

Bertini I, Fragai M, Lee Y-M, Luchinat C, Terni B (2004b) Paramagnetic metal ions in ligand screening: the Co^{II} matrix metalloproteinase 12. Angew Chem Int Ed 43:2254–2256

Bertini I, Luchinat C, Parigi G, Pierattelli R (2005) NMR of paramagnetic metalloproteins. ChemBioChem 6:1536–1549

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, Luchinat C, Parigi G, Pierattelli R (2008) Perspectives in NMR of paramagnetic proteins. Dalton Trans 2008:3782–3790

Bertini I, Kursula P, Luchinat C, Parigi G, Vahokoski J, Willmans M, Yuan J (2009) Accurate solution structures of proteins from X-ray data and minimal set of NMR data: calmodulin peptide complexes as examples. J Am Chem Soc 131:5134–5144

Bertini I, Bhaumik A, De Paepe G, Griffin RG, Lelli M, Lewandowski JR, Luchinat C (2010a) High-resolution solid-state NMR structure of a 17.6 kDa protein. J Am Chem Soc 132:1032–1040

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

Bertini I, Calderone V, Cerofolini L, Fragai M, Geraldes CFGC, Hermann P, Luchinat C, Parigi G, Teixeira JMC (2012a) The catalytic domain of MMP-1 studied through tagged lanthanides. Dedicated to Prof. A.V. Xavier. FEBS Lett 586:557–567

Bertini I, Ferella L, Luchinat C, Parigi G, Petoukhov MV, Ravera E, Rosato A, Svergun DI (2012b) MaxOcc: a web portal for maximum occurence analysis. J Biomol NMR 53:271–280

Bertini I, Luchinat C, Nagulapalli M, Parigi G, Ravera E (2012c) Paramagnetic relaxation enhancements for the characterization of the conformational heterogeneity in two-domain proteins. Phys Chem Chem Phys 14:9149–9156

Blackledge M (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. Prog NMR Spectrosc 46:23–61

Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. Science 313:1638–1642

Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition (vol 5, pg 789, 2009). Nat Chem Biol 5:954

Bonvin AM, Brunger AT (1996) Do NOE distances contain enough information to assess the relative populations of multi-conformer structures? J Biomol NMR 7:72–76

Bothe JR, Nikolova EN, Eichhorn CD, Chugh J, Hansen AL, Al Hashimi HM (2011) Characterizing RNA dynamics at atomic resolution using solution-state NMR spectroscopy. Nat Methods 8:919–931

Burgi R, Pitera J, Van Gunsteren WF (2001) Assessing the effect of conformational averaging on the measured values of observables. J Biomol NMR 19:305–320

Camilloni C, Vendruscolo M (2015) A tensor-free method for the structural and dynamical refinement of proteins using residual dipolar couplings. J Phys Chem B 119:653–661

Cerofolini L, Fields GB, Fragai M, Geraldes CFGC, Luchinat C, Parigi G, Ravera E, Svergun DI, Teixeira JMC (2013) Examination of matrix metalloproteinase-1 (MMP-1) in solution: a preference for the pre-collagenolysis state. J Biol Chem 288:30659–30671

Chen Y, Campbell SL, Dokholyan NV (2007) Deciphering protein dynamics from NMR data using explicit structure sampling and selection. Biophys J 93:2300–2306

Chou JJ, Li S, Klee CB, Bax A (2001) Solution structure of Ca^{2+} calmodulin reveals flexible hand-like properties of its domains. Nat Struct Biol 8:990–997

Choy W-Y, Forman-Kay JD (2001) Calculation of ensembles of structures representing the unfolded state of an SH3 domain. J Mol Biol 308:1011–1032

Chuang GY, Mehra-Chaudhary R, Ngan CH, Zerbe BS, Kozakov D, Vajda S, Beamer LJ (2010) Domain motion and interdomain hot spots in a multidomain enzyme. Protein Sci 19:1662–1672

Clore GM, Schwieters CD (2004) How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J Am Chem Soc 126:2923–2938

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

Diaz-Moreno I, Diaz-Quintana A, De la Rosa MA, Ubbink M (2005) Structure of the complex between plastocyanin and cytochrome f from the cyanobacterium nostoc Sp. PCC 7119 as determined by paramagnetic NMR. J Biol Chem 280:18908–18915

Fisher CK, Stultz CM (2011) Constructing ensembles for instrinsically disordered proteins. Curr Opin Struct Biol 21:426–431

Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. J Am Chem Soc 132:14919–14927

Fragai M, Luchinat C, Parigi G (2006) "Four-dimensional" protein structures: examples from metalloproteins. Acc Chem Res 39:909–917

Gaponenko V, Sarma SP, Altieri AS, Horita DA, Li J, Byrd RA (2004) Improving the accuracy of NMR structures of large proteins using pseudocontact shifts as long/range restraints. J Biomol NMR 28:205–212

Gardner RJ, Longinetti M, Sgheri L (2005) Reconstruction of orientations of a moving protein domain from paramagnetic data. Inverse Probl 21:879–898

Gempf KL, Butler SJ, Funk AM, Parker D (2013) Direct and selective tagging of cysteine residues in peptides and proteins with 4-nitropyridyl lanthanide complexes. Chem Commun (Camb) 49:9104–9106

Gochin M, Roder H (1995a) Protein structure refinement based on paramagnetic NMR shifts: applications to wild-type and mutants forms of cytochrome c. Protein Sci 4:296–305

Gochin M, Roder H (1995b) Use of pseudocontact shifts as a structural constraint for macromolecules in solution. Bull Magn Reson 17:1–4

Guerry P, Salmon L, Mollica L, Ortega Roldan JL, Markwick P, van Nuland NA, McCammon JA, Blackledge M (2013) Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. Angew Chem Int Ed Engl 52:3181–3185

Hansen MR, Mueller L, Pardi A (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. Nat Struct Biol 5:1065–1074

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

Huang J, Grzesiek S (2010) Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin. J Am Chem Soc 132:694–705

Hulsker R, Baranova MV, Bullerjahn GS, Ubbink M (2008) Dynamics in the transient complex of plastocyanin-cytochrome f from Prochlorothrix hollandica. J Am Chem Soc 130:1985–1991

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

Jensen MR, Hansen DF, Ayna U, Dagil R, Hass MA, Christensen HE, Led JJ (2006) On the use of pseudocontact shifts in the structure determination of metalloproteins. Magn Reson Chem 44:294–301

John M, Otting G (2007) Strategies for measurements of pseudocontact shifts in protein NMR spectroscopy. ChemPhysChem 8:2309–2313

Jones E, Oliphant E, Peterson P et al (2001) SciPy: Open source scientific tools for Python

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

Kobashigawa Y, Saio T, Ushio M, Sekiguchi M, Yokochi M, Ogura K, Inagaki F (2012) Convenient method for resolving degeneracies due to symmetry of the magnetic susceptibility tensor and its application to pseudo contact shift-based protein-protein complex structure determination. J Biomol NMR 53:53–63

Korzhnev DM, Kay LE (2008) Probing invisible, low-populated states of protein molecules by relaxation dispersion NMR spectroscopy: an application to protein folding. Acc Chem Res 41:442–451

Kuffner JJ (2004) Effective sampling and distance metrics for 3D rigid body path planning. In: Proceedings IEEE international conference on Robotics and Automation (ICRA), vol 4, p 3993

Kukic P, Camilloni C, Cavalli A, Vendruscolo M (2014) Determination of the individual roles of the linker residues in the interdomain motions of calmodulin using NMR chemical shifts. J Mol Biol 426:1826–1838

Kurland RJ, McGarvey BR (1970) Isotropic NMR shifts in transition metal complexes: calculation of the Fermi contact and pseudocontact terms. J Magn Reson 2:286–301

Lakomek NA, Walter KF, Fares C, Lange OF, de Groot BL, Grubmuller H, Bruschweiler R, Munk A, Becker S, Meiler J, Griesinger C (2008) Self-consistent residual dipolar coupling based model-free analysis for the robust determination of nanosecond to microsecond protein dynamics. J Biomol NMR 41:139–155

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

Latham MP, Hanson P, Brown DJ, Pardi A (2008) Comparison of alignment tensors generated for native tRNA(Val) using magnetic fields and liquid crystalline media. J Biomol NMR 40:83–94

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

Liu WM, Keizers PH, Hass MA, Blok A, Timmer M, Sarris AJ, Overhand M, Ubbink M (2012) A pH-sensitive, colorful, lanthanide-chelating paramagnetic NMR probe. J Am Chem Soc 134:17306–17313

Loh CT, Ozawa K, Tuck KL, Barlow N, Huber T, Otting G, Graham B (2013) Lanthanide tags for site-specific ligation to an unnatural amino acid and generation of pseudocontact shifts in proteins. Bioconjug Chem 24:260–268

Lohman JAB, Maclean C (1978) Alignment effects on high resolution NMR spectra induced by the magnetic field. Chem Phys 35:269–274

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

Losonczi JA, Prestegard JH (1998) Improved dilute bicelle solutions for high-resolution NMR of biological macromolecules. J Biomol NMR 12:447–451

Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138:334–342

Luchinat C, Nagulapalli M, Parigi G, Sgheri L (2012a) Maximum occurence analysis of protein conformations for different distributions of paramagnetic metal ions within flexible two-domain proteins. J Magn Reson 215:85–93

Luchinat C, Parigi G, Ravera E, Rinaldelli M (2012b) Solid state NMR crystallography through paramagnetic restraints. J Am Chem Soc 134:5006–5009

Maltsev AS, Grishaev A, Roche J, Zasloff M, Bax A (2014) Improved cross validation of a static ubiquitin structure derived from high precision residual dipolar couplings measured in a drug-based liquid crystalline phase. J Am Chem Soc 136:3752–3755

Man B, Su XC, Liang H, Simonsen S, Huber T, Messerle BA, Otting G (2010) 3-Mercapto-2,6-pyridinedicarboxylic acid: a small lanthanide-binding tag for protein studies by NMR spectroscopy. Chem Eur J 16:3827–3832

Montalvao R, Camilloni C, De SA, Vendruscolo M (2014) New opportunities for tensor-free calculations of residual dipolar couplings for the study of protein dynamics. J Biomol NMR 58:233–238

Musiani F, Rossetti G, Capece L, Gerger TM, Micheletti C, Varani G, Carloni P (2014) Molecular dynamics simulations identify time scale of conformational changes responsible for conformational selection in molecular recognition of HIV-1 transactivation responsive RNA. J Am Chem Soc 136:15631–15637

Nesterov Y (2012) Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM J Optim 22:341–362

Nodet L, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. J Am Chem Soc 131:17908–17918

O'Leary DP (2009) Scientific computing with case studies. SIAM, Bangkok

Pickford AR, Campbell ID (2004) NMR studies of modular protein structures and their interactions. Chem Rev 104:3557–3566

Pintacuda G, John M, Su XC, Otting G (2007) NMR structure determination of protein-ligand complexes by lanthanide labeling. Acc Chem Res 40:206–212

Prestegard JH, Al-Hashimi HM, Tolman JR (2000) NMR structures of biomolecules using field oriented media and residual dipolar couplings. Q Rev Biophys 33:371–424

Ramirez BE, Bax A (1998) Modulation of the alignment tensor of macromolecules dissolved in a dilute liquid crystalline medium. J Am Chem Soc 120:9106–9107

Ravera E, Salmon L, Fragai M, Parigi G, Al-Hashimi HM, Luchinat C (2014) Insights into domain-domain motions in proteins and RNA from solution NMR. Acc Chem Res 47:3118–3126

Rinnenthal J, Buck J, Ferner J, Wacker A, Furtig B, Schwalbe H (2011) Mapping the landscape of RNA dynamics with NMR spectroscopy. Acc Chem Res 44:1292–1301

Rodriguez-Castañeda F, Haberz P, Leonov A, Griesinger C (2006) Paramagnetic tagging of diamagnetic proteins for solution NMR. Magn Reson Chem 44:S10–S16

Russo L, Maestre-Martinez M, Wolff S, Becker S, Griesinger C (2013) Interdomain dynamics explored by paramagnetic NMR. J Am Chem Soc 135:17111–17120

Ryabov YE, Fushman D (2006) Analysis of interdomain dynamics in a two-domain protein using residual dipolar couplings together with 15 N relaxation data. Magn Reson Chem 44:S143–S151

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

Saio T, Ogura K, Shimizu K, Yokochi M, Burke TR Jr, Inagaki F (2011) An NMR strategy for fragment-based ligand screening utilizing a paramagnetic lanthanide probe. J Biomol NMR 51:395–408

Schmitz C, Vernon R, Otting G, Baker D, Huber T (2012) Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol 416:668–677

Schroeder R, Barta A, Semrad K (2004) Strategies for RNA folding and assembly. Nat Rev Mol Cell Biol 5:908–919

Sgheri L (2010a) Conformational freedom of proteins and the maximal probability of sets of orientations. Inverse Probl 26:035003-1–035003-19

Sgheri L (2010b) Joining RDC data from flexible protein domains. Inverse Probl 26:115021-1–115021-12

Sicheri F, Kuriyan J (1997) Structures of Src-family tyrosine kinases. Curr Opin Struct Biol 7:777–785

Simin M, Irausquin S, Cole CA, Valafar H (2014) Improvements to REDCRAFT: a software tool for simultaneous characterization of protein backbone structure and dynamics from residual dipolar couplings. J Biomol NMR 60:241–264

Stelzer AC, Frank AT, Bailor MH, Andricioaei I, Al Hashimi HM (2009) Constructing atomic-resolution RNA structural ensembles using MD and motionally decoupled NMR RDCs. Methods 49:167–173

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

Su XC, Huber T, Dixon NE, Otting G (2006) Site-specific labelling of proteins with a rigid lanthanide-binding tag. ChemBioChem 7:1599–1604

Su XC, Man B, Beeren S, Liang H, Simonsen S, Schmitz C, Huber T, Messerle BA, Otting G (2008a) A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy. J Am Chem Soc 130:10486–10487

Su XC, McAndrew K, Huber T, Otting G (2008b) Lanthanide-binding peptides for NMR measurements of residual dipolar couplings and paramagnetic effects from multiple angles. J Am Chem Soc 130:1681–1687

Svergun DI, Petoukhov MV, Koch MHJ (2001) Determination of domain structure of proteins from X-ray solution scattering. Biophys J 80:2946–2953

Swarbrick JD, Ung P, Chhabra S, Graham B (2011a) An iminodiacetic acid based lanthanide binding tag for paramagnetic exchange NMR spectroscopy. Angew Chem Int Ed Engl 50:4403–4406

Swarbrick JD, Ung P, Su XC, Maleckis A, Chhabra S, Huber T, Otting G, Graham B (2011b) Engineering of a bis-chelator motif into a protein alpha-helix for rigid lanthanide binding and paramagnetic NMR spectroscopy. Chem Commun (Camb) 47:7368–7370

Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a diluite liquid crystalline medium. Science 278:1111–1114

Tjandra N, Kuboniwa H, Ren H, Bax A (1995) Rotational dynamics of calcium-free calmodulin studied by ^{15}N-NMR relaxation measurements. Eur J Biochem 230:1014–1024

Tolman JR (2001) Dipolar couplings as a probe of molecular dynamics and structure in solution. Curr Opin Struct Biol 11:532–539

Tolman JR, Ruan K (2006) NMR residual dipolar couplings as probes of biomolecular dynamics. Chem Rev 106:1720–1736

Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. Proc Natl Acad Sci USA 92:9279–9283

Tonks NK (2006) Protein tyrosine phosphatases: from genes, to function, to disease. Nat Rev Mol Cell Biol 7:833–846

Torchia DA (2015) NMR studies of dynamic biomolecular conformational ensembles. Prog Nucl Magn Reson Spectrosc 84–85:14–32

Valafar H, Prestegard JH (2004) REDCAT: a residual dipolar coupling analysis tool. J Magn Reson 167:228–241

Wang H, Eberstadt M, Olejniczak ET, Meadows RP, Fesik SW (1998) A liquid crystalline medium for measuring residual dipolar couplings over a wide range of temperatures. J Biomol NMR 12:443–446

Wöhnert J, Franz KJ, Nitz M, Imperiali B, Schwalbe H (2003) Protein alignment by a coexpressed lanthanide-binding tag for the measurement of residual dipolar couplings. J Am Chem Soc 125:13338–13339

Yagi H, Maleckis A, Otting G (2013a) A systematic study of labelling an alpha-helix in a protein with a lanthanide using IDA-SH or NTA-SH tags. J Biomol NMR 55:157–166

Yagi H, Pilla KB, Maleckis A, Graham B, Huber T, Otting G (2013b) Three-dimensional protein fold determination from backbone amide pseudocontact shifts generated by lanthanide tags at multiple sites. Structure 21:883–890

Zhang Y, Zuiderweg ER (2004) The 70-kDa heat shock protein chaperone nucleotide-binding domain in solution unveiled as a molecular machine that can reorient its functional subdomains. Proc Natl Acad Sci USA 101:10272–10277

Zhang Q, Throolin R, Pitt SW, Serganov A, Al Hashimi HM (2003) Probing motions between equivalent RNA domains using magnetic field induced residual dipolar couplings: accounting for correlations between motions and alignment. J Am Chem Soc 125:10530–10531

Zhuang T, Lee HS, Imperiali B, Prestegard JH (2008) Structure determination of a Galectin-3-carbohydrate complex using paramagnetism-based NMR constraints. Protein Sci 17:1220–1231

Zweckstetter M (2008) NMR: prediction of molecular alignment from structure using the PALES software. Nat Protoc 3:679–690

Zweckstetter M, Bax A (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. J Am Chem Soc 122:3791–3792

Zweckstetter M, Bax A (2001) Characterization of molecular alignment in aqueous suspensions of Pf1 bacteriophage. J Biomol NMR 20:365–377

Supplementary Material

Table S1. MaxOR for the 5 regions with centers shown in Figure 8C and their pairwise combinations. The colors refer to the color of the conformations depicted in Fig. 8C.

	A (red)	B (green)	C (blue)	D (yellow)	E (magenta)
A	0.54	0.90	0.90	0.64	0.63
B	-	0.53	0.58	0.78	0.96
C	-	-	0.55	0.77	1.00
D	-	-	-	0.50	0.68
E	-	-	-	-	0.60

Table S2. MaxOR for the 4 regions with centers shown in Figure S3 and their pairwise combinations. The colors refer to the color of the conformations depicted in Fig. S3.

	A (red)	B (green)	C (blue)	D (yellow)
A	0.58	0.91	0.61	0.79
B	-	0.56	0.72	0.95
C	-	-	0.51	0.79
D	-	-	-	0.57

Table S3. MaxOR for the 4 regions with centers shown in Figure S4 and their pairwise combinations. The colors refer to the color of the conformations depicted in Fig. S4.

	A (red)	B (green)	C (blue)	D (yellow)
A	0.62	0.66	0.73	0.74
B	-	0.60	0.71	0.73
C	-	-	0.60	0.96
D	-	-	-	0.60

Table S4. Exchange between two states (extended and compact) with unequal populations: also in this case both SES and the MaxOcc/MaxOR analysis can recover the two conformations used to calculate the averaged data. In this simulation, the averaged datasets were calculated from ensembles composed of two CaM conformations, one compact and one extended, with weight of 80% or 20%. Using the SES approach, both the major and the minor conformations could be identified in all cases (using diamagnetic or paramagnetic data, with 80% weight for the extended or the compact conformation). The results of the MaxOcc/MaxOR analysis are provided in table: for each simulation the spatial and angular displacement of the centers of the ensembles of the best scoring structures from the centers of the actually sampled distribution is reported (in all cases, the major state was recovered with MaxOcc calculations, and the minor state required the use of the MaxOR approach).

Simulation	Restraint	Error	recovered major state (80%)		recovered minor state (20%)	
			translation [Å]	rotation [deg]	translation [Å]	rotation [deg]
80% closed state 20% open state	dRDC	1 Hz	0.0	0.0	2.45	5.2
	pRDC+PCS	1 Hz/0.01 ppm	1.12	2.6	5.4	5.1
80% open state 20% closed state	dRDC	1 Hz	0.0	0.0	3.65	10.7
	pRDC+PCS	1 Hz/0.01 ppm	2.70	7.8	1.24	2.0

Fig S1. SES recovery using pRDC. Color code: green - simulated structures in the centers of the regions, red, blue - reconstructed structures with higher and lower weight, respectively. In all cases the correct orientations are recovered, yet the positions are inaccurate. A) Simulation1, 3 Hz error, B) Simulation 2, 2 Hz error.

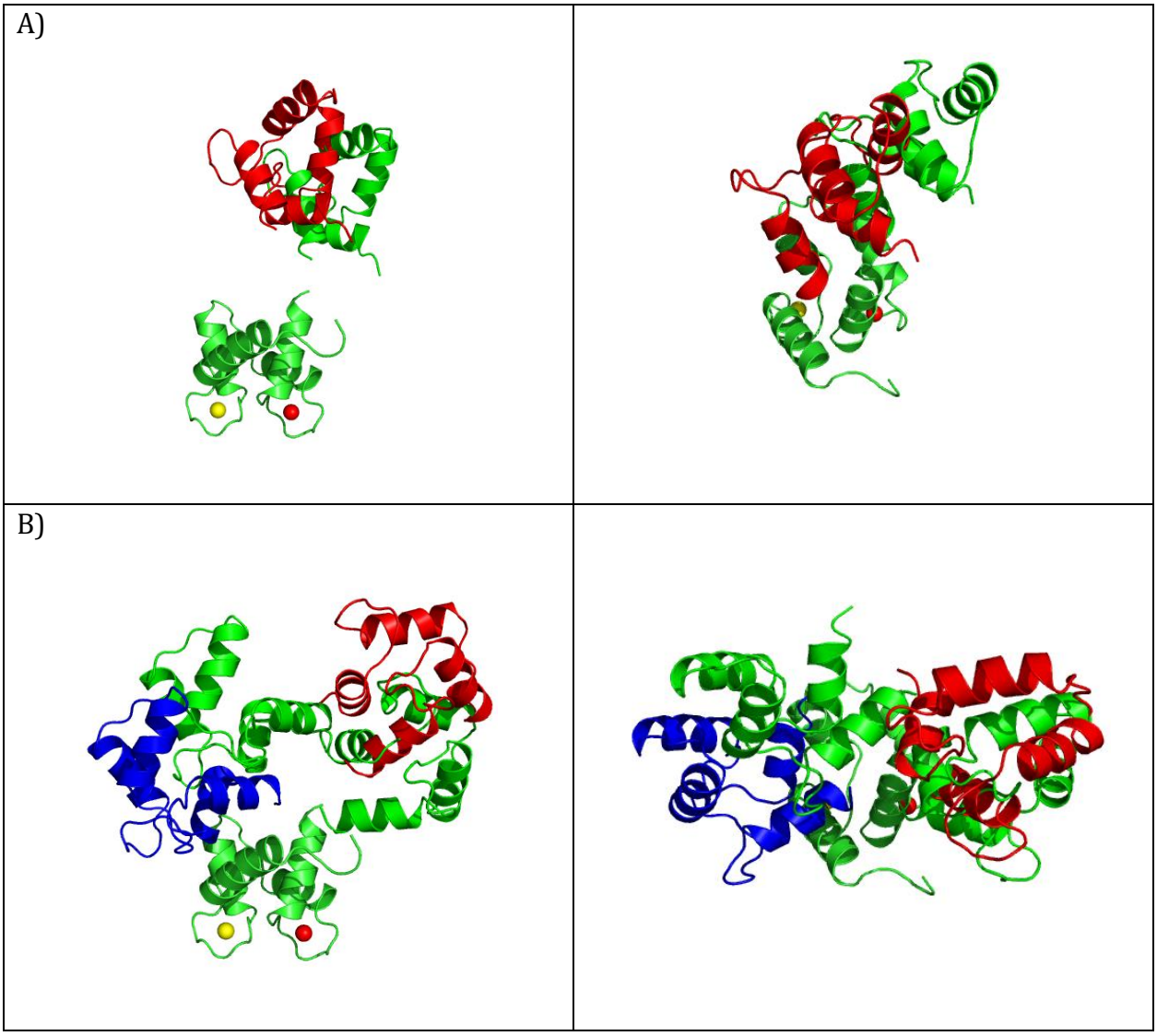


Fig S2. SES recovery using PCS. Color code: green - simulated structures in the centers of the regions, red, blue - reconstructed structures with higher and lower weight, respectively. A) Simulation2, 0.02 ppm error – the blue recovered structure exactly coincides with one of the original centers , B) Simulation 3, 0.01 ppm error- one 'average' state recovered instead of the two centers.

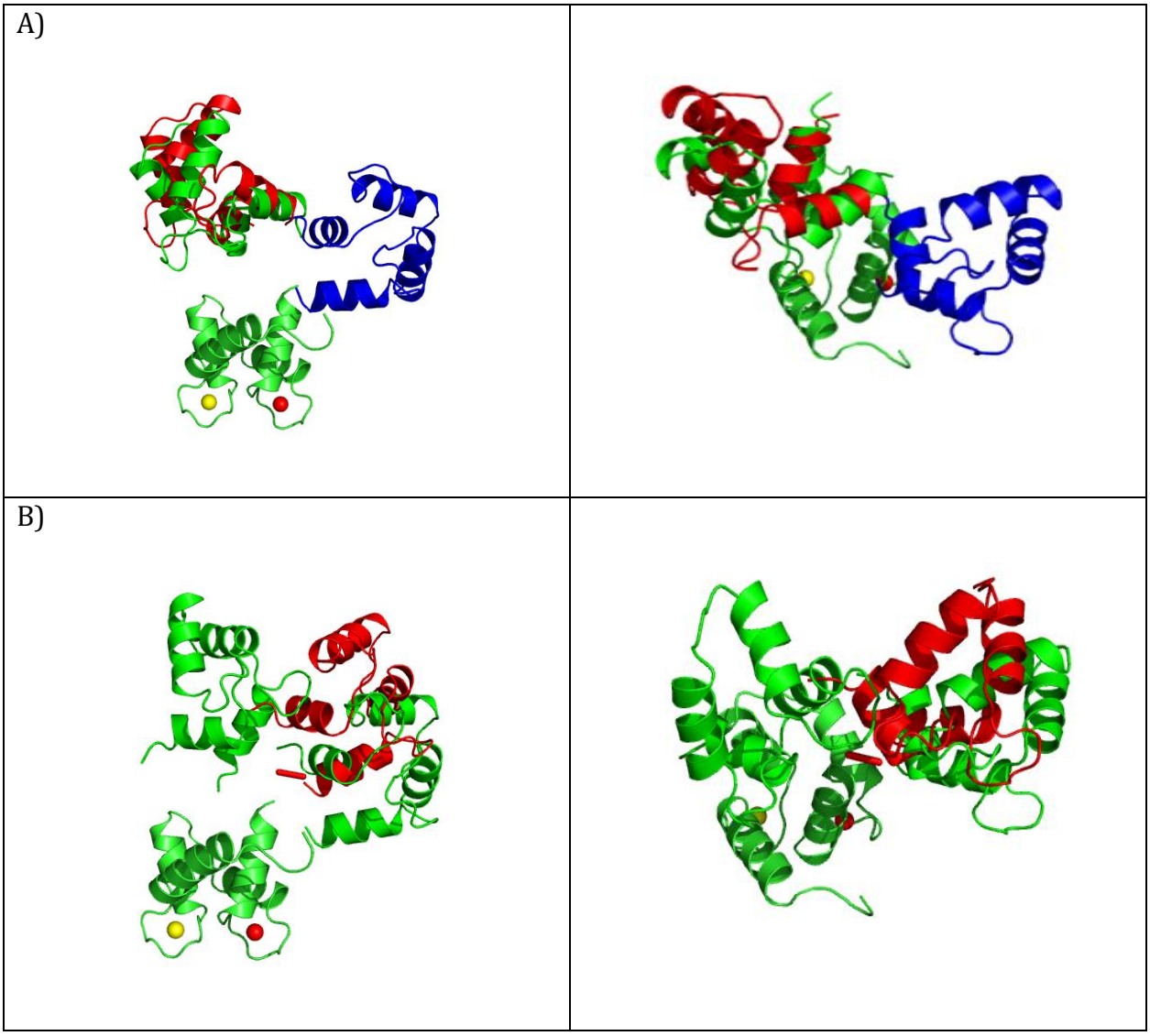
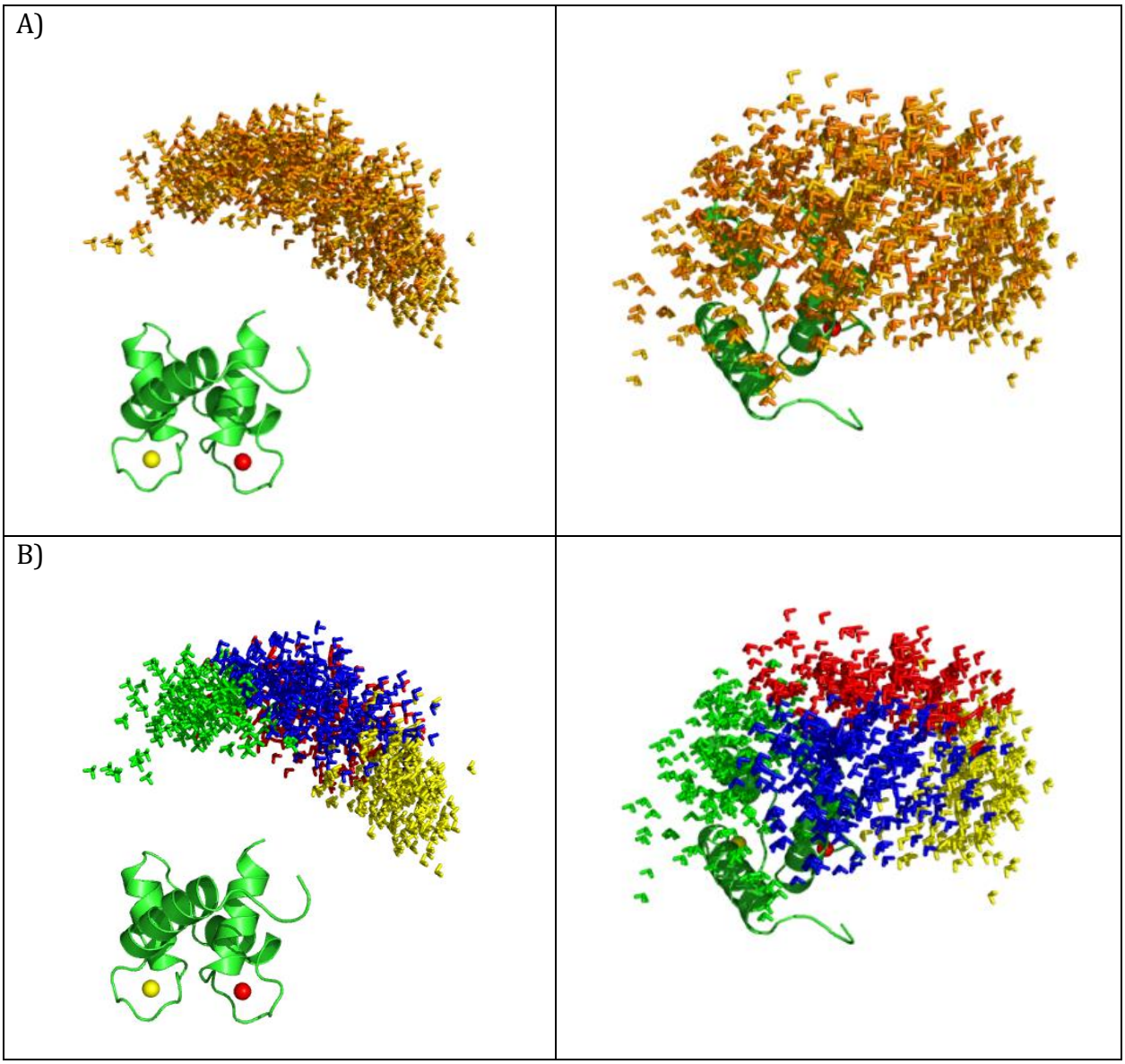
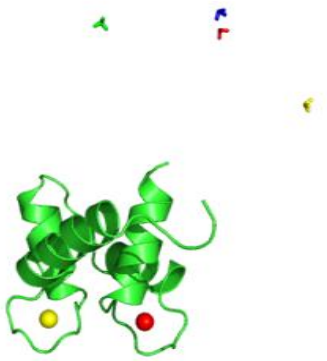
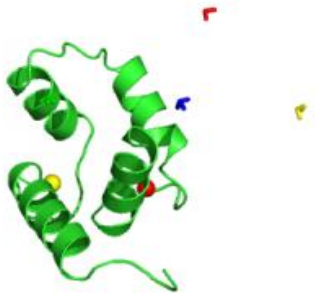


Fig S3. MaxOcc/MaxOR results for Simulation 3 with paramagnetic data (1 Hz error for pRDC and 0.01 ppm error for PCS). Each conformation is represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain, A) The structures with the highest MaxOcc, color code – according to the MaxOcc value (0.0-blue, 0.6-red) B) The clusters formed by the structures in panel A, C) The centers of the clusters, D) The centers of clusters chosen with MaxOR versus the centers of the simulated regions (black).

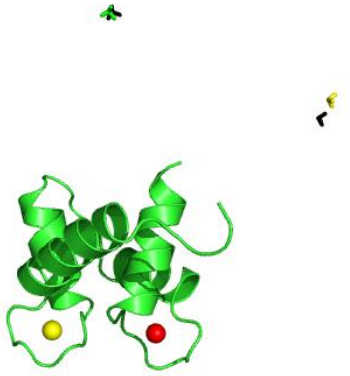


C)





D)



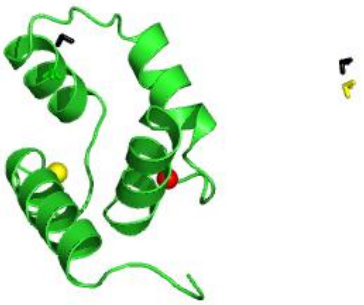
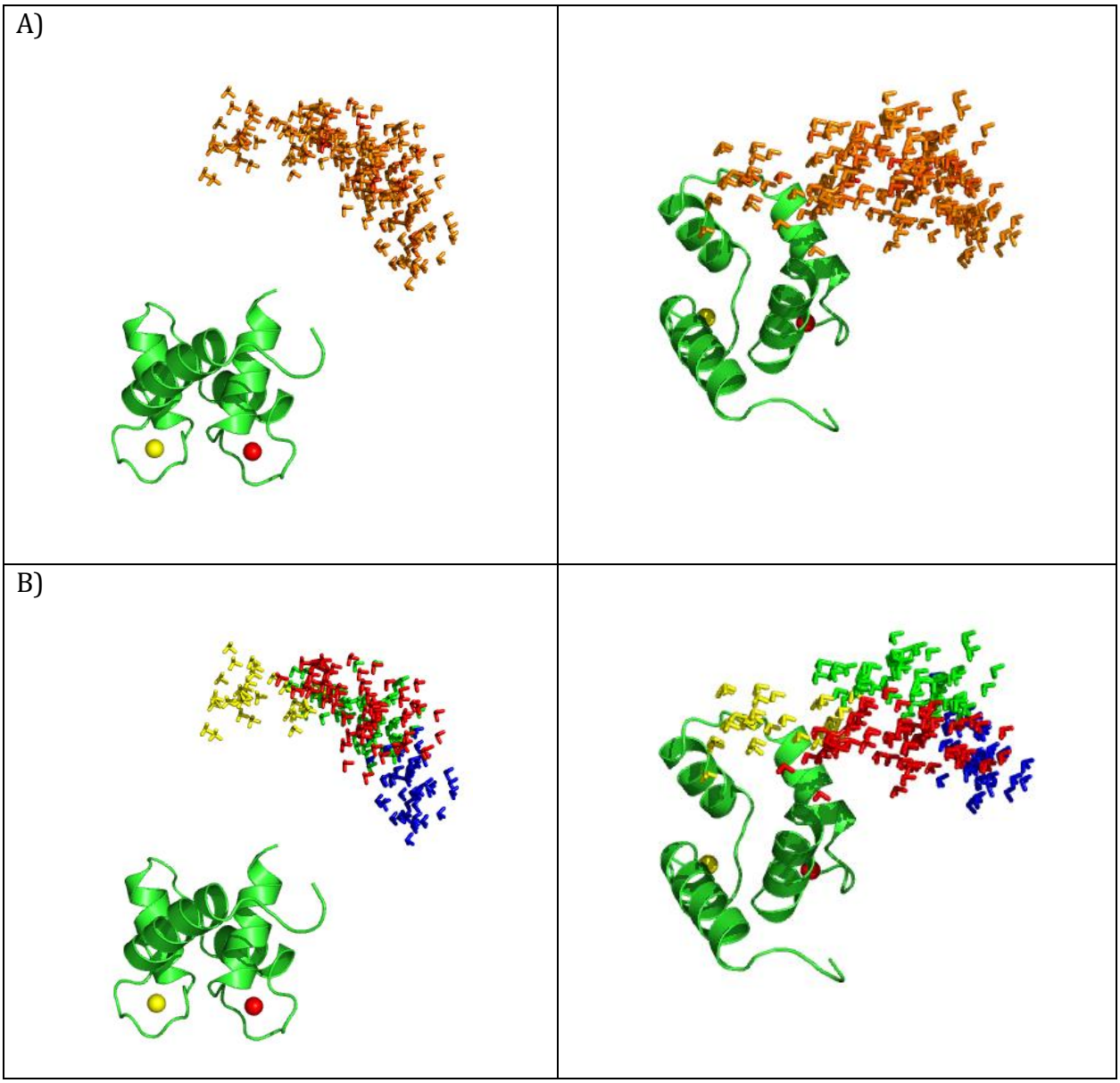
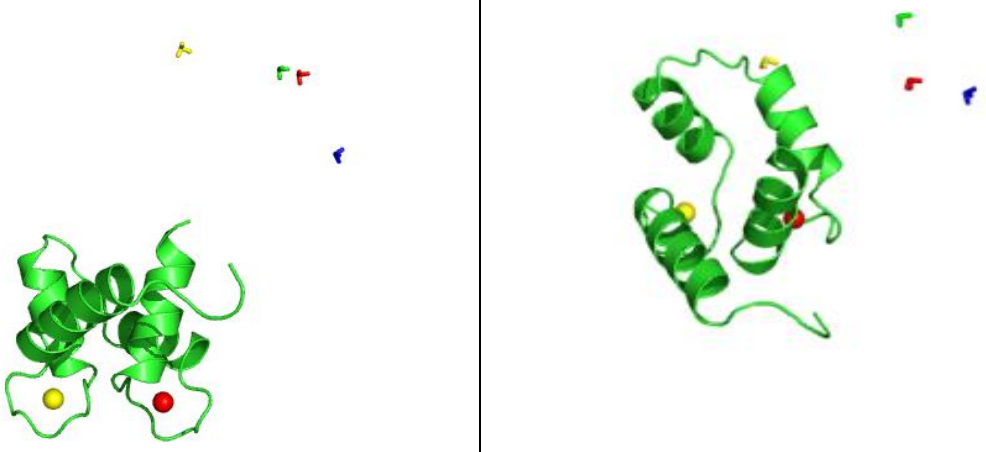


Fig S4. MaxOcc/MaxOR results for Simulation 3 with dRDC (1 Hz error). Each conformation is represented by a triad of Cartesian axes, centered at the center of mass of the C-terminal domain, A) The structures with the highest MaxOcc, color code – according to the MaxOcc value (0.0-blue, 0.6-red) B) The clusters formed by the structures in panel A, C) The centers of the clusters, D) The centers of clusters chosen with MaxOR versus the centers of the simulated regions (black).



C)



D)

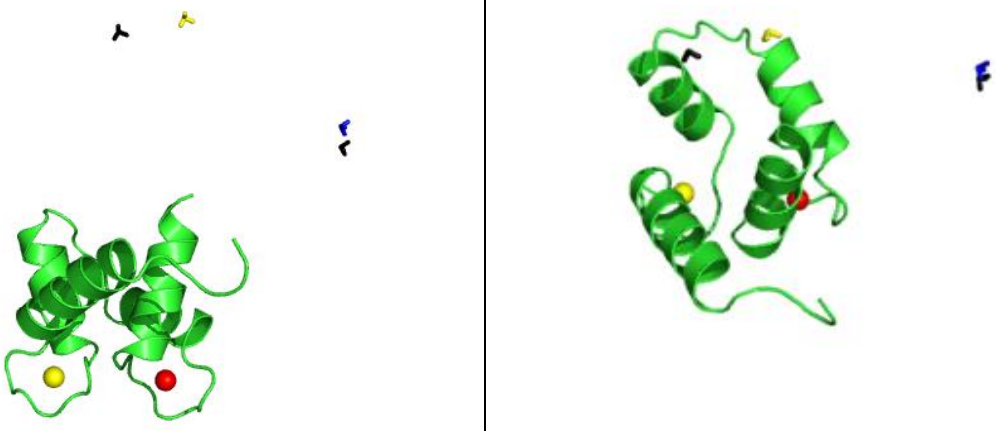


Fig S5. Any set of averaged RDC and PCS measurements can be seen as a point ($\overline{P}$) which can be reconstructed by a convex combination (i.e. a linear combination with weights summing up to 1) of sets of RDC and PCS values corresponding to a number of different protein conformations (P_1, P_2, P_3, in the figure, panel a). 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 $\overline{P}$. The SES method consists in reconstructing $\overline{P}$ using a smaller number of vertices, 2 in our example (panel b) that can still yield a solution. In fact, we can obtain a good approximation by letting the sum of the probabilities be less than 1: the combination of $\overline{P}_2$ and $\overline{P}_3$ reconstructing $\overline{P}$ as shown in panel b may be seen as a combination of $\overline{P}_2$ and $\overline{P}_3$ with weights scaled down according to the linear scaling factor between the triangles $O\overline{P}_2\overline{P}_3$ and OP_2P_3 (O is the origin).

