RNA Folding Algorithms with G-Quadruplexes

Ronny Lorenz[1], Stephan H. Bernhart[2], Fabian Externbrink[2], Jing Qin[3], Christian Höner zu Siederdissen[1], Fabian Amman[1], Ivo L. Hofacker[1,4], and Peter F. Stadler[2,1,3,4,5,6]

[1]Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany; [3]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; [4]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark; [5]Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany; [6]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Abstract. G-quadruplexes are abundant locally stable structural elements in nucleic acids. The combinatorial theory of RNA structures and the dynamic programming algorithms for RNA secondary structure prediction are extended here to incorporate G-quadruplexes. Using a simple but plausible energy model for quadruplexes, we find that the overwhelming majority of putative quadruplex-forming sequences in the human genome are likely to fold into canonical secondary structures instead.

Key words: Dynamic programming, RNA folding algorithms, `ViennaRNA Package`

1 Introduction

Guanosine-rich nucleic acid sequences readily fold into four-stranded structures known as G-quadruplexes. DNA quadruplexes are, for instance, an important component of human telomeres [1], they appear to be strongly overrepresented in the promoter regions of diverse organisms, and they can associate with a variety of small molecule ligands, see [2, 3] for recent reviews. SNPs in G-quadruplexes, finally, have been implicated as a source variation of gene expression levels [4]. RNA quadruplexes have also been implicated in regulatory functions. Conserved G-quadruplex structures within the 5'-UTR of the human TRF2 mRNA [5] and eukaryotic MT3 matrix metalloproteinases, for example, repress translation [6]. Another well-studied example is the interaction of the RGG box domain fragile X mental retardation protein (FMRP) to a G-quartet-forming region in the human semaphorin 3F (S3F) mRNA [7, 8]. A recent review of G-quadruplex-based translation regulation is [9]. A functional RNA G-quadruplex in the 3' UTR was recently described as a translational repressor of the proto-oncogene PIM1 [10]. A mechanistic study of this effect, which seems to be widely used in

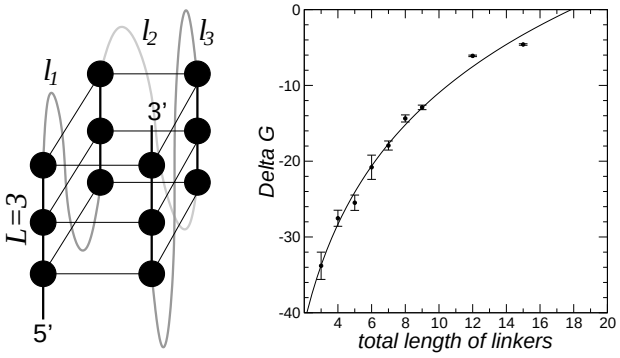


Fig. 1. RNA quadruplexes form parallel arrangements with $L = 2...5$ layers. Folding energies for $L = 3$ depend mostly on the total length ℓ of the linker sequences: the data from ref. [16] fit well to an energy model of the form $\Delta G = a + b \ln \ell$ (solid line).

the cell [11, 12] can be found e.g. in [13]. Most recently, G-quadruplexes were also reported in several long non-coding RNAs [14]. G-quadruplexes are potentially of functional importance in the 100 to 9000 nt G-rich telomeric repeat-containing RNAs (TERRAs). These pol-II transcripts are produced from telomers and seem to be important for the regulation of telomerase activity [15].

Quadruplex structures consist of stacked associations of G-quartets, i.e., planar assemblies of four Hoogsteen-bonded guanines. As in the case of base pairing, the stability of quadruplexes is derived from π-orbital interactions among stacked quartets. The centrally located cations that are coordinated by the quartets also have a major influence on the stability of quadruplex structures, which are composed of at least two and typically not more than 5 stacked quartets.

DNA quadruplexes are structurally heterogeneous: depending on the glycosidic bond angles there are 16 possible structures and further combinatorial complexity is introduced by the relative orientations of the backbone along the four edges of the stack [17]. RNA quadruplexes, in contrast, appear to be structurally monomorphic forming parallel-stranded conformations (Fig. 1, left) independently of surrounding conditions, i.e., different cations and RNA concentration [18]. In this contribution we restrict ourselves to the simpler case of RNA quadruplexes.

Bioinformatically, G-quadruplex structures have been investigated mostly as genomic sequence motifs. The `G4P Calculator` searches for four adjacent runs of at least three Gs. With its help a correlation of putative quadruplex forming sequences and certain functional classes of genes was detected [19]. Similarly, `quadparser` [20] recognizes the pattern (1) below. It was used e.g. in [21] to demonstrate the enrichment of quadruplexes in transcriptional regulatory regions. A substantial conservation of such sequence patterns in mammalian promoter regions is reported in [22]. The web service `QGRS Mapper` uses a similar pattern and implements a heuristic scoring system [23], see also [24] for a review. A Bayesian prediction framework based on Gaussian process regression was recently introduced to predict melting temperatures of quadruplex sequences [25].

The formation of RNA quadruplexes necessarily competes with the formation of canonical secondary structures. Hence they cannot be fully understood in iso-

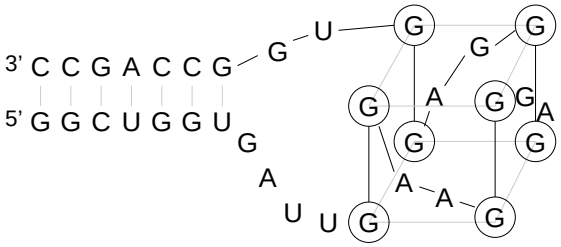


Fig. 2. Structure of the G-quadruplex in a hairpin of human semaphorin 3F RNA that binds the RGG box domain of fragile X mental retardation protein (FMRP). Redrawn based on [7].

lation. In this contribution we therefore investigate how G-quadruplex structures can be incorporated into RNA secondary structure prediction algorithms.

2 Energy Model for RNA Quadruplexes

Thermodynamic parameters for RNA quadruplexes can be derived from measurements of UV absorption as a function of temperature [26], analogous to melting curves of secondary structures. While the stability of DNA G-quadruplexes strongly depends on the arrangement of loops [27, 28] this does not appear to be the case for RNA. RNA not only forms mostly parallel-stranded stacks for G-quartets but their stability also exhibits a rather simple dependence of the loop length [16]. In further contrast to DNA [29], they appear to be less dependent on the nucleotide sequence itself.

