Measuring similarity between dynamic ensembles of biomolecules

Shan Yang[1], Loïc Salmon[2] & Hashim M Al-Hashimi[3]

We present a simple and general approach termed REsemble for quantifying population overlap and structural similarity between ensembles. This approach captures improvements in the quality of ensembles determined using increasing input experimental data—improvements that go undetected when conventional methods for comparing ensembles are used—and reveals unexpected similarities between RNA ensembles determined using NMR and molecular dynamics simulations.

There is growing interest in moving beyond a static description of biomolecules toward a dynamic description in terms of conformational ensembles[1–3] in which a biomolecule is represented as a population-weighted distribution of many conformations. Studies indicate that biomolecules employ this broad pool of conformations during folding and when carrying out their biological functions[4]. An ensemble description of biomolecules can also help quantify thermodynamically important conformational entropy[5] and define a broad range of receptors that can be targeted in drug discovery[6].

Methods to assess similarity between static structures are well developed and widely used in classifying biomolecules, understanding evolutionary relationships between them and predicting their structures and functions[7]. New methods are needed to compare dynamic ensembles of biomolecules[8–10] and are important not only for helping establish dynamics-function relationships[4] but also in assessing the quality of ensembles determined using biophysical methods[3,8–10]. Among many approaches for comparing probability distributions[11], the Jensen-Shannon divergence (Ω^2)[8,9] and S-score (S)[10] have been used to compare dynamic ensembles of biomolecules[3]. Although these approaches provide quantitative information regarding ensemble similarity, particularly with regard to the population overlap between two distributions, they do not quantify the extent of structural similarity for nonoverlapping conformations.

For example, on the basis of Ω^2 or S, two similar yet nonoverlapping conformational ensembles are measured as having zero similarity, whereas the same level of similarity is assigned to two conformational ensembles that differ much more substantially (**Fig. 1a**). The underlying problem is that nonoverlapping conformations in two distributions contribute to Ω^2 and S in a manner independent of the extent of structural similarity (Online Methods). Other common measures of similarity or distance between probability distributions suffer from the same limitation, including the χ^2 and the Bhattacharyya distance[11]. In addition, in application to ensembles, Ω^2 and S are typically reported for an arbitrarily chosen bin size used to describe a given structural variable. However, these measures of similarity are highly dependent on bin size or method used to cluster conformations in an ensemble[8–10]. An alternative approach, eRMSD, which compares ensembles by computing the pairwise r.m.s. deviation in atomic positions between every pair of conformations in two ensembles[12], does not capture the population overlap, cannot be generally used to dissect individual structural degrees of freedom and can be obscured by outliers.

We developed an approach for simultaneously quantifying population overlap and structural similarity between ensembles: REsemble. Here, the overlap between two distributions is evaluated using methods such as Ω^2 and S as a function of increasing the bin size used to build the histogram describing a given structural variable, such as a torsion angle or distance. Increasing the bin size effectively reduces the 'structural resolution' with which a given structural variable is defined and thereby increases the probability of binning conformations in two ensembles into common bins (**Fig. 1a**). Ensembles that differ substantially in structural terms will require larger bin sizes to overlap. We assess overlap using the square root of Ω^2 because it provides several desirable properties, including being a proven metric[9,11]. The value of Ω comparing two ensembles either stays constant (barring statistical noise) or decreases with increasing bin size, and it always plateaus at $\Omega = 0$ at some bin size cutoff. The plot of Ω versus bin size then provides a rich two-dimensional description of ensemble similarity that captures population overlap and structural similarity simultaneously, with the latter encoded in the steepness with which Ω drops with bin size (**Supplementary Fig. 1**). The approach readily accommodates outliers, which result in long-lasting near-zero Ω plateaus, without compromising the ability to detect similarity in other regions of the ensemble (**Supplementary Fig. 2**). Summing the values of Ω over K bin sizes and normalizing relative to values expected for zero overlap yields a single-value metric $\Sigma_K \Omega(w^T, w^P)$ that ranges between 0 and 1 for perfect and zero similarity, respectively (Online Methods).