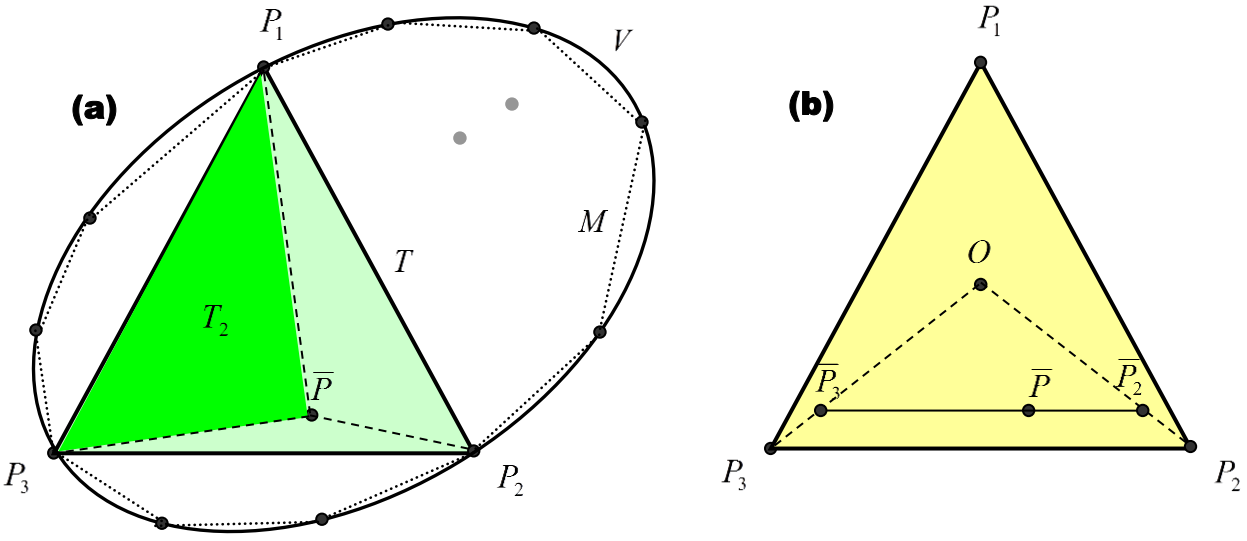
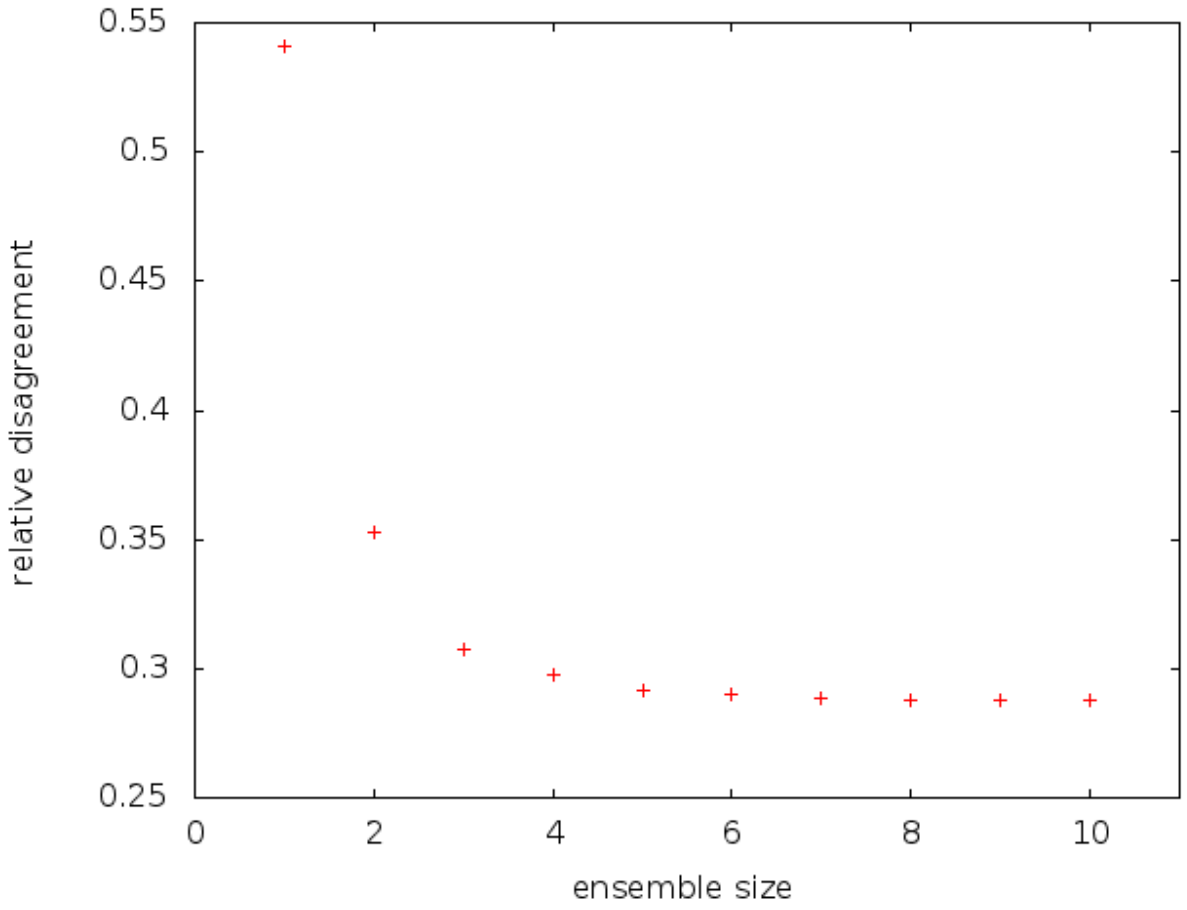


Figure S6. The number of conformations included in the ensemble determined with the SES method is determined using the L-curve method: a smoothing spline is fitted through the points and then the position of the maximum curvature is computed. In this example, the optimal size is given by 2 conformations.



3.3 Inter-helical conformational preferences of HIV-1 TAR-RNA from Maximum Occurrence Analysis of NMR data and molecular dynamics simulations

Witold Andrałojć,[1] Enrico Ravera,[1,2] Loïc Salmon,[3] Giacomo Parigi,[1,2] Hashim M. Al-Hashimi,[4] Claudio Luchinat[1,2]

[1]Magnetic Resonance Center "CERM", University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino (FI), Italy

[2]Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino (FI), Italy

[3] Department of Molecular, Cellular and Developmental Biology and Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI 48109

[4] Department of Biochemistry and Department of Chemistry, Duke University School of Medicine, 307 Research Drive, Durham, North Carolina 27710, United States

*Phys. Chem. Chem. Phys, **2016**, 18(8), pp 5743-5752*

This page intentionally left blank

PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2016, **18**, 5743

Inter-helical conformational preferences of HIV-1 TAR-RNA from maximum occurrence analysis of NMR data and molecular dynamics simulations†

Witold Andrałojć,[a] Enrico Ravera,[ab] Loïc Salmon,[c] Giacomo Parigi,*[ab] Hashim M. Al-Hashimi[d] and Claudio Luchinat[ab]

Received 9th July 2015,
Accepted 7th September 2015

DOI: 10.1039/c5cp03993b

www.rsc.org/pccp

Detecting conformational heterogeneity in biological macromolecules is a key for the understanding of their biological function. We here provide a comparison between two independent approaches to assess conformational heterogeneity: molecular dynamics simulations, performed without inclusion of any experimental data, and maximum occurrence (MaxOcc) distribution over the topologically available conformational space. The latter only reflects the extent of the averaging and identifies regions which are most compliant with the experimentally measured NMR Residual Dipolar Couplings (RDCs). The analysis was performed for the HIV-1 TAR RNA, consisting of two helical domains connected by a flexible bulge junction, for which four sets of RDCs were available as well as an 8.2 μs all-atom molecular dynamics simulation. A sample and select approach was previously applied to extract from the molecular dynamics trajectory conformational ensembles in agreement with the four sets of RDCs. The MaxOcc analysis performed here identifies the most likely sampled region in the conformational space of the system which, strikingly, overlaps well with the structures independently sampled in the molecular dynamics calculations and even better with the RDC selected ensemble.

Introduction

The fundamental importance of extensive conformational dynamics for allowing non-coding RNAs to carry out a wide variety of regulatory functions is well recognised.[1–4] RNA secondary structure consists of stable A-form helical domains that are connected by bulges, internal loops, and higher order junctions. Such helix–junction–helix (HJH) motifs play essential roles in the folding and biological function of non-coding RNAs. They are often points of significant flexibility that guide large adaptive changes in the orientation of helical domains and RNA global structure during folding, ribonucleoprotein assembly, and catalysis. HJH motifs also serve as binding sites for proteins, small molecules, and metal ions. Characterizing the extent and nature of inter-helical flexibility across HJH motifs is of primary importance for understanding the physical principles underlying RNA folding and recognition.[5] However, due to the biophysical properties of RNA it remains a major challenge. First collecting rich NMR datasets such as residual dipolar couplings (RDCs) is limited by the difficulties of obtaining significantly independent alignment.[6] Then, the presence of large internal motions, couples the internal dynamics to the overall diffusive or alignment properties of the RNA, complicating the interpretation of NMR spin relaxation[7–10] or RDC.[11–17] Finally, due to the potentially complex conformational dynamics, recovering an ensemble from experimental data remains an under-determined problem.[4,15,16,18–23]

The transactivation response element (TAR) RNA from the HIV-1 virus is a well-studied RNA drug target that plays essential roles during viral replication.[17,24,25] TAR consists of two A-form helical domains connected by a flexible three residue bulge linker. In previous work, each of the two TAR helices were independently elongated as a means of decoupling internal and overall motions.[7,17] This made it possible to interpret RDCs in terms of inter-helical motions since the elongated helix dominates the overall alignment. In particular, the measured RDCs could be interpreted in terms of motions of the short helix relative to the elongated one. The RDCs measured on two independently elongated TAR samples made it possible to characterize inter-helical motions with 3D orientation sensitivity.[15,17,26]

[a] *Magnetic Resonance Center "CERM", University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino (FI), Italy. E-mail: parigi@cerm.unifi.it*
[b] *Department of Chemistry "Ugo Schiff", University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino (FI), Italy*
[c] *Department of Molecular, Cellular and Developmental Biology and Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI 48109, USA*
[d] *Department of Biochemistry and Department of Chemistry, Duke University School of Medicine, 307 Research Drive, Durham, North Carolina 27710, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5cp03993b

More recently, we showed the feasibility of using a shape-based prediction[27–30] of the alignment tensor approach for treating couplings between internal and overall motions.[31] This made it possible to integrate additional RDCs measured in partially elongated TAR samples in the determination of atomic resolution ensembles. Such ensembles were composed of conformations selected from a conformational pool obtained using an 8.2 μs molecular dynamics calculation[31] computed using the CHARMM36 force-field.[32–34] From this long MD trajectory conformational ensembles in agreement with the experimental RDC data were selected.[31] This approach permitted to extract from the whole pool of structures determined by MD calculation, the conformations which may better represent the conformational variability of the system. The selected structures clustered into three distinct states, separated by large transitions in inter-helical orientations, coupled to local melting of base-pairs near the junction. The RDC-selected ensemble included conformations that bear strong resemblance to the ligand bound conformations of TAR.

We here apply a different approach for the analysis of the averaged experimental RDCs, based on the compliance of each and any sterically-allowed conformation with respect to the average experimental data. The method, called maximum occurrence (MaxOcc),[35,36] aims at identifying conformations that can exist for a large share of the time; this is done by assigning to each conformation the maximum time that it can exist and be in agreement with the experimental observation,[37–39] when taken together with an arbitrary number of other conformations. Thus it is possible to identify the conformations, which must necessarily have a negligibly small weight and those which may have a large weight, whatever the real ensemble of conformations experienced by the RNA is. We have previously demonstrated by synthetic tests that the conformations used to construct synthetic ensembles are found to have a high MaxOcc.[35,40,41] The analysis was performed without taking advantage of the MD calculations, *i.e.* without restricting the possible RNA conformations to the pool of structures sampled by the MD trajectory. Strikingly, the RNA structures with large MaxOcc define a conformational region in substantial overlap with the structures sampled in the MD calculations, indicating good convergence between the MD results and the MaxOcc analysis. Furthermore, the previously determined structural ensembles selected from the MD trajectory[31] is on average even closer to the most likely region of the MaxOcc landscape.

Materials and methods

Experimental RDC datasets

The experimental RDC data measured using the Pf1 phage alignment medium for four constructs of HIV-1 TAR RNA (non-elongated, with the first helix elongated by 3 base pairs, with either the first or the second helix elongated by 22 base pairs) were previously published.[11,17,31,42] The helix elongation causes a strong modulation of the alignment of the RNA strand, leading to a high degree of independence of the different sets of RDCs.

In the study we analyzed the one bond couplings measured between the sugar C1′–H1′, C2′–H2′, C3′–H3′, and C4′–H4′ and base C2–H2, C5–H5, C6–H6, C8–H8, C5–C6, N1–H1, and N3–H3 pairs of atoms for nucleotides in both helical regions. The data measured for the A22–U40 base pair was omitted in the current analysis due to previously reported conformational flexibility of this base pair.[31]

Generation of the pool of conformers and prediction of RDCs

The MaxOcc analysis of HIV-1TAR was performed using the broadest possible topologically allowed conformational space obtained through exhaustive sampling of inter-helical Euler angles[43] in increments of 5°, excluding the orientations violating loose sterical and stereochemical restraints.[17] The two separately well-folded regions were assumed to adopt idealized A-form helical structures and the bulge nucleotides were not explicitly modelled in this study. For each conformer, the 4 sets of RDCs were predicted using the PALES software.[27] A steric description was used based on the cylindrical wall model with an effective low concentration (0.022 g mL^{-1}) as no significant improvement of the alignment tensor prediction was observed for nucleic acids when the electrostatic model is used.[44] To model the alignment of constructs that feature elongation of one of the helices, the proper number of base pairs was added to the initial structure assuming idealized A-form geometry. The helix II is capped by a UUCG apical loop corresponding to the sequence of the experimentally used TAR constructs.

Euler angle definition

The Euler angles were defined as previously described.[43] In this definition α_h, γ_h and β_h varies from −180 to 180. Other common Euler angle conventions may have β_h restricted to only positive values. The degeneracy introduced by this broader definition of β_h is lifted by choosing the solution that minimizes $\delta = \sqrt{{\alpha_h}^2 + {\beta_h}^2 + {\gamma_h}^2}$.

MaxOcc calculations

The calculation of MaxOcc of each selected conformer is performed by finding optimized ensembles that yield the best agreement with experimental observables, while containing the selected conformer with a given weight. The calculation is repeated for a different weight of the same conformer. As this weight is increased, the agreement with the experimental data may start to deteriorate. The weight at which the quality of the fit reaches a fixed threshold corresponds to the MaxOcc of that conformer, *i.e.* to the highest weight that it could have in any ensemble that explains the experimental data. The target function used in the fit has the form of the quality factor Q.[45] The best fit obtainable without applying any restraint to the weight of the conformers had a Q of 0.22 (corresponding to $\chi^2 \approx 1.55$). A fit was considered good if the corresponding Q was below a threshold defined 20% higher than the lowest Q of 0.22, that is 0.264 (this corresponds to a χ^2 close to 2.0; as it is only Q, not χ^2, that is optimized, the latter rises slightly faster).

When external alignment RDCs are used as experimental observables, the problem of finding an optimized ensemble

with one structure (labelled j) present at a fixed weight (x_{MO}) can be expressed as

$$\underset{x,c_1,\ldots c_K}{\operatorname{argmin}}\left\{\|\mathbf{A}(c_1,\ldots,c_k)\mathbf{x}-\mathbf{y}\|_2{}^2+\lambda\left[\left(x_{MO}-x_j\right)^2-\left(1-x_{MO}-\sum_{i=1,i\neq j}^{N}x_i\right)^2\right]\right\}\quad\text{subject to }\mathbf{x}\geq 0 \tag{1}$$

where $\mathbf{x}$ is the vector of the weights of the N structures composing the pool, $\mathbf{y}$ is the vector of M experimentally observed RDC values, normalized by their norm, $c_1,\ldots c_k$ are the scaling factors between the experimental and back-calculated RDC for each of the k constructs (required because the magnitude of alignment induced by the anisotropic solution is not known exactly, and may differ from the one assumed in the PALES calculation), λ is a weighting factor, and $\mathbf{A}(c_1,\ldots c_k)$ is the M × N matrix whose columns contain the RDC values back-calculated for each of the conformers, again normalized by the norm of the experimental RDC data. The $\mathbf{A}(c_1,\ldots c_k)$ matrix is created by stacking the submatrices $\mathbf{A}_n$ containing back-calculated RDCs of single constructs multiplied by the appropriate scaling factors c_n:

$$\mathbf{A}(c_1,\ldots,c_K)=\begin{bmatrix}c_1\mathbf{A}_1\\ c_1\mathbf{A}_2\\ \vdots\\ c_k\mathbf{A}_k\end{bmatrix}.$$

The $\mathbf{y}$ vector and the $\mathbf{A}$ matrix were normalized in such a way that the $\|\mathbf{Ax}-\mathbf{y}\|_2{}^2$ term corresponds to the square of the Q factor between the experimental and back-calculated data. The value of λ, set to 10 in the present calculations, is found with the L-curve method, as a compromise between a good fit of the experimental observables and the proximity of the sum of the weights to 1.[41]

The problem as expressed in eqn (1) would require a non-linear minimization due to the presence of the unknown $c_1,\ldots,c_k$ factors. It becomes linear if the scaling factors $c_1,\ldots,c_k$ are fixed to constant values. The optimal values of $c_1,\ldots,c_k$ for a given back-calculated data vector $\mathbf{y_{calc}} = \mathbf{Ax}$ (arising either from a single structure or an ensemble) can be readily calculated as $c_{opt}=\frac{y_{calc}\cdot y}{y\cdot y}$. A minimization procedure was thus applied which involved making an initial guess of the values of scaling factors $c_1,\ldots,c_k$, solving eqn (1) for $\mathbf{x}$ (with fixed $c_1,\ldots,c_k$) using a non-negative least squares method (a frugal coordinate descent algorithm,[46] combined with random coordinate search[47]), then calculating the optimal $c_1,\ldots,c_k$ for the present $\mathbf{x}$ vector, and finally using them as the fixed scaling factors in the next iteration of non-negative least squares minimization, in an iterative fashion, until convergence of the $c_1,\ldots,c_k$ values was reached.

MaxOR calculations

The MaxOcc analysis of interdomain mobility can yield additional insights into the details of the sampled conformational subspace if it is supplemented by maximum occurrence of regions (MaxOR) calculations.[40] This method, which is the natural extension of the MaxOcc approach for single conformations to conformational regions, aims at determining the maximum amount of time for which a group of conformers can collectively exist in agreement with the averaged experimental data. To achieve this goal the algorithm described above is somewhat modified, according to eqn (2). Instead of fixing the weight of one conformer to the desired value x_{MO}, it is the sum of the weights of all conformers composing the chosen group that is fixed to x_{MO}:

$$\underset{x,c_1,\ldots c_k}{\operatorname{argmin}}\left\{\|\mathbf{A}(c_1,\ldots,c_k)\mathbf{x}-\mathbf{y}\|_2{}^2+\lambda\left[\left(x_{MO}-\sum_{i\in C}x_i\right)^2-\left(1-x_{MO}-\sum_{i\in D}x_i\right)^2\right]\right\}\quad\text{subject to }\mathbf{x}\geq 0 \tag{2}$$

where C and D indicate the structures within and outside the selected group, respectively.

Results and discussion

Maximum occurrence of single conformers – a comparison with extensive MD

A pool containing all sterically-allowed RNA structures was generated by sampling all topologically allowed combinations of the inter-helical Euler angles α_h, β_h and γ_h, defining the inter-domain orientation of the two RNA domains, in steps of 5° for each angle separately. The three angles (see Fig. 1) represent

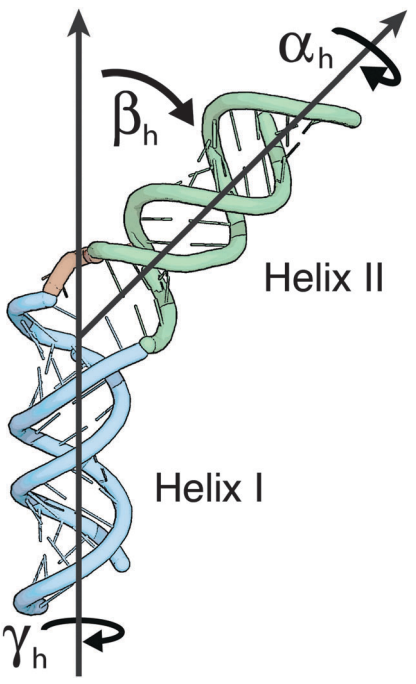


Fig. 1 Angles (α_h, β_h, γ_h) inter-helical Euler angles defining the inter-domain orientation of the two RNA domains: α_h and γ_h report on the twisting of the first and second helices around their respective axis, respectively, and β_h on the inter-helical bending.

the twisting of the upper (helix II) and lower (helix I) helices around their respective axes (α_h and γ_h) and the inter-helical bending (β_h).[43] For each of the conformations in the generated pool (37 005 structures) the MaxOcc value was calculated using the implementation of the maximum occurrence method described in Materials and methods section. The obtained MaxOcc values show a considerable spread over the pool of conformations (from 17% to 70%) indicating that indeed specific structures are much more compliant with the experimental data than others. The fine sampling of the conformational space permits to observe that MaxOcc is a smooth function of the three inter-helical Euler angles. Fig. 2 and 4a show the 2D projections of the MaxOcc function on different pairs of inter-helical angles. It can be easily appreciated that the structures with the highest MaxOcc are grouped into a single well-defined conformational region, with a peak at around $-10 < \alpha_h < 5^\circ$, $45 < \beta_h < 55^\circ$, $-15 < \gamma_h < 5^\circ$, centered at $\alpha_h = -5^\circ$, $\beta_h = 50^\circ$ and $\gamma_h = -5^\circ$. To ease the understanding of the 3D shape of the high MaxOcc region, a 3D representation is given in Fig. 2d. Additional structures with intermediate-high MaxOcc values (up to 50%) appear at close to $\beta_h = 180^\circ$. They correspond most likely to a non-physical solution,[48] whose high MaxOcc value arises from inherent degeneracy of the RDC data.[16,49–51]

In a previous work[31] the HIV-1 TAR RNA was studied by means of an 8.2 μs MD simulation. Interestingly when the coordinates of the structures constituting the MD are superimposed to the MaxOcc profile it appears that practically the entire MD trajectory falls inside the identified high MaxOcc region (Fig. 3 and 4b). It is a very encouraging result that two completely independent approaches suggest similar conformational sampling for the system in question.

Even though the geometric center of the MD trajectory (the averaged Euler angles over the whole MD simulation are $\alpha_h = -22^\circ$, $\beta_h = 32^\circ$ and $\gamma_h = -57$) is somewhat shifted with respect to the peak of the MaxOcc profile, one has to keep in mind that the MD trajectory taken as such fits the experimental RDC data rather poorly ($\chi^2 = 6.03$). It is actually possible that, despite the overall sampling of conformations is correctly reconstructed by the MD, the populations of the specific conformational regions are not correctly represented, as already pointed out,[4,31] owing to a lack of convergence or to imperfection in the applied force field. It is worth noting that the MD trajectory treats both local and global degrees of freedom, while the approach proposed

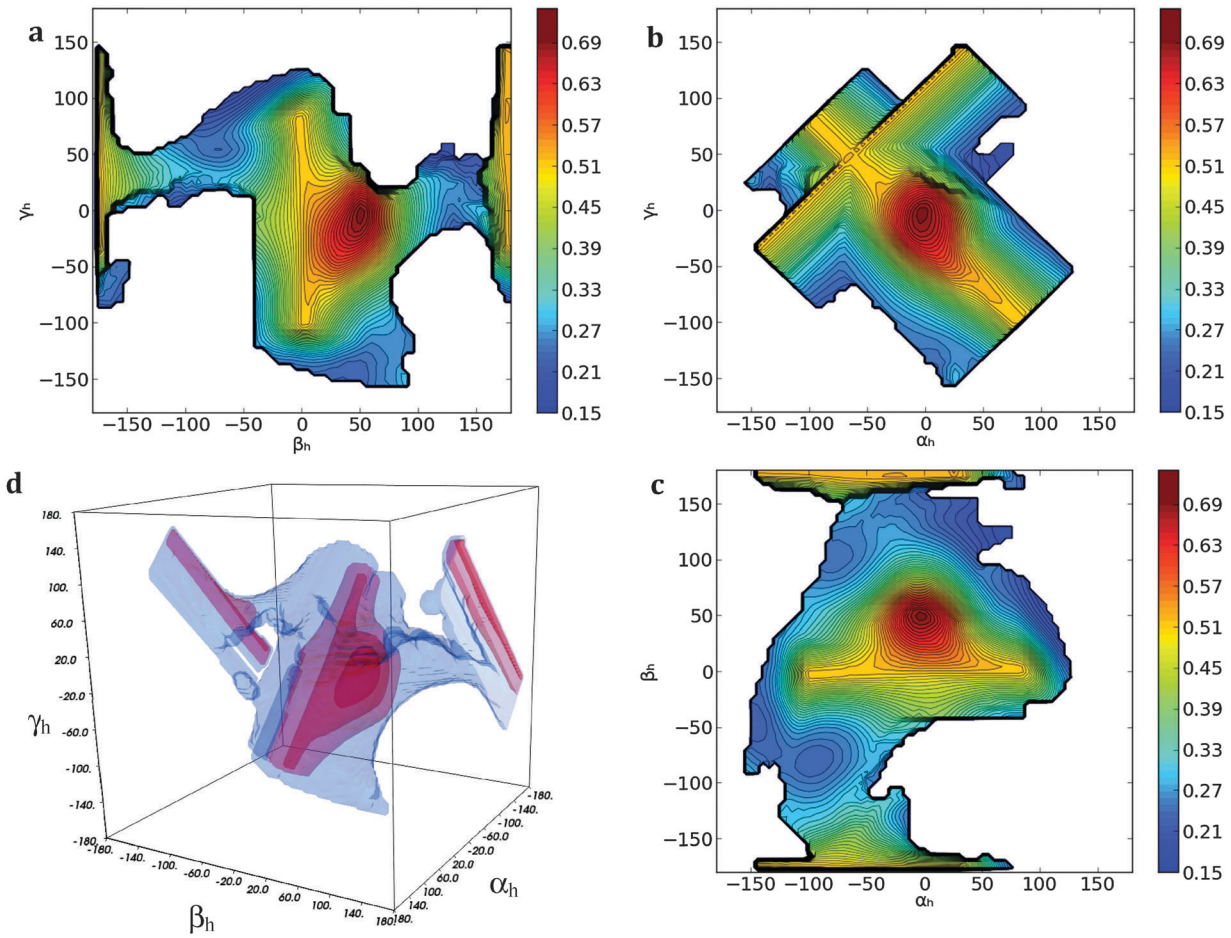


Fig. 2 (a–c) The MaxOcc landscape (MaxOcc values color coded) as a function of α_h, β_h and γ_h angles (2D projections). White areas correspond to not sampled regions. (d) 3D representation of the full sampled space (blue) and of the area which encompasses high MaxOcc conformations (outer red surface, MaxOcc > 0.4; middle red surface, MaxOcc > 0.5; inner red surface, MaxOcc > 0.6).

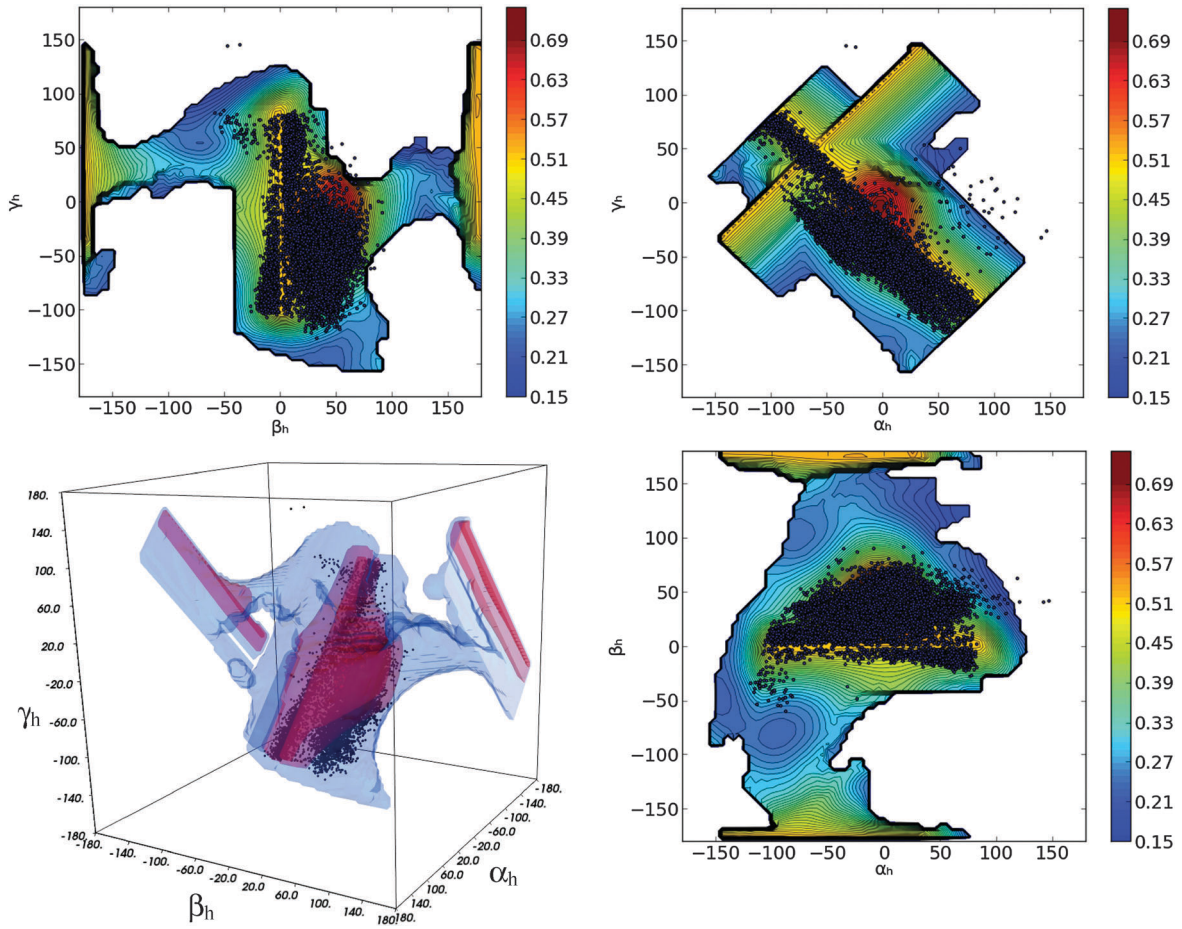


Fig. 3 Superimposition of MD trajectory (dark dots) to the MaxOcc landscape (color coded) as a function of α_h, β_h and γ_h angles (2D projections and 3D representation). White areas correspond to not sampled regions.

here only considers the conformational dynamics of the bulge. The possibility of imperfect weighing of the MD trajectory was already explored using a sample and select (SAS) approach[4,26,31] to properly reweight different regions of the MD trajectory. Small ensembles that fit well the experimental RDCs were repeatedly selected from the original trajectory and then combined to provide the 'RDC reweighted ensemble'. Interestingly, the geometric center of this reweighted trajectory is located much closer to the MaxOcc peak (the average values of the Euler angles for the SAS ensemble are $\alpha_h = -15^\circ$, $\beta_h = 52^\circ$ and $\gamma_h = -28^\circ$) than the original MD trajectory (Fig. 5). The improved agreement between the MaxOcc analysis and the MD sampling when the latter is adjusted using experimental information may not seem surprising, yet it should not to be taken for granted due to the under-determination of the recovery problem, the differences in the assumptions used in the two approaches, and the different physical meaning of the conformations selected by the two approaches. The fact that the MaxOcc and SAS methods actually favor a similar region of the conformational space can be considered an additional cross-validation of the ensemble previously extracted from the MD[31] and further suggests that indeed the structures located in this part of the conformational space are crucial for explaining the HIV-1 TAR conformational sampling in solution.

Seven distinct structures of HIV-1 TAR RNA bound to different small molecule ligands are available in the PDB.[17] When their coordinates are superimposed to the MaxOcc profile (Fig. 6a and Fig. S1, ESI†) it appears that also these structures are located either close to the peak of the MaxOcc function or on its shoulder towards lower values of β_h. This finding may suggest that ligand binding occurs by taking advantage of pre-existing conformations of HIV-1 TAR RNA, which are already highly populated in the conformational ensemble of the free nucleic acid.

Maximum occurrence of conformational regions

The MaxOcc analysis identified the part of the conformational space which contains the single structures that can explain the largest share of the experimental observables by themselves. However, even the structures with the highest MaxOcc can contribute only up to 70% to the conformational ensemble sampled by HIV-1 TAR RNA. The next question to ask is what is the smallest compact ensemble or the simplest mobility scheme which can account for the experimental observables. One of the simplest mobility schemes that one can conceive consists of a motion around a single center. The MaxOR approach was applied to quantify the smallest amount of conformational heterogeneity that has to occur around the peak of the MaxOcc profile in order

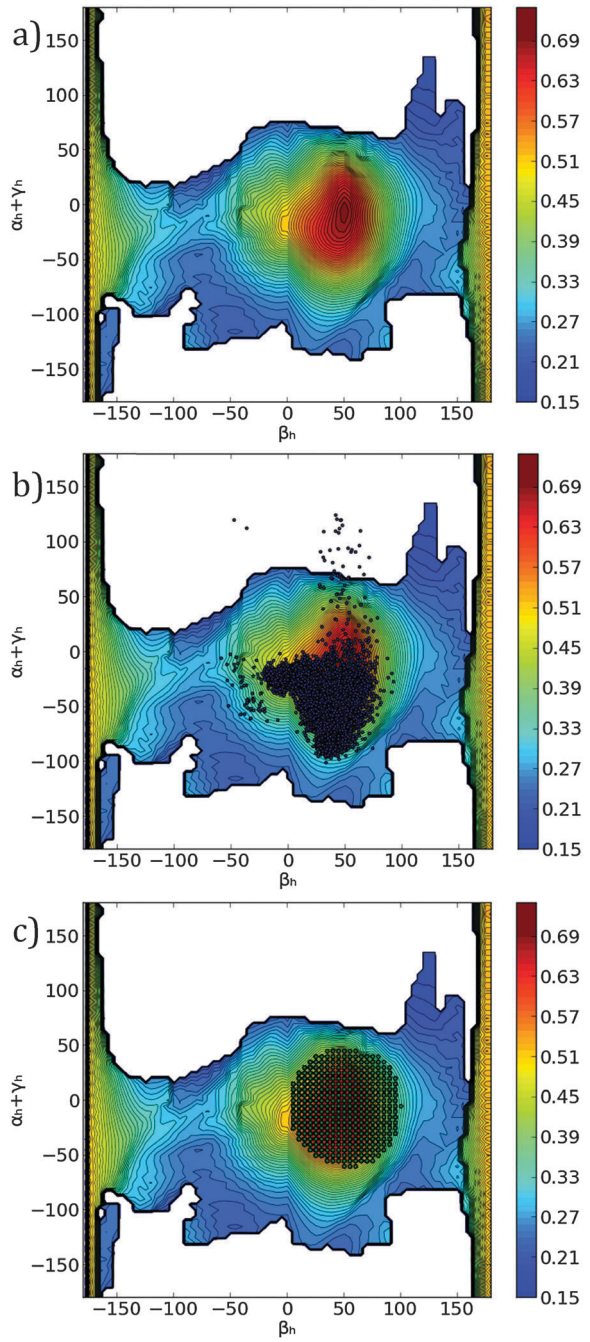


Fig. 4 (a) The MaxOcc landscape (MaxOcc values color coded) as a function of the β_h and $\alpha_h + \gamma_h$ coordinates. White areas correspond to not sampled regions. (b) Superimposition of the MD trajectory (dark dots) to the MaxOcc landscape. (c) The smallest region centered at the MaxOcc peak with MaxOR = 1 (green dotted area) superimposed to the MaxOcc landscape (color coded) in the (β_h, $\alpha_h + \gamma_h$) coordinates.

to obtain an ensemble which fully reproduces the experimental data. For this purpose, several regions were built around the conformation with the highest MaxOcc comprising all structures that can be obtained from the central conformation by changing the inter-helical orientation through a single axis rotation in any direction by less than a fixed angle (the quaternion representation of rotations was used at this step, because the Euler angle representation is not the best way to define distances between two structures). By increasing the maximum allowed rotation in steps of 10° and calculating the MaxOR of the corresponding regions, it was found that rotations up to 50° from the central conformer have to occur in order to obtain a MaxOR of 1 (*i.e.*: full agreement with the experimental data) (Fig. 4c). Thus if mobility in a symmetric region around a single center is assumed, inter-helical motions of high amplitude (the most distant conformations are 100° of rotation apart from one another) have to be considered to explain the experimental RDC values, in good agreement with initial studies of TAR dynamics.[42]

The size of the conformational space to be sampled by the system can likely be reduced if instead of an isotropic distribution around a single center, two or more separated, yet compact, regions are allowed to be explored.[40,41,52] In order to identify other compact regions in the conformational space that can best complement the MaxOcc peak, a broad series of MaxOR calculations was performed. In each calculation the considered region was composed of two parts: the structures composing the peak of the MaxOcc profile ($-10 < \alpha_h < 5°$, $45 < \beta_h < 55°$, $-15 < \gamma_h < 5°$) and another group of structures constituting a 5°·5° square in the (β_h, $\alpha_h + \gamma_h$) 2D projection of the conformational space. The second part of the region was changed in the different calculations in a systematic way in order to cover the whole (β_h, $\alpha_h + \gamma_h$) space. The results of the whole procedure, shown in Fig. S2 (ESI†), demonstrate that there exist only two compact areas in the (β_h, $\alpha_h + \gamma_h$) space which, when added to the MaxOcc peak, lead to a considerable increase of MaxOR. These two areas are located around $\beta_h = -40$, $\alpha_h + \gamma_h = -15$, and $\beta_h = 165$, $\alpha_h + \gamma_h = -20$. Because these regions are separated by an almost 180° rotation, it is probable that one of them arises from the inherent degeneracy of the RDC data (*i.e.* it is just a 'ghost'[51] of the other). As the region with high values of β_h is located close to the edge of the available conformational space, possibly hardly sterically allowed if a more physically accurate modelling of the bulge was applied,[48] it is quite safe to assume that this region is indeed a 'ghost' of the other region. Thus the MaxOR analysis shows that conformers situated around $\beta_h = -40°$, $\alpha_h + \gamma_h = -15°$ are the best suitable to complement the structures located close to the peak of MaxOcc, and when the two are taken together they are nearly enough to explain the whole experimental observables (MaxOR of the pair is 99%).

The size of the complementing region is, as said above, a 5°·5° square in the (β_h, $\alpha_h + \gamma_h$) 2D projection, yet it has the shape of a long rod in the whole (α_h, β_h, γ_h) 3D conformational space. In order to locate more precisely the actual structures responsible for the high MaxOR, such rod can be thus further subdivided into 5°·5°·5° cubes in the full 3D Euler angle space with the centers at $\alpha_h = x$, $\beta_h = -40$, $\gamma_h = -x - 15°$, where x runs over all the values of α_h sterically allowed at this point of space, in steps of 5°. Fig. S3 (ESI†) presents the MaxOR values of each cube as a function of the α_h angle. The MaxOR function has a single maximum at $\alpha_h = -15°$ (and $\gamma_h = 0°$) and its value at this point is only slightly lower (MaxOR = 97%) than when the whole rod is considered. The volume occupied by these regions is much smaller than the volume occupied by the single region with MaxOR equal to 1 identified before. Therefore, we have identified a pair of compact regions in

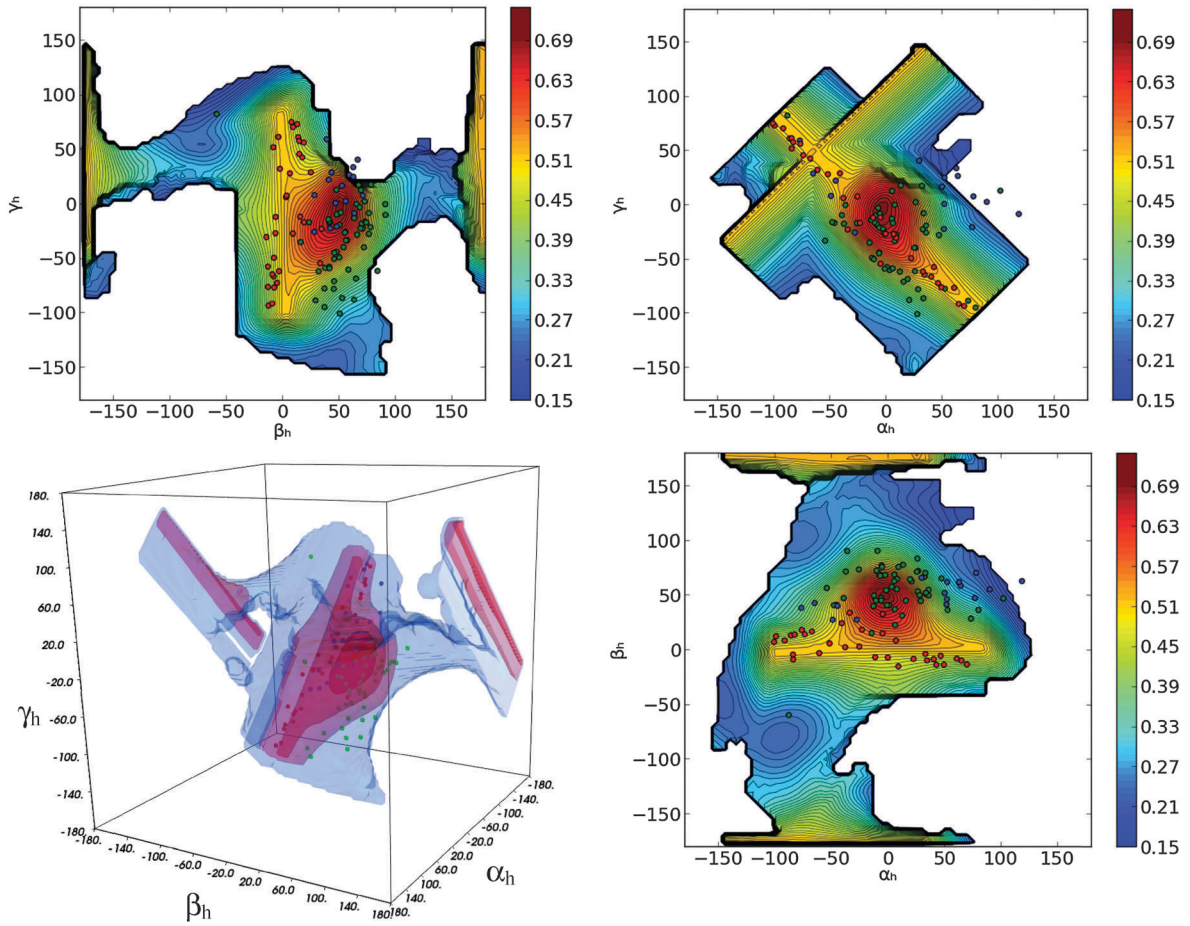


Fig. 5 Ensemble selected from the MD trajectory by SAS ('RDC reweighted trajectory'), divided into three clusters after the original paper (cluster 1 in green, cluster 2 in red and cluster 3 in blue), superimposed to the MaxOcc landscape (color coded) as a function of α_h, β_h and γ_h angles (2D projections and 3D representation). White areas correspond to not sampled regions.

the Euler angle space, one located at the peak of the MaxOcc profile and another at $\alpha_h = -15$, $\beta_h = -40$ $\gamma_h = 0^\circ$ (Fig. 6d), that constitute a compact conformational sampling able to fit the experimental data (MaxOR of 100% is easily obtainable with this pair by slightly increasing the size of either region).

The possible existence of other, clearly distinct, two-centered ensembles not containing a region close to the peak of the MaxOcc profile was examined by performing a series of additional calculations over all pairs of 2D regions of size of $20^\circ \cdot 20^\circ$ (Fig. 6b). Interestingly, all the two-centered ensembles with the highest maxOR (>95%) are composed of a region located in proximity of the MaxOcc peak (with the coordinates of their centers in the range $50 < \beta_h < 90$ and $-30 < \alpha_h + \gamma_h < 30$) and of another region very close either to the identified minor state ($-50 < \beta_h < 10$ and $-30 < \alpha_h + \gamma_h < -10$) or to its ghost solution described above. Therefore, although the positions of the two states may be subject to some uncertainty, yet the existence of any other distinct two-centered ensemble with high maxOR value can be excluded.

