PCCP

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/pccp

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

Corresponding Author:

Prof. Giacomo Parigi: parigi@cerm.unifi.it
parigi@cerm.unifi.it

Abstract

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]

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 (0.022g/mL) 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, β_h and γ_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

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

where **x** is the vector of the weights of the N structures composing the pool, **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 sub-matrices $\mathbf{A}_n$ containing back-calculated RDCs of single constructs multiplied by the appropriate scaling factors c_n:

$$\mathbf{A}(c_1,\ldots,c_K) = \begin{bmatrix} c_1\mathbf{A}_1 \\ c_2\mathbf{A}_2 \\ \vdots \\ c_k\mathbf{A}_k \end{bmatrix}.$$

The **y** vector and the **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 eq. 1 would require a non-linear minimization due to the presence of the unknown *c* factor. It becomes linear if the scaling factor *c* is fixed to a constant value. The optimal value of *c* 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 $c_{opt} = \frac{\boldsymbol{y_{calc} \cdot y}}{\boldsymbol{y \cdot y}}$. A minimization procedure was thus applied which involved making an initial guess of the value of scaling factor *c*, solving eq.1 for **x** (with fixed *c*) using a non-negative least squares method (a frugal coordinate descent algorithm[46], combined with random coordinate search[47]), then calculating the optimal *c* for the present **x** vector, and finally using it as the fixed scaling factor in the next iteration of non-negative least squares minimization, in an iterative fashion, until convergence of the *c* value 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 Eq. 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} :

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

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

Figure 1) represent the twisting of the first and second 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 (37005 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. Figures 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 $-10<\alpha_h<5°$, $45<\beta_h<55°$, $-15<\gamma_h<5°$, centered at α_h = -5°, β_h =50° and γ_h = -5°. 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 β_h =180°. 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 Fig.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 α_h =-22°, β_h =32° and γ_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 (χ^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 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 α_h =-15°, β_h =52° and γ_h =-28°) 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 S1) 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 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, 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 β_h =-40, $\alpha_h+\gamma_h$ =-15, and β_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 β_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 α_h =x, β_h =-40, γ_h =-x-15°, where x runs over all the values of α_h sterically allowed at this point of space, in steps of 5°. Figure S3 presents the MaxOR values of each cube as a function of the α_h angle. The MaxOR function has a single maximum at α_h=-15° (and γ_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 the Euler angle space, one located at the peak of the MaxOcc profile and another at α_h=-15, β_h=-40 γ_h=0° (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°·20° (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 reference [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 1, 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 figures 6c and 5. 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 figures 6c and 5) and the main state identified by MaxOcc/MaxOR, and between the SAS cluster 2 (in red in figures 6c and 5) 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 figure 6b. Also our qualitative identification of the major and minor states is in line with the relative 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 figure 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).

Figure 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.

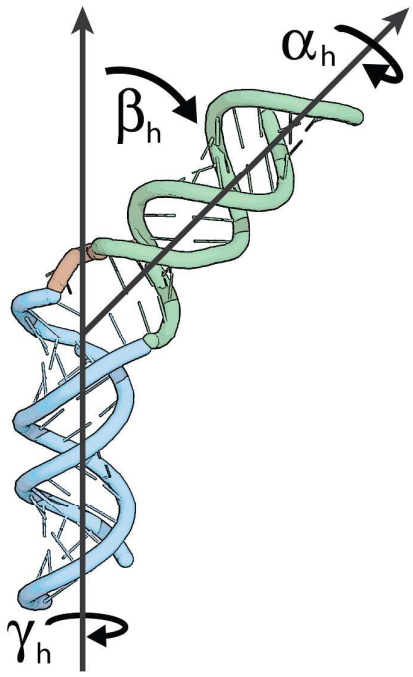


Figure 2. a-c) The MaxOcc landscape (MaxOcc values color coded) as a function of α_h , β_h and γ_h angles (2D projections). White areas correspond to not sampled regions. d) 3D representation of the full sampled space (blue) and of the area which encompasses high MaxOcc conformations (outer red surface,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

