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

A. Sáinz-Agost[1,2], F. Falo[1,2], 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 study the different conformations of these structures through their Free Energy Landscape at different temperatures. All-atom simulations are adopted according to a mesoscopic G-quadruplex model previously developed by our group. We use a small number of significant reaction coordinates to analyze the evolution of the system by applying two dimensionality reduction techniques: Principal Component Analysis (PCA) and time-Independent Component Analysis (tICA). The data of the trajectories of the system in this reduced space are encoded into a Complex Markov Network which, in conjunction with an Stochastic Steepest Descent, provides an hierarchical organization of the different nodes into basins of attraction, so revealing the main intermediate states and the most relevant transitions the system undergoes in its denaturation path.

Author summary

G-quadruplexes are fascinating structures formed in guanine-rich regions of DNA and RNA, and they are key players in gene regulation and genome stability. Their ability to promote or block biological processes makes them potential targets for therapies.

In this study, we explored how G-quadruplexes evolve and melt at different temperatures. Through different data analyses of our molecular simulations, we mapped out their observed denaturation pathways together with their occurrence of their intermediate states. Specifically, with the use of Complex Markov Networks we encoded the trajectories into networks able to clearly reveal the main intermediates and quantify the most relevant connections between them, providing detailed information about the structural denaturation pathway.

1 Introduction

G-quadruplexes (G4s) are secondary, non-canonical structures arising from either one or multiple guanine rich DNA or RNA chains, and relatively 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 relatively 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 the over or under exposure of different binding sites of interest upon adoption of a G-quadruplex topology. The formation of these structures at the moment of DNA replication can lead to the interruption of this process, destabilising the helicases [11] and combining with other factors, can lead to epigenetic instability [12].

Similarly, overabundance on guanine in mRNA can lead to difficulties in protein translation in the ribosome. Several works have researched 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 molecule, the latter called intermolecular G4s) and the particular twists and turns present in their backbones. 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. Conformations of these unimolecular quadruplexes are characterized by the relative orientation of the guanine backbone; in the case of the parallel structure, all guanine tracts are oriented in the same direction, since the loops connect the top of a guanine track with the bottom of the next. This conformation is depicted on Fig. 1.

The behavior under force and mechanical stability of these structures 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 interest, due to the involvement of these structures in genetic regulation.

Molecular Dynamics (MD) simulations have proven to be a valid and powerful tool in examining the thermal and mechanical stability of biomolecular structures, such as G-quadruplexes, revealing complex energy landscapes and characterizing slow folding kinetics [22, 23].

Building on this approach, we used *Replica Exchange Molecular Dynamics* (REMD) [24, 25] to study the evolution of our system through the parallel simulation of 8 copies of the human parallel G-quadruplex (PDB: 1KF1) [26] in a temperature range close to its 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 previously published mesoscopic model [17, 18] was applied, followed by dimensionality reduction techniques, specifically Principal Component Analysis (PCA) [27, 28] and time-lagged Independent Component Analysis (tICA) [29]. 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) [30–33] combined with an stochastic steepest descent algorithms, we constructed a series of networks describing the unfolding process. These networks outlined the main conformations the system adopts as it unfolds, identifying stable and transitory intermediates. The description provided by

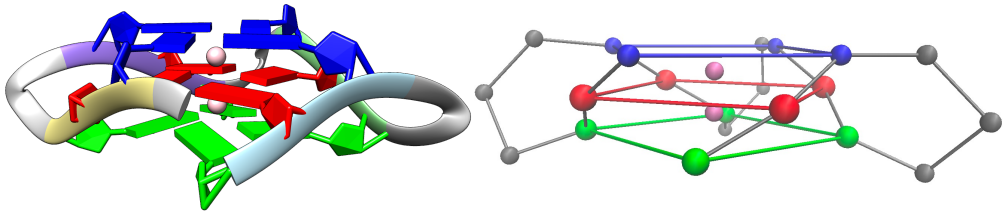

Fig 1. On the left, parallel conformation sourced from the PDB (1KF1, ribbon representation in Chimera). On the right, the resulting structure after the coarse graining procedure. The different G-tetrads on the structures are marked in different colors, the loops in grey and the ions pink. For the PDB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

tICA proved to be more straightforward and 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, as well as the different dimensionality reduction approaches taken, 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, and additional information regarding both methods and results can be found in the Appendix, beginning with Section 5.1.

2 Methods.

2.1 All-atom simulations with Gromacs.

The initial structure for the simulation was retrieved from the Protein Data Bank; the intramolecular parallel DNA G-quadruplex (1KF1) [26], formed by TTA repeat units, as 5′-AGGGTTAGGGTTAGGGTTAGGG-3′.

The simulations were carried out using the software Gromacs [34, 35]. For the force field we took the amber modification Parmbsc0 [36], which has been extensively used for both DNA [37] and G-quadruplex analysis [6] and proven to be the best option for our calculations [38], although other options, such as OL15 [39] or Parmbsc1 [40], can be considered. Control measures to verify the stability of the complexes in normal conditions under the force field of interest were carried, with the expected results.

The structures were solvated using Tip3P water model [41] in a periodic cubic box 1 nm larger than the diameter of the DNA body. Afterwards, more K^+ ions were introduced in the simulation box to counteract the negative charges present in the DNA chains, resulting in a concentration of 0.14 M on average for the potassium ions (K^+), with no negative ions added to the mix.

The energy of the solvent was minimized using the 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.

In order to be able to correctly observe the different unfolding intermediates of the structures, the simulations are performed at temperatures near the melting point of the parallel G4, being around 65°C [42]. At these temperatures, the unfolding kinetics were notoriously slow. In fact, achieving a successful unfolding process in a single run has proven to be quite challenging, given the typical extension of all-atom simulations.

In an effort to significantly increase the sampling of the free energy landscape characterizing the structures, thus enhancing the likelihood of observing an unfolding process, we introduced the *Replica Exchange Molecular Dynamics* (REMD) method, also called Parallel Tempering [24, 25], used to circumvent similar issues [43]. 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) [44–46], 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, such as the Metropolis criterion [47] in our specific case. This interchange of conformations at diverse temperatures aids the system in overcoming high energy barriers that would otherwise impede to explore of the free energy landscape.

We constructed eight replicas, each 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 surpasses that of the original measurements, a fact that has been shown to increase the melting temperature T_m [48]. 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 [44–46], which indicates the autocorrelation decay time of the potential energy of the whole system as the minimum exchange time. To estimate this time, 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 Annex.

2.2 Mesoscopic discretization of the model.

The aim of this work is to extract the main conformations the G-quadruplexes adopt along their 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 try to focus on only 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 single identical beads. The positions of these beads were determined as the center of mass for 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 [49, 50]. The resulting structures are depicted on the left side of Fig. 1.

2.3 Dimensionality reduction techniques.

Following the adjustment in coordinates, 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 [51, 52] 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, firstly introduced in [27, 28], relies on the solution of an eigenvalue problem involving the covariance matrix of the input data.

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

Where $C(0)$ is the covariance matrix $C(0)_{ij} = \sum_t x_i(t)x_j(t)/N$ of the normalized (mean 0, unit variance) data, V contains the eigenvectors v_i by columns, and Λ 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 Appendix, Section 5.1.

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 big reduction in the magnitude of the eigenvalue λ_{n+1} when compared to λ_n, meaning that most of the relevant system information (which corresponds to the majority of the variance) resides within the preceding 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 system. It is crucial to note that this assumption holds true only when these transitions represent the most significant alterations in the system's variance, 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, which we present here below.

2.3.2 time-Independent Component Analysis

tICA is a method first introduced in [29] 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 $\mathrm{t} + \tau$, 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 Section 5.1.

The method relies on the resolution of a generalized eigenvalue problem involving the covariance matrix of the data and the time-lagged correlation matrix $C(\tau)$, similar

to the former but correlating the data at times t and t + τ:

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

where Γ is the diagonal eigenvalue matrix, and W the eigenvectors matrix, with each of the different w vectors as columns. This general 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 [53] is typically used in its place, detailed in Section 5.1.

It has been shown [54] 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 based on their timescales, and may uncover certain intermediates that PCA does not consider, since the change in variance they induce is not remarkable.