Comparison of MaxOR results and previous results

Having identified such a two-region scheme as the most compact ensemble capable of explaining the experimental averaged RDCs, one can re-examine Fig. 6a, where the positions of the ligand bound structures are shown. It can be noted that these structures (all except one) are either located within the regions defined by the two-center MaxOR calculations or in the conformational space between them.

In ref. 31, the conformations that were selected by the SAS algorithm from the MD trajectory could be clustered into three main states, on the basis of the bending angle and the inter-nucleotide distance between A22, the last nucleotide in helix I, and U23, the first nucleotide in the bulge. Whereas the present pool lacks the information about the inter-nucleotide distance, we could compare the location of the three clusters in the Euler angles space. The results of such a comparison are shown in Fig. 5 and 6c. Although the clusters selected from the MD are more spread than the MaxOcc peak, there is a clear similarity between the SAS cluster 1 (in green in Fig. 5 and 6c) and the main state identified by MaxOcc/MaxOR, and between the SAS cluster 2 (in red in Fig. 5 and 6c) and the minor state found by MaxOR. This correspondence is particularly striking if we extend the comparison to the generalized positions of the MaxOR regions shown in Fig. 6b. Also our qualitative identification of the major and minor states is in line with the relative

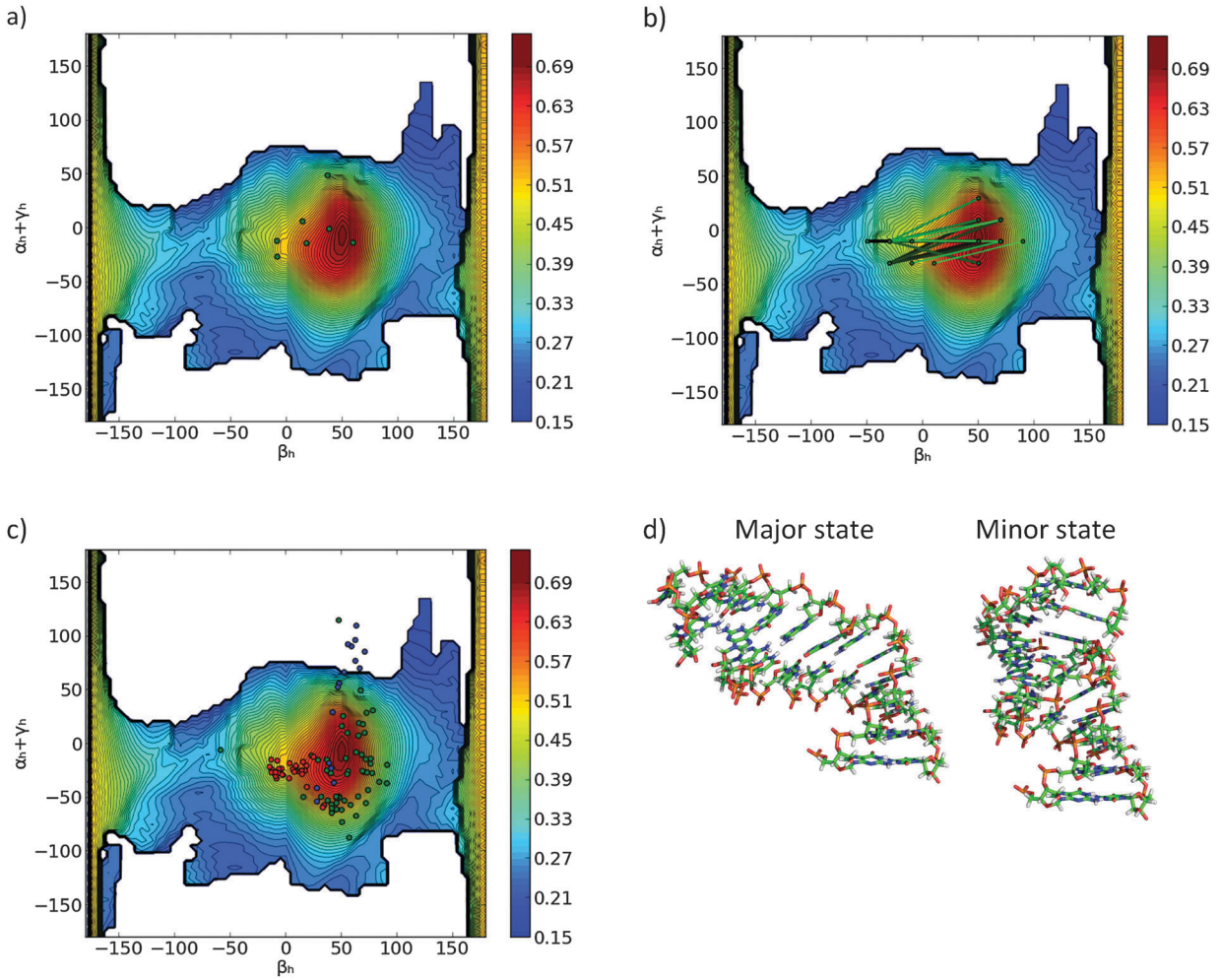


Fig. 6 Superposition of the MaxOcc landscape (MaxOcc values color coded), as a function of the β_h and $\alpha_h + \gamma_h$ coordinates, and (a) the ligand bound structures available in the PDB (green dots), (b) the set of pairs of 20°·20° regions with MaxOR > 95% (depicted as dots located in the centers of the regions, connected by a line; pairs including the 'ghost' of the minor state are omitted for clarity), (c) the ensemble selected from the MD trajectory by SAS ('RDC reweighted trajectory') divided into three clusters as in the original paper (cluster 1 in green, cluster 2 in red and cluster 3 in blue). White areas correspond to not sampled regions. (d) Representative RNA conformations of the two compact regions, one located at the peak of the MaxOcc profile (major state) and another at $\alpha_h = -15$, $\beta_h = -40$, $\gamma_h = 0°$ (minor state), able to fit the experimental data.

importance of the clusters found by SAS, as cluster 1 was sampled for 66% of time and cluster 2 for 19%. The third cluster, representing 15% of weight in the SAS ensemble and located approximately in between the two others states, does not find its counterpart in the current analysis. A possible explanation can be found from the analysis of the structural details of the conformers composing this third cluster. The latter cluster features the melting of the A22–U40 base pair (the last base pair of the first helix), which allows them to sample inter-helical angles which are sterically disallowed when the helices are modelled as rigid bodies, like in the current MaxOcc analysis. Furthermore, the SAS ensemble actively incorporates experimental data within the bulge, potentially requiring a more complex model of motion to be adequately explained. A glance at Fig. 6c reveals that a significant fraction of the structures from the SAS cluster 3 is indeed located outside of what was considered the sterically allowed space for the MaxOcc analysis, while the remaining part is practically within the ranges of the Euler angles of the other two identified states.

Finally, we note that if only conformations in the first or second half of the MD trajectory were considered, either the conformations in the MaxOcc peak or in the minor state are scarcely sampled, thus suggesting that significantly shorter MDs would not be able to capture the structural variability detected by the RDC data. This is in line with the previous observation that the quality of the RDC fit deteriorates considerably when applying SAS to a shorter 80 ns MD trajectory.[31]

Conclusions

We have applied the MaxOcc and MaxOR approaches to analyse the RDC datasets previously acquired by some of us for the HIV-1 TAR RNA strand. Our analysis shows that all conformations which can provide the highest contributions to the experimental averaged data are clustered into one broad but well-defined peak in the conformational space defined by the three Euler angles

providing the inter-domain orientation of the two RNA strands. Very interestingly many of the known ligand bound structures of HIV-1 TAR RNA turn out to be very similar to the conformers with the highest MaxOcc suggesting that known ligands may actually bind to a HIV-1 TAR conformation that is already highly present in the free RNA ensemble. A comparison of the present analysis with the MD simulation previously performed for this system shows that the MD sampling largely covers the medium-high MaxOcc regions. It is intriguing to observe how two completely different approaches tend to converge to a common result: molecular dynamics is in fact only based on the driving force of a general force field, whereas the MaxOcc results only reflect the regions of the conformational space which mostly comply with the experimental data. Moreover the agreement between the two approaches is significantly improved when the MD trajectory is reweighted based on averaged experimental RDCs, suggesting the validity of the SAS approach used for that purpose.[31]

Finally, another compact region of conformations, apart from the MaxOcc peak, was identified, which is the best suitable to complement the latter in a two centered conformational ensemble. We have also shown that this pair of regions constitutes the simplest conformational ensemble capable of reproducing the experimental RDC values and that they resemble the two principal states determined by selecting conformational ensembles from the MD trajectory.

Acknowledgements

This work was supported by Ente Cassa di Risparmio di Firenze, MIUR PRIN 2012SK7ASN, the European FP7 ITN contract pNMR No. 317127, and Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI). HMA acknowledges support from the US National Institutes of Health (R01AI066975 and PO1GM0066275).

References

1 E. A. Dethoff, J. Chugh, A. M. Mustoe and H. M. Al Hashimi, *Nature*, 2012, **482**, 322–330.
2 A. M. Mustoe, C. L. Brooks and H. M. Al Hashimi, *Annu. Rev. Biochem.*, 2014, **83**, 441–466.
3 J. Rinnenthal, J. Buck, J. Ferner, A. Wacker, B. Furtig and H. Schwalbe, *Acc. Chem. Res.*, 2011, **44**, 1292–1301.
4 L. Salmon, S. Yang and H. M. Al Hashimi, *Annu. Rev. Phys. Chem.*, 2014, **65**, 293–316.
5 M. H. Bailor, X. Sun and H. M. Al Hashimi, *Science*, 2010, **327**, 202–206.
6 M. P. Latham, P. Hanson, D. J. Brown and A. Pardi, *J. Biomol. NMR*, 2008, **40**, 83–94.
7 Q. Zhang, X. Sun, E. D. Watt and H. M. Al Hashimi, *Science*, 2006, **311**, 653–656.
8 A. L. Hansen and H. M. Al-Hashimi, *J. Am. Chem. Soc.*, 2007, **129**, 16072–16082.
9 Y. E. Ryabov and D. Fushman, *Magn. Reson. Chem.*, 2006, **44**, S143–S151.
10 Y. E. Ryabov and D. Fushman, *J. Am. Chem. Soc.*, 2007, **129**, 3315–3327.
11 E. A. Dethoff, A. L. Hansen, Q. Zhang and H. M. Al Hashimi, *J. Magn. Reson.*, 2010, **202**, 117–121.
12 G. Lipari and A. Szabo, *J. Am. Chem. Soc.*, 1982, **104**, 4546–4559.
13 R. Brüschweiler, B. Roux, M. Blackledge, C. Griesinger, M. Karplus and R. R. Ernst, *J. Am. Chem. Soc.*, 1992, **114**, 2289–2302.
14 J. Iwahara and G. M. Clore, *J. Am. Chem. Soc.*, 2010, **132**, 13346–13356.
15 E. Ravera, L. Salmon, M. Fragai, G. Parigi, H. M. Al-Hashimi and C. Luchinat, *Acc. Chem. Res.*, 2014, **47**, 3118–3126.
16 M. Fragai, C. Luchinat, G. Parigi and E. Ravera, *Coord. Chem. Rev.*, 2013, **257**, 2652–2667.
17 Q. Zhang, A. C. Stelzer, C. K. Fisher and H. M. Al-Hashimi, *Nature*, 2007, **450**, 1263–1267.
18 L. Salmon, G. Nodet, V. Ozenne, G. Yin, M. R. Jensen, M. Zweckstetter and M. Blackledge, *J. Am. Chem. Soc.*, 2010, **132**, 8407–8418.
19 P. Guerry, L. Salmon, L. Mollica, J. L. Ortega Roldan, P. Markwick, N. A. van Nuland, J. A. McCammon and M. Blackledge, *Angew. Chem., Int. Ed. Engl.*, 2013, **52**, 3181–3185.
20 A. Cavalli, C. Camilloni and M. Vendruscolo, *J. Chem. Phys.*, 2013, **138**, 094112.
21 K. Lindorff-Larsen, K. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen and M. Vendruscolo, *J. Am. Chem. Soc.*, 2004, **126**, 3291–3299.
22 W. Boomsma, J. Ferkinghoff-Borg and K. Lindorff-Larsen, *PloS Comput. Biol.*, 2014, **10**, e1003406.
23 P. Bernadò, E. Mylonas, M. V. Petoukhov, M. Blackledge and D. I. Svergun, *J. Am. Chem. Soc.*, 2007, **129**, 5656–5664.
24 F. Musiani, G. Rossetti, L. Capece, T. M. Gerger, C. Micheletti, G. Varani and P. Carloni, *J. Am. Chem. Soc.*, 2014, **136**, 15631–15637.
25 S. Jager, P. Cimermancic, N. Gulbahce, J. R. Johnson, K. E. McGovern, S. C. Clarke, M. Shales, G. Mercenne, L. Pache, K. Li, H. Hernandez, G. M. Jang, S. L. Roth, E. Akiva, J. Marlett, M. Stephens, I. D'Orso, J. Fernandes, M. Fahey, C. Mahon, A. J. O'Donoghue, A. Todorovic, J. H. Morris, D. A. Maltby, T. Alber, G. Cagney, F. D. Bushman, J. A. Young, S. K. Chanda, W. I. Sundquist, T. Kortemme, R. D. Hernandez, C. S. Craik, A. Burlingame, A. Sali, A. D. Frankel and N. J. Krogan, *Nature*, 2012, **481**, 365–370.
26 A. T. Frank, A. C. Stelzer, H. M. Al-Hashimi and I. Andricioaei, *Nucleic Acids Res.*, 2009, **37**, 3670–3679.
27 M. Zweckstetter and A. Bax, *J. Am. Chem. Soc.*, 2000, **122**, 3791–3792.
28 M. Zweckstetter and A. Bax, *J. Biomol. NMR*, 2001, **20**, 365–377.
29 M. Zweckstetter, *Nat. Protoc.*, 2008, **3**, 679–690.
30 K. Berlin, D. P. O'Leary and D. Fushman, *J. Magn. Reson.*, 2009, **201**, 25–33.
31 L. Salmon, G. Bascom, I. Andricioaei and H. M. Al Hashimi, *J. Am. Chem. Soc.*, 2013, **135**, 5457–5466.
32 E. J. Denning and A. D. Mackerell, Jr., *J. Am. Chem. Soc.*, 2011, **133**, 5770–5772.

33 N. Foloppe and A. D. Mackerell, Jr., *J. Comput. Chem.*, 2000, **21**, 86–104.
34 A. D. Mackerell, Jr., N. Banavali and N. Foloppe, *Biopolymers*, 2000, **56**, 257–265.
35 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.
36 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.
37 R. J. Gardner, M. Longinetti and L. Sgheri, *Inv. Probl.*, 2005, **21**, 879–898.
38 M. Longinetti, C. Luchinat, G. Parigi and L. Sgheri, *Inv. Probl.*, 2006, **22**, 1485–1502.
39 I. Bertini, Y. K. Gupta, C. Luchinat, G. Parigi, M. Peana, L. Sgheri and J. Yuan, *J. Am. Chem. Soc.*, 2007, **129**, 12786–12794.
40 W. Andralojc, C. Luchinat, G. Parigi and E. Ravera, *J. Phys. Chem. B*, 2014, **118**, 10576–10587.
41 W. Andralojc, K. Berlin, D. Fushman, C. Luchinat, G. Parigi, E. Ravera and L. Sgheri, *J. Biomol. NMR*, 2015, **62**, 353–371.
42 H. M. Al-Hashimi, Y. Gosser, A. Gorin, W. Hu, A. Majumdar and D. J. Patel, *J. Mol. Biol.*, 2012, **315**, 95–102.
43 M. H. Bailor, A. M. Mustoe, C. L. Brooks, III and H. M. Al Hashimi, *Nat. Protoc.*, 2011, **6**, 1536–1545.
44 M. Zweckstetter, G. Hummer and A. Bax, *Biophys. J.*, 2004, **86**, 3444–3460.
45 G. Cornilescu, J. Marquardt, M. Ottiger and A. Bax, *J. Am. Chem. Soc.*, 1998, **120**, 6836–6837.
46 V. K. Potluru, Frugal Coordinate Descent for Large-Scale NNLS., 2012, Ref Type: Conference Proceeding.
47 Y. Nesterov, *SIAM J. Control*, 2012, **22**, 341–362.
48 A. M. Mustoe, H. M. Al Hashimi and C. L. Brooks, III, *J. Phys. Chem. B*, 2014, **118**, 2615–2627.
49 H. M. Al-Hashimi, H. Valafar, M. Terrell, E. R. Zartler, M. K. Eidsness and J. H. Prestegard, *J. Magn. Reson.*, 2000, **143**, 402–406.
50 I. Bertini, M. Longinetti, C. Luchinat, G. Parigi and L. Sgheri, *J. Biomol. NMR*, 2002, **22**, 123–136.
51 M. Longinetti, G. Parigi and L. Sgheri, *J. Phys. A: Math. Gen.*, 2002, **35**, 8153–8169.
52 J. R. Tolman, H. M. Al-Hashimi, L. E. Kay and J. H. Prestegard, *J. Am. Chem. Soc.*, 2001, **123**, 1416–1424.

Electronic Supplementary Information

Figure S1 Ligand bound structures of HIV1TAR deposited in the PDB (green dots) superimposed to the MaxOcc landscape (color coded) as function of α_h , β_h and γ_h angles (2D projections). White areas correspond to not sampled regions.

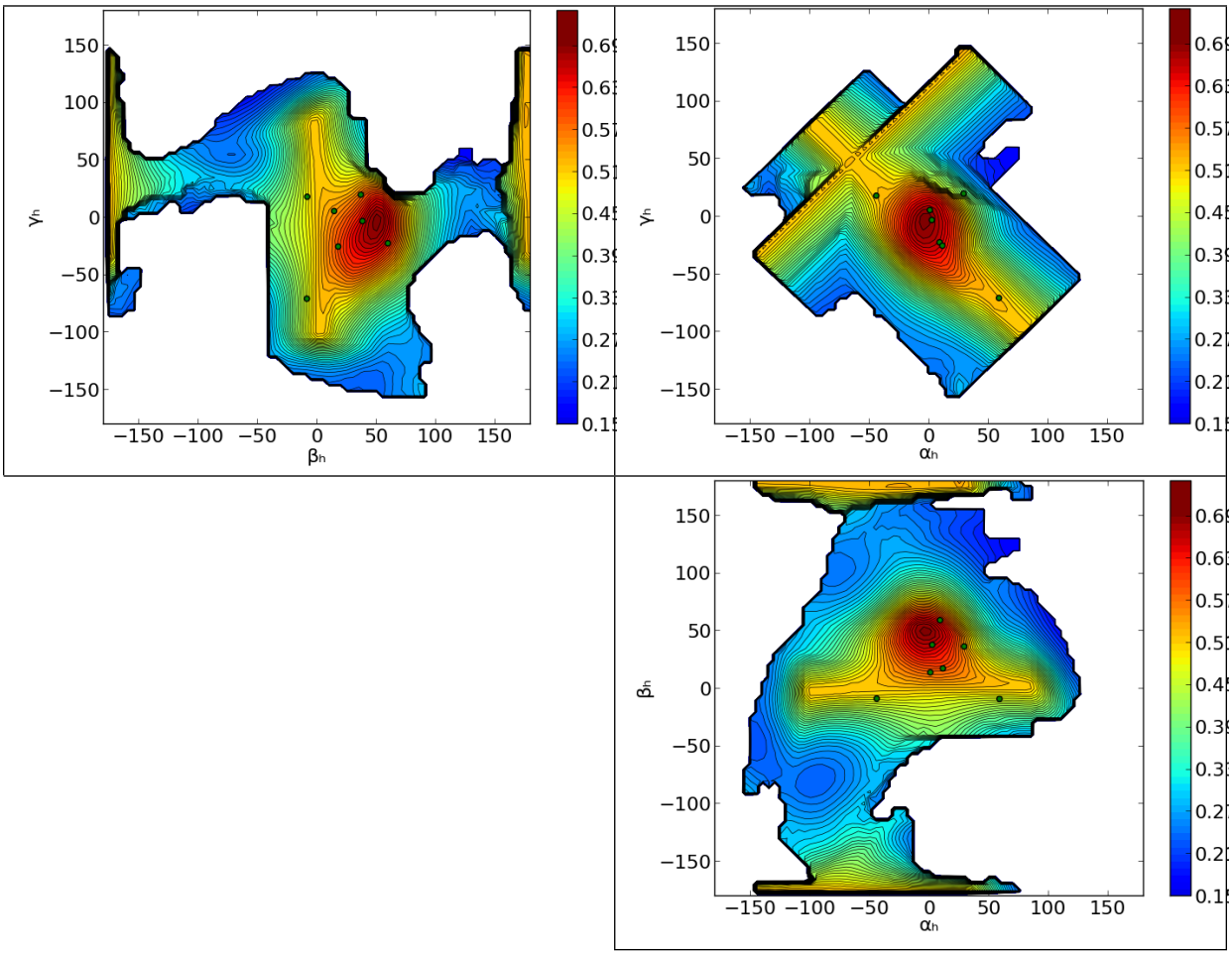


Figure S2. In order to identify two-centered compact regions which can explain the experimental data, two subregions are selected in the conformational space. The subregion composing the peak of the MaxOcc profile ($-10<\alpha_h<5°$, $45<\beta_h<55°$, $-15<\gamma_h<5°$) is complemented by another subregion of size of 5°·5° in the (β_h ,$\alpha_h+\gamma_h$) 2D projection of the conformational space. This second subregion was changed in the different calculations in a systematic way in order to cover the whole (β_h ,$\alpha_h+\gamma_h$) space. The figure shows the MaxOR of the two-centered regions as a function of the (β_h , $\alpha_h+\gamma_h$) position of the second subregion. The red rectangle indicates the position of the first subregion, at the peak of the MaxOcc profile, used together with the second subregion.

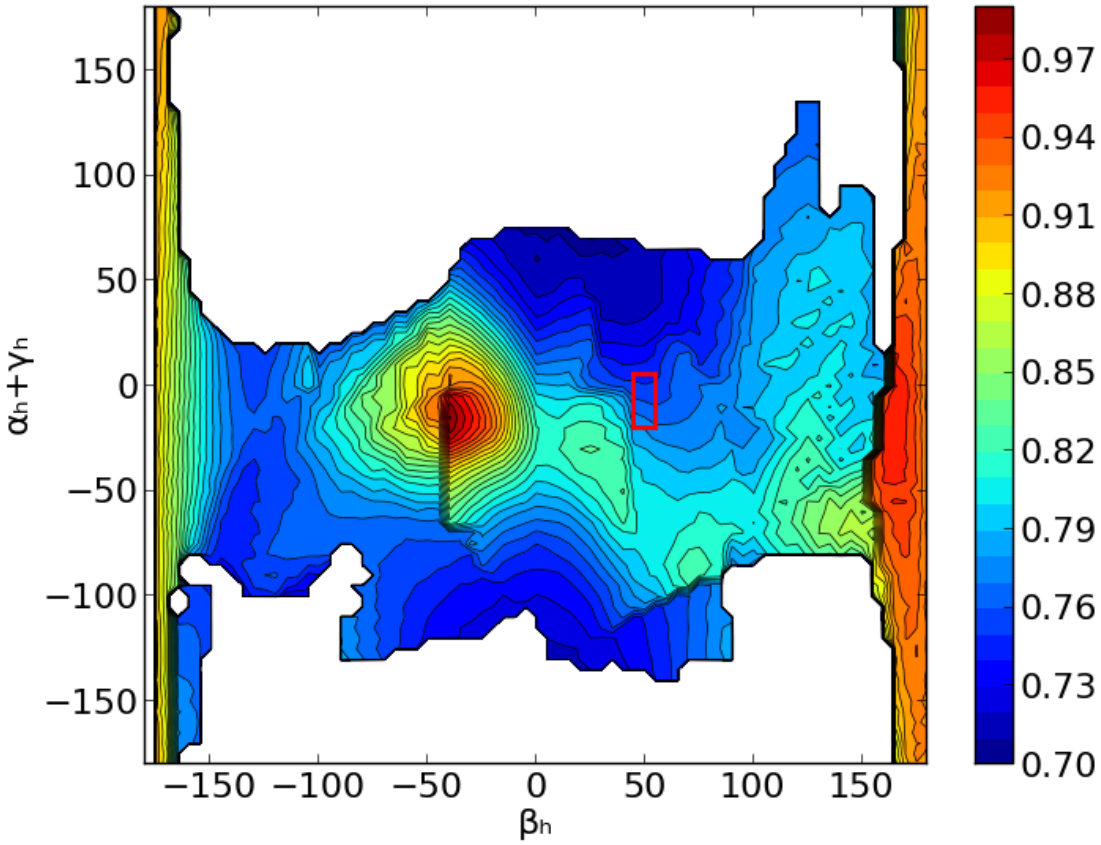
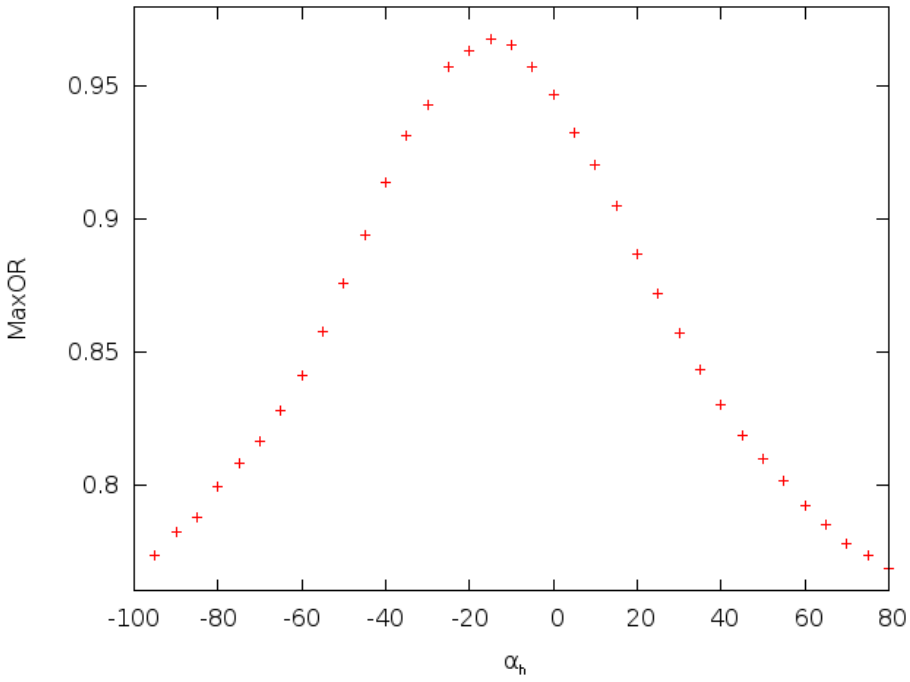


Figure S3. MaxOR of the two-centered ensembles as a function of the α_h coordinate of the region complementing that at the peak of the MaxOcc profile (see text) when the other coordinates are fixed to β_h =-40 and γ_h =-α_h -15°.



This page intentionally left blank

3.4 Identification of productive and futile encounters in an electron transfer protein complex

Witold Andrałojć[a,e] , Yoshitaka Hiruma[b] , Wei-Min Liu[b] , Enrico Ravera[a] , Masaki Nojiri[c,d] , Giacomo Parigi[a,e] , Claudio Luchinat[a,e,1], Marcellus Ubbink[b,1]

[a]Magnetic Resonance Center (CERM), University of Florence, and Interuniversity Consortium for Magnetic Resonance of Metalloproteins (CIRMMP), Sesto Fiorentino, Italy;

[b]Leiden Institute of Chemistry, Leiden University, Leiden, The Netherlands;

[c]Department of Chemistry, Osaka University, Japan;

[d]RIKEN SPring-8 Center, Japan;

[e]Department of Chemistry, University of Florence, Sesto Fiorentino, Italy

submitted

This page intentionally left blank

Identification of productive and futile encounters in an electron transfer protein complex

Witold Andrałojć[a,e] , Yoshitaka Hiruma[b] , Wei-Min Liu[b] , Enrico Ravera[a] , Masaki Nojiri[c,d] , Giacomo Parigi[a,e] , Claudio Luchinat[a,e,1], Marcellus Ubbink[b,1]

[a]*Magnetic Resonance Center (CERM), University of Florence, and Interuniversity Consortium for Magnetic Resonance of Metalloproteins (CIRMMP), Sesto Fiorentino, Italy;* [b]*Leiden Institute of Chemistry, Leiden University, Leiden, The Netherlands;* [c]*Department of Chemistry, Osaka University, Japan;* [d]*RIKEN SPring-8 Center, Japan;* [e]*Department of Chemistry, University of Florence, Sesto Fiorentino, Italy*

[1]To whom correspondence may be addressed. Email: luchinat@cerm.unifi.it or m.ubbink@chem.leidenuniv.nl

Abstract.

Well-defined, stereospecific states in protein complexes are often in exchange with an ensemble of more dynamic orientations, the encounter states. The structure of the stereospecific complex between cytochrome P450cam and putidaredoxin was solved recently by X-ray diffraction as well as paramagnetic NMR spectroscopy. Besides the stereospecific complex, the NMR data clearly demonstrate the presence of additional states in the complex in solution. In these encounter states, populated for a few percent of the time, putidaredoxin assumes multiple orientations and samples a large part of the surface of cytochrome P450cam. To characterize the nature of the encounter states, an extensive paramagnetic NMR dataset has been analyzed using the Maximum Occurrence of Regions methodology. The analysis reveals the location and maximal spatial extent of the additional states needed to fully explain the NMR data. Under the assumption of sparsity of the size of the conformational ensemble, several minor states can be located quite precisely. The distribution of these minor states correlates with the electrostatic potential map around cytochrome P450cam. Whereas some minor states are on isolated positively charged patches, others are connected to the stereospecific site via positively charged paths. The existence of electrostatically favorable pathways between the stereospecific interaction site and the different minor states, or lack thereof, suggests a means to discriminate between productive and futile encounter states.

Keywords: paramagnetic NMR, encounter complex, cytochrome P450cam, putidaredoxin, maximum occurrence, PREs, minor states

Significance Statement.

Paramagnetic NMR spectroscopy is exquisitely sensitive for sparsely populated states in protein-protein interactions, and thus it can provide important information on how protein-protein complexes form and evolve toward their productive state. However, the description of ensembles of protein-protein orientations is non-trivial, and great care must be taken when attempting to derive biologically-relevant results. We have applied an algorithm that restricts the conformational space sampled by the two partners to the maximum allowed for by the data. These ensembles can then be reduced assuming the principle of scarcity. We found that some of these states are linked to the main state through electrostatic pathways. Such paths help to identify those minor states that are able to evolve into the productive complex.

Introduction

Crystal structures suggest that proteins assume unique, stereospecific orientations within protein-protein complexes. However, a number of studies in solution have made clear that encounter states are an inherent element of protein complexes (1–8), especially in electron transfer where the interactions are often extremely fast (9). In the encounter complex, the proteins assume multiple other orientations, often in equilibrium with the major, stereospecific state. In low-affinity complexes, with dissociation constant (K_d) values > 10 μM, the encounter complex can represent a sizeable fraction, and in some cases a well-defined, stereospecific complex may even be absent (10–15). The presence of encounter states may be a consequence of the chemical nature of proteins. In non-obligate stereospecific complexes, the interface represents a small fraction of the total protein surface, so it is reasonable to assume that weak interactions occur also elsewhere. In the case in which protein pairs have evolved to exhibit a high association rate by using electrostatic pre-orientation, electrostatic patches appear to enhance the presence of encounter states (16, 17). On the one hand the pre-orientation reduces the surface area that is visited by the partner, thus enhancing the number of productive encounters, but on the other hand, highly charged patches can bind the oppositely charged protein in many orientations with about equal energy, a situation that has been compared with Velcro binding (18).

Cytochrome P450cam (cytP450cam) from *Pseudomonas putida* is the most comprehensively studied member of the ubiquitous superfamily of cytochromes P450 – heme-containing monooxygenases involved in a plethora of chemical reactions, including drug metabolism, oxidation of xenobiotics and the synthesis of steroids, that receive electrons from electron transfer proteins (19). CytP450cam catalyzes the stereo- and regio-specific hydroxylation of camphor to 5-exo-hydroxycamphor. The reaction involves two one-electron reduction steps, with electrons transferred from NADH via putidaredoxin reductase (PdxR) and putidaredoxin (Pdx). Studies *in vitro* have elucidated an intriguing difference between these two electron transfer steps by demonstrating that the first can be achieved by a variety of reducing agents, while the second strictly requires the presence of Pdx (20–22). This behavior has raised much interest in the molecular details of the cytP450cam-Pdx interaction over the last three decades. Yet, the crystal and solution structures of the electron transfer (ET) complex formed by the two molecules were solved only very recently by X-ray crystallography and paramagnetic NMR, respectively (23, 24). The relative orientation of the proteins within the complex in the solution and crystalline states are practically identical, consistent with the information available from the mutagenesis studies of the system and favorable for effective ET (21, 25–29). It is still under debate whether Pdx induces opening of substrate-bound cytP450cam. The crystal structures showed that in the complex cytP450cam was present in the open state, with the F and G helices moved aside to allow substrate access to the active site. However, a recent solution study using paramagnetic NMR provided evidence that at ambient temperature in solution cyt450cam remains closed upon binding to Pdx (30). Low-temperature EPR studies suggested that Pdx binding leads to a mixture of open and closed states for the oxidized cytP450cam, but not for the reduced, CO-bound state (31, 32). Modelling work supports a view that Pdx binding does affect the active site, enabling the catalytic reaction, perhaps leading to partial opening of the substrate access channel (33).

The stereospecific complex that was found in solution did not explain all the paramagnetic NMR data measured for the system, as was already pointed out by Hiruma *et al.* (24). One of the

paramagnetic probes introduced into cytP450cam induces much stronger paramagnetic relaxation enhancements (PREs) for multiple amide protons in Pdx than predicted by the stereospecific structure, suggesting that the latter does not represent a complete picture of the interaction of cytP450cam with Pdx in solution. The PREs unaccounted for by both the solution and crystal structures (called, from now on, the main state or the stereospecific complex) thus report on the presence of either a minor binding site or an encounter complex. At the same time, most NMR observables - pseudocontact shifts (PCSs), residual dipolar couplings (RDCs) and a significant part of PREs - are very well reproduced by a single structure. This strongly suggests that the population of the additional state(s) is small, but this does not necessarily imply the absence of a functional role. The goal of the present work was to shed more light on the presence of these minor state(s) by making use of a significantly enlarged paramagnetic NMR dataset obtained by introducing additional paramagnetic probes on the surfaces of the two proteins and of recently developed methods to analyze multiple conformational states (34, 35) by exploiting paramagnetism derived restraints.

When a paramagnetic tag is attached to one of the interacting proteins, the PCSs, RDCs and PREs measured for the partner protein report on the reciprocal position and orientation of the two macromolecules. If some mobility is present between the two proteins, then the measured PCSs, RDCs and, under specific assumptions, also PREs are simple population-weighted averages of the values that would have been measured for the individual sampled states, provided the states interconvert rapidly on all the different timescales determined by the PCS, RDC and PRE. (36, 37) (fast exchange regime). We have treated this as a fast exchange system because binding and dissociation of Pdx and cytP450cam are fast on the chemical shift timescale, and internal mobility within the complex is likely to be much faster. We have also assumed the correlation time for the PREs to be the reorientation correlation time of the whole complex, which represents a safe choice because it leads to overestimation of the PRE effect that, in turn, reflects into an overestimation of the expected size of the regions discussed in the section “Identifying the minor states by minOR calculations” (see SI). The exquisite sensitivity of PREs to the presence of even very sparsely populated states - as long as they are located close to the paramagnetic probe - is well recognized and is a direct consequence of the particularly strong distance dependence of the PRE effect ($\sim r^{-6}$). Conformers located very close to the paramagnetic probe, in fact, can account for PREs orders of magnitude higher than when they are located further away, and thus even a marginally small population of such states can easily dominate the measured (ensemble averaged) PREs. A similar behavior does not apply to RDCs (which are independent of the distance between the nuclei and the paramagnetic tag) and is much weaker for the PCSs (which depend on r^{-3}). Therefore, PREs can reveal the existence of very small populations of conformers that may remain hidden when analyzing PCSs and RDCs only.

Even though the combined use of PCSs, RDCs and PREs yields a wealth of highly complementary pieces of information on the reciprocal position and orientation of the two interacting proteins, in the presence of mobility the problem of recovering a conformational ensemble from averaged data remains highly underdetermined because an infinite number of different ensembles can reproduce the measured data equally well. This important issue has been widely studied, and a variety of methodologies have been developed by different groups (38–41) to extract information (i.e. plausible conformational ensembles) from such conformationally averaged data (see review (42)). One of such

methods consists in assigning to each conformer a value, called Maximum Occurrence (MaxOcc), which is defined as the highest fraction of time for which this conformer can exist *in any optimized conformational ensemble* without causing violations of the experimental averaged data (34, 37, 43–47). It thus provides an upper bound for the population of each conformer in the real conformational ensemble of the system.

The concept of Maximum Occurrence was subsequently generalized from single structures to groups of conformers (or regions in the conformational space) giving rise to the MaxOR (Maximum Occurrence of Regions) and minOR (minimum Occurrence of Regions) analysis (35). The minOR calculations permit to identify conformational regions that *must* necessarily be sampled by the system in order to reproduce the experimental data. In fact, the minOR is the *smallest* population that the considered region must have in any conformational ensemble capable of reproducing the averaged data. In this work the motionally averaged PREs are analyzed, in conjunction with PCSs, RDCs and the available protein structures, to extract the information that they contain on the sampling of minor states in the complex between cytP450cam and Pdx. So far the MaxOR/minOR methods have been applied to characterize relative motions between protein domains linked by flexible parts of the polypeptide chain and a helix-bulge-helix RNA element (48). It is the first time that this approach is used to characterize encounter states. In an encounter state the proteins can in theory assume any relative orientation, making the number of possibilities much larger than in the case of domain motions. The strategy is based on the identification of the conformational regions, distinct from the stereospecific complex, that must be sampled to fulfill the data (i.e., with minOR > 0). Our analysis shows that over 10 such regions exist, in several cases spread over parts of the conformational space that are very distant from the main interaction site. We also demonstrate that, if sparsity assumptions about the nature of the conformational sampling are made, the minor states become much better defined, and plausible models for the internal motion in the complex can be proposed. The locations of the identified minor states correlate with the surface charge distributions of the two proteins, suggesting that they represent an electrostatically driven encounter complex that may have mechanistic implications.

Results and discussion

Assessing the contribution of the stereospecific interaction site

The dataset used in the current study is greatly enlarged with respect to the original paper (24). The double armed paramagnetic lanthanoid probe CLaNP-7 (49) was introduced in two new positions, 272/276 (probe 2) and 195/199 (probe 5) on the surface of cytP450cam (see Figure 1), yielding two additional sets of PREs and one additional set of PCSs and RDCs (as no intermolecular PCSs were observed for probe 5). Another set of PREs was also recorded by labeling the Pdx with a MTSL spin label on C73 (probe 6). The data recorded previously consisted of PCSs, RDCs and PREs obtained by placing the CLaNP-7 probe in positions 126/130 (probe 1) and 333/337 (probe 3) of cytP450cam, as well as position 6/12 of Pdx (probe 4). A summary of the experimental restraints is given in Table S1 and their values are listed in Tables S6, S7 and S8.

A global analysis of the data was first performed to determine the compatibility with the previously identified main state (PDB: 2M56) (24). The goal is threefold: i) to check whether it remains the most

favored conformer when the new data are included in the analysis, ii) to identify the experimental PREs that cannot be accounted for by the main state (even allowing for some local mobility), and iii) to determine the upper and lower limits of its population. The compatibility of all available restraints with the proposed main state for the complex was monitored by calculating the MaxOcc of each structure in a densely sampled, broad pool of sterically allowed conformations of the complex, possibly relevant for the fit of the experimental data (see *Materials and methods* and SI). The largest MaxOcc value calculated for a single conformer is 99.4%, and the corresponding structure has the center of mass of Pdx translated by only 1.3 Å and rotated by 6° with respect to the position of the best scoring conformer of the previously published NMR family (PDB: 2M56), as calculated by superimposing the cytP450cam structures. This conformation is practically within the breadth of the NMR family itself. This result demonstrates that the additional paramagnetic NMR data confirm the structure of the stereospecific complex (2M56) determined previously as the 'main state'. It also indicates that this state can represent the structure of the complex with a weight up to more than 99% without causing any violations of the experimental data. At the same time, however, several tens of PREs are strongly misfitted - some of them by more than 100 s^{-1} - if the weight of the main state is increased to 100% (Q_{PRE} = 0.69), as depicted in Figures 1 and 2. These large violations thus leave no doubt that additional conformations are sampled by the system. It should be recalled that the real weight of the main state can be actually smaller than 99.4%, because MaxOcc represents the upper limit for the occurrence of a conformation (see below).

To check whether the PREs in disagreement with the main state can be explained by a limited, local mobility, the MaxOR of a conformational region comprising all the conformers with Pdx within 5 Å translation and 10° rotation from the main state was calculated. The MaxOR of this region is 99.5%, and it reproduces the PREs only marginally better (Q_{PRE} = 0.60) than the single conformer with largest MaxOcc, demonstrating that local mobility cannot be the main cause of the unexpectedly high PREs. More than 50 PREs (caused by four different paramagnetic probes) are actually significantly underestimated (by more than 30 s^{-1} or, if the PRE itself is smaller than that, by 20 s^{-1}). To obtain a first, crude estimate of how broad the conformational sampling should be, the MaxOR calculations were performed by gradually increasing the size of the region built around the main state. It was found (see Table S3) that the smallest region with MaxOR of 100.0% that can yield a good fit of all data must contain conformers with Pdx translated up to 55 Å and rotated up to 100° from the main state, i.e. must span a significant fraction of the whole conformational space available to the complex. This finding does not imply that this whole region must be populated, but demonstrates that the system must sample conformers located far from the main state.

Identifying the minor states by minOR calculations

A PRE which cannot be accounted for by the main state (see Figures 1 and 2) reports the proximity of a nucleus and a paramagnetic center in a conformation that exists for a fraction of time. The PRE is essentially isotropic and, thus, depends only on the metal-nuclear distance, so the PRE leaves a large degeneracy in the position of the nucleus relative to the center. If conformational regions are generated by translation and rotation of Pdx (in the fixed frame of cytP450cam) with respect to the main state, as was done above, these regions will comprise a large number of structures not fulfilling the distance constraint in order to include all conformations that do. The presence of so many structures makes it

unclear which conformations are really important to fit the data. This is why more informative conformational regions were introduced. Below we describe the steps to filter out the conformers most relevant to describe the experimental data, with the underlying assumptions made at each step.

First, a minimal spherical region around a given probe was defined that still contains conformers that cannot all be left out to explain the restraints produced by that probe (minOR > 0). Given the relation between the probe-nucleus distance in a conformer and the population of that conformer (weight), in principle that region becomes very large if the population of the conformer goes up, because a larger population allows for conformers at a larger distance to experience a given PRE. However, in defining these regions, all restraints were evaluated. The PCSs, RDCs and PREs from other probes restrict the weight that a region as a whole can attain, thus limiting the size of the region. Furthermore, only Pdx molecules located close to the cytP450cam surface were considered to represent physically possible complexes (see SI). Given the large population of the main state, each encounter region is expected to have quite a low population and for this reason, all regions with minOR > 0 were selected.