When this approach is applied to our previous examples (**Fig. 1a**), the structurally similar but nonoverlapping ensembles start with $\Omega = 1$ for small bin sizes, implying zero similarity, but Ω rapidly drops to 0 with increasing bin size, indicating strong structural similarity (**Fig. 1b**). The drop in Ω with increasing bin size is far less steep for the structurally more dissimilar ensembles. $\Sigma\Omega$ is clearly different in the two cases (0.05 and 0.40; **Fig. 1b**) and captures the structural differences between the two ensembles.

[1]Department of Chemistry, University of Michigan, Ann Arbor, Michigan, USA. [2]Biophysics, University of Michigan, Ann Arbor, Michigan, USA. [3]Department of Biochemistry and Chemistry, Duke University School of Medicine, Durham, North Carolina, USA. Correspondence should be addressed to H.M.A.-H. (hashim.al.hashimi@duke.edu).

RECEIVED 4 NOVEMBER 2013; ACCEPTED 4 MARCH 2014; PUBLISHED ONLINE 6 APRIL 2014; DOI:10.1038/NMETH.2921

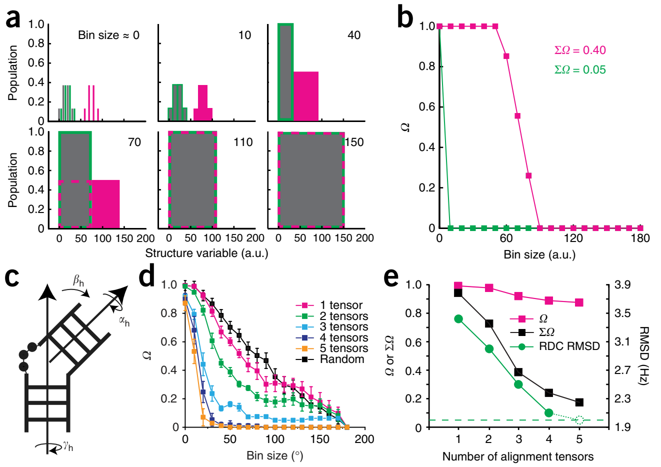

Figure 1 | Measuring population overlap and structural similarity between ensembles. (**a**) Three discrete ensembles (gray, green and magenta) described in terms of an arbitrary structural variable are shown as a function of increasing bin size used to build the histogram distribution. Dashed magenta and solid green boxes around the gray ensemble indicate the portion of magenta and green ensemble, respectively, that are binned together with the gray ensemble. (**b**) Ω as a function of increasing bin size, comparing the gray vs. green (green line) and gray vs. magenta (magenta line) ensembles. (**c**) The relative orientation of two helices (or domains) is defined using three Euler angles (α_h, β_h, γ_h). Shown are two RNA helices linked by a trinucleotide bulge. (**d**) Ω vs. bin size, comparing the interhelical angle distributions about a trinucleotide bulge linker between a target ensemble (N = 5) and ensembles (N = 5) that are selected from the pool randomly (black) or using an increasing number of input RDC data sets in SAS selections (color coded). The s.d. of Ω at each bin size over the 50 repetitions of each prediction is shown as error bars (Online Methods). (**e**) Values of Ω at a bin size of ~0° (magenta squares) and $\Sigma\Omega$ (black squares) as a function of the number of RDC data sets used in ensemble reconstruction. Also shown is the r.m.s. deviation (RMSD) in leave-out cross-validation in which a constructed ensemble is used to predict a common left-out set of RDCs (green circles). The dashed circle represents the optimum RMSD when the left-out data set itself is included in the selection, and the horizontal dashed line denotes the assigned 2-Hz RDC uncertainty.