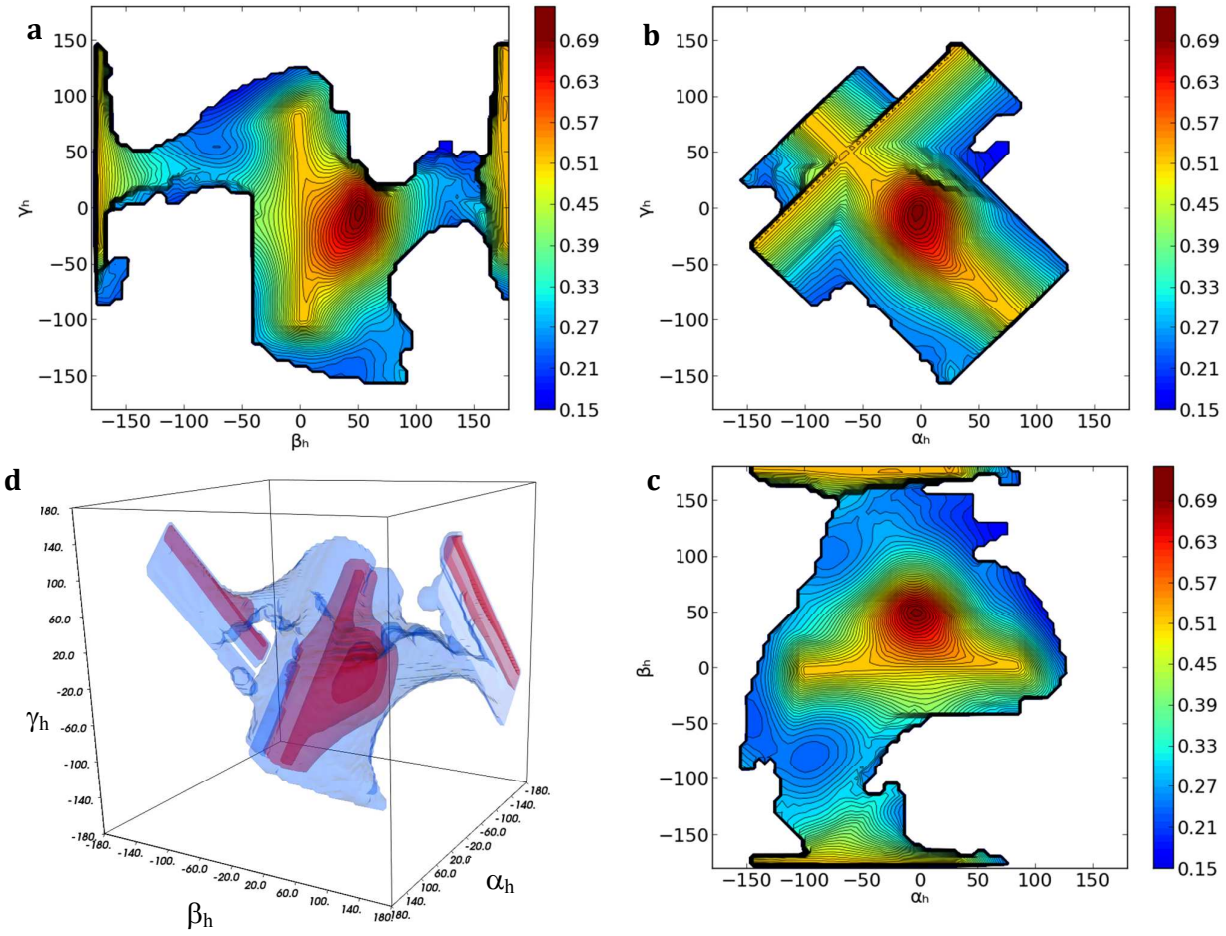


Figure 3 Superimposition of MD trajectory (dark dots) to the MaxOcc landscape (color coded) as a function of α_h , β_h and γ_h angles (2D projections and 3D representation).White areas correspond to not sampled regions.