To define a limited number of regions, we assumed that PREs of structurally close nuclei, induced by the same paramagnetic probe, are likely probing the same minor state(s), and those PREs were clustered for the definition of the regions (see SI for the details on the clustering procedure, and Figure S1 depicting the positions on the protein structures of the nuclei of which the PREs were clustered). In this way, 12 clusters were defined, containing between 1 and 9 observations each. The clusters were labeled using letter codes (A-L), preceded by the number of the probe which caused the specific PREs.

The sizes of the smallest regions with minOR>0, as well as their MaxOR, are listed in Table S4. The table also reports the size that each region would need to have to include the main interaction site. This indicates how distant each cluster is from the main state. These distances confirm that none of the defined regions contain the main interaction site (as expected), although some of them have their borders very close to it. Examples of regions are shown in Figure 3, panels a (region 2.C) and d (region 6.L) and all the identified regions are depicted in Fig. S2. Many regions exhibit some overlap, and in some cases the structures in the overlap areas can fit the PREs of both regions simultaneously, suggesting that in principle all these PREs could report on the same minor state (see the discussion in the SI for details). Table S5 identifies these overlapping regions. This observation is important for the discussion of the number of the detected minor states.

The three regions constructed from the PREs induced by probe 2 on Pdx nuclei share no conformations that explain all PREs simultaneously (see Table S5). This means that at least three clearly distinct orientations of Pdx located next to this probe have to be sampled in the Pdx-cytP450cam encounter complex. These three regions (2.A, 2.B and 2.C), although spanning a similar range of positions of Pdx with respect to cytP450cam (Fig. S2a-c), are in fact, defined by Pdx conformers with different orientations. This is illustrated in Fig 3b, in which the frame of reference is that of Pdx. For probe 3, also on cytP450cam, region 3.E overlaps with both region 3.D and 3.F, which in turn are mutually exclusive, meaning that at least two different groups of structures close to that probe have to be sampled. However, several overlaps exist between regions related to probes 2 and 3, so that the three states required by probe 2 could also simultaneously fit all data related to probe 3. Thus, three distinct minor states are sufficient to fit all the data originating from probes 2 and 3. Of course, this does

not demonstrate that only three states must exist; it only shows that if one makes the *explicit assumption* of looking for a solution with the smallest number of states, then three would be enough.

Analogously, for the other two probes (probe 4 and probe 6, located on the Pdx) no overlaps are present between the six different regions related to the corresponding PREs. As far as the regions induced by the same probe are concerned, this may have been expected because, due to the large size of cytP450cam, the clusters of nuclei defining the different regions are well separated from one another (Fig. 1). The lack of overlap between the regions of probe 4 and probe 6 indicates that the two probes report on distinct orientations of Pdx.

The presence of overlaps between regions related to probes 2 & 3 (labelled A-F) on one side and 4 & 6 (G-L) on the other suggests that some of the data related to probes 4 and 6 can actually be explained by conformations present in the regions defined from probes 2 & 3. In particular, region 4.I overlaps with most of the regions of probes 2 & 3. This is not surprising as the cytP450cam residue central in defining this region for probe 4 (I275) is located very close to the attachment site of probe 2 on the surface of cytP450cam (C272/C276).

Finding the most representative minor states by applying an assumption of sparsity

The regions considered in the previous section contain all the structures that can possibly be used to account for specific subsets of PREs. Their borders, sizes and relations (overlaps) constitute the safe information that can be obtained from the available PREs without making any assumption on the characteristics of the conformational ensemble of the complex, besides the rigidity of the proteins. The cost of this generality is that most regions are very broad and do not offer a detailed description of the system in terms of sampled conformational states.

We are now going to consider whether making some reasonable assumptions about the characteristics of the real conformational ensemble can significantly restrict the number of possible solutions and provide more precise information on the ensemble itself. We apply the Occam's razor principle (an assumption with precedents in NMR conformational analyses (50)), i.e. we aim to describe the experimental data with a minimum set of solutions. Reaching this goal could involve exploiting the overlaps between the regions, discussed in the previous section. Moreover, the assumption of sparsity allows one to limit the size of the regions themselves.

Small sets of conformers needed to fit all the PREs related to a given cluster were determined by splitting the regions into cubes of 5 Å x 5 Å x 5 Å and then fitting the experimental data using the main state complemented by the conformers present in one of the cubes. If the fit of the PREs related to the probe in this cluster is almost as good as that obtained when the whole region is taken (as calculated from the Q factor), then the cube represents a possible sampled state, otherwise it is discarded. Analogous to the determination of the regions discussed above, the combined weight of the conformers within the cube was limited by fitting all restraints for all nuclei, and including the main state. As fitting the PCSs and RDCs requires a large weight for the main state, this is an efficient way to limit the weight of any conformer in an encounter state.

With this approach, several regions (mostly those for probes 4 and 6) were shrunk considerably, and in four cases practically reduced to well-defined states (regions 4.G, 4.I, 6.J and 6.L, see Figures 3 and S2)). Each cube in these 12 shrunk regions has minOR = 0, but still, when any of the cubes, one for

each region, are taken together and complemented with the main state, they are in good agreement with the experimental data.

A model for futile and productive encounter states

At this point, it is possible to analyze the relation between the various regions in functional terms. It is clear that the encounter complex in the Pdx-cytP450cam complex represents a small fraction, but covers a large surface area. Encounter states close to the main state are thought to enhance the association rate by reducing the diffusional search dimensions (51). The farther an encounter conformer is from the main state, the larger the chance that the complex will dissociate before the partners have reached the main state and the association is complete. Such states are called futile encounter complexes (52, 53). If they represent a significant fraction, they reduce the association rate, because they effectively represent a competitive binding site that is inactive. Thus, we wondered whether the encounter complexes described here are isolated or connected to the main state, which could provide evidence that Pdx can easily diffuse to (and from) the main state. In the latter case, the minor states could represent productive rather than futile encounter states, and enhance the association rate. The possible connections were hypothesized by inspecting whether electrostatically favorable pathways for the exchange with the main state exist, as well as whether the PRE data indicate the presence of intermediate states along such pathways. The Pdx surface is mostly negatively charged, as is a significant part of the cytP450cam surface. The most prominent positively charged zone of the cytP450cam surface is located at the main binding site for Pdx.

We first consider the four minor states that are already relatively well-defined. Most of them are located far from the main state, and if they were to represent productive encounter states several connecting intermediate states should exist and be detected by PRE measurements. Out of these well-defined minor states, one (region 6.L) lays on the surface of cytP450cam opposite to the main state (Figure 4). Interestingly, this minor state is positioned directly over a small positively charged patch on cytP450cam. This patch extends toward the attachment site of probe 3 and reaches the main interaction site. The entire pathway is located in the zone of conformational space where the structures necessary to fit the PREs sensed by probe 3 are located (yellow mesh in Figure 4). On the basis of this electrostatic path to the main state and the likely presence of conformers on this path, it can be hypothesized that region L of probe 6 represents a productive encounter location. Two other well-defined minor states (region 4.G and region 6.J) are located very close to each other (but Pdx has clearly distinct orientations) and their exact positions can be selected such that they are placed over positively charged zones of cytP450cam. In this case however, the positive patch does not extend toward the main state and also no obvious intermediate conformations can be selected from the minOR derived regions, suggesting that the observed interactions at this site are futile. The last well-defined state (region 4.I) is located on the same side of cytP450cam as the main state, but close to the attachment site of probe 2. The surface potential map suggests that it can either directly exchange with the main state or do so through intermediates that can be selected from the regions defined using probe 2 (orange mesh). Finally, two other regions (region 4.H and region 6.K) contain conformers which are very close to the main state and thus can directly exchange with it. The positive patch around the main interaction site is actually somewhat broader than the interface of the stereospecific complex and, interestingly, it extends exactly

towards the part of space in which regions 4.H and 6.K are located, and in the opposite direction towards probe 2 and region 4.I, while ending right after the main interaction site in the direction of probe 1 (for which no additional PREs were observed). To complete the discussion of correlations between the positions of the minor states and the electrostatic potential map of cytP450cam, it is worth noting that no minor states were identified in the whole zone of helices F and G (covering the substrate entry channel), which is strongly negatively charged. This finding is not due to lack of paramagnetic probe coverage in this area because probe 5 is attached to helix G, and so it shows that no relevant interactions are occurring in this area.

The surface of cytP450cam features one other zone of positive potential covering a part of its surface exactly opposite to the main interaction site, but none of the identified regions is located next to it. Actually almost no PREs were sensed by the residues in that zone (the highest PRE being 11 s^{-1}), suggesting that also in this part of the protein no interaction occurs. In general, however, the PRE derived regions and the surface charge distribution of cytP450cam coincide surprisingly well.

Conclusions

The very good sensitivity of PREs to minor states, coupled to the application of the Maximum Occurrence methodology to tackle the analysis of motionally averaged data, allowed us to identify a series of lowly populated states that together are populated a few percent at maximum, in the presence of a strongly dominating major conformation of cytP450cam in complex with Pdx. The MaxOR/minOR method alone traces the outer borders of the possible spatial extent of these states as reported by the available PRE data. Due to the large indetermination of the ensemble recovery problem, as well as to the intrinsic insensitivity of PREs to certain type of motions, this result represents all the information that the available experimental data can provide without assumptions on the type of conformational heterogeneity exploited by the system. However, under a single regularizing assumption, the MaxOR/minOR results permit to propose a plausible model for the interactions of the two proteins. The minor states determined within this model point towards an electrostatic nature of the underlying interaction, as the center of mass of the mostly negative Pdx seems to follow the positive patches present on the surface of cytP450cam. With the approach described here it has been possible to formulate hypotheses about which families of conformers represent productive encounters and which ones are futile. These hypotheses can be tested by modification of the surface using site-directed mutagenesis or with computational approaches.

Materials and methods

Chemicals. CLaNP-7 was synthesized and loaded with Lu^{3+}, Gd^{3+} and Tm^{3+} ions as described (49). MTSL ((1-oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl)-methanesulfonothioate) and MTS ((1-acetyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl)-methanesulfonothioate) were purchased from Toronto Research Chemicals, North York, Ontario, Canada.

Mutagenesis. To prepare double cysteine mutants of Pdx and cytP450cam, site-directed mutagenesis was carried out using QuikChange protocol (Stratagene, La Jolla, CA).

Protein production. Pdx and cytP450cam variants were produced as previously described (24).

Paramagnetic probe attachments. Pdx and cytP450cam variants were labeled with Ln^{3+}-CLaNP-7 as previously described (24). For MTS(L) tagging**,** Pdx WT was incubated with 5 mM DTT in 20 mM potassium phosphate, pH 7.0, 50 mM KCl for 30 min on ice. DTT was removed by using a PD-10 column. The protein solution was mixed with five molar equivalents of MTS(L) and incubated for one hour at 4 °C. Oligomers and surplus of spin labels were removed by two steps of chromatography. The MTS(L) incubated sample was filtered and loaded on a HiTrap Q HP anion-exchange column (2 mL). The protein was eluted with a linear gradient of 0.1–0.4 M KCl in 20 mM potassium phosphate, pH 7.4 and the first peak fractions were collected and subsequently loaded on a Superose 12 size exclusion column pre-equilibrated with 50 mM Tris-HCl, pH 7.4, 100 mM KCl and 1 mM camphor, 1% (v/v) MeOH, 6% (v/v) D_2O. The eluted brown fractions were pooled and concentrated. Approximately, 80–90% of protein was lost during the labeling procedure, presumably because MTSL labeling reduces the stability of the iron sulfur cluster.

NMR samples and experiments. NMR samples contained 100–200 μM [^{2}H, ^{15}N] Pdx (WT) or [^{2}H, ^{15}N] cytP450cam (C334A) with two molar equivalents of Ln^{3+}-CLaNP-7 labeled cytP450cam or Ln^{3+}-CLaNP-7/MTS(L) tagged Pdx mutants, respectively, in 50 mM Tris-HCl, pH 7.4, 100 mM KCl, 1 mM camphor and 7% (v/v) D_2O. 2D ^{15}N-^{1}H HSQC and ^{15}N-^{1}H TROSY spectra (54) were recorded at 290 K on a Bruker Avance III 600 MHz spectrometer equipped with a TCI-Z-GRAD cryoprobe.

NMR assignment. All NMR data were processed in NmrPipe (55) and analyzed in CCPNMR (56). The amide resonances of oxidized Pdx and cytP450cam were assigned based on previous works (24, 57).

Data analysis. PCS and RDC measurements were performed as previously described (24). PRE datasets of Gd^{3+}-CLaNP-7 and MTSL were analyzed as reported before (10, 24). Briefly, the peak heights of the amide resonances of the proteins in the presence of partner protein with either a diamagnetic probe (Lu^{3+}-CLaNP-7 or MTS) or paramagnetic probe (Gd^{3+}-CLaNP-7 or MTSL) were represented as I_{dia} and I_{para}, respectively. The I_{para}/I_{dia} ratios were normalized by dividing them by the averaged values of the 10 largest I_{para}/I_{dia} values (0.96, 1.15, 1.09, 0.95, 1.0 and 0.97 for cytP450cam probe position 1, 2, 3, 5, and Pdx probe position 4 and MTSL, respectively). The PREs values ($R_{2,para}$) were calculated as described (58).

MaxOcc and MaxOR/minOR calculations. The MaxOR of a given group of conformers is calculated by finding ensembles optimized to simultaneously (i) fit the experimental data within a certain threshold and (ii) include the highest population of the conformers in question. The appropriate ensembles can be found by solving the linear optimization problem defined by:

$$\underset{\mathbf{x}}{\operatorname{argmin}}\left\{\left\|\mathbf{Ax}-\mathbf{y}\right\|_2^2-\gamma\sum_{i\in C}x_i+\lambda(1-\sum_{i=1}^{N}x_i)^2\right\}\text{s.t.}\,\mathbf{x}\geq 0 \qquad (1)$$

where **x** is the vector of the weights of the N structures composing the considered pool of conformers, **y** is the vector of M experimentally observed values of PCSs, RDCs and PREs, normalized by their respective norms, **A** is the M by N matrix whose columns contain the PCS, RDC and PRE values back-calculated for each of the conformers (see SI), again normalized by the norm of the experimental data. Note that this normalization of the **y** vector and the **A** matrix, makes the term $\|\mathbf{Ax}-\mathbf{y}\|_2^2$ to correspond to the sum of squares of the Q-factors (59) between the experimental and back-calculated data for PCSs, RDCs and PREs. This term reflects the fit of the experimental data. The maximization of the population of the considered region is achieved by the second term in which C denotes the set of the structures within the considered region. The third term constrains that the sum of populations of all the structures to 1.

The second term is a regularization and, as such, it has an adverse effect on the quality of reproduction of the experimental data (it will tend to increase the weight of a certain group of conformers not necessarily well suited to fit the experimental data); thus the weighting factor γ is introduced to control the influence of this term.

During a MaxOR calculation γ is first set to 0 and Eq. 1 is solved in order to obtain the best possible fit of the available experimental data (lowest sum of Q-factors or $\|\mathbf{Ax}-\mathbf{y}\|_2^2$). Every ensemble which can fit the experimental data within a threshold set 30% higher than the minimum value, obtained from Eq. 1 in the best possible fit, is considered to reproduce the experimental data in a satisfactory matter.

After the best achievable value of $\|\mathbf{Ax}-\mathbf{y}\|_2^2$ is established, the calculations are repeated for increasing γ values. By increasing γ, the population of the considered region increases and, at the same time, the agreement with the experimental data decreases. The value of γ at which the quality of the fit matches the 30% threshold is found through a one-dimensional bisection method. The total population of the considered region ($\sum_{i\in C} x_i$) found for this value of γ is thus the highest that can exist without causing a misfit of the experimental data – and as such it is the MaxOR of this region.

The other weighting factor λ depends on the specific experimental dataset and its value is found with the L-curve method, as a compromise between a good fit of the experimental observables and the proximity of the sum of the weights to 1 (in a minimization in which term 2 is not present) (60). A value of 10 was used in the present calculations.

The calculation of minOR proceeds in the same way with the only difference that the second term is now taken with a positive sign. In this way the population of the considered region is forced to be as *small* as possible.

Also for calculating the MaxOcc of a given single conformation, eq. 1 is used, with the set C containing only the single conformer under examination.

Acknowledgements

Fruitful discussions with Prof. Charalampos Kalodimos and Prof. Gottfried Otting are acknowledged.

Funding Information

This research was supported by Ente Cassa di Risparmio di Firenze, MIUR PRIN 2012SK7ASN, European Commission projects BioMedBridges No. 284209, pNMR No. 317127, and Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions. Specifically, we thank the EU ESFRI Instruct Core Centre CERM, Italy. ER holds a FIRC triennial fellowship "Gino Mazzega and Guglielmina Locatello" (17941). M.U. and W.M.L. were supported by a VICI grant from the Netherlands Organisation for Scientific Research, grant nr. 700.58.441 to M.U. The X-ray diffraction studies were performed at the beamline 44XU in the synchrotron radiation facility SPring-8 (Japan) (Proposal Nos. 2013A6841 and 2013B6841), and supported by JSPS KAKENHI Grant Number 23750190 and Grants from Japan Foundation for Applied Enzymology and Takeda Science Foundation.

Author contributions

C.L., G.P. and M.U. designed research; Y.H and W.M.L performed experimental research; M.N. provided the new crystallographic data, W.A and E.R. performed computational research, and W.A., C.L., G.P., E.R. and M.U. wrote the paper.

Figure 1. Paramagnetic lanthanide/MTSL tag positions and distribution of PREs unaccounted for by the main state of the complex. For each paramagnetic probe, the amide protons of the reporter protein with PREs unaccounted for by the main state are indicated by gray spheres - large spheres signify strongly misfitted PREs (used to construct the clusters), and smaller spheres moderately (~20 s^{-1}) misfitted ones. a) Probes placed on cytP450cam (PREs measured on Pdx); b) probes placed on Pdx (PREs measured on cytP450cam).

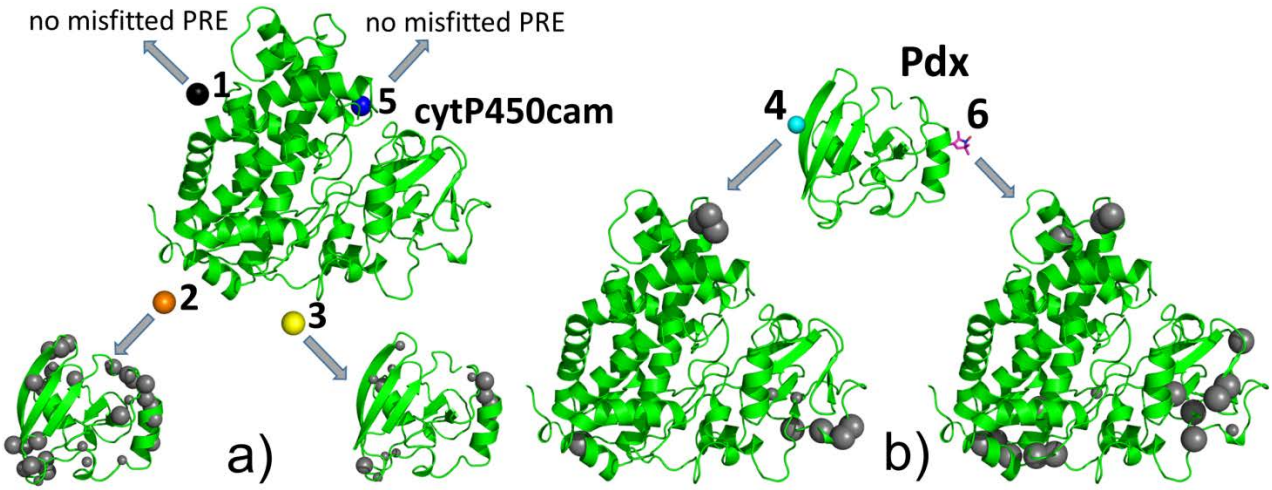


Figure 2. Observed and calculated PREs. The experimental PREs are shown as blue dots, the PRE backcalculated from the main state alone are shown as a red line and those from an ensemble containing 99.4% of the main state complemented with 0.6% of other conformers optimized to fit the experimental data as a green line. The complementing structures were selected from the broad conformational pool (see text) and their specific weights were chosen through the ensemble optimization algorithm of the MaxOR/minOR program. The error bars represent propagated errors based on the noise level in the NMR spectra and the fitted error in the determination of the diamagnetic transverse relaxation rate. The errors for probes 4 and 6 are noticeably higher due to the lower intensity of the cytP450cam detected spectra.

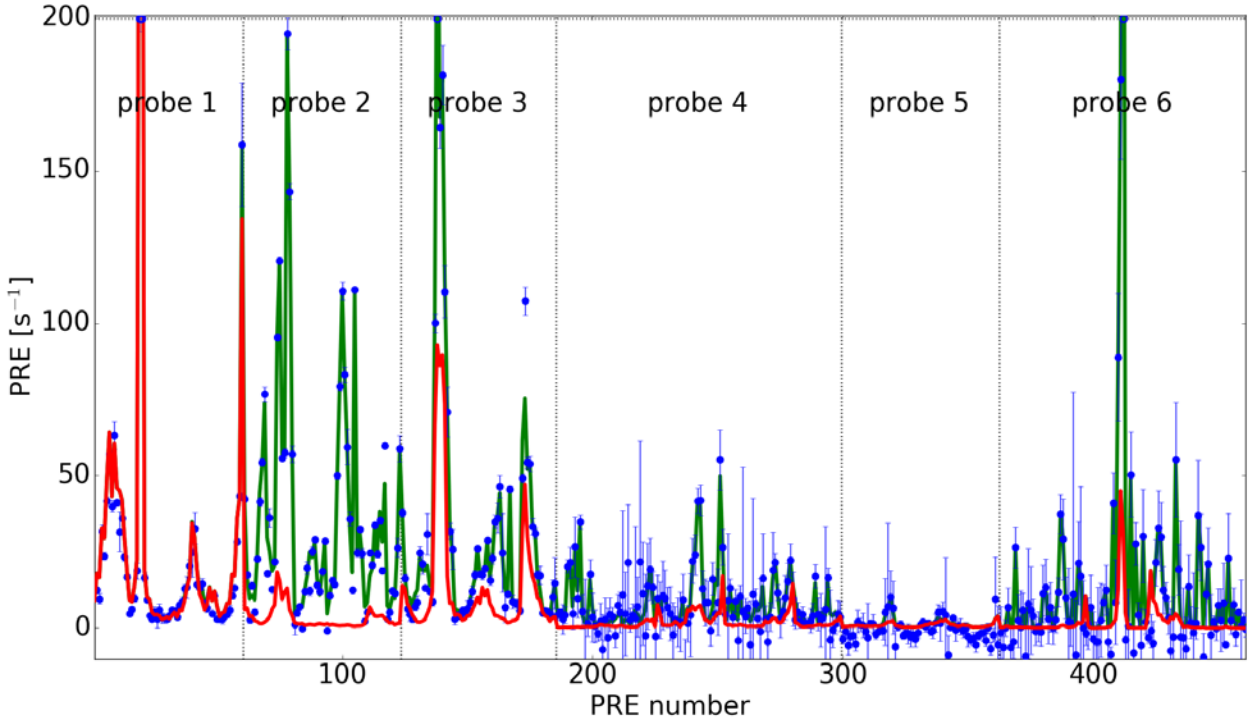


Figure 3. Representative minOR regions and their counterparts shrunk by assuming sparsity. a) The smallest region with minOR > 0 for cluster 2.C; cytP450cam is shown as green cartoon with the positions of probes 1, 2 and 3 indicated as black, orange and yellow axis frames, respectively, the blue dots represent the centers of mass of Pdx in all the conformations belonging to the region and the grey sphere indicates the position of Pdx in the stereospecific complex; b) The three regions with minOR>0 for probe 2, clusters A (red), B (orange) and C (blue), represented in the Pdx frame of reference; Pdx is shown in green cartoon and the blue dots mark the positions of probe 2 (on cytP450cam) in all the conformations composing the regions; c) The subsections of the region shown in panel (a) (cluster 2.C) obtained by applying an assumption of sparsity (see text): each sphere marks the center of a 5 Å by 5 Å by 5 Å cube, the different shades of blue denote the quality of the fit (from dark blue Q = 0 to light blue Q=0.2); d) Region 6.L before (left side) and after (right side) applying the assumption of sparsity, represented as in panels (a) and (c).

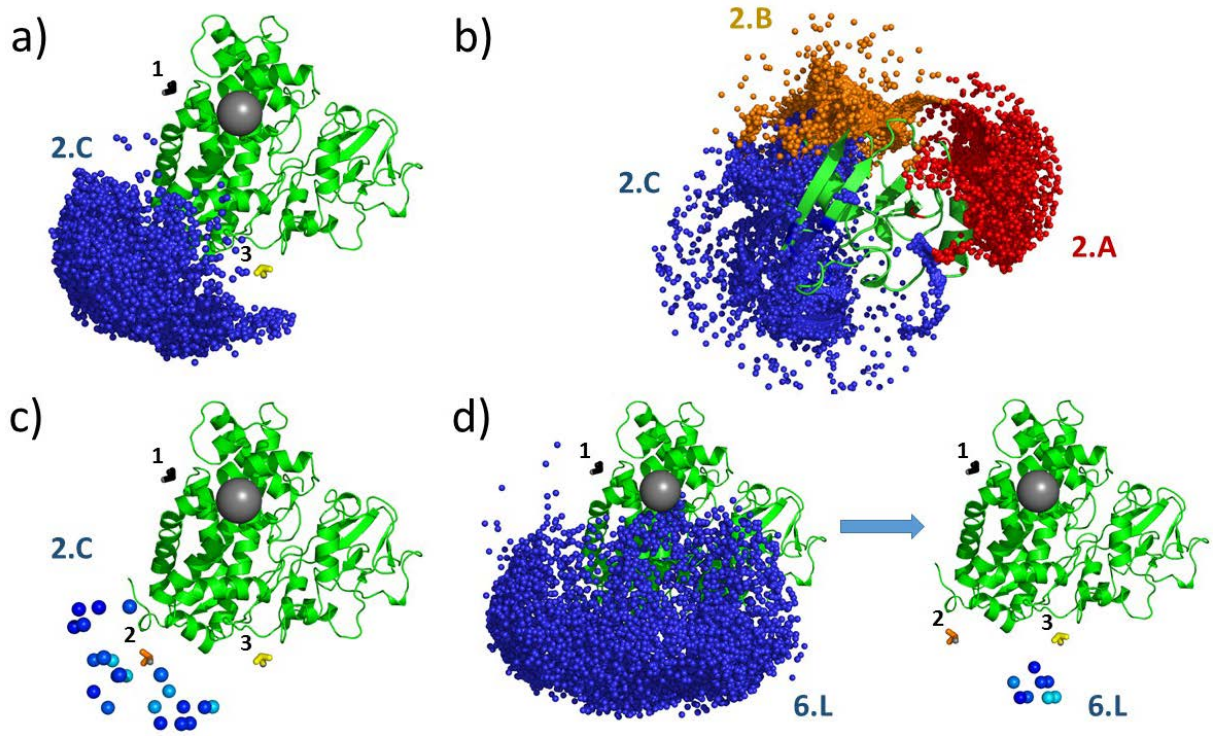
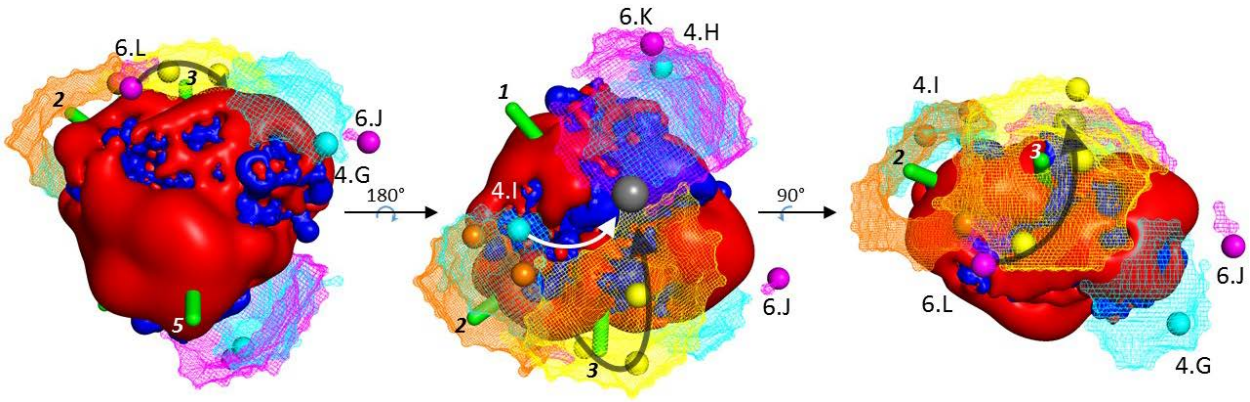


Figure 4. The positive (blue) and negative (red) electrostatic potential isosurfaces (obtained using APBS (61) and traced in PyMol (62) at 0.7 kT/e) of cytP450cam are depicted together with the center of mass positions of Pdx in the encounter regions obtained after applying the scarcity assumption (see text), shown as a mesh with the colors representing the states related to the different probes; probe 2, orange; probe 3, yellow; probe 4, cyan; probe 6, magenta. The colored spheres are representatives of the respective encounter regions placed above positively charged surface areas. The gray sphere represents the main state. The green bars indicate the probe positions on cytP450cam. The curved arrows indicate possible paths between the minor states and the main binding site, suggesting these states could represent productive encounter sites. Regions 4.G and 6.J illustrate isolated states that may represent futile encounter sites.



References

1. Anthis NJ, Clore GM (2015) Visualizing transient dark states by NMR spectroscopy. *Q Rev Biophys* 48(1):35–116.

2. Bashir Q, Scanu S, Ubbink M (2011) Dynamics in electron transfer protein complexes. *FEBS J* 278(9):1391–1400.

3. Water KV de, Nuland NAJ van, Volkov AN (2014) Transient protein encounters characterized by paramagnetic NMR. *Chem Sci* 5(11):4227–4236.

4. Schilder J, Ubbink M (2013) Formation of transient protein complexes. *Curr Opin Struct Biol* 23(6):911–918.

5. Ubbink M (2012) Dynamics in transient complexes of redox proteins. *Biochem Soc Trans* 40(2):415–418.

6. Clore GM, Iwahara J (2009) Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev* 109(9):4108–4139.

7. Clore GM (2008) Visualizing lowly-populated regions of the free energy landscape of macromolecular complexes by paramagnetic relaxation enhancement. *Mol Biosyst* 4(11):1058–1069.

8. Schreiber G, Keating AE (2011) Protein binding specificity versus promiscuity. *Curr Opin Struct Biol* 21(1):50–61.

9. Gray HB, Winkler JR (1996) Electron transfer in proteins. *Annu Rev Biochem* 65(1):537–561.

10. Scanu S, et al. (2012) The complex of cytochrome f and plastocyanin from Nostoc sp. PCC 7119 is highly dynamic. *ChemBioChem* 13(9):1312–1318.

11. Scanu S, Foerster JM, Ullmann GM, Ubbink M (2013) Role of hydrophobic interactions in the encounter complex formation of the plastocyanin and cytochrome f complex revealed by paramagnetic NMR spectroscopy. *J Am Chem Soc* 135(20):7681–7692.

12. Scanu S, Foerster JM, Timmer M, Ullmann GM, Ubbink M (2013) Loss of electrostatic interactions causes increase of dynamics within the plastocyanin–cytochrome f complex. *Biochemistry* 52(38):6615–6626.

13. Xu X, et al. (2008) Dynamics in a pure encounter complex of two proteins studied by solution scattering and paramagnetic NMR spectroscopy. *J Am Chem Soc* 130(20):6395–6403.

14. Liang ZX, et al. (2002) Dynamic docking and electron transfer between Zn-myoglobin and cytochrome b(5). *J Am Chem Soc* 124(24):6849–6859.

15. Worrall JAR, et al. (2002) Myoglobin and cytochrome b5: a Nuclear Magnetic Resonance study of a highly dynamic protein complex. *Biochemistry* 41(39):11721–11730.

16. Schreiber G, Haran G, Zhou H-X (2009) Fundamental aspects of protein−protein association kinetics. *Chem Rev* 109(3):839–860.

17. Ubbink M (2009) The courtship of proteins: Understanding the encounter complex. *FEBS Lett* 583(7):1060–1066.

18. McLendon G (1991) Control of biological electron transport via molecular recognition and binding: The "velcro" model. *Long-Range Electron Transfer in Biology*, Structure and Bonding. (Springer Berlin Heidelberg), pp 159–174.

19. Ahuja S, et al. (2013) A model of the membrane-bound cytochrome b5-cytochrome P450 complex from NMR and mutagenesis data. *J Biol Chem* 288(30):22080–22095.

20. Lipscomb JD, Sligar SG, Namtvedt MJ, Gunsalus IC (1976) Autooxidation and hydroxylation reactions of oxygenated cytochrome P-450cam. *J Biol Chem* 251(4):1116–1124.

21. Tosha T, Yoshioka S, Ishimori K, Morishima I (2004) L358P mutation on cytochrome P450cam simulates structural changes upon putidaredoxin binding. The structural changes trigger electron transfer to oxy-P450cam from electron donors. *J Biol Chem* 279(41):42836–42843.

22. Pochapsky SS, Pochapsky TC, Wei JW (2003) A model for effector activity in a highly specific biological electron transfer complex: the cytochrome P450cam−putidaredoxin couple. *Biochemistry* 42(19):5649–5656.

23. Tripathi S, Li H, Poulos TL (2013) Structural basis for effector control and redox partner recognition in cytochrome P450. *Science* 340(6137):1227–1230.

24. Hiruma Y, et al. (2013) The structure of the cytochrome P450cam–putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography. *J Mol Biol* 425(22):4353–4365.

25. Holden M, Mayhew M, Bunk D, Roitberg A, Vilker V (1997) Probing the interactions of putidaredoxin with redox partners in camphor P450 5-monooxygenase by mutagenesis of surface residues. *J Biol Chem* 272(35):21720–21725.

26. Unno M, Shimada H, Toba Y, Makino R, Ishimura Y (1996) Role of Arg112 of cytochrome P450cam in the electron transfer from reduced putidaredoxin. Analyses with site-directed mutants. *J Biol Chem* 271(30):17869–17874.

27. Hiruma Y, et al. (2014) Hot-spot residues in the cytochrome P450cam–putidaredoxin binding interface. *ChemBioChem* 15(1):80–86.

28. Koga H, et al. (1993) Essential role of the Arg112 residue of cytochrome P450cam for electron transfer from reduced putidaredoxin. *FEBS Lett* 331(1–2):109–113.

29. Yoshioka S, Takahashi S, Ishimori K, Morishima I (2000) Roles of the axial push effect in cytochrome P450cam studied with the site-directed mutagenesis at the heme proximal site. *J Inorg Biochem* 81(3):141–151.

30. Skinner SP, et al. (2015) Delicate conformational balance of the redox enzyme cytochrome P450cam. *Proc Natl Acad Sci* 112(29):9022–9027.

31. Myers WK, Lee Y-T, Britt RD, Goodin DB (2013) The conformation of P450cam in complex with putidaredoxin is dependent on oxidation state. *J Am Chem Soc* 135(32):11732–11735.

32. Liou S-H, Mahomed M, Lee Y-T, Goodin DB (2016) Effector roles of putidaredoxin on cytochrome P450cam conformational states. *J Am Chem Soc* 138(32):10163–10172.

33. Hollingsworth SA, Batabyal D, Nguyen BD, Poulos TL (2016) Conformational selectivity in cytochrome P450 redox partner interactions. *Proc Natl Acad Sci* 113(31):8723–8728.

34. Bertini I, et al. (2010) Conformational space of flexible biological macromolecules from average data. *J Am Chem Soc* 132(38):13553–13558.

35. Andrałojć W, Luchinat C, Parigi G, Ravera E (2014) Exploring regions of conformational space occupied by two-domain proteins. *J Phys Chem B* 118 (36):10576–10587.

36. 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(18):5879–5896.

37. Bertini I, Luchinat C, Nagulapalli M, Parigi G, Ravera E (2012) Paramagnetic relaxation enhancement for the characterization of the conformational heterogeneity in two-domain proteins. *Phys Chem Chem Phys* 14(25):9149–9156.

38. Berlin K, et al. (2013) Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J Am Chem Soc* 135(44):16595–16609.

39. Salmon L, Bascom G, Andricioaei I, Al-Hashimi HM (2013) A general method for constructing atomic-resolution RNA ensembles using NMR residual dipolar couplings: the basis for interhelical motions revealed. *J Am Chem Soc* 135(14):5457–5466.

40. Camilloni C, Vendruscolo M (2015) Using pseudocontact shifts and residual dipolar couplings as exact NMR restraints for the determination of protein structural ensembles. *Biochemistry* 54(51):7470–7476.

41. Chen J-L, et al. (2016) 3D structure determination of an unstable transient enzyme intermediate by paramagnetic NMR spectroscopy. *Angew Chem Int Ed*. doi:10.1002/anie.201606223.

42. Ravera E, Sgheri L, Parigi G, Luchinat C (2016) A critical assessment of methods to recover information from averaged data. *Phys Chem Chem Phys* 18(8):5686–5701.

43. Bertini I, et al. (2007) Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. *J Am Chem Soc* 129(42):12786–12794.

44. Longinetti M, Luchinat C, Parigi G, Sgheri L (2006) Efficient determination of the most favoured orientations of protein domains from paramagnetic NMR data. *Inverse Probl* 22(4):1485–1502.

45. Bertini I, et al. (2012) MaxOcc: a web portal for Maximum Occurence Analysis. *J Biomol NMR* 53(4):271–280.

46. Luchinat C, Nagulapalli M, Parigi G, Sgheri L (2012) Maximum occurrence analysis of protein conformations for different distributions of paramagnetic metal ions within flexible two-domain proteins. *J Magn Reson* 215(1):85–93.

47. Das Gupta S, et al. (2011) Narrowing the conformational space sampled by two-domain proteins with paramagnetic probes in both domains. *J Biomol NMR* 51(3):253–263.

48. Andrałojć W, et al. (2016) Inter-helical conformational preferences of HIV-1 TAR-RNA from maximum occurrence analysis of NMR data and molecular dynamics simulations. *Phys Chem Chem Phys* 18(8):5743–5752.

49. Liu W-M, et al. (2012) A pH-sensitive, colorful, lanthanide-chelating paramagnetic NMR probe. *J Am Chem Soc* 134(41):17306–17313.

50. Clore GM, Schwieters CD (2004) How much backbone motion in ubiquitin Is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? *J Am Chem Soc* 126(9):2923–2938.

51. M. Delbrück, G. Adam (1968) Reduction of dimensionality in biological diffusion processes. *Structural Chemistry and Molecular Biology*, eds A. Rich, N. Davidson (Freeman, San Francisco), pp 198–215.

52. Harel M, Spaar A, Schreiber G (2009) Fruitful and futile encounters along the association reaction between proteins. *Biophys J* 96(10):4237–4248.

53. Fawzi NL, Doucleff M, Suh JY, Clore GM (2010) Mechanistic details of a protein-protein association pathway revealed by paramagnetic relaxation enhancement titration measurements. *Proc Natl Acad Sci* 107(4):1379–1384.

54. Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci* 94(23):12366–12371.

55. Delaglio F, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX Pipes. *J Biomol NMR* 6(3):277–293.

56. Vranken WF, et al. (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline. *Proteins Struct Funct Bioinforma* 59(4):687–696.

57. Lyons TA, Ratnaswamy G, Pochapsky TC (1996) Redox-dependent dynamics of putidaredoxin characterized by amide proton exchange. *Protein Sci* 5(4):627–639.

58. Battiste JL, Wagner G (2000) Utilization of site-directed spin labelling and high-resolution heteronuclear nuclear magnetic resonance for global fold determination of large proteins with limited Nuclear Overhauser Effect data. *Biochemistry* 39(18):5355–5365.

59. Cornilescu G, Marquardt J, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J Am Chem Soc* 120(27):6836–6837.

60. Andrałojć W, et al. (2015) Information content of long-range NMR data for the characterization of conformational heterogeneity. *J Biomol NMR* 62(3):353–371.

61. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc Natl Acad Sci* 98(18):10037–10041.

62. Schrödinger, LLC (2015) *The PyMOL Molecular Graphics System, Version 1.8*.

Supplementary Information to

Identification of productive and futile encounters in an electron transfer protein complex

Witold Andrałojć[a,e] , Yoshitaka Hiruma[b] , Wei-Min Liu[b] , Enrico Ravera[a] , Masaki Nojiri[c,d] , Giacomo Parigi[a,e] , Claudio Luchinat[a,e,1], Marcellus Ubbink[b,1]

[a]Magnetic Resonance Center (CERM), University of Florence, and Interuniversity Consortium for Magnetic Resonance of Metalloproteins (CIRMMP), Sesto Fiorentino, Italy; [b]Leiden Institute of Chemistry, Leiden University, Leiden, The Netherlands; [c]Department of Chemistry, Osaka University, Japan; [d]RIKEN SPring-8 Center, Japan; [e]Department of Chemistry, University of Florence, Sesto Fiorentino, Italy

[1]To whom correspondence may be addressed. Email: luchinat@cerm.unifi.it or m.ubbink@chem.leidenuniv.nl

REFMAC refinement of the proteins

As the fit of several sets of RDCs to any tensor was far from satisfactory, a refinement of a new crystal structure of the complex (1.7 Å resolution, submitted to the PDB under entry 5GXG, details will be published elsewhere) with REFMAC5-NMR (1) was performed before the MaxOcc calculations. The structure of the complex in the new crystals is virtually identical to the one deposited previously as PDB entry 3W9C. The refinement was performed against the (intermolecular) RDC data and the X-ray reflections, with the magnitudes and orientations of each tensor left free to be optimized. This refinement allowed to significantly improve the fitting of all 4 sets of RDC (see Table S2), without worsening the agreement of the structure with the X-ray data or causing significant deviations from the optimal geometric parameters (Table S2). The resulting structures were used in the MaxOcc calculations.

Back-calculation of the paramagnetic NMR observables

To back-predict the PCSs and RDCs induced in the partner protein for any static conformation of the complex, the magnetic susceptibility anisotropy tensors $\Delta\boldsymbol{\chi}$ must be known for each metal position. In the current case, they were obtained experimentally from the fitting of *intramolecular* PCSs (2). With these tensor parameters, the PCSs and RDCs of a given structure can be calculated through the following equations:

$$\Delta\delta^{pcs} = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3\cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta\cos 2\phi\right] \quad \text{eq. (S1)}$$

where $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ are the axial and rhombic components of $\Delta\boldsymbol{\chi}$, r is the electron-nucleus distance and θ and ϕ are the spherical angles describing the orientation of the electron-nucleus vector in the main frame of the $\Delta\boldsymbol{\chi}$ tensor;

$$rdc(Hz) = -\frac{1}{4\pi}\frac{B_0^2}{15kT}\frac{\gamma_A\gamma_B h}{4\pi^2 r_{AB}^3}\times\left[\Delta\chi_{ax}\left(3\cos^2\theta - 1\right) + \frac{3}{2}\Delta\chi_{rh}\left(\sin^2\theta\cos 2\phi\right)\right] \quad \text{eq. (S2)}$$

where B_0 is the magnetic field strength, *k* the Boltzmann constant, *T* the absolute temperature, *h* the Planck constant, γ_A and γ_B are the gyromagnetic ratios of the nuclei involved in the dipolar interaction, r_{AB} is their distance and θ and ϕ are the spherical angles defining the orientation of the A-B vector in the main frame of the $\Delta\boldsymbol{\chi}$ tensor.

