Telomeric G-quadruplex Intermediates unveiled by Complex Markov Network Analysis.

A. Sáinz-Agost[1,2], F. Falo[1,2], and A. Fiasconaro[1,2,*]

[1]Departamento de Física de la materia condensada, Universidad de Zaragoza, 50009, Zaragoza, Spain
[2]Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, 50009, Zaragoza, Spain
[*]Corresponding author: afiascon@unizar.es

ABSTRACT

G-quadruplexes are secondary, non-canonical RNA/DNA structures formed by guanine-rich sequences assembled into four-stranded helical structures by the progressive stacking of G-Tetrads, planar arrangements of guanines stabilised by monovalent ions such as K^+ or Na^+. Their stability plays a very important role in the prevention of DNA degradation, leading to the promotion or inhibition of specific biological pathways upon formation. In this work, we explore the occurrences of intermediates originating from the unfolding of these structures by using all-atom simulations, analyzing a small number of significant reaction coordinates to follow the evolution of the system by applying a mesoscopic simplification of the structures followed by two different dimensionality reduction techniques: Principal Component Analysis (PCA) and time-Independent Component Analysis (tICA). The data of the reduced trajectories are then encoded into a Complex Markov Network which, in conjunction with an Stochastic Steepest Descent, provides a hierarchical organization of the different nodes into basins of attraction. This procedure is able to reveal the main intermediates and the most relevant transitions the system undergoes in its denaturation path.

1 Introduction

G-quadruplexes (G4s) are secondary, non-canonical structures arising from either one or multiple guanine rich DNA or RNA chains, and are very common in the human genome[1,2]. These conformations emerge from the gradual stacking of *G-tetrads*, planar arrangements of four guanines stabilized by Hoogsteen hydrogen bonds[3–5]. The stacking process is further aided by the presence of either monovalent or divalent cations along the central channel these tetrads define[6], with typically K^+ or Na^+ being involved.

The relevance and recent interest in these structures comes from their biological function. G-quadruplexes have an important role in DNA stability[7] as well as in regulation processes[8–10], the latter associated to either the over- or under-exposure of different binding sites of interest upon adoption of a G-quadruplex topology. The presence of these structures at the moment of DNA replication can lead to the interruption of this process, destabilizing the helicases[11] and, combining with other factors, can lead to epigenetic instability[12]. Similarly, overabundance of guanine in mRNA can lead to difficulties in protein translation in the ribosome. Several works have investigated their gene promotion and repression capabilities[13,14], as well as its relevance as a potential therapeutic binding site for cancer treatment, given that the telomeric regions of the chromosomes have a tendency to form these structures[15,16].

G-quadruplexes can adopt diverse topologies[3], primarily influenced by two key factors: the number of single-stranded DNA/RNA chains participating in their formation (from either a single or several molecules, the latter called intermolecular G4s) and the particular twists and turns present in their backbones, dividing them into parallel (all guanine strands oriented in the same direction), antiparallel (neighboring strands oriented in opposite directions) or hybrid. Our research focuses specifically on unimolecular quadruplexes, formed by a single stranded DNA (ssDNA) chain. Within this context, we examined one of the multiple structures available: the *parallel* G-quadruplex, which has all guanine tracts are oriented in the same direction, in which the loops connect the top of a guanine track with the bottom of the next. This conformation is depicted on Fig. 1.

The mechanical stability of G-quadruplexes under force has been extensively documented[6,17–19]. Thus, the studies in this field have turned towards the study of their thermodynamical properties[20,21]. In particular, the observation and subsequent characterization of different conformations and folding intermediates of G-quadruplexes has become an object of large interest, due to the involvement of these structures in genetic regulation.[22].

The goal of this work is to focus into the thermal unfolding of G-quadruplexes and investigate the presence of intermediate structures in the unfolding pathway. The stability of the G4 is affected by several factors both external, such as ionic concentration[23,24], presence of other elements in the solvent[25,26] and bath temperature, or internal, such as strand orientation (parallel, antiparallel or hybrid), length and structure of the loops[24,27,28], and chain rigidity[18].

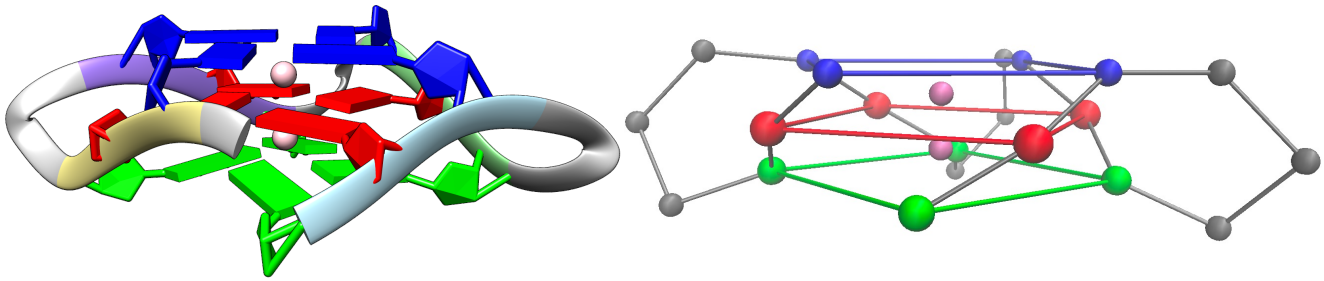

Figure 1. On the left, parallel conformation sourced from the PDB (1KF1, ribbon representation in Chimera). On the right, the resulting structure after the coarse graining procedure. The different G-tetrads on the structures are marked in different colors, the loops in grey and the ions in pink. As concerns the PDB structure, the different guanine tracts are also represented in homogeneous colors: yellow (first tract, 5′-end), purple (second tract), lime (third tract) and ice blue (fourth tract, 3′-end), respectively.

Molecular Dynamics (MD) simulations have proven to be a valid and powerful tool in examining both the thermal and mechanical stability of biomolecular structures, also G-quadruplexes, revealing complex energy landscapes and characterizing their folding kinetics[29,30]. In the context of MD, we have used the *Replica Exchange Molecular Dynamics* (REMD)[31,32] method to study the evolution of our system through the parallel simulation of 8 copies of the human parallel G-quadruplex (PDB: 1KF1)[33] in a temperature range close to its reported melting point (65°), observing an unfolding event in only one of the eight replicas. To efficiently analyze the resulting high-dimensional data, a coarse-graining approach based on a consolidated mesoscopic model[17,18] was applied, followed by the dimensionality reduction techniques, specifically Principal Component Analysis (PCA)[34,35] and time-lagged Independent Component Analysis (tICA)[36]. These techniques allowed us to reduce the complexity of the data and retain only the most relevant information about the system. Using Complex Markov Networks (CMN)[37–40] combined with a stochastic steepest descent algorithm, we constructed a series of networks describing the unfolding process. These networks outlined the main conformations the system adopts as it unfolds, identifying both stable and transitory intermediates. The description provided by tICA proved to be clearer than that of PCA, sometimes identifying a temporal evolution of the denaturation pathway.

This work is structured as follows: in Section 2 we elaborate on the simulation setup, the different dimensionality reduction approaches, and the description and construction of the Complex Markov Networks. Section 3 contains a discussion of the simulation findings, the effect of the presence of the cations inside the G4 structure on the unfolding dynamics, the results arising from the application of the two dimensionality reduction procedures considered, as well as the resulting CMNs describing the unfolding. The final conclusions are contained in Section 4. Additional information regarding both methods and results can be found in the Supplementary Material associated to this publication.

2 Methods.

2.1 All-atom simulations with Gromacs.

The initial structure for the simulation, the intramolecular parallel DNA G-quadruplex (1KF1)[33], formed by GGGTTA repeat units, as 5′-AGGGTTAGGGTTAGGGTTAGGG-3′, was retrieved from the Protein Data Bank (PDB).

The simulations were carried out using the software Gromacs[41,42], with the help of the force field amber modification Parmbsc0[43], which has been used for both DNA[44] and G-quadruplex analysis[6] and has been proven to be the best option for our calculations[45], although other options, such as OL15[46] or Parmbsc1[47], could also be considered. Control measures to verify the stability of the complexes in normal conditions under the force field of interest were carried out, obtaining the expected results.

The structures were solvated using the Tip3P water model[48] in a periodic cubic box 1 nm larger than the diameter of the DNA body. Afterwards, more potassium K^+ ions were introduced in the simulation box to counteract the negative charges present in the DNA chains, resulting in an average potassium concentration of 0.14 M, with no negative ions added to the mix.

The energy of the solvent was minimized using a steepest descent algorithm for up to 50000 iterations, leaving the G4s frozen. This was followed by equilibration in the NVT and NPT ensembles, both lasting for 100 ps using a leap-frog integrator, with a Berendsen thermostat and a Parrinello-Rahman barostat. The electrostatic interactions were treated via the particle mesh Ewald method, establishing its cutoff, as well as the one for the Lennard-Jones interactions, at 1 nm. All simulations performed used a time step of 2 fs, and simulation data was recorded every 20 ps.

To observe the unfolding intermediates of the structures, our simulations have been performed at temperatures near the melting point of the parallel G4, being slightly higher than 65°C[49]. At these temperatures, the unfolding kinetics are notoriously

slow.

To increase the unfolding probability of the G4s and avoid that the system may get stuck in a metastable state[50], we applied in our simulations the *Replica Exchange Molecular Dynamics* (REMD) method, also called Parallel Tempering[31,32]. This methodology consists in running concurrent simulations of multiple copies of the entire system (structure, solvent and ions), each of them at different temperatures. We refer to the simulation boxes containing a copy of the system at a given temperature as *replicas*. Once the simulation is started, after a sufficient time lapse (below described)[51–53], attempts are made to exchange the content between replicas at different temperatures, re-scaling the velocities of the particles in the process. The probability of such exchanges is determined by a prescribed algorithm, in our case the Metropolis criterion[54]. This interchange of conformations at diverse temperatures aids the system in overcoming high energy barriers.

We constructed eight replicas, with assigned temperatures exceeding the nominal melting point of the structures (65°C). This adjustment was made in consideration of the ionic strength found in our simulations, which is higher than that of the experimental measurements, a fact that has been shown to increase the melting temperature T_m[23]. The concrete values were in the interval $[343,\ 345,\ \ldots, 357]$ K.

As refers to the exchange time between replicas, we followed the criteria reported in[51–53], which indicates the autocorrelation decay time of the potential energy of the whole system as the minimum exchange time. To estimate this quantity, we performed specific simulations at the lowest temperature (343 K), which presents the larger autocorrelation time, finding it to be approximately equal to 50 ns. To ensure a successful exploration of the energy landscape, we considered an exchange time of 0.2 μs. The evolution of the potential energy as well as the fit to an exponential function can be found in the Supplementary Material.

2.2 Mesoscopic discretization of the model.

The aim of this work is to extract the main conformations the human parallel G-quadruplex adopts along its unfolding trajectory. The method employed for the analysis relies partially in the diagonalization of different correlation matrices, whose dimension corresponds to the number of coordinates involved in the description of the all-atom trajectories of these structures, as explained in later paragraphs. In order to reduce the size of these matrices as well as to only focus on the conformations that lead to significant changes in the backbone of the G-quadruplexes, we decided to coarse-grain the structures before this analysis.