A G-quadruplex with $2 \leq L \leq 5$ stacked G-quartets and three linkers of length $l_1, l_2, l_3 \geq 1$ has the form

$$G_L N_{l_1} G_L N_{l_2} G_L N_{l_3} G_L \tag{1}$$

It is commonly assumed that $1 \leq l_i \leq 7$ [25], although *in vitro* data for DNA suggest that longer linkers are possible [30]. For $L = 2$, the existence of quadruplexes with $1 \leq \ell_i \leq 2$ was reported [31]. For $L = 3$ detailed thermodynamic data are available only for the 27 cases $1 \leq l_1, l_2, l_3 \leq 3$ and for some longer symmetric linkers $l_1 = l_2 = l_3$ [16], see Figure 1b. To our knowledge, no comprehensive data are available for $L \geq 4$. It appears reasonable to assume that the stacking energies are additive. The energetic effect of the linkers appears to be well described in terms of the total linker length ℓ [16]. As shown in Figure 1b the free energy depends approximately logarithmically on ℓ. In this contribution we are mostly concerned with the algorithmic issues of including G-quadruplexes into thermodynamic folding programs. In particular we ignore here the strong dependence of quadruplex stability on the potassium concentration, see e.g. [32]. We thus resort to the simplified energy function

$$E[L, \ell] = a(L-1)g_0 + b\ln(\ell - 2) \tag{2}$$

with parameters $a = -18$ kcal/mol and $b = 12$ kcal/mol if the pattern (1) is matched, and $E = \infty$ otherwise.

G-quadruplex structures can be located within loops of more complex secondary structures. Fig. 2, for instance, shows the $L = 2$, $l_1 = l_2 = l_3 = 2$ quadruplex in a hairpin of the semaphorin 3F RNA [7]. It seems natural to treat G-quadruplexes inside multiloops similar to their branching helices: each unpaired base incurs a penalty a and each G-quadruplex within a loop is associated with an additional "loop strain" b. For the interior-loop case of Fig. 2, only stabilizing mismatch contributions of the enclosing pair and a penalty for the stretches of unpaired bases are used. Sterical considerations for this case suggest that a G-quadruplex is flanked by a stretch of at least three unpaired nucleotides or has at least one unpaired nucleotide on either side.

3 Combinatorics of Structures with Quadruplexes

RNA secondary structures consist of mutually non-crossing base pairs and unpaired positions. Thus they can be represented as strings composed of matching parentheses (base pairs) and dots. This "dot-parenthesis" notation is used by the `ViennaRNA Package` [33]. G-quadruplexes constitute an extra type of structural element. The semaphorin hairpin, Fig. 2, can therefore be written as

```
GGCUGGUGAUUGGAAGGGAGGGAGGUGGCCAGCC
(((((((....++..++..++..++..)))))))
```
(3)

using the symbol `+` to mark the bases involved in G-quartets. This string representation uniquely identifies all G-quartets since the first run of `+` symbols determines L for the 5'-most quadruplex, thus determining the next three G-stacks which are separated by at least one '`.`' and must have the same length. It follows immediately that the number of secondary structures with G-quadruplexes is still smaller than 4^n, an observation that is important for the evolvability of RNAs [34]. In order to get a tighter bound on the number of structures we use here, for the sake of presentation, a simplified model in which we omit the restrictions of a minimal size of a hairpin loop and allow quadruplexes with any value of $L \geq 2$ and $l_i \geq 1$. A refinement with a more realistic parametrization can be found in the supplement.

Let $\mathbf{g}_n$ denote the number of secondary structures with G quadruplexes on a sequence of length n. The corresponding generating function is $\mathbf{G}(x) = \sum_{n\geq 0} \mathbf{g}_n x^n$. Similarly, let $\mathbf{q}_n$ be the number of quadruplexes on length n. As derived in the appendix, its generating function is

$$\mathbf{Q}(x) = \sum_{n\geq 0} \mathbf{q}_n x^n = \frac{x^{11}}{(1-x)^3(1-x^4)} \tag{4}$$

The basic idea is now to consider a structure consisting of b base pairs, u unpaired bases and k quadruplexes. Then there are $\binom{2b+k}{k}$ ways to insert k quadruplexes into each of the $C_b = \frac{1}{b+1}\binom{2b}{b}$ possible arrangements of b matching pairs of parentheses, see [35] about the Catalan numbers C_b. Into each of these arrangements

we can insert u unpaired bases in $\binom{2b+k+u}{u}$ different ways. Thus we have

$$\begin{aligned}\mathbf{G}(x) &= \sum_k \sum_b \sum_u \frac{1}{b+1}\binom{2b}{b}\binom{2b+k}{k}\binom{2b+k+u}{u} x^{2b+u}\mathbf{Q}(x)^k \\ &= \frac{2}{1-x-\mathbf{Q}(x)+\sqrt{(1-3x-\mathbf{Q}(x))(1+x-\mathbf{Q}(x))}}\end{aligned} \tag{5}$$

Following [36] we find that the coefficients of $\mathbf{G}(x)$ are asymptotically given by $\mathbf{g}_n \sim k_0 n^{-3/2}\gamma^n$, where k_0 is a positive constant and $\gamma \approx 3.00005$. A more detailed model accounting for minimal stack and loop lengths yields $\gamma \approx 2.2903$ if isolated base pairs are allowed, and $\gamma \approx 1.8643$ for canonical secondary structures. Details are given in the Supplemental Material.

4 RNA Folding Algorithms

Energy Minimization. Dynamic programming algorithms for secondary structure prediction are based on a simple recursive decomposition: any feasible structure on the interval $[i, j]$ has the first base either unpaired or paired with a position k satisfying $i < k \leq j$. The condition that base pairs do not cross implies that the intervals $[i+1, k-1]$ and $[k+1, j]$ form self-contained structures whose energies can be evaluated independent of each other. In conjunction with the standard energy model [37], which distinguishes hairpin loops, interior loops (including stacked base pairs), and multi-loops, this leads to the recursions diagrammatically represented in Fig. 3 (ignoring the cases involving black blocks). This algorithmic approach was pioneered e.g. in [38, 39] and is also used in the `ViennaRNA Package` [33].