The PREs are back-calculated according the following relationship:

$$R_{2\mathrm{PRE}} = \frac{\gamma_A^2 g_e^2 \mu_B^2 \mu_0^2 S(S+1)}{240\pi^2 r^6}\left(4\tau_r + \frac{3\tau_r}{1+\omega_A^2\tau_r^2}\right) \quad \text{eq. (S3)}$$

where g_e is the electron *g* factor, μ_B the Bohr magneton, μ_0 the magnetic permeability of vacuum, *S* the spin quantum number, ω_A the Larmor frequency and τ_r the rotational correlation time of the complex, estimated to be 30 ns using HYDRONMR (3). The correlation time for dipole-dipole relaxation is assumed

to be equal to τ_r, because gadolinium relaxation time is expected to be slower (4, 5). Curie spin relaxation was neglected, as it was estimated to be around 2% of the Solomon term.

Generation of a broad pool of conformers

As no linkers are present between the two components of the complex, Pdx is able to approach the cytP450cam surface from any direction and in any orientation. In order to uniformly sample such a vast conformational space with satisfactory resolution one would have to consider a pool composed of many millions of conformers. Performing the analysis using such a huge number of structures is impossible for the required memory and computational time, and allowing for a much worse resolution would be highly detrimental for the fit of such strongly distance dependent data as the PREs. A way around this problem can be found by noting that due to the underlined strong distance dependence of the PRE, the vast majority of the sterically available conformers will actually sense only negligible relaxation enhancements, and thus they can be disregarded. Also, as the PCS and RDC data are subject to only marginal averaging (as certified by their good fit to a single conformation) there will exist only a limited set of Pdx orientations and positions that can possibly be used to account for these data (in other words all the conformers significantly different from the main state and its 'ghosts' will have a MaxOcc close to 0). This suggests that only a rather limited fraction of all the sterically allowed conformations of the complex have the features needed to be considered even 'possibly useful' to fit the experimental observables. The idea for generating a pool of conformations that on one hand would be computationally tractable and on the other hand would have enough resolution *where it is needed*, thus consists in identifying all the parts of the conformational space which can potentially be useful to reproduce the experimental data and then sample them extensively.

One method to achieve this goal consists in repeatedly fitting the experimental data by docking an ensemble of Pdx copies to cytP450cam, and collecting the structures obtained at each iteration. After each run, all the structures generated so far are taken and for each of them the closest neighbor distance (or better the density of neighbors within a given radius) is monitored. As long as there are numerous conformers for which this distance is high (or the number of neighbors is low) it is clear that part of the conformational space possibly useful for explaining the experimental observables is under-sampled, and thus further iterations of the docking minimization are performed. At the end, a pool with fine sampling of the conformational space around each generated conformer is obtained. Indeed, in our case, only a limited fraction of the conformers has the potential to significantly contribute to the fitting of the experimental observables, and at this stage we expect that there will be still large conformational regions devoid of structures (actually arriving at such a situation would per se strongly suggest the 'uselessness' of the remaining non-sampled regions).

For the current system the ensemble docking step was achieved through a simulated annealing protocol using a FORTRAN90 program developed *ad hoc*. In short, the minimization starts from taking N copies of Pdx (with equal weights summing up to 1) in random positions on the surface of cytP450cam (the different copies of Pdx are not visible to each others and can thus intersect freely) and back-calculating the experimental observables for this initial ensemble. This provides an initial value of the target function, defined as the sum of the Q-factors of the different types of experimental observables. In

addition, the target function contains terms which penalize the approach of a hydrogen atom to the unpaired electron closer than what the steric hindrance of the specific probe allows (6 Å for CLaNP and 2 Å for the MTSL). A structure is then chosen and the 7 variables describing its positions (the 3 Cartesian coordinates of the center of mass and the 3 Euler angles providing the orientation) and weight are randomly changed (by a value proportional to the current temperature). In order to always fix the sum of the weights to 1, the weight of another randomly chosen structure is modified accordingly). If the new position is either clashing with cytP450cam (closest CA-CA distance below 3 Å) or not in contact with it (closest CA-CA distance above 7 Å), it is modified by translation along the line connecting the (new) centers of masses of the two proteins, until a contact without clashes is achieved (this approach allows for accounting the 2D character of the translational search without introducing a high rejection rate at high temperatures, when the displacements are big and often lead to clashes or losses of contact). After each step the contribution to the target function is calculated, and the step is accepted or rejected according to the Metropolis criterion. The temperature schedule and the ensemble size N were carefully optimized to achieve the best fitting of the data without including too many redundant structures at each minimization. The optimal ensemble size N was found to be 80. After each minimization, the non-redundancy of the new structure is monitored by checking whether its removal from the optimized ensemble leads to an increase of the target function by more than 0.1%. From 2 to 19 structures were removed in this way at each iteration. The described simulated annealing minimization protocol was run over 850 times, providing a pool of 59000 conformations.

In order to double check that this reduced "simulated annealing pool" of conformations does not influence significantly the obtained results, some MaxOR/minOR calculations were repeated twice, one time using this pool and another time using an 'extended pool' composed of the former complemented by a sparse geometric one (composed of a few thousands of structures). The geometric pool was obtained by selecting 100 angular positions of the center of mass of Pdx in the frame of cytP450cam, optimized to cover approximately uniformly the whole solid angles (these positions were taken from the source code of the simulation tool SIMPSON, which uses them to obtain a best sampling of crystallite orientations for ssNMR simulations). For each angular position of the center of mass a set of 100 orientations was sampled by taking 20 optimal (α,β) Euler angle pairs to uniformly cover the angular space (once again taken from SIMPSON), and sampling for each of them 5 uniformly spaced values of γ. For each of the 10000 conformations obtained in this way, the optimal distance between the centers of mass of the two proteins was found (using the method described for the simulated annealing pool), and then all conformations with clashes with the tags were removed (leading to a final number of ~7000 conformers). In all the tested cases, practically no change in the MaxOR/minOR was observed upon including the geometric pool in the calculations, reinforcing the claim that the simulated annealing pool indeed contains all the structures which are possibly important for fitting the experimental observables.

Defining the clusters of misfitted PREs

The number of regions to be considered in the fit the experimental PREs can be greatly reduced by clustering the HNs of nearby residues which miss PREs induced by the same probe, and constructing regions based on the distances of the corresponding probe from the centers of such clusters (instead of

from single nuclei). The clustering was performed using the k-medoids method and 12 satisfactorily narrow (~10 Å distance between furthest members in most cases) clusters were obtained. They were composed of the following residues:

probe2, cluster A: residues 69 72 73 74 75 76 78 81 106 (of Pdx)
probe2, cluster B: residues 9 10 11 13 31 33 35 (of Pdx)
probe2, cluster C: residues 2 18 19 20 21 95 97 (of Pdx)
probe3, cluster D: residue 19 (of Pdx)
probe3, cluster E: residue 31 (of Pdx)
probe3, cluster F: residues 73, 74, 78 (of Pdx)
probe4, cluster G: residues 20 21 23 25 (of cytP450cam)
probe4, cluster H: residues 228, 229, 230 (of cytP450cam)
probe4, cluster I: residue 275 (of cytP450cam)
probe6, cluster J: residues 52 54 65 307 308 314 (of cytP450cam)
probe6, cluster K: residues 224 228 230 (of cytP450cam)
probe6, cluster L: residues 261 271 275 339 340 381 (of cytP450cam)

Figure S1 depicts the clustering obtained for probes 2 and 6 (which feature the highest number of misfitted PREs) allowing to appreciate the compactness and uniqueness of the resulting clusters.

The significance of the overlaps between the regions with minOR>0

Many of the obtained regions are quite large and thus one may expect overlaps between several of them. Such overlaps are actually very numerous, and thus many conformers may be able to fit simultaneously the PREs belonging to different clusters.

It should be noted, however, that a region which has a minOR>0 contains not only the conformers that fit the considered subset of PREs very well, but also all the conformers that fit the data rather poorly (or only fit part of them), as they all have to be excluded in order to see an important misfit of the data. To check which overlaps are actually important, for each pair of overlapping regions the full set of the experimental observables was fitted using only the structures constituting their common part, complemented by a small region around the main state to allow for the fitting of PCSs and RDCs. The quality of the fitting of the specific PREs used to define the two regions was quantified using the Q-factors calculated only from these subsets of PREs. Only the pairs of regions with $Q_{selected}<0.2$ were considered to have a meaningful overlap. Table S5 lists all such meaningful overlaps.

Table S1. Summary of experimental restraints

Probe	Probe position	Types of restraints	Number of restraints
Probe 1: CLaNP-7	P450cam K126C/R130C	PCS	54
		RDC	48
		PRE	60
Probe 2: CLaNP-7	P450cam Q272C/Q276C	PCS	61
		RDC	57
		PRE	63
Probe 3: CLaNP-7	P450cam A333C/H337C	PCS	43
		RDC	37
		PRE	62
Probe 4: CLaNP-7	Pdx V6C/R12C	PCS	103
		RDC	94
		PRE	114
Probe 5: CLaNP-7	P450cam E195C/A199C	PRE	63
Probe 6: MTSL	Pdx C73	PRE	99

Table S2. Quality of the RDC fits (in terms of Q-factors), R-factor and deviations from ideal geometry before and after refinement with REFMAC-NMR.

	original structure	refined structure
Q-factor Probe 1 (P450cam K126C/R130C/C334A)	0.262	0.152
Q-factor Probe 2 (P450cam Q272C/Q276C/C334A)	0.693	0.395
Q-factor Probe 3 (P450cam A333C/H337C/C334A)	0.427	0.287
Q-factor Probe 4 (Pdx V6C/R12C/C73S/C85)	0.423	0.130
R-factor	0.1695	0.1699
R free	0.1964	0.1970
RMS of bond length [Å]	0.0088	0.0116
RMS of bond angle [deg]	1.4630	1.9814
RMS of chiral volume	0.0862	0.1063

Table S3. MaxOR values for regions composed of all the structures located within specific translation and rotation thresholds from the lowest energy conformer in PDB entry 2M56. Regions that comprise sufficient conformers to fully explain the experimental data are shaded.

		Maximum translation from the main state [Å]												
		5	10	15	20	25	30	35	40	45	50	55	60	65
Rotation from the main state [deg]	10	0.995	0.995	0.996	0.996	0.996	0.996	0.996	0.996	0.997	0.997	0.997	0.997	0.997
	20	0.995	0.996	0.996	0.996	0.997	0.996	0.996	0.997	0.997	0.997	0.997	0.997	0.998
	30	0.996	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.998	0.998	0.998	0.998
	40	0.995	0.996	0.996	0.997	0.997	0.997	0.998	0.998	0.998	0.998	0.998	0.998	0.998
	50	0.996	0.996	0.996	0.997	0.997	0.997	0.998	0.998	0.998	0.998	0.998	0.999	0.999
	60	0.996	0.996	0.997	0.997	0.997	0.997	0.998	0.998	0.998	0.999	0.999	0.999	0.999
	70	0.996	0.996	0.996	0.997	0.997	0.998	0.998	0.998	0.998	0.999	0.999	0.999	0.999
	80	0.996	0.996	0.997	0.997	0.997	0.998	0.998	0.999	0.999	0.999	0.999	**1.000**	**1.000**
	90	0.995	0.996	0.997	0.997	0.997	0.998	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**
	100	0.995	0.996	0.997	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	110	0.996	0.996	0.997	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	120	0.996	0.996	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	130	0.995	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	140	0.996	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	150	0.996	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
	160	0.996	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**
	170	0.995	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**
	180	0.995	0.996	0.997	0.997	0.998	0.998	0.999	0.999	0.999	**1.000**	**1.000**	**1.000**	**1.000**

Table S4. Size of smallest regions with minOR > 0 around each of the considered clusters of misfitted PREs and their respective MaxOR values. For comparison, the distance between the corresponding probe and the center of the cluster in the major state is also given.

		The radius of the smallest region with minOR > 0 (Å)	The MaxOR of that region (%)	Distance of the probe to the center of cluster in main state (Å)
probe 2	cluster A	16	2.4	54.2
	cluster B	18	3.5	44.7
	cluster C	22	14.6	38.1
probe 3	cluster D	24	99.8	25.1
	cluster E	30	84.3	33.8
	cluster F	26	17.1	44.2
probe 4	cluster G	25	8.5	64.8
	cluster H	32	56.6	38.7
	cluster I	26	8.1	42.3
probe 6	cluster J	20	36.3	23.3
	cluster K	15	37.8	22.8
	cluster L	21	17.2	45.4

Table S5. Overlaps between different regions with minOR > 0 that can be used to fit both subsets of data (1- fitting is possible, 0 – not possible). The smallest region around the main state with minOR > 0 (10 Å translation and 40 degrees rotation) is also included for comparison.

	main state		Probe 2			Probe 3			Probe 4			Probe 6		
			A	B	C	D	E	F	G	H	I	J	K	L
main state		-	0	0	0	0	0	0	0	0	0	0	1	0
Probe 2	cluster A	0	-	0	0	1	1	0	0	0	0	0	0	1
	cluster B	0	0	-	0	1	1	1	0	0	1	0	0	0
	cluster C	0	0	0	-	1	1	1	0	0	1	0	0	0
Probe 3	cluster D	0	1	1	1	-	1	0	1	0	1	0	0	0
	cluster E	0	1	1	1	1	-	1	1	0	1	0	1	1
	cluster F	0	0	1	1	0	1	-	1	0	1	0	0	1
Probe 4	cluster G	0	0	0	0	1	1	1	-	0	0	0	0	0
	cluster H	0	0	0	0	0	0	0	0	-	0	0	0	0
	cluster I	0	0	1	1	1	1	1	0	0	-	0	0	0
Probe 6	cluster J	0	0	0	0	0	0	0	0	0	0	-	0	0
	cluster K	1	0	0	0	0	1	0	0	0	0	0	-	0
	cluster L	0	1	0	0	0	1	1	0	0	0	0	0	-

Table S6. PCS values measured for amide protons of cytochrome P450cam (probe 4) and Pdx (probes 1-3). The PCS were induced by the presence of a Tm^{3+} ion. The experimental uncertainty associated with each measured PCS value was estimated as 0.02 ppm.

Probe number	Residue number	PCS value [ppm]
1	3	0.47
1	4	0.32
1	5	0.43
1	6	0.18
1	7	0.14
1	9	-0.14
1	10	-0.05
1	11	0.04
1	12	0.22
1	14	0.56
1	15	0.58
1	18	0.71
1	19	0.49
1	20	0.40
1	21	0.49
1	22	0.54
1	51	0.00

1	52	0.01
1	53	0.09
1	54	0.04
1	55	0.07
1	56	0.09
1	57	0.06
1	58	0.06
1	59	0.10
1	60	0.10
1	62	0.08
1	63	0.02
1	64	0.02
1	65	-0.02
1	66	-0.03
1	68	-0.04
1	69	-0.10
1	72	-0.19
1	73	-0.27
1	74	-0.31
1	77	-0.31
1	78	-0.24
1	79	-0.14
1	81	-0.04
1	82	-0.08
1	83	-0.03
1	89	0.23
1	90	0.39
1	91	0.31
1	93	0.25
1	94	0.23
1	95	0.27
1	96	0.30
1	97	0.26
1	99	0.21
1	100	0.06
1	101	0.03
1	103	-0.34
2	2	0.24
2	3	0.18
2	4	0.14
2	5	0.16
2	6	0.13
2	7	0.12
2	9	0.11
2	10	0.09
2	11	0.10
2	12	0.09
2	13	0.13
2	14	0.14
2	15	0.19
2	18	0.46

2	19	0.34
2	20	0.31
2	31	0.40
2	33	0.26
2	35	0.16
2	50	0.15
2	51	0.13
2	52	0.12
2	53	0.11
2	54	0.07
2	55	0.07
2	56	0.08
2	57	0.08
2	58	0.07
2	59	0.08
2	60	0.11
2	62	0.10
2	63	0.12
2	64	0.15
2	65	0.12
2	66	0.14
2	68	0.13
2	69	0.13
2	72	0.12
2	73	0.14
2	74	0.12
2	75	0.15
2	76	0.14
2	77	0.10
2	78	0.11
2	79	0.10
2	81	0.08
2	82	0.11
2	83	0.13
2	89	0.22
2	90	0.28
2	91	0.18
2	93	0.10
2	94	0.13
2	95	0.11
2	96	0.12
2	99	0.13
2	100	0.12
2	101	0.13
2	103	0.12
2	105	0.17
2	106	0.18
3	5	-0.32
3	6	-0.27
3	7	-0.22
3	9	-0.16

3	10	-0.17
3	11	-0.16
3	12	-0.20
3	13	-0.19
3	14	-0.30
3	19	-0.81
3	33	-0.27
3	51	-0.26
3	52	-0.27
3	53	-0.28
3	54	-0.21
3	55	-0.22
3	56	-0.26
3	57	-0.26
3	58	-0.27
3	59	-0.34
3	63	-0.31
3	65	-0.28
3	66	-0.25
3	68	-0.24
3	69	-0.18
3	72	-0.16
3	73	-0.14
3	74	-0.15
3	75	-0.15
3	76	-0.17
3	77	-0.13
3	78	-0.13
3	79	-0.18
3	81	-0.20
3	82	-0.22
3	83	-0.28
3	95	-0.42
3	99	-0.30
3	100	-0.26
3	101	-0.24
3	103	-0.16
3	105	-0.15
3	106	-0.14
4	12	0.09
4	14	0.07
4	17	0.06
4	18	0.08
4	20	0.06
4	21	0.06
4	22	0.10
4	23	0.07
4	24	0.09
4	25	0.08
4	26	0.10
4	27	0.09

4	30	0.11
4	34	0.11
4	35	0.10
4	36	0.09
4	37	0.12
4	39	0.14
4	40	0.13
4	44	0.11
4	46	0.10
4	47	0.10
4	48	0.08
4	49	0.06
4	50	0.09
4	52	0.09
4	54	0.11
4	55	0.09
4	56	0.10
4	59	0.10
4	60	0.10
4	61	0.10
4	62	0.10
4	65	0.13
4	67	0.14
4	68	0.18
4	69	0.18
4	70	0.17
4	74	0.22
4	91	0.12
4	109	0.56
4	137	0.15
4	145	0.06
4	147	0.08
4	148	0.09
4	149	0.11
4	150	0.11
4	171	0.13
4	172	0.11
4	189	0.11
4	207	0.13
4	209	0.12
4	216	0.17
4	226	0.27
4	260	0.15
4	261	0.12
4	267	0.11
4	271	0.11
4	273	0.12
4	274	0.15
4	279	0.21
4	301	0.15
4	302	0.10

4	304	0.10
4	306	0.10
4	307	0.11
4	308	0.12
4	309	0.09
4	310	0.09
4	311	0.08
4	312	0.11
4	313	0.10
4	314	0.09
4	315	0.13
4	316	0.11
4	317	0.13
4	329	0.14
4	330	0.12
4	332	0.19
4	333	0.15
4	334	0.10
4	336	0.12
4	338	0.14
4	339	0.19
4	340	0.17
4	342	0.21
4	380	0.07
4	381	0.08
4	384	0.08
4	387	0.08
4	388	0.08
4	389	0.09
4	390	0.10
4	391	0.12
4	398	0.18
4	401	0.11
4	402	0.12
4	409	0.05
4	410	0.06
4	411	0.11
4	412	0.07
4	413	0.05
4	414	0.07

Table S7. RDC values measured for amide N-H pairs in cytochrome P450cam (probe 4) and Pdx (probes 1-3). The RDC were induced by the presence of a Tm^{3+} ion and measured at 600 MHz. The experimental uncertainty associated with each measured RDC value was estimated as 4 Hz.

Probe number	Residue number	RDC value [Hz]
1	3	-14

1	4	-16
1	5	-28
1	6	-12
1	7	-20
1	9	-8
1	10	-14
1	11	7
1	12	-18
1	14	-25
1	15	-18
1	18	16
1	19	1
1	20	-13
1	22	-3
1	51	-12
1	52	-19
1	53	-22
1	54	18
1	55	-1
1	56	34
1	57	5
1	58	-10
1	59	21
1	60	-6
1	62	20
1	63	-7
1	64	-15
1	65	17
1	66	20
1	68	22
1	69	16
1	72	8
1	73	19
1	74	7
1	77	-7
1	78	6
1	79	-13
1	82	-13
1	83	-12
1	89	-9
1	91	1
1	93	-2
1	95	23
1	96	-4
1	99	-18
1	100	-18
1	103	-10
2	2	-29
2	3	1
2	4	-3
2	5	14

2	6	3
2	7	3
2	9	-2
2	10	-11
2	11	4
2	12	-10
2	13	13
2	14	-2
2	15	6
2	20	-13
2	33	6
2	35	3
2	50	-4
2	51	-8
2	52	3
2	53	1
2	54	-4
2	55	-7
2	56	2
2	57	-9
2	58	-15
2	59	1
2	60	14
2	62	-9
2	63	-5
2	64	2
2	65	-6
2	66	5
2	68	-5
2	69	-4
2	72	-11
2	73	15
2	74	-9
2	76	2
2	77	-9
2	78	16
2	79	11
2	81	-6
2	82	18
2	83	-10
2	89	-3
2	90	-15
2	91	-7
2	93	-4
2	94	5
2	95	-1
2	96	2
2	99	4
2	100	2
2	101	-3
2	103	10

2	105	6
2	106	-9
3	6	-3
3	7	4
3	9	-13
3	10	-2
3	11	22
3	12	3
3	13	-11
3	14	-3
3	51	-9
3	52	7
3	53	-13
3	54	-10
3	56	6
3	57	-11
3	58	-1
3	59	-10
3	63	-16
3	66	2
3	68	-11
3	69	-11
3	72	-16
3	73	1
3	74	-4
3	75	-17
3	76	-16
3	77	-8
3	79	-13
3	81	-13
3	82	-13
3	83	-16
3	93	3
3	99	-1
3	100	-6
3	101	-2
3	103	-11
3	105	6
3	106	-1
4	12	15
4	14	4
4	17	-21
4	18	16
4	20	-9
4	21	-9
4	23	-4
4	24	-10
4	25	-4
4	26	-9
4	27	-8
4	30	-5

4	34	27
4	35	9
4	36	-10
4	37	16
4	39	-8
4	40	-5
4	44	-18
4	46	-15
4	47	5
4	48	-6
4	49	-6
4	50	8
4	52	-11
4	55	-9
4	56	-6
4	59	12
4	60	9
4	62	-3
4	68	-5
4	69	-9
4	70	-12
4	74	-11
4	91	20
4	137	-3
4	145	-9
4	147	1
4	148	-15
4	149	-14
4	150	-25
4	171	6
4	172	-2
4	189	17
4	207	13
4	209	14
4	216	7
4	226	10
4	260	-18
4	261	-4
4	267	-11
4	273	16
4	274	34
4	279	-9
4	301	19
4	302	-2
4	304	-4
4	306	6
4	307	32
4	308	15
4	309	4
4	310	3
4	311	0

4	312	20
4	313	-8
4	315	24
4	316	-4
4	317	13
4	329	20
4	330	15
4	332	29
4	333	21
4	334	-10
4	336	-14
4	338	-1
4	339	-14
4	340	-2
4	342	1
4	380	-21
4	381	-15
4	384	-9
4	387	9
4	388	0
4	389	-11
4	390	-12
4	391	-9
4	398	8
4	401	-22
4	409	7
4	410	19
4	411	-1
4	412	-12
4	413	-8
4	414	8

Table S8. PRE values measured for amide protons of cytochrome P450cam (probes 4 and 6) and Pdx (probes 1-3 and 5).

Probe number	Residue number	PRE value [s^{-1}]	PRE error [s^{-1}]
1	2	12.9	0.8
1	3	12.6	0.9
1	4	10	2
1	5	32	2
1	6	24	1
1	7	42	1
1	9	57	3
1	10	40	2
1	11	63	4
1	12	41.2	0.8
1	13	32	7
1	14	36	0.6
1	15	23	1
1	18	16.8	0.6

1	19	4.7	0.3
1	20	6.3	0.7
1	21	10	1
1	22	19	3
1	31	210	20
1	33	220	10
1	51	17	1
1	52	9.3	0.9
1	53	4.1	0.6
1	54	2.8	0.5
1	55	5	2
1	56	1	3
1	57	6	2
1	58	3	2
1	59	3	2
1	60	-0.2	0.9
1	62	4.1	0.3
1	63	5.9	0.4
1	64	6	1
1	65	3.7	0.6
1	68	7	1
1	69	9	1
1	72	13	1
1	73	13	4
1	74	20	2
1	75	25	8
1	76	33	5
1	77	14.2	0.4
1	78	13.2	0.5
1	79	15	2
1	81	8	1
1	82	7	1
1	83	11	1
1	89	4.8	0.8
1	90	11	1
1	91	0	1
1	93	-1	1
1	94	1	2
1	95	3.9	0.7
1	96	5	1
1	97	6	2
1	99	11	1
1	100	13.3	0.8
1	101	28	1
1	103	43	1
1	106	160	20
2	2	42.4	0.6
2	3	17.5	0.8
2	4	3	1
2	5	14	1
2	6	5.3	0.7

2	7	22.7	0.7
2	9	41	1
2	10	55	1
2	11	77	2
2	12	18	0.4
2	13	36	3
2	14	12.8	0.4
2	15	21.8	0.7
2	18	95.4	0.9
2	19	120.6	0.8
2	20	55.8	0.8
2	21	57	1
2	31	195	5
2	33	143	3
2	35	57	3
2	50	1	3
2	51	5	1
2	52	7.2	0.6
2	53	-0.2	0.5
2	54	12.1	0.5
2	55	20	2
2	56	12	1
2	57	25	1
2	58	29.1	0.9
2	59	14	1
2	60	12.1	0.9
2	62	18.7	0.3
2	63	28.5	0.4
2	64	-0.8	0.6
2	65	10.6	0.6
2	66	16	1
2	68	14	1
2	69	50	1
2	72	79	1
2	73	111	3
2	74	83	3
2	75	59	6
2	76	36	2
2	77	12.5	0.4
2	78	111.1	0.7
2	79	25	1
2	81	32.3	0.8
2	82	25	2
2	83	2.2	0.6
2	89	5.2	0.5
2	90	24.7	0.8
2	91	21	1
2	93	33.8	0.9
2	94	24	1
2	95	35	1
2	96	19	1

2	97	59.7	0.6
2	99	3	0.7
2	100	5.3	0.6
2	101	12.2	0.6
2	103	11.3	0.6
2	105	26	3
2	106	59	4
3	2	38	1
3	3	16	1
3	4	6	2
3	5	4	2
3	6	2	1
3	7	7	1
3	9	9	2
3	10	25	2
3	11	21	3
3	12	13.3	0.6
3	13	31	7
3	14	12.1	0.5
3	15	9	1
3	18	100	3
3	19	244	7
3	20	164	7
3	21	180	10
3	22	110	9
3	31	71	8
3	33	32	3
3	35	26	7
3	51	2	2
3	52	4.4	0.9
3	53	4	0.6
3	54	5.9	0.7
3	55	5	2
3	57	6	3
3	58	8	2
3	59	12	2
3	60	17	1
3	62	26.2	0.4
3	63	17.8	0.5
3	64	18	1
3	65	19.7	0.7
3	66	28.8	0.7
3	68	16	2
3	69	23	2
3	72	35	2
3	73	36	5
3	74	46	4
3	75	20	10
3	76	7	5
3	77	11.2	0.5
3	78	45.7	0.8

3	79	9	2
3	81	8	1
3	82	8	2
3	83	6	1
3	89	49	1
3	90	107	5
3	91	54	2
3	93	54	2
3	94	33	3
3	95	31	2
3	96	17	1
3	97	17	3
3	99	5	0.9
3	100	5.8	0.9
3	101	5	1
3	103	4.9	0.9
3	105	10	6
3	106	15	8
4	12	4	2
4	14	5	1
4	17	3	5
4	18	4	3
4	20	20	2
4	21	21	5
4	22	5	5
4	23	30	10
4	24	10	6
4	25	35	2
4	26	5	6
4	27	-10	30
4	30	2	4
4	34	18	5
4	35	4	8
4	36	1	5
4	37	-2	3
4	39	2	6
4	40	-10	10
4	44	3	7
4	46	0	10
4	47	4	6
4	48	4	3
4	49	-2	4
4	50	8	4
4	52	5	7
4	54	10	30
4	55	0	10
4	56	20	20
4	59	3	4
4	60	4	4
4	61	0	30
4	62	10	10

4	65	20	40
4	67	10	20
4	68	0	20
4	69	10	10
4	70	20	10
4	74	13	5
4	91	-3	9
4	109	10	8
4	137	6	9
4	145	10	10
4	147	10	9
4	148	7	7
4	149	5	9
4	150	0	10
4	171	3	4
4	172	2	8
4	189	0	10
4	207	9	8
4	209	-20	10
4	216	2	6
4	224	0	10
4	226	22	7
4	228	24	7
4	229	42	4
4	230	42	5
4	233	13	6
4	260	9	8
4	261	8	7
4	267	-1	5
4	271	16	5
4	273	5	3
4	274	10	10
4	275	60	10
4	279	27	6
4	299	5	2
4	301	10	20
4	302	13	3
4	304	4	2
4	306	9	4
4	307	-4	7
4	308	10	4
4	309	10	50
4	310	4	6
4	311	7	2
4	312	-10	10
4	313	10	30
4	314	8	5
4	315	3	3
4	316	0	20
4	317	20	10
4	323	3	4

4	329	-4	8
4	330	4	5
4	332	20	10
4	333	22	3
4	334	12	3
4	336	3	6
4	338	6	4
4	339	10	7
4	340	15	6
4	342	22	5
4	360	10	9
4	380	3	3
4	381	5	3
4	384	2	2
4	387	5	2
4	388	2	2
4	389	1	1
4	390	2	4
4	391	4	4
4	392	17	4
4	395	5	5
4	398	3	9
4	401	3	3
4	402	10	30
4	409	17	3
4	410	0	20
4	411	-1	6
4	412	1	3
4	413	5	4
4	414	4	5
5	2	-4	2
5	3	-1	2
5	4	-6	3
5	5	-3	4
5	6	-3	3
5	7	-5	2
5	9	2	4
5	10	1	2
5	11	1	4
5	12	-1	1
5	13	0	10
5	14	2	1
5	15	2	3
5	18	-2	2
5	19	-1.2	0.7
5	20	0	2
5	21	-1	3
5	22	10	10
5	31	10	30
5	33	10	9
5	35	7	9

5	51	-6	5
5	52	1	1
5	53	1	1
5	54	-2	1
5	55	-2	1
5	56	-2	3
5	57	-3	3
5	58	-2	2
5	59	-3	3
5	60	-1	2
5	62	1.2	0.5
5	63	2.1	0.6
5	64	1	2
5	65	2	1
5	66	-1	1
5	68	-1	2
5	69	2	3
5	72	5	3
5	73	2	6
5	74	5	3
5	75	0	20
5	76	-9	6
5	77	0.3	0.6
5	78	3.6	0.6
5	79	-1	4
5	81	1	2
5	82	0	3
5	83	-1	2
5	89	0	2
5	90	-3	3
5	91	-4	2
5	93	-3	2
5	94	-5	3
5	95	-2	1
5	96	-2	2
5	97	-2	3
5	99	-4	2
5	100	0	2
5	101	-7	2
5	103	-2	2
5	105	10	10
5	106	-10	10
6	12	-1	8
6	14	-6	7
6	17	0	3
6	18	7	6
6	20	-6	7
6	21	-4	5
6	22	26	7
6	23	2	9
6	24	0	10

6	25	6	6
6	26	-10	20
6	27	10	20
6	30	10	10
6	34	4	5
6	35	-1	5
6	36	1	5
6	37	-1	5
6	39	11	9
6	40	13	9
6	44	3	8
6	47	7	8
6	48	0	5
6	49	3	4
6	50	12	5
6	52	37	6
6	54	30	10
6	56	10	7
6	59	0	20
6	60	2	8
6	61	10	70
6	62	-10	20
6	65	20	10
6	67	20	30
6	69	0	20
6	74	-2	4
6	91	1	8
6	137	10	10
6	145	-4	9
6	148	-4	7
6	150	0	20
6	171	0	10
6	172	8	6
6	189	13	8
6	209	0	10
6	216	-8	3
6	224	40	10
6	226	10	10
6	228	90	20
6	229	180	30
6	230	420	30
6	233	0	10
6	260	0	10
6	261	50	10
6	267	-8	9
6	271	28	7
6	273	4	4
6	274	-10	10
6	275	30	20
6	279	2	4
6	299	-3	2

6	302	-1	5
6	304	-5	7
6	306	22	8
6	307	30	10
6	308	30	10
6	309	13	8
6	310	10	5
6	311	3	5
6	312	-10	10
6	313	0	10
6	314	60	20
6	316	19	8
6	317	0	40
6	329	0	20
6	330	-10	10
6	332	10	10
6	333	2	5
6	334	0	20
6	338	2	5
6	339	40	20
6	340	30	10
6	342	-10	8
6	380	10	10
6	381	20	10
6	384	0	10
6	387	2	2
6	388	0	4
6	389	7	5
6	390	0	10
6	392	10	10
6	395	-5.3	0.6
6	398	20	10
6	401	3	5
6	402	8	6
6	409	-2	4
6	411	5.3	0.6
6	412	0	20
6	413	3	2
6	414	0	3

Figures

Figure S1. Clustering of residues for which PREs unaccounted for by the main state were measured. The residues in green are those for which PREs were not measured or were successfully fitted by the main state; other residues are colored according to the clusters they belong to. The solid spheres of a given color depict the geometric center of the corresponding cluster. a) probe 2 (PREs measured on Pdx), cluster A – red, cluster B – blue, cluster C - orange; b) probe 3 (PREs measured on Pdx), cluster D – red, cluster E – blue, cluster F - orange; c) probe 4 (PREs measured on cytP450cam), cluster G – red, cluster H – blue, cluster I – orange; d) probe 6 (PREs measured on cytP450cam), cluster J – red, cluster K – blue, cluster L – orange. CytP450cam orientation is as in Figure 1.

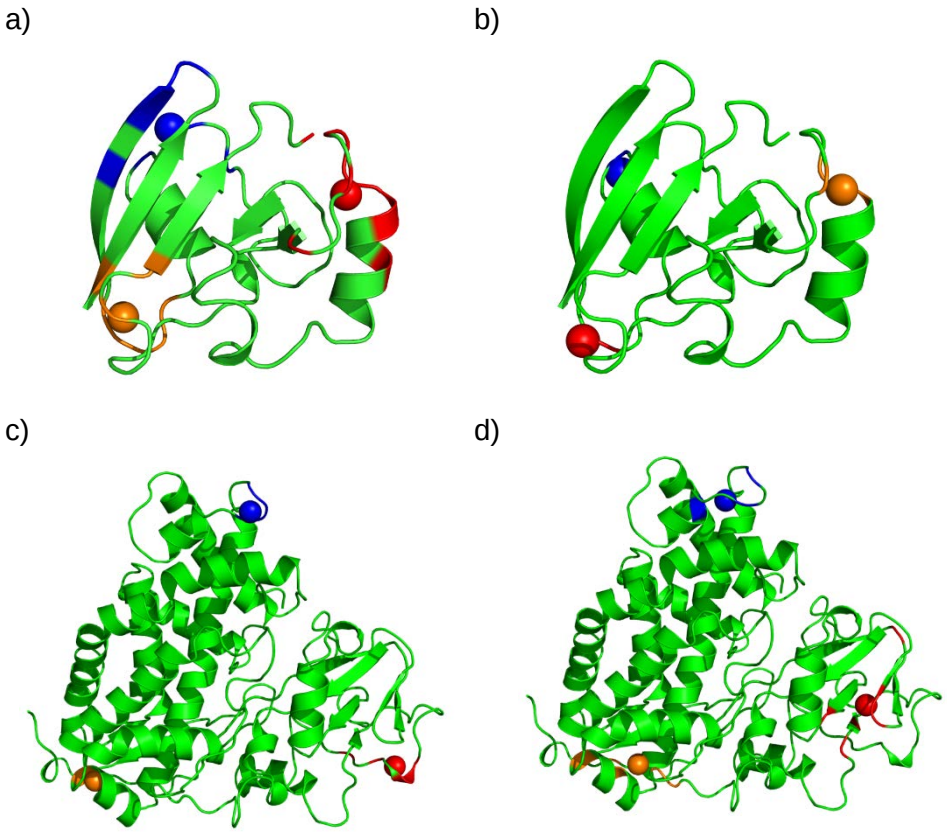
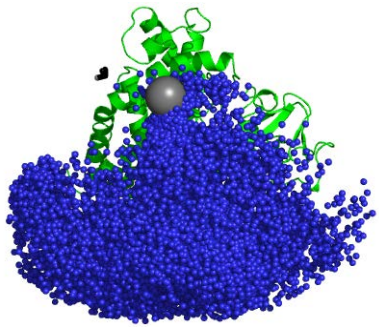
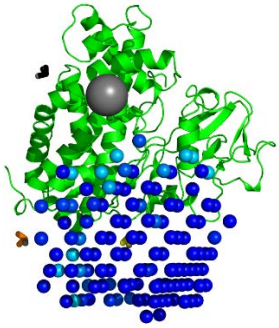
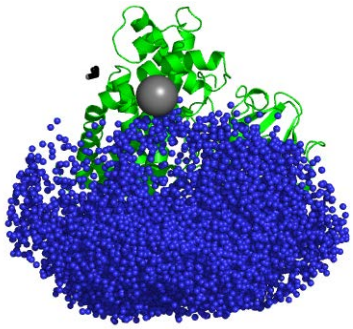
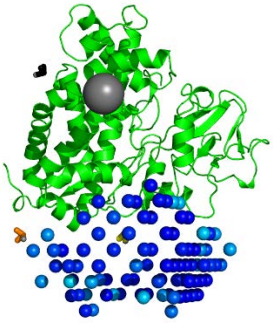
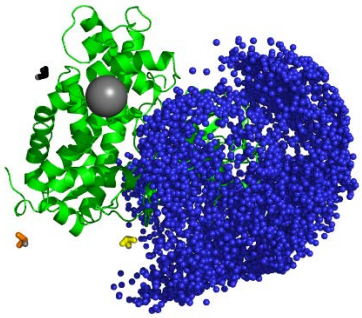
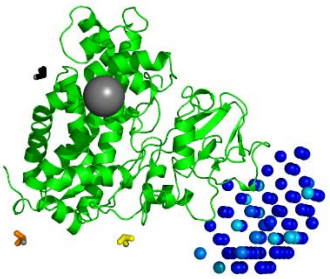
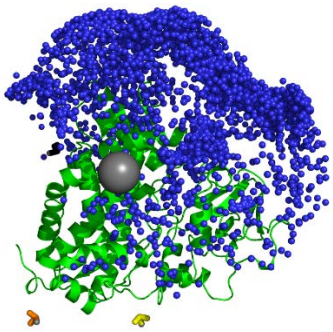
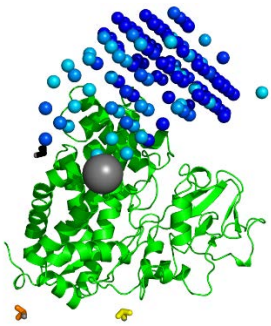
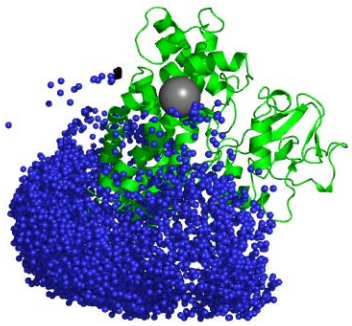


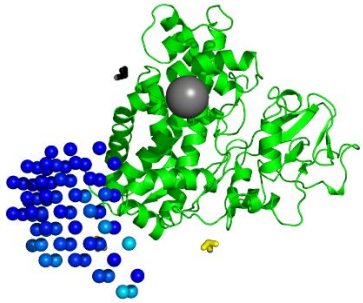
Figure S2. The smallest regions with minOR > 0 (left side of each panel) and all the 5 Å by 5 Å by 5 Å subsections of these regions containing structures that fit all the PREs in the given cluster simultaneously to a Q < 0.2 (right side of each panel). The cytP450cam is shown as green cartoon with the positions of probes 1, 2 and 3 indicated as black, orange and yellow axis frames respectively. The blue dots represent either the centers of mass of putidaredoxin (left side of the panel) or of the 5 Å by 5 Å by 5 Å subregions (right side of the panel). The different shade of blue in the right side panels denotes the quality of the fit (from dark blue Q = 0 to light blue Q=0.2) – see main text.

a) probe 2, region A	
b) probe 2, region B	
c) probe 2, region C	
d) probe 3, region D	

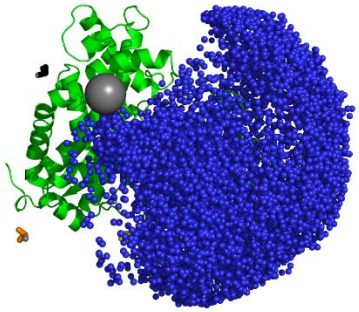
e) probe 3, region E 	
f) probe 3, region F 	
g) probe 4, region G 	
h) probe 4, region H 	

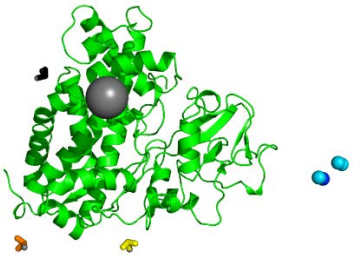
i) probe 4, region I



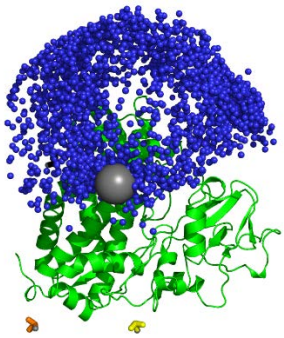


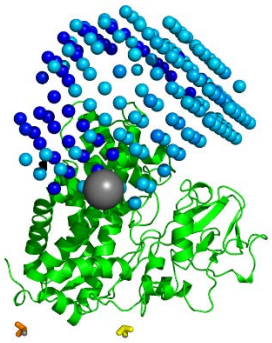
j) probe 6, region J

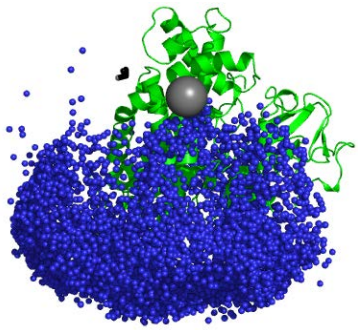




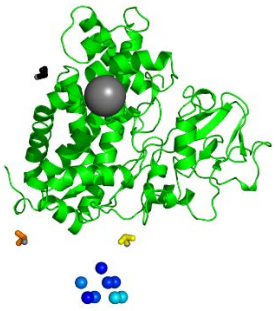
k) probe 6, region K







l) probe 6, region L



References

1. Rinaldelli M, et al. (2014) Simultaneous use of solution NMR and X-ray data in REFMAC5 for joint refinement/detection of structural differences. *Acta Crystallogr D Biol Crystallogr* 70(4):958–967.

