Article

Determination of a structural ensemble representing the dynamics of a G-quadruplex DNA

Guillem Portella, Modesto Orozco, and Michele Vendruscolo

Biochemistry, **Just Accepted Manuscript** • DOI: 10.1021/acs.biochem.9b00493 • Publication Date (Web): 09 Dec 2019

Determination of a structural ensemble representing the dynamics of a G-quadruplex DNA

Guillem Portella[1,2,3], Modesto Orozco[2,3,4], Michele Vendruscolo[1,*]

[1]*Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK,*
[2]*Institute for Research in Biomedicine (IRB Barcelona), Barcelona Institute for Science and Technology (BIST), 08028 Barcelona, Spain,*
[3]*Joint BSC-CRG-IRB Research Program in Computational Biology, 08028 Barcelona, Spain,*
[4]*Department of Biochemistry and Biomedicine, University of Barcelona, 08028 Barcelona, Spain*

[*]Correspondence to: mv245@cam.ac.uk

Abstract

It is increasingly recognised that the structure and dynamics of G-quadruplex DNA are dictated by its sequence and greatly affected by environmental factors. The core guanine tetrads (G-tetrads) coordinate cations and display a strong conformational rigidity compared with the connecting loops. Although long loops linking the G-tetrads are typically disfavoured, when present, they provide a striking view of the dynamics of short, single-stranded DNA regions. In addition to their role in determining the stability of the G-quadruplex state, these loops are potential drug targets. In order to characterise accurately the dynamics of this DNA state, we apply the principles of structural ensemble determination developed in the last two decades for protein molecules to DNA molecules. We thus performed extensive molecular dynamics simulations restrained with NMR residual dipolar couplings to determine a structural ensemble of the human CEB25 minisatellite G-quadruplex, which contains a connecting loop of 9 nucleotides. This structural ensemble displays a wide set of arrangements for the loop and a compact, well-defined G-quadruplex core. Our results show the importance of stacking interactions in the loop and strengthen the ability of the closing base pairs to confer a large thermodynamic stability to the structure.

Introduction

Beyond the canonical duplex form, DNA can populate a wide range of states, from single-stranded conformations to four-stranded arrangements. The remarkable plasticity of DNA is particularly well exemplified by the large array of G-quadruplex (G4) structures known to date[1-9]. These assemblies are stabilised by Watson-Crick/Hoogsteen base pairs and reinforced by cation coordination[4]. The Hoogsteen base pairs involve the N7 nitrogen of the purine base and the C6 amino group of an interacting pyrimidine base, and are among the most common non-canonical base pairs, allowing the formation of intramolecular and intermolecular triplexes and quadruplexes *in vitro*. Increasing evidence indicates that these structures also exist *in vivo* and have important biological roles[10]. G4s are abundant in telomeric regions, where they can exist as an RNA/DNA hybrid, and G4-forming sequences are enriched in near promoter regions and transcription start sites[11]. Their presence in gene promoters has identified G4s as druggable targets[11-14], and there is strong evidence that G4s are stable and detectable across the human genome[15]. The ligands developed to date that target G4s are primarily designed to bind the G4 guanine tetrads[16-17], or its wide grooves[18-19]. Most of these ligands show large affinities for G4s, but low sequence specificity. It has thus been proposed that binding specificity might be achieved by targeting the loop regions of G4s[11].

A typical consensus sequence for G-quadruplex, d($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$), where N can be any base[10], indicates that long intra-guanine loops are typically not favoured. A recently published NMR structure of a propeller-type parallel stranded G-quadruplex (CEB25 G-quadruplex) containing a 9-nucleotide loop shows a remarkably large thermal stability (2LPW)[20]. The loop appears somewhat structurally confined in the 2LPW set of conformers, with a portion (residues 11-17) showing larger structural variability. Notably, residues A2 and T18 form a base pair clipping together chain 2 and 4, which has been shown to give a large thermal stability to quadruplex[20]. Interestingly, mutations along the loop region 10-16 do not affect the overall stability of the G4.

Overall, this G-quadruplex provides a unique opportunity to visualize the conformational dynamics of relatively long DNA loops. In order to achieve this goal, we exploit the principles of structural ensemble determination that have been established for proteins[21-22] and RNA[23-27], and are beginning to be applied to DNA[28-29]. In our approach, we perform restrained molecular dynamics (MD) simulations with NMR residual dipolar couplings

(RDCs) to characterise the conformational fluctuations of this G-quadruplex. The resulting set of conformations, together with their statistical weights, define a 'Boltzmann ensemble', which represents the range of structures that are populated during the conformational fluctuations of the G-quadruplex. This ensemble is not simply made up by multiple models, each one of which is consistent individually with the available experimental data, which could be called an 'uncertainty ensemble' and that would reflect the errors in the determination of the average structure of the G-quadruplex itself.

Materials and Methods

Molecular dynamics simulations

The sequence of G4 is d(AAGGGTGGGTGTAAGTGTGGGTGGGT), where residues 10-18 form a 9-nucleotide loop. The first structure in the PDB entry 2LPW[20] was used as initial structure. The structure was positioned in a previously equilibrated mixture of water using an octahedral cage such that the box boundaries are placed at least 2 nm away from any DNA atom. The amount of potassium ions added guarantees the overall charge neutrality. One potassium ion was placed at the center of each guanine tetrad at the beginning of the simulation. After the systems were prepared, they were subject to energy minimization and 10 ns of molecular dynamics simulations to randomize the conformations and obtain a constant system density. We performed four types of simulations starting from the equilibrated G-quadruplex system (**Table S1**):

- Three independent molecular dynamics simulations (MD-BSC1), starting from three frames randomly obtained from the last 5 ns of the 10 ns initial equilibrium simulation. These simulations amount to a total of 5.2 μs (two of 1.7 μs and one of 1.8 μs).
- One molecular dynamics simulations of 1 μs with RDC-ensemble restraints using 8 replicas (M8-BSC1+RDC) (see below for further details), which results in a cumulated trajectory of 8 μs for the analysis.
- One bias-exchange (BE) metadynamics simulation of 1 μs (see below for details). Due to the number of unbiased replicas used in the BE we obtained a concatenated trajectory of 4 μs for the analysis.

- Three BE metadynamics with RDC-ensemble restraints (BE-BSC1+RDC), two of 2 μs and one of 0.6 μs. Due to the number of unbiased replicas used in the BE we obtained a concatenated trajectory of 18.4 μs for the analysis

All simulations were carried out using the Gromacs-4.6 software[30], with periodic boundary conditions and the particle mesh Ewald method[31] for the long-range electrostatics, using a cut-off of 1.0 nm for the short-range repulsive and attractive dispersion interactions, which were modeled via a Lennard-Jones potential. We used the Settle algorithm[32] to constrain bond lengths and angles of water molecules, and P-Lincs[33] was used for all other bond lengths, allowing a time step of 2 fs for the integration of the Newton's equations of motion. The temperature was held constant at 300 K by using a thermostat.[34] The pressure was controlled by coupling the simulation box to a pressure bath of 1 atm.[35] The force field describing the interactions for DNA was generated based on the parmBSC1 parameters,[36-37] and we used the SPC/E[38] model to describe the water molecule, sodium and potassium ions using the parameters of Smith and Dang.[39]

Molecular dynamics simulations with RDC restraints

We obtained the RDCs for atoms for the C1'-H1' bonds and C8-H8/C6-H6 of each residues from data files deposited in the BMRB database[40]. We employed the recently introduced θ-method to incorporate the RDC information as restraints to our molecular dynamics simulations[41]. This method does not require determination of the molecule alignment tensor, as it restraints the orientation of individual bonds along the Z-axis of the reference system. In our method, at least 8 copies of the system are simulated simultaneously[42] (M8-BSC1 simulations), and the restraints are imposed on the overall ensemble of structures. Each individual RDC is calculated from the structure using

$$D_{XY} = -\frac{\gamma_X \gamma_Y \mu_0 h}{(2\pi r_{XY})^3} \left\langle \frac{3cos\theta - 1}{2} \right\rangle \quad (1)$$

where θ is the angle of the pair of atoms XY with respect to the Z-axis, γ_a is the gyromagnetic constant of the atom a, h is the Planck constant, μ_0 the vacuum permeability, and r_{XY} the distance between the pair of atoms. The overall Pearson correlation between computed and experimental RDCs is initially restrained to a value of 1 using a harmonic potential with a force constant of 10000 kJ/mol. From this simulation, we determine the scaling factor α from

the linear relation $D_{exp}=\alpha D_{calc}$, making sure that the value does not change significantly as we increase the simulation time, and the correlation has reached values above 0.99. Next, we perform the actual RDC-restrained simulations using the determined α value. In these simulations, the quality factor (Q-factor), $q = \sqrt{\sum\left(P_{calc} - P_{exp}\right)^2} \Big/ \sqrt{\sum P_{exp}^2}$, computed as a weighted root-mean square deviation between the experimental and ensemble averaged calculated RDCs, is restrained to a value of zero with a harmonic potential with a force constant of 2000 kJ/mol.