G-quadruplexes form closed structural elements on well-defined sequence intervals. Thus they can be treated just like substructures enclosed by a base pair, so that the additional ingredients in the folding algorithms are the energies G_{ij} (free energy of the most stable quadruplex so that the pattern (1) matches exactly the interval $[i, j]$) and the partition functions Z^G_{ij} (defined as the sum of the Boltzmann factors of all distinct quadruplexes on the interval $[i, j]$). As a consequence of (1) we have $G_{ij} < \infty$ and $Z^G_{ij} > 0$ only if $|j - i| < 4L_{\max} + \ell_{\max}$. All possible quadruplexes on the interval $[i, j]$ can be determined and evaluated in $\mathcal{O}(L^2_{\max}\ell^2_{\max})$ time so that these arrays can be precomputed in $\mathcal{O}(n(L_{\max} + \ell_{\max})L^2_{\max}\ell^2_{\max})$, i.e., in linear time.

The standard recursions for RNA secondary structure prediction can now be extended by extra terms for quadruplexes, see Fig. 3. The simplest strategy would be to add G-quadruplexes as an additional type of base-pair enclosed structures. This would amount to using standard interior loop parameters also for cases such as Fig. 2. Hence we use the somewhat more elaborate grammar of Fig. 3, which introduces the quadruplexes in the form of additional cases into the multi-loop decomposition. An advantage of this method is that one can use different parameter values to penalize the inclusion of quadruplexes and helical components into a multiloop. Clearly the grammar is still unambiguous, i.e.,

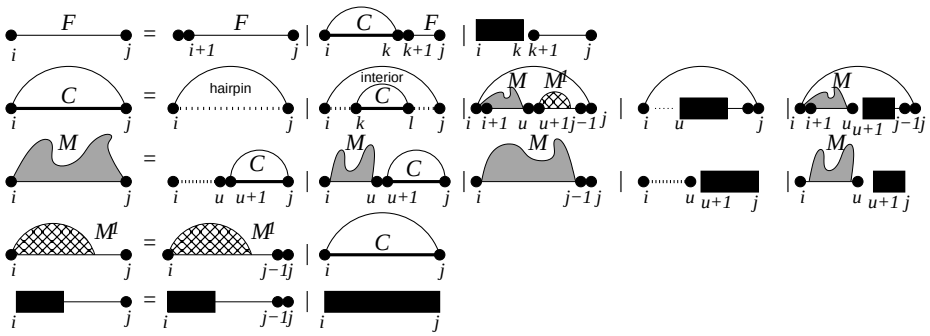


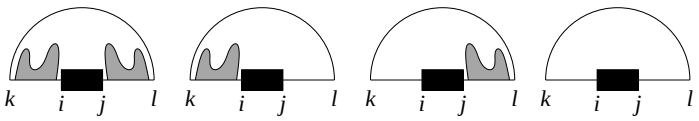
Fig. 3. Extension of recursions of the `ViennaRNA Package` to accomodate G-quadruplexes. This grammar treats G-quadruplexes with multi-loop like energies also in an interior-loop-like context.

every structure has an unique parse. Thus it can be used directly to compute partition functions.

Base Pairing Probabilities. A straightforward generalization of McCaskill's algorithm can be used to compute the probabilities P_{ij} of all possible base pairs (i, j). The probability P_{ij}^G of finding a G-quadruplex delimited by positions i and j then can be written as

$$P_{ij}^G = \frac{Z_{1,i-1} Z_{ij}^G Z_{j+1,n}}{Z} + \sum_{\substack{k<i-1 \\ l>j+1}} P_{kl} \mathbb{P}\left\{\text{quadruplex}[i,j] \middle| (k,l)\right\} \tag{6}$$

The conditional probabilities $\mathbb{P}\{\dots\}$ in turn are composed of the four individual cases depending on the placement of the components of the generalized multiloop enclosed by (k, l) relative to the interval $[i, j]$:



(7)

This decomposition translates to the recursion $\mathbb{P}\left\{\text{quadruplex}[i,j] \middle| (k,l)\right\} =$

$$\frac{Z_{k+1,i-1}^M Z_{ij}^G Z_{j+1,l-1}^M}{Z_{kl}^B} + \frac{Z_{k+1,i-1}^M Z_{ij}^G \hat{b}^{l-j-1}}{Z_{kl}^B} + \frac{\hat{b}^{i-k-1} Z_{ij}^G Z_{j+1,l-1}^M}{Z_{kl}^B} + \frac{\hat{b}^{i-k-1} Z_{ij}^G \hat{b}^{l-j-1}}{Z_{kl}^B}$$

where $\hat{b} = \exp(-b/RT)$. From the P_{ij}^G it is straightforward to compute the probability of a particular quadruplex as

$$p([i, L, l_1, l_2, j]) = \frac{\exp(-E[L, \ell])}{Z_{ij}^G} P_{ij}^G \tag{8}$$

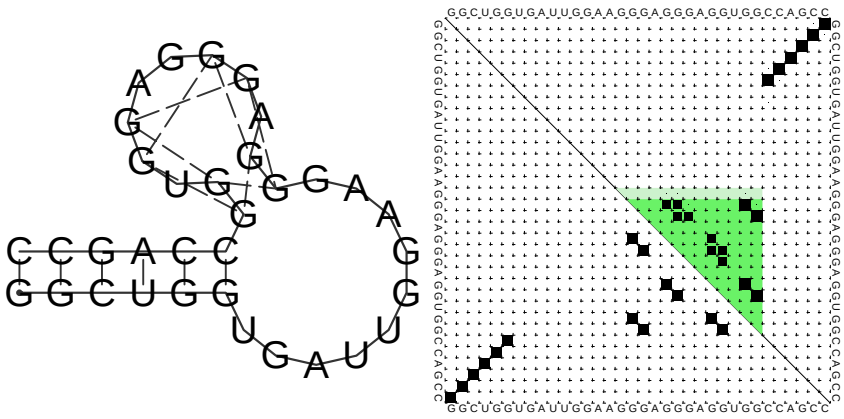


Fig. 4. Representation of minimum free energy structure (l.h.s.) and base pairing probability matrix (r.h.s.) of the semaphorin hairpin (see Fig. 2) respectively.

where $l_3 = j-i+1-4L-l_1-l_2$. Summing up the probabilities of all quadruplexes that contain a particular contact $i' : j'$ of two guanosines in a layer finally yield the probability of the G:G contact $i' : j'$.