Specifically, the coarse-graining procedure, consistent with prior investigations in our research group[17,18], involved the removal of water and ions from the simulation box, retaining only the DNA bases, which were subsequently reduced to unique identical beads. The positions of these beads were determined as the center of mass for the guanines and the phosphate position for the loops. The transformation of gromacs coordinates into a coarse-grained system was executed using Python 3.9.18 in conjunction with the MDAnalysis package[55,56]. The resulting structure is depicted on the right side of Fig. 1.

2.3 Dimensionality reduction techniques.

Following the coarse-graining, our system now comprises 21 beads, each characterized by its position along the three Cartesian axes, resulting in a total of 63 degrees of freedom. Extensive research[57,58] has demonstrated that the majority of these degrees of freedom do not contribute significantly to our understanding of the system. Instead, the essential aspects can be effectively captured using a reduced set of reaction coordinates derived from combinations of the original ones.

In this study, we employ two distinct methods to diminish the system's dimensionality while retaining a substantial amount of relevant information. These methodologies are recognized as "Principal Component Analysis" (PCA) and "time-Independent Component Analysis." (tICA).

2.3.1 Principal Component Analysis

PCA is a dimensionality reduction procedure which produces a series of orthogonal reaction coordinates from linear combinations of the input data $x_i,\ i = 1,\ \ldots 63$. Its aim is to retain and explain as much of the variance of the original data as possible. This procedure, first introduced in[34,35], relies on the solution of an eigenvalue problem involving the covariance matrix of the input data

$$\mathbf{C}(0)\mathbf{V} = \mathbf{\Lambda}\mathbf{V}, \tag{1}$$

where $\mathbf{C}(0)$ is the covariance matrix $C(0)_{ij} \propto \sum_t x_i(t)x_j(t)$ of the normalized data (mean 0, unit variance), $\mathbf{V}$ contains the eigenvectors v_i by columns, and $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix. The *principal components* are therefore obtained by projecting our original data into the different eigenvectors. A more detailed look into the derivation of the method can be found in the Supplementary Material.

The relative magnitude of the resulting eigenvalues is a measure of the proportion in which the original variance is projected onto the corresponding PC. Thus, depending on the distribution of the eigenvalues λ_i, a subset of coordinates $n \leq N$ can be selected for the ongoing system description. The appropriate value for n corresponds to a large reduction in the magnitude of

the eigenvalue λ_{n+1} when compared to λ_n, meaning that most of the relevant information of the system (which corresponds to the majority of the variance) resides within the first n degrees of freedom.

In our specific context, PCA is employed to eliminate extraneous degrees of freedom while retaining those that contribute to insights into conformational changes within the G4 structure. It is crucial to note that this assumption holds true only when these transitions represent the most significant alterations in the variance of the system, a condition that may not always be met. To selectively filter changes based on their kinetics and retain only those with a timescale exceeding a predefined threshold, we employ tICA, here below presented.

2.3.2 time-Independent Component Analysis

tICA is a method, first introduced in[36], which produces a series of reaction coordinates from linear combinations of the input data that maximise not the variance, but the autocorrelation of the projected data between times t and $t+\tau$, with τ being a time window selected by the user, known as the *lag time*. The mathematical framework of tICA is similar in its nature to that one of PCA, and is presented in the Supplementary Material.

The method relies on the resolution of a generalized eigenvalue problem involving the covariance matrix of the data and the time-lagged correlation matrix $C(\tau)_{ij} \propto \sum_t x_i(t)x_j(t+\tau)$, similar to the former but correlating the data at times t and $t+\tau$:

$$\mathbf{C}(\tau)\mathbf{W} = \mathbf{C}(0)\mathbf{\Gamma}\mathbf{W}, \tag{2}$$

where $\mathbf{\Gamma}$ is the diagonal eigenvalue matrix, and $\mathbf{W}$ the eigenvectors matrix, with each of the different w_i vectors as columns. This general equation (Eq. (2)) is typically not solvable due to the small value of the determinant of the matrices involved, leading to numerical errors in the calculations. The AMUSE algorithm[59] is typically used in its place, detailed in the Supplementary Material.

It has been shown[60] that the eigenvalues γ_i correspond to the value of autocorrelation of the i-th component and that the cross-correlation between the i-th and the j-th coordinates vanishes at the lag time $t = \tau$. Furthermore, if we assume the autocorrelation of the signal to have an exponential decay, the associated constant is given by the lag time and the eigenvalue as:

$$t_i = -\frac{\tau}{\log|\gamma_i|} \tag{3}$$

Therefore, the value of the lag-time acts as a threshold for the time-scales detected in our system: if a certain timescale is bigger than τ it will be detected and included in our coordinates, otherwise it will be discarded, since its autocorrelation has had time to decay to 0: it is a kinetic filter. With this property in mind, tICA should be able to distinguish important conformational changes and intermediates that PCA could not consider if they are not affected by a remarkable variance value.

2.4 Conformational Markov Network and conformational basins.

Conformational Markov Networks (CMNs) are a tool that has been previously introduced to study the Free Energy Landscape of a diverse array of physical systems[37–40]. It consists in building a series of nodes connected via directed links.

For the construction of a CMN, each of the n reduced coordinates, obtained from either PCA or tICA, are discretized into m intervals. Every possible combination of the intervals of the coordinates that has been occupied (i.e. the trajectory of our real system has been in that precise combination at an arbitrary time t) will constitute a *node* in our system. Therefore, we have a number of nodes $N_{\text{nodes}} \leq m^n$. The weight of each node is given by the number of times its associated combination of intervals is visited by the trajectory, normalized with the total number of time frames the trajectory is divided in. The links between them are directed and described by P_{ij}, denoting the transition probability from node j to node i, normalized such that $\sum_i P_{\text{ij}} = 1$.

The resulting network size can become challenging to analyze depending on the chosen number of intervals m, which is directly correlated to the number of nodes. In our specific case, $n = 3$ coordinates and $m = 10$ intervals were chosen, resulting in up to 1000 distinct nodes. To reduce this number and eliminate potential redundancies where different nodes essentially represent the same conformation, a Stochastic Steepest Descent algorithm was employed. This algorithm, described in detail in Ref[37], groups the nodes into basins, which represent different attractors of the trajectory. Additionally, a filtering based on a cutoff of the weight of the nodes has been introduced with the goal to retain only fundamental basins, and eliminating the ones less representative ($P_i < 10^{-5}$).

3 Results.

3.1 Direct GROMACS simulations.

One of the magnitudes we investigated in our study is the *Root Mean Square Deviation* (RMSD) of the G-quadruplex structure: the difference at each time between the G4 positions and the equivalent native configuration, calculated along all replica's trajectories. This measure is the simplest way to indicate how much the structure changes during the time evolution, and it has been calculated both for the complete G4-system and the guanine tetrads only (*i.e.* the piled guanine structure without

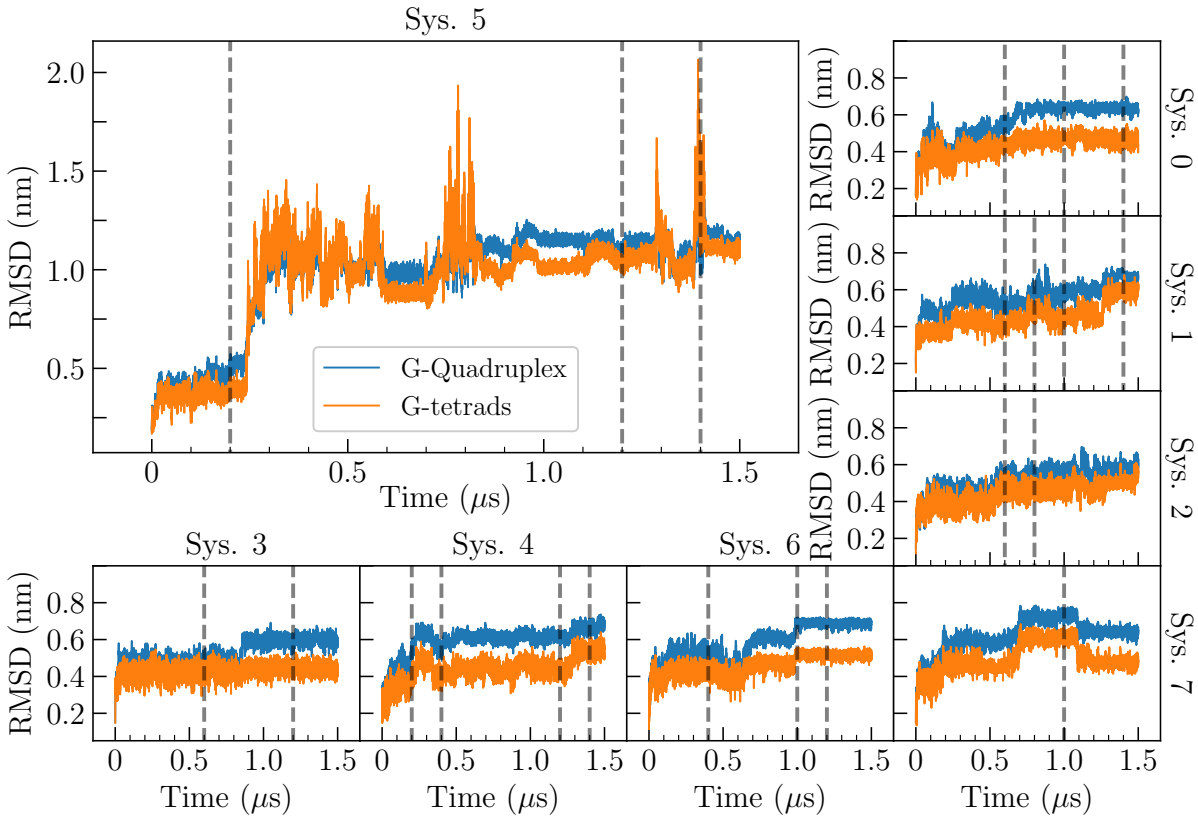

Figure 2. RMSD calculated over the different starting configurations (Sys.) of the parallel G-quadruplexes. In blue, the RMSD of the whole structure. In orange, the RMSD taking into account only the guanines forming the planar arrangement. The vertical dashed lines correspond to successful exchanges between replicas. Sys. 5, since its the one experiencing unfolding and used for the analysis, has been enhanced.

the external loops), the latter being the most important guideline to confirm that changes in the RMSD actually correspond to unfolding processes. In fact, a significant change in loop conformation could lead to an increase in the total RMSD, while not necessarily being associated to the denaturation of the G-quadruplex. Thus the confirmation of the unfolding relies on the G-tetrads structure only.

This information is contained in Fig. 2, that shows both the whole G-quadruplex and G-tetrads. To display the RMSD values properly, we used the following analysis: instead of tracking the behavior of a replica whose contents are continuously changing over time, we analyze the evolution of a specific G4 structure as it passes through between replicas. This approach prevents the misinterpretation of changes in RMSD that are actually due to exchanges between replicas as real unfolding events. We refer to these trajectories as *Systems* (Sys.), as they represent the same nucleic structure over all the replicas. The successful exchanges between replicas, when a Sys. changes temperature, are marked by gray vertical lines in Fig. 2. The RMSD curves focusing on the replicas can be found in the Supplementary Material, corresponding to Supplementary Figs. 1 & 2.

