2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction

Ronny Lorenz[1], Stephan H. Bernhart[23], Jing Qin[34], Christian Höner zu Siederdissen[1], Andrea Tanzer[15], Fabian Amman[2], Ivo L. Hofacker[167], Peter F. Stadler[123478]

[1]Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria [2]Bioinformatics Group of the Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany [3]Interdisciplinary Center for Bioinformatics of the University of Leipzig [4]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany [5]Center for Genomic Regulation (CRG), Dr. Aiguader, 88, 08003 Barcelona, Spain [6]Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria [7]Center for RNA in Technology and Health, Univ. Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark [8]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, D-04103 Leipzig, Germany [9]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA

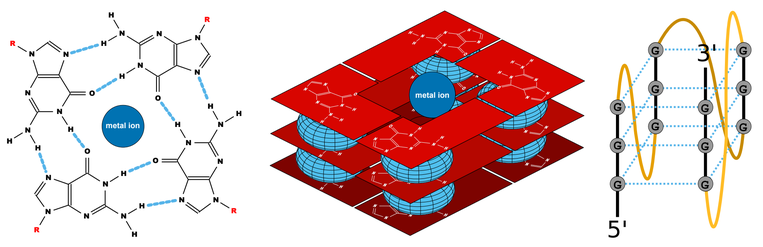
Contact: ronny@tbi.univie.ac.at - http://www.tbi.univie.ac.at

1. Introduction

Nucleic acid sequences are known to fold into four-stranded structures called G-quadruplexes if they are guanosine-rich. In DNA, these locally stable structures are present, for instance, as an important component of human telomers. They appear strongly over-represented in promotor regions of diverse organisms and can interact with several small molecule ligands.

RNA G-quadruplexes also have been found to exhibit regulatory functions. Within the 5'- and 3'-UTR of several protein coding genes they control (*usually repress*) translation. In coding regions they act as a specific recognition site for proteins, e.g. the RGG box domain in fragile X mental retardation protein (*FMRP*) in human semaphorin 3F mRNA. Non-coding RNAs also exhibit functionally important G-quadruplexes, e.g. G-rich telomeric repeat-containing RNAs (*TERRAs*) and several long non-coding RNAs.

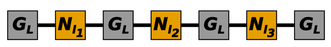
Chemically, a G-quadruplex consists of stacked planar assemblies of four Hoogsteen-bonded guanosines, so called **G-quartets**. It draws its free energy from π-orbital interactions between the stacked quartets which is further stabilized by a centrally located cation. While DNA G-



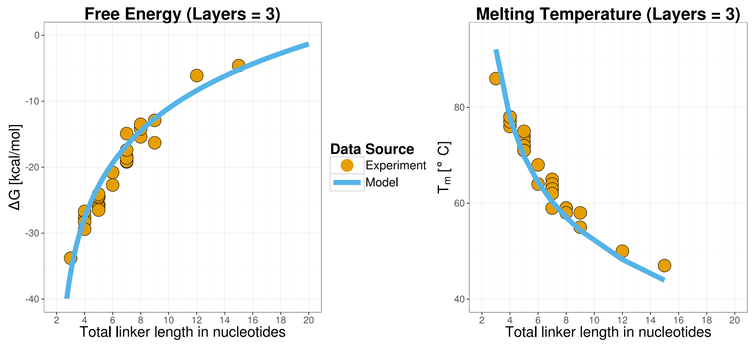
quadruplexes seem to appear in parallel stranded conformations only. Here, we restrict ourselves to the case of RNA quadruplexes to derive a plausible model to incorporate G-quadruplexes into thermodynamic secondary structure prediction algorithms.