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 [30–33]. They consist of a series of nodes connected via directed links.

For the construction of a CMN the n reduced coordinates, obtained from either PCA or tICA, are discretized each 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 will given by the number of times its associated combination of intervals is visited by the trajectory of the PCAs corresponding to each interval i_{m}, 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 [30], groups the nodes into basins, which represent different attractors of the trajectory. Additionally, a filtering based in a node weight cutoff can be introduced with the goal to retain only fundamental basins, a topic to be discussed in more detail in the Result section.

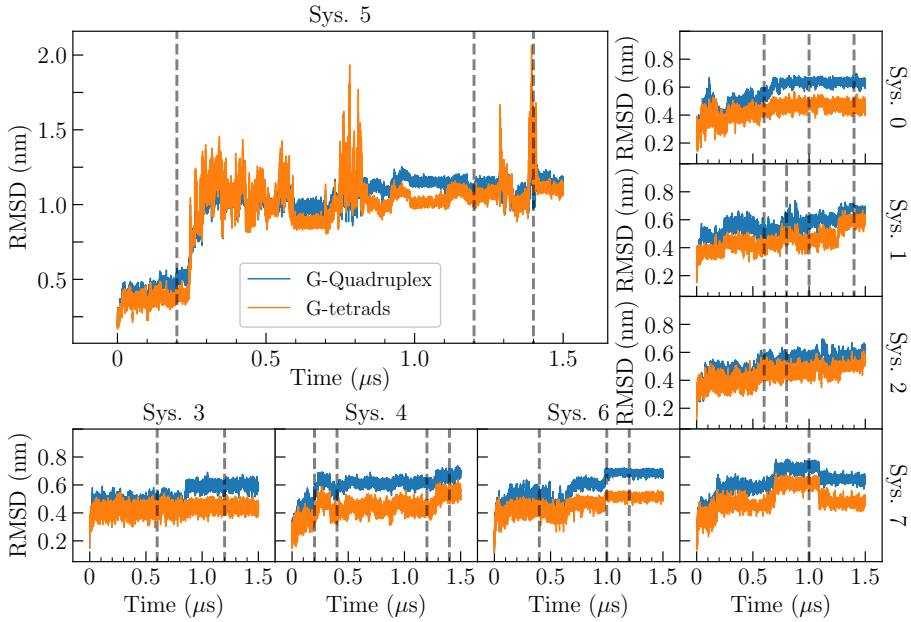

Fig 2. RMSD calculated over the different starting configurations (Sys.) of the parallel G-quadruplexes. In blue, the RMSD of the whole structure. In orange, the RMSD taking into account only the guanines forming the planar arrangement. The vertical dashed lines correspond to successful exchanges between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

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 indicates 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 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: we choose to, instead of tracking the behavior of a replica whose contents are continuously changing over time, 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 succesful exchanges between replicas, when a Sys. changes temperature, are marked by grey vertical lines in Fig. 2. The RMSD curves by replica can be found in the Appendix 5.1.

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

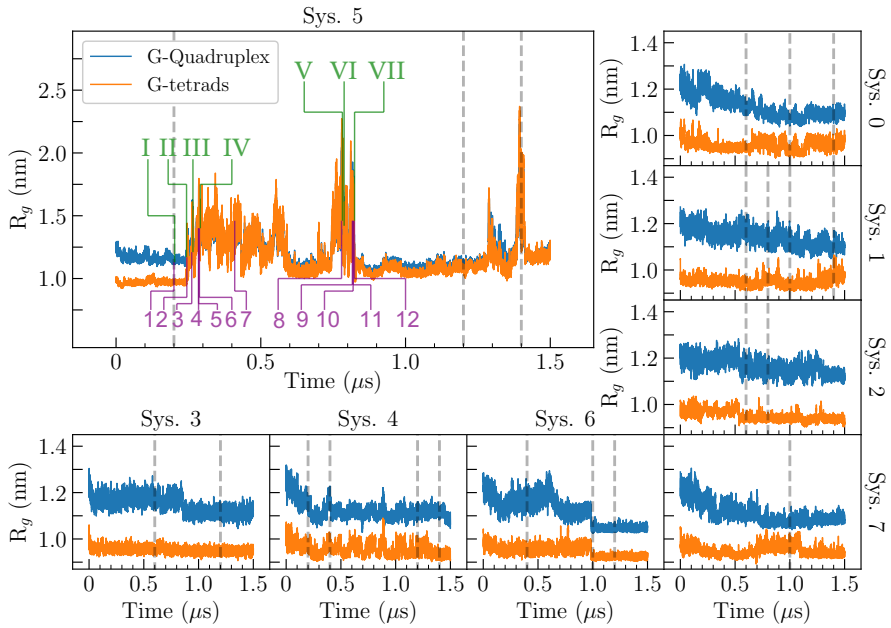

Fig 3. Radius of gyration computed over the different starting configurations (Sys.) of the parallel G-quadruplex. In blue, the R_g of the whole structure. In orange, the R_g for the G-tetrads. Sys. 5 has been enlarged, since it is the one considered for posterior analysis due to the presence of unfolding.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

line in the panels) or absence (G-tetrads, orange line) of the loops in the calculation. Sys. 4 undergoes great deviations from the norm, reaching values up to 2 nm independently of the loops, thus confirming an unfolding event. These results further confirm 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 the ions from the central channel, occurrence described in the next Section 3.1.1.

These same results can be understood through second metrics used in this study, the *radius of gyration* (R_g) that 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 Sys. 4, indicating the unfolding of the structure already seen in Fig. 2.

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 resides 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

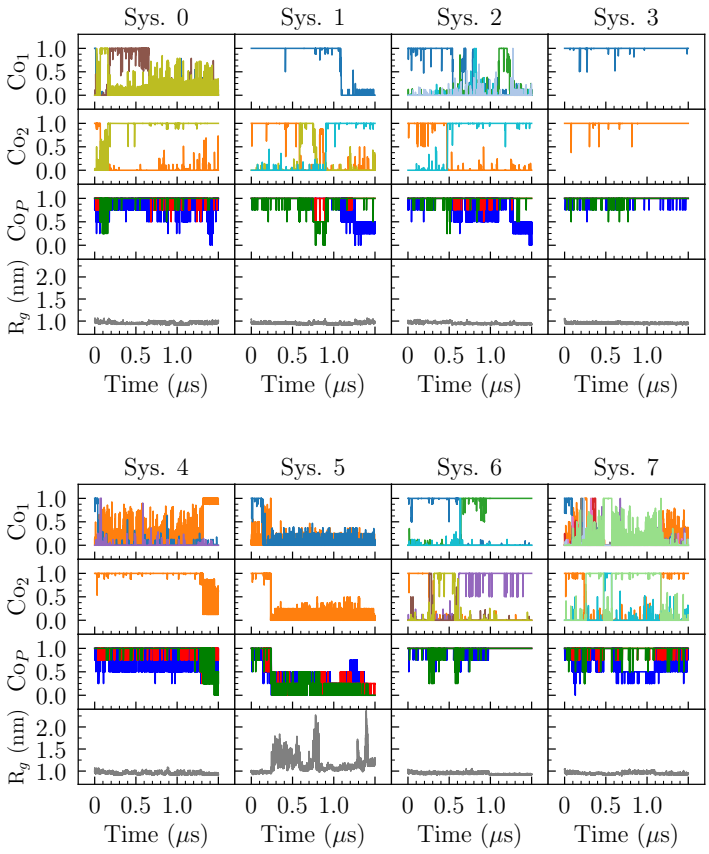

Fig 4. Coordination numbers of the top (Co_1) and bottom (Co_2) positions available for the ions, coordination number of the three G-tetrads (Co_P, blue for the top plane, red for the middle plane and green for the bottom plane) and radius of gyration of the G-tetrads (loops excluded). Only ions which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

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 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 System 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 manage to reincorporate other ions present in the simulation box. Typically, these 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 $\mathrm{R_g}$. 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 displays 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 $\mathrm{Co_{P1}}$. 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 $\mathrm{Co_{P2}}$. Subfigures **C** and **D**, taken 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, $\mathrm{Co_{P1}}$. 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 $\mathrm{Co_{P3}}$. 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 $\mathrm{Co_{P3}}$, with a new maximum value of $\mathrm{Co_{P3,\ max}} = 0.75$.