All but one of the Sys. present an RMSD that quickly stabilizes at either approximately 0.6 nm or 0.4 nm, depending either on the inclusion (complete G4, blue line in the panels) or absence (G-tetrads, orange line) of the loops in the calculation, respectively. Sys. 5 undergoes great deviations from the norm, reaching values up to 2 nm independently of the loops, thus confirming an unfolding event. These results further emphasize the correctness of the choice of the REMD method; in fact, the unfolding processes of the G-quadruplex have an average lifetime typically larger than the scope of the simulations, thus making them difficult to observe without the use of replicas at different temperatures.

The other G4 structures remain in a relatively stable configuration, though some of them present the loss of one ion from the central channel, occurrence described in the next Section 3.1.1.

These same results can be understood through a second metric, the *radius of gyration* (R_g), which measures the average square distance of the monomers respect to the center of mass of the structure. Fig. 3 shows the values of (R_g), as a function of time, where we can observe its low variability in the majority of systems, while a rapid and sudden deviation with respect to the stable value occurs in Sys. 5, indicating the unfolding of the structure already seen in Fig. 2.

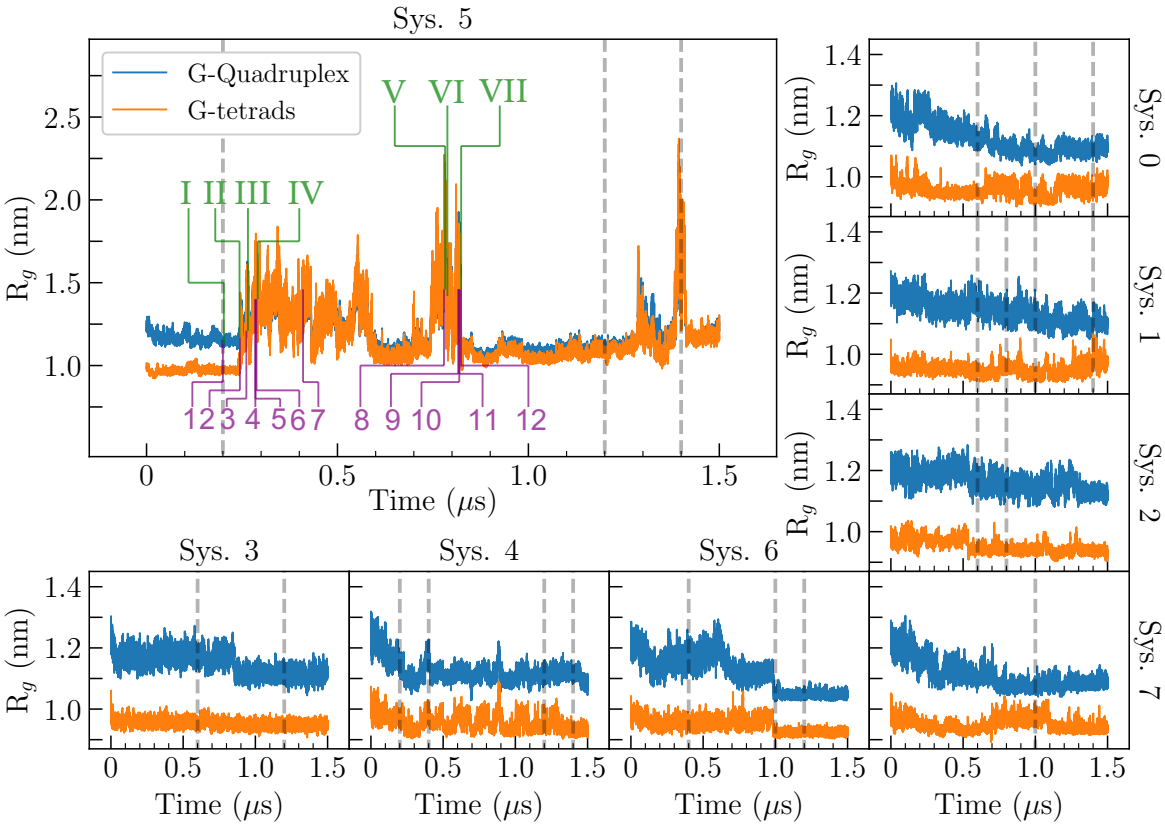

Figure 3. Radius of gyration computed over the different starting configurations (Sys.) of the parallel G-quadruplex. In blue, the R_g of the whole structure. In orange, the R_g for the G-tetrads. Sys. 5 has been enhanced, since it is the one considered for posterior analysis due to the presence of unfolding. The vertical grey dashed lines correspond to successful exchanges between replicas.

3.1.1 Effect of the monovalent ions in the stability of the system.

The stability of G-quadruplex structures is highly influenced by the presence or absence of ions, typically monovalent, within the central channel[6]. These ions, which stably reside in the G4 because of their positive electrostatic interactions with the negatively charged G-tetrads and the spatial arrangement of guanines in the native conformation, may escape if a structural conformation with a sufficiently wide central channel occurs.

The replica analysis is once again inverted, focusing on a single structure as it traverses between replicas to prevent confusing exchanges between replicas with ion loss events. To characterize the ions' position, two coordination numbers were employed: one for the site corresponding to the top and intermediate planes, denoted as Co_1, and the other for that of the intermediate and bottom planes, denoted as Co_2. These metrics quantify the number of bonds formed between each ion (filtered according to the closeness to the G4 structure) and the guanine bases constituting the tetrads, with values ranging from 0.0 (indicating no bonds) to 1.0 (indicating all 8 possible bonds) in increments of 0.125. A bond is considered established if the distance between the O6′ atom of guanine and the ion is less than 4.5 Å.

Since the middle tetrad is involved in the calculation of both coordination numbers, an ion exhibiting a coordination number of 1.0 in one of the two possible placements would also present a value of 0.5 in the other. A situation like this could be mistaken with the presence of two ions within the structure, one fully coordinated (hence $Co_i = 1.0$) and another partially coordinated ($Co_j \leq 1.0,\ j \neq i$). To prevent this specific confusion, once an ion achieves $Co_i = 1.0$ in one position, its coordination number is set to 0 in the other position. Monitoring the width of the central channel itself involves defining an additional set of coordination numbers for the three guanine tetrads, referred to as Co_P. These numbers, similar to those used for ions, quantify the distances between the four guanines forming each tetrad, based on the Hoogsteen bonds present in the structure. The coordination number ranges from 0.0 to 1.0 in increments of 0.25, with a bond considered established if the distance between two neighboring guanines is less than 5.0 Å.

Fig. 4 illustrates the evolution of the two coordination numbers for the ions (Co_1 and Co_2 in the first two rows), along with the three coordination numbers of the G-tetrads (third row, each color represents a plane as depicted in Fig. 1: blue for the top

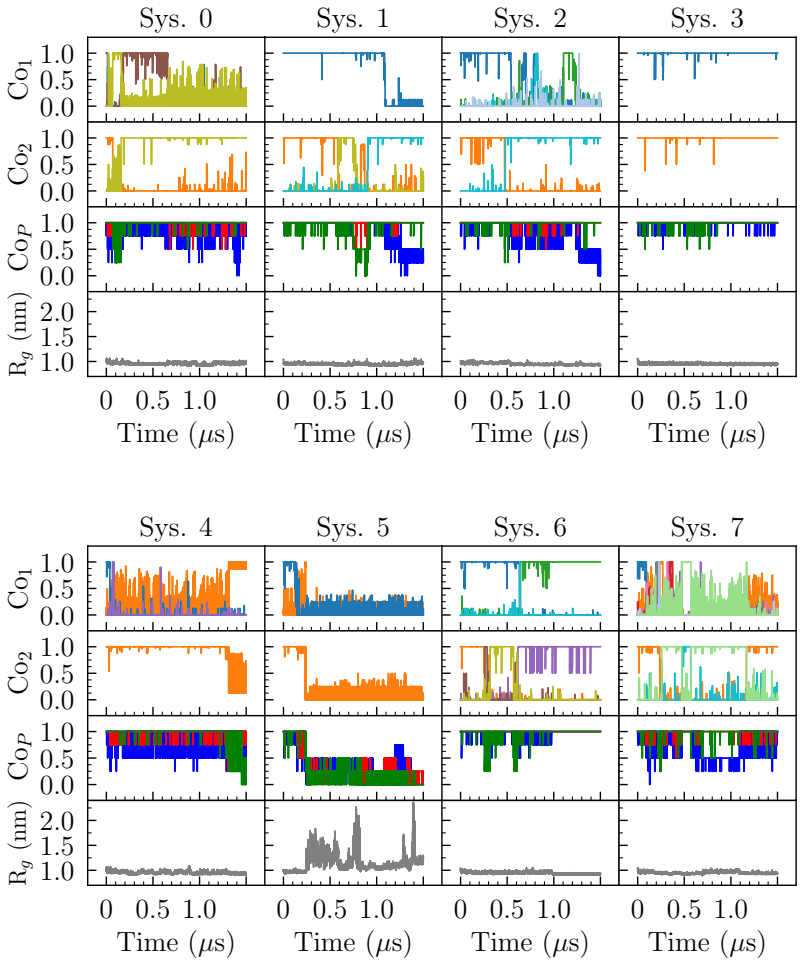

Figure 4. Coordination numbers of the top (Co_1) and bottom (Co_2) positions available for the ions, coordination number of the three G-tetrads (Co_P, blue for the top plane, red for the middle plane and green for the bottom plane) and radius of gyration of the G-tetrads (loops excluded). Only ions which achieve a value of 1.0 in each system are represented.

plane, red for the middle plane and green for the bottom plane) and the radius of gyration of the G-tetrads (loops excluded). All systems, except Sys. 3, experience the loss of one or both ions at some point of their trajectories. Excluding Sys. 5, which unfolds and consequently lacks a definable central channel, the systems either recover both ions or maintain one ion while losing the other, as observed in Systems 0, 1, 2, 4 and 7. It should be noted that the recovered ions are not necessarily those initially lost; in fact all the ions in the simulation box are equivalent, thus color changes in Fig. 4 can eventually occur. The departure of an ion from the channel increases the system's susceptibility to destabilization. This phenomenon is reflected in Fig. 4, where the loss of an ion from either position in the channel leads to significant fluctuations in both the coordination number of the corresponding planes and the radius of gyration R_g. An example of this can be seen in Sys. 7 in the time window [0.75 - 1.0] μs, showing fluctuation in the plane coordination numbers and deviations in R_g. The same deviations can be seen in Rep. 7 in Figs. 2 & 3 during the same time interval.

Ion escape and reabsorption are facilitated by slight changes in the central channel's width. The primary causes of these deformations includes the bending of G-tetrads, the motion of single guanines unbinding/drifting away from the structure. These deformations are not limited to the top and bottom tetrads. Disruptions in the intermediate plane can also result in the transfer of an ion from one position in the ionic channel to another, as observed in Sys. 0 at 0.2 μs , Sys. 4 at 1.3 μs and Sys. 7 at 0.5 μs.