Bias-exchange metadynamics simulations

The RDC restraints within θ method can be combined with enhanced sampling techniques, such as bias-exchange metadynamics[43], which results in the type of simulation we termed BE-BSC1-RDC (**Table S1**). We performed well-temperate bias-exchange metadynamics to efficiently explore the conformational landscape of the G-quadruplex 9-nt long loop. In these calculations we subjected the α, ε, χ and ζ dihedral angles of each residue between residues 10 to 18 to a time-dependent biasing potential. Each dihedral was biased in a different replica of the system, and an exchange step between pairs of replicas was attempted every 2 ps with a Monte Carlo acceptance criteria. Four extra non-biasing replicas were added to the simulation, totaling 40 replicas. In each replica, the dihedral under consideration was biased with a time-dependent Gaussian potential deposited every 1 ps, with σ=0.1 and height h=0.15. The well-tempered bias factor was set to 8 in all simulations.

We performed three independent runs with RDC-restrained bias exchange metadynamics, starting from different initial structures of the G4, one of 0.6 μs and two of 1.8 μs in length. Note that for each run we obtain 4 unbiased replicas that can be readily analyzed without need of post-processing, but which benefited from exchanging conformations with biased replicas. Therefore, the cumulated BE-BSC1-RDC trajectory amounts to 18.4 μs.

Convergence measures

The required simulation time for the convergence depends on the simulation type and the metric utilized. We found that, not surprisingly, the use of BE significantly speeds up the convergence and increases the sampling, and that those with RDC-restraints require longer times. We evaluated the convergence of our simulations by monitoring local fluctuations via time dependent dihedral free energy profiles and residue-averaged root mean square

fluctuations (RMSF), as well correlated motions measured by principal component analysis (PCA) and the Schlitter conformational entropy[44].

The free energy profiles for each biased dihedral as the simulation time increases show that differences between local minima converge after ~10 microseconds aggregate time (**Fig. S3**) for the case of the BE-BSC1-RDC trajectories (**Table S1**), whereas BE-BSC1 trajectories converge very quickly, in less than 3 microseconds. M8-BSC1-RDC and MD-BSC1 seem reasonably converged after 4 microseconds (**Fig. S3**). The residue averaged RMSF provide a similar picture (**Fig. S4**).

The conformational space of the G-quadruplex around its native state measured using PCA converges in a somewhat similar fashion as described for the torsional free energy profiles. It is clear from the PCA that BE-BSC1 ensembles cover a larger conformational space (**Fig. 4D,E**), with the NMR models lying nearly in the averaged conformation (**Fig. 4B,C**). This larger exploration of the conformational landscape is well reflected in the extent of the conformational entropy (**Fig. 4F**). The dependence of the conformational entropy on the number of structures included in the PCA shows that all our simulations appear to be relatively well converged after ~60% of the total simulation time. The longer BE-BSC1+RDC trajectories had more time to explore the conformational landscape, and thus shows better converged entropy.

The normalized overlap between the covariance matrices of atomic positions from the full trajectory and subsets of increasing number of structures has been suggested as means of checking the convergence of PCA, one being identical matrices. We find that BE-BSC1 conformational space converges very quickly, after 20% of the trajectory (1 microsecond) we reach an overlap above 90%, whereas the BE-BSC1+RDC requires up to 9 microseconds (~50% of the overall number of frames) (**Fig. S5** and **Table S2**). The M8-BSC1+RDC and MD-BSC1 require more than 5 microseconds to reach 90% of overlap with the full trajectory, indicating that more sampling would be desirable when using these simulation protocols.

Back-calculation of RDCs

We assessed the quality of the molecular dynamics simulations by comparing experimental RDCs and NOE-derived interatomic distances, available from 2LPW, against RDCs and distances obtained from our simulations. The quality factor of computed RDCs (Q-factor), a

root mean square deviation between experimental and calculated RDCs weighted by the dispersion of experimental values, depends on the number of structures included in the ensemble. Around 5% of the available NOE distances are not fulfilled during the plain molecular dynamics simulations, which compares to ~1% in the NMR-derived 2LPW set of conformers (**Table S5**). Back-calculation of RDCs from our BE-BSC1 and RDC-restrained BE-BSC1 ensembles reveals a Q-factor lower than that of the 2LPW set of conformers using the same number of structures, which decreases ~0.1 using 15 structures in the ensemble averaging (**Fig. S6**). Reassuringly, the percentage of NOE violations in our RDC-restrained ensemble is similar to the 2LPW set of conformers, ~1% (**Table S5**).

We quantified the accuracy the force field calculations and RDC-restrained molecular dynamics simulations to reproduce the experimental RDCs by means of the Q-factor. Back-calculation of RDCs from a trajectory generated by the θ method is straightforward using equation 1, as the molecule has been already aligned during the simulation. However, direct application of eq. 1 on any structure or ensemble of structures requires finding first the optimal orientation that best reproduce the observed RDCs. For a given ensemble of size *n*, we compute the optimal ensemble averaged alignment using a Monte Carlo (MC) approach. At each iteration step, one random structure from the ensemble is randomly rotated and the new set of RDCs is computed. If the correlation of the ensemble-averaged RDCs with the experimental values increases we accept the move. If the correlation decreases, we accept the move with a probability proportional to $e^{(-\Delta q \beta)}$, where Δq is the difference in correlation coefficient and β an inverse temperature. We use a simulated annealing protocol to decrease the acceptance ratio as the number of MC grows by increasing β, which should drive the search towards the global optimal. The MC iterations are stopped after reaching a given prefixed number, which should be at least ~100 times larger than n, or after reaching a correlation of 0.999. From the aligned ensemble we can extract the scaling factor α and compute the Q-factor. If the size of the ensemble *n* is smaller than the number of structures to consider in the calculation, e.g. generated by a molecular dynamics simulation, we average the back-calculated RDCs over all sets of *n* commensurable within the set of structures, and generate a trajectory/ensemble averaged Q-factor.

NOE distance restraints

We used the Gromacs utility g_disre to determine the ensemble-averaged distances and their violations from the NOE distance restraints deposited for the PDB entry 2LPW. The distance averaged used has the $\langle r^{-6} \rangle$ dependency characteristic of NOE signals.

Structural cluster analysis

We performed structural clustering on the G-quadruplex using a single-linkage algorithm using the root mean square deviation (RMSD) of atomic positions on the BE-MetD+RDC ensemble. As reference atoms, we selected all the heavy atoms in the sugar/phosphate backbone, as well as the N1, N3 and C5 of pyrimidines, and N9, N1 and N7 of purines. We performed three different clustering methods, using three subsets of atoms: using all residues except loop residue 10-16, using the guanine core of the quadruplex (residues 3-5, 7-9, 19-31 and 23-25), and using the loop residues (residues 10 to 18). We used a 0.15 nm cut-off for the single-linkage algorithm for the guanine core, and 0.2 nm for the other two subsets.

Base stacking calculations

We determined the number of stacking bases by a combination of three criteria: inter-base center of mass distance (set to a maximum of 0.5 nm), base co-planarity (angle between the two base plane normal vectors, max 45 deg) and base displacement (angle between the two bases centre of mass and one base plane normal, max 45 deg). The stacking number for a given pair of bases is determined as the product of three step function: $(1-(A/A_0)^6)/(1-(A/A_0)^{12})$, where A is the distance or angle under consideration, and A_0 is the cut-off value. Each function drops quickly from 1 to 0 at the cut-off value, allowing fractional contributions to the overall stacking number.