2. Hiruma Y, et al. (2013) The structure of the cytochrome P450cam–putidaredoxin complex determined by paramagnetic NMR spectroscopy and crystallography. *J Mol Biol* 425(22):4353–4365.

3. 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(1):138–146.

4. Clarkson RB, et al. (1998) Multi-frequency EPR determination of zero field splitting of high spin species in liquids: Gd(III) chelates in water. *Mol Phys* 95(6):1325–1332.

5. Hudson A, Lewis JWE (1970) Electron spin relaxation of 8S ions in solution. *Trans Faraday Soc* 66:1297–1301.

3.5 How to tackle protein structural data from solution and solid state: An integrated approach

Azzurra Carlon,[a] , Enrico Ravera,[a] Witold Andrałojć,[a] Giacomo Parigi,[a] Garib N. Murshudov,[b] Claudio Luchinat[a]

[a]Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy1

[b]MRC Laboratory for Molecular Biology, Francis Crick Ave, Cambridge CB2 0QH, UK

Prog. Nucl. Magn. Reson. Spectrosc., ***2016***, *92-93, 54-70*

This page intentionally left blank

Progress in Nuclear Magnetic Resonance Spectroscopy 92–93 (2016) 54–70

Contents lists available at ScienceDirect

Progress in Nuclear Magnetic Resonance Spectroscopy

journal homepage: www.elsevier.com/locate/pnmrs

How to tackle protein structural data from solution and solid state: An integrated approach

Azzurra Carlon [a], Enrico Ravera [a], Witold Andrałojć [a], Giacomo Parigi [a], Garib N. Murshudov [b], Claudio Luchinat [a,*]

[a] Magnetic Resonance Center (CERM) and Department of Chemistry "Ugo Schiff", University of Florence, Italy[1]
[b] MRC Laboratory for Molecular Biology, Francis Crick Ave, Cambridge CB2 0QH, UK

Edited by David Neuhaus and Gareth Morris

ARTICLE INFO

Article history:
Received 15 September 2015
Accepted 13 January 2016
Available online 21 January 2016

Keywords:
Paramagnetic restraints
Residual dipolar couplings
Pseudo-contact shifts
Structural refinement
Integrated structural biology

ABSTRACT

Long-range NMR restraints, such as diamagnetic residual dipolar couplings and paramagnetic data, can be used to determine 3D structures of macromolecules. They are also used to monitor, and potentially to improve, the accuracy of a macromolecular structure in solution by validating or "correcting" a crystal model. Since crystal structures suffer from crystal packing forces they may not be accurate models for the macromolecular structures in solution. However, the presence of real differences should be tested for by simultaneous refinement of the structure using both crystal and solution NMR data. To achieve this, the program REFMAC5 from CCP4 was modified to allow the simultaneous use of X-ray crystallographic and paramagnetic NMR data and/or diamagnetic residual dipolar couplings. Inconsistencies between crystal structures and solution NMR data, if any, may be due either to structural rearrangements occurring on passing from the solution to solid state, or to a greater degree of conformational heterogeneity in solution with respect to the crystal. In the case of multidomain proteins, paramagnetic restraints can provide the correct mutual orientations and positions of domains in solution, as well as information on the conformational variability experienced by the macromolecule.

Contents

* Corresponding author at: Via L. Sacconi 6, 50019 Sesto Fiorentino (FI), Italy. Tel.: +39 0554574296; fax: +39 0554574924.
E-mail addresses: carlon@cerm.unifi.it (A. Carlon), ravera@cerm.unifi.it (E. Ravera), andralojc@cerm.unifi.it (W. Andrałojć), parigi@cerm.unifi.it (G. Parigi), garib@mrc-lmb.cam.ac.uk (G.N. Murshudov), claudioluchinat@cerm.unifi.it (C. Luchinat).
[1] URL: http://www.cerm.unifi.it/.

http://dx.doi.org/10.1016/j.pnmrs.2016.01.001
0079-6565/

1. Introduction

The most widely used techniques for elucidation of molecular structures at atomic resolution are X-ray crystallography and NMR spectroscopy, accounting as of July 2015 for 90% and 9.4% of all deposited protein structures, respectively, and 58% and 41% of all deposited nucleic acid structures. Besides new experimental techniques gaining more and more resounding success (e.g. cryo-EM, with more than 390 new entries in the last 3 years and resolution down to 2.2 Å [1]), X-ray and NMR still play a key role in answering many unresolved questions in the structural biology field. The unique importance of the integration of these two techniques has been recognized, taking advantage of the distinctive features of each. The strength of X-ray crystallography lies in a precise determination of a unique macromolecular structure (or a unique set of different structures present in the asymmetric unit of the crystal), whereas NMR spectroscopy has the power to probe the dynamics experienced in solution. On the other hand, neither X-ray nor NMR, if used as stand-alone tools, can provide a complete, precise and accurate picture of the biological system under investigation and of its interactions with other complexes or biomolecules. It also should be mentioned that X-ray crystallographic diffraction and NMR data are intrinsically different – the former gives information that progresses from the overall shape of the molecule up to individual atom positions as the resolution increases, whereas NMR provides immediate information about short-range inter-atom distances and bond orientations, which progresses to overall shape of the molecule with increasing number and quality of restraints. Therefore, the two techniques are highly complementary, because the combination of the two yields valuable information throughout the spectrum of distance scales, even in the presence of suboptimal X-ray and/or NMR data.

Despite its great success, there are some intrinsic limitations of X-ray crystallography: molecules in crystals experience crystal packing forces that may change their conformation and/or reduce conformational heterogeneity. NMR data are usually very accurate, but the collection of a large number of long-range interatomic distances is often very difficult, resulting in a lower precision of the NMR model with respect to the X-ray structure. Furthermore, NMR restraints are usually too few for solving molecular structures without strongly relying on prior knowledge defined by geometrical constraints based on covalent bonding. Therefore, it has long been known that X-ray and NMR data provide complementary information, which can be profitably analysed together for a more accurate description of biomolecules. Moreover, the complementarity of X-ray and NMR resides in the different types of information provided by these techniques, since X-ray relies mostly on the contribution given by the heavy atoms to the electron diffraction pattern, while for NMR the vast majority of restraints involve the hydrogen nuclei. Even more importantly, as anticipated above, at low and medium resolution, X-ray data contain information on overall shape and long-range structural details, whereas short-range structural details, of the order of the interatomic distances, are accessible only at very high resolutions, which are not always achievable. In contrast, NMR data mainly provide direct information on local details, in the form of interatomic distances or orientations of vectors connecting chemically bound nuclei. Therefore, information from NMR and X-ray data is perfectly complementary.

Among the structural restraints which can be obtained in NMR spectroscopy, pseudo-contact shifts (PCSs) [2] and residual dipolar coupling (RDCs) [3] have attracted increasing interest during the last decades for their intrinsic long-range nature. They can in fact provide structural information on the relative positions or orientations of pairs of atoms throughout the whole macromolecule or a large part of it. When the molecule is paramagnetic, dipolar interactions arise between the nuclei and the residual electron polarization, which is proportional to the magnetic susceptibility. If the magnetic susceptibility is anisotropic, these dipolar interactions do not average to zero upon rotation and PCSs arise (see later). An anisotropic magnetic susceptibility is usually associated with metal ions coordinated to the molecule [4–6] which, if not originally present, can be included by substitution of a diamagnetic metal ion [7–16] or, alternatively, rigidly attached through tags [17–41]. Other paramagnetic centres, such as organic radicals, have too little anisotropy to cause PCS or alignment effects. Magnetic susceptibility anisotropy also causes partial alignment of the molecule. In turn, partial alignment prevents internuclear dipolar interactions to be completely abolished by rotation, causing RDCs. This self-orientation is an alternate way to generate RDCs without using an external alignment medium [4,42–54]. PCSs and paramagnetic RDCs depend on the molecular nuclear coordinates in a common frame defined by the magnetic susceptibility anisotropy tensor associated with the paramagnetic metal. Self-orientation RDCs can also be obtained in the case of molecules for which the diamagnetic susceptibility is anisotropic [3,55–59] although in this case PCSs are not present.

It is interesting to observe that the presence of self-alignment also affects the chemical shifts of the observed species, if the chemical shielding of the nucleus is anisotropic. In the case of paramagnetic systems, the observed shift (not to be confused with the PCS) will be a combination of the effects of chemical shielding anisotropy and of the interaction with the electron average magnetic moment [60–62]; in diamagnetic systems it will reflect the chemical shielding anisotropy [63].

PCSs and RDCs contain structural information that has proved very helpful for solving protein structures [5,10,64–70], and they have therefore been included as structural restraints in the most commonly used programs for protein structure determination from NMR data [65,69,71–75].

PCSs and RDCs are even more precious restraints in the investigation of proteins constituted by multiple domains, and of protein–protein complexes. In the case of rigid systems, in which the structure of each single unit is known, PCSs and RDCs can be

used to determine the relative arrangements of the individual units [76–83]. An easy way to recover information on the relative orientations of rigid units is to estimate the magnetic susceptibility tensors for each of the units and then superimpose them. The presence of degenerate solutions can be removed by the use of an adequate number of metal ions or orienting media [84,85].

PCSs and RDCs are even more useful for retrieving important information about the investigated biological system when mobility is present. Especially RDCs, due to their intrinsic sensitivity to small structural changes, are widely used to probe conformational rearrangements of proteins occurring on timescales up to 10^{-2} s. Also in this case, the comparison of the anisotropy tensors or alignment tensors estimated for the single units constituting a multi-domain protein or a protein–protein complex can provide interesting insights into the mobility of the system.

Finally, PCSs and RDCs have been used for validation of existing molecular models and for their refinement. Several protocols have been presented for calculating refined structures based on restraining the backbone dihedral angles and/or the nuclear coordinates to undergo minimal deviations from the values of crystal models and simultaneously to improve the agreement with the experimental PCSs and/or RDCs, or also by allowing for sizable global changes [35,64,86–93].

More recently, PCSs and RDCs have been used for structure refinement in combination with X-ray data, by implementing them as structural restraints in the program REFMAC5 [94]. Differently from the above-mentioned approaches, the refinement protocol implemented in REFMAC-NMR takes advantage of the structural information contained not only in the available crystallographic model, but also directly in the primary X-ray data (i.e. electron diffraction pattern) and in the NMR data. This permits the limitations of the direct use of crystallographic models, whose accuracy in the atomic coordinates is related to the resolution of X-ray data, to overcome, obtaining improved structural models with accuracy determined equally by the contributions of both X-ray and NMR data.

2. Combining X-ray and solution NMR data

2.1. General aspects

In several cases, a large number of violations of solution NMR data have been noticed with respect to structural models obtained from X-ray crystallography. Historically, such discrepancies have also occasionally led to correction of the X-ray structure on the basis of the NMR information [95]. However, such discrepancies may either reflect real differences between the true structures of the molecules in crystals and in solution, or may be due to the different but complementary information contained in the X-ray and NMR data, thus the true structures may be the same but the process of deriving the models by analysing the different types of data may lead to inconsistencies. In a number of cases, refinements performed using combined X-ray and NOE-derived restraints revealed large consistency of the data, resulting in the improvement of the geometry of the model in terms of Ramachandran plot with respect to the structure calculated without NMR data [96–100]. The few violations were interpreted as real differences between the structures in crystals and in solution, mostly ascribable to a limitation of the conformational freedom for some flexible parts and to the presence of crystal packing forces in the solid state. In other cases, the joint refinement provided more accurate models, in the sense that they better fulfil independent sets of experimental observables, especially in the presence of poorly determined regions of the electron density maps, due to packing disorder within the crystal [101], or in the presence of medium or low resolution diffraction data [102].

As an example, some discrepancies were observed between a very large NMR dataset collected for ubiquitin, a protein frequently used as a model system for the development of new NMR methodologies, and its crystal structures [103]. These discrepancies can be interpreted as mobility effects and used to derive motional information on the system [103–105]. On the other hand, they can also be explained by an intrinsic inaccuracy in the crystallographic models, so that a single refined structure could adequately explain the NMR data [94,106]. This points out some critical aspects in the interpretation of potential inconsistencies observed between solution data and crystal models, which requires a careful analysis before any further action.

Hence, a joint refinement of the crystal structural model with solution data may represent a first step that should be undertaken for the correct interpretation of any inconsistencies between X-ray and NMR data. In particular, the assessment of the significance of the possible discrepancies represents a crucial point, which must take into account the inaccuracies characteristic of the two types of data.

Zweckstetter and Bax [107] examined the effect of the presence of structural inaccuracy, also referred to as "structural noise", demonstrating through a number of simulations and Monte Carlo analysis that the agreement between the experimental and back-calculated RDCs is strictly related to the amount of structural noise present in the model. Moreover, in many cases the agreement of the NMR data was observed to depend on the resolution of the X-ray structure, such that it improves significantly for crystal structures solved with higher resolution [90,108–110]. Indeed, the uncertainty in heavy atom positions necessarily affects the positions of ^{1}H nuclei; because of the small contribution hydrogens make to the diffraction pattern, information on their positions from the experimental data is usually insufficient, and the different computational tools used to position them indirectly on the basis of ideal geometrical values often give different results. On the other hand, ultrahigh resolution X-ray structures show that the positions of the resolved ^{1}H nuclei can deviate from ideal geometry [111]. Indeed, as extensively demonstrated by many authors [111–114], the geometries of molecules may often deviate from standard values for various types of protein bond length and bond angles, due to effects that are conformation-dependent, explained by avoidance of atomic clashes or optimization of favourable electrostatic interactions.

Regarding the positions of hydrogen atoms, attention should be paid to the fact that the hydrogen electrons (seen by X-ray data) are not centred on the positions of the nuclei (seen by NMR data) but are closer to the atoms to which they are attached. Therefore, the hydrogen "positions" must differ for the evaluation of the X-ray and NMR restraints, to take into account the different distances of hydrogen nuclei and their electron clouds from the atoms to which they are bound.

In summary, it is of primary importance to take advantage of X-ray and NMR data in an efficient and unbiased way, and to combine them carefully so as to avoid any possible misinterpretation of the data. As already mentioned, REFMAC-NMR performs a structural refinement by the simultaneous use of primary X-ray and NMR data, mostly PCSs and/or RDCs, and thus can be used to detect the presence of significant discrepancies, outside the uncertainty related to the experimental dataset, between solution and crystal structures; in cases where all data are consistent, it can provide a more reliable structural model.

In the sections that follow, a systematic approach to the investigation of possible structural differences between crystal and solution data, and possibly to the calculation of refined models, is proposed, and applied to a number of different proteins.

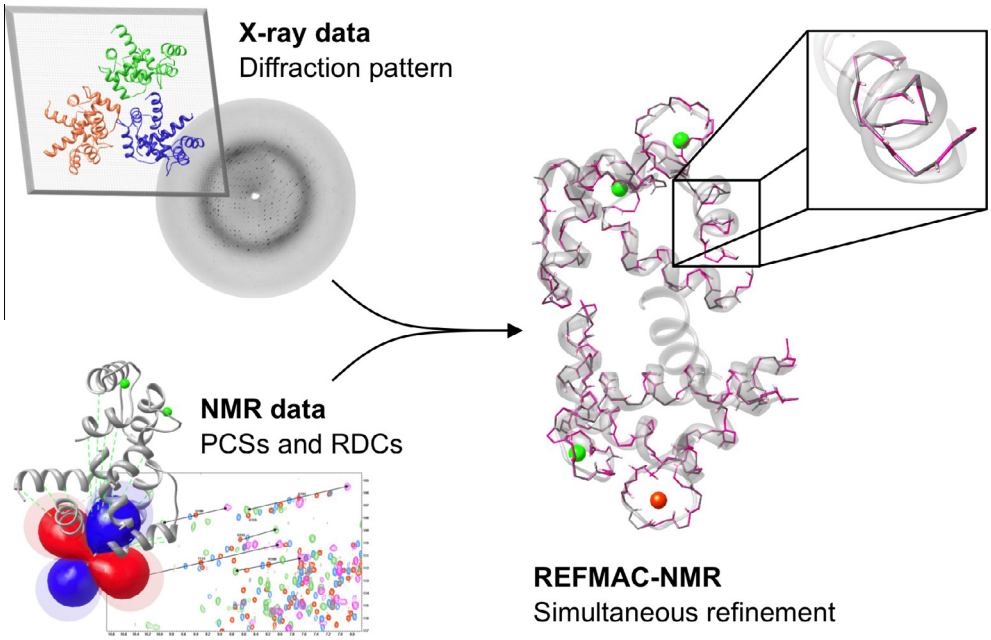


Fig. 1. Schematic representation of the information used in REFMAC-NMR structure refinement.

2.2. Structure refinement using REFMAC-NMR

REFMAC-NMR consists of an extension of the crystallographic refinement program REFMAC5 [115], a part of the CCP4 suite [116] and commonly used for medium-to-low resolution structures, by the inclusion of PCS and RDC NMR data as structural restraints in addition to the X-ray diffraction pattern (Fig. 1).

In the case of medium-to-low resolution X-ray data, the uncertainty related to the heavy atoms coordinates, as determined from X-ray data alone, may directly affect the positions of those remaining atoms which cannot be observed in the X-ray diffraction pattern (i.e. those with very weak electron density) and which are then included in the model according to the covalent geometry. This effect mostly involves the detection of hydrogen positions and often results in a non-negligible disagreement with experimental NMR data collected for them. Therefore, the inaccuracy related to the structural model can be taken into account through joint refinement against both NMR and X-ray data.

The general approach used by REFMAC-NMR for the inclusion of NMR data in the structure refinement consists of: (i) a first minimization against the X-ray data alone, with an automatic setting of relative weights of geometry and X-ray violations, possibly followed by manual adjustment of the weights to reduce the calculated rms deviations of bond lengths, bond angles, and chiral volumes, if too large; and (ii) a second minimization performed with the same weight settings including the NMR data (i.e. PCSs, RDCs), in order to decrease the discrepancy between experimental and back-calculated data, often expressed by means of the *Q*-factor value defined as:

$$Q = \sqrt{\frac{\sum_i (\exp_i - \mathrm{calc}_i)^2}{\sum_i \exp_i^2}}$$

As detailed in [94], in REFMAC-NMR the NMR restraint contribution (t) to the total minimized function was defined as:

$$t = k_{\mathrm{PCS}} \sum_i w_i \left[\max \left(|\mathrm{PCS}_i^{\mathrm{calc}} - \mathrm{PCS}_i^{\mathrm{obs}}| - T_i, 0 \right)^2 \right] + k_{\mathrm{RDC}} \sum_i w_i \left[\max \left(|\mathrm{RDC}_i^{\mathrm{calc}} - \mathrm{RDC}_i^{\mathrm{obs}}| - T_i, 0 \right)^2 \right]$$

where T_i is the tolerance on each PCS or RDC value, w_i is the weight, and k_{PCS} and k_{RDC} are the overall weighting factors for PCSs and RDCs respectively. Besides the already optimized weight matrix values, the second minimization requires optimization of the weights of the NMR data and of the additional torsion angle restraints *pep1*, *pep2* and ω. Three further torsion angles were in fact introduced in the REFMAC library to restrain the planarity of the proper (O_i, C_i, N_{i+1}, C^{α}_{i+1}), of the improper (C_{i-1}, N_i, C^{α}_i, H_i) (out of plane bending of HN), and of the proper C^{α}_i–C_i–N_{i+1}–C^{α}_{i+1} dihedral angles (*pep1*, *pep2* and ω, respectively). This is needed to avoid departure of geometric parameters from ideality being caused by the inclusion of the NMR data in the refinement. Furthermore, overall weighting parameters are also introduced to enforce the contribution given by the ideal geometries on all the atoms involved ('*weight refined_atoms*') or not involved ('*weight other_atoms*') in the calculation of gradients and of the second derivatives corresponding to X-ray reflections.

In this second minimization, together with the refined structure, the best-fit magnetic susceptibility anisotropy tensors or the alignment tensors associated to the PCSs and paramagnetic RDCs or to the diamagnetic RDCs, respectively, are also calculated.

Comparison between the results of refinement calculations performed with and without the NMR restraints provides the first indication of consistency between X-ray and NMR datasets and, thus, whether significant differences are present between the crystal and solution structures.

If the inclusion of NMR data in the structure refinement does not produce evident worsening in the agreement of X-ray data with the structural model (indicated by the free *R* factor), and provides at the same time a good agreement of the PCS/RDC data (indicated by the *Q*-factors), it can be concluded that the molecule has essentially the same structure in the crystal and in solution. The local adjustments driven by the NMR data may then reveal whether or not the discrepancies initially observed with the crystal structures can be explained within the uncertainty related to the X-ray data, and in this case, REFMAC-NMR can provide more reliable structural models.

In cases where it is not possible to obtain a good agreement between NMR (high *Q* factors) and X-ray (high free *R* factor) data with the same refined structure, it can be concluded that there

are significant differences between the structures of the molecule in solution and in the solid state. Still, a top-down approach can be used by performing independent refinement for different units/sub-units of the investigated system, or by discarding the residues responsible for the mismatch. Obviously, a limitation of this approach is the reduced number of NMR measurements available for the different units/sub-units, which may result in a larger uncertainty in the anisotropy tensors associated with the experimental PCSs and RDCs.

For multi-domain protein cases, the number of PCSs and RDCs available is usually sufficient to be used for an independent structure refinement of each domain. From these data, the anisotropy tensors are determined independently for each domain. It should be mentioned that, while RDCs provide valuable information both in the absence and in the presence of inter-domain mobility (see Section 2.4), PCSs collected for domains moving with respect to the metal ion also report on the variation of the metal-nucleus distance and, with some caveats (see Section 2.4), they are better used for calculation of the anisotropy tensors for the domain bearing the paramagnetic metal.

2.3. Tensor calculation

PCSs arising in the presence of a paramagnetic metal depend on the magnetic susceptibility anisotropy tensor and on the nuclear coordinates according to the following equation [2,61,117,118]:

$$\mathrm{PCS}=\frac{1}{4\pi r^3}\left[\chi_{zz}\frac{2z^2-x^2-y^2}{2r^2}+(\chi_{xx}-\chi_{yy})\frac{x^2-y^2}{2r^2}+\chi_{xy}\frac{2xy}{r^2}+\chi_{xz}\frac{2xz}{r^2}+\chi_{yz}\frac{2yz}{r^2}\right] \quad (1)$$

where x, y, z are the coordinates of the nucleus when the metal ion is defined to be at the origin of the coordinate system, r is the distance between the observed nucleus and the metal ion, and χ_{ij} are the components of the symmetric susceptibility anisotropy tensor.

Dipolar couplings are averaged in solution according to the following relation

$$\mathrm{RDC}=\frac{\mu_0\gamma_A\gamma_B\hbar}{4\pi^2 r_{AB}^3}\langle 3\cos^2\theta-1\rangle \quad (2)$$

where r_{AB} is the distance between the two coupled nuclei A and B and θ is the angle between the interspin vector $\boldsymbol{r_{AB}}$ and the external magnetic field; therefore, upon free rotation, they average to zero. However, when a partial orientation is imposed, the average is no longer equal to zero. For instance, when orientation arises because of the presence of a paramagnetic centre, and thus depends on the same χ tensor components present in Eq. (1), Eq. (2) is recast into the form [4,61]:

$$\mathrm{RDC}=3k\left[\chi_{zz}\frac{2z_{AB}^2-x_{AB}^2-y_{AB}^2}{2r_{AB}^2}+(\chi_{xx}-\chi_{yy})\frac{x_{AB}^2-y_{AB}^2}{2r_{AB}^2}+\chi_{xy}\frac{2x_{AB}y_{AB}}{r_{AB}^2}+\chi_{xz}\frac{2x_{AB}z_{AB}}{r_{AB}^2}+\chi_{yz}\frac{2y_{AB}z_{AB}}{r_{AB}^2}\right] \quad (3)$$

where

$$w_{AB}^2=(w_A-w_B)^2 \quad (w=x,y,z)$$

$$K=-\frac{S_{LS}}{4\pi}\frac{B_0^2}{15kT}\frac{\gamma_A\gamma_B\hbar}{2\pi r_{AB}^3}$$

and S_{LS} is the model-free order parameter, introduced to take account of some average local mobility of the coupled nuclei vectors. RDCs do not depend on the position of the coupled nuclei with respect to the metal ion.

Diamagnetic RDCs are described by an equation of the same form as that for paramagnetic RDCs (Eq. (2)):

$$\mathrm{RDC}=-\frac{3\mu_0 S_{LS}}{8\pi^2}\frac{\gamma_a\gamma_b\hbar}{r_{AB}^2}\left[A_{zz}\frac{2z_{AB}^2-x_{AB}^2-y_{AB}^2}{2r_{AB}^2}+(A_{xx}-A_{yy})\frac{x_{AB}^2-y_{AB}^2}{2r_{AB}^2}+A_{xy}\frac{2x_{AB}y_{AB}}{r_{AB}^2}+A_{xz}\frac{2x_{AB}z_{AB}}{r_{AB}^2}+A_{yz}\frac{2y_{AB}z_{AB}}{r_{AB}^2}\right] \quad (4)$$

where A_{ij} are the components of the molecular alignment tensor. The magnetic susceptibility anisotropy values are defined by axial and rhombic components given by

$$\Delta\chi_{ax}=\chi_{zz}-\frac{\chi_{xx}+\chi_{yy}}{2}=\frac{3}{2}\chi_{zz}$$

and

$$\Delta\chi_{rh}=\chi_{xx}-\chi_{yy}$$

where χ_{ii} are the components of the χ tensor in the frame in which it is diagonal. Analogously, the anisotropies of the alignment tensor are described by the fraction of alignment along the z axis (A) and by the rhombicity (R), [119]

$$A=\frac{3}{2}A_{zz}$$

$$R=\frac{A_{xx}-A_{yy}}{A}$$

where A_{ii} are the components of the A tensor, again in the frame in which it is diagonal.

Anisotropy tensors and alignment tensors can be calculated from the agreement between experimental and back-calculated PCS and/or RDC data against a structural model. Several programs are available that can do this, among them the web-based interface FANTEN [76].

Recently, a method for structure calculations was proposed that does not require the calculation of a tensor describing RDC (and PCS) data. This tensor-free method, called the theta-method, relies on determining the molecular structure from MD trajectories and replica averaging, to obtain the best agreement between the experimental RDCs and theoretical RDCs directly evaluated using Eq. (2) [120].

2.4. Motionally averaged data

For systems composed of multiple units or domains, RDCs, which are independent of the distance to the metal ion, can provide a first estimate of the presence and extent of mobility. Indeed, whatever the conformational sampling experienced by the system, RDCs collected from single units or domains can always be described by a tensor, though averaged and reduced with respect to the real susceptibility anisotropy tensor of the paramagnetic ion (Fig. 2c). In the case of proteins composed of two domains, only one of which contains a paramagnetic metal, experiencing mobility with respect to one another, the RDCs collected for the domains without a metal ion are reduced by averaging in proportion to the extent of the motion. Therefore, the anisotropies of the tensors back-calculated from these data (let us call them $\Delta\tilde{\chi}_{ax(\text{domain 2})}$) are reduced with respect to the anisotropies of the tensors back-calculated from the RDCs of the metal-bearing domain ($\Delta\chi_{ax(\text{domain 1})}$). The extent of the motion can be easily quantified using the ratio of the magnitudes of the anisotropies of the tensors calculated for the two domains, i.e. from $\Delta\tilde{\chi}_{ax(\text{domain 2})}/\Delta\chi_{ax(\text{domain 1})}$. This is not valid in principle for PCSs, which cannot be represented by any mean tensor in the case of motion, unless all nuclei have a fixed distance from the metal ion during the motion [121], which is never the case in practice.

Once it has been assessed that significant mobility affects the system, PCSs and RDCs can be used to further characterize the resulting conformational variability. In this respect, a multitude

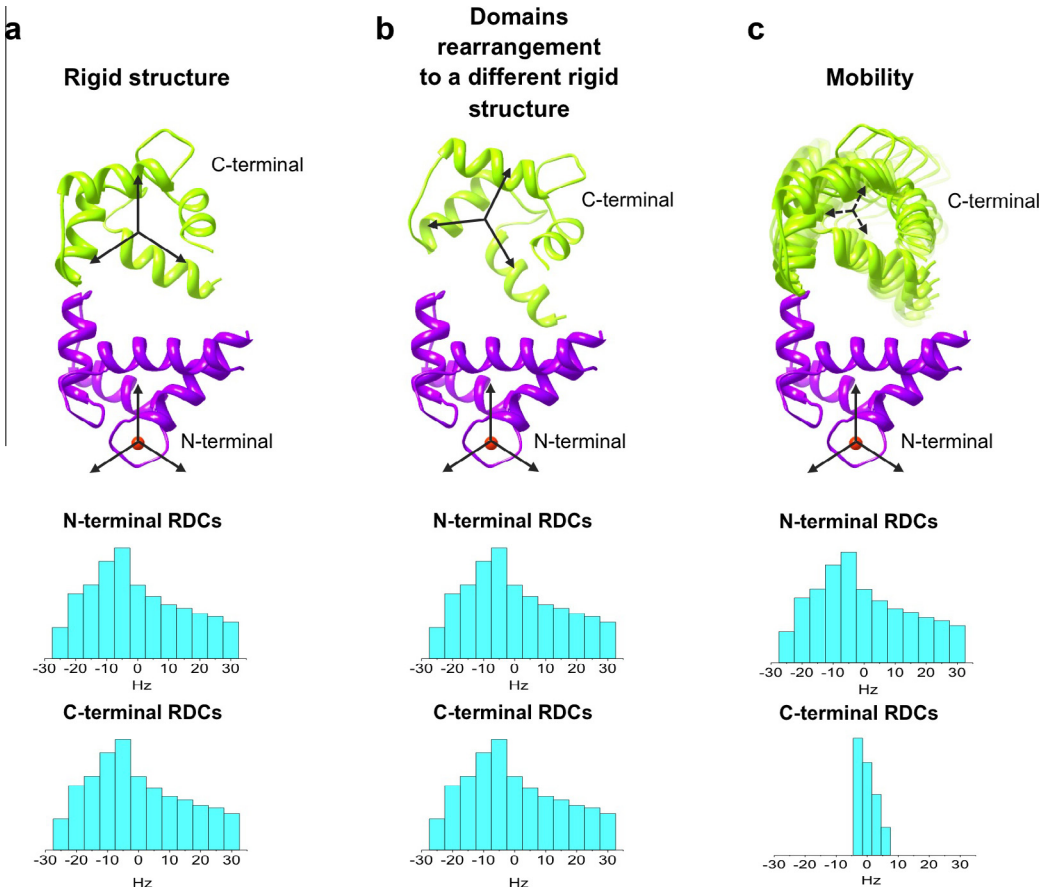


Fig. 2. The different effects of domain rearrangement and mobility on the range of observed RDCs.

of methods have been implemented to reconstruct the ensembles of putative conformations sampled by the system. This is, by definition, an ill-posed inverse problem, allowing for an infinite number of solutions (see Section 3.2.1) [57,81,103–105,122–152].

2.5. A step-by-step approach for the refinement of multidomain systems

When all domains composing a system have been satisfactorily refined against X-ray and NMR data, the best-fit tensors obtained for each domain should be compared (Fig. 2a–c). As already mentioned, this comparison might provide a very useful insight into the presence of inter-domain mobility or conformational rearrangements for the solution structure with respect to the crystal structure.

Scheme 1 reports a general approach that can be used to analyse the data and results that may be obtained under various different circumstances.

2.5.1. Check 1: Are tensor magnitudes equal?

As seen, mobility affects the measured RDCs, which are averaged over the values corresponding to all sampled conformations; this always results in a reduction of the observed values with respect to those that would have been observed in the absence of conformational variability. This, in turn, corresponds to a reduction of the associated anisotropy/alignment tensor (Fig. 2c).

In the case of paramagnetic RDCs, the averaged tensor obtained from the best fit of the experimental data against the structural model of a protein domain corresponds to an averaged magnetic susceptibility anisotropy tensor: the average is performed over all orientations sampled by the tensor with respect to the protein domain, and results in a reduced tensor.

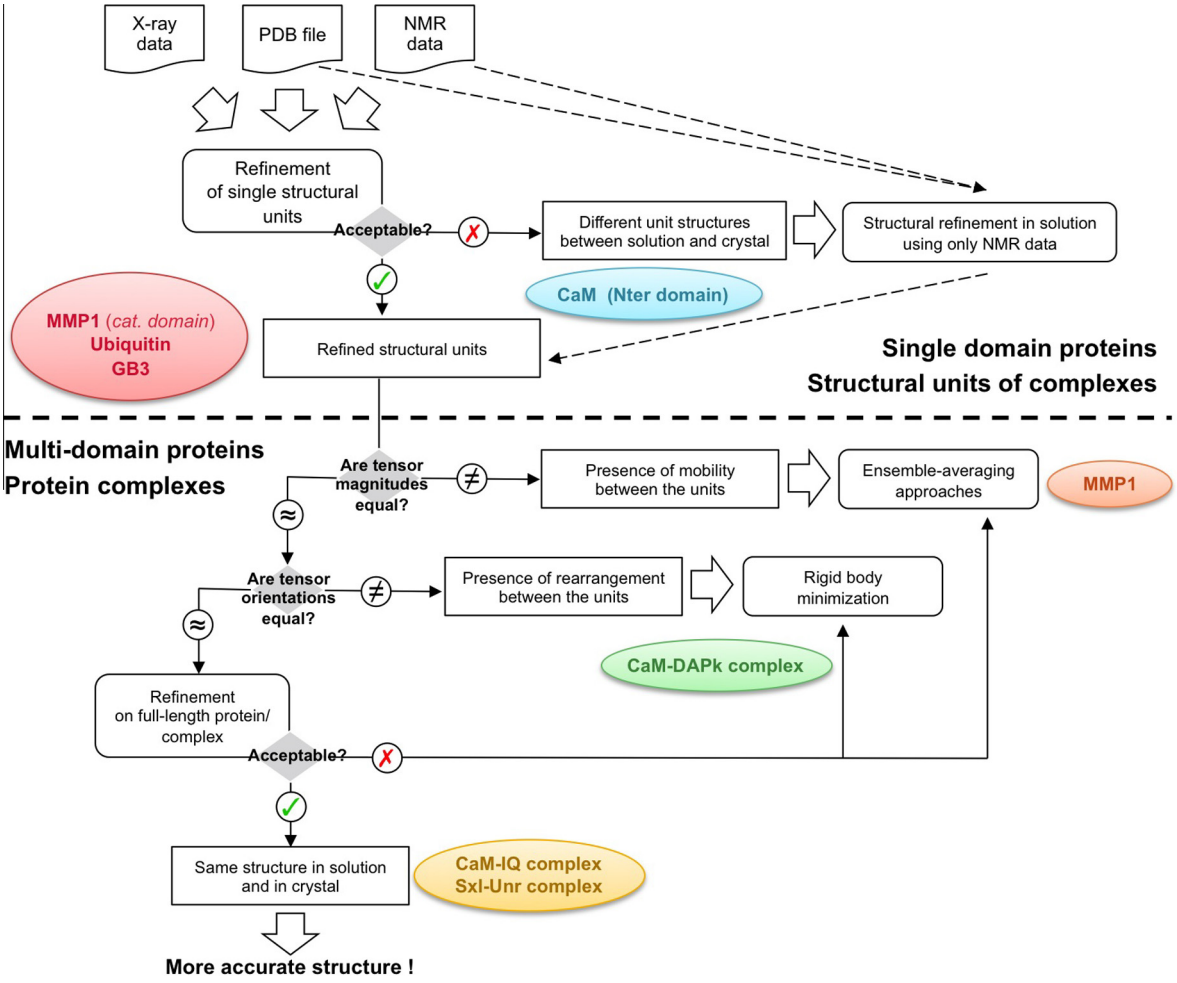
In the case of diamagnetic RDCs, an averaged tensor can still be calculated, although in this case it results from the averaging of alignment tensors, which differ from one another depending on the shape of the protein conformations (see Section 2.6). Therefore, a sizable difference in the RDC-determined tensors as calculated for the different domains is a clear indication of inter-domain mobility.

PCSs, which for rigid systems are usually employed to obtain a robust estimation of the tensor, are not scaled correctly in a dynamic system. For this reason, in principle they cannot be used in the structure refinement of the protein domain if the latter is mobile with respect to the position of the metal ion, unless the geometry of the system suggests that the variability in the metal position is small compared to the metal-nuclear distances, and some inaccuracy in the results can be tolerated [121].

Because of their sensitivity to global long-range conformational changes, RDCs and PCSs can be usefully used jointly to explore the conformational space sampled by the system in the case of conformational variability. Various approaches have been conceived for such a purpose (see Section 3.2.1), to generate and analyse ensembles of conformations in agreement with the averaged data.

2.5.2. Check 2: Are tensor orientations equal?

If the system can be stated to be rigid because the magnitudes of the anisotropy tensors calculated for the different domains are very similar, still the domains can be arranged differently in solution and in the solid state. This results in metal positions and tensor orientations that differ for the various domains (Fig. 2b).



Scheme 1.

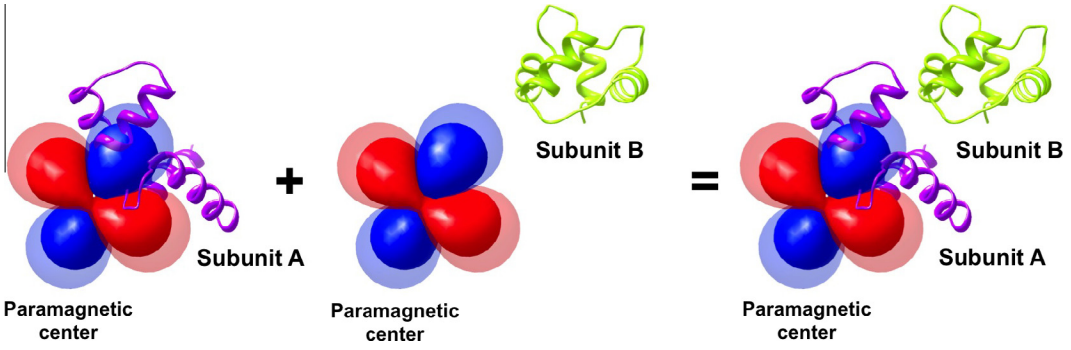


Fig. 3. Rigid body minimization based on the use of magnetic susceptibility anisotropy tensors. The tensors are first computed for each domain, then superimposed to reconstruct the reciprocal position of the two domains.

Therefore, the relative arrangements of the protein domains in solution can be retrieved by applying a rigid rototranslation to one domain in order to superimpose the sets of estimated tensors (i.e. the centre positions and axes defining their orientations) on those determined for the other domain (Fig. 3). This corresponds to recovering a structure in agreement with a set of tensors able to correctly describe the whole system.

2.5.3. Check 3: Is the full-length protein refinement acceptable?

If both the magnitudes and the orientations of the anisotropy/alignment tensors calculated for the protein domains are similar (Fig. 2a), as well as the metal positions, a full-length protein refinement can be attempted through REFMAC-NMR using a single tensor for each dataset. If this final refinement is also of good quality (low free *R*-factor and *Q*-factor), it can be stated that all data (X-ray and NMR) are in agreement with a single structural model and that the refined structure obtained represents a more accurate description of the protein. If not, as may happen in the presence of some differences in the anisotropy/alignment tensors obtained for the distinct protein domains, the discrepancies can find explanations in structural rearrangements or in conformational heterogeneity, which could not be clearly detected in the previous steps.

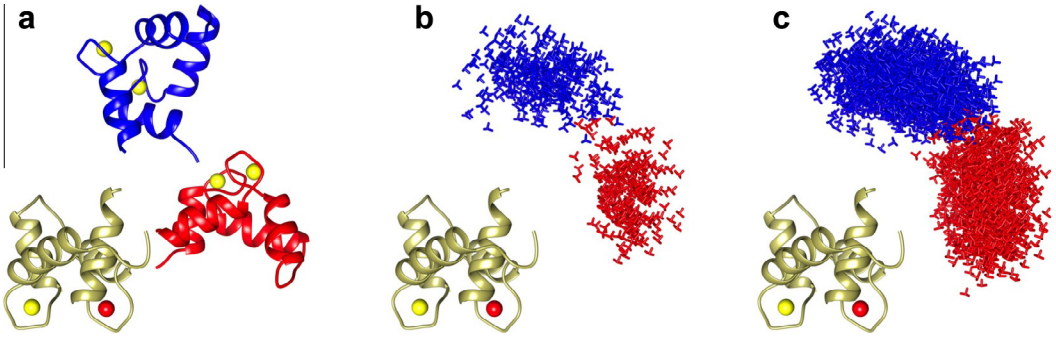


Fig. 4. (a) The two selected conformations of the two domain protein CaM used for the synthetic test: the N-terminal domain is shown in yellow, the C-terminal domain in blue (elongated conformation) or in red (compact conformation); (b and c) mobility schemes around the two centres: the C-terminal domain (replaced for each conformation with a 3-axes system centred on its centre-of-mass for a better visualization) is allowed to move within 20 Å and 20° (b) or within 20 Å and 40° (c) of translation and rotation, around the central conformations shown in (a).

Table 1
Axial components of the alignment tensors (A_a) and axial magnetic susceptibility anisotropy ($\Delta\chi_{ax}$)[a] calculated for ensembles of structures generated around the extended and compact conformations shown in Fig. 4a (blue and red, respectively).

d (Å)	δ (°)	# structures	N-terminal A_a[a]	C-terminal A_a[a]	N-terminal $\Delta\chi_{ax}$[a,d]	C-terminal $\Delta\chi_{ax}$[a,d]
Extended conformation						
0	0	1	1	1	1	1
20[b]	20[b]	321[b]	0.89	0.86	1	0.88
20[c]	40[c]	1418[c]	0.75	0.84	1	0.61
30	50	3872	0.73	0.69	1	0.39
Whole pool of sterically allowed conformations		32,723	0.52	0.41	1	0.025
Compact conformation						
0	0	1	1	1	1	1
20[b]	20[b]	247[b]	1.13	1.03	1	0.87
20[c]	40[c]	1264[c]	0.67	0.86	1	0.63
30	50	3393	0.81	0.62	1	0.43
Whole pool of sterically allowed conformations		32,723	0.63	0.49	1	0.025

[a] With respect to the single central conformation.
[b] Ensembles shown in Fig. 4b.
[c] Ensembles shown in Fig. 4c.
[d] With the metal in the N-terminal domain.

As already discussed, this approach can be applied with some flexibility in the definition of the subunits/domains composing the biological system investigated. The approach is limited by the number and variability of the experimental data available for the individual parts, and thus by the uncertainty in the corresponding estimated tensors. Moreover, the quality of the achieved agreements intrinsically depends on the experimental error as well as on the approximations in the applied model. Thus, a careful examination of all possible cases is generally recommended.

This step-by-step approach can also be applied in a top-down fashion: in this case, an overall refinement of the whole system is first attempted and, in case of unsatisfactory results (large Q-factor and/or R-free value), refinements of the individual subunits/domains can be performed.