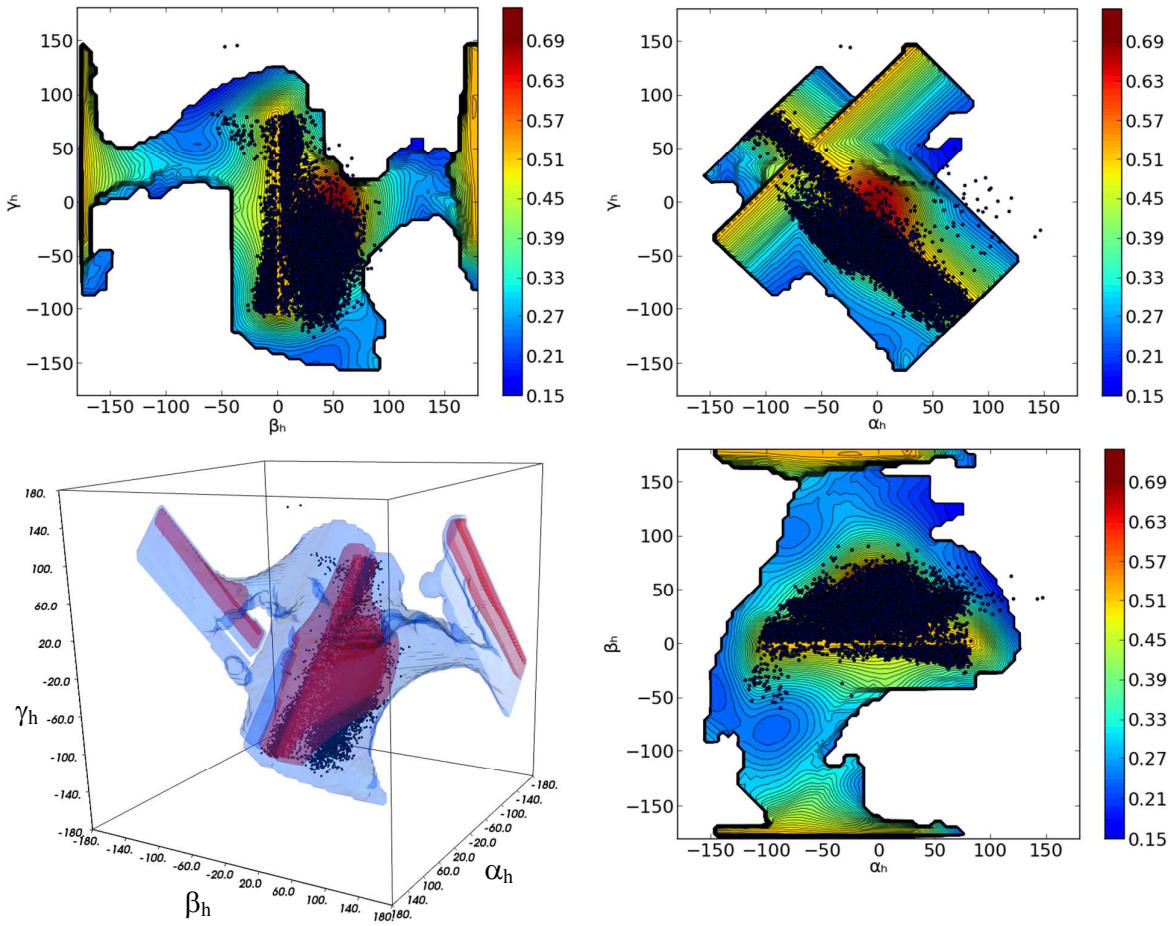


Figure 4 a) The MaxOcc landscape (MaxOcc values color coded) as a function of the β_h and $\alpha_h + \gamma_h$ coordinates. White areas correspond to not sampled regions. b) Superimposition of the MD trajectory (dark dots) to the MaxOcc landscape. c) The smallest region centered at the MaxOcc peak with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