Stacking energy calculations

We have analyzed the stacking interaction energies between canonical DNA bases in a duplex DNA using the Ascona B-DNA consortium (ABC) database of molecular dynamics simulations[45], which contains 39 different DNA oligomers that include all 136 unique tetranucleotides. In this database the molecular dynamics simulations extend to 100 ns and were carried out using the parm99-BSC0 force field. We computed the stacking energies as the sum of Coulomb and van der Waals energies using the last 85 ns of simulations and a time step of 100 ps.

The averaged stacking energies, in kJ/mol, between adjacent bases are TA=-40.5+/-0.1, AT=-40.9+/-0.1, GA=-48+/-0.2, AG=-50.9 +/- 0.4.

Circular correlation calculations

We computed the circular correlation r_c between pairs of dihedral angles α and β following the equation proposed by Jammalamadaka and SenGupta [46] for n data points

$$r_{c,n} = \frac{\sum_{i=1}^{n} sin\,(\alpha_i - \bar{\alpha}) sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{n} sin^2\,(\alpha_i - \bar{\alpha}) sin^2(\beta_i - \bar{\beta})}}$$

Here $\bar{\alpha}$ and $\bar{\beta}$ are the circular averages of the pair of dihedrals. We used a frequency of 1 ns to obtain the dihedral values from the simulation. The P values associated to the correlation were computed from a test statistic under a null hypothesis of no-correlation, which follows a normal distribution for $n \to \infty$

$$\sqrt{n}\sqrt{\frac{\widehat{\lambda_{20}}\widehat{\lambda_{02}}}{\widehat{\lambda_{22}}}}\, r_{c,n}$$

where

$$\widehat{\lambda_{ij}} = \frac{1}{n}\sum_{k=1}^{n} sin^i\,(\alpha_k - \bar{\alpha}) sin^j(\beta_k - \bar{\beta})$$

Results and Discussion

A structural ensemble representing the dynamics of the G-quadruplex

The large conformational space available to the loop region of the G-quadruplex makes it problematic to represent it as a single, average structure. The approach that we applied here enables a range of structures to be determined that correspond to the conformational fluctuations of this form of DNA (**Fig. 1**). In this approach, NMR measurements are incorporated as structural restraints in molecular dynamics simulations in order to obtain an extensive sampling of the conformational space compatible with the experimental observations[21-23]. As this method implements the maximum entropy principle[47] it offers a statistical mechanics representation of the conformational fluctuations of the G-quadruplex.

The convergence in the simulations of the relevant backbone degrees of freedom (including α/γ transitions, rotation about the χ angle and sugar repuckering), which is fundamental for an accurate description of the G-quadruplex dynamics and thermodynamics, requires timescales and extend beyond the microsecond regime. Indeed G-quadruplexes with much shorter loops, such as the human telomeric G-quadruplex[48] and require microsecond molecular dynamics over the millisecond timescale to converge, as shown for the *c-kit* promoter G4[49]. Enhanced sampling methodologies, such as metadynamics[50], enable a correct statistical mechanics description and are therefore particularly helpful to circumvent the necessity for extremely long simulations times caused by the infrequent crossing of high free energy barriers. In this work we considered various combinations of enhanced sampling methods (**Table S1**) to extensively explore the conformational space available to the G4 loop region (see **Supporting Information**). In particular, in the bias-exchange metadynamics simulations[43] (BE) simulations, which we used to determine the structural ensemble of G4 (**Fig. 1**), four dihedral angles (α, ζ, ϑ and χ) of each residue in the loop were biased using well-tempered metadynamics with a different replica for each dihedral angle. We modelled the G-quadruplex in solution using the Parmbsc1 force field[37], and the molecular dynamics simulations were carried out using Gromacs 4.6[30] (see **Supporting Information**).

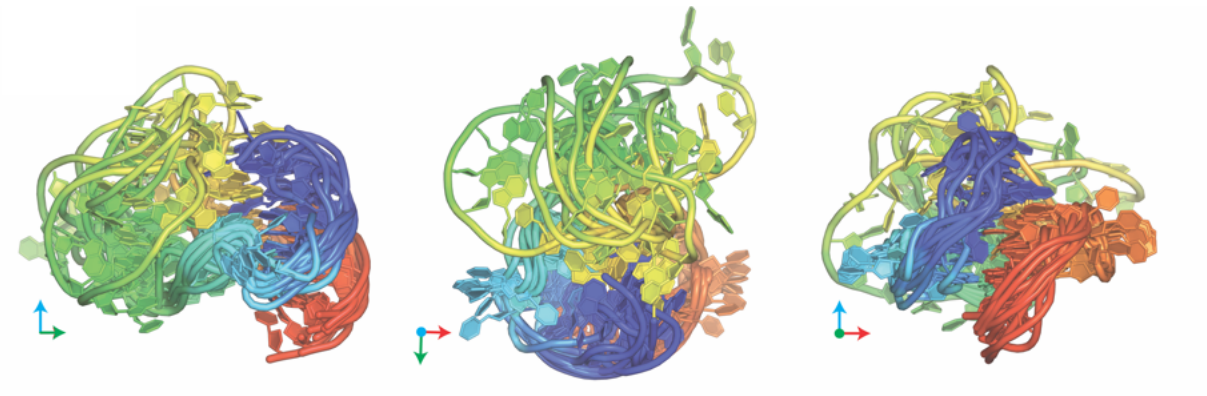


Figure 1. Structural ensemble of the G-quadruplex. We show three different views of the conformational ensemble (BE-BSC1-RDC simulations, see **Table S1**) that we determined in this work for the human CEB25 minisatellite G-quadruplex.

As structural restraints in the molecular dynamics simulations we used previously published residual dipolar couplings[40] (RDCs), since this type of NMR observable has been shown to be particularly powerful to report on the dynamics of protein and RNA molecules[21-23]. These restraints are enforced via the θ method[41], which does not require the determination of an alignment tensor, and that can be combined with enhanced sampling methods to provide a statistical thermodynamically correct description. As the simulation progresses, the replica-averaged RDC restraints are enforced, and replicas of the system with different bias are exchanged following a Monte Carlo scheme. We have also used the θ method without enhanced sampling within the minimal under-restraining minimal over-restraining framework[42] (MUMO) with eight replicas (M8), as a reference to illustrate the importance of enhanced sampling methods (**Table S1**).

G4 exhibits two separate regions in terms of dynamics

The structural ensemble that we determined illustrates the extreme robustness of the G4 core, which narrowly fluctuates around an average conformation (**Figs. 1** and **2**). By contrast, the loop regions are extremely heteromorphic, in particular the long loop between residues 10 and 16, as shown by the conformational fluctuations (**Fig. S1A**) and S^2 order parameters (**Fig. S1B**). Overall, the structural ensemble shows a much larger degree of conformational flexibility as compared to a previously set of conformers (PDB code 2LPW[20]). This result is expected, as the 2LPW set represents an uncertainty ensemble, which captures the errors in determining the average structure, while our ensemble represents a statistical ensemble, which describes the structures and corresponding statistical weights explored during the conformational fluctuations of G4[22]. For example, the C1'-N9/C1'-N1 bonds order parameters are ~0.1, 0.2 below the values obtained from the 2LPW set of conformers.