Fig. 5 shows various conformations associated with the dynamics of the ions in the different systems. Subfigure **A** illustrates the escape of an ion from the top position, corresponding in Fig. 4 to the fall of the blue line in Sys. 1 around 1.1 μs, which triggers a decrease in Co_{P1}. Subfigure **B** shows the reverse process, where an ion is integrated into the structure, associated with the increase of the olive Co_1 line at around 0.2 μs in Sys. 0 and a corresponding rise in Co_{P2}. Subfigures **C** and **D**, taken

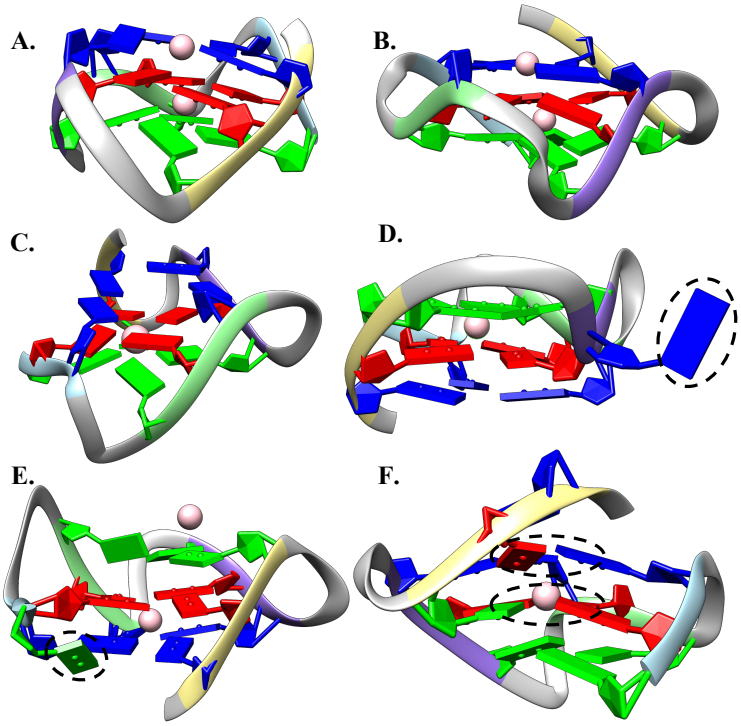

Figure 5. Relevant conformations associated to ion dynamics in the simulation. **A.** Escape of an ion from the top plane. **B.** Absorption of an ion through the top plane. **C.** Deformation of the top plane after ion loss. **D.** Separation of a single guanine from the tetrad of the top plane (indicated by black dashed outline). **E.** Separation of a single guanine from the bottom tetrad (indicated by black dashed outline), leading to ion loss. **F.** Slip-stranded conformation leading to the transference of an ion from the bottom to the top parts of the central channel. Slip-stranded contacts circled by dashed lines.

from Sys. 2 and 7, respectively, depict the separation of a guanine from the top (blue) plane after ion escape around 1.3 μs and 0.7 μs respectively, leading to a significant drop in the coordination number of that plane, $\mathrm{Co}_{\mathrm{P1}}$. Subfigure **E** shows a similar phenomenon, but with the displaced guanine in the bottom plane. This separation occurs while an ion remains bound, eventually leading to its loss around 0.76 μs in Sys. 2, accompanied by a decrease in $\mathrm{Co}_{\mathrm{P3}}$. Lastly, Subfigure **F** shows a slip-stranded conformation in Sys. 4 at 1.3 μs. The slip-stranding effect reduces the maximum possible coordination number to 0.875 and widens the central channel, facilitating the transfer of the ion from the bottom to the top position, altering the trends of the coordination numbers. This slip-stranding also causes a decrease in the coordination number $\mathrm{Co}_{\mathrm{P3}}$, with a new maximum value of $\mathrm{Co}_{\mathrm{P3,\,max}} = 0.75$.

From these results, it is evident that the presence of ions significantly influences the G-quadruplex stability, generally leading to partial guanine separation and other structural changes following their loss. In Sys. 5, which undergoes unfolding, both ions are lost before the process begins: one is lost early in the simulation, briefly replaced by another, which is lost again around 0.1 μs, and the other ion is lost at 0.24 μs, with unfolding commencing around 0.26 μs.

In the remaining systems, ion loss typically occurs in only one of the two available cavities. In this sense the loss of only one ion make the G4 structure less stable as it appears like a metastable state that maintains a certain stability. The G4 structure remains not completely compact, as visible in Fig. 5, where the structures are partially modified with the loss of one ion. We understand that these structures can maintain relative stability over time up to the loss of the second ion, a condition that definitely triggers the unfolding process. This observation may explain why only one system fully unfolded, as it required the loss of both ions before the unfolding process could initiate.

3.1.2 Unfolding process.

The unfolding process occurs in Sys. 5, shown in Figs. 2 and 3. The identification of peculiar states in the trajectory will be largely developed with the analysis with tICA and PCA, later on presented.

In Sys. 5, the initial configuration after the replica exchange occuring at $t = 0.2\,\mu$s corresponds to a folded G4 which has already lost one of its ions, as visible in Fig. 4. Afterwards, it quickly loses the other ion, at $t = 0.24\,\mu$s, therefore triggering the unfolding itself, as reflected by the increase of RMSD and $\mathrm{R_g}$ in Figs. 2 & 3 at $t = 0.25\,\mu$s.

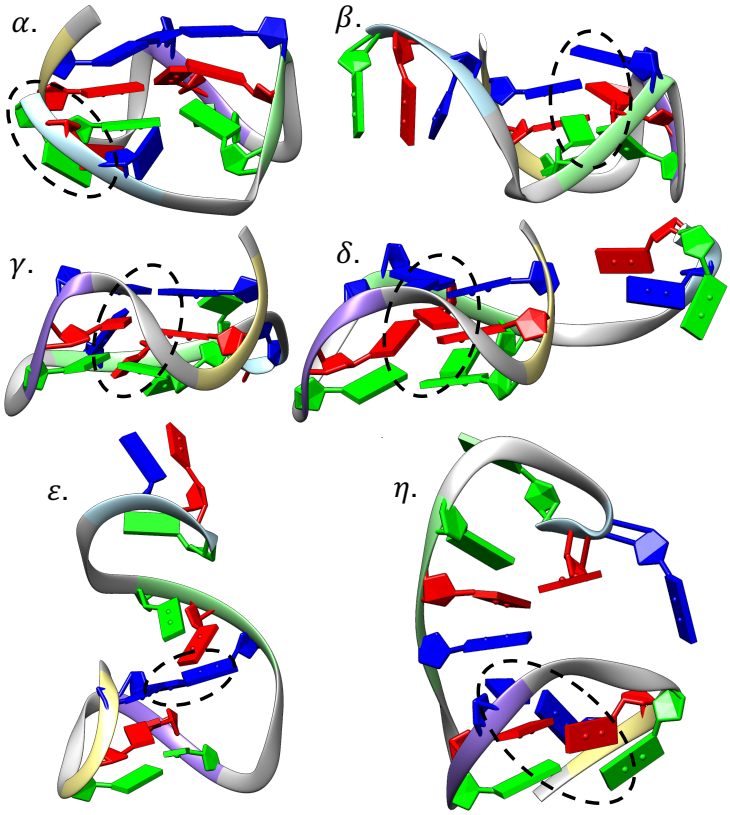

Figure 6. Highlighted conformations achieved during the evolution of the system, with the dotted circles highlighting the features of interest. $\boldsymbol{\alpha}$. Initial detachment of the fourth (icy blue) tract of guanines. $\boldsymbol{\beta}$. Example of a G-triplex conformation with a slip-stranded effect between tracts 1 (yellow) and 3 (green). $\boldsymbol{\gamma}$. Formation of a G-hairpin between tracts 1 (yellow) and 2 (magenta). $\boldsymbol{\delta}$. Refolded triplex after re- attachment of tract 1. $\boldsymbol{\varepsilon}$. Stable intermediate state of G- hairpin (tracts 1 and 2) with stacking interaction of the third tract. $\boldsymbol{\eta}$. Final state of the system, with tracts 1 and 2 in a cross hairpin conformation.

Fig. 6$\boldsymbol{\alpha}$ presents a G4 conformation in which the fourth guanine tract (ice blue, see Fig. 1) separates from the structure with the remaining three segments, thus forming a G-triplex[61,62], one common intermediates in G-quadruplex unfolding[63,64]. During the next 30 ns several attempts at refolding are made unsuccessfully. Moreover the dached circle in Fig. 6$\boldsymbol{\beta}$ shows, in the remaining G-triplex, the so-called “slip-stranding", *i.e.* a guanine tract that moves upwards or downwards respect to the others[65,66]. Other transient conformations can appear, such as G-hairpins[63,67–69], i.e. the bond of two guanine tracts only (see Fig. 6$\boldsymbol{\gamma}$), or even temporary refolding events into a G-triplex (Fig. 6$\boldsymbol{\delta}$).

However, after the successful refolding into a triplex, the third G4 tract detaches again, rotates over itself and forms a single Hoogsteen bond with the top plane of the G-hairpin formed by the first and second tract, as shown in Fig. 6$\boldsymbol{\varepsilon}$. Remarkably, this latter hybrid between a G-triplex and a G-hairpin remains stable for approximately $\Delta t \approx 0.6\,\mu$s of the simulation. The intense fluctuations observed in Fig. 3 in the interval $[0.3, 0.9]\,\mu$s correspond to changes in the distance between the third and fourth tracts. Eventually, the fourth strand rotates over itself and binds with the third strand through the bottom guanine (green), corresponding to the plateau between $t = 0.6\,\mu$s and $t = 0.75\,\mu$s. After instant $t = 0.75\,\mu$s, the bond joining the G-hairpin with the third tract breaks. The chain elongates, pulling in the second tract and separating it from the first, breaking the G-hairpin. It is quickly reformed but with the two tracts rotated, forming a “cross-hairpin"[69]. During this time, the third and fourth tracts rearrange themselves into two main conformations: one characterized by the stacking interactions between both tracts leading to a high R_g, and another presenting the formation of Hoogsteen bonds between their guanines, with smaller R_g. The interchange between them forms the peaks observed around $0.8\,\mu$s.

The final conformation of the system, stable and leading to the plateau in Fig. 3 starting at $t = 0.8\,\mu$s, depicts the formation of a *cross hairpin* (an arrangement of guanine rich strands in a cruciform shape, as circled in Fig. 6$\boldsymbol{\eta}$) between the first and the second tract, with the third and fourth stacks remaining close, accompanied by the formation of transient Hoogsteen bonds between them.

3.2 Dimensionality reduction.

Both PCA and tICA are techniques quite sensible to the characteristics of the data of interest.

The application of the two methods in our trajectories has required to clean the data as follows: i. first of all, we corrected the effects of the periodic boundary conditions on the GROMACS trajectories by removing the artificial discontinuities; ii. in order to prevent unrealistic distances of the trajectory coordinates emerging from global rotations and/or displacements of the structure as a whole, the coordinates of the structure have been rescaled by applying both a translation, superimposing the centers of mass with that of native structure, and rotations of the whole structure to recover the native orientation. A mean-square distance method, taking the initial state of the system as a reference, has been applied. iii. Finally, the mean was subtracted from each of the trajectory coordinates x_i, which have been used as input values for both PCA and tICA analysis.

3.2.1 Eigenvalues.

The eigenvalues obtained from solving Eqs. (1) and (2) give the possibility to reduce the information provided by the complete degrees of freedom of the system into a smaller set of variables containing its most significant part.