From this 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.

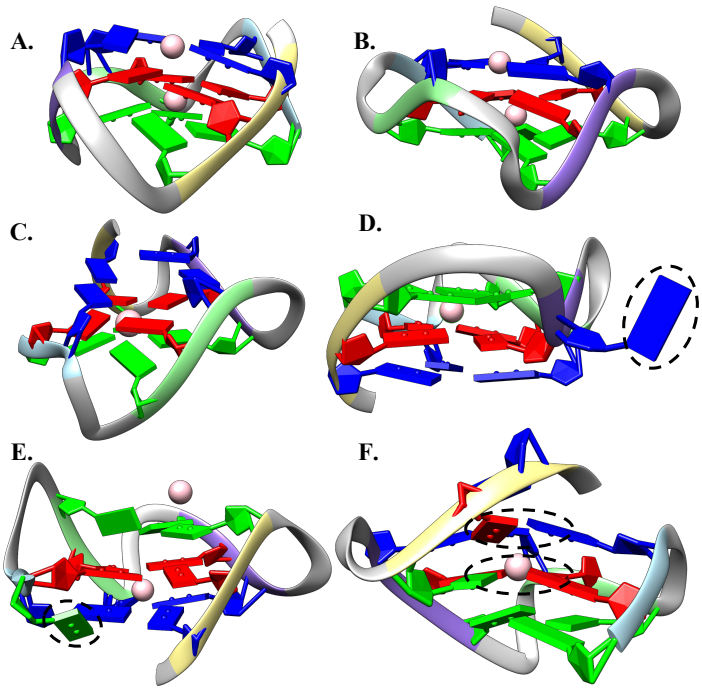

Fig 5. Relevant conformations associated to ion dynamics in the simulation. **A.** Escape of an ion from the top plane. **B.** Absorption of an ion through the top plane. **C.** Deformation of the top plane after ion loss. **D.** Separation of a single guanine from the tetrad of the top plane (indicated 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.

In the remaining systems, ionic loss typically occurs in only one of the two available cavities, and losses of both ions are usually brief. 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 is mainly depicted by means of the Sys 4, shown in Figs. 2 and 3. The identification of peculiar states of the trajectory will be largely developed with the analysis with tICA and PCA, later on presented.

In Sys. 4, 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 R_g and RMSD in Figs. 2 & 3 at $t = 0.25\,\mu$s.

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 [55, 56], one of the common intermediates in G-quadruplex unfolding [57, 58]. During the next 30 ns several attempts at refolding are made unsuccessfully. Moreover the arrow 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 [59, 60]. Other transient conformations can appear, such as G-hairpins [57, 61–63], 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 distance between the third and fourth tracts. Eventually the fourth strand rotates over itself and binds with the third through the bottom guanine (green), corresponding to the plateau between $t = 0.6\,\mu$s and $t = 0.75\,\mu$s.

After the time $t = 0.75\,\mu$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" [63]. 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 $\mathrm{R_g}$, and another presenting the formation of Hoogsteen bonds between their guanines, with smaller $\mathrm{R_g}$. 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 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 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 has been rescaled by applying a translation making a superposition of the centre of mass with the native structure, and applying a 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, 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 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 are combinations of the initial coordinates. Thus, the bigger the eigenvalues, the more relevant that direction is to the overall description of the dynamics. Ideally, one should find an eigenvalue or a 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 a few reaction coordinates. The products of the diagonalization of the correlation matrix are contained in the left side of Fig. 7.

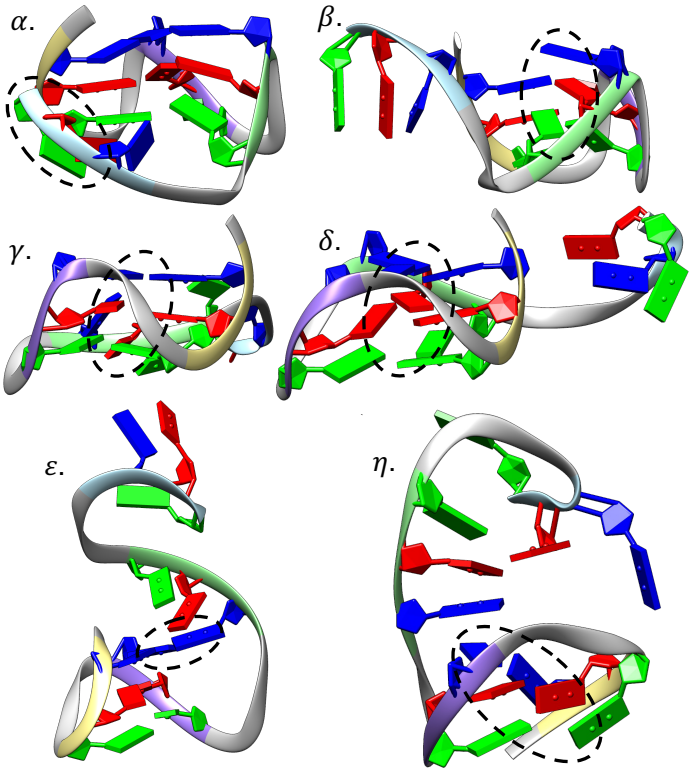

Fig 6. Highlighted conformations achieved during the evolution of the system. ***α***. Initial detachment of the fourth (icy blue) tract of guanines. ***β***. Example of a G-triplex conformation with a slip-stranded effect between tracts 1 (yellow) and 3 (green). ***γ***. Formation of a G-hairpin between tracts 1 (yellow) and 2 (magenta). ***δ***. Refolded triplex after re- attachment of tract 1. ***ε***. Stable intermediate state of G-hairpin (tracts 1 and 2) with stacking interaction of the third tract. ***η***. Final state of the system, with tracts 1 and 2 in a cross hairpin conformation.

The first four eigenvalues are clearly bigger 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) of the Appendix.

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

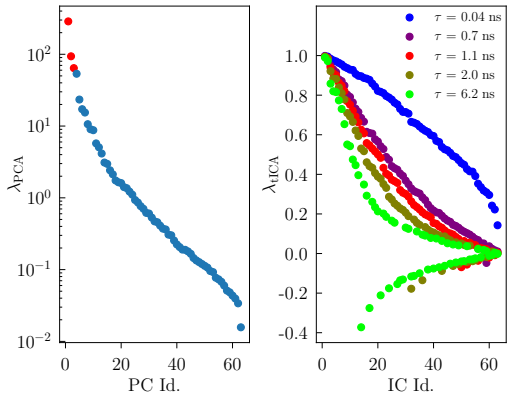

Fig 7. On the left, PCA eigenvalues obtained from solving (1), with the vertical axis in log-scale. The three eigenvalues in red correspond to the combinations chosen for data projection. On the right, tICA eigenvalues obtained from solving (2) for different values of the time window τ, seen in the legend.

into a few simulation instants 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 projected on the eigenvectors associated to the three chosen eigenvalues. On the right side, the the histogram of the corresponding coordinates occupation for both methods: PCA and tICA. The analysis of Fig. 8 reveals

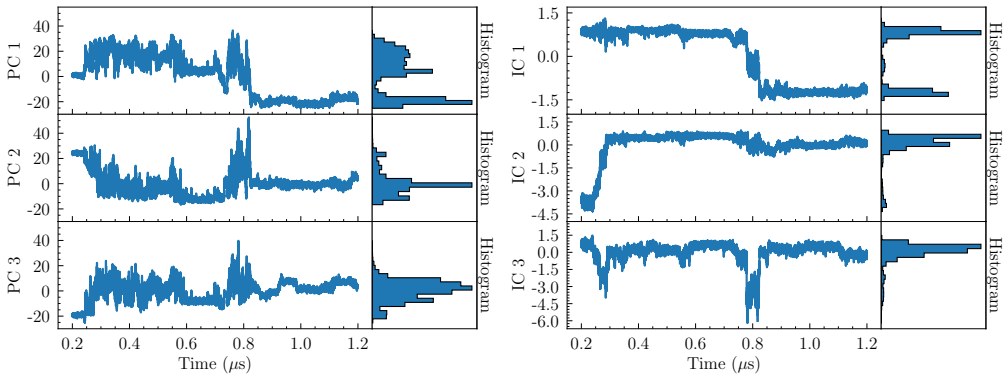

Fig 8. Projected trajectory from Rep. 4 between its exchanges at 0.2 and 1.2 μs onto the first three eigenvectors (left), along with the histogram of the coordinates (right). PCA on the right, tICA with $\tau = 0.7$ ns on the left.