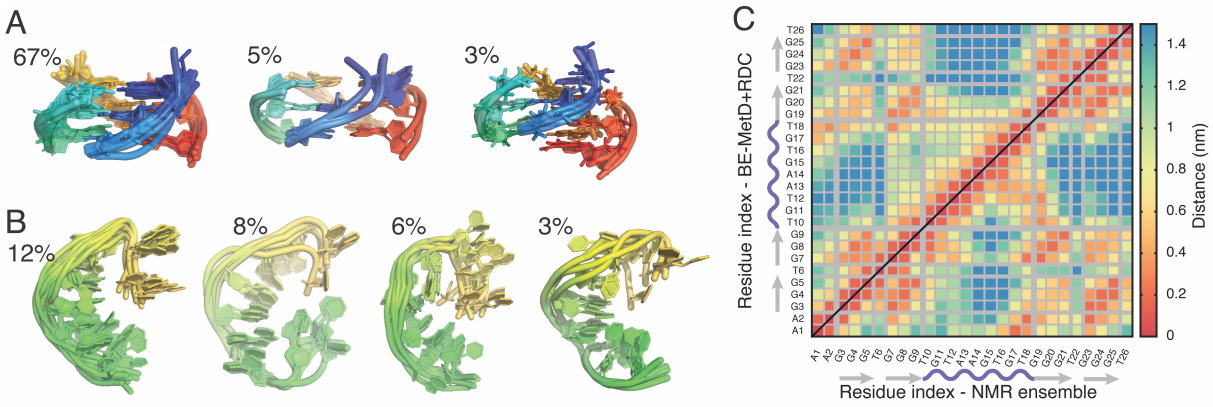


Figure 2. Most populated states in the structural ensemble of the G-quadruplex. Since the loop region (residues 10 to 18) is largely disordered (see also **Fig. S1**), we performed separate clustering analysis on different subparts of the G4; see methods for details on the clustering procedure. **(A)** Most populated clusters obtained by using the G4 core guanine tetrads. The largest cluster corresponds to a well-defined quadruplex core, and the smaller clusters show cases where either G3 is flipped out, or there is a registry change between the first and the second consecutive pairs of G-triplets (G3-G5/G7-G9 and G19-G21/G23-G25). **(B)** Most populated clusters obtained by analysis of the loop regions (residue 10-18). **(C)** Ensemble averaged contact matrix between residues of G4, computed as the average of minimum distance between pairs of residues. The upper left diagonal shows the results from the BE-MetD+RDC simulations, and the lower-right diagonal the average values from the NMR model 2LPW.

G-core. In order to visualize and quantify the structural ensemble of G4, we grouped the different structures according to their structural similarity based on their root mean square deviation (RMSD, see **Supporting Information**). The very large number of microstates that the loop region can adopt is mostly, although not entirely, insensitive of the particular arrangement of the rest of the G4. We performed, instead, a clustering analysis on three subsets of the G4: the guanine core region, the loop region, and all the G4 except the most disordered loop region (residues 10 to 16). The results of the clustering on the G4-core reveal that the most populated structure (with nearly 70% population) corresponds to a conformation in which all the guanine tetrads are perfectly formed (**Fig. 2A**). Such arrangement is clearly captured by an ensemble-averaged distance matrix, showing close contacts in off-diagonal guanine triplets (**Fig. 2C**). The next most populated clusters only represent 5% and 3% of all structures (**Fig. 2A**). They contain, respectively, structures with the G3

base flipped out from its low free energy position, and a registry change between the first and the second consecutive pairs of G-triplets (G3-G5/G7-G9 and G19-G21/G23-G25)[29, 51].The remaining 25% are scattered in relatively low populated clusters, well below 2% populations, showing small deviations from the lowest free energy conformation, which depend on the particular ion distribution around the G4.

Loop. The heterogeneity of the structural ensemble that we determined reveals that the loop region (region 10-18) lacks persistent intramolecular contacts (**Fig. 2C**, above the diagonal), some of which were present in the 2LPW set of conformers (**Fig. 2C**, below the diagonal). The number and pattern of stacked bases in the loop region is variable, with an average value of about 50% of the maximum possible (**Fig. S2**). The bending angle between adjacent residues, defined as the angle between consecutive C1' atoms, shows that on average the loop region draws a smooth curve from residues 10 to 15 (an angle of ~120 degrees) with a clear bend at residues 16 to 18 (**Fig. 3A**). The large standard deviation (~30 degrees) in the loop region, together with the histogram of bend angles, however, indicate that this averaged picture is only a part of the story, and that in fact the loop is found with marked bends in about one third of the structures (**Fig. 3B**). Taken together, these structural descriptors (the contact map and the distributions of stacked bases and of bending angles) strongly indicate that the loop region is rather disordered. Indeed, only a few conformations are found with populations above 5% (**Fig. 2B**). The largest cluster involves a well-defined stacking arrangement of residue G17 and T18, with the rest of the loop presenting around ~50% of the maximum number of stacked bases (**Fig. S2**).

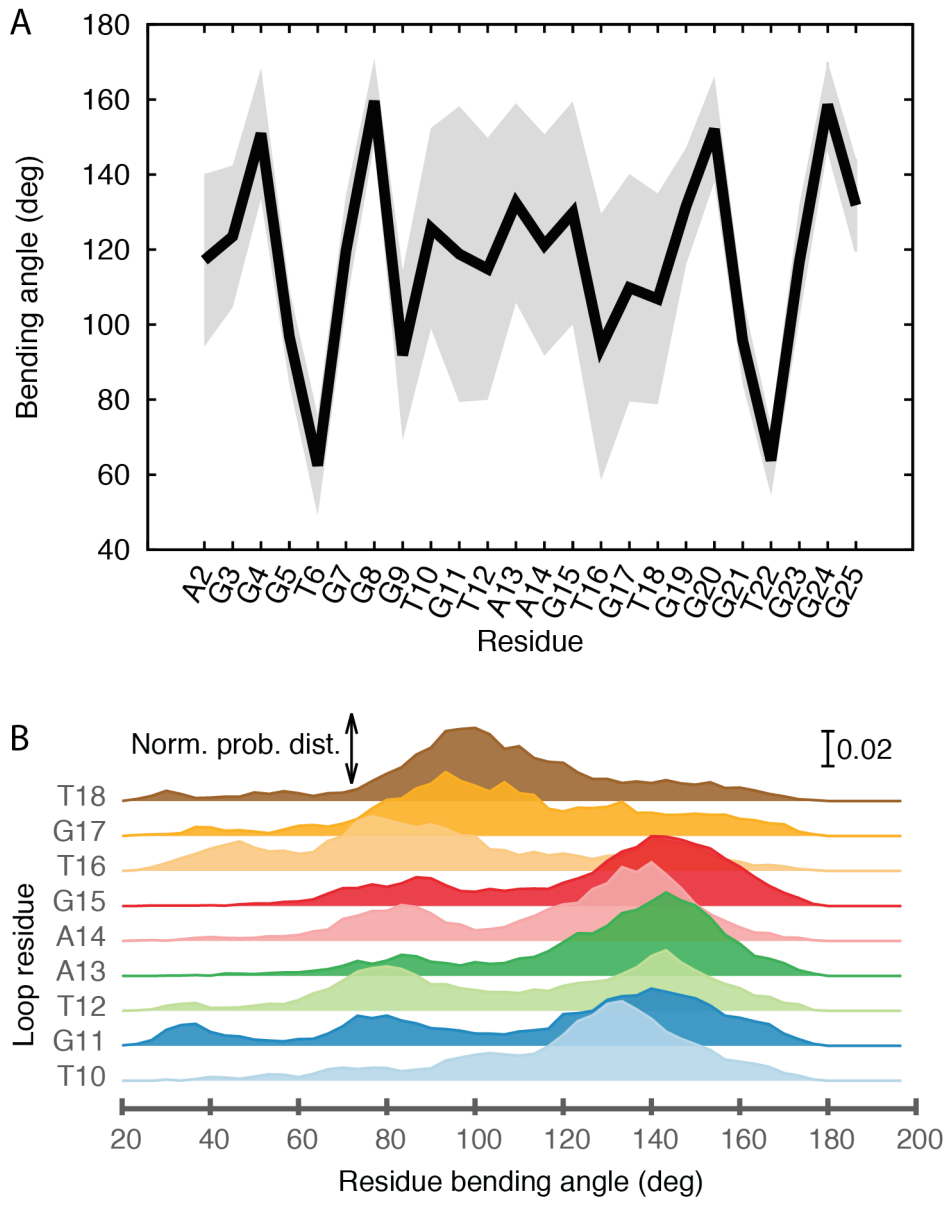


Figure 3. (A) Bending angle along the G4 sequence, defined as the angle between consecutive C1' atoms (i.e. from adjacent residues, we report the angle centered at the mid residue). The gray area around the black curve indicates the standard deviation of the mean angle. Small angles, below 120, can be considered bent. **(B)** Normalized probability distribution of bending angles for the loop region.