In PCA, the relative magnitude of the eigenvalues (with respect to their total sum) is related to the percentage of the total variance projected onto the corresponding eigenvector which is combination of the initial coordinates. Thus, the larger the eigenvalues, the more relevant that particular direction is to represent the overall description of the dynamics. Ideally, one should find an eigenvalue or a small series of them with magnitudes clearly superior to the rest, indicating that the majority of information of the variance of the system is contained in one reaction coordinate. The outcomes of the diagonalization of the correlation matrix are contained in the left side of Fig. 7.

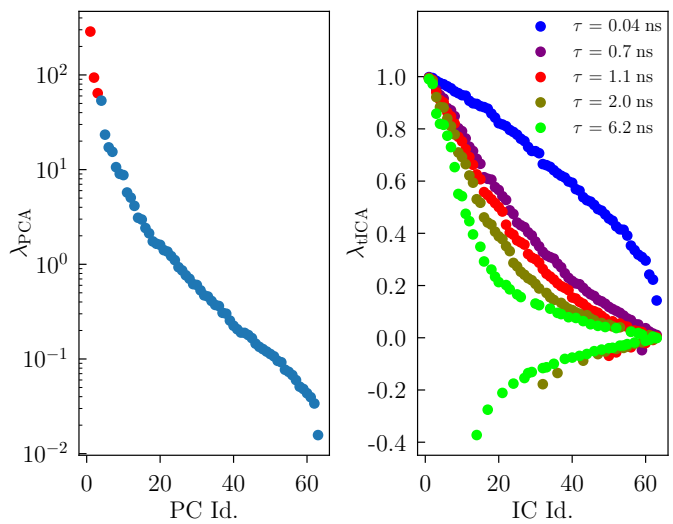

Figure 7. On the left, PCA eigenvalues obtained from solving (1), with the vertical axis in log-scale. The three eigenvalues in red correspond to the combinations chosen for data projection. On the right, tICA eigenvalues obtained from solving (2) for different values of the time window τ, seen in the legend.

The first four eigenvalues are clearly larger than all the rest, specially the first. We choose to use the first three eigenvectors associated to them as the basis for the PCA procedure, accounting for 70.9% of the variance.

For tICA, the interpretation of the eigenvalues is not as straightforward; their magnitude indicates the minimum timescale its associated coordinate is able to discern, calculated as in Eq. (3).

The eigenvalues of tICA are constrained between -1 and 1, with the negative values only appearing for large values of the lag time τ (see $\tau = 2.04$ ns in Fig. 7). These eigenvalues indicate modes in the system (described by their corresponding eigenvectors) that decay over time, typically corresponding to fast transitions or back-and-forth fluctuations that do not contribute to the stable, meaningful changes tICA aims to capture. Thus, the appearance of negative eigenvalues at a certain lag-time τ imposes an effective ceiling onto this magnitude, $\tau = 1.1$ ns in our particular case. The right side of Fig. 7 contains the eigenvalues for tICA calculated at different lag-times, showing the appearance of negative eigenvalues from $\tau = 1.1$ ns onwards.

The analysis of the trajectories was carried out with different values of τ but, for the rest of this document, a lag-time of $\tau = 0.7$ ns has been chosen. The reason for this particular value relies on the compromise we found which consists, on the one hand, in eliminating as many fast irrelevant motions of the system as possible that are translated into a few simulation instants

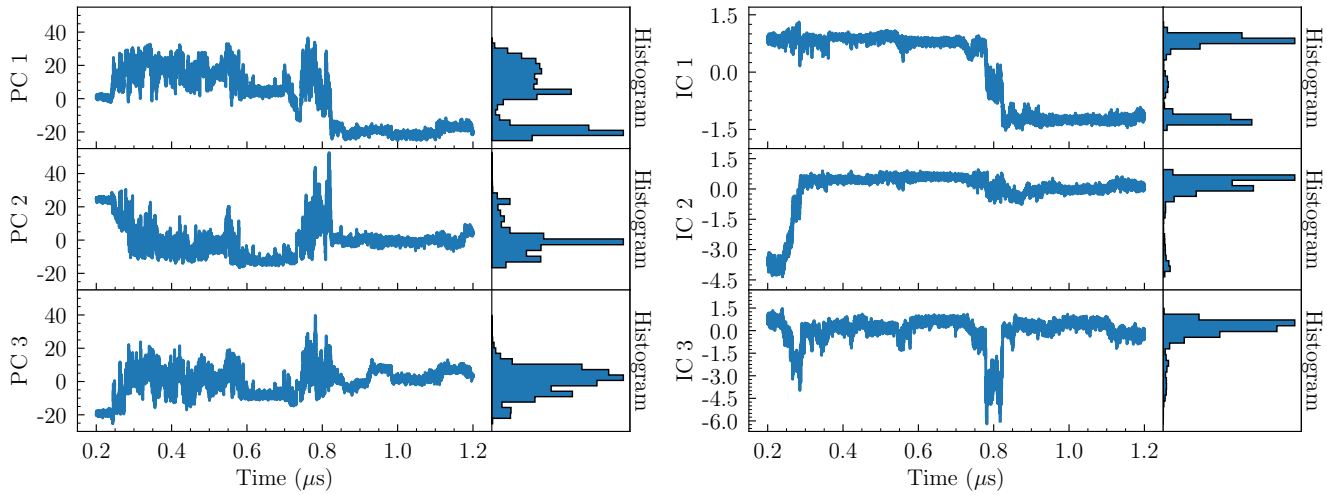

Figure 8. Projected trajectory from Sys. 5 between 0.2 and 1.2 μs onto t he first three eigenvectors (left), along with the histogram of the coordinates (right). PCA on the right, tICA with $\tau = 0.7$ ns on the left.

and thus undetected by higher values of tICA, while, and on the other hand, in protecting the information contained in the longer-lived states.

3.2.2 Projected trajectories.

Fig. 8 shows the trajectories of both methods, PCA and tICA, projected on the eigenvectors associated to the three chosen eigenvalues with, on the right hand side, the histogram of the corresponding coordinates occupation. The analysis of Fig. 8 reveals appreciable differences between the two methodologies. PCA yields trajectories characterized by histograms exhibiting broad Gaussian-like peaks, whereas tICA produces trajectories with narrower distributions, able to better distinguish between different states of the system with improved clarity. Commonalities are observed in both methods, visible, for example in the first coordinate PC1 and IC1 which show a similar trend: high value for the first interval between $t = 0.25$ and $t = 0.8\,\mu$s, and after that a narrow transition to another constant value. Both trajectories describe the general behavior observed in the root-mean-square deviation (RMSD) (see Fig. 2), with tICA exhibiting less fluctuations than PCA, that, instead, almost reproduces the same shape as the original trajectory.

The different capabilities of the methods in detecting independent and well-defined states throughout the system's trajectory become evident when plotting the coordinates against each other as in Fig. 9. The image depicts a 3D plot of the coordinates extracted from PCA (left) and tICA (right), with the color scale corresponding to the free energy differences in the trajectory determined by the relative occupation of specific coordinate combinations during evolution, and computed as $-\log(P/P_0)$, with P_0 being the lowest occupation probability ($\neq 0$) in the states ensemble. On the bottom of each 3D plot there are three heatmaps plotting the coordinates against each other by pairs.

In Fig. 9 is well visible that PCA exhibits greater variance in the evolution of its coordinates featuring bright states surrounded with clouds of relatively populated spots which, when 3D plotted, seem to collapse in a central globule. Conversely, tICA produces distinct bright spots that are well separated, with additional small areas of low population in between, leading to a 3D plot characterized by small cloud separated from one another.

3.3 Complex Markov Network results.

Once a suitable value for τ for tICA is chosen and the trajectories are projected onto the reduced coordinate space we can proceed with their encoding into a complex network. As previously explained, the trajectory undergoes segmentation into equidistant subsets, each constituting a node when populated by the system. Subsequently, a Stochastic Steepest Descent algorithm is applied to facilitate the classification of all nodes into distinct basins of attraction.

The resulting networks of basins depicting the unfolding process are illustrated in Figs. 10 and 11.

The results from tICA are presented first, since they provide a more straightforward depiction of the unfolding process. The PCA results are then analysed subsequently.

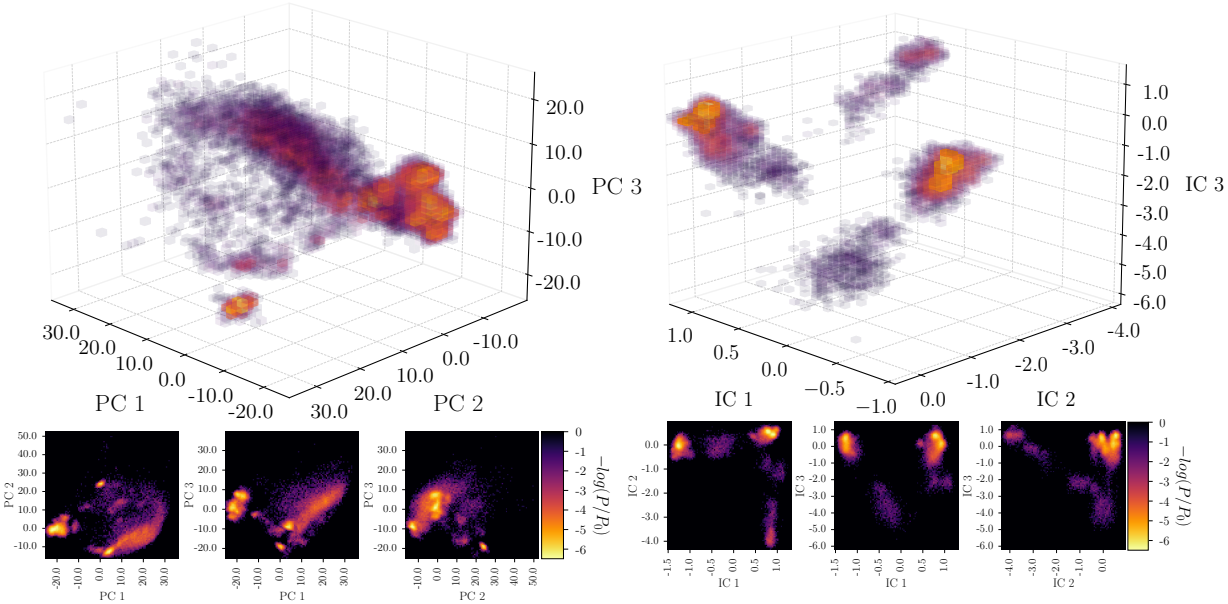

Figure 9. Colormaps depicting the correlation of the different coordinates emerging from PCA (left) and tICA (right). Within each subfigure, the bottom three panels depict heatmaps illustrating the pairwise relationships between the projected coordinates, with the color representing the relative occupancy of the corresponding coordinate pairs throughout the trajectory. The three-dimensional plots illustrates the combined coordination of the coordinates. Each data point in this plot corresponds to a specific conformation sampled during the trajectory, with the color representing the relative occupancy of a particular combination during the simulation.

3.3.1 tICA analysis.

