A pilgrim's guide to G-quadruplex nucleic acid folding

Valérie Gabelica

This is the author version (post-print version) of a commentary article published in
Biochimie 105 (2014) 1—3, http://dx.doi.org/10.1016/j.biochi.2014.08.007

A Pilgrim's Guide to G-quadruplex Nucleic Acid Folding

Valérie Gabelica[1,2]

1 Univ. Bordeaux, IECB, ARNA laboratory, F-33600 Pessac, France
2 INSERM, U869, ARNA laboratory, F-33000 Bordeaux, France

valerie.gabelica@inserm.fr

Every day, when I bike back home a little southwards of Bordeaux, I pass by a discrete post symbolizing "The Way of St James". If I would bike, from post to post, for a thousand more kilometers, I would finally arrive in Santiago de Compostela in Spain, going down a pathway followed by pilgrims from medieval times to this day. Translated to French, "The Way of St James" becomes "Les Chemins de St Jacques": it's a shift from the singular to the plural. And the routes to Compostela are indeed many, as pilgrims were coming from all over Europe to converge to their final destination. It is, among other significations, the convergence of these multiple routes towards a final goal that is symbolized by the shell of St James. Of course, depending on where you were coming from, there are preferred routes, nodes where several routes converge, and recommended landmarks to stop over for a rest, some of which being located just before major difficulties such as the crossing of the mountains between France and Spain.

Biomolecule folding routes are a lot like The Way of St James. The goal – the final state – is the folded structure. But the map – the free energy landscape – shows many different routes, converging to the final goal like a funnel. Protein folding pathways [1], and now RNA folding pathways [2], are nowadays usually described in this framework of folding funnels. The folding funnel concept is not to be opposed with the concept of "intermediates". However, for biopolymers, intermediates should not be viewed as single states (as a chemist would depict chemical reaction intermediates), but as ensembles of states sharing structural and kinetic properties. Intermediates are regions of the free energy landscape where a fraction of the molecular population gathers for a while before crossing a potential energy barrier.

Over the years, DNA folding had attracted less attention than protein and RNA folding, until other folding motifs that the Watson-Crick double helix came under focus. For example, G-quadruplex structures are formed by guanine-rich strands thanks to hydrogen bonding between four guanines that form base tetrads (the G-quartets), and thanks to cation-assisted stacking of successive G-quartets. Importantly, G-quadruplexes can interfere with replication, transcription and genome maintenance. However, for a given G-rich sequence, it remains difficult to predict which guanines form G-quartets with which ones and, with a few exceptions [3], to predict the resulting strand topology.

In particular, the human telomeric sequence, consisting of (TTAGGG) repeats, one of the earliest and most studied one given its importance for cell division and cancer, turned out to be one of the most compound DNA folding problems, even for oligonucleotides containing barely two dozens of nucleotides. To this day, no less than six intramolecular folding topologies (as distinguished by relative strand orientation, loop placement and syn/anti conformation of the guanines, see Figure 1,

south of the map) have been solved at high resolution. Which of these structures is the most stable in long telomeric repeats *in vivo* is still debated. Nor is it known whether the thermodynamically most stable structure is the relevant one in the cellular environment, or whether kinetically controlled structures could prevail. As a consequence, it is important to explore not only the structural and thermodynamic, but also the kinetic aspects of G-quadruplex folding. In other words, it is important to learn something about all plausible forms of G-rich sequences on their folding pathways towards a G-quadruplex. We need to learn more about the North part of the map as well.