Correlated motions in the loop region

A principal component analysis (PCA) of the G-quadruplex ensemble reveals that the first five main collective modes mainly involve large loop-motions, which capture ~50% of the overall

variance of atomic displacements (**Fig. 4**). The first large amplitude mode describes oscillations of the loop as an almost-rigid entity with respect to the main axis of the G-quadruplex, with the loop bases almost-perfectly stacked, whereas the next three modes show motions localized in different regions of the loop (2^{nd} and 3^{rd} modes display one node, and the 4^{th} reveal three nodes in the vibration of the loop. Taken together, these results indicate that the motion of the loop is reminiscent of the harmonic vibration of a string with immobilized ends.

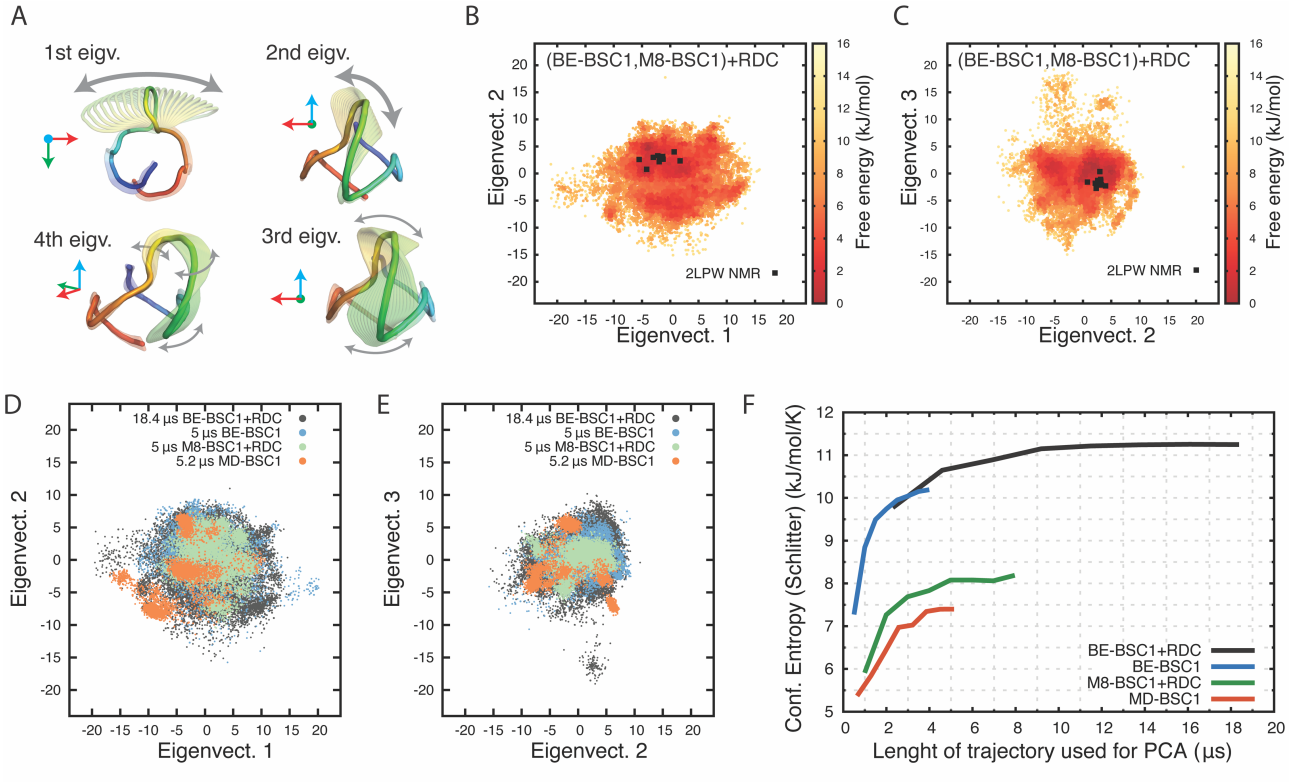


Figure 4. Principal component analysis (PCA) for the 26.4 µs of RDC-restrained molecular dynamics simulations using the parmBSC1 force field. (A) Cartoon illustration of the interpolated projections between the two extremes along the trajectory on the average structure (only the first four principal modes are shown). The first modes mainly involve motions of the loop; notably the first mode shows the collective motion of the loop with respect to the main axis of the G-quadruplet core. (**B,C**) Projection of 26.4 µs of parmBSC1+RDC-restrained simulations on the first three PC eigenvectors, colour-coded based on the associated free energy. The structures were obtained from concatenated bias-exchange RDC-restrained parmBSC1 (BE-BSC1+RDC), and parmbsc1+ ensemble averaged RDC restraints (M8-BSC1+RDC) simulations. The projection of the ten deposited NMR structures of 2LPW are overlaid as black squares. (**D,E**) Comparison of sampling along essential subspaces. PCA was performed on a trajectory containing: 18.4 µs of BE-

BSC1+RDC simulations (dark gray), 5 μs of BE-BSC1 (blue), 8 μs of M8-BSC1+RDC simulations (green), 5 μs of free molecular dynamics using parmBSC1 force field (orange). The graphs show the projection of each set of trajectories along the first 3 eigenvectors from the PCA on the concatenated trajectories. **(F)** Conformational entropy of the G4 computed via the Schlitter approximation (which involves the eigenvalues of the PCA of each trajectory), as function of the length of the simulation.

Correlations of backbone dihedrals

We analysed the circular correlations of α/γ dihedrals as a measure of the conformational freedom of the loop region. These correlations are present between α and γ dihedrals of the same residue in the case of base paired or conformationally constrained DNA (see[52] for a discussion on α/γ dihedrals correlated motion). Indeed, anticorrelations can be readily and significantly detected in the G4 core (**Fig. 5**), and are notably absent in its loop region. We observed a positive correlation for the α/γ dihedrals in residues A13 and G15, which hints at the formation of structural elements in the central region of the loop. However, the ensemble-averaged distance matrix does not show such interactions (**Fig. 2C**), and neither do our clustering calculations. These positions do, nevertheless, agree well with vibrational nodes along the second, third and fourth PCA modes involving the loop region. We also observe that measurements of other NMR observables, including J coupling constant and 31P chemical shifts may provide further validation for the results that we have presented.