The tICA analysis summarized in Fig. 10 shows a linear representation of the unfolding process, with 7 states identified as basins of attraction in the CMN analysis. The representative node of each basin is labeled from 1 to 7 in the figure, and its corresponding time of appearance in the system trajectory is marked in Fig. 3 with green lines.

Starting from a fully folded conformation (basin #I), the next state (basin #II) portrays the system as already having severed its Hoogsteen bonds with the fourth guanine tract at the 3′-end of the chain. In this basin tICA is able to recognize the "slip-stranded" conformations between the remaining tracts in the G-triplex, reminiscing of Fig. 6$\boldsymbol{\beta}$.

Basin #III depicts the moment after which the fourth guanine tract, previously disconnected from the main triplex but still proximal to it due to crowding effects, has separated. Moreover, another detail can be observed, consisting in the slip-stranding of tract 2 from the remaining triplex (slip-stranded contacts circled in dashed line in Fig. 10, basin #3).

In Basin #IV the third tract has separated from the triplex, rotating over itself and forming a single Hoogsteen bond with the top guanines (blue) of the first and second tracts, as described in Fig. 6$\boldsymbol{\varepsilon}$. This basin represents the first relatively stable state of the system, corresponding to the time interval [0,3-0,7] μs in Figs. 2 & 3. The fluctuations of these magnitudes in that time interval correspond to waving motions between the third and fourth guanine tracts.

Basins #V and #VI describe very similar situations. The last bond between the third tract and the remaining G-hairpin breaks. The third tract drifts away and pulls the second tract, breaking the G-hairpin and making it into the so called "cross-hairpin" (Fig. 6$\boldsymbol{\eta}$). There is a slight difference between the two states, consisting in the relative position of tracts 3 and 4, which in basin #VI are a little closer due to a stacking interaction created by the top guanines (green) of both tracts (dashed circle in Fig. 10).

Basin #VII corresponds to the final state of the system: tracts 1 and 2 remain in a cross-hairpin, while the third and fourth remain close to each other, with two Hoogsteen bonds formed between them, almost constituting a complete hairpin. This configuration is very stable, with the presence of fluctuations due to the separation of the two blocks (see Figs. 2 and 3 from $t = 0.8\,\mu$s onwards).

With these results in mind, it becomes evident that tICA is able to provide us, upon selection of a proper value for the time window τ, a clear denaturation path. The states identified through the basins exhibit a high degree of dissimilarity, capturing the significant structural transformations and avoiding the oversampling of both transient and short-lived conformations.

In the figure, the links between states are bidirectional (two arrows) and weighted. They represent multiple transitions between the different basins with no information about a particular time sequence. This is due to the fact that, in principle, the

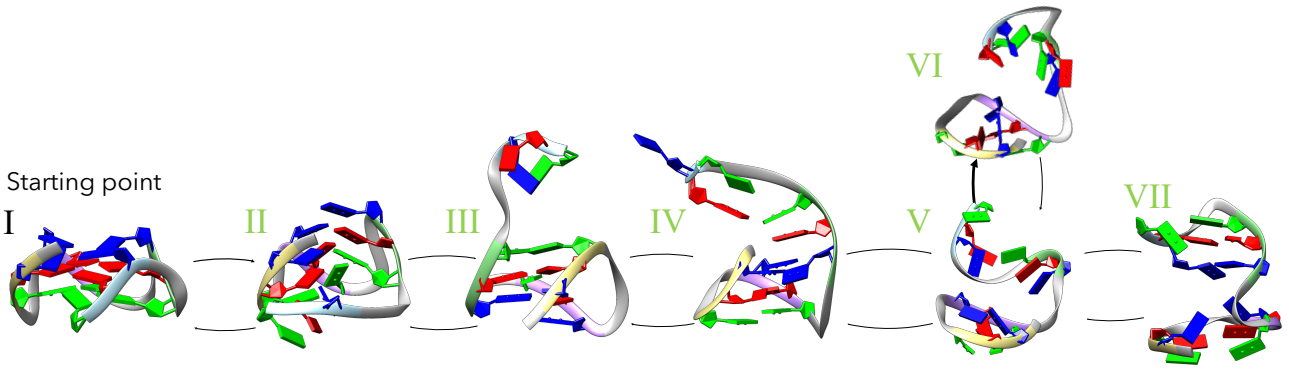

Figure 10. Network of basins found by tICA, $\tau = 0.7$ ns. The green numbers refer to the equally colored lines in Fig. 3, showing the position of the basins along the trajectory.

system investigated lies in a dynamical equilibrium condition. This means that if the trajectories were long enough, they could be analyzed independently of time, with the occurrence of the same intermediates.

To study the stability of the basins found by tICA, the average escape time of each state can be calculated. These times reflect the average time the G-quadruplex remains in a given conformation before undergoing a transition. By analyzing escape times, insights into the stability and transition kinetics of different conformational states can be gained. We calculate escape times by using two approaches: one based on Complex Markov Network self-loops and another analyzing the time spent in a given conformation, providing a an estimation of the free energy difference between the different basins. The detailed methodology as well as the results from these calculations can be found in the Supplementary Material; they reveal that the structure depicted in basin #IV has a larger lifetime than the rest, thus classifying it as a stable intermediate of the unfolding.

Additionally, the above analysis has been performed under a mesoscopic reduction of data, as previously explained. Nevertheless, the same tICA procedure can be applied without that intermediate step. In that case an even broader spectrum of states is resolved, with the resulting network still describing a clear denaturation path. This analysis is also included in the Supplementary Material.

3.3.2 PCA results.

Fig. 11 shows the network of states generated by the PCA procedure. It is evident that the interpretation of the basins becomes severely more challenging here when compared to the one produced by tICA. PCA aims to build coordinates containing the maximum variance possible which, when applied to the identification of unique basins of attraction, leads to the labeling of some equivalent states as different, overestimating the differences between equivalent configurations.

The initial state corresponds to the same starting conformation as in tICA, designated as basin #1 in Fig. 11. Subsequently, basin #2 marks the point at which the $3'$-end disengages from the primary structure, forming the G-triplex. Following this event, a number of tightly connected states is revealed (from basin #3 to #8). Basin #3 shows a slip-stranded G-triplex, with the fourth guanine tract away from the structure. Up to this point, the states analyzed by PCA are equivalent to those of tICA, namely basins #I, #II and #III.

Basin #4 shows a G-hairpin formed by the first and second tracts, with one of the guanines from the third tract in contact with the hairpin, while in basin #5 that contact has been broken, leaving a G-hairpin and two drifting guanine tracts (third and fourth).

The remaining basins (numbers #6, #7 and #8) in this heavily interconnected section of the network are classified as erroneously different by PCA respect to tICA. In particular, basins #6 and #7 depict a G-hairpin accompanied by an additional third guanine tract forming a single Hoogsteen bond. The difference between these conformations fundamentally lies in the relative orientation of bonds on the guanine columns forming the hairpin. Basin #8, on the other hand, is analogous to basin #5, represents a G-hairpin no longer in contact with the third tract. All these states (basins #5, #6, #7 and #8) identified by PCA are thus equivalent to basin #IV in tICA. Remarkably, the state identified by PCA only (basin #7) is then completely superfluous in describing the unfolding process because it is contained in basin #IV in tICA.

States #9, #10, and #11 and #12 show the final conformation of the system. They correspond to basins #V, #VI and #VII in tICA. They reveal the formation of transient and relatively short-lived bonds reminiscent of G-hairpins, resulting each of them in the division of the G-quadruplex into two well-separated guanine tracts: a cross-hairpin (tracts 1 and 2) and a disordered clump (tracts 3 and 4). However, while the three states found by tICA were distinguishable by the presence or absence of certain interactions (stacking between guanines in basin #VI and guanine-guanine bonds in basin #VII), in the case of PCA they are functionally equivalent: basins #10, #11 and #12 show the cross-hairpin in the first two tracts, along with an additional

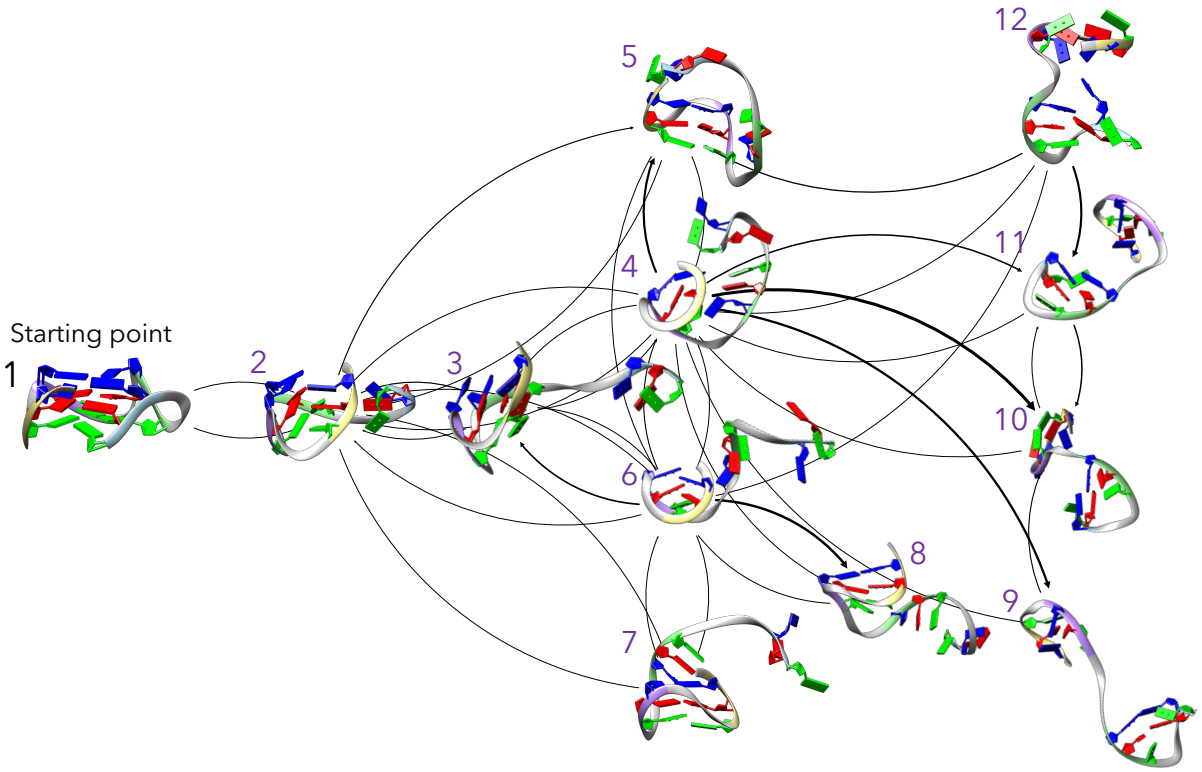

Figure 11. Network of basins found by PCA. The purple numbers refer to the equally colored lines of Fig. 3, indicating the position of the basins along the trajectory.

Hoogsteen bond between the third and fourth, while basin #9 presents an additional bond between them, with no stacking interaction being captured.

To conclude, our simulations reveal that PCA falls short in providing a minimal unfolding path akin to that offered by tICA, instead providing a multitude of interconnected states, some of them functionally equivalent.