appreciable differences between the two methodologies. PCA yields trajectories characterized by histograms exhibiting broad Gaussian-like peaks, whereas tICA produces trajectories with narrower distributions than PCA, 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: initial high value for $t = 0$ and $t = 0.6\,\mu$s, and after that a narrow transition to another constant value. Both trajectories describe the general behavior of the fluctuations in the root-mean-square deviation (RMSD) trajectory (see Fig. 2) with the difference that tICA presents a small variability with respect to PCA,

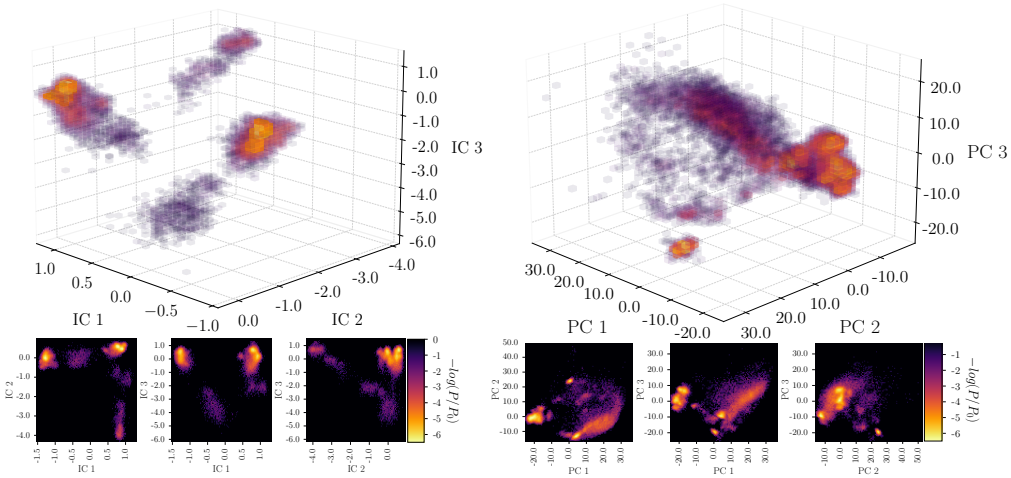

Fig 9. Colormaps depicting the correlation of the different coordinates emerging from PCA (left) and tICA (right). Within each subfigure, the bottom three panels depict heatmaps illustrating the pairwise relationships between the projected coordinates, with the color representing the relative occupancy 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.

that, instead, almost reproduces the same shape as the original behavior, just reduced.

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 tICA (top) and PCA (bottom), with the colored spots corresponding to occupied states in the trajectory. The colors themselves reference the basin each point belongs to, which is further expanded in later paragraphs. On the bottom of each 3D plot there are three heatmaps plotting the coordinates against each other by pairs. Their color scale is determined by the relative occupation of specific coordinate combinations during evolution, computed as $-\log(P/P_0)$, with P_0 being the sum of the occupation of all the states.

From Fig. 9 it can be seen 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 resultant networks of basins depicting the unfolding process are illustrated in

Figs. 10 and 11.

3.3.1 tICA analysis.

The tICA analysis summarized in Fig. 10 shows a linear representation of the unfolding process, summarized in 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 formed a single Hoogsteen bond with the top guanines (blue) of t, as described in Fig. 6$\boldsymbol{\epsilon}$. This basin represents the first relatively stable state of the system, corresponding to the time interval [0,3-0,7] μs in Figs. 3 & 2. The fluctuations of these magnitudes in that time interval correspond to waving movement 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, braking the G-hairpin and making it he 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 typically one or more Hoogsteen bonds formed between them. This configuration is relatively stable, with presence fluctuations caused by 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 an 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 and weighted. In fact they represents the multiple transitions between the different basins with no information about a particular time sequence. This is due to the fact that, in principle, the system investigated lies in a dynamical equilibrium condition. This means that if the trajectories are long enough, they can be analyzed backwards in time, and the results would reproduce exactly the same pattern configurations.

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 using two approaches: one based on Complex Markov Network self-loops and another analyzing the time spent in a given

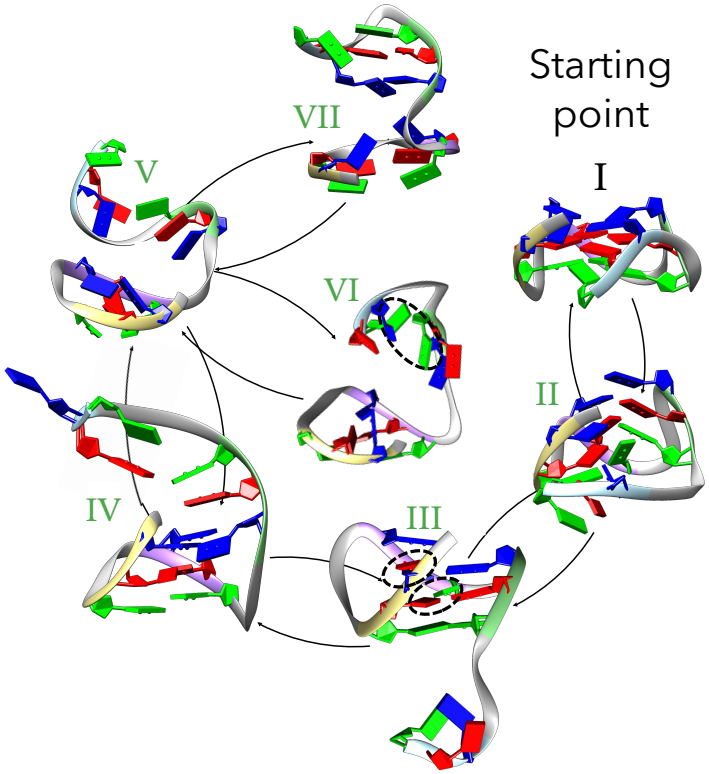

Fig 10. Network of basins found by tICA, $\tau = 0.7$ ns. The green numbers refer to the equally colored lines in Fig. 3, showing the position of the basins along the trajectory.

conformation before transition in the trajectories. The detailed methodology as well as the results from these calculations can be found in Appendix 5.3; 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 included in Appendix 5.4.

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 here severely more challenging 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 identification of some equivalent states as different, so 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 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 minimum unfolding path akin to that offered by tICA, resulting in 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 Appendix 5.3. Furthermore, the PCA analysis without the mesoscopic reduction is also found in the Appendix 5.4, leading to a higher number of basins with increased degeneration between them.

4 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

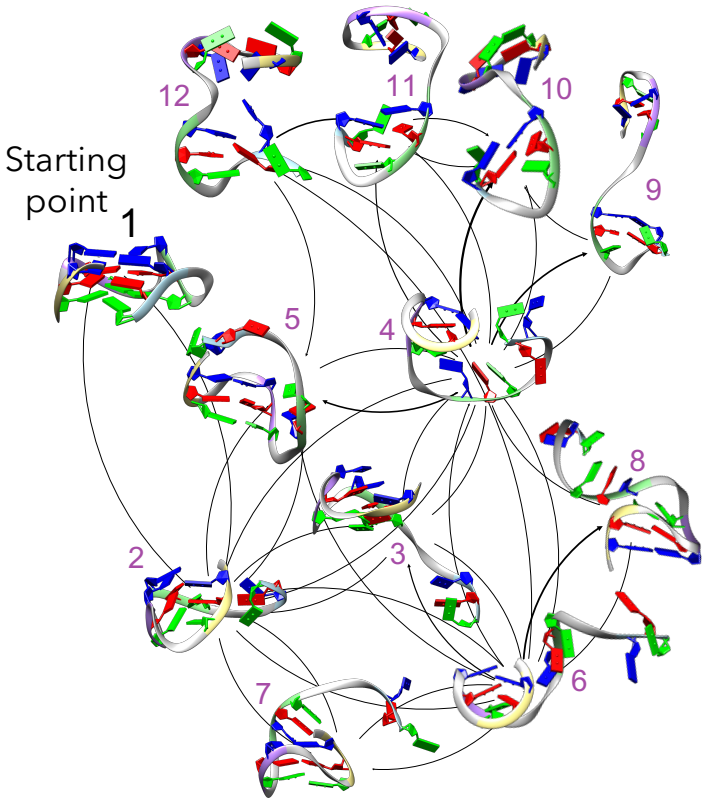

Fig 11. Network of basins found by PCA. The purple numbers refer to the equally colored lines of Fig. 3, indicating the position of the basins along the trajectory.

performed with the full system coordinates reports none significant change (Appendix 5.5). 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 [64].