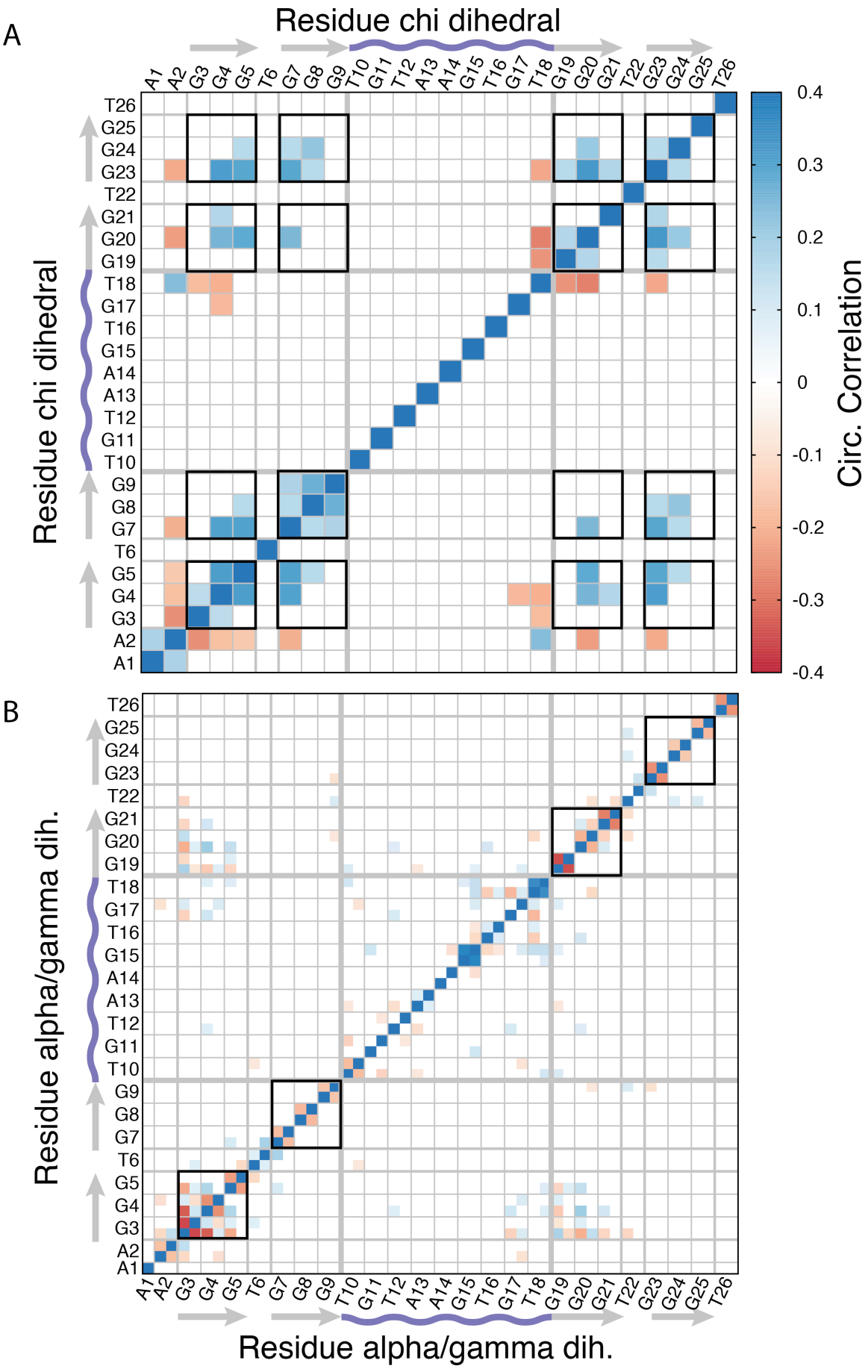

A
Residue chi dihedral
Residue chi dihedral
Circ. Correlation
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
B
Residue alpha/gamma dih.
Residue alpha/gamma dih.
A1 A2 G3 G4 G5 T6 G7 G8 G9 T10 G11 T12 A13 A14 G15 T16 G17 T18 G19 G20 G21 T22 G23 G24 G25 T26

Figure 5. Circular correlations between dihedral angles. (A) χ dihedral angles and **(B)** α and γ dihedral angles. For each pair of residues, the color map reports the value for the coefficient of correlation between a pair of dihedral angles. Only correlations with an associated P-value smaller than 0.0001 are reported (see SI Methods). The dihedral angles were obtained from 18.4 μs of the BE-BSC1+RDC simulation using a 1 ns sampling frequency. The triplets of consecutive guanine residues that form the G-quadruplex core are underlined with small gray arrows, and a wiggling deep purple line indicates the region corresponding to the loop residues. Within the color map the G4 regions are highlighted using a black square, in **(A)** the areas corresponding to correlations among different triplets of consecutive guanines, in **(B)** only those between these same triplets. Note the correlation (or anticorrelation) of adjacent α and γ dihedrals in the G-quadruplex core, and the clear correlation between residues 3-5 with residues 19-21 (diagonally opposed in the structure). Notably, these correlations between consecutive (adjacent) G-triplets are very small. The loop region is characterized by very weak adjacent inter-residue correlation, except for residue G15 and T18. Residue G15 is typically found in a steep turn towards the G-core, while residue T18 is almost part of the G-core itself, interacting with A2.

Analysis of the interactions that stabilise the G-quadruplex conformation

A well-defined arrangement of loop residues G17 and T18, together with their close-contacts between with residues A1 and A2 (**Fig. 6**), indicates that these pairs of residues extend the structured region of the G4 beyond the guanine core. We therefore performed a clustering analysis including all residues in the G4 except the loop of residues 10 to 16. The most populated clusters report mainly on differences between G17-T18 and A1-A2 interactions (**Fig. 6B**). The most populated clusters, around 50% of the overall populations, show the formation of A2-T18 rWC base pair (this base pair is present in 50% of the population), which is also present in the 2LPW set of conformers, while A1 is stacking either on top of A2 and/or T18 (**Fig. 6C**). The differences amongst the most populated clusters are in the specific arrangement of residues A1 and G17. In the largest cluster A1 is found stacking on top of both T18 and A2, with G17 stacking on top of A1. The remaining conformations mainly illustrate that A1 favourably interacts with T18-A2 base pair, even forming a low populated triplet A1-T18-A2 (5% population).

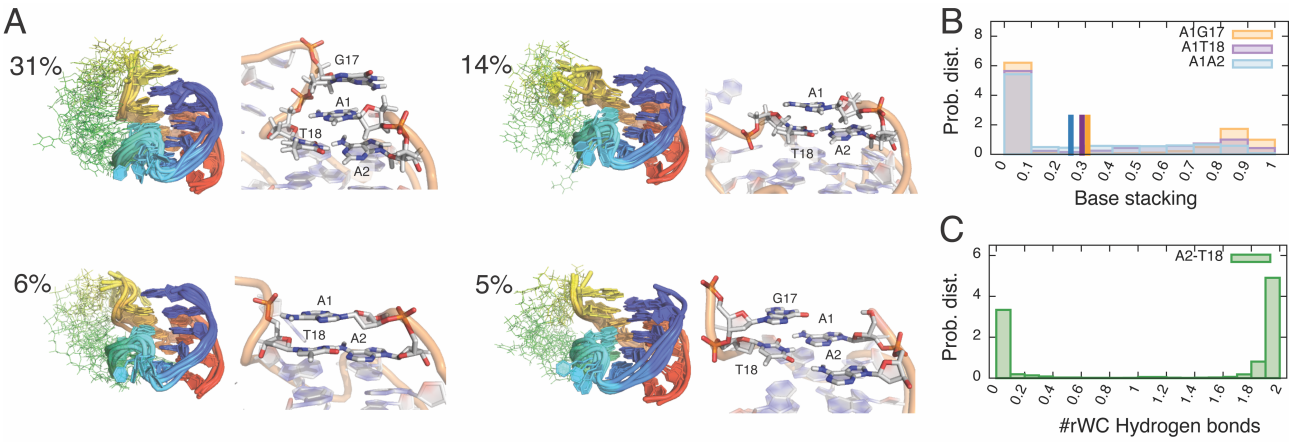


Figure 6. Analysis of the interactions stabilise the G-quadruplex conformation. **(A)** Most populated clusters obtained by using all residues except loop residues 10-16, which mainly reports differences due to the closing base pairs, exemplified by the cluster centroid in the close up shown on the r.h.s. of each cluster (5' region against residues G17 and T18). **(B)** Probability distribution of the formation of base stacking between A1G7, A1T18 and A1A2 (see the methods section for our definition of base stacking). A vertical thick line indicates the expected value of the number of stacking interactions. **(C)** Probability to form the closing reversed Hoogsteen base pair A2-T18.

Dynamics of 5'end and loop interactions and their role in G4 stability

The high thermal stability of this G4 conformation has been partially ascribed to the formation of A2-T18 base pairs, along with a non-canonical base pair between A1-G17. Indeed, thermodynamic studies reveals that removing A1 or A1 and A2, mutating them to T or mutation T18 to C reduces the melting temperature of the wild-type G4 by at least 2 degrees, whereas little to no effect occur if the mutations take place in the rest of the loop region. The non-canonical interactions between A1-G17 had been suggested based on the appearance of weak NOEs between A1(H2) and G17(H1'), A1(H2)-G17(H4'), as well as between A1(H8) and T18(CH_3).

These NOEs are well reproduced in all our molecular dynamics simulations (**Table S2**), except for a small violation of 0.02 nm occurring for A1H8-T18CH3 in the M8-BSC1-RDC ensemble, and 0.15 nm for the A1(H2)-G17(H4') in MD-BSC1. These results confirm the conclusion that even several microseconds of standard molecular dynamics simulations are not sufficient to sample the conformational space of G4. Although these NOEs are well reproduced, especially in our BE-BSC1-RDC simulations, they do not necessarily imply the formation of stable base pairs between

A1-G17. Actually, our simulations show that while the A2-T18 base pair is fully present 50% of the time, and there is substantial circular correlation between the χ angle of both bases (**Fig. 5**), the A1-G17 non-canonical base pair proposed in the NMR structures is low populated (under 2%).