Approaches for constructing ensembles of biomolecules using experimental data are being developed[1,2,13,14]. In testing these methods, the ability to measure ensemble similarity is fundamentally important. A common ensemble construction approach uses 'sample and select' (SAS)[15] (Online Methods) or a similar scheme[3] to guide selection of conformations from a computationally generated pool and construct ensembles that satisfy experimental data. Methods such as cross-validation[1–3] have been used to show that the quality of constructed ensembles generally improves with increasing input experimental data; however, no study has directly quantified the extent or nature of the improvement.

We used our approach to measure the similarity between a known target ensemble (N = 5) constructed by randomly selecting five conformations from a pool of ~40,000 conformations and ensembles reconstructed using SAS and up to five independent sets of synthetic residual dipolar couplings (RDCs)[16,17] (Online Methods). For simplicity, we focused on determining ensembles describing the relative orientation of two chiral domains (in this case, A-form RNA helices) as defined using three Euler angles (**Fig. 1c**). Here the conformational pool represents the topologically allowed orientations of two A-form helices linked by a trinucleotide bulge[18]. As described previously[18], we measure similarity in terms of the amplitude of single-axis rotations (Online Methods).

The conventional Ω value computed between the target and SAS-reconstructed ensemble at the smallest bin size of ~0° (Online Methods) ranges between 0.87 and 0.99 (**Fig. 1d**). This implies a very poor level of similarity that is comparable to that observed when comparing the target ensemble with an ensemble (N = 5) constructed by randomly selecting conformations from the same pool without guidance from RDC data (Ω = 0.99). Moreover, Ω changes insubstantially when the number of RDC data sets used to reconstruct the ensemble is increased (**Fig. 1d**). Similar results are obtained using the S-score, χ^2 (**Supplementary Fig. 3**) and Bhattacharyya distance (data not shown). These results are at odds with cross-validation analysis (Online Methods), which shows substantial improvements in the quality of ensembles determined with increasing RDC data sets as judged by their ability to predict a common fifth RDC data set that is left out from the ensemble construction. The r.m.s. deviation between measured and predicted RDCs approaches the assigned RDC uncertainty when four RDC data sets are used, a result implying strong similarity between the target and reconstructed ensembles (**Fig. 1e**). This improvement in ensemble construction with increasing RDC data sets is perfectly captured when Ω is computed as a function of increasing bin size. Ω decreases with increasing bin size, and this reduction occurs more rapidly when a larger number of RDC data sets is used in the ensemble construction (**Fig. 1d**). This decrease is much less steep for the randomly selected ensemble (**Fig. 1d**), resulting in $\Sigma\Omega$ values that decrease with increasing input RDC data sets, in excellent agreement with the cross-validation results (**Fig. 1e**). Similarly, our approach captures improvements in the constructed ensembles upon decreasing RDC uncertainty that go undetected with direct application of Ω (**Supplementary Fig. 4**).

We also used our approach to assess the quality of an ensemble determined for the transactivation response element (TAR) RNA (**Fig. 2a**) from the human immunodeficiency virus type 1 (HIV-1) using molecular dynamics (MD) simulations. We previously reported[19] poor agreement (r.m.s. deviation = 8.6 Hz; experimental uncertainty ≈ 2 Hz) between four independent sets of RDCs measured in TAR (**Supplementary Fig. 5**) and RDCs predicted for a TAR ensemble obtained from an 8.2-μs MD simulation computed on Anton supercomputer using the CHARMM36 force field[20]. The specific degrees of structural freedom that underlie this disagreement remain unclear and are difficult to resolve given that RDCs report on both local and global aspects of structure[16,17].