Fig. 4 shows an example of the graphical output of `RNAfold`. In the minimum energy case we use a very simple modification of the standard layout [40] treating each quadruplex like a local hairpin structure, explicitly indicating the G-G pairs. Quadruplexes are shown in addition to the individual G-G pairs as shaded triangles in the base pair probability dot plots. From the base pairing probabilities we also compute MEA [41] and centroid structures.

By definition the centroid structure X minimizes the expected base pair distance to the other structures within the Boltzmann-weighted ensemble. In the absence of G-quadruplexes X consists of all base pairs (i, j) with $p_{ij} > 1/2$. A certain ambiguity arises depending on whether X is interpreted as a list of base pairs that may contain incomplete quadruplexes, or whether quadruplexes are treated as units. Here, we insert a quadruplex if $P^G_{ij} > 0.5$, and represent it by the most stable quadruplex with endpoints i and j. The same representation is used for MEA structures where we extend the maximized expected accuracy to $\mathrm{EA} = \sum_{(i,j)\in S} 2\gamma(P_{i,j} + P^G_{ij}) + \sum_i P^u_i$ with $P^u_i = 1 - \sum_j P_{ij} - \sum_{k\le i\le l} P^G_{kl}$, accordingly.

Consensus Structures can be readily obtained for a given multiple sequence alignment. The idea is to apply the dynamic programming recursions to alignment columns. The energy contributions are determined as the average of the corresponding contributions to the individual sequences [43]. In addition small contributions are added to favor pairs of columns with consistent (e.g. GC→GU) and compensatory mutations (AU→GC) since these provide direct evidence for selection acting to preserve base pairing. Similarly, penalties are added if one or a few sequences cannot form a base pair. We refer to [44] for details of the scoring model implemented in `RNAalifold`. Here, we extend it by a simple system

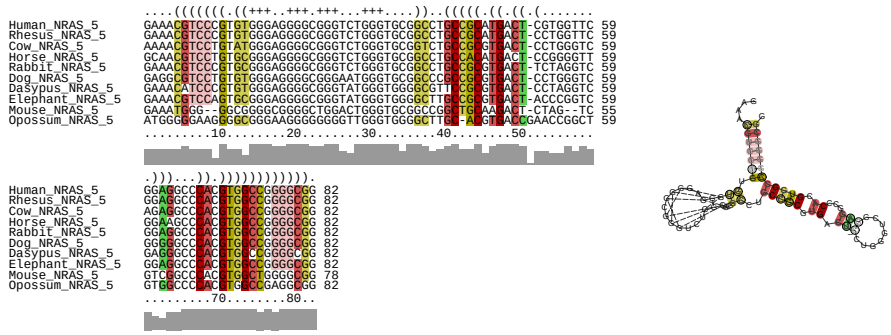


Fig. 5. Consensus structure of the 5'-most part of the 5'UTR of the NRAS mRNA, exhibiting a conserved G-quadruplex with $L = 3$ that modulates translation of the *NRAS* proto-oncogene [42]. Colors indicate the number (red 1, ochre 2, green 3) of different types of basepairs in a pair of alignment columns, unsaturated colors indicate basepairs that cannot be formed by 1 or 2 sequences. Substitutions in stem regions are indicated by circles in the secondary structure drawing.

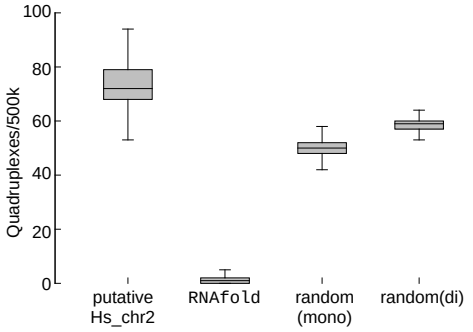
of penalties for mutations that disrupt quadruplexes. Non-G nucleotides incur an energy E' in the outer layers of the quadruplex and $2E'$ in the inner layers as they affect one or two stacking interactions, respectively. An example of a consensus structure prediction is shown in Fig. 5.

Implementation Details. The implementation of G-quadruplex folding in `RNAfold` and `RNAalifold` essentially follows the extended grammar shown in Fig. 3, distinguishing the energy contribution of unpaired bases in the external loop from those enclosed by base pairs. The energies of all possible G-quadruplexes are pre-computed, storing the energy of the most stable quadruplex for each pair of endpoints in the triangular matrix G. As this matrix will be very sparse for most inputs, a sparse matrix optimization is possible, but not yet implemented. In the backtracing part we re-enumerate quadruplexes with given endpoints whenever necessary. Base pairing probabilities are computed as outlined above. Since there cannot be a conflict with canonical base pairs, we store P_{ij}^G as part of the base pairing probability matrix. The probabilities of individual G-G contacts are computed by enumeration as a post-processing step. We also adapted the `RNAeval` and `RNAplot` programs so that sequence/structure pairs can be parsed and re-evaluated according to the extended grammar.

Availability. The source code can be downloaded from `www.tbi.univie.ac.at/~ronny/programs/`.

5 Evaluation

Runtime Performance. The runtime of `RNAfold` with the extended grammar of Fig. 3 was compared to the implementation of the standard model. For both, energy minimization and partition function, virtually no difference was observed.



L	putative	folding	fraction
2	56730	782	0.014
3	14121	225	0.016
4	1361	23	0.017
5	110	2	0.018

Fig. 6. Abundance and stability of putative G-quadruplexes. L.h.s.: Box plot showing the number of potential G-quadruplexes in human chromosome 2 within sliding windows of 500 000 nucleotides. For comparison, the same information for a random sequence with the same mono- or di-nucleotide composition than chr. 2 is presented as well. Both, the mono- and di-nucleotide distribution have been generated from chromosome 2. `RNAfold` denotes the number of putative G-quadruplexes stable enough to occur in a predicted structure of 100 nucleotides up- and downstream of the putative G-quadruplex (with median=1, interquartile range=0–2). R.h.s.: fraction of stable quadruplexes as function of L. for human chromosome 2.

For short sequences of about 200 nt the additional pre-processing steps incur a minor but negligible runtime overhead.