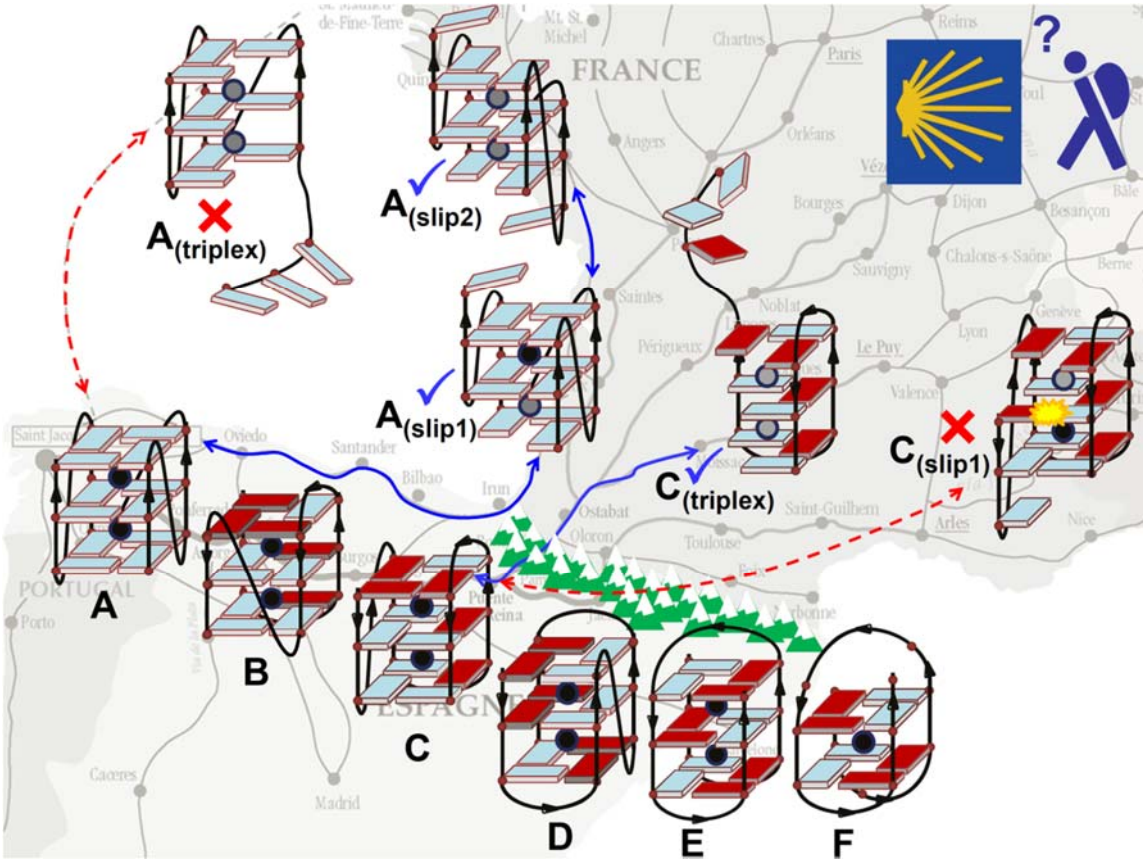

Figure 1: The folding pathways of G-quadruplexes are like the Way of St James: it's an extremely multi-pathways map with multiple possible stop-overs. **South of the map (A-F)**: The six high resolution stable structures of 4-repeat human telomeric G-quadruplexes reported to date. Each guanine base is represented by a rectangle, with dark red/light blue color code representing the syn/anti conformation, respectively. Four bases in a square plane form a G-quartet. Structures A-E have three G-quartets and structure F has two G-quartets. **North of the map:** some possible stop-overs explored by unbiased molecular dynamics: slipped structures were found plausible points of interest for the fully parallel, all-anti structure A whereas they were found impossible for hybrid structures due to steric clash between bases (see for example strand slippage from structure C) [4]. In contrast, triplex structures were found plausible for structures containing both syn and anti bases (see for example the triplex of C), but unstable for the parallel, all-anti structure derived from structure A [5]. The map of the Way of St James and the shell-shaped logo were reproduced from Wikimedia Commons [12].

There are several ways to explore the folding landscape and find points of interest, where the molecules would stop for different amounts of time (a pleasant meadow in a basin, a safe shelter before a difficult mountain pass, a detour to a dead-end maybe), or definitely (the final destination). The approach that made possible the characterization of six folded structures was to modulate slightly the base sequence or the environment (nature of monovalent cations, crowding conditions for X-ray cristallography), so as to slightly tilt the free energy landscape and drive nearly 100% of the population to the same end point. In such conditions, atomic structural investigation by NMR or X-ray crystallography is possible. Of course, whether these tilted landscapes represent the landscape really at stake *in vivo* is a matter of debate.

Now let's explore the folding routes at some further distance from the final destination. A first possibility is to start from the final destination and then tilt the landscape a little more drastically to see where the population ends up. This is done in denaturation or unfolding experiments: the free energy landscape is changed by increasing the temperature, by adding chemical denaturing agents, or by applying a force to separate the two ends of the molecule. It's important to realize that the route taken will depend on the method used to drag the molecule away from its final state, and that it is by exploring a sufficiently wide variety of roads that the map will be built.