We previously showed[19] that, using the SAS approach, we could construct a TAR ensemble that much better satisfies the four sets of RDCs from the MD-generated pool (**Supplementary Fig. 5**). To assess the source of discrepancy between the MD simulation and measured RDCs, we used our approach to directly compare the MD trajectory and the SAS-based RDC-selected ensemble. We observed substantial differences ($\Sigma\Omega$ = 0.40) in the interhelical angle distributions between the two ensembles (**Fig. 2b**). This discrepancy alone is expected to affect all RDCs measured in TAR because changes in interhelical orientation lead to changes in the global structure and overall alignment of the molecule. The observed differences in interhelical angle distributions are

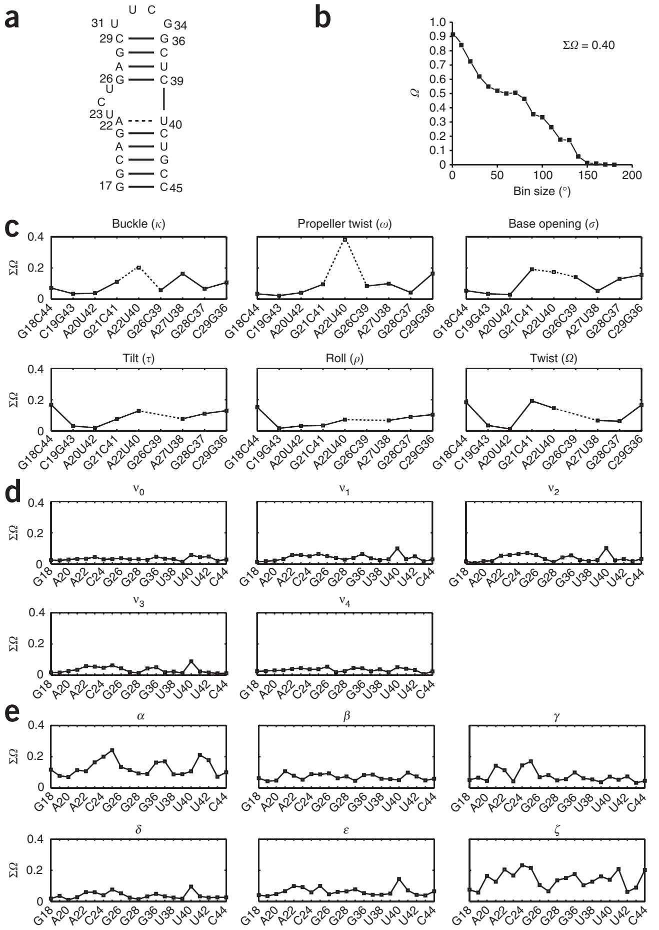

Figure 2 | Comparing MD-generated and NMR RDC–selected ensembles of HIV-1 TAR. (**a**) Secondary structure of HIV-1 TAR RNA. The highly flexible junction A22-U40 base pair is indicated by a dashed line. (**b**) Ω vs. bin size plot comparing the interhelical angle distribution in the MD and RDC-selected (N = 20) ensembles. The binning is performed in terms of single-axis rotation amplitudes (Online Methods). (**c**–**e**) $\Sigma\Omega$ value comparing the distributions of base-pair parameters (**c**) and sugar (**d**) and backbone (**e**) torsion angles between the MD and the RDC-selected ensemble. The intra-base-pair parameters for the flexible junction A22-U40 base pair are shown using open symbols, but dashed lines and inter-base-pair parameters are not shown for the junction G26-C39 base pair are ill defined owing to the presence of the bulge between G26-C39 and A22-U40.

not surprising given that longer simulations are likely needed to properly sample conformational space and that the TAR interhelical orientation strongly depends on ionic strength[18].

In contrast, we observed much better agreement for local angle parameters, including base-pair parameters (**Fig. 2c**) and sugar (**Fig. 2d**) and phosphodiester-backbone torsion angles (**Fig. 2e**), for which $\Sigma\Omega < 0.20$ on average. Cases with $\Sigma\Omega$ of >0.2 are rare and tend to be concentrated in the junction A22-U40 base pair and bulge residues, which have previously been shown to be flexible by NMR spin relaxation[13], and the phosphodiester backbone torsion angles α and ζ, which show broad distributions in the MD ensemble (**Supplementary Fig. 6**). The deviations in α and ζ at the bulge linker, and in base-pair parameters for residues surrounding the bulge, are likely linked to the deviations observed in the interhelical angle distributions (**Fig. 2b**). The ability of RDCs to define all the above angles during the SAS selection was confirmed by simulation tests (**Supplementary Fig. 7**). It is interesting to note that by defining interhelical orientation and helical parameters, RDCs indirectly help define phosphodiester-backbone torsion angles in and around the bulge[19]. These results suggest that even though the MD trajectory yields poor agreements with RDCs measured throughout TAR, the main source of disagreement is the interhelical angle distribution.