Occurrance and stability of G-quadruplexes in genomes. Sequence motifs of the form (1) that can in principle form quadruplex structures are very abundant in most genomes, see e.g. [19–21]. The number of putative quadruplex-forming sequences is even slightly larger than expected from random sequences with the same mono- or dinucleotide distributions, Fig. 6. The overwhelming majority of these quadruplex candidates, however, is unstable compared to canonical secondary structures that use some or all of Gs in canonical base pairs. We observe that less than 2% of the putative quadruplexes are thermodynamically stable. Interestingly, this effect is nearly independent of the number of layers (L). This data is, however, preliminary in that it reflects the occurrance of putative G-quadruplexes on the human chromosome 2 only. More comprehensive and accurate data that are not restricted to a single fixed window length will be computed with our forthcoming local folding algorithm that extends `RNALfold` [45].

6 Discussion

We have shown in this contribution that structural elements such as G-quadruplexes that correspond to uninterrupted sequence intervals can be included in a rather straightforward way into the standard dynamic programming recursions – provided a corresponding extension of the energy model can be devised. The G-quadruplex-aware programs are currently available as a separate branch 2.0.3g

of the `ViennaRNA Package` using a very simple energy function for the quadruplexes that reproduces the few available experimental data at least semi-quantitatively. Following further optimization of the code the algorithmic extensions will be integrated in the main version of the package in the near future. The extensions in Fig. 3 can also be applied to local folding algorithms such as `RNALfold` and `RNAplfold` or the the exhaustive enumeration of suboptimal structures in `RNAsubopt`. This is ongoing work, as is a comprehensive set of tools for genome-wide scans for putative G-quadruplexes.

It is less obvious how to handle quadruplexes in RNA-RNA interactions since our recursions consider local G-quadruplexes only. At least it is clear that they can be included in all those parts of the structure that are not involved in intermolecular contacts. Some quadruplex structures, however, are formed *in trans*. The binding of G-rich small RNAs to G-rich regions in reporter mRNAs leads to the formation of an intermolecular RNA G-quadruplex that in turn can inhibit translation in living cells [46]. One can use `RNAup` [47] to compute the probabilities $p^{(1)}$ and $p^{(2)}$ that the G-rich regions are unpaired. From these, we obtain the free energies $G^{(i)} = -RT \ln p^{(i)}$ to make the binding site accessible. It remains to compute the interaction energy itself.

The main problem for practical applications of quadruplex-aware RNA folding tools is our limited knowledge of the energy function in particular for $L \neq 3$ and for asymmetric linkers. Even with the crude energy function employed here it becomes clear that the overwhelming majority of putative genomic quadruplex sequences will fold into a canonical secondary structure rather than G-quadruplex structures.

Acknowledgements. This work was supported in part by the German Research Foundation (STA 850/7-2, under the auspicies of SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes"), the Austrian GEN-AU projects "regulatory non coding RNA", "Bioinformatics Integration Network III" and the Austrian FWF project "SFB F43 RNA regulation of the transcriptome".

References

1. Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D., Lipps, H.J.: Telomere end-binding proteins control the formation of G-quadruplex DNA structures *in vivo*. Nature Struct. Mol. Biol. **12** (2005) 847–854
2. Johnson, J.E., Smith, J.S., Kozak, M.L., Johnson, F.B.: *In vivo veritas*: using yeast to probe the biological functions of G-quadruplexes. Biochimie **90** (2008) 1250–1263
3. Wong, H.M., Payet, L., Huppert, J.L.: Function and targeting of G-quadruplexes. Curr Opin Mol Ther. **11** (2009) 146–155
4. Baral, A., Kumar, P., Halder, R., Mani, P., Yadav, V.K., Singh, A., Das, S.K., Chowdhury, S.: Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. Nucleic Acids Res. (2012) doi: 10.1093/nar/gkr1258.
5. Gomez, D., Guédin, A., Mergny, J.L., Salles, B., Riou, J.F., Teulade-Fichou, M.P., Calsou, P.: A G-quadruplex structure within the 5'-UTR of TRF2 mRNA represses translation in human cells. Nucleic Acids Res. **38** (2010) 7187–7198

6. Morris, M.J., Basu, S.: An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. Biochemistry **48** (2009) 5313–5319
7. Menon, L., Mihailescu, M.R.: Interactions of the G quartet forming semaphorin 3F RNA with the RGG box domain of the fragile X protein family. Nucleic Acids Res. **35** (2007) 5379–5392
8. Bensaid, M., Melko, M., Bechara, E.G., Davidovic, L., Berretta, A., Catania, M.V., Gecz, J., Lalli, E., Bardoni, B.: FRAXE-associated mental retardation protein (FMR2) is an RNA-binding protein with high affinity for G-quartet RNA forming structure. Nucleic Acids Res. **37** (2009) 1269–1279
9. Bugaut, A., Balasubramanian, S.: 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res. (2012) doi: 10.1093/nar/gks068.
10. Arora, A., Suess, B.: An RNA G-quadruplex in the 3' UTR of the proto-oncogene PIM1 represses translation. RNA Biology **8** 802–805
11. Huppert, J.L., Bugaut, A., Kumari, S., Balasubramanian, S.: G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Res. **36** (2008) 6260–6268
12. Beaudoin, J.D., Perreault, J.P.: 5'-UTR G-quadruplex structures acting as translational repressors. Nucleic Acids Res. **38** (2010) 7022–7036
13. Wieland, M., Hartig, J.S.: RNA quadruplex-based modulation of gene expression. Chem. Biol. **14** (2007) 757–763
14. Jayaraj, G.G., Pandey, S., Scaria, V., Maiti, S.: Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biolog **9** (2012) 81–86
15. Luke, B., Lingner, J.: TERRA: telomeric repeat-containing RNA. EMBO J. **28** (2009) 2503–2510
16. Zhang, A.Y., Bugaut, A., Balasubramanian, S.: A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. Biochemistry **50** (2011) 7251–7258
17. Webba da Silva, M.: Geometric formalism for DNA quadruplex folding. Chemistry **13** (9738-9745) 2007
18. Zhang, D.H., Zhi, G.Y.: Structure monomorphism of RNA G-quadruplex that is independent of surrounding condition. J. Biotechnol. **150** (2010) 6–10
19. Eddy, J., Maizels, N.: Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res. **34** (2006) 3887–3896
20. Huppert, J.L., Balasubramanian, S.: Prevalence of quadruplexes in the human genome. Nucleic Acids Res. **33** (2005) 2908–2916
21. Zhao, Y., Du, Z., Li, N.: Extensive selection for the enrichment of G4 DNA motifs in transcriptional regulatory regions of warm blooded animals. FEBS Letters **581** (2007) 1951–1956
22. Verma, A., Halder, K., Halder, R., Yadav, V.K., Rawal, P., Thakur, R.K., Mohd, F., Sharma, A., Chowdhury, S.: G-quadruplex DNA motifs as conserved cis-regulatory elements. J. Med. Chem. **51** (2008) 5641–5649
23. Kikin, O., D'Antonio, L., Bagga, P.S.: QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. **34** (2006) W676–W682
24. Todd, A.K.: Bioinformatics approaches to quadruplex sequence location. Methods **43** (2007) 246–251
25. Stegle, O., Payet, L., Mergny, J.L., MacKay, D.J.C., Huppert, J.L.: Predicting and understanding the stability of G-quadruplexes. Bioinformatics **25** (2009) i374–i382
26. Mergny, J.L., Lacroix, L.: Uv melting of g-quadruplexes. Curr Protoc Nucleic Acid Chem. **Unit 17.1.** (2009)