A second possibility is to start from an unfolded ensemble in a non-native environment, and then initiate the folding by changing the environment to a native-like one. This is the equivalent of suddenly transforming the map from flat to 3D and let the pilgrims find their way. The molecular population coming from all over the free energy map is then monitored, and intermediates are distinguished based both on structural properties (region of the conformational landscape) and kinetic properties (one is more likely to identify as intermediates those kinetic bottlenecks on the folding pathway, like are like the valleys at the bottom of mountain passes.

The most challenging aspect is to *characterize* the intermediates that have been distinguished based on kinetic and/or physical (i.e. spectral, hydrodynamic, or distance) properties. For structural ensembles that are neither totally unfolded nor totally stable, it's almost impossible to obtain structural information with atomistic detail. Among the five topologies fully characterized by NMR to date, the one having two G-quartets instead of three attracts attention as a possible intermediate: not each and every guanine has found its partners, and one G-tract is slipped by one step (see Figure 1F). However strand slippage involves breaking and reforming G-quartets. Also, not all slippage mechanisms are sterically allowed, because the guanine syn-anti patterns in the quartets before and after the strand slippage should remain unchanged, but must remain compatible mutually with other bases of the new quartets formed. As a result, strand slippage is kinetically hampered for all structures except the fully parallel, all-anti structure A. Therefore, such stable structures with slipped strands can also be dead ends on the folding route, and are not necessarily possible to reach on an unfolding route.

A faster folding/unfolding route would be the stepwise assembly of quartets strand per strand, a route on which partial dimers containing G-G base pairs and partial trimers containing G-triplets would be plausible intermediates. Unfortunately the structural interpretation of physical properties monitored in folding/unfolding studies is not straightforward. Force-extension distributions in single-molecule unfolding are compatible with an intermediate having one G-tract removed from the core, which would remain a so-called G-triplex [6, 7]. Intermediates having a circular dichroism spectrum

with a maximum at 265 nm and a minimum at 242 nm were identified both upon thermal unfolding [8] and K^+-jump folding [9] were also assigned as G-triplexes, based on the similarity with the CD spectrum of a 3-repeat telomeric sequence [6]. Intriguingly, these CD spectra are identical to those expected for fully parallel G-quadruplexes with homo-stacking between successive guanines [10].

This intriguing G-triplex therefore raises the following questions: What is (are) their structure(s)? (e.g., are they parallel-stranded or antiparallel-stranded?) Are they stable for a sufficient amount of time to be called intermediates? In other words, are G-triplexes local minima on the folding routes of telomeric DNA and if so, where are these points exactly on the conformational map?

To answer these questions, molecular modeling must come into play. A state-of-the-art molecular mechanics (MM) force field is like a faithful local 3D map for the molecule, which will wander around valleys and hills, with more or less energy depending on its temperature. Also, force fields describe the molecular structure with atomistic resolution, providing exquisite details on the structure and how it changes over time. Molecular dynamics (MD) simulations allow one to sample a trajectory of the molecule in its force field. The drawback is that limitations in computer power do limit the time given for the molecule to explore its landscape, and hence the explored area.

There are ways to expand the sampling (the explored zone on the landscape), using for example metadynamics or replica exchange between multiple trajectories at different temperatures. The first atomistic structure of a G-triplex was provided by a 80-ns metadynamics MD simulation of the folding of the thrombin binding aptamer sequence, a short DNA sequence forming an antiparallel structure with two G-quartets: in one of the intermediates the G-tract at the 3'-end protruded out in the solvent [11]. The stability of the rest of the structure, the G-triplex, was experimentally confirmed by NMR, CD, and thermal unfolding studies on the truncated sequence missing the 3'-end. A drawback of enhanced sampling methods is that the sampling is biased: it is like if the landscape was tilted so as to drive exploration further away from the starting point.

To allow the molecule to explore the natural landscape, *unbiased* sampling is required, and as a consequence, long simulations are needed to allow the molecule to explore the landscape in a meaningful way. In a previous work, Stadlbauer et al. investigated strand-slipped structures as late-stage folding intermediates [4]. In a feature paper of the present issue [5], the authors report the longest to date unbiased sampling MD simulations of G-triplexes derived from the human telomeric sequence, with three tracts of three guanines each.