In conclusion, we report a simple and robust method, REsemble, to measure the similarity between dynamic ensembles that overcomes limitations in conventional methods that primarily capture population overlap at a single bin size and thereby fail to measure structural similarity. The approach can be used in conjunction with many other appropriate metrics to compare any structural variable of interest and can broadly be applied to measure ensemble similarity in proteins, nucleic acids and other polymers determined using a wide variety of biophysical methods.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank members of the Al-Hashimi laboratory for critical comments on the manuscript. This work was supported by the US National Institutes of Health (R01AI066975 and P01 GM0066275).

AUTHOR CONTRIBUTIONS

S.Y., L.S. and H.M.A.-H. conceived the idea; S.Y. and L.S. carried out the data analysis with help from H.M.A.-H.; S.Y., L.S. and H.M.A.-H. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Jensen, M.R. *et al. Structure* **17**, 1169–1185 (2009).
2. Clore, G.M. & Schwieters, C.D. *Biochemistry* **43**, 10678–10691 (2004).
3. Salmon, L., Yang, S. & Al-Hashimi, H.M. *Annu. Rev. Phys. Chem.* doi:10.1146/annurev-physchem-040412-110059 (16 December 2013).
4. Boehr, D.D., Nussinov, R. & Wright, P.E. *Nat. Chem. Biol.* **5**, 789–796 (2009).
5. Wand, A.J. *Curr. Opin. Struct. Biol.* **23**, 75–81 (2013).
6. Stelzer, A.C. *et al. Nat. Chem. Biol.* **7**, 553–559 (2011).
7. Richardson, J.S. & Richardson, D.C. *Annu. Rev. Biophys.* **42**, 1–28 (2013).
8. Lindorff-Larsen, K. & Ferkinghoff-Borg, J. *PLoS ONE* **4**, e4203 (2009).
9. Fisher, C.K., Huang, A. & Stultz, C.M. *J. Am. Chem. Soc.* **132**, 14919–14927 (2010).
10. De Simone, A., Richter, B., Salvatella, X. & Vendruscolo, M. *J. Am. Chem. Soc.* **131**, 3810–3811 (2009).
11. Cha, S.-H. *Int. J. Math. Models Methods Appl. Sci.* **1**, 300–307 (2007).
12. Brüschweiler, R. *Curr. Opin. Struct. Biol.* **13**, 175–183 (2003).
13. Frank, A.T., Stelzer, A.C., Al-Hashimi, H.M. & Andricioaei, I. *Nucleic Acids Res.* **37**, 3670–3679 (2009).
14. Marsh, J.A., Teichmann, S.A. & Forman-Kay, J.D. *Curr. Opin. Struct. Biol.* **22**, 643–650 (2012).
15. Chen, Y., Campbell, S.L. & Dokholyan, N.V. *Biophys. J.* **93**, 2300–2306 (2007).
16. Tolman, J.R., Flanagan, J.M., Kennedy, M.A. & Prestegard, J.H. *Proc. Natl. Acad. Sci. USA* **92**, 9279–9283 (1995).
17. Tjandra, N. & Bax, A. *Science* **278**, 1111–1114 (1997).
18. Bailor, M.H., Mustoe, A.M., Brooks, C.L. III. & Al-Hashimi, H.M. *Nat. Protoc.* **6**, 1536–1545 (2011).
19. Salmon, L., Bascom, G., Andricioaei, I. & Al-Hashimi, H.M. *J. Am. Chem. Soc.* **135**, 5457–5466 (2013).
20. Denning, E.J., Priyakumar, U.D., Nilsson, L. & Mackerell, A.D. Jr. *J. Comput. Chem.* **32**, 1929–1943 (2011).