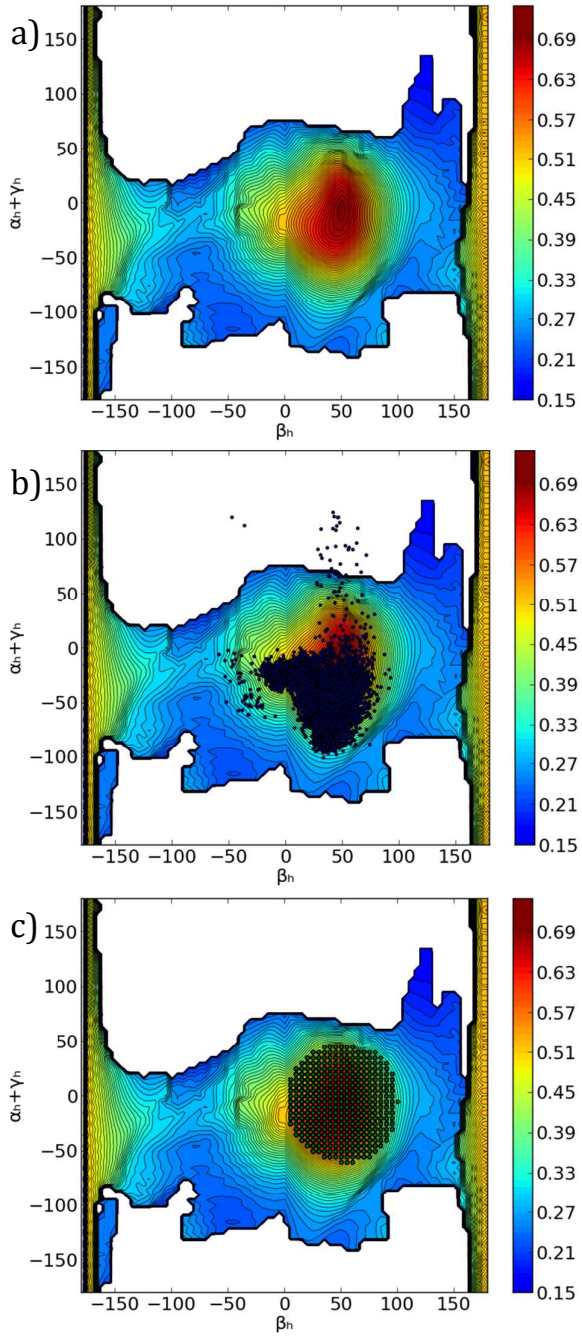


Figure 5 Ensemble selected from the MD trajectory by SAS ('RDC reweighted trajectory'), divided into three clusters after the original paper (cluster 1 in green, cluster 2 in red and cluster 3 in blue), superimposed to the MaxOcc landscape (color coded) as a function of α_h , β_h and γ_h angles (2D projections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