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-quadruplexs unfolding pathway.

These denaturation pathways have revealed a series of either transient or stable intermediates, displaying elements previously observed in the literature, such as G-triplexes [55–58], G-hairpins [61–63], slip-stranding effects [59, 60] and cross-hairpins [63], further validating our results.

A particularly long lasting intermediate containing a G-hairpin (0.4 μs out of 1.5μs total) has been revealed by our simulations (basin #IV in tICA, basins #7 and #8 in PCA). 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.

Importantly, as the simulations have been performed in biologically significant range of temperature around 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.

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 [65, 66], 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 appears of a G-quadruplex conformational switch between parallel and either antiparallel or hybrid shapes, due to the unfolding process. This possibility remains to be explored both *in silico* and *in vitro*.

5 Appendix.

5.1 Mathematical description of PCA/tICA.

PCA. Consider a dataset $\mathbf{r}(\mathrm{t})$, encompassing information regarding the system's positions at various time instances, with the mean already subtracted from each coordinate. Our goal is to derive a linear combination of this input data that maximizes the variance.

$$\mathrm{Var}(\mathbf{r}v) = \mathrm{Var}\left(\sum_i v_i r_i\right) \text{ is maximal}$$

Where v is the vector containing the coefficients of the linear combination that we aim to obtain. From the properties of the variance, this expression can be rewritten as

$$\begin{aligned}\mathrm{Var}(Ra) = v_1^2\mathrm{Var}(x_1) + v_2^2\mathrm{Var}(x_2) + \\ 2v_1v_2\mathrm{Covar}(x_1x_2) + \cdots = v^TC(0)v\end{aligned}$$

Where $C(0)$ corresponds to the covariance matrix of $\mathbf{r}(\mathrm{t})$, the 0 denoting that the covariance is calculated between the values of $\mathbf{r(t)}$ at the same time instances. This notation is relevant, and will come in handy later. The problem of obtaining the optimal values for v can be solved employing Lagrange multipliers. Introducing the restriction that v must be a unit vector, i.e. $v^Tv = 1$, we have:

$$v^TC(0)v - \lambda(v^Tv - 1) = 0 \overset{\partial/\partial v}{\Longrightarrow} C(0)v = \lambda v \tag{4}$$

Thus, the problem of obtaining the values of v is, in fact, an eigenvalue problem of the covariance matrix, in which v is the eigenvector and λ the eigenvalue. The latter also coincides with the variance of the data projected into the direction marked by v. Since the covariance matrix is symmetric and real, all of its eigenvalues will be real, and we can extend this procedure to obtain N vectors. Generalising:

$$C(0)V = \Lambda V \tag{5}$$

Where V is matrix containing the eigenvectors v_i by columns, and Λ is the diagonal eigenvalue matrix. Then, projecting our input data into the new coordinates, we obtain the *Principal Components*: $\mathbf{x}^T(t) = \mathbf{r}^T(t)V$.

tICA. Given a value of τ, we want the autocorrelation of the projection between t and $t+\tau$ to be maximal, defined as

$$\begin{aligned}\mathrm{Autocorr}(\tau, \mathbf{r}w) = \langle \mathbf{r}w(t)\mathbf{r}w(t+\tau)\rangle = w_1^2\langle r_1(t)r_1(t+\tau)\rangle \\ +w_2^2\langle r_2(t)r_2(t+\tau)\rangle + w_1w_2\langle r_1(t)r_2(t+\tau)\rangle \\ +w_1w_2\langle r_2(t)r_1(t+\tau)\rangle + \cdots = w^TC(\tau)w\end{aligned}$$

Where $C(\tau)$ is the time-lagged correlation matrix, whose element ij is given by

$$C_{ij}(\tau) = \frac{1}{N-\tau-1}\sum_t^{N-\tau} r_i(t)r_j(t+\tau)$$

In order to eliminate the effect of the variance in the eigenvalues of the system and be able to select modes based on timescale alone, our restriction is to force our coordinates to have unit variance: $w^TC(0)w = 1$, leaving

$$w^TC(\tau)w - \gamma(w^TC(0)w - 1) = 0 \tag{6}$$

Corresponding to the generalized eigenvalue problem of

$$C(\tau)W = C(0)W\Gamma \tag{7}$$

Where Γ is the diagonal eigenvalue matrix, and W the eigenvectors matrix, with each of the different w vectors as columns.

It has been shown [54] 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:

$$t_i = -\frac{\tau}{\log|\gamma_i|} \tag{8}$$

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.

The general 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 [53] is typically used in its place, consisting of several steps:

1. Subtract the mean from the data: $\mathbf{r}'(t) = \mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t$.
2. Compute the PCA of the data, and project it into the obtained eigenvectors:

$$C(0)V = \Lambda V \;\rightarrow\; \mathbf{x}^T = \mathbf{r}^T V \tag{9}$$

3. Normalise the newly obtained principal components by their corresponding eigenvalue: $\mathbf{x}'(t) = \Lambda^{-1}\mathbf{x}(t)$
4. Build the time-lagged covariance matrix of $\mathbf{x}$'(t), symmetrize and diagonalise it:

$$C(\tau)U = \Gamma U \tag{10}$$

   Note that the eigenvalues obtained in this step coincide with those in (2), but the eigenvectors do not. Since the latter are the coordinates of W in the space defined by $V\Lambda^{-1}$, it can be verified that $W = V\Lambda^{-1}U$.
5. Project $\mathbf{x}'$ onto these new eigenvectors, $\mathbf{y}^T = \mathbf{x}'^T U$. This is the projection of our original data into the tICA-defined coordinate system.

Identically to PCA, a few components of $\mathbf{y}(t)$, the ones corresponding to the larger timescales, can be employed to study the whole system.

5.2 Simulation figures by replica ID.

Figs. 12 & 13 contain the evolution of the RMSD and R_g respectively, for the different replicas. The results, as the ones in the main text, are shown using the full G-quadruplex in the calculation (blue lines) and using only the G-tetrads, excluding the loops (orange lines). The grey vertical dashed lines indicate successful exchanges between them, with the destination replica shown above them.

The unfolding event characterized by a sudden increase in RMSD and R_g begins in Rep. 4, corresponding to Sys. 5 after the exchange at 0.2 μs. The high values in the graph remain up to 1.2 μs, in which an exchange between Reps. 4 & 3 takes place, leading to the sudden increases in the latter. This information could be mistakenly interpreted as an additional unfolding event in Rep. 3, thus the choice in presentation in Figs. 2 & 3.