2.6. Averaged tensors from diamagnetic and paramagnetic RDCs

As seen in Section 2.4, in case of paramagnetic (self-alignment) RDCs measured for two-domain proteins (or for other systems composed of two, or more, internally rigid domains), one of which contains a paramagnetic metal, conformational averaging greatly reduces the magnitude of the alignment tensor of the domain without the metal with respect to that of the metal-bearing domain. This reduction is dependent on the degree of interdomain mobility and, indeed, the ratio of the anisotropies of the alignment tensors for the two domains is sometimes used as a generalized order parameter describing interdomain mobility Fig. 2c [127,153]. The situation is very different for diamagnetic RDCs induced by external alignment media. In that case, none of the domains is fixed in the frame of the alignment tensor, and thus the RDCs measured for both domains are subject to motional averaging (except in the case where one domain is much bigger and/or more charged than the other and thus completely dominates the alignment properties of the system [154]). For this reason, the averaged tensors obtained for the two domains can have very similar magnitudes even in the presence of extensive interdomain mobility. Furthermore, as each conformer aligns to a different extent (according to its shape and charge distribution), different alignment tensors should be considered for the different conformations sampled by the system. This may have the result that, on passing from a single rigid state to an ensemble of conformations, an increase rather than a decrease of the alignment tensor can be observed, if some of the newly populated states align more strongly than the original structure.

It is instructive to illustrate the statements made above by means of some specific numerical simulations. For this purpose, two conformers (one extended and one compact, see Fig. 4) were chosen from a broad pool of sterically allowed conformations of

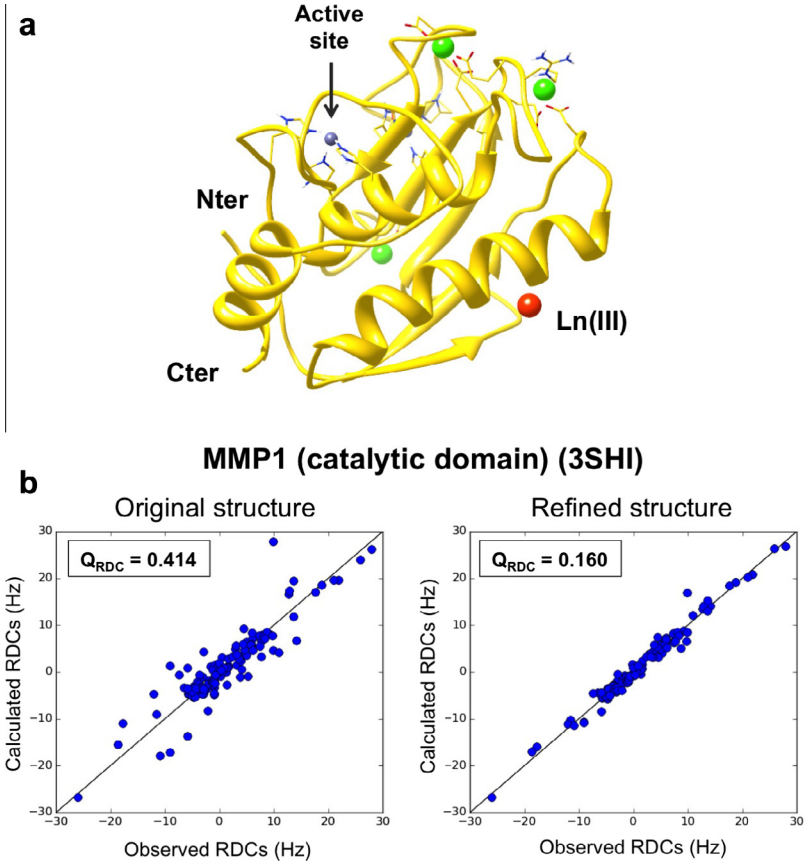


Fig. 5. (a) Structure of the catalytic domain of MMP1, showing the active site Zn(II) ion and the lanthanide tag position; (b) correlation plot between calculated and observed RDCs before and after the refinement.

the protein calmodulin (CaM, see Sections 3.1.4 and 3.2.2) [140]. The diamagnetic RDCs for the two conformers (and all the other conformers in the pool), induced by a fully steric alignment medium, were calculated using the program PALES [119,155]. The corresponding alignment tensors differ for the two conformers in such a way that the ratio of tensor magnitudes between the compact and the extended conformation is 0.83. The presence of conformational variability around a central structure was then simulated by constructing ensembles comprised of all the conformers with Cartesian and angular displacement from either the extended or the compact structure within defined thresholds (see Fig. 4 and Table 1). Averaged RDC values (and corresponding averaged tensors) were calculated by averaging (with equal weight) the PALES-derived RDCs.

Table 1 provides a summary of the obtained results in terms of the ratios of the magnitudes of the alignment tensors corresponding to the different ensembles with respect to those for the central (rigid) structures. Clearly, in both cases the magnitudes of the alignment tensors for the C-terminal and the N-terminal domains remain comparable even when a very large mobility is present. Therefore, it is not possible to exclude the presence of extensive conformational variability, or to quantify its extent, from the observation that different domains have alignment tensors of similar magnitude. This is in contrast with the average tensors determined from the paramagnetic RDCs, which are progressively reduced as conformational variability increases (last column of Table 1). It is also worth noting that for an ensemble around a compact conformer, the presence of limited mobility actually leads to an increase in the molecular alignment determined from diamagnetic RDCs, as discussed above.

3. Case studies

3.1. Single domain proteins

3.1.1. The catalytic domain of matrix metalloproteinase 1 (MMP1)

Matrix metalloproteinases are a family of multidomain proteins that mediate the hydrolysis of extracellular protein substrates such as elastin and collagen [156], and are putatively involved in cancer invasion. The interdomain orientation, which is crucial for catalysis, will be discussed in Section 3.2.1. These proteases have one catalytic Zn(II) ion in one domain, the catalytic (cat) domain. The structures of the cat domains of several MMPs have been solved by NMR and X-ray crystallography, mainly for the purpose of rational drug design [157].

REFMAC-NMR was used to refine the catalytic domain of MMP1 [94] (PDB code: 3SHI), for which PCSs of H^N nuclei and RDCs of H^N–N coupled nuclei for three paramagnetic lanthanides (Tb^{3+}, Tm^{3+}, and Yb^{3+}) bound to the protein through the CLaNP-5 tag were available [35]. Residues revealing extensive mobility through relaxation measurements were discarded from the refinement calculation. The original crystal structure (resolution: 2.2 Å) showed modest agreement for PCSs and RDCs, initially pointing to possible discrepancies between the solution and the crystal structure. The REFMAC-NMR refinement protocol was applied and a significant overall improvement in the agreement of NMR data was obtained, especially for RDCs, for which the *Q*-factor decreases from 0.414 to 0.160 (Fig. 5). The local adjustment to the crystallographic structure carried out by PCSs and RDCs resulted in a rms deviation for the backbone atoms with respect to the original model of only 0.039 Å. These small displacements in heavy atom positions, which

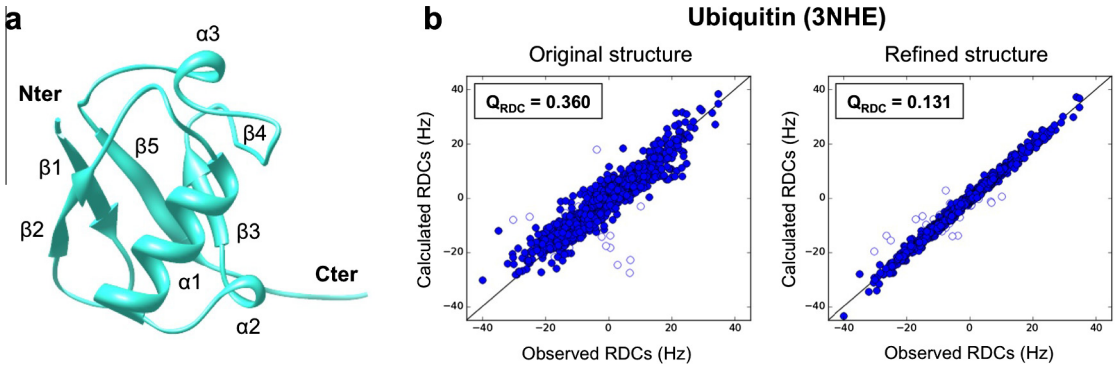


Fig. 6. (a) Structure of ubiquitin; (b) correlation plot between calculated and observed RDCs before and after the refinement (empty dots refer to residues 8 and 72).

are as compliant with the X-ray reflections as the original structure, are sufficient to improve the positioning of H^N nuclei in the structure and, thus, the agreement with the NMR data.

3.1.2. Ubiquitin

Ubiquitin is a small regulatory protein (76 residues, 8.5 kDa), which is found in almost all eukaryotic cells as well as in bacteria [158,159]. Attachment of this protein to other proteins exerts a signalling effect that depends on the location, the residue through which ubiquitin is coordinated, and how many ubiquitin molecules are bound [160]; a detailed discussion is given in Refs. [161,162]. A prototypical example is the covalent binding of ubiquitin to proteins targeted for proteasomal degradation [163,164].

The structure of ubiquitin has been solved as many as 178 times, alone or in complexes, by X-ray (129 PDB entries), NMR (46 solution and 2 solid-state entries) and electron microscopy (1 entry).

A sizable range of RDC datasets has been available for ubiquitin for a long time, and this wealth of RDC data has been interpreted differently by different groups. Prestegard and co-workers developed a methodology to extract simultaneously both structural and dynamic properties for ubiquitin, obtaining a solution structure (in agreement, on average, with the X-ray structure) and a characterization of its anisotropic internal motions, quantified by the generalized degree of order [153]. Analysis of the data was carried out by Clore and Schwieters looking for the smallest possible ensemble which could account for the experimental data, yielding a two-site jump model [104]. In the same paper it was also noted that the structural impact of residues showing the largest difference between the two states is minor, as they can be accounted for by compensatory movement of the backbone angles. Vendruscolo and co-workers developed a protocol for simultaneous determination of ubiquitin structure and dynamics by combination of molecular dynamics simulations and experimental information derived from NMR spectroscopy, finding a considerable conformational heterogeneity throughout the protein structure [105]. Along the same lines, Griesinger and co-workers postulated that ubiquitin follows mainly a single concerted motion, which allows for sampling many of the conformations observed in X-ray structures of complexes and at the same time explains the discrepancy between the experimental and the back-calculated data from any individual conformations [103]. In contrast, Bax and co-workers recently showed that it is possible to generate a single low-energy conformation that exhibits similar agreement to the experimental data (both the working set and the cross-validation set) as the previously discussed ensemble. This implies that the better fit of the ensemble to the experimental data with respect to any individual X-ray structure "is likely dominated by averaging of small errors in the backbone coordinates and not from true dynamics" [106]. It is worth noting that these conclusions do not to detract from the intrinsic ability of RDCs to reveal motions, but rather question the accuracy of reconstructed ensembles.

In this example, H^N–N RDCs measured from as many as 36 different alignment media [54,103] were used [94] for the refinement of an X-ray structure of ubiquitin (PDB code: 3NHE), which displayed a 1.26 Å resolution. The initial disagreement shown for the original structure decreases, in terms of Q_{RDC}, from 0.360 to 0.121 after performing REFMAC-NMR refinement (Fig. 6) [94]. Notably, a few violating RDCs were still found for the refined structure, mostly belonging to residues located in regions experiencing extensive mobility, as revealed by relaxation experiments [103,106], although not showing larger B-factors in the crystal structure. The possibility of achieving a similarly good agreement with the experimental data using different structures does not undermine the validity of the approach, but rather reflects the need to acquire as many experimental restraints as possible to reduce the uncertainty in the structural model.

3.1.3. IgG-binding domain of protein G (GB3)

Protein G is of particular interest, as it binds to a wide range of antibodies and its interactions have been well-characterized. Protein G has been shown to bind strongly to the Fc fragment and weakly to the Fab fragment of human immunoglobulin G (IgG) [165]. Protein–protein interactions between various protein G domains and isolated fragments of IgG have been studied by both solution NMR spectroscopy and X-ray crystallography [90,166–168].

For GB3 (PDB code: 1IGD), RDCs for five alignment media were collected for H^N–N, C^α–H^α, and C–C^α nuclear pairs [90]. Despite the acceptable agreement with experimental datasets already shown for the original model (with resolution 1.1 Å), REFMAC-NMR produced a further improvement in the Q_{RDC}, which decreased from 0.146 to 0.081.

A refinement of the same X-ray structure has been also performed by Bax and co-workers and used to draw interesting conclusions about the peptide bond H^N–N orientations as a result of pyramidalization of peptide N and fluctuations orthogonal to the peptide plane [90]. More recently, the same group presented a newly refined solution structure of GB3 using further C^α–C^β RDCs collected for three different alignment media, and validated the obtained structure using the prediction of three-bond J couplings by Karplus equations [168].

3.1.4. The N-terminal domain of calmodulin (CaM)

CaM is a 16 kDa, highly conserved protein, implicated in many biological functions. It is composed of two very similar domains. Each domain is composed of two EF-hand motifs, and can bind up to 2 calcium(II) ions, so that the full length protein can bind

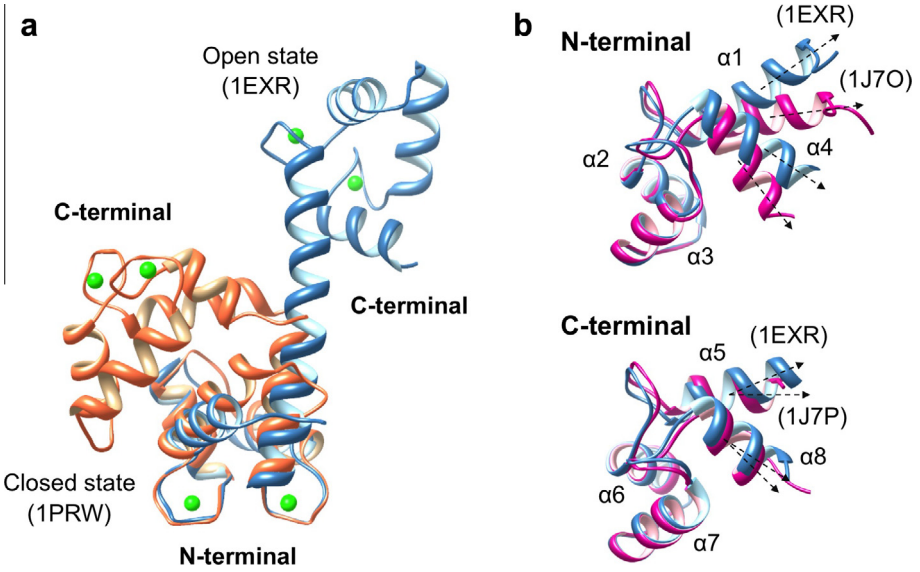


Fig. 7. (a) X-ray structures of free calmodulin, showing open and closed conformation; (b) differences between the solution structure and the X-ray structure of the individual domains [170].

up to 4 calcium(II) ions. In resting cells the protein binds one calcium(II) ion only. Upon increase of calcium concentration, all four sites are occupied, and the protein exposes two hydrophobic patches that can bind to several targets and is thus a prototypical intermediate messenger [169]. The interdomain mobility, which is relevant for the biological function, will be discussed in Section 3.2.1.

A set of diamagnetic H^N–N, C^α–H^α, C–C^α, and C–H^α RDCs [170] were used as input in an attempted refinement with REFMAC-NMR of the N-terminal domain of CaM (PDB code: 1EXR) [94]. However, in this case no satisfactory agreement was achieved for the NMR data without a significant increase in the *R*-free values and in the violations of geometrical parameters. This result confirms the findings by Bax and co-workers [170], showing substantial differences in the inter-helical orientations between the crystal and the solution structures of CaM N-terminal domain, mainly due to reorientation of helices 1 and 4. The REFMAC-NMR calculations indicate that the detected differences are outside the uncertainty of X-ray data (resolution: 1.0 Å) (Fig. 7).

3.2. Multi-domain proteins and complexes

3.2.1. Full length matrix metalloproteinase 1 (MMP1)

Matrix metalloproteinases (MMPs) are enzymes involved in extracellular matrix degradation, which is a fundamental step in many physiological processes such as tissue remodelling and repair [171,172]. Most MMPs comprise a prodomain that is removed upon activation, a catalytic domain responsible for hydrolytic activity, and a hemopexin-like domain that probably plays a role in substrate recognition. In several pathologies, overexpression of MMPs, or misregulation of their activity, is related to disease progression [173–175]. MMPs are thus validated pharmaceutical targets. Selectively targeting these proteins has turned out to be a difficult task [176,177], and structure-based drug design appears to be of great importance [178–180]. For this reason, the structural characterization of MMPs has received much attention [181].

After binding of the CLaNP-5 tag into the catalytic N-terminal domain of MMP-1 (see Section 3.1.1), PCS and paramagnetic RDCs induced by three different lanthanides (Tb^{3+}, Tm^{3+}, and Yb^{3+}) were also measured for the C-terminal hemopexin-like domain of the protein. The analysis of the collected RDCs reveals that the magnitudes of RDC-derived magnetic susceptibility anisotropy tensors for the catalytic domain are around 3–4 times higher than those calculated for the hemopexin domain. Such a reduction of the RDC-derived tensor observed for the metal-free domain is, as described above, a clear indication of conformational averaging. Therefore, a joint refinement against X-ray and NMR data was not performed. On the other hand, the system clearly samples only a limited subset of the sterically allowed conformations, as a uniform population of the entire conformational space would result in a much more drastic reduction of the tensor magnitudes (up to approximately 30 times).

As already mentioned, the problem of recovering the sampled conformers from averaged experimental observables is an ill-posed inverse problem. This means that an infinite number of, often very different, ensembles of structures can satisfy the experimental restraints equally well. Still the averaged experimental data clearly contain information about the conformational preferences of the system. In order to extract as much information as possible without the risk of over-interpreting the data, we developed an approach called Maximum Occurrence (MaxOcc) [134,140,182,183]. The method permits one to assign to each sterically allowed conformation a value, called MaxOcc, defined as the highest weight which the conformation can have in any ensemble in best agreement with the experimental observables. The MaxOcc of a conformation thus represents the upper bound that the averaged data impose on the statistical weight of that structure. Numerous synthetic tests have shown that in many cases the conformers with the highest MaxOcc correspond to those that are the most prominent in the actual conformational ensemble [31,85,140,151].

The MaxOcc analysis performed using averaged PCS and paramagnetic RDC data, complemented by the SAXS profile measured for the protein, could identify structures with MaxOcc values up to 47%. Interestingly, only a very small fraction of sterically allowed conformers have high MaxOcc (only 0.3% of the structures have MaxOcc > 40%), and all of them are clustered in a well-defined region of the conformational space of the protein, so that they are characterized by an interdomain orientation and position that can be defined relatively well (see Fig. 8). In these highest MaxOcc conformations the residues of the hemopexin domain responsible for collagen binding are well exposed to the solution; after collagen binding by the hemopexin domain, the catalytic domain is already

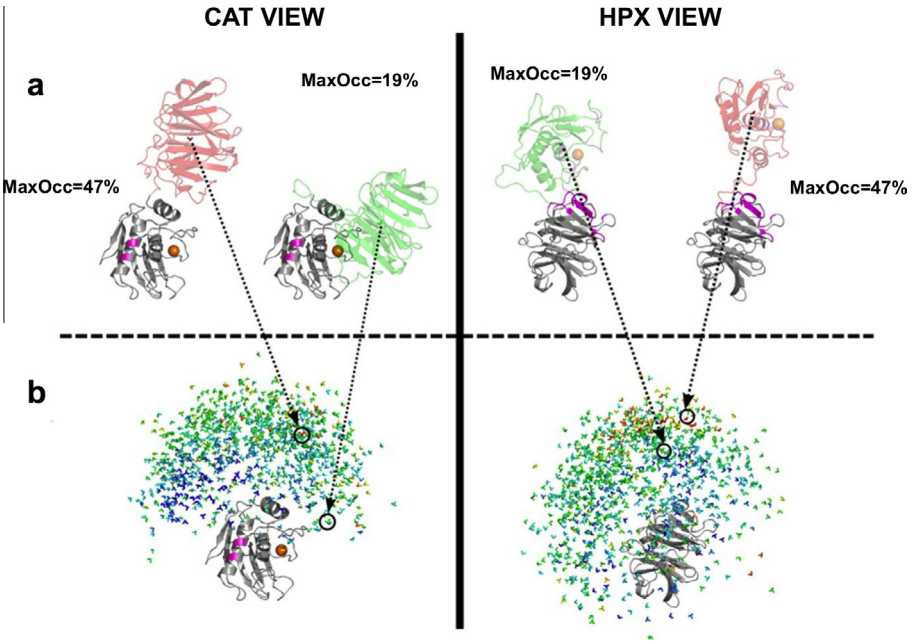


Fig. 8. Visualization of the results of MaxOcc calculations for 1000 randomly selected MMP-1 conformations. The conformations are displayed superimposed on the CAT domain (left) and on the HPX domain (right). The catalytic metal is represented as an orange sphere. Colours from blue (<5%) to red (47%) represent the MaxOcc values of the various structures. (A) The structure with the highest MaxOcc (47%) and the X-ray crystallographic structure 2CLT are coloured according to their MaxOcc values. (B) Each conformation is represented, for graphical simplicity, as a colour-coded 3-axis system, positioned at the centre of mass of the HPX (left column) or CAT (right column) domain. Reproduced with permission from Ref. [38] © of the American Society for Biochemistry and Molecular Biology.

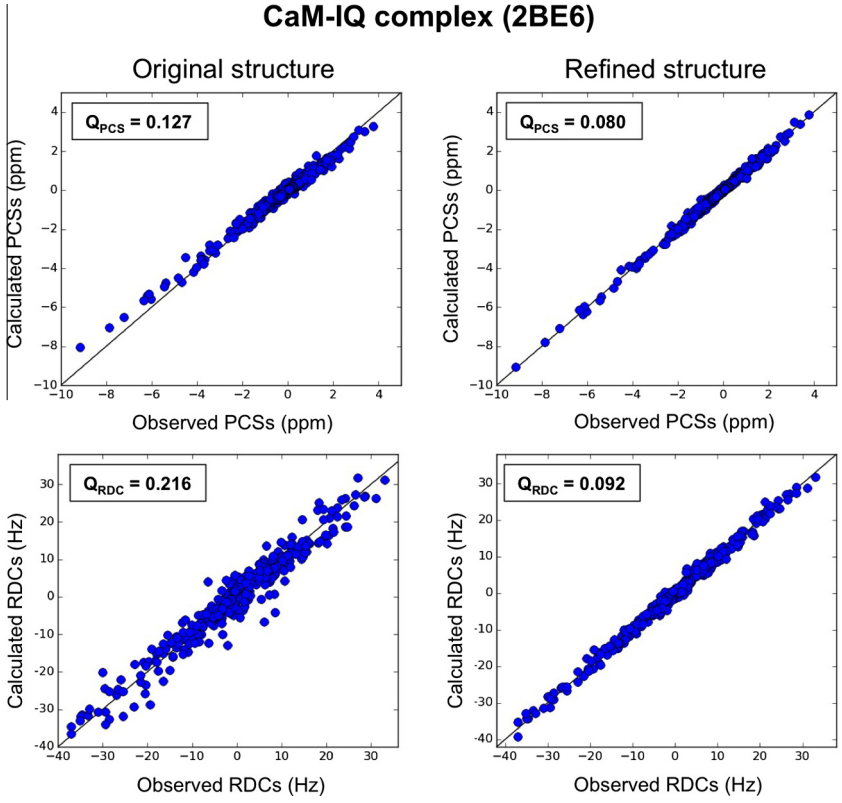


Fig. 9. Correlation plot of the PCS and RDCs calculated for chain B of the CaM–IQ 2BE6 structure before and after refinement.

preferentially oriented in such a way as to easily access the collagen as well, and a rotation by around 50° is sufficient to recover the previously proposed conformation responsible for the first step of collagenolysis [184]. Thus the MaxOcc analysis suggests that already in its free form in solution the full-length MMP-1 shows a preference for conformations which are poised for interaction with collagen and thus for catalytic activity [38]. Notably, the highest MaxOcc conformations differ largely from the closed MMP-1 structures obtained by X-ray crystallography, with a MaxOcc of about 20%.

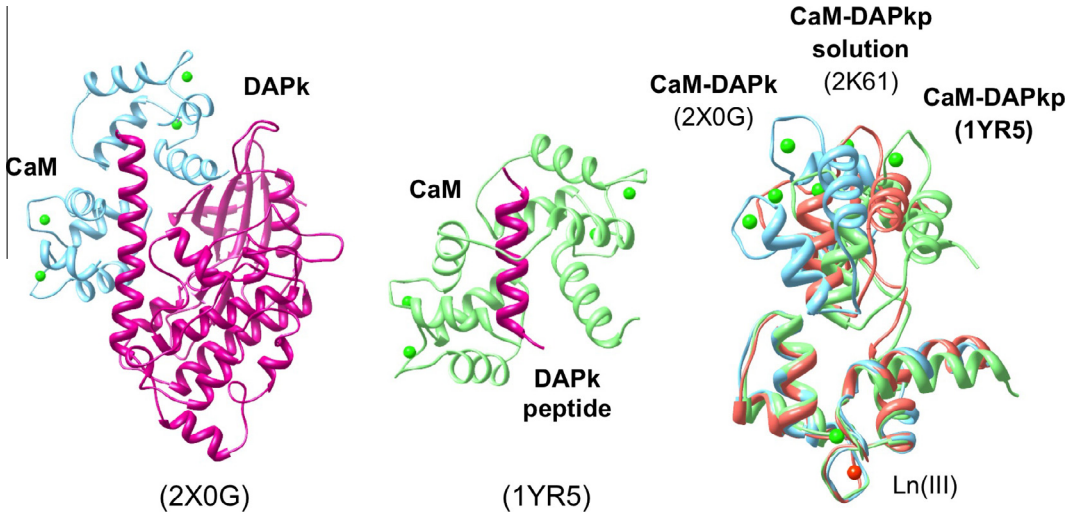


Fig. 10. Comparison between the X-ray structures of CaM bound to full-length DAPK (2X0G) [191], a peptide derived from DAPK–DAPKp– (1YR5) [92], and the solution structure in the CaM–DAPKp complex (2K61) [92].

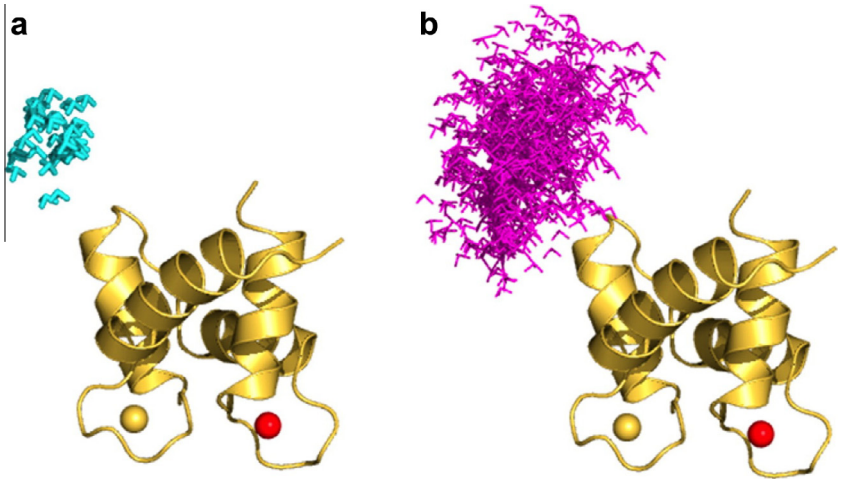


Fig. 11. (a) Smallest region with MaxOR = 1 and (b) smallest region with MinOR (Minimum Occurrence of Regions) = 0.54 of the CaM–DAPKp complex. Reproduced with permission from Ref. [151] © of the American Chemical Society.

3.2.2. Full length CaM

As already mentioned, CaM represents an interesting system as it comprises two almost rigid domains able to rearrange freely with respect to one another, when the protein is free in solution, because they are connected by a flexible linker. This high degree of flexibility is of course absent in the solid state, so that no joint refinement can be performed for the full-length structure using X-ray and NMR data. The protein has been engineered with the mutation N60D so as to make it possible for it to bind one lanthanide(III) ion selectively in the second calcium(II) binding site of the N-terminal domain [11].

The very different spread of the RDC data measured for the two domains when a paramagnetic metal is placed in one domain was earlier used to deduce the presence of a very large conformational variability [31,185–187]. Protein conformations can be ranked according to their compliance with PCSs and paramagnetic RDCs, using the MaxOcc analysis described for MMP-1 (see Section 3.2.1) [140,187,188]. This analysis takes advantage of the availability of the solution structures of the individual domains of the protein, previously obtained by refining the crystal structures using diamagnetic RDC data [170].

Interdomain flexibility is the basis of the functionally relevant role of CaM, for which a large number of binding partners have been identified. The formation of complexes induces CaM to assume a compact conformation, and depending on the specific binding interactions, the mobility between the two domains can be reduced to different extents, ranging from still very flexible [134,189] up to almost complete relative immobilization.

3.2.3. Calmodulin–IQ peptide complex (CaM–IQ)

Calmodulin is also involved in the regulation of voltage-gated calcium channels (Ca(V)s), and drives two self-regulatory calcium-dependent feedback processes that require interaction between Ca^{2+}–CaM and a Ca(V) channel consensus isoleucine–glutamine (IQ) motif: calcium-dependent inactivation (CDI) and calcium-dependent facilitation (CDF) [190].

The structure of CaM in complex with the IQ-recognition motif (PDB code: 2BE6) was refined using H^N, N PCSs and H^N–N RDCs measured for six different lanthanides (Tb^{3+}, Dy^{3+}, Ho^{3+}, Er^{3+}, Tm^{3+}, and Yb^{3+}) [144]. No reasonably good fit was observed for any of the three models (chains A, B, and C) contained in the asymmetric unit of the crystal. The stepwise approach described in

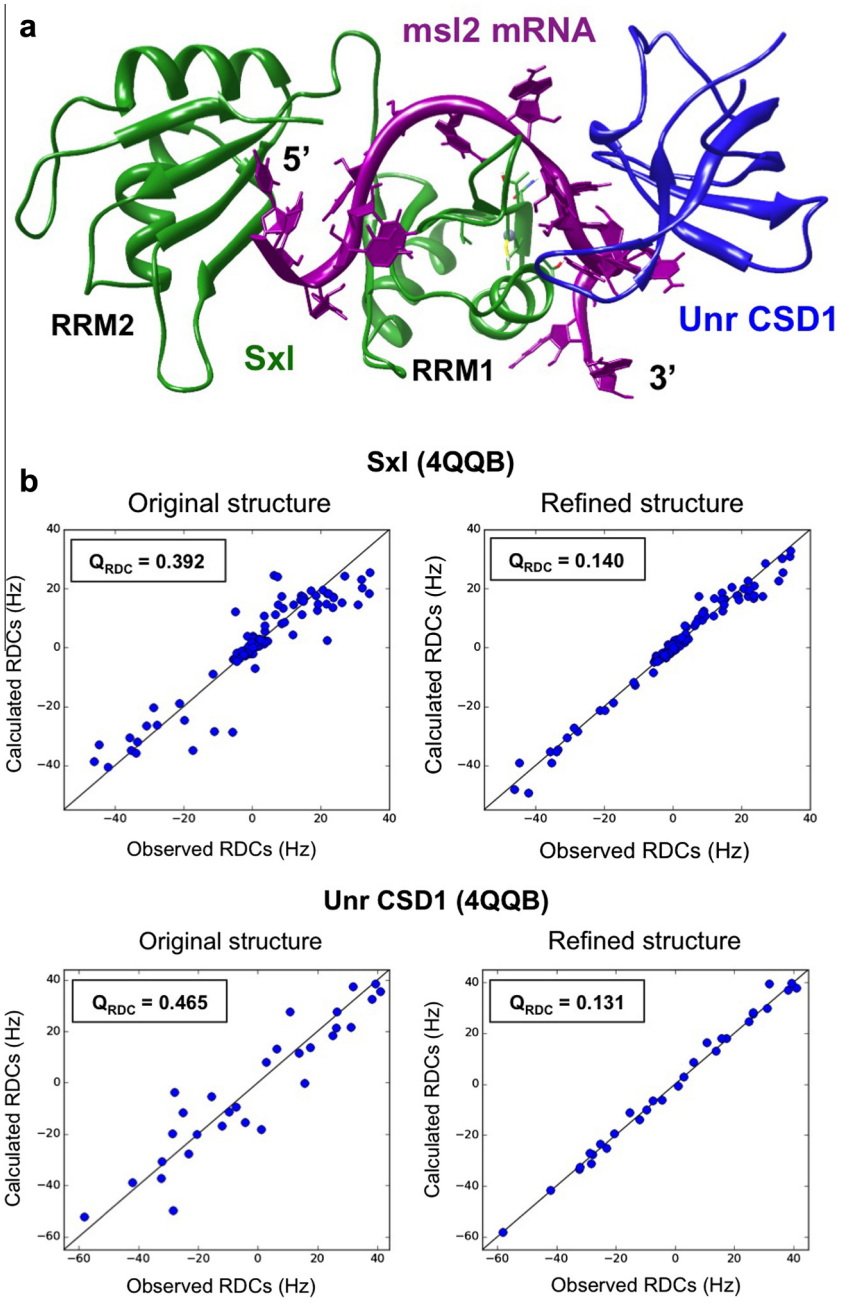


Fig. 12. (a) Structure of the ternary Sxl-Unr-*msl2*-mRNA regulatory complex; (b) correlation plot of the RDCs calculated for chains A and X of the 4QQB structure before and after refinement.

Scheme 1 was therefore applied so as to try to understand the origin of the discrepancies between the NMR data and the crystal structures. A REFMAC-NMR refinement of the individual CaM domains was first performed using both PCSs and RDCs for the N-terminal domain (the one bearing the metal ion), and RDCs only for the C-terminal domain. Note that the exclusive use of RDCs permits one to avoid any kind of assumption regarding the presence of interdomain mobility. Since the comparison of tensor magnitudes and alignments revealed the absence of extensive mobility or conformational rearrangement, a refinement of the full-length crystallographic structure was performed using both PCSs and RDCs. The refinement succeeded in reaching low *Q*-factors without increasing the free *R*-factor, thus providing a structure in agreement with both X-ray and NMR data (Fig. 9) (unpublished observations from the authors' laboratory).

This result should be compared with a previously performed analysis [144] where the initial discrepancies were explained in terms of mobility, so that the NMR data could be reproduced using a conformational ensemble. A careful assessment of the accuracy of the back-calculated tensors, and of their differences, outside the experimental error, is thus needed to reveal interdomain mobility.

3.2.4. Calmodulin–DAPK peptide complex (CaM–DAPKp)

Death-associated protein kinase (DAPK) provides a model for calcium-bound calmodulin (CaM)-dependent protein kinases (CaMKs). When DAPK is in a complex with CaM, the DAPK autoregulatory domain forms a seven-turn helix, through which it interacts with the DAPK catalytic domain. When bound to DAPK, CaM adopts a more extended conformation [191] than found in the crystal structures of CaM–CaMK peptide complexes.

In the case of CaM in complex with the DAPK peptide (DAPKp), neither the crystal structure of the complex (PDB code: 1YR5) [92] nor that with the full-length DAPK (PDB code: 2X0G) [191] exhibited satisfactory agreement with the NMR data. REFMAC-NMR refinement calculations could not provide any good agreement with PCSs and RDCs [94], even excluding residues affected by significant mobility (identified by relaxation experiments). Exclusion of the first helix of the N-terminal domain of CaM drastically reduced the disagreement, but not sufficiently to obtain an acceptable result. This indicates that in this complex the arrangement of the two CaM domains is globally different in solution compared to the crystal. Each CaM domain in the CaM–DAPKp complex could actually be refined using Xplor-NIH and NMR data only, by restraining the backbone dihedral angles to be as close as possible to those of the X-ray structure. In this case the REFMAC-NMR refinement would not be possible even for the single domains because of differences in the inter-helical orientations (see Section 3.1.4). The two domains could subsequently be positioned with respect to one another in order to superimpose the corresponding anisotropy tensors [92]. Interestingly, the final model shows a relative position of the two domains that is intermediate between the positions in the crystal structures determined for the full-length DAPK and the DAPK peptide (Fig. 10).

A more accurate analysis of the data reveals that the anisotropy tensors determined for the three metals from the C-terminal domain are marginally (a few percent) smaller than their N-terminal domain counterparts. The system was therefore subjected to an additional analysis using MaxOcc and its latest extension MaxOR (Maximum Occurrence of Regions) [151,192]. The previously refined structure was shown to have a MaxOcc of 92%, thus implying that other conformations, although with small weight, must be considered as contributors to a structural ensemble for a best fit of the data. The MaxOR analysis allowed us to identify the most compact structural ensemble (composed of rigid domain conformations) which can fully explain the experimental observables [151] (Fig. 11).

3.2.5. Sxl-Unr translation complex

An example of a system composed of multiple domains that has been successfully refined with REFMAC-NMR is provided by the ternary Sxl-Unr-*msl2*-mRNA regulatory complex, which consists of both RNA recognition motifs (RRMs) of Sxl, the first of five cold shock domains of Unr (CSD1), and an 18-mer single-stranded RNA derived from *msl2*-mRNA. Assembly of this complex is vital for female viability in fruit flies, as repression of *msl2*-mRNA by Sex-lethal (Sxl) and Upstream-of-N-Ras (Unr) prevents the formation of the dosage compensation complex, resulting in normal transcription of X-linked genes.

The crystal structure of the complex (PDB code: 4QQB) showed a modest agreement with the RDCs induced by Pf1 phage alignment medium [193]. This modest agreement is probably consistent with the extent of structural noise expected for a crystal structure at 2.8 Å resolution. The refinement revealed an almost invariant arrangement of the two domains constituting Sxl, and of the entire complex, with a relevant decrease in Q_{RDC} from 0.440 in the crystal structure to 0.144 in the refined structure. Moreover, the reduction in the structural noise of the Sxl-Unr complex by REFMAC-NMR allowed the identification of putatively mobile residues, and of minor discrepancies due to packing forces experienced in the crystal lattice [194] (Fig. 12).

Notably, the REFMAC-NMR refinement produced an effective improvement of the structural model with NMR data. As a proof that the major contribution to the improvement was not due to in-plane or out-of-plane distortions of the H^N–N bonds (even if within the standard limits), the protons were removed from the refined structure and added back using automatic methods available from common software (Molprobity [195]). Evaluation of the "reprotonated" structure showed that the agreement with NMR data was clearly maintained, whereas adding protons with the same program to the original structure did not provide any improvement.

4. Conclusions

The results of structural approaches based on the use of individual techniques such as X-ray and NMR have long provided the basis for the understanding of complex macromolecules. However, they increasingly appear somewhat limited; the results summarized here prove that an integrated approach based on the combination of data from these two radically different sources not only yields structures with improved quality (i.e., which agree with both sets of data), but also allows for a deeper understanding of the behaviour of the biomolecules in solution, revealing changes that may occur upon crystallization.

Acknowledgements

This work has been supported by Ente Cassa di Risparmio di Firenze, MIUR PRIN 2012SK7ASN, European Commission projects BioMedBridges No. 284209, pNMR No. 317127, and Instruct, part of the European Strategy Forum on Research Infrastructures (ESFRI) and supported by national member subscriptions. Specifically, we thank the EU ESFRI Instruct Core Centre CERM, Italy. GNM was funded by the Medical Research Council (grant No. MC_US_A025_0102).

References

[1] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, J.L. Milne, S. Subramaniam, Science 348 (2015) 1147–1151.
[2] R.J. Kurland, B.R. McGarvey, J. Magn. Reson. 2 (1970) 286–301.
[3] J.A.B. Lohman, C. Maclean, Chem. Phys. 35 (1978) 269–274.
[4] J.R. Tolman, J.M. Flanagan, M.A. Kennedy, J.H. Prestegard, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 9279–9283.
[5] L. Banci, I. Bertini, K.L. Bren, M.A. Cremonini, H.B. Gray, C. Luchinat, P. Turano, J. Biol. Inorg. Chem. 1 (1996) 117–126.
[6] L. Banci, I. Bertini, G. Gori Savellini, A. Romagnoli, P. Turano, M.A. Cremonini, C. Luchinat, H.B. Gray, Proteins Struct. Funct. Genet. 29 (1997) 68–76.
[7] M. Allegrozzi, I. Bertini, M.B.L. Janik, Y.-M. Lee, G. Liu, C. Luchinat, J. Am. Chem. Soc. 122 (2000) 4154–4161.
[8] I. Bertini, M.B.L. Janik, G. Liu, C. Luchinat, A. Rosato, J. Magn. Reson. 148 (2001) 23–30.
[9] I. Bertini, M.B.L. Janik, Y.-M. Lee, C. Luchinat, A. Rosato, J. Am. Chem. Soc. 123 (2001) 4181–4188.
[10] I. Bertini, A. Donaire, B. Jiménez, C. Luchinat, G. Parigi, M. Piccioli, L. Poggi, J. Biomol. NMR 21 (2001) 85–98.
[11] I. Bertini, I. Gelis, N. Katsaros, C. Luchinat, A. Provenzani, Biochemistry 42 (2003) 8011–8021.
[12] I. Bertini, M. Fragai, Y.-M. Lee, C. Luchinat, B. Terni, Angew. Chem. Int. Ed. 43 (2004) 2254–2256.
[13] S. Balayssac, I. Bertini, A. Bhaumik, M. Lelli, C. Luchinat, Proc. Natl. Acad. Sci. U.S.A. 105 (2008) 17284–17289.
[14] I. Bertini, A. Bhaumik, G. De Paepe, R.G. Griffin, M. Lelli, J.R. Lewandowski, C. Luchinat, J. Am. Chem. Soc. 132 (2010) 1032–1040.
[15] C. Luchinat, G. Parigi, E. Ravera, M. Rinaldelli, J. Am. Chem. Soc. 134 (2012) 5006–5009.
[16] A. Bhaumik, C. Luchinat, G. Parigi, E. Ravera, M. Rinaldelli, Cryst. Eng. Commun. 15 (2013) 8639–8656.
[17] K. Barthelmes, A.M. Reynolds, E. Peisach, H.R.A. Jonker, N.J. DeNunzio, K.N. Allen, B. Imperiali, H. Schwalbe, J. Am. Chem. Soc. 133 (2011) 808–819.
[18] J. Wöhnert, K.J. Franz, M. Nitz, B. Imperiali, H. Schwalbe, J. Am. Chem. Soc. 125 (2003) 13338–13339.
[19] F. Rodriguez-Castañeda, P. Haberz, A. Leonov, C. Griesinger, Magn. Reson. Chem. 44 (2006) S10–S16.
[20] X.C. Su, T. Huber, N.E. Dixon, G. Otting, ChemBioChem 7 (2006) 1599–1604.
[21] M. John, G. Otting, ChemPhysChem 8 (2007) 2309–2313.
[22] G. Pintacuda, M. John, X.C. Su, G. Otting, Acc. Chem. Res. 40 (2007) 206–212.
[23] T. Zhuang, H.S. Lee, B. Imperiali, J.H. Prestegard, Protein Sci. 17 (2008) 1220–1231.
[24] X.C. Su, K. McAndrew, T. Huber, G. Otting, J. Am. Chem. Soc. 130 (2008) 1681–1687.