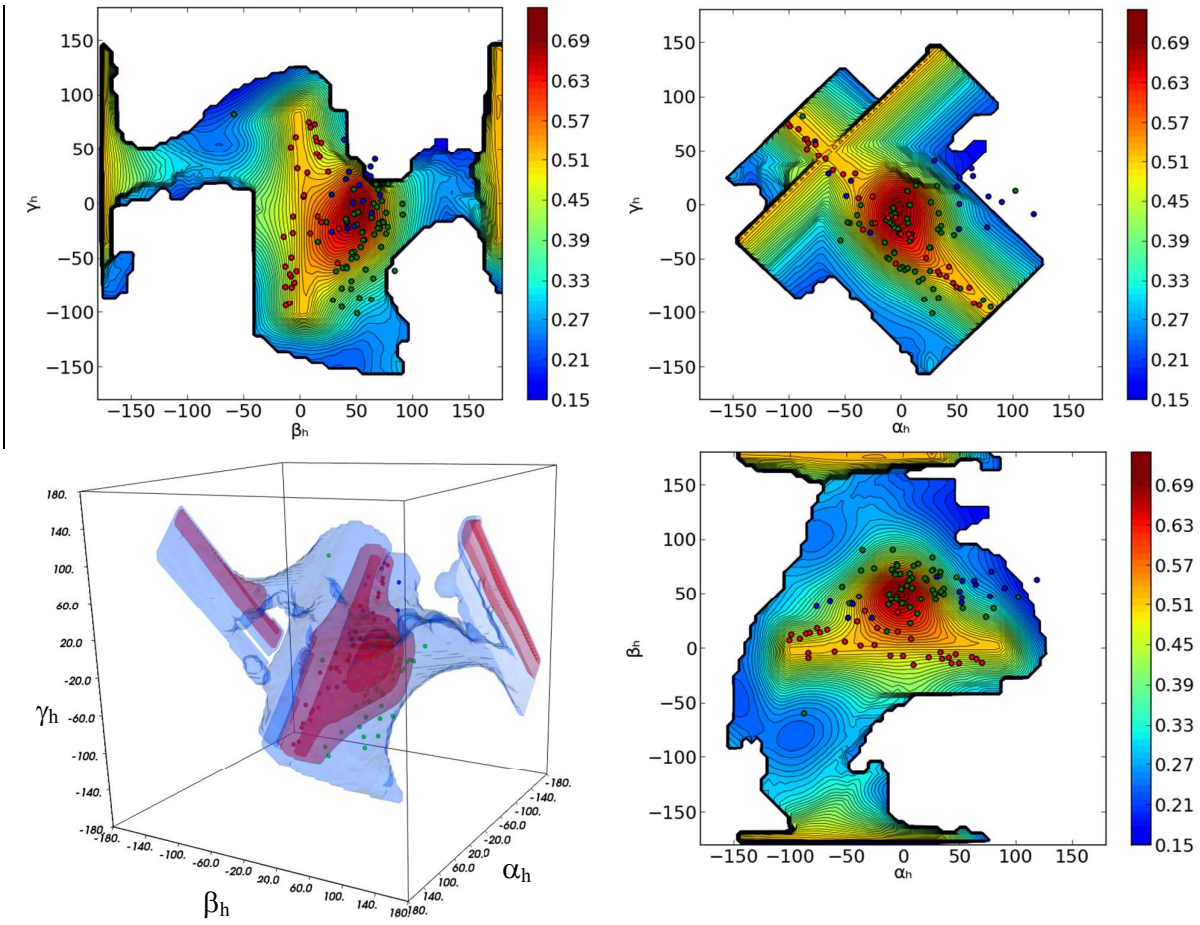
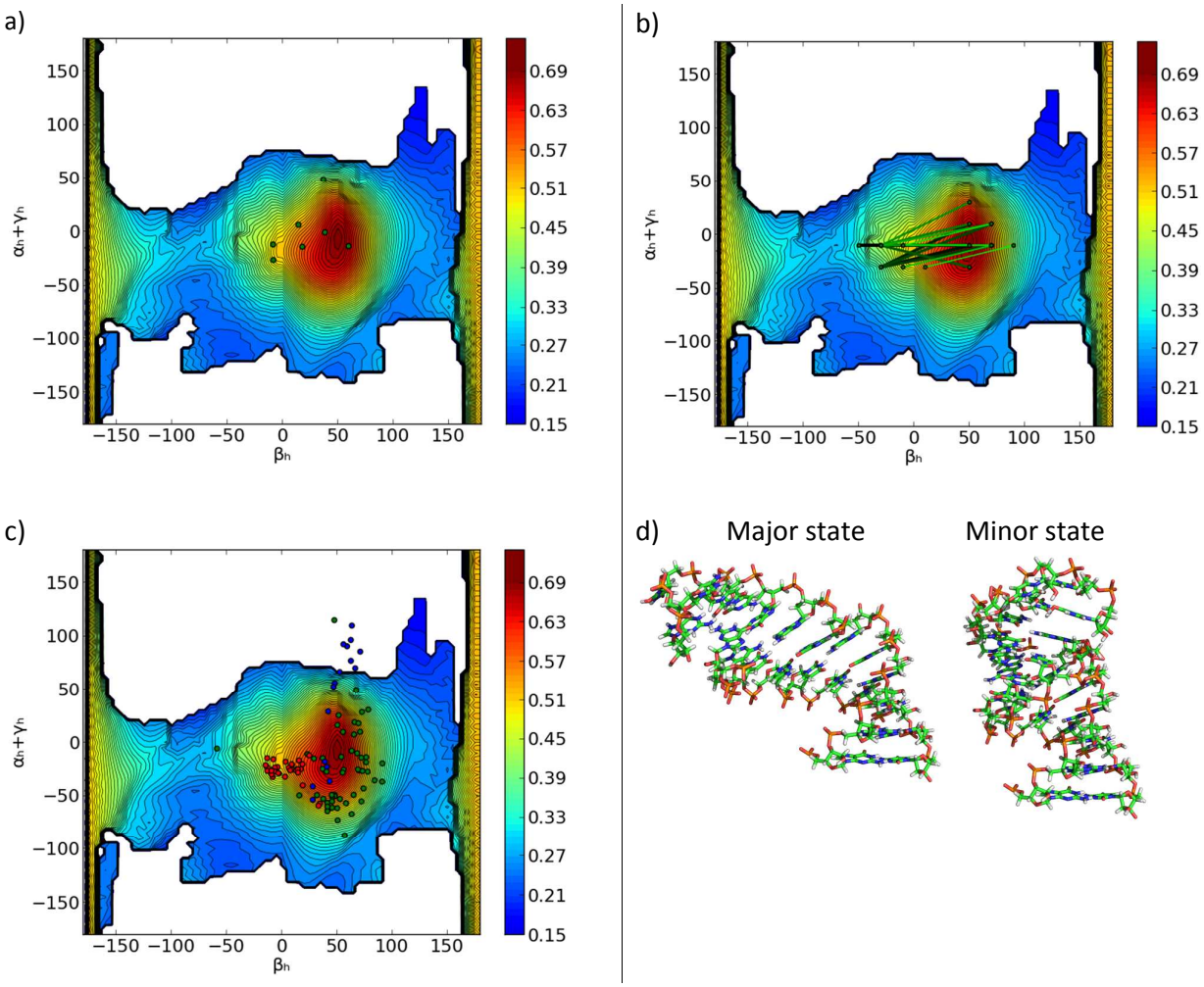


Figure 6 Superposition of the MaxOcc landscape (MaxOcc values color coded), as a function of the β_h and $\alpha_h + \gamma_h$ coordinates, and a) the ligand bound structures available in the PDB (green dots), b) the set of pairs of 20°·20° regions with MaxOR>95% (depicted as dots located in the centers 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 α_h=-15, β_h=-40, γ_h=0° (minor state), able to fit the experimental data.



Reference List

(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.Comp.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, DOI 10.1007/s10858-015-9951-6.

(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, *Biophysical Journal*, 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 Journal on Optimization*, 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.