However, our enhanced sampling simulations, especially BE-BSC1-RDC, reveal that the 5' end and the end of the loop region frequently interact mainly via stacking with each other in a highly dynamic way (**Fig. 2**). Indeed, A1 stacks with G17 with ~31% population, and with T18 stacking with ~29% population (**Table S3**). The base stacking justifies well the ~2 degrees difference between wild-type G4 and the m1T mutation reported[20], since the interacting energy between T and A are typically ~9 kJ/mol less favorable than between G and A (computed from the ABC-consortium database of molecular dynamics simulations of 136 unique tetranucleotides[45], see **Supporting Information**). We have as well detected one A1-T18 hydrogen bond (A1H61/H62-T18O2 or A1N1-T18H3), at least with ~50% population, which might contribute towards the stability of the G4 quadruplex. Albeit the A1-G17 stacking momentarily breaks the contacts G17 and T18, and the very dynamic nature of this region, the NOEs between the G17 and T18 are all very well reproduced in our BE-BSC1-RDC and BE-BSC1 simulations (zero violations, **Table S4**).

Conclusions

By using NMR residual dipolar couplings as structural restraints in metadynamics simulations, we have characterised the dynamics of a G-quadruplex by determining a structural ensemble at high resolution. This structural ensemble reveals the remarkable extent and complexity of the conformational fluctuations of this form of DNA, and identifies specific stacking interactions between the core and loop regions that enhance the stability of this state, which could be in principle targeted pharmacologically. We note that the strategy that we have described in this work is general and can be applied to other conformational states of nucleic acids, especially those characterised by the presence flexible single strands for which the structural data available are very scarce.

Acknowledgements

Funding for this work was provided by the Centre for Misfolding Diseases, University of Cambridge, UK.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: XXX. Five supporting tables and six supporting figures (PDF).

References

1. Aviñó, A.; Portella, G.; Ferreira, R.; Gargallo, R.; Mazzini, S.; Gabelica, V.; Orozco, M.; Eritja, R. (2014) Specific loop modifications of the thrombin-binding aptamer trigger the formation of parallel structures. *FEBS J. 281* (4), 1085-1099.

2. Saneyoshi, H.; Mazzini, S.; Aviñó, A.; Portella, G.; González, C.; Orozco, M.; Marquez, V. E.; Eritja, R. (2009) Conformationally rigid nucleoside probes help understand the role of sugar pucker and nucleobase orientation in the thrombin-binding aptamer. *Nucl. Acids Res. 37* (17), 5589-601.

3. Smargiasso, N.; Rosu, F.; Hsia, W.; Colson, P.; Baker, E. S.; Bowers, M. T.; De Pauw, E.; Gabelica, V. (2008) G-quadruplex DNA assemblies: loop length, cation identity, and multimer formation. *J. Am. Chem. Soc. 130* (31), 10208-16.

4. Neidle, S.; Balasubramanian, S. (2006) *Quadruplex Nucleic Acids*. RSC Publishing: Cambridge, p 301.

5. Phan, A. T. (2010) Human telomeric G-quadruplex: structures of DNA and RNA sequences. *FEBS J. 277* (5), 1107-17.

6. Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem. 5* (3), 182-6.

7. Burge, S.; Parkinson, G. N.; Hazel, P.; Todd, A. K.; Neidle, S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucl. Acids Res. 34* (19), 5402-15.

8. Rhodes, D.; Lipps, H. J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucl. Acids Res. 43* (18), 8627-37.

9. Siddiqui-Jain, A.; Grand, C. L.; Bearss, D. J.; Hurley, L. H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. USA 99* (18), 11593-8.

10. Bochman, M. L.; Paeschke, K.; Zakian, V. A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet. 13* (11), 770-80.

11. Balasubramanian, S.; Hurley, L. H.; Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Disc. 10* (4), 261-75.

12. Neidle, S.; Read, M. A. (2001) G-quadruplexes as therapeutic targets. *Biopolymers 56* (3), 195-208.

13. Read, M.; Harrison, R. J.; Romagnoli, B.; Tanious, F. A.; Gowan, S. H.; Reszka, A. P.; Wilson, W. D.; Kelland, L. R.; Neidle, S. (2001) Structure-based design of selective and potent G quadruplex- mediated telomerase inhibitors. *Proc. Natl. Acad. Sci. USA. 98* (9), 4844-4849.

14. Neidle, S. (2016) Quadruplex nucleic acids as novel therapeutic targets. *J. Med. Chem. 59* (13), 5987-6011.

15. Hansel-Hertsch, R.; Di Antonio, M.; Balasubramanian, S. (2017) DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat. Rev. Mol. Cell Bio. 18* (5), 279-284.

16. Collie, G. W.; Campbell, N. H.; Neidle, S. (2015) Loop flexibility in human telomeric quadruplex small-molecule complexes. *Nucl. Acids Res. 43* (10), 4785-4799.

17. Ma, D. L.; Ma, V. P. Y.; Chan, D. S. H.; Leung, K. H.; Zhong, H. J.; Leung, C. H. (2012) In silico screening of quadruplex-binding ligands. *Methods 57* (1), 106-114.

18. Cosconati, S.; Marinelli, L.; Trotta, R.; Virno, A.; De Tito, S.; Romagnoli, R.; Pagano, B.; Limongelli, V.; Giancola, C.; Baraldi, P. G.; Mayol, L.; Novellino, E.; Randazzo, A. (2010) Structural and conformational requisites in DNA quadruplex groove binding: another piece to the puzzle. *J. Am. Chem. Soc. 132* (18), 6425-33.

19. Cosconati, S.; Marinelli, L.; Trotta, R.; Virno, A.; Mayol, L.; Novellino, E.; Olson, A. J.; Randazzo, A. (2009) Tandem Application of Virtual Screening and NMR Experiments in the discovery of brand new DNA quadruplex groove binders. *J. Am. Chem. Soc. 131* (45), 16336-16337.

20. Amrane, S.; Adrian, M.; Heddi, B.; Serero, A.; Nicolas, A.; Mergny, J.-L.; Phan, A. T. (2012) Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite. *J. Am. Chem. Soc. 134* (13), 5807-5816.

21. Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. (2005) Simultaneous determination of protein structure and dynamics. *Nature 433* (7022), 128-32.

22. Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. (2017) Principles of protein structural ensemble determination. *Curr. Op. Struct. Biol. 42*, 106-116.

23. Borkar, A. N.; Bardaro, M. F., Jr.; Camilloni, C.; Aprile, F. A.; Varani, G.; Vendruscolo, M. (2016) Structure of a low-population binding intermediate in protein-RNA recognition. *Proc. Natl. Acad. Sci. USA. 113* (26), 7171-6.

24. Bottaro, S.; Bussi, G.; Kennedy, S. D.; Turner, D. H.; Lindorff-Larsen, K. (2018) Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. *Sci. Adv. 4* (5), eaar8521.

25. Bergonzo, C.; Hall, K. B.; Cheatham, T. E., 3rd (2016) Divalent ion dependent conformational changes in an RNA stem-loop observed by molecular dynamics. *J. Chem. Theory Comput. 12* (7), 3382-9.

26. Bochicchio, A.; Krepl, M.; Yang, F.; Varani, G.; Sponer, J.; Carloni, P. (2018) Molecular basis for the increased affinity of an RNA recognition motif with re-engineered specificity: A molecular dynamics and enhanced sampling simulations study. *PLoS Comput. Biol. 14* (12), e1006642.

27. Sponer, J.; Bussi, G.; Krepl, M.; Banas, P.; Bottaro, S.; Cunha, R. A.; Gil-Ley, A.; Pinamonti, G.; Poblete, S.; Jurecka, P.; Walter, N. G.; Otyepka, M. (2018) RNA Structural dynamics as captured by molecular simulations: A comprehensive overview. *Chem. Rev. 118* (8), 4177-4338.