ONLINE METHODS

Jensen-Shannon divergence and *S*-score. Mathematical expressions for the Jensen-Shannon divergence (Ω^2) and S-score (S) are given by equations (1) and (2), respectively

$$\Omega^2(w_i^T(m), w_i^P(m)) = S\left(\frac{w_i^T(m) + w_i^P(m)}{2}\right) - \frac{1}{2}[S(w_i^T(m)) + S(w_i^P(m))] \tag{1}$$

$$S(w_i^T(m), w_i^P(m)) = \frac{1}{2}\sum_{i=1}^{N} | w_i^T(m) - w_i^P(m) | \tag{2}$$

in which $w_i^T(m)$ and $w_i^P(m)$ represent the population weights for the ith bin in ensemble T and P, respectively, for a given bin size m. $S(w_i) = -\Sigma w_i(m)\log_2 w_i(m)$ in equation (2) is the information entropy. Ω^2 and S vary between 0 and 1 for maximum and minimum similarity and are equal to 0 if and only if $\{w_i^T(m)\} = \{w_i^P(m)\}$. Equations (1) and (2) show that for nonoverlapping regions in two distributions, defined as cases in which $w_i^T(m) = 0$; $w_i^P(m) \neq 0$ or $w_i^T(m) \neq 0$; $w_i^P(m) = 0$, the contribution to Ω^2 and S is independent of the extent of structural similarity.

The sum of population overlap over all bin sizes (K) normalized relative to values expected for zero overlap ($\Omega = 1$ for all bin sizes) provides a convenient single-value measure of population overlap and structural similarity that we refer to as $\Sigma_K\Omega(w^T, w^P)$ and that ranges between 0 and 1 for perfect and zero similarity, respectively,

$$\sum_K \Omega(w^T, w^P) = \frac{\sum_m \Omega(w_i^T(m), w_i^P(m))}{K} \tag{3}$$

Note that $\Sigma_K\Omega(w^T, w^P)$ is also a metric and therefore symmetric: $\Sigma_K\Omega(w^T, w^P) = \Sigma_K\Omega(w^P, w^T)$, and it is equal to 0 if and only if two distributions are identical at all bin sizes or $w^T = w^P$.

Sample and select (SAS) approach. In the SAS approach[13,15,19], experimental RDCs are used to guide construction of an ensemble by selecting N conformations from a conformational pool that minimize the following χ^2 function:

$$\chi^2 = \sum_{i=1}^{L} (D_i^{\text{calc}} - D_i^{\text{exp}})^2 / L \tag{4}$$

in which L is the total number of RDCs used in SAS; D_i^{calc} and D_i^{exp} are calculated and experimentally measured RDCs, respectively. In our implementation of SAS, first an initial ensemble of N conformations is randomly selected from the pool. Then at each step (k) of the selection procedure, one conformation in the ensemble is randomly chosen and replaced by a conformation randomly selected from the rest of the pool. The change from step k to $k + 1$ is accepted if $\chi^2(k + 1) < \chi^2(k)$; if $\chi^2(k + 1) \geq \chi^2(k)$ with a probability $P = \exp((\chi^2(k) - \chi^2(k + 1))/T)$, where T is an effective temperature that is linearly decreased using a simulated-annealing scheme[13]. The initial effective temperature is set sufficiently high so that >99% of the conformations can be replaced and slowly decreased until the acceptance probability is smaller than 10^{-5}. At each effective temperature, 200,000 steps were implemented followed by a decrease of effective temperature using $T_{i+1} = 0.9T_i$. A Matlab script (**Supplementary Software**) was used to implement this SAS-based ensemble construction.