27. Bugaut, A., Balasubramanian, S.: A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. Biochemistry **47** (2008) 689–697
28. Zhang, D.H., Fujimoto, T., Saxena, S., Yu, H.Q., Miyoshi, D., Sugimoto, N.: Monomorphic RNA G-quadruplex and polymorphic DNA G-quadruplex structures responding to cellular environmental factors. Biochemistry **49** (2010) 4554–4563
29. Guédin, A., De Cian, A., Gros, J., Lacroix, L., Mergny, J.L.: Sequence effects in single-base loops for quadruplexes. Biochimie **90** (2008) 686–696
30. Guédin, A., Gros, J., Patrizia, A., Mergny, J.L.: How long is too long? Effects of loop size on G-quadruplex stability. Nucleic Acids Res. **38** (2010) 7858–7868
31. Lauhon, C.T., Szostak, J.W.: RNA aptamers that bind flavin and nicotinamide redox cofactors. J Am Chem Soc **117** (1995) 1246–1257
32. Joachimi, A., Benz, A., Hartig, J.S.: A comparison of DNA and RNA quadruplex structures and stabilities. Bioorg Med Chem **17** (2009) 6811–6815
33. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L.: ViennaRNA Package 2.0. Alg. Mol. Biol. **6** (2011) 26
34. Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L.: From sequences to shapes and back: A case study in RNA secondary structures. Proc. Roy. Soc. Lond. B **255** (1994) 279–284
35. Koshy, T.: Catalan Numbers with Applications. Oxford Univ. Press, Oxford UK (2008)
36. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, New York (2009)
37. Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., Turner, D.H.: Incorporating chemical modification constraints into a dynamic programming algorithm for prediction ofRNA secondary structure. Proc. Natl. Acad. Sci. USA **101** (2004) 7287–7292
38. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9** (1981) 133–148
39. McCaskill, J.S.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29** (1990) 1105–1119
40. Bruccoleri, R.E., Heinrich, G.: An improved algorithm for nucleic acid secondary structure display. Computer Appl. Biosci. **4** (1988) 167–173
41. Do, C.B., Woods, D.A., Batzoglou, S.: CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics **22**(14) (2006) e90–e98
42. Kumari, S., Bugaut, A., Huppert, J.L., Balasubramanian, S.: An RNA G-quadruplex in the 5'UTR of the NRAS proto-oncogene modulates translation. Nat. Chem. Biol. **3** (2007) 218–221
43. Hofacker, I.L., Fekete, M., Stadler, P.F.: Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. **319** (2002) 1059–1066
44. Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F.: `RNAalifold`: improved consensus structure prediction for RNA alignments. BMC Bioinformatics **9** (2008) 474
45. Hofacker, I.L., Priwitzer, B., Stadler, P.F.: Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics **20** (2004) 191–198
46. Ito, K., Go, S., Komiyama, M., Xu, Y.: Inhibition of translation by small RNA-stabilized mRNA structures in human cells. J. Am. Chem. Soc. **133** (2011) 19153–19159
47. Mückstein, U., Tafer, H., Hackermüller, J., Bernhard, S.B., Stadler, P.F., Hofacker, I.L.: Thermodynamics of RNA-RNA binding. Bioinformatics **22** (2006) 1177–1182

Supplemental Material
A Combinatorics of RNA Structures with Quadruplexes

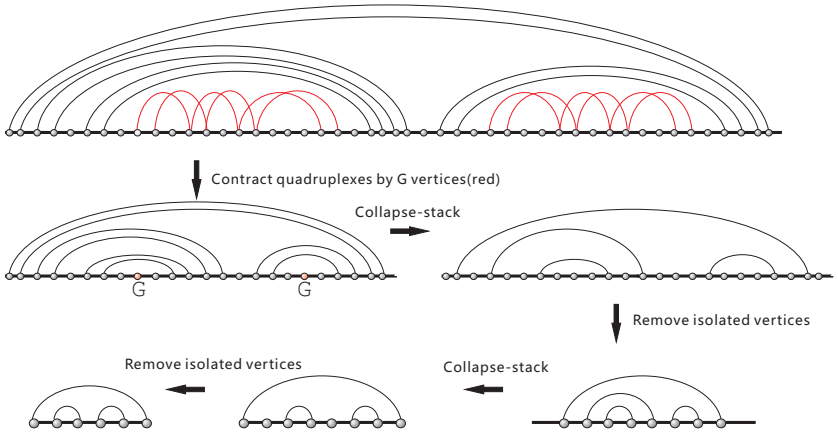
Here we describe a more detailed combinatorial model of secondary structures with G-quadruplexes than the simplified version outlined in the main text.

A1 Model

A secondary structure of length n is a noncrossing partial matching (matching with isolated vertices) such that each base pair is of length at least 3. For simplicity, we consider an arbitrary number of $L \geq 2$ stacked G-quartets and three linkers of length l_1, l_2 and $l_3 \geq 1$. We allow quadruplexes in any context with the following exception: If (i, j) is a base pair that encloses a single quadruplexes, then at least one of three conditions is satisfied: (1) $i + 1$ and $j - 1$ are both unpaired; (2) $i + 1$, $i + 2$, $i + 3$ are unpairedm or (3) $j - 3$, $j - 2$, $j - 1$ are unpaired. We call such structures G-structures in the following.