Analogously as done with tICA analysis, the relative stability of the basins identified can be studied through the escape times, whose results can be found in the Supplementary Material. Furthermore, the PCA analysis without the mesoscopic reduction is also found in the Supplementary Material, leading to a higher number of basins with increased degeneration between them. They correspond to Supplementary Figures 6 & 7.

4 Discussion and conclusions.

The study of unfolding pathways in biological systems is essential to understanding their stability and functional roles within their environments. In this work, we combined molecular dynamics simulations and multiple analytical techniques to explore the unfolding process of the human G-quadruplex in its parallel configuration (PDB: 1KF1). By characterizing key structural changes during unfolding, we have gained insight into the underlying factors that contribute to the stability and flexibility of this biologically relevant structure.

We employed Replica Exchange Molecular Dynamics (REMD) to accelerate and improve the sampling of the G-quadruplex, allowing us to observe a distinct unfolding event in only one of the eight replicas near unfolding temperatures.

A crucial aspect of our study involved analyzing ion dynamics. We observed that total ion loss appears to be a necessary step to start the unfolding process, which also agrees with mechanical unfolding studies[17]. Systems that experienced only partial ion loss maintained their structural integrity, even reabsorbing ions from the environment and returning to their initial configurations, as evidenced in Fig. 4.

We demonstrate that the mesocopic reduction of the coordinates[18] allowed us to efficiently simplify the system while preserving key dynamics. In fact, the analysis performed with the full system coordinates reports none significant change (Supplementary Material, figs. 6 & 7). This opens the door for the creation and development of realistic coarse-grained models capable of reproducing the unfolding of these structures, similar to those applied to DNA chains[70].

Moreover, dimensionality reduction techniques (PCA and tICA) have been applied to the coarse-grained system, providing

a smaller number of output trajectories, which were later transformed into a series of Complex Markov Networks to identify key states and transitions in the G-quadruplex unfolding pathway.

A series of molecular dynamics simulations have predicted the existence of intermediates in the G4 denaturation pathways. Examples are the G-triplex [22,61,62,79,80], G-hairpins[22,67–69,73,74,79,80], slip-stranding effects[65,66] and cross-hairpins[69]. Interestingly, the presence of the G-hairpin in the G4 loops, is known to increase the stability of the G-quadruplex with the loop length[28,77].

More recently, the experimental observation of these intermediates has been satisfactorily confirmed with circular dichroism (CD) and smFRET (single molecule Fluorescence Resonance Energy Transfer) techniques [63,64]. The authors claim that both anti-parallel and (probably) hybrid G4 structures, in either K^+ or Na^+ rich environment, unfold to a ssDNA through the formation of G-triplex and G-hairpins intermediates. The alternative pathway, which would avoid the passage to G-triplex (by forming at the same time two G-hairpins) has not been found in those experiments, nor in our simulations. The difference of these experiments with our simulations resides in the fact that we simulate here the unfolding of a fully parallel G4 structures, whose intermediates are, evidently, again parallel.

Experimentally there is no clear evidence about the existence of such structures, because of their elusive nature and their low lifetime. The only G-triplex structure successfully identified is the anti-parallel thrombin binding aptamer (TBA)[75], by means of NMR (Nuclear Magnetic Resonance) and CD spectroscopy. Other works have been able to reveal the existence of similar conformations through the use of multiple techniques, such as DNA origami[22,80] or microfluidic mixers[79].

In another unfolding experiment Ref.[76], the parallel G-triplex remains unrevealed. The authors comment there that the entire G4 parallel structure is unlikely to be observed under their experimental conditions. In a more recent series of experiments with parallel G4s where the photochemical trapping strategy is used[64,78], the occurrence of the parallel G-triplex again appears elusive.

The reason why the experimental occurrence of this structure is unlikely may reside in the fact that the parallel G-triplex intermediates have a very short lifetime that make it their experimental identification difficult. The lifetime registered in our simulations, appears one order of magnitude lower that the more stable G-hairpin. In fact, our simulations have revealed a particularly long lasting (0.4 μs out of 1.5μs total) intermediate containing a G-hairpin (basin #IV in tICA, basins #7 and #8 in PCA), while the G-triplex remains at a maximum of 40 ns. The different lifetimes of the structures can be coded in the probability to belong to a certain basin, and correspondly to have a higher or lower free energy state, calculations that can be found in the "Basin escape times" of the Supplementary Material.

Interestingly, the final state of the system (basin #VII in tICA, basin #9 to #12 in PCA) also presents a combination of again a G-hairpin with a cross-hairpin, which confirm to be the most stable intermediates remaining out of the parallel G-quadruplex.

The unfolding process is triggered by the loss of both ions in the G4 cavities. Many occurrences have registered the loss of only one ion, without the complete G4 unfolding. These structures remain not completely compact but they resist over time, so they appear as metastable states that maintain a certain stability. We have no notice of the registered observation of such structures in the literature or if they may have some biological importance. The loss of the second ion definitely triggers the unfolding process.

Importantly, as the simulations have been performed in biologically significant range of temperature slightly above the denaturation point, we trust that the conformations found in our simulations do reflect real biological conditions, separated by thermally accessible barriers in the unfolding process.

It is important to note that under the above-described thermal conditions, our simulations hardly reveal denaturation processes. That is why we have shown only one successful unfolding trajectory, able to identify one clear unfolding pathway.

Regarding the particular methods employed for the analysis, both tICA and PCA lead to networks capable of describing the most meaningful unfolding trajectory. We found that tICA is particularly effective in separating the most relevant conformations, while PCA identifies a larger number of states, which however turn out to be equivalent to each other in many cases.

Future investigations could benefit from exploring alternative dimensionality reduction techniques, particularly those that do not rely on linear combinations of input data, such as non-linear methods or machine learning-based approaches[71,72], which, however, would require the generation of a larger number of unfolding trajectories, thus leading to an increased computational cost.

As a last remark, the final configurations obtained in our simulation (basin #VII in tICA) leave open the refolding occurrence not necessarily into the same parallel G-quadruplex, but to a different G4 structure, either antiparallel or hybrid ones. In this sense, the possibility of a conformational switch between parallel and either antiparallel or hybrid shapes seems likely, due to the unfolding process. This possibility remains to be explored both *in silico* and *in vitro*.

Funding.

The authors acknowledge the Grant No. PID2020-113582GB-I00 and the Grant No. PID2023-147734NB-I00 funded by MCIN/AEI/10.13039/501100011033, the support of the Aragon Government to the Recognized group 'E36_23R Física Estadística y no-lineal (FENOL)'. AS-A also acknowledges the support of the predoctoral FPI fellowship PRE2021-100456 funded by MCIN/AEI/10.13039/501100011033.

Author contributions statement

AF & FF conceived the project. AS-A conducted the simulations and the data analysis. AF, FF & AS-A wrote and reviewed the manuscript.

Additional information

No competing interest is declared.

Data Availability

All data generated or analyzed during this study are included in the published article.

References

1. Hans J. Lipps & D. Rhodes. G-quadruplex Structures: In Vivo Evidence and Function. *Trends in Cell Biology* **19** 414 (2009)
2. E. Y. Lam, D. Beraldi, D. Tannahill & S. Balasubramanian. G-quadruplex structures are stable and detectable in human genomic DNA *Nat. Commun.* **4** 1796 (2013)
3. Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic acids research* **34** 5402 (2006)
4. G. Louit, A. Hocquet, M. Ghomi, M. Meyer & J. Sühnel. Are guanine tetrads stabilised by bifurcated hydrogen bonds? An AIM topological analysis of the electronic density. *PhysChemComm* **5** 94 (2002)
5. J. Li, J. J. Correia, L. Wang, J. O. Trent & J. B. Chaires. Not so crystal clear: the structure of the human telomere G-quadruplex in solution differs from that present in a crystal. *Nucleic acids research* **33** 4649 (2005)
6. A. E. Bergues-Pupo, J. R. Arias-Gonzalez, M. C. Morón, A. Fiasconaro & F. Falo. Role of the central cations in the mechanical unfolding of DNA and RNA G-quadruplexes *Nucleic Acids Research* **43** 7638 (2015)
7. P. Murat and S. Balasubramanian. Existence and Consequences of G-quadruplex Structures in DNA. *Current Opinion in Genetics & Development* **25** (2014)
8. Hans J. Lipps & D. Rhodes. G-quadruplexes and Their Regulatory Roles in Biology. *Nucleic Acids Research* **43** 8627 (2015)
9. R. Hänsel-Hertsch, M. Di Antonio & S. Balasubramanian. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential *Nat. Rev. Mol. Cell Biol.* **18** 279 (2017)
10. T. Endoh & N. Sugimoto. Mechanical insights into ribosomal progression overcoming RNA G-quadruplex from periodical translation suppression in cells. *Sci. Rep.* **6** 1 (2016)
11. J. Lopes, A. Piazza, R. Bermejo, B. Kriegsman, A. Colosio, M. P. Teulade-Fichou, M. Foiani & A. Nicolas. G-quadruplex-induced instability during leading-strand replication. *The EMBO journal*, **30** 4033 (2011)
12. C. Papadopoulou, G. Guilbaud, D. Schiavone & J. E. Sale. Nucleotide Pool Depletion Induces G-quadruplex-Dependent Perturbation of Gene Expression. *Cell reports* **13** 2491 (2015)
13. A. Siddiqui-Jain, C. L. Grand, D. J. Bearss & L. H. Hurley. 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** 11593 (2002)
14. T. Tian, Y. Chen, S. Wang & X. Zhou. G-quadruplex: A Regulator of Gene Expression and Its Chemical Targeting *Chem*, **4**, 1314 (2018)
15. N. Kosiol, S. Juranek, P. Brossart, A. Heine & K. Paeschke. G-quadruplexes: a promising target for cancer therapy. *Mol Cancer.* **20** 40 (2021)

16. D. Sen, W. Gilbert. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334** 364 (1988).

17. A. E. Bergues-Pupo, I. Gutiérrez, J. R. Arias-Gonzalez, F. Falo & A. Fiasconaro. Mesoscopic model for DNA G-quadruplex unfolding. *Sci Rep* **7** 11756 (2017)

18. A. E. Bergues-Pupo, F. Falo & A. Fiasconaro. Modelling the DNA Topology: The Effect of the Loop Bending on G-quadruplex Stability. *Journal of Statistical Mechanics: Theory and Experiment* **9**(2019)

19. Y. Cheng, Y. Zhang, & Y. Huijuan. Characterization of G-quadruplexes Folding/Unfolding Dynamics and Interactions with Proteins from Single-Molecule Force Spectroscopy. *Biomolecules* **11** 1579 (2021)

20. J. Jana & K. Weisz. Thermodynamic Stability of G-Quadruplexes: Impact of Sequence and Environment. *ChemBioChem* **22** 19 (2021)

21. A. N. Lane, J. B. Chaires, R. D. Gray & J. O. Trent. Stability and Kinetics of G-quadruplex Structures. *Nucleic Acids Research* **36** 17 (2008)

22. A. Rajendran, M. Endo, K. Hidaka, M-P. Teulade-Fichou, J-L. Mergny & H. Sugiyama. Small Molecule Binding to a G-Hairpin and a G-Triplex: A New Insight into Anticancer Drug Design Targeting G-Rich Regions. *Chemical Communications* **51** (2015)

23. C. Schildkraut & S. Lifson. Dependence of the Melting Temperature of DNA on Salt Concentration. *Biopolymers* **3** 2 (1965)