Evaluating quality of interhelical ensembles determined with increasing input RDCs. The capability of RDCs to reconstruct interhelical ensembles using the SAS approach was investigated using synthetic RDC data, using up to five RDC data sets corresponding to five perfectly orthogonal alignment tensors. In these simulations, a given conformation is represented using three interhelical Euler angles (α_h, β_h, γ_h) describing the relative orientation of the two idealized A-form helices representing the TAR helices connected by a trinucleotide bulge (**Fig. 1c**). The conformational pool necessary for the SAS selection was generated by using the corresponding topologically allowed space. This space corresponds to all possible interhelical orientations that satisfy basic steric and connectivity restraints imposed by the bulge[18]. The pool was generated using a 5° resolution grid (i.e., each conformation differs from its closest neighbor by a 5° change in one of the three Euler angles). For a trinucleotide bulge, the pool represents ~10% of the total possible interhelical orientations. A target ensemble containing five distinct conformations ($N = 5$) was then randomly selected from this topologically allowed pool. Five orthogonal alignment tensors arbitrarily fixed on the reference helix were then generated using the Gram-Schmidt procedure[21]. For each of the five alignment tensors, all possible one-bond CH RDCs were computed for the target ensemble. For each alignment tensor, the RDCs for the five conformations were averaged and error corrupted assuming 2-Hz RDC uncertainty.

The SAS approach was then implemented to select an ensemble of $N = 5$ distinct conformations using one, two, three, four and five sets of input RDCs to guide selection. The target and the predicted ensemble were then compared using similarity measurements including Ω, S-score, χ^2 and Bhattacharyya distance at various bin sizes as described below. The same process was repeated 50 times, and the similarity between target and predicted ensembles was averaged over these 50 comparisons at each bin size. s.d. was calculated to estimate the variation in repeated simulations using different numbers of input alignments. The s.d. was similar across the groups and was small compared to the observed differences between them (**Fig. 1d**). For the RDC cross-validation analysis, ensembles determined using one, two, three and four RDC data sets in the SAS selection were used to predict a fifth RDC data set that was not used in the selection. The resultant r.m.s. deviation between the RDCs for this fifth data set and values back-calculated from the predicted ensemble was then computed[19].

Binning interhelical orientations. The Cartesian distance in the Euler space $((\alpha_{hA} - \alpha_{hB})^2 + (\beta_{hA} - \beta_{hB})^2 + (\gamma_{hA} - \gamma_{hB})^2)^{1/2}$ between two sets of Euler angles A and B defining two distinct interhelical orientations does not provide a measure of structural similarity between the two conformations[18]. First, there are inherent degeneracies ($\alpha_h' = \alpha_h + 180$, $\beta_h' = -\beta_h$, $\gamma_h' = \gamma_h + 180$; $\alpha_h' = \alpha_h - 180$, $\beta_h' = -\beta_h$, $\gamma_h' = \gamma_h - 180$; $\alpha_h' = \alpha_h + 180$, $\beta_h' = -\beta_h$, $\gamma_h' = \gamma_h - 180$; $\alpha_h' = \alpha_h - 180$, $\beta_h' = -\beta_h$, $\gamma_h' = \gamma_h + 180$) that map several sets of distinct interhelical Euler angles to the same conformation[18]. This problem was overcome by using a restricted grid of Euler angles devoid of any degeneracy[18]. Second, even after taking into account the above degeneracy, the Cartesian distance between two sets of Euler angles does not provide a faithful measurement of structural similarity. For example, the Cartesian distances between (0, 0, 0) and (5, 5, 5) is ~9° in the Euler space, whereas the two conformations differ by single-axis rotation with amplitude of ~11°. Likewise, the conformations (5, 5, 0) and (170, –10, 170) differ by a Cartesian distance of ~237°, but the two conformations differ by a single-axis rotation with amplitude of ~25°. More generally, the Cartesian distance between Euler angles can be smaller than, equal to or larger than the actual difference between two conformations. Therefore, we used the amplitude of single-axis rotation to bin interhelical orientations together and measure similarity between ensembles[18] (see below).