A stack of length τ consists of exactly τ "parallel" arcs $((i, j), (i + 1, j - 1), \ldots, (i + (\tau - 1), j - (\tau - 1)))$. We say that a G-structure is τ-canonical if all stacks consist of at least τ arcs.

The enumeration is based on the notion of shapes, that is, matchings in which each stack consists of exactly one arc. The shape of an arbitrary G-structure s is obtained by (1) contracting each G-quadruplex to a single vertex labelled 'G', and (2) iteratively collapsing each stack to a single arc and then removing any isolated vertices from the resulting diagram as in the following example:



A2 Generating functions

Let $\mathbf{s}_{n,t}$ denote the number of all noncrossing shapes over $2n$ vertices with t arcs of length 1 (1-arcs) and its corresponding generating function $\mathbf{S}(u, e) = \sum_n \sum_t \mathbf{s}_{n,t} u^n e^t$. Denote by $\mathbf{m}_t$ the number of noncrossing matchings with t

arcs. Note that $\mathbf{m}_{2t}$ is the well-known t-th Catalan number. Using the generating function $\mathbf{M}(u) = \frac{1-\sqrt{1-4u}}{2u}$ for the matchings we have

$$\mathbf{S}(u,e) = \frac{1+u}{1+2u-ue}\mathbf{M}\left(\frac{u(1+u)}{(1+2u-ue)^2}\right).$$

In the following we will make use of several auxiliary functions:

$$\mathbf{Q}(x) = \frac{x^{11}}{(1-x^4)(1-x)^3}, \qquad \mathbf{P}_0(x) = \frac{(1-x)\,\mathbf{Q}(x)}{1-x-x\mathbf{Q}(x)}, \qquad \mathbf{R}(x) = \frac{x^2}{(1-x)^2},$$

$$\mathbf{T}(x) = \frac{1}{(1-x)^2}, \qquad \mathbf{L}(x) = \frac{2x^3}{1-x}, \qquad \mathbf{P}_1(x) = \frac{x}{1-x} + \mathbf{T}(x)\mathbf{P}_0(x),$$

$$\mathbf{P}_2(x) = \frac{x^3}{1-x} + (\mathbf{R}(x)+\mathbf{L}(x))\mathbf{P}_0(x), \qquad \mathbf{P}_3(x) = \frac{1}{1-x} + \mathbf{T}(x)\mathbf{P}_0(x).$$

Our main result is

Theorem 1. *The generating function of τ-canonical G-structures is*

$$\boldsymbol{G}^{\tau}(x) = \sum_n \boldsymbol{g}_n^{\tau} x^n = \boldsymbol{P}_3(x)\boldsymbol{S}\left(\frac{x^{2\tau}\cdot \boldsymbol{P}_3^2(x)}{(1-x^2)-x^{2\tau}(2\boldsymbol{P}_1(x)+\boldsymbol{P}_1^2(x))}, \frac{\boldsymbol{P}_2(x)}{\boldsymbol{P}_3(x)}\right).$$

Proof. We utilize the following combinatorial classes: $\mathcal{E}$ (neutral class, consisting of a single element of size 0), $\mathcal{Z}$ (vertices, with size 1), $\mathcal{U}$ (arcs, comprising two vertices thus having size 2), and $\mathcal{W}$ (quadruple arcs taking 4 vertices).
Claim 1. The generating function of the numbers $\mathbf{q}_n$ of quadruplexes on length n is

$$\mathbf{Q}(x) = \frac{x^{11}}{(1-x^4)(1-x)^3}.$$

Let $\mathcal{Q}$ denote the combinatorial class of G-quadruplexes. By construction, each quadruplex consists of $L \geq 2$ stacked G-quartets and three linkers of length at least 1. Thus we have $\mathcal{Q} = \mathcal{W}^2 \times \mathbf{SEQ}(\mathcal{W}) \times (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^3$. This implies the Claim 1.

Denote by $\mathbf{p}_n$ the number of G-structures of length n without base pairs outside quadruplexes with two additional restrictions: (1) its first and last vertices are part of a quadruplex and (2) if there exist two consecutive G-quartets, then there exists at least one isolated vertex between them.
Claim 2. The generating function of $\mathbf{p}_n$ is $\mathbf{P}_0(x) = \frac{(1-x)\,\mathbf{Q}(x)}{1-x-x\mathbf{Q}(x)}$.

We proceed by induction on the number of G-quadruplexes. Let $\mathbf{p}_n^k$ denote the number of single stranded secondary structures with k G-quadruplexes of length n, then we have its corresponding generating function $\mathbf{P}_0^k(x) = \mathbf{Q}(x)^k \cdot (\frac{x}{1-x})^{k-1}$. The claim follows by summing all $k \geq 1$.
Claim 3. Let λ be a fixed noncrossing shape with $s \geq 1$ arcs and $m \geq 0$ 1-arcs (arcs of length 1). Then the generating function of τ-canonical G-structures containing arc length at least 3 that have shape λ is given by

$$\mathbf{Q}_{\tau}^{\lambda}(x) = \mathbf{P}_3(x) \cdot \left(\frac{x^{2\tau}\cdot \mathbf{P}_3^2(x)}{(1-x^2)-x^{2\tau}(2\mathbf{P}_1(x)+\mathbf{P}_1^2(x))}\right)^s \left(\frac{\mathbf{P}_2(x)}{\mathbf{P}_3(x)}\right)^m$$

In order to prove the claim, we use the additional notations $\mathcal{Z}_{r/b/g}$ for red/blue/green vertices. We can construct an arbitrary τ-canonical G-structure with arc-length at least 3 and shape λ in the following way: Starting from the shape λ, insert at most one red isolated vertex into the $(2s+1)$ intervals except the interval $[i, i+1]$ for which that $(i, i+1)$ is an 1-arc in λ. The corresponding combinatorial class is $\mathcal{M}_1 = \mathcal{U}^s \times (\mathcal{E} + \mathcal{Z}_r)^{2s-m+1}$. Next insert exactly one green isolated vertex after each vertex j such that $(j, j+1)$ forms an 1-arc in λ. This yields the class $\mathcal{M}_2 = \mathcal{M}_1 \times \mathcal{Z}_g^m$.