[25] X.C. Su, B. Man, S. Beeren, H. Liang, S. Simonsen, C. Schmitz, T. Huber, B.A. Messerle, G. Otting, J. Am. Chem. Soc. 130 (2008) 10486–10487.
[26] P.H.J. Keizers, A. Saragliadis, Y. Hiruma, M. Overhand, M. Ubbink, J. Am. Chem. Soc. 130 (2008) 14802–14812.
[27] D. Häussinger, J. Huang, S. Grzesiek, J. Am. Chem. Soc. 131 (2009) 14761–14767.
[28] X.C. Su, G. Otting, J. Biomol. NMR 46 (2010) 101–112.
[29] B. Man, X.C. Su, H. Liang, S. Simonsen, T. Huber, B.A. Messerle, G. Otting, Chem. Eur. J. 16 (2010) 3827–3832.
[30] M.A.S. Hass, P.H.J. Keizers, A. Blok, Y. Hiruma, M. Ubbink, J. Am. Chem. Soc. 132 (2010) 9952–9953.
[31] S. Das Gupta, X. Hu, P.H.J. Keizers, W.-M. Liu, C. Luchinat, M. Nagulapalli, M. Overhand, G. Parigi, L. Sgheri, M. Ubbink, J. Biomol. NMR 51 (2011) 253–263.
[32] T. Saio, K. Ogura, K. Shimizu, M. Yokochi, T.R. Burke Jr., F. Inagaki, J. Biomol. NMR 51 (2011) 395–408.
[33] J.D. Swarbrick, P. Ung, X.C. Su, A. Maleckis, S. Chhabra, T. Huber, G. Otting, B. Graham, Chem. Commun. (Camb.) 47 (2011) 7368–7370.
[34] J.D. Swarbrick, P. Ung, S. Chhabra, B. Graham, Angew. Chem. Int. Ed. Engl. 50 (2011) 4403–4406.
[35] I. Bertini, V. Calderone, L. Cerofolini, M. Fragai, C.F.G.C. Geraldes, P. Hermann, C. Luchinat, G. Parigi, J.M.C. Teixeira, FEBS Lett. 586 (2012) 557–567.
[36] W.M. Liu, P.H. Keizers, M.A. Hass, A. Blok, M. Timmer, A.J. Sarris, M. Overhand, M. Ubbink, J. Am. Chem. Soc. 134 (2012) 17306–17313.
[37] Y. Kobashigawa, T. Saio, M. Ushio, M. Sekiguchi, M. Yokochi, K. Ogura, F. Inagaki, J. Biomol. NMR 53 (2012) 53–63.
[38] L. Cerofolini, G.B. Fields, M. Fragai, C.F.G.C. Geraldes, C. Luchinat, G. Parigi, E. Ravera, D.I. Svergun, J.M.C. Teixeira, J. Biol. Chem. 288 (2013) 30659–30671.
[39] H. Yagi, A. Maleckis, G. Otting, J. Biomol. NMR 55 (2013) 157–166.
[40] K.L. Gempf, S.J. Butler, A.M. Funk, D. Parker, Chem. Commun. (Camb.) 49 (2013) 9104–9106.
[41] C.T. Loh, K. Ozawa, K.L. Tuck, N. Barlow, T. Huber, G. Otting, B. Graham, Bioconjug. Chem. 24 (2013) 260–268.
[42] A. Bax, N. Tjandra, Nat. Struct. Biol. 4 (1997) 254–256.
[43] A. Bax, N. Tjandra, J. Biomol. NMR 10 (1997) 289–292.
[44] N. Tjandra, J.G. Omichinski, A.M. Gronenborn, G.M. Clore, A. Bax, Nat. Struct. Biol. 4 (1997) 732–738.
[45] N. Tjandra, A. Bax, Science 278 (1997) 1111–1114.
[46] M.R. Hansen, L. Mueller, A. Pardi, Nat. Struct. Biol. 5 (1998) 1065–1074.
[47] J.A. Losonczi, J.H. Prestegard, J. Biomol. NMR 12 (1998) 447–451.
[48] P.S. Prosser, J.A. Losonczi, I.V. Shiyanovskaya, J. Am. Chem. Soc. 120 (1998) 11010–11011.
[49] B.E. Ramirez, A. Bax, J. Am. Chem. Soc. 120 (1998) 9106–9107.
[50] H. Wang, M. Eberstadt, E.T. Olejniczak, R.P. Meadows, S.W. Fesik, J. Biomol. NMR 12 (1998) 443–446.
[51] H.M. Al-Hashimi, H. Valafar, M. Terrell, E.R. Zartler, M.K. Eidsness, J.H. Prestegard, J. Magn. Reson. 143 (2000) 402–406.
[52] J.H. Prestegard, H.M. Al-Hashimi, J.R. Tolman, Q. Rev. Biophys. 33 (2000) 371–424.
[53] M. Zweckstetter, A. Bax, J. Biomol. NMR 20 (2001) 365–377.
[54] N.A. Lakomek, K.F. Walter, C. Fares, O.F. Lange, B.L. de Groot, H. Grubmuller, R. Bruschweiler, A. Munk, S. Becker, J. Meiler, C. Griesinger, J. Biomol. NMR 41 (2008) 139–155.
[55] Q. Zhang, R. Throolin, S.W. Pitt, A. Serganov, H.M. Al Hashimi, J. Am. Chem. Soc. 125 (2003) 10530–10531.
[56] M.P. Latham, P. Hanson, D.J. Brown, A. Pardi, J. Biomol. NMR 40 (2008) 83–94.
[57] E. Ravera, L. Salmon, M. Fragai, G. Parigi, H.M. Al-Hashimi, C. Luchinat, Acc. Chem. Res. 47 (2014) 3118–3126.
[58] L. Yao, J. Ying, A. Bax, J. Biomol. NMR 43 (2009) 161–170.
[59] M. Assfalg, I. Bertini, P. Turano, A.G. Mauk, J.R. Winkler, B.H. Gray, Biophys. J. 84 (2003) 3917–3923.
[60] I. Bertini, I.C. Felli, C. Luchinat, J. Magn. Reson. 134 (1998) 360–364.
[61] I. Bertini, C. Luchinat, G. Parigi, Progr. NMR Spectrosc. 40 (2002) 249–273.
[62] M. John, A.Y. Park, G. Pintacuda, N.E. Dixon, G. Otting, J. Am. Chem. Soc. 127 (2005) 17190–17191.
[63] G. Cornilescu, A. Bax, J. Am. Chem. Soc. 122 (2000) 10143–10154.
[64] M. Gochin, H. Roder, Protein Sci. 4 (1995) 296–305.
[65] L. Banci, I. Bertini, J.G. Huber, C. Luchinat, A. Rosato, J. Am. Chem. Soc. 120 (1998) 12903–12909.
[66] V. Gaponenko, S.P. Sarma, A.S. Altieri, D.A. Horita, J. Li, R.A. Byrd, J. Biomol. NMR 28 (2004) 205–212.
[67] I. Diaz-Moreno, A. Diaz-Quintana, M.A. De la Rosa, M. Ubbink, J. Biol. Chem. 280 (2005) 18908–18915.
[68] M.R. Jensen, D.F. Hansen, U. Ayna, R. Dagil, M.A. Hass, H.E. Christensen, J.J. Led, Magn. Reson. Chem. 44 (2006) 294–301.
[69] C. Schmitz, R. Vernon, G. Otting, D. Baker, T. Huber, J. Mol. Biol. 416 (2012) 668–677.
[70] H. Yagi, K.B. Pilla, A. Maleckis, B. Graham, T. Huber, G. Otting, Structure 21 (2013) 883–890.
[71] L. Banci, I. Bertini, M.A. Cremonini, G. Gori Savellini, C. Luchinat, K. Wüthrich, P. Güntert, J. Biomol. NMR 12 (1998) 553–557.
[72] L. Banci, I. Bertini, G. Cavallaro, A. Giachetti, C. Luchinat, G. Parigi, J. Biomol. NMR 28 (2004) 249–261.
[73] C. Schmitz, A.M. Bonvin, J. Biomol. NMR 50 (2011) 263–266.
[74] P. Güntert, Methods Mol. Biol. 278 (2004) 353–378.
[75] C.D. Schwieters, J. Kuszewski, N. Tjandra, G.M. Clore, J. Magn. Reson. 160 (2003) 65–73.
[76] M. Rinaldelli, A. Carlon, E. Ravera, G. Parigi, C. Luchinat, J. Biomol. NMR 61 (2015) 21–34.
[77] G.M. Clore, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 9021–9025.
[78] P. Dosset, J.C. Hus, D. Marion, M. Blackledge, J. Biomol. NMR 20 (2001) 223–231.
[79] H. Valafar, J.H. Prestegard, J. Magn. Reson. 167 (2004) 228–241.
[80] G. Pintacuda, A.Y. Park, M.A. Keniry, N.E. Dixonj, G. Otting, J. Am. Chem. Soc. 128 (2006) 3696–3702.
[81] R. Hulsker, M.V. Baranova, G.S. Bullerjahn, M. Ubbink, J. Am. Chem. Soc. 130 (2008) 1985–1991.
[82] B. Simon, T. Madl, C.D. Mackereth, M. Nilges, M. Sattler, Angew. Chem. Int. Ed. 49 (2010) 1967–1970.
[83] I. Bertini, C. Luchinat, G. Parigi, Coord. Chem. Rev. 255 (2011) 649–663.
[84] M. Longinetti, G. Parigi, L. Sgheri, J. Phys. A: Math. Gen. 35 (2002) 8153–8169.
[85] W. Andralojc, K. Berlin, D. Fushman, C. Luchinat, G. Parigi, E. Ravera, L. Sgheri, J. Biomol. NMR 62 (2015) 353–371.
[86] J.J. Chou, S. Li, A. Bax, J. Biomol. NMR 18 (2000) 217–227.
[87] N.R. Skrynnikov, N.K. Goto, D. Yang, W.-Y. Choy, J.R. Tolman, G.A. Mueller, L.E. Kay, J. Mol. Biol. 295 (2000) 1265–1273.
[88] F. Tian, H. Valafar, J.H. Prestegard, J. Am. Chem. Soc. 123 (2001) 11791–11796.
[89] J.J. Chou, D.A. Case, A. Bax, J. Am. Chem. Soc. 125 (2003) 8959–8966.
[90] T.S. Ulmer, B.E. Ramirez, F. Delaglio, A. Bax, J. Am. Chem. Soc. 125 (2003) 9179–9191.
[91] J.H. Prestegard, K.L. Mayer, H. Valafar, G.C. Benison, Methods Enzymol. 394 (2005) 175–209.
[92] I. Bertini, P. Kursula, C. Luchinat, G. Parigi, J. Vahokoski, M. Wilmanns, J. Yuan, J. Am. Chem. Soc. 131 (2009) 5134–5144.
[93] D. Gottstein, D.K. Kirchner, P. Güntert, J. Biomol. NMR 52 (2012) 351–364.
[94] M. Rinaldelli, E. Ravera, V. Calderone, G. Parigi, G.N. Murshudov, C. Luchinat, Acta Crystallogr. D D70 (2014) 958–967.
[95] A. Arseniev, P. Schultze, E. Wörgötter, W. Braun, G. Wagner, M. Vasak, J.H. Kägi, K. Wüthrich, J. Mol. Biol. 201 (1988) 637–657.
[96] B. Shaanan, A.M. Gronenborn, G.H. Cohen, G.L. Gilliland, B. Veerapandian, D.R. Davies, G.M. Clore, Science 257 (1992) 961–964.
[97] C.A. Schiffer, R. Huber, K. Wüthrich, W.F. Van Gunsteren, J. Mol. Biol. 241 (1994) 588–599.
[98] M. Miller, J. Lubkowski, J.K.M. Rao, A.T. Danishefsky, J.G. Omichinski, K. Sakaguchi, H. Sakamoto, E. Appella, A.M. Gronenborn, G.M. Clore, FEBS Lett. 399 (1996) 166–170.
[99] M.L. Raves, J.F. Doreleijers, H. Vis, C.E. Vorgias, K.S. Wilson, R. Kaptein, J. Biomol. NMR 21 (2001) 235–248.
[100] M. Tang, L.J. Sperling, D.A. Berthold, C.D. Schwieters, A.E. Nesbitt, A.J. Nieuwkoop, R.B. Gennis, C.M. Rienstra, J. Biomol. NMR 51 (2011) 227–233.
[101] D.W. Hoffman, C.S. Cameron, C. Davies, S.W. White, V. Ramakrishnan, J. Mol. Biol. 264 (1996) 1058–1071.
[102] J.A. Chao, J.R. Williamson, Structure 12 (2004) 1165–1176.
[103] O.F. Lange, N.-A. Lakomek, C. Farès, G.F. Schröder, K.F.A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, B.L. de Groot, Science 320 (2008) 1471–1475.
[104] G.M. Clore, C.D. Schwieters, J. Am. Chem. Soc. 126 (2004) 2923–2938.
[105] K. Lindorff-Larsen, R.B. Best, M.A. DePristo, C.M. Dobson, M. Vendruscolo, Nature 433 (2005) 128–132.
[106] A.S. Maltsev, A. Grishaev, J. Roche, M. Zasloff, A. Bax, J. Am. Chem. Soc. 136 (2014) 3752–3755.
[107] M. Zweckstetter, A. Bax, J. Biomol. NMR 23 (2002) 127–137.
[108] S. Vijay-Kumar, C.E. Bugg, W.J. Cook, J. Mol. Biol. 194 (1987) 531–544.
[109] M. Ottiger, F. Delaglio, A. Bax, J. Magn. Reson. 131 (1998) 373–378.
[110] J.P. Derrick, D.B. Wigley, J. Mol. Biol. 243 (1994) 906–918.
[111] D.S. Berkholz, M.V. Shapovalov, R.L. Dunbrack Jr., P.A. Karplus, Structure 17 (2009) 1316–1325.
[112] M.W. MacArthur, J.M. Thornton, J. Mol. Biol. 264 (1996) 1180–1195.
[113] L. Esposito, S.A. De, A. Zagari, L. Vitagliano, J. Mol. Biol. 347 (2005) 483–487.
[114] D.S. Berkholz, C.M. Driggers, M.V. Shapovalov, R.L. Dunbrack Jr., P.A. Karplus, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) 449–453.
[115] G.N. Murshudov, P. Skubàk, A.A. Lebedev, N.S. Pannu, R.A. Steiner, R.A. Nicholls, M.D. Winn, F. Long, A.A. Vagin, Acta Crystallogr. D 67 (2011) 355–367.
[116] M.D. Winn, C.C. Ballard, K.D. Cowtan, E.J. Dodson, P. Emsley, P.R. Evans, R.M. Keegan, E.B. Krissinel, A.G.W. Leslie, A. McCoy, S.J. McNicholas, G.N. Murshudov, N.S. Pannu, E.A. Potterton, H.R. Powell, R.J. Read, A.A. Vagin, K. S. Wilson, Acta Crystallogr. D 67 (2011) 235–242.
[117] H.M. McConnell, R.E. Robertson, J. Chem. Phys. 29 (1958) 1361–1365.
[118] M.D. Kemple, B.D. Ray, K.B. Lipkowitz, F.G. Prendergast, B.D.N. Rao, J. Am. Chem. Soc. 110 (1988) 8275–8287.
[119] M. Zweckstetter, Nat. Protoc. 3 (2008) 679–690.
[120] C. Camilloni, M. Vendruscolo, J. Phys. Chem. B 119 (2015) 653–661.
[121] D. Shishmarev, G. Otting, J. Biomol. NMR 56 (2013) 203–216.
[122] A.M. Bonvin, A.T. Brunger, J. Biomol. NMR 7 (1996) 72–76.
[123] W.-Y. Choy, J.D. Forman-Kay, J. Mol. Biol. 308 (2001) 1011–1032.
[124] D.I. Svergun, M.V. Petoukhov, M.H.J. Koch, Biophys. J. 80 (2001) 2946–2953.
[125] R. Burgi, J. Pitera, W.F. Van Gunsteren, J. Biomol. NMR 19 (2001) 305–320.
[126] J. Iwahara, C.D. Schwieters, G.M. Clore, J. Am. Chem. Soc. 126 (2004) 5879–5896.
[127] I. Bertini, C. Del Bianco, I. Gelis, N. Katsaros, C. Luchinat, G. Parigi, M. Peana, A. Provenzani, M.A. Zoroddu, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 6841–6846.
[128] M. Blackledge, Progr. NMR Spectrosc. 46 (2005) 23–61.

[129] M. Fragai, C. Luchinat, G. Parigi, Acc. Chem. Res. 39 (2006) 909–917.
[130] D.D. Boehr, R. Nussinov, P.E. Wright, Nat. Chem. Biol. 5 (2009) 954.
[131] Y.E. Ryabov, D. Fushman, Magn. Reson. Chem. 44 (2006) S143–S151.
[132] Y. Chen, S.L. Campbell, N.V. Dokholyan, Biophys. J. 93 (2007) 2300–2306.
[133] P. Bernadò, E. Mylonas, M.V. Petoukhov, M. Blackledge, D.I. Svergun, J. Am. Chem. Soc. 129 (2007) 5656–5664.
[134] I. Bertini, Y.K. Gupta, C. Luchinat, G. Parigi, M. Peana, L. Sgheri, J. Yuan, J. Am. Chem. Soc. 129 (2007) 12786–12794.
[135] L. Nodet, L. Salmon, V. Ozenne, S. Meier, M.R. Jensen, M. Blackledge, J. Am. Chem. Soc. 131 (2009) 17908–17918.
[136] A.C. Stelzer, A.T. Frank, M.H. Bailor, I. Andricioaei, H.M. Al Hashimi, Methods 49 (2009) 167–173.
[137] J. Huang, S. Grzesiek, J. Am. Chem. Soc. 132 (2010) 694–705.
[138] C.K. Fisher, A. Huang, C.M. Stultz, J. Am. Chem. Soc. 132 (2010) 14919–14927.
[139] Q. Bashir, A.N. Volkov, G.M. Ullmann, M. Ubbink, J. Am. Chem. Soc. 132 (2010) 241–247.
[140] I. Bertini, A. Giachetti, C. Luchinat, G. Parigi, M.V. Petoukhov, R. Pierattelli, E. Ravera, D.I. Svergun, J. Am. Chem. Soc. 132 (2010) 13553–13558.
[141] J. Rinnenthal, J. Buck, J. Ferner, A. Wacker, B. Furtig, H. Schwalbe, Acc. Chem. Res. 44 (2011) 1292–1301.
[142] J.R. Bothe, E.N. Nikolova, C.D. Eichhorn, J. Chugh, A.L. Hansen, H.M. Al Hashimi, Nat. Methods 8 (2011) 919–931.
[143] K. Berlin, C.A. Castañeda, D. Schneidman-Dohovny, A. Sali, A. Nava-Tudela, D. Fushman, J. Am. Chem. Soc. 135 (2013) 16595–16609.
[144] L. Russo, M. Maestre-Martinez, S. Wolff, S. Becker, C. Griesinger, J. Am. Chem. Soc. 135 (2013) 17111–17120.
[145] P. Guerry, L. Salmon, L. Mollica, J.L. Ortega Roldan, P. Markwick, N.A. van Nuland, J.A. McCammon, M. Blackledge, Angew. Chem. Int. Ed. Engl. 52 (2013) 3181–3185.
[146] L. Salmon, G. Bascom, I. Andricioaei, H.M. Al Hashimi, J. Am. Chem. Soc. 135 (2013) 5457–5466.
[147] A. Cavalli, C. Camilloni, M. Vendruscolo, J. Chem. Phys. 138 (2013) 094112.
[148] P. Kukic, C. Camilloni, A. Cavalli, M. Vendruscolo, J. Mol. Biol. 426 (2014) 1826–1838.
[149] M. Sanchez-Martinez, R. Crehuet, Phys. Chem. Chem. Phys. 16 (2014) 26030–26039.
[150] W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, PLoS Comput. Biol. 10 (2014) e1003406.
[151] W. Andralojc, C. Luchinat, G. Parigi, E. Ravera, J. Phys. Chem. B 118 (2014) 10576–10587.
[152] D.A. Torchia, Prog. Nucl. Magn. Reson. Spectrosc. 84–85 (2015) 14–32.
[153] J.R. Tolman, H.M. Al-Hashimi, L.E. Kay, J.H. Prestegard, J. Am. Chem. Soc. 123 (2001) 1416–1424.
[154] Q. Zhang, H.M. Al-Hashimi, Nat. Methods 5 (2008) 243–245.
[155] M. Zweckstetter, A. Bax, J. Am. Chem. Soc. 122 (2000) 3791–3792.
[156] I. Bertini, M. Fragai, C. Luchinat, Curr. Pharm. Des. 15 (2009) 3592–3605.
[157] W. Bode, K. Maskos, Biol. Chem. 384 (2003) 863–872.
[158] G. Goldstein, M. Scheid, U. Hammerling, D.H. Schlesinger, H.D. Niall, E.A. Boyse, Proc. Natl. Acad. Sci. U.S.A. 72 (1975) 11–15.
[159] K.D. Wilkinson, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 15280–15282.
[160] D.C. Schwartz, M. Hochstrasser, Trends Biochem. Sci. 28 (2003) 321–328.
[161] D. Komander, M. Rape, Annu. Rev. Biochem. 81 (2012) 203–229.
[162] D. Fushman, K.D. Wilkinson, F1000 Biol. Rep. 3 (2011) 26.
[163] A. Hershko, A. Ciechanover, H. Heller, A.L. Haas, I.A. Rose, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1783–1786.
[164] A. Ciechanover, H. Heller, S. Elias, A.L. Haas, A. Hershko, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1365–1368.
[165] G.C. Stone, U. Sjobring, L. Bjorck, J. Sjoquist, C.V. Barber, F.A. Nardella, J. Immunol. 143 (1989) 565–570.
[166] A.M. Gronenborn, G.M. Clore, J. Mol. Biol. 233 (1993) 331–335.
[167] J.P. Derrick, D.B. Wigley, Nature 359 (1992) 752–754.
[168] F. Li, J.H. Lee, A. Grishaev, J. Ying, A. Bax, ChemPhysChem 16 (2015) 572–578.
[169] F.C. Stevens, Can. J. Biochem. Cell Biol. 61 (1983) 906–910.
[170] J.J. Chou, S. Li, C.B. Klee, A. Bax, Nat. Struct. Biol. 8 (2001) 990–997.
[171] H. Birkedal-Hansen, Curr. Opin. Cell Biol. 7 (1995) 728–735.
[172] G.J. Murphy, G. Murphy, J.J. Reynolds, FEBS Lett. 89 (1991) 4–7.
[173] W.G. Stetler-Stevenson, R. Hewitt, Sem. Cancer Biol. 7 (1996) 147–154.
[174] N. Borkakoti, J. Mol. Med. 78 (2000) 261–268.
[175] L. Steinman, Cell 85 (1996) 299–302.
[176] L.M. Coussens, B. Fingleton, L.M. Matrisian, Science 295 (2002) 2387–2392.
[177] M. Pavlaki, S. Zucker, Cancer Metastasis Rev. 22 (2003) 177–203.
[178] V. Lukacova, Y.F. Zhang, M. Mackov, P. Baricic, S. Raha, J.A. Calvo, S. Balaz, J. Biol. Chem. 279 (2004) 14194–14200.
[179] N. Borkakoti, Biochem. Soc. Trans. 32 (2004) 17–20.
[180] I. Bertini, V. Calderone, M. Cosenza, M. Fragai, Y.-M. Lee, C. Luchinat, S. Mangani, B. Terni, P. Turano, Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 5334–5339.
[181] K. Maskos, W. Bode, Mol. Biotechnol. 25 (2003) 241–266.
[182] M. Longinetti, C. Luchinat, G. Parigi, L. Sgheri, Inv. Probl. 22 (2006) 1485–1502.
[183] M. Fragai, C. Luchinat, G. Parigi, E. Ravera, Coord. Chem. Rev. 257 (2013) 2652–2667.
[184] I. Bertini, M. Fragai, C. Luchinat, M. Melikian, M. Toccafondi, J.L. Lauer, G.B. Fields, J. Am. Chem. Soc. 134 (2012) 2100–2110.
[185] G. Barbato, M. Ikura, L.E. Kay, R.W. Pastor, A. Bax, Biochemistry 31 (1992) 5269–5278.
[186] J.L. Baber, A. Szabo, N. Tjandra, J. Am. Chem. Soc. 123 (2001) 3953–3959.
[187] I. Bertini, L. Ferella, C. Luchinat, G. Parigi, M.V. Petoukhov, E. Ravera, A. Rosato, D.I. Svergun, J. Biomol. NMR 53 (2012) 271–280.
[188] I. Bertini, C. Luchinat, M. Nagulapalli, G. Parigi, E. Ravera, Phys. Chem. Chem. Phys. 14 (2012) 9149–9156.
[189] 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, A.G. Tzakos, Structure 20 (2012) 522–533.
[190] F. Van Petegem, F.C. Chatelain, D.L. Minor Jr., Nat. Struct. Mol. Biol. 12 (2005) 1108–1115.
[191] D. de Diego, I.J. Kuper, N. Bakalova, P. Kursula, M. Wilmanns, Sci. Signal 3 (2010) ra6.
[192] W. Andralojc, E. Ravera, L. Salmon, G. Parigi, H.M. Al-Hashimi, C. Luchinat, Phys. Chem. Chem. Phys. (2015), http://dx.doi.org/10.1039/C5CP03993B.
[193] J. Hennig, C. Militti, G.M. Popowicz, I. Wang, M. Sonntag, A. Geerlof, F. Gabel, F. Gebaur, M. Sattler, Nature 515 (2014) 287–290.
[194] A. Carlon, E. Ravera, J. Hennig, G. Parigi, M. Sattler, C. Luchinat, J. Am. Chem. Soc. (2016), http://dx.doi.org/10.1021/jacs.5b11598.
[195] V.B. Chen, W.B. Arendall III, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, D.C. Richardson, Acta Crystallogr. D: Biol. Crystallogr. 66 (2010) 12–21.

Glossary of abbreviations

A: Alignment tensor
CaM: Calmodulin
CCP4: Collaborative computational project no. 4 Software for Macromolecular Crystallography (www.ccp4.ac.uk)
DAPk: Death-associated protein kinase
GB3: IgG-binding domain 3 of protein G
IQ: IQ-recognition motif of voltage-gated calcium channel
MaxOcc: Maximum Occurrence
MaxOR: Maximum Occurrence of Regions
MMP1: Matrix-metalloproteinase 1
NOE: Nuclear Overhauser Effect
PCS: Pseudocontact shift
PDB: RCSB Protein DataBank (www.rcsb.org/pdb/home/home.do)
RDC: Residual dipolar coupling
REFMAC5: REFinement of MACromolecular structures (software for crystallography, www2.mrc-lmb.cam.ac.uk/groups/murshudov/)
SAXS: Small-angle X-ray Scattering
Sxl-Unr: Sex-lethal and Upstream-of-N-Ras
χ: Magnetic susceptibility tensor in molecular coordinate frame

This page intentionally left blank

4 Conclusions

The study of internal mobility in multidomain biological systems is a field whose importance becomes more and more appreciated and for which paramagnetic NMR appears a particularly promising tool. Sophisticated data analysis protocols are however required in order to connect the motionally averaged NMR data with the most plausible ensembles of sampled structures. The current thesis was aimed at vastly expanding the capabilities of one of such procedures: the Maximum Occurrence approach and at clearly demonstrating its capabilities of dealing not only with selected 'proof of principle' systems, but also with actual challenging cases of biological interest.

The first contribution, which opened the path to all the following research, consisted of implementing for the first time a computer program based on the idea of Maximum Occurrence of Regions. The new software uses a linear minimization algorithm in the ensemble optimization step of the calculation, allowing for both fast and robust optimization. In parallel, a new software for the calculation of Maximum Occurrences of single conformations was also developed based on the new algorithm. It demonstrated an over hundredfold decrease in computational time with respect to the original implementation of MaxOcc. We thus now dispose of tools to calculate the maximum occurrences of both single structures and conformational regions in a high-throughput manner.

The introduction of MaxOR strongly broadens the spectrum of information one can extract form motionally averaged experimental data, allowing us to go far beyond just identifying the most sampled conformations which was the domain of MaxOcc. The Maximum Occurrence of Regions approach (also dubbed Maximum Occupancy Limit of Regions) have already demonstrated its capability to provide insights into the following types of questions:

1) Quantifying the smallest amount of mobility that has to be invoked to explain the experimental

observations (as it was shown for calmodulin complex with DAPk peptide and further confirmed with simulated data)

2) Distinguishing between the sampling of a broad continuous spectrum of structures and an exchange between two distinct sites (demonstrated with various sets of simulated data)
3) Identification of sets of structures best suited to complement a major state (found for example through MaxOcc) or, in other words, identification of the simplest mobility scheme capable of explaining the available data (see the work on HIV1TAR)
4) Confirming the presence and setting bounds on the spatial extent of even very sparsely populated minor states (as was done for the cytP450cam complex with Pdx)

The capabilities of the MaxOR approach are still being explored and thus the complied list very likely does not cover all the potential of the method.

My work explored also another fundamental aspect of the study of multi-domain proteins with long range NMR observables, namely the actual type and amount of information that the different available observables provide. In this regard the paramagnetic restraints (PCS and RDC) were compared to another commonly used NMR observable: the RDC originating from the presence of an alignment medium. Extensive studies through computer simulations suggest that both kinds of experimental data are equally capable of identifying the main sampled conformations for simple motional models (motion in a cone, exchange between two sites). They both also show a similar, high robustness against experimental noise. This direct comparison thus demonstrates that both paramagnetic and diamagnetic long range restraints are equally viable tools for the study of interdomain mobility in biological systems. This study can also be seen from another angle, as a direct comparison between two conceptually different approaches to the analysis of ensemble averaged data: the MaxOcc and SES methods. In all the preformed simulations, both the methods were able to recover the main sampled states with a comparable and high accuracy which constitutes yet another confirmation of the viability of both the approaches.

The ultimate motivation behind the development of any computational tool in structural biology is to answer questions related to the behavior of real biologically important systems. Three real systems were studied with MaxOcc in the framework of my thesis – calmodulin bound to the DAPk peptide, HIV1TAR RNA element and protein complex between cytochrome P450cam and putidaredoxin. For the first of them the performed study allowed to detect and quantify residual mobility around a central structure, resembling the one known from the X-ray studies. It had also confirmed (through the application of minOR) that high MaxOcc value calculated for that structure is not just an artifact of the ensemble averaging. These results, although interesting from theoretical point of view, carry relatively few new

insights into the biology of the interaction and thus this project can be viewed more as a proof of principle of the MaxOR/minOR approach for real experimental data. For the other two systems on the other hand, the information extracted through our analysis may help to verify hypotheses about the mechanism of their function or to formulate new ones. In the case of HIV1TAR our study has shown that an exchange between two states with unequal populations is the simplest mobility scheme capable of explaining the experimental average RDCs. We have also demonstrated that within such a model the possible exact locations of the two states are quite limited. Interestingly when one analyzes the known ligand bound conformations of the system (HIV1TAR is actually an intensively studied drug target) it becomes clear that almost all of them lay on the shortest conformational path between the two proposed states. Such a result is thus a suggestion that ligand binding to HIV1TAR occurs through the mechanism of conformational selection (from the states visited during the exchange between the two postulated states) and not through an induced fit. Similar states were earlier proposed based on the analysis of a MD simulation performed for the system[27]. For the cytP450cam-Pdx complex our study has led to the identification of several lowly populated minor states, complementing the previously characterized specific complex. The repartition of the minor states correlates with the electrostatic potential map around cytP450cam suggesting that these states constitute an electrostatically driven encounter complex. Moreover based on the presence (or absence) of electrostatically favorable pathways between a given minor states and the specific interaction site one can formulate hypotheses on which of the encounter states are productive (able to evolve into the specific complex) and which ones are futile.

To conclude, my thesis have contributed to the field of the characterization of internal mobility in multidomain biological systems in two different respects. On one hand, my work provided a new, yet already thoroughly tested, computational tool to analyze motionally averaged data. On the other, specific biologically important systems were studied with the developed method, uncovering connections between their conformational heterogeneity and the mechanism of their function/inhibition. My work has also demonstrated the generality of the MaxOcc approach and its applicability to a range of systems far beyond multidomain proteins, for which it was originally devised. Concurrently, the assortment of experimental observables that our method is compatible with is also continuously growing (for example the usage of DEER data in MaxOcc is currently under implementation) leading to the possibility of using MaxOcc as an integrative structural biology tool. As a mature and proven methodology MaxOcc has thus the potential to become an important instrument for any structural biologist dealing with interdomain conformational variability.

5 Bibliography

(1) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438* (7064), 117.

(2) Huang, Y. J.; Montelione, G. T. *Nature* **2005**, *438* (7064), 36.

(3) Fragai, M.; Luchinat, C.; Parigi, G. *Acc. Chem. Res.* **2006**, *39* (12), 909.

(4) Kern, D.; Zuiderweg, E. R. *Curr. Opin. Struct. Biol.* **2003**, *13* (6), 748.

(5) Hammes, G. G. *Biochemistry* **2002**, *41* (26), 8221.

(6) Tzeng, S.-R.; Kalodimos, C. G. *Nature* **2009**, *462* (7271), 368.

(7) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433* (7022), 128.

(8) Jarymowycz, V. A.; Stone, M. J. *Chem. Rev.* **2006**, *106* (5), 1624.

(9) Valentine, E. R.; Palmer, A. G. *Biochemistry* **2005**, *44* (9), 3410.

(10) Lundström, P.; Mulder, F. A. A.; Akke, M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (47), 16984.

(11) Schanda, P.; Forge, V.; Brutscher, B. *Proc. Natl. Acad. Sci.* **2007**, *104* (27), 11257.

(12) Corazza, A.; Rennella, E.; Schanda, P.; Mimmi, M. C.; Cutuil, T.; Raimondi, S.; Giorgetti, S.; Fogolari, F.; Viglino, P.; Frydman, L.; Gal, M.; Bellotti, V.; Brutscher, B.; Esposito, G. *J. Biol. Chem.* **2010**, *285* (8), 5827.

(13) Mulder, F. A. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. *Nat. Struct. Mol. Biol.* **2001**, *8* (11), 932.

(14) Wang, L.; Pang, Y.; Holder, T.; Brender, J. R.; Kurochkin, A. V.; Zuiderweg, E. R. P. *Proc. Natl. Acad. Sci.* **2001**, *98* (14), 7684.

(15) Bhabha, G.; Ekiert, D. C.; Jennewein, M.; Zmasek, C. M.; Tuttle, L. M.; Kroon, G.; Dyson, H. J.; Godzik, A.; Wilson, I. A.; Wright, P. E. *Nat. Struct. Mol. Biol.* **2013**, *20* (11), 1243.

(16) Mittermaier, A.; Kay, L. E. *Science* **2006**, *312* (5771), 224.

(17) Palmer, A. G. *Chem. Rev.* **2004**, *104* (8), 3623.

(18) Ishima, R.; Torchia, D. A. *Nat. Struct. Mol. Biol.* **2000**, *7* (9), 740.

(19) Rosenblum, G.; Van den Steen, P. E.; Cohen, S. R.; Grossmann, J. G.; Frenkel, J.; Sertchook, R.; Slack, N.; Strange, R. W.; Opdenakker, G.; Sagi, I. *Structure* **2007**, *15* (10), 1227.

(20) Mackereth, C. D.; Madl, T.; Bonnal, S.; Simon, B.; Zanier, K.; Gasch, A.; Rybin, V.; Valcárcel, J.; Sattler, M. *Nature* **2011**, *475* (7356), 408.

(21) Diego, I. de; Kuper, J.; Bakalova, N.; Kursula, P.; Wilmanns, M. *Sci. Signal.* **2010**, *3* (106), ra6.

(22) Babu, Y. S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1988**, *204* (1), 191.

(23) Chattopadhyaya, R.; Meador, W. E.; Means, A. R.; Quiocho, F. A. *J. Mol. Biol.* **1992**, *228* (4), 1177.

(24) Barbato, G.; Ikura, M.; Kay, L. E.; Pastor, R. W.; Bax, A. *Biochemistry* **1992**, *31* (23), 5269.

(25) Kay, L. E. *Biochem. Cell Biol.* **1998**, *76* (2–3), 145.

(26) Frank, A. T.; Stelzer, A. C.; Al-Hashimi, H. M.; Andricioaei, I. *Nucleic Acids Res.* **2009**, *37* (11), 3670.

(27) Salmon, L.; Bascom, G.; Andricioaei, I.; Al-Hashimi, H. M. *J. Am. Chem. Soc.* **2013**, *135* (14), 5457.

(28) Boisbouvier, J.; Delaglio, F.; Bax, A. *Proc. Natl. Acad. Sci.* **2003**, *100* (20), 11333.
(29) Bertini, I.; Luchinat, C.; Parigi, G.; Pierattelli, R. *Dalton Trans.* **2008**, No. 29, 3782.
(30) Bertini, I.; Donaire, A.; Jim?nez, B.; Luchinat, C.; Parigi, G.; Piccioli, M.; Poggi, L. *J.Biomol.NMR* **2001**, *21*, 85.
(31) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. *J. Am. Chem. Soc.* **2010**, *132* (38), 13553.
(32) Huang, J.; Warner, L. R.; Sanchez, C.; Gabel, F.; Madl, T.; Mackereth, C. D.; Sattler, M.; Blackledge, M. *J. Am. Chem. Soc.* **2014**, *136* (19), 7068.
(33) Rieping, W.; Habeck, M.; Nilges, M. *Science* **2005**, *309* (5732), 303.
(34) Berlin, K.; Castañeda, C. A.; Schneidman-Duhovny, D.; Sali, A.; Nava-Tudela, A.; Fushman, D. *J. Am. Chem. Soc.* **2013**, *135* (44), 16595.
(35) Camilloni, C.; Vendruscolo, M. *Biochemistry* **2015**, *54* (51), 7470.
(36) Iwahara, J.; Schwieters, C. D.; Clore, G. M. *J. Am. Chem. Soc.* **2004**, *126* (18), 5879.
(37) Ravera, E.; Sgheri, L.; Parigi, G.; Luchinat, C. *Phys Chem Chem Phys* **2016**, *18* (8), 5686.
(38) Luchinat, C.; Nagulapalli, M.; Parigi, G.; Sgheri, L. *J. Magn. Reson.* **2012**, *215* (1), 85.
(39) Bernado, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129* (17), 5656.
(40) Lee, J.; Lee, D.; Park, H.; Coutsias, E. A.; Seok, C. *Proteins* **2010**, *78* (16), 3428.
(41) Volkov, A. N.; Ubbink, M.; Nuland, N. A. J. van. *J. Biomol. NMR* **2010**, *48* (4), 225.
(42) Bertini, I.; Luchinat, C.; Parigi, G.; Ravera, E. *NMR of Paramagnetic Molecules: Applications to Metallobiomolecules and Models*; Elsevier, 2016.
(43) Bertini, I.; Luchinat, C.; Parigi, G. *Prog. Nucl. Magn. Reson. Spectrosc.* **2002**, *40*, 249.
(44) Rouf, S. A.; Mareš, J.; Vaara, J. *J. Chem. Theory Comput.* **2015**, *11* (4), 1683.
(45) Bertini, I.; Bianco, C. D.; Gelis, I.; Katsaros, N.; Luchinat, C.; Parigi, G.; Peana, M.; Provenzani, A.; Zoroddu, M. A. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (18), 6841.
(46) Keizers, P. H. J.; Saragliadis, A.; Hiruma, Y.; Overhand, M.; Ubbink, M. *J. Am. Chem. Soc.* **2008**, *130* (44), 14802.
(47) Liu, W.-M.; Keizers, P. H. J.; Hass, M. A. S.; Blok, A.; Timmer, M.; Sarris, A. J. C.; Overhand, M.; Ubbink, M. *J. Am. Chem. Soc.* **2012**, *134* (41), 17306.
(48) Barthelmes, K.; Reynolds, A. M.; Peisach, E.; Jonker, H. R. A.; DeNunzio, N. J.; Allen, K. N.; Imperiali, B.; Schwalbe, H. *J. Am. Chem. Soc.* **2011**, *133* (4), 808.
(49) Barthelmes, D.; Gränz, M.; Barthelmes, K.; Allen, K. N.; Imperiali, B.; Prisner, T.; Schwalbe, H. *J. Biomol. NMR* **2015**, *63* (3), 275.
(50) Bertini, I.; Kursula, P.; Luchinat, C.; Parigi, G.; Vahokoski, J.; Wilmanns, M.; Yuan, J. *J. Am. Chem. Soc.* **2009**, *131* (14), 5134.
(51) Koehler, J.; Meiler, J. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *59* (4), 360.
(52) Longinetti, M.; Luchinat, C.; Parigi, G.; Sgheri, L. *Inverse Probl.* **2006**, *22* (4), 1485.
(53) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104* (17), 4546.
(54) Bax, A.; Grishaev, A. *Curr. Opin. Struct. Biol.* **2005**, *15* (5), 563.
(55) Zweckstetter, M. *Nat. Protoc.* **2008**, *3* (4), 679.
(56) Berlin, K.; O'Leary, D. P.; Fushman, D. *J. Magn. Reson.* **2009**, *201* (1), 25.
(57) Zhang, Q.; Stelzer, A. C.; Fisher, C. K.; Al-Hashimi, H. M. *Nature* **2007**, *450* (7173), 1263.

(58) Ravera, E.; Salmon, L.; Fragai, M.; Parigi, G.; Al-Hashimi, H.; Luchinat, C. *Acc. Chem. Res.* **2014**, *47* (10), 3118.
(59) Bertini, I.; Luchinat, C.; Nagulapalli, M.; Parigi, G.; Ravera, E. *Phys. Chem. Chem. Phys.* **2012**, *14* (25), 9149.
(60) Clore, G. M.; Szabo, A.; Bax, A.; Kay, L. E.; Driscoll, P. C.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1990**, *112* (12), 4989.
(61) Baber, J. L.; Szabo, A.; Tjandra, N. *J. Am. Chem. Soc.* **2001**, *123*, 3953.
(62) Iwahara, J.; Clore, G. M. *J. Am. Chem. Soc.* **2010**, *132* (38), 13346.
(63) Cornilescu, G.; Marquardt, J.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120* (27), 6836.

6 Acknowledgments

I would like to express my sincere gratitude to Prof Giacomo Parigi, the supervisor of my thesis, for all the support and encouragement I constantly received from him. I could always count on his advice and assistance both regarding my scientific work and everyday aspects of studying and living in Italy.
I am also very grateful to Prof Claudio Luchinat who, despite all his obligations as the director of CIRMMP, was always readily available for in-depth scientific discussions and actively followed the development of my research projects.
I would like to thank Dr Enrico Ravera who was also closely involved in my research projects, for all the useful discussions and inspiring scientific input.
I am also sincerely thankful to all my colleagues with whom I had the pleasure to work during these three years, especially to Azzurra Carlon whose friendly assistance with the computational tools she developed was invaluable for my work and to Dr Linda Cerofolini who was always willing to share her vast knowledge on NMR methods and data analysis.
Special thanks to the members of my 'lunch group', for the great atmosphere and all the useful Italian vocabulary that I learned from you :-)
I would also like to warmly thank Prof Marcellus Ubbink, Dr Gregg Siegal and all the people in their groups for making my two secondments in Leiden a time in which I could not only develop scientifically, but also which I will remember with joy.
My gratitude goes also to the members of the research groups of Prof David Fushman and Prof Hashim al-Hashimi, which whom I had the luck to collaborate on two very inspiring projects.
And finally, last but certainly not least, I would like to warmly thank my family, especially my parents, for their constant loving support.

Funding acknowledgment
The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n°317127.