The binning grid points are constructed by picking a binning origin, defined by the minimum value of each of the three Euler angles in the two ensembles upon comparison, and then incrementing each Euler angle by an amount defined by the bin size to cover the entire nondegenerate 3D Euler space. Changing the binning origin has minimal effects on the resulting analysis (data not shown). Next, the amplitude of a single-axis rotation (ω) connecting a given conformation in the ensemble defined by Euler angles (α_{h1}, β_{h1}, γ_{h1}) and a point on the grid (α_{h2}, β_{h2}, γ_{h2}) is computed[18]

$$R(\alpha_{h1},\beta_{h1},\gamma_{h1}) = O(x,y,z,\omega)R(\alpha_{h2},\beta_{h2},\gamma_{h2}) \tag{5}$$

where $R(\alpha_{hi}, \beta_{hi}, \gamma_{hi})$ is the Euler rotation matrix that transforms the coaxial interhelical orientation (0, 0, 0) into the orientation defined by (α_{hi}, β_{hi}, γ_{hi}) (ref. 18), and $O(x, y, z, \omega)$ represents a single-axis rotation about a unit vector (x, y, z) with amplitude ω. $O(x, y, z, \omega)$ can also be expressed by a 3 × 3 matrix in terms of x, y, z and ω

$$O(x,y,z,\omega) = \begin{pmatrix} \cos\omega + x^2(1-\cos\omega) & xy(1-\cos\omega) - z\sin\omega & xz(1-\cos\omega) + y\sin\omega \\ xy(1-\cos\omega) + z\sin\omega & \cos\omega + y^2(1-\cos\omega) & yz(1-\cos\omega) - x\sin\omega \\ xz(1-\cos\omega) - y\sin\omega & xy(1-\cos\omega) + x\sin\omega & \cos\omega + z^2(1-\cos\omega) \end{pmatrix} \tag{6}$$

And the rotation amplitude ω is given by

$$\omega = \arccos\left(\frac{O_{11} + O_{22} + O_{33} - 1}{2}\right) \tag{7}$$

in which O_{11}, O_{22} and O_{33} are the three diagonal elements of $O(x, y, z, \omega)$.

In this manner, the amplitude of the single-axis rotation connecting a given conformation in an ensemble to every grid point is computed, and the conformation is binned to the grid point that leads to the minimum single-axis rotation amplitude ω. The population of each grid point is then calculated to be the number of conformations binned divided by the total number of conformations in the ensemble. In our case, binning of the target and the predicted ensemble led to two population distributions on the same binning grid for a given bin size, and the value of Ω between the two ensembles at the given bin size is then calculated using equation (1). This procedure was repeated as a function of increasing bin size. This analysis was performed using a Matlab script (**Supplementary Software**).

Analysis of MD trajectory–based ensembles. An in-house Perl script (**Supplementary Software**) was used to compute interhelical angles (α_h, β_h, γ_h) describing the relative orientation of two A-form helices[18]. All intra- and inter-base-pair parameters were computed using Curves+[22], and all the local torsion angles defining the sugar and backbone geometry were computed using an in-house C script (**Supplementary Software**). The resulting interhelical orientations defined by three Euler angles were binned and analyzed as described above. Distributions of base-pair parameters and sugar and backbone torsion angles were directly binned to a binning grid ranging between 0° and 360° with evenly distributed increments defined by the bin size. The value of Ω was calculated at each given bin size for each parameter/angle distribution using equation (1), and the values of $\Sigma\Omega$ were calculated using equation (3) for distributions of interhelical orientation, base-pair parameter and sugar and backbone torsion angles.

21. Fisher, C.K., Zhang, Q., Stelzer, A. & Al-Hashimi, H.M. *J. Phys. Chem. B* **112**, 16815–16822 (2008).
22. Lavery, R. & Sklenar, H. *J. Biomol. Struct. Dyn.* **6**, 655–667 (1989).

doi:10.1038/nmeth.2921