28. Islam, B.; Stadlbauer, P.; Krepl, M.; Havrila, M.; Haider, S.; Sponer, J. (2018) Structural dynamics of lateral and diagonal loops of human telomeric G-quadruplexes in extended MD simulations. *J. Chem. Theory Comput. 14* (10), 5011-5026.

29. Stadlbauer, P.; Krepl, M.; Cheatham, T. E., 3rd; Koca, J.; Sponer, J. (2013) Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucl. Acids Res. 41* (14), 7128-43.

30. Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics 29* (7), 845-854.

31. Darden, T.; York, D.; Pedersen, L. (1993) Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys. 98*, 10089-10092.

32. Miyamoto, S.; Kollman, P. A. (1992) SETTLE: An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comp. Chem. 13*, 952-962.

33. Hess, B. (2008) P-LINCS: A parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput. 4* (1), 116-122.

34. Bussi, G.; Donadio, D.; Parrinello, M. (2007) Canonical sampling through velocity rescaling. *J Chem. Phys. 126* (1), 014101.

35. Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys. 81*, 3684-3690.

36. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins 65(3)*, 712-725.

37. Ivani, I.; Dans, P. D.; Noy, A.; Perez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goni, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Gelpi, J. L.; Gonzalez, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. (2016) PARAMBSC1: a refined force field for DNA simulations. *Nat. Methods 13* (1), 55-58.

38. Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. (1987) The missing term in effective pair potentials. *J. Phys. Chem. 91* (24), 6269-6271.

39. Smith, D. E.; Dang, L. X. (1994) Computer simulations of NaCl association in polarizable water. *J. Chem. Phys. 100* (5), 3757-3766.

40. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. (2008) BioMagResBank. *Nucl. Acids Res. 36* (suppl 1), D402-D408.

41. Camilloni, C.; Vendruscolo, M. (2014) A tensor-free method for the structural and dynamical refinement of proteins using residual dipolar couplings. *J. Phys. Chem. B 119(3)*, 8225-8226.

42. Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR 37* (2), 117-35.

43. Piana, S.; Laio, A. (2007) A bias-exchange approach to protein folding. *J. Phys. Chem. B 111* (17), 4553-9.

44. Schlitter, J. (1993) Estimation of absolute and relative entropies of macromolecules using the covariance matrix. *Chem. Phys. Lett. 215*, 617-621.

45. Lavery, R.; Zakrzewska, K.; Beveridge, D.; Bishop, T. C.; Case, D. a.; Cheatham, T.; Dixit, S.; Jayaram, B.; Lankas, F.; Laughton, C.; Maddocks, J. H.; Michon, A.; Osman, R.; Orozco, M.; Perez, A.; Singh, T.; Spackova, N.; Sponer, J. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucl. Acids Res. 38* (1), 299-313.

46. Jammalamadaka, J. R.; SenGupta, A. (2001) *Topics in Circular Statistics*. World Scientific Publishing Co.: London.

47. Cavalli, A.; Camilloni, C.; Vendruscolo, M. (2013) Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys. 138* (9), 094112.

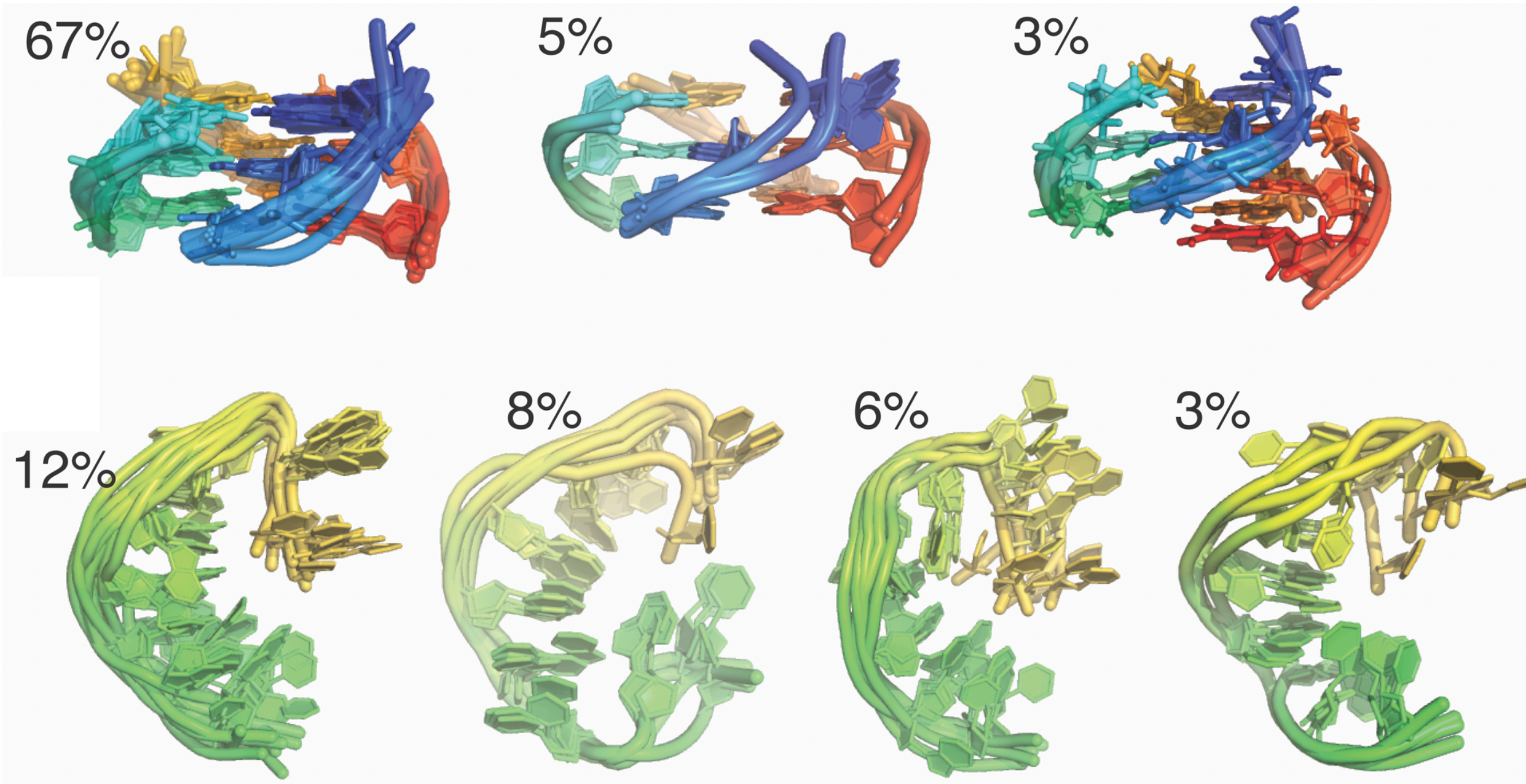
48. Islam, B.; Sgobba, M.; Laughton, C.; Orozco, M.; Sponer, J.; Neidle, S.; Haider, S. (2013) Conformational dynamics of the human propeller telomeric DNA quadruplex on a microsecond time scale. *Nucl. Acids Res. 41* (4), 2723-2735.

49. Islam, B.; Stadlbauer, P.; Krepl, M.; Koca, J.; Neidle, S.; Haider, S.; Sponer, J. (2015) Extended molecular dynamics of a c-kit promoter quadruplex. *Nucl. Acids Res.* 43(18), 8673-8693

50. Laio, A.; Parrinello, M. (2002) Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA. 99* (20), 12562-12566.

51. Limongelli, V.; De Tito, S.; Cerofolini, L.; Fragai, M.; Pagano, B.; Trotta, R.; Cosconati, S.; Marinelli, L.; Novellino, E.; Bertini, I.; Randazzo, A.; Luchinat, C.; Parrinello, M. (2013) The G-triplex DNA. *Angew. Chem. Int. Ed. 52* (8), 2269-73.

52. Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham 3rd, T. E.; Laughton, C. A.; Orozco, M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Bioph. J. 92* (11), 3817-3829.


67%
5%
3%
12%
8%
6%
3%