24. T. Q. N. Nguyen, K. W. Lim & A. T. Phan. Folding Kinetics of G-Quadruplexes: Duplex Stem Loops Drive and Accelerate G-Quadruplex Folding. *The Journal of Physical Chemistry B* **124** (2020)

25. H. Han, D. R. Langley, A. Rangan & L. H. Hurley. Selective Interactions of Cationic Porphyrins with G-Quadruplex Structures. *Journal of the American Chemical Society* **123** (2001)

26. D. Miyoshi, H. Karimata & N. Sugimoto. Hydration Regulates Thermodynamics of G-Quadruplex Formation under Molecular Crowding Conditions. *Journal of the American Chemical Society* **128** (2006)

27. A. T. Phan, K. N. Luu & D. J. Patel. Different Loop Arrangements of Intramolecular Human Telomeric (3+1) G-Quadruplexes in K^+ Solution. *Nucleic Acids Research* **34** (2006)

28. M. Nakata, N. Kosaka, K. Kawauchi & D. Miyoshi. Quantitative Effects of the Loop Region on Topology, Thermodynamics, and Cation Binding of DNA G-Quadruplexes. *ACS Omega* **9** (2024)

29. S. Haider & S. Neidle. Molecular modeling and simulation of G-quadruplexes and quadruplex-ligand complexes. *Humana Press* **17** (2010) ISO 690

30. H. Zhu, S. Xiao & H. Liang. Structural dynamics of human telomeric G-quadruplex loops studied by molecular dynamics simulations. *PloS one* **8** e71380 (2013)

31. K. Hukushima & Koji Nemoto. Exchange Monte Carlo Method and Application to Spin Glass Simulations. *Journal of the Physical Society of Japan* **65** 6 (1996)

32. Y. Sugita & Y. Okamoto. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical Physics Letters* **314** 1–2 (1999)

33. G. N. Parkinson, M. P. H. Lee, & S. Neidle. Crystal Structure of Parallel Quadruplexes from Human Telomeric DNA. *Nature* **417** 6891 (2002)

34. K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **11** 559 (1901)

35. H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* **24** 417 (1933)

36. L. Molgedey & H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters* **72** 3634 (1994)

37. D. Prada-Gracia, J. Gómez-Gardeñes, P. Echenique & F. Falo. Exploring the free energy landscape: from dynamics to networks and back. *PLoS computational biology* **5** e1000415 (2009)

38. D. J. Wales. Energy Landscapes: Some New Horizons. *Current Opinion in Structural Biology* **20** 1 (2010)

39. D. J. Wales. Exploring Energy Landscapes. *Annual Review of Physical Chemistry* **69** 1 (2018)

40. X. Zeng, L. Zhang, X. Xiao, Y. Jiang, Y. Guo, X. Yu, X. Pu & M. Li. Unfolding Mechanism of Thrombin-Binding Aptamer Revealed by Molecular Dynamics Simulation and Markov State Model. *Scientific Reports* **6** (2016)

41. H. Bekker, H.J.C. Berendsen, E.J. Dijkstra, S. Achterop, R. van Drunen, D. van der Spoel, A. Sijbers & H. Keegstra et al. Gromacs: A parallel computer for molecular dynamics simulations. *Physics computing* **92** 252-256 (1993)

42. H.J.C. Berendsen, D. van der Spoel & R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91** 43–56 (1995)

43. A. Pérez, I. Marchán, D. Svozil, J. Šponer, T. E. Cheatham, C. A. Laughton & M. Orozco. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of $\alpha\gamma$ Conformers. *Biophysical Journal* **92** 11 (2007)

44. J. Šponer, X. Cang & T. E. Cheatham. Molecular Dynamics Simulations of G-DNA and Perspectives on the Simulation of Nucleic Acid Structures. *Methods* **57** 1 (2012)

45. N .Li, Y. Gao, F. Qiu & T. Zhu. Benchmark Force Fields for the Molecular Dynamic Simulation of G-quadruplexes. *Molecules* **26** 5379 (2021)

46. M. Zgarbová, J. Šponer, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo & P. Jurečka. Refinement of the Sugar–Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation* **11** 12 (2015)

47. I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, J. Walther et al. Parmbsc1: A Refined Force Field for DNA Simulations. *Nature Methods* **13** 1 (2016)

48. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey & M. L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* **79** 2 (1983)

49. J. L. Mergny, A. T. Phan & L. Lacroix. Following G-quartet formation by UV-spectroscopy. *FEBS letters* **435** 74 (1998)

50. P. Pokorná, V. Mlýnský, G. Bussi, J. Šponer & P. Stadlbauer. Parallel G-quadruplex Folds via Multiple Paths Involving G-Tract Stacking and Structuring from Coil Ensemble. *Preprint, bioArXiv* (2023)

51. R. Qi, G. Wei, B. Ma & R. Nussinov. Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example. In *Peptide Self-Assembly*. Springer New York (2018)

52. M. J. Abraham & J. E. Gready. Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation* **4** 7 (2008)

53. X. Periole & A. E. Mark. Convergence and Sampling Efficiency in Replica Exchange Simulations of Peptide Folding in Explicit Solvent. *The Journal of Chemical Physics* **126** 1 (2007)

54. C. P. Robert & G. Casella. Monte Carlo Statistical Methods. *Springer Texts in Statistics. New York, NY: Springer New York* (2004)

55. R. Gowers, M. Linke, J. Barnoud, T. Reddy, M. Melo, S. Seyler, J. Domański, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. **98–105** (2016)

56. N. Michaud-Agrawal, E. J. Denning, T. B. Woolf & O. Beckstein. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *Journal of Computational Chemistry* **32** 10 (2011)

57. A. Amadei, A. B. Linssen & H. J. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **17** 412 (1993)

58. M. U. Ali, S. Ahmed, J. Ferzund, A. Mehmood & A. Rehman. Using PCA and factor analysis for dimensionality reduction of bio-informatics data. *arXiv preprint arXiv* 1707.07189 (2017)

59. A. Hyvarinen, J. Karhunen & E. Oja. Independent component analysis. *Studies in informatics and control* **11** 205 (2002)

60. G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis & F. Noé. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **139** 07B604_1 (2013)

61. P. Stadlbauer, L. Trantírek, T. E. Cheatham, J. Koča & J. Šponer. Triplex Intermediates in Folding of Human Telomeric Quadruplexes Probed by Microsecond-Scale Molecular Dynamics Simulations. *Biochimie* **105** (2014)

62. M. Bončina, J. Lah, I. Prislan & G. Vesnaver. Energetic Basis of Human Telomeric DNA Folding into G-quadruplex Structures. *Journal of the American Chemical Society* **134** 23 (2012)

63. X. Hou, Y. Fu, W. Wu, L. Wang, F. Teng, P. Xie, P. Wang, & X. Xi. Involvement of G-Triplex and G-Hairpin in the Multi-Pathway Folding of Human Telomeric G-quadruplex. *Nucleic Acids Research* **45** 19 (2017)

64. J. Tassilo Grün & H. Schwalbe. Folding Dynamics of Polymorphic G-quadruplex Structures. *Biopolymers* **113** 1 (2022)

65. P. Stadlbauer, M. Krepl, T. E. Cheatham, J. Koca, & J. Šponer. Structural Dynamics of Possible Late-Stage Intermediates in Folding of Quadruplex DNA Studied by Molecular Simulations. *Nucleic Acids Research* **41** 14 (2013)

66. I. Kejnovská, P. Stadlbauer, L. Trantírek, D. Renčiuk, M. Gajarský, D. Krafčík, J. Palacký, et al. G-Quadruplex Formation by DNA Sequences Deficient in Guanines: Two Tetrad Parallel Quadruplexes Do Not Fold Intramolecularly. *Chemistry – A European Journal* **27** 47 (2021)

67. M. Gajarský, M. Lenarčič Živković, P. Stadlbauer, B. Pagano, R. Fiala, J. Amato, L. Tomáška, J. Šponer, J. Plavec & L. Trantírek. Structure of a Stable G-Hairpin. *Journal of the American Chemical Society* **139** 10 (2017)

68. P. Stadlbauer, P. Kührová, P. Banáš, J. Koča, G. Bussi, L. Trantírek, M. Otyepka & J. Šponer. Hairpins Participating in Folding of Human Telomeric Sequence Quadruplexes Studied by Standard and T-REMD Simulations. *Nucleic Acids Research* (2015)

69. P. Stadlbauer, P. Kührová, L. Vicherek, P. Banáš, M. Otyepka, L. Trantírek & J. Šponer. Parallel G-Triplexes and G-Hairpins as Potential Transitory Ensembles in the Folding of Parallel-Stranded DNA G-quadruplexes. *Nucleic Acids Research* **47** 14 (2019)

70. T. E. Ouldridge, L. A. Ard & J. P. K. Doye. Structural, Mechanical, and Thermodynamic Properties of a Coarse-Grained DNA Model. *The Journal of Chemical Physics* **134** 8 (2011)

71. A. Moses. Statistical Modeling and Machine Learning for Molecular Biology. *Chapman and Hall/CRC* (2017)

72. P. Mehta, M. Bukov, C. Wang, A.G.R. Day, C. Richardson, C.K. Fisher & D. J. Schwab. A High-Bias, Low-Variance Introduction to Machine Learning for Physicists. *Physics Reports* **810** (2019)

73. T. Mashimo, H. Yagi, Y. Sannohe, A. Rajendran, H. Sugiyama. Folding pathways of human telomeric Type-1 and Type-2 G-quadruplex structures. *J. Am. Chem. Soc.* **132** 14910 (2010).

74. R.D. Gray, J.O. Trent and J.B. Chaires, Folding and unfolding pathways of the human telomeric G-quadruplex. *J. Mol. Biol.* **426** 1629 (2014).

75. L. Cerofolini, J. Amato, A. Giachetti, V. Limongelli, E. Novellino, M. Parrinello, M. Fragai, A. Randazzo, C. Luchinat, G-Triplex structure and formation propensity. *Nucleic Acids Res.* **42** 13393 (2014).

76. M. Aznauryan, S. Sondergaard, S.L. Noer, B. Schiott, V. Birkedal, A direct view of the complex multi-pathway folding of telomeric G-quadruplexes. *Nucleic Acids Res.* **44** 11024 (2016).

77. K.W. Lim, Z.J. Khong, A.T. Phan, Thermal stability of DNA quadruplex-duplex hybrids. *Biochemistry*. **53** 247 (2014).

78. J. T. Grün, A. Blümler, I. Burkhart, J. Wirmer-Bartoschek, A. Heckel & H. Schwalbe. Unraveling the Kinetics of Spare-Tire DNA G-Quadruplex Folding. *Journal of the American Chemical Society* **143** (2021)

79. Y. Li, C. Liu, X. Feng, Y. Xu & B-F. Liu. Ultrafast Microfluidic Mixer for Tracking the Early Folding Kinetics of Human Telomere G-Quadruplex. *Analytical Chemistry* **86** (2014)

80. A. Rajendran, M. Endo, K. Hidaka & H. Sugiyama. Direct and Single-Molecule Visualization of the Solution-State Structures of G-Hairpin and G-Triplex Intermediates. *Angewandte Chemie International Edition* **53** (2014)