Next, we inflate each arc into a stack of size $t \geq 0$. In case of $t \geq 1$, between the arcs of the obtained stack we insert a blue isolated vertex to the left or the right, or on both sides in order to separate the arcs and for each such insertion exactly one blue isolated vertex is used. This results in the combinatorial class $\mathcal{M}_3$ from $\mathcal{M}_2$ by the substitution

$$\mathcal{U} \to \sum_{t \geq 1} \mathcal{U}^t \times (2\mathcal{Z}_b + \mathcal{Z}_b^2)^{t-1}.$$

Now we inflate each arc in the resulting structure into a stack of size at least τ. The combinatorial class $\mathcal{M}_4$ results from $\mathcal{M}_3$ via the substitution $\mathcal{U} \to \frac{\mathcal{U}^\tau}{1-\mathcal{U}}$.

Next we inflate each red isolated vertex into either a sequence of isolated vertices of length at least one or a $\mathcal{P}_0$-structure ϑ_1 in addition with two sequences of isolated vertices (at least 1) at both ends of ϑ_2 or a sequence of isolated vertices (at least 3) at one of the ends of ϑ_2. The corresponding class $\mathcal{M}_6$ is symbolically obtained from $\mathcal{M}_5$ by the substitution $\mathcal{Z}_g \to \mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) + (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0 + 2\mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) \times \mathcal{P}_0$.

We then inflate each green isolated vertex into either a sequence of isolated vertices of length at least three or a $\mathcal{P}_0$-structure ϑ_2 in addition with two sequences of isolated vertices (at least 1) at both ends of ϑ_2. The corresponding class $\mathcal{M}_6$ is symbolically obtained from $\mathcal{M}_5$ by the substitution $\mathcal{Z}_g \to \mathcal{Z}^3 \times \mathbf{SEQ}(\mathcal{Z}) + (\mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0$.

We finally inflate each blue isolated vertex into either a sequence of isolated vertices of length at least one or a $\mathcal{P}_0$-structure ϑ_3 in addition with two sequences of isolated vertices at both ends of ϑ_3. The corresponding combinatorial class $\mathcal{M}_7$ is symbolically obtained from $\mathcal{M}_6$ by the substitution $\mathcal{Z}_b \to \mathcal{Z} \times \mathbf{SEQ}(\mathcal{Z}) + (\mathbf{SEQ}(\mathcal{Z}))^2 \times \mathcal{P}_0$.

Combine the steps together, the claim follows. The procedure is illustrated in Fig. 7.

In particular, $\mathbf{Q}_\tau^\lambda(x)$ depends only upon the number of arcs and 1-arcs in λ. Then by definition of the generating function $\mathbf{S}(u, e)$, we obtain $\mathbf{G}^\tau(x)$ by summing over all the possible shapes and the theorem follows. □

A3 Asymptotics

Let us briefly recall some facts concerning the singularity analysis of functional composition [36]. Suppose $f(x)$ and $g(x)$, with $g(0) = 0$, have non-negative coeffients and are analytic at the origin. We consider the composition

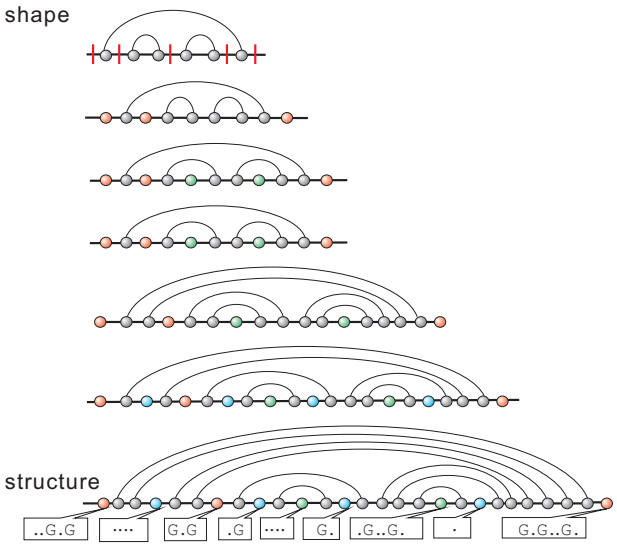


Fig. 7. From a shape to a secondary structure with G-quadruplexes. Each G-quadruplex is shown as a vertex labelled by G.

$h = f(g(x))$. Let ρ_f, ρ_g, and ρ_h be the corresponding radii of the convergence, and let $\tau_g = g(\rho_g)$. The asymptotic behavior of h then depends on the comparison of τ_g and ρ_f:

1. $\tau_g > \rho_f$ (supercritical case) the singularity type is that of the external function f;
2. $\tau_g < \rho_f$ (subcritical case) the singularity of $f(g)$ is driven by that of the inside function g;
3. $\tau_g = \rho_f$ (critical case) the singularity type is a mix of the types of the internal function and the external function and needs special attention.

Theorem 2. *Let $\boldsymbol{g}_n^{\tau}$ denote the number of τ-canonical G-structures on length n. Then we have for $\tau = 1, 2$*

$$\boldsymbol{g}_n^{\tau} \sim k_\tau \, n^{-3/2} \, (\rho_\tau^{-1})^n.$$

Here, $\rho_1^{-1} \approx 2.2903$, $\rho_2^{-1} \approx 1.8643$, and k_1, k_2 are positive constants.

Proof. Combining the expressions for $\mathbf{S}(u, e)$ and $G^{\tau}(x)$ we arrive at

$$\mathbf{G}^{\tau}(x) = \frac{A_1(x)}{A_2(x)} \cdot \mathbf{M}\left(\frac{B_1(x)}{B_2(x)}\right),$$

where $A_1(x)$, $A_2(x)$, $A_3(x)$, and $A_4(x)$ are fixed polynomials. Clearly $\mathbf{Q}^{\tau}(x)$ is algebraic. Furthermore, since the composition scheme is supercritical [36] for the cases $\tau = 1$ and $\tau = 2$, the singularity type is that of the external function, i.e., $\mathbf{M}(x)$. In particular, we have $\rho_1^{-1} \approx 2.2903$ and $\rho_2^{-1} \approx 1.8643$. □

For the corresponding structures without any G-quadruplex, we obtain the results immediately by setting $\mathbf{Q}(x) = 0$ in the above derivation. Thus, we obtain $\hat{\rho}_1^{-1} \approx 2.2887$ and $\hat{\rho}_2^{-1} \approx 1.8489$. Numerical values were obtained with `Maple`, version 11.