Importantly for unbiased sampling, the choice of the starting structures is crucial, and must be driven by a scientific question. Here nine different starting structures of three-triad telomeric sequences were tried out, corresponding to different loop placements, with the underlying question: which of these triplexes are stable on the microsecond timescale and which are not? The MD simulations revealed which proposed starting structures remain around their starting point, or on the contrary whether they move away and how quick they do so. This must be tested for a long enough time scale and repeated a few times for each starting point to check consistency. The present study reveals that several antiparallel triplexes (for example, $C_{(triplex)}$ on Figure 1) remained "stable" on the microsecond time scale (i.e. the postulated triplex starting structure was indeed close to a local minimum on the landscape, and molecules tended to stay around), whereas fully parallel triplexes (e.g., $A_{(triplex)}$ on Figure 1) were unstable on that time scale.

The lessons learned are many. Molecular dynamics simulations help to validate intermediates on the folding energy landscape: the simulations determine whether they are points of interest by observing whether molecules placed in that state remain in this local minimum for a significant amount of time. Importantly, they also provide refined details at the atomistic level both on their structure and dynamics within these regions of interest. Also, interplay between theory and experiment is the key to untangle folding problems. Without modeling, structural interpretation of physical measurements is only speculation. And in current state of the art and hardware, without thoughtful speculations simulations would wander about the landscape with no sense of direction. Mutual understanding between specialists of both communities is therefore crucial, and gaining understanding in the advantages and limitations of state-of-the-art MD simulations is another very good reason for experimentalists to read the present feature article from Stadlbauer et al. [5].

[1] J.N. Onuchic, N.D. Socci, Z. Luthey-Schulten, P.G. Wolynes, Protein folding funnels: the nature of the transition state ensemble, Fold.Des 1 (1996) 441-450.

[2] D. Thirumalai, C. Hyeon, RNA and protein folding: Common themes and variations, Biochemistry 44 (2005) 4957-4970.

[3] M. Webba da Silva, M. Trajkovski, Y. Sannohe, N. Ma'ani Hessari, H. Sugiyama, J. Plavec, Design of a G-quadruplex topology through glycosidic bond angles, Angew Chem Int Ed Engl 48 (2009) 9167-9170.

[4] P. Stadlbauer, M. Krepl, T.E. Cheatham, 3rd, J. Koca, J. Sponer, Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations, Nucleic Acids Res 41 (2013) 7128-7143.

[5] P. Stadlbauer, L. Trantirek, T.E. Cheatham III, J. Koca, J. Sponer, Triplex intermediates in folding of human telomeric quadruplexes probed by microsecond-scale molecular dynamics simulations, Biochimie XXX (2014) YYY.

[6] D. Koirala, T. Mashimo, Y. Sannohe, Z. Yu, H. Mao, H. Sugiyama, Intramolecular folding in three tandem guanine repeats of human telomeric DNA, Chem. Commun. 48 (2012) 2006-2008.

[7] T. Mashimo, H. Yagi, Y. Sannohe, A. Rajendran, H. Sugiyama, Folding pathways of human telomeric type-1 and type-2 G-quadruplex structures, J. Am. Chem. Soc. 132 (2010) 14910-14918.

[8] R. Buscaglia, R.D. Gray, J.B. Chaires, Thermodynamic characterization of human telomere quadruplex unfolding, Biopolymers 99 (2013) 1006-1018.

[9] R.D. Gray, J.O. Trent, J.B. Chaires, Folding and unfolding pathways of the human telomeric G-quadruplex, J Mol Biol 426 (2014) 1629-1650.

[10] A.I. Karsisiotis, N. Ma'ani Hessari, E. Novellino, G.P. Spada, A. Randazzo, M. Webba da Silva, Topological characterization of nucleic acid G-quadruplexes by UV absorption and circular dichroism, Angew. Chem. Int. Ed. 50 (2011) 10645-10648.

[11] V. Limongelli, S. De Tito, L. Cerofolini, M. Fragai, B. Pagano, R. Trotta, S. Cosconati, L. Marinelli, E. Novellino, I. Bertini, A. Randazzo, C. Luchinat, M. Parrinello, The G-triplex DNA, Angew Chem Int Ed Engl 52 (2013) 2269-2273.

[12] M. Zentgraf, Map of the way of St James In Europe, http://en.wikipedia.org/wiki/Way_of_St._James#mediaviewer/File:Stjacquescompostelle1.png (accessed July 30, 2014)