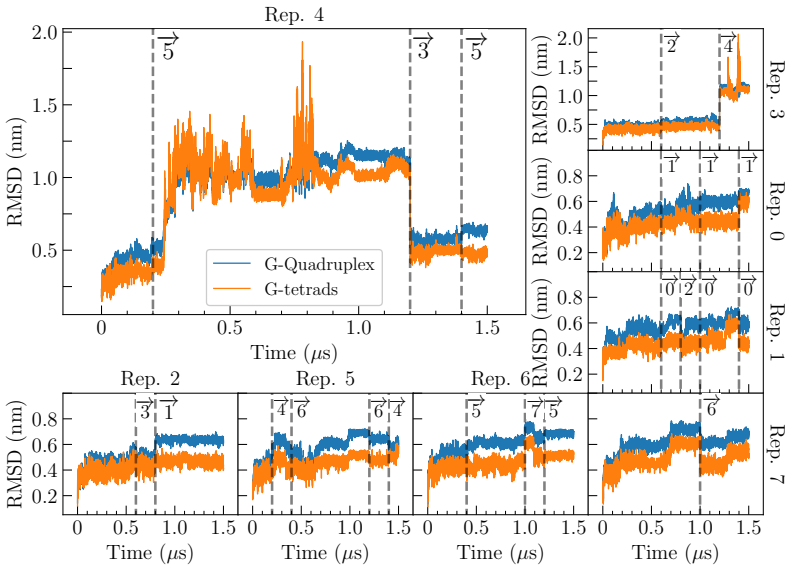

Fig 12. RMSD calculated over the different replicas (Rep. ID in each graph) of the parallel G-quadruplexes. In blue, the RMSD of the whole structure. In orange, the RMSD taking into account only the guanines forming the planar arrangement. The vertical dashed lines correspond to successful exchanges, with their destination replica shown on their right under an arrow. Rep. 4, since its the one used for the analysis, has been enhanced.

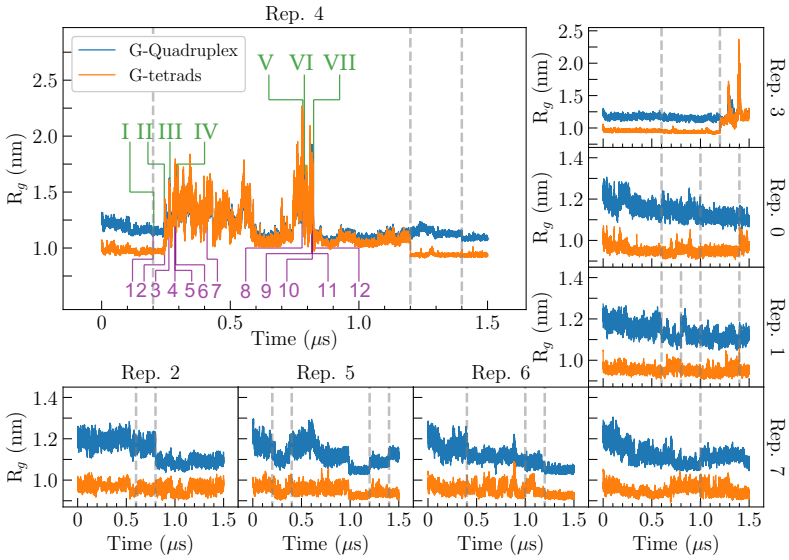

Fig 13. Radius of gyration computed over the different replicas of the parallel G-quadruplex. In blue, the R_g of the whole structure. In orange, the R_g for the G-tetrads. The fourth replica has been enlarged, since it is the one considered for posterior analysis. The vertical grey dashed lines correspond to successful exchanges, with their destination replica shown with a number under an arrow on the right of the lines. The purple and green lines correspond to the states identified by PCA and tICA respectively, and are further explained in later paragraphs.

5.3 Basin escape times.

Each of the basins found during our analysis corresponds to a significant conformation the system adopts at some point in its unfolding trajectory. The stability of this conformations is a matter of interest, allowing us to classify them into either stable intermediate steps or fast-lived (although necessary) transition states, contributing information to the energy landscape of G-quadruplexes.

One possible measure to study the relative stability of these basins is the average escape time, *i.e.* the typical time the G-quadruplex remains in a particular conformation before transitioning into another state. There are two different methods available for the estimation of this quantity, resulting in two different values of the escape time.

The first method calculates the escape time from the results of the Complex Markov Networks. Specifically, the escape time (t_{CMN} for this method) can be estimated using the self-loop probabilities associated with each basin, whose weight represents the probability of remaining in the same state during the next timestep,

$$t_{\mathrm{CMN}} = \frac{\Delta t}{1 - P_{\alpha\alpha}} \tag{11}$$

where Δt is the time resolution of the trajectory used to construct the CMN (20 ps), and $P_{\alpha\alpha}$ represents the self-loop probability of basin α.

The second method relies on computing the times the trajectory remains into a given basin along the simulation (named t_{traj}). A series of histograms (one for each basin) of those times can be build. Assuming that the underlying distribution of the histograms is exponential, the average escape time should correspond to the mean of the data points. Thus, for basin β,

$$t_{\mathrm{traj},\beta} = \sum_{i=1}^{N_\beta} t_{i\beta} \tag{12}$$

where N_β represents the number of transitions from basin β to other basins, and $t_{i\beta}$ the time spent in basin β before changing to basin i.

The values of the escape times were calculated using both methods for the PCA and tICA results.

tICA. The results of the escape times for tICA can be found in Table 1. Both estimations of the escape times show increased escape time values on the first, fourth and seventh basins. This is in line with our results: the system remains an initial time in the original conformation (basin #I) before undergoing unfolding, a long time in the final conformation (basin #VII) and basin #IV was shown to be an stable intermediate, surviving for a large portion of the simulation. For the rest of the basins, a general hierarchy is difficult to stablish, since values are relatively higher or lower depending on the method used.

Basin	#I	#II	#III	#IV	#V	#VI	#VII
t_{CMN}(ns)	1000.0	0.17	0.75	1.23	0.50	0.97	24.69
t_{traj}(ns)	1.23	0.50	0.98	24.61	0.75	0.16	125.82

Table 1. Average escape time for each of the basins detected by tICA, with $\tau = 0.7$ ns. t_{CMN} is the time calculated from the autoloops of each basin, while t_{traj} is measured from the trajectory itself.

The difference between the values of both methods and possible underestimation of the escape times themselves are fundamentally linked to a lack of statistical data. Changes between basins are not numerous, and their values tend to be concentrated around low times. This affects both the times themselves and the self-loops.

Additionally, basins #I and #VII are affected by our choice beginning and end points of the trajectory. Since our analysis starts after the replica exchange at 0.2 μs and unfolding begins around 0.25 μs, the stability of the initial state of the trajectory is underestimated. The same happens for the final state of the trajectory, although in a lesser extent: trajectory analysis ends at 1.2 μs, but the system remains in that state an additional 0.3 μs, not taken into account due to the replica exchange taking place in between.

PCA. The results of the escape times for tICA can be found in Table 2. Similarly to tICA, both t_{CMN} and t_{traj} exhibit relatively high values on some basins. Particularly, basin #1, corresponding to the initial state of the simulations and, depending on the calculation of the escape time, basins #7 or #8, corresponding to the stable intermediate (basin #IV for tICA).

Basin	#1	#2	#3	#4	#5	#6
$\mathrm{t_{CMN}}$(ns)	63.50	0.09	0.09	0.08	0.50	0.16
$\mathrm{t_{traj}}$(ns)	8.91	0.23	0.15	0.16	0.62	0.09
Basin	#7	#8	#9	#10	#11	#12
$\mathrm{t_{CMN}}$(ns)	8.89	0.18	0.23	0.62	0.15	0.14
$\mathrm{t_{traj}}$(ns)	0.50	10.04	0.14	0.10	0.18	0.08

Table 2. Average escape time for each of the basins detected by PCA. $t_{\mathrm{auto.}}$ is the time calculated from the self-loops of each basin, while t_{traj} is measured from the trajectory itself.

Contrary from tICA, the final state of the system (basin #VII for tICA, basins #9 to #12 for PCA) is not identified properly, having a escape time virtually indistinguishable from the others. The cause for this phenomenon takes root on the fact that PCA, as explained in the Section 3.3, tends to identify different states as unique, when in reality they are functionally equivalent. Basins #9 to #12 correspond to the same conformation but are classified as different by PCA. This leads to 4 well interconnected states, that have reduced self-loops and lifetimes, leading to a decrease in t_{CMN} and t_{traj}.

This effect, in combination with the previously mentioned lack of statistical data regarding transitions between states, leads to a general underestimation of escape times, and both methods do not find a common hierarchy; some states are classified as more stable in one method respect to the other.

5.4 Relation between the states skipped and the resulting networks.

As commented previously in Subsection 3.3, a valid method to reduce the final number of basins identified in the building of the Complex Markov Network consist on, instead of taking every frame of the trajectory to build such network, only processing one of every X frames.

Note that this step is applied after the application of the dimensionality reducing algorithms in order to have a common ground on which to compare the results of the final network, and to prevent the creation of artificial coordinates not representative of the system at hand.

The idea of the method is that, when changing the number of frames read, we may avoid the oversampling of certain regions not of interest for the unfolding and/or conformational changes of the system (i.e. regions of relatively long-lived transient states) and thus potentially enhance the presence of states that, while short-lived, may be of importance to the observed phenomena. Obviously, a risk one can incur into when

applying this technique is to skip more steps than necessary; this eventually leads to the detection of a single basin of attraction. Fig. 14 contains the number of detected basins as a function of the number of skipped states in the trajectory for both PCA and tICA ($\tau = 0.7$ ns).

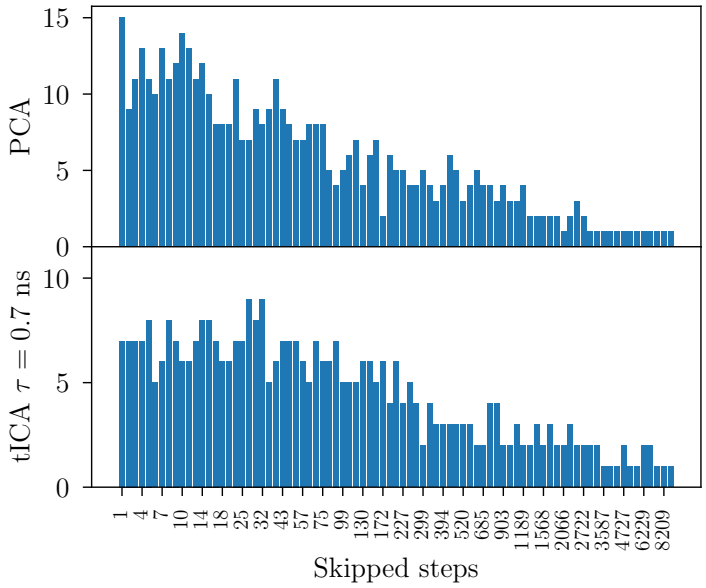

Fig 14. Number of basins in the final Complex Markov Network as a function of the frames skipped in its construction for both PCA (top) and tICA ($\tau = 0.7$ ns, bottom). As stated in the text, both methods eventually detect a single basin. The top limit for the number of skipped frames is set arbitrarily; a single frame was detected from then onwards.

Additionally, for both the PCA and tICA methods the final single basin detected corresponds to a native conformation. This is due to the fact that, when we skip enough steps, all of the subsequent recorded nodes of the network are connected to the first step of the trajectory, which is a native conformation.

However, analyzing low values of the number of skipped frames for both methods quickly reveals that the effect of this parameter is the opposite: while we do indeed reduce the number of basins detected, the native state of the system is heavily over-represented. This may be due to how the stochastic steepest descent works, making connections between unrelated nodes and grouping them under a single representative basin, corresponding to the native state. After seeing this results, no states were skipped in the construction of any of the networks.

5.5 Effect of the coarse-graining of the trajectories.

In our analysis method, the trajectories of the system obtained from the all-atom simulations in Gromacs are converted into a reduced number of coordinates, following the description of the G-quadruplex mesoscopic model [17]. However, given how later either PCA or tICA are applied onto said trajectories, the mesoscopization of the coordinates is not a necessary stage, but rather one we introduced as to reduce the possible influence of the relative orientations of the side chains onto the final results of PCA and tICA. Here we present the final networks for both methods without the mesoscopic discretization applied. The networks contained in Figs. 15 & 16 represent the results of the analysis without the application of the mesoscopic coarse-graining.

5.5.1 tICA ($\tau = 0.7$ ns)

The network uncovered by tICA is essentially the same as the one encountered in the main text: starting from the native folded conformation (basin #1) we observe the formation of the G-triplex, with a rogue guanine chain separated from the rest (nodes #2 and #3). That G-triad is further split, leading to a G-hairpin core and drifting guanines. However, different to the results after mesoscopization, here we have the coexistence of 3 interconnected states depicting the formation of said G-hairpin.

Node #4 describes the G-hairpin with a single Hoogsteen bond with the rotated guanine tract 3, corresponding to the top plane, node #5 describes the G-hairpin with that Hoogsteen bond already broken, and finally, node #6 shows the rotation of one of the guanine chains forming the G-hairpin into a situation similar to the one encountered in the previous state. Node #4 is only connected with node #6 since they are equivalent between each other.

After that, the network returns to more familiar territory; nodes 2 and 1 show the collapse of the G-quadruplex into a disordered lump after its complete denaturation, corresponding to the final state of the recorded trajectories.

Thus, in the case of tICA, no notable differences between the networks are recorded, save for the interconnection of the G-hairpin like states, which is not contemplated after mesoscopization. However, the links between nodes #4 and #6 seem rather fictitious; one would expect to first break the remaining Hoogsteen bond like the situation found in #5, and may reflect the effect that considering the side chains for the analysis may have in our system. All nodes identified in this network were around the same time intervals than the ones detected before by tICA.

5.5.2 PCA.

For PCA, the network increases significantly in complexity, having 17 different nodes well interconnected between them. Fig. 16 contains the pertaining network for this case.

This network fundamentally contains the same unfolding process (although with more nodes present) found in previous paragraphs. However, given the degree of interconection between the nodes, trying to extract an evolution over time of the state of system is extremely difficult, even more so than in the previous PCA network.

The evolution of the system begins with the native state at node 11. Afterwards, node 6, 5, 4 and 2 describe the transformation of the native state into the G-triplex, with every one of those nodes progressively describing the breaking of said G-triplex. Node 10 is the final stage of that structure, in which a slip has taken place (one chain of guanines presenting a relative displacement in the bonds between one another, leading to the destruction of overall one Hoogsteen bond).

Subsequently nodes 1, 3, 9 and 8 show the dynamics of the G-hairpin. The first three show the system of two guanine columns in various states of arrangement respect of the angle between bonds and the state of the rogue chains, while the last one shows the moment in which the hairpin is broken and one chain has rotated over itself, forming a series of Hoogsteen bonds but with the pairs interacting exchanges.

The rest of the nodes fall in some category in between those two, representing transient states found in the system. An example of this is node 17, which represents a hairpin with a single bond with another guanine, transitioning into a full hairpin.

The end of the simulation belongs to node 19, representing the moment in which the disordered G-quadruplex collapses onto itself, and remains there for an extended period.

This confirms that for PCA, which already had problems elucidating a clear unfolding path when the mesoscopic reduction was applied, only leads to worse results when omitting this step. Additionally, outside of the range around the beginning and the end of the simulation, no additional intermediate states were detected.

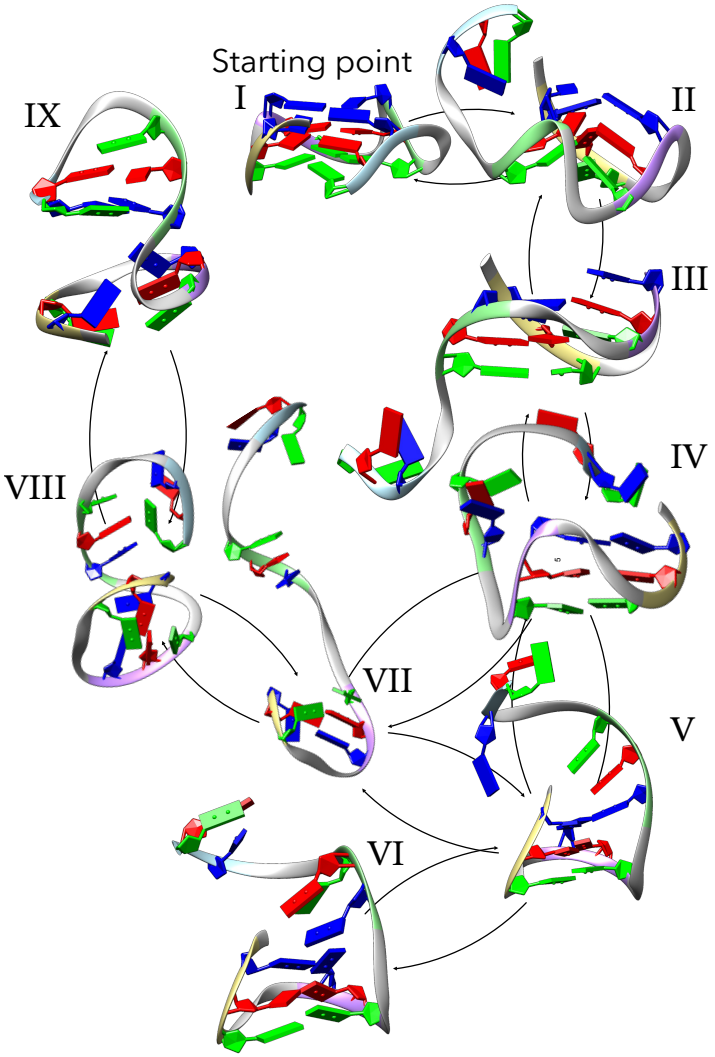

Fig 15. Network of basins found for the tICA method, with $\tau = 0.7$ ns, without the mesoscopic discretization of the trajectories. Note that the IDs on the nodes do not correspond to the ones shown before.

5.6 Autocorrelation time.

Simulations were performed to study the evolution and decay of the potential function of the lowest temperature replica (343 K) in order to establish the appropriate exchange rate. The decay of the potential function was fitted to an exponential and the results are contained in Fig. 17.

After the initial sharp decrease the potential energy decays at a much slower rate. The fit to an exponential reveals a decay constant $\tau_{\text{decay}} \approx 50$ ns, with the fit in clear agreement with the data. Thus, our choice of exchange rate between replicas of 0.2 μs is validated, and the Replica Exchange procedure is correctly performed, allowing for the equilibration of each copy before a swap is performed.

6 Competing interests

No competing interest is declared.

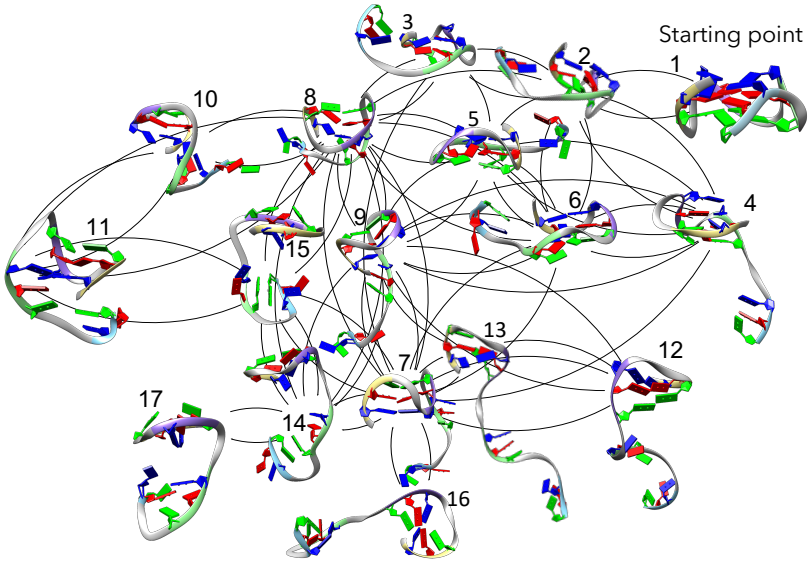

Fig 16. Network of basins found for the PCA method without the mesoscopic discretization of the trajectories. Note that the IDs on the nodes do not correspond to the ones shown before.

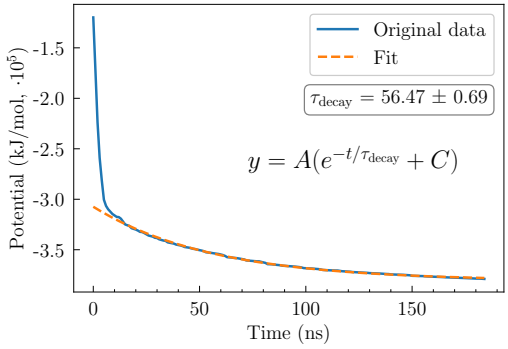

Fig 17. Evolution of the potential energy of the lowest temperature replica (343 K). In blue, the original data. In orange, the fit to an exponential function, whose expression and decay rate τ_{decay} are contained in the figure.

7 Author contributions statement

A.F. & F.F. conceived the project. A.S-A. conducted the simulations and the data analysis. A.F, F.F & A.S-A wrote and reviewed the manuscript.

8 Acknowledgments

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)'. A.S-A. also acknowledges the support of the predoctoral FPI fellowship PRE2021-100456 funded by the spanish Ministerio de Educación y Ciencia.

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. Shnel. 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. Morn, 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. Hnsel-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. Gutirrez, 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 GQuadruplexes: 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. S. Haider & S. Neidle. Molecular modeling and simulation of G-quadruplexes and quadruplex-ligand complexes. *Humana Press* **17** (2010) ISO 690

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

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

25. Y. Sugita & Y. Okamoto. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical Physics Letters* **314** 12 (1999)

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

27. 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)

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

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

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

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

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

33. 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)

34. 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)

35. H.J.C. Berendsen, D. van der Spoel & R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91** 4356 (1995)

36. A. Prez, I. Marchn, D. Svozil, J. poner, 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)

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

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

39. M. Zgarbov, J. poner, M. Otyepka, T. E. Cheatham, R. Galindo-Murillo & P. Jureka. Refinement of the SugarPhosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation* **11** 12 (2015)

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

41. 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)

42. J. L. Mergny, A. T. Phan & L. Lacroix. Following Gquartet formation by UVspectroscopy. *FEBS letters* **435** 74 (1998)

43. P. Pokorn, V. Mlnsk, G. Bussi, J. poner & P. Stadlbauer. Parallel G-quadruplex Folds via Multiple Paths Involving G-Tract Stacking and Structuring from Coil Ensemble. *Preprint, bioArXiv* (2023)

44. 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)

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

46. 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)

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

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

49. R. Gowers, M. Linke, J. Barnoud, T. Reddy, M. Melo, S. Seyler, J. Domaski, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. **98105** (2016)

50. N. MichaudAgrawal, 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)

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

52. 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)

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

54. G. Prez-Hernndez, F. Paul, T. Giorgino, G. De Fabritiis & F. No. Identification of slow molecular order parameters for Markov model construction. *The Journal of chemical physics* **139** 07B604_1 (2013)

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

56. M. Bonina, 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)

57. 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)

58. J. Tassilo Grn & H. Schwalbe. Folding Dynamics of Polymorphic Gquadruplex Structures. *Biopolymers* **113** 1 (2022)

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

60. I. Kejnovsk, P. Stadlbauer, L. Trantrek, D. Reniuk, M. Gajarsk, D. Krafk, J. Palack, et al. GQuadruplex Formation by DNA Sequences Deficient in Guanines: Two Tetrad Parallel Quadruplexes Do Not Fold Intramolecularly. *Chemistry A European Journal* **27** 47 (2021)

61. M. Gajarsk, M. Lenari ivkovi, P. Stadlbauer, B. Pagano, R. Fiala, J. Amato, L. Tomka, J. poner, J. Plavec & L. Trantrek. Structure of a Stable G-Hairpin. *Journal of the American Chemical Society* **139** 10 (2017)

62. P. Stadlbauer, P. Khrov, P. Ban, J. Koa, G. Bussi, L. Trantrek, M. Otyepka & J. poner. Hairpins Participating in Folding of Human Telomeric Sequence Quadruplexes Studied by Standard and T-REMD Simulations. *Nucleic Acids Research* (2015)

63. P. Stadlbauer, P. Khrov, L. Vicherek, P. Ban, M. Otyepka, L. Trantrek & J. poner. 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)

64. 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)

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

66. 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)