2. Energy Model

Based on measurements of UV absorption as function of temperature, thermodynamic parameters can be inferred for G-quadruplexes. While the stability of DNA quadruplexes strongly depends on the arrangement of the interspacing loop regions (linkers) and their sequence content, RNA quadruplexes do not show this behavior. They rather exhibit a simple dependence on the total length of their loops. Thus, the number of stacked G-quartets $2 \le L \le 5$ and the three linker lenghts $l_1,l_2,l_3 \ge 1$ suffice to describe the G-quadruplex:



For $L = 3$ thermodynamic data from Zhang et al. [2] shows that the free energy depends approximately logarithmically on the total linker length $l = l_1 + l_2 + l_3$.



Assuming that the stacking energies are additive and ignoring the strong dependence of stability on the potassium concentration we apply the following energy function

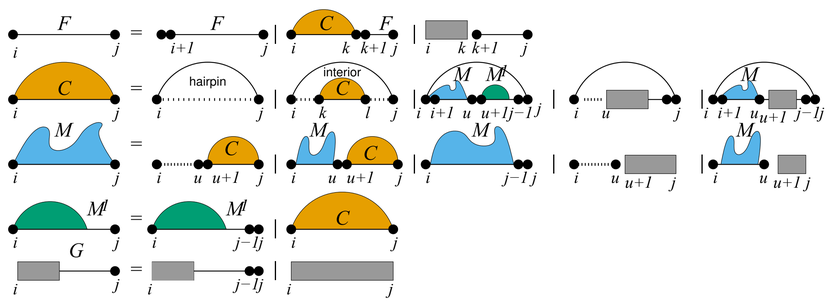
$$E[L,l,T] = a(T)(L-1) + b(T)\ln(l-2)$$

with

$$a(T) = H_a + TS_a$$
$$b(T) = H_b + TS_b$$

3. RNA Folding Algorithms

Thermodynamic RNA secondary structure prediction is usually based on a dynamic programming (DP) algorithm. In detail, the algorithm performs a simple recursive decomposition to assess all configurations on a sequence interval *[i,j]*. For each interval position *i* can be either unpaired or paired with a position $i < k \le j$. Furthermore, the standard energy model distinguishes between several loop types enclosed by the pair *(i,k)*: *hairpin loops*, *interior loops* and *multibranch loops*.

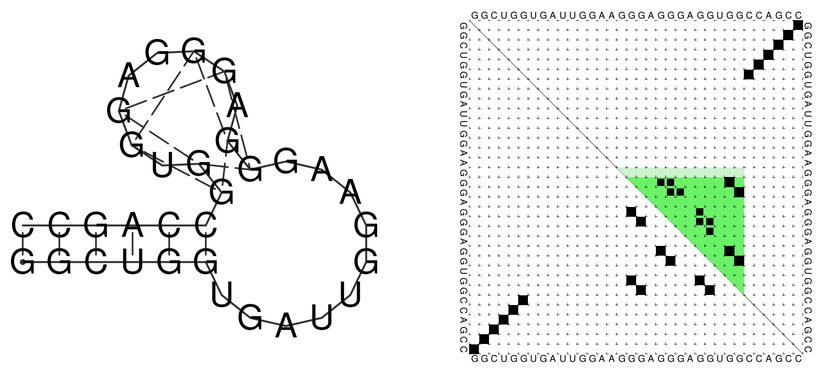


Since G-quadruplexes form a closed structure on a short well-defined sequence interval they can be treated as substructures enclosed by a base pair. Thus, it suffices to add an auxilary energy table G_{ij} to store (a) the free energy of the most stable quadruplex in the interval *[i,j]* for MFE prediction or (b) the partition function Z^G_{ij} to capture the complete Boltzmann ensemble of distinct G-quadruplexes. This modification does not influence the asymptotic time complexity of the algorithm. It is even possible to precompute G_{ij} in $\mathcal{O}(n(L_{max}+l_{max})L^2_{max}l^2_{max})$, i.e. **linear time**. Instead of embedding a quadruplex as an additional type of base pair enclosed structure we rather treat it as an additional component within a multibranch or exterior loop. This allows for penalizing the inclusion of a quadruplex and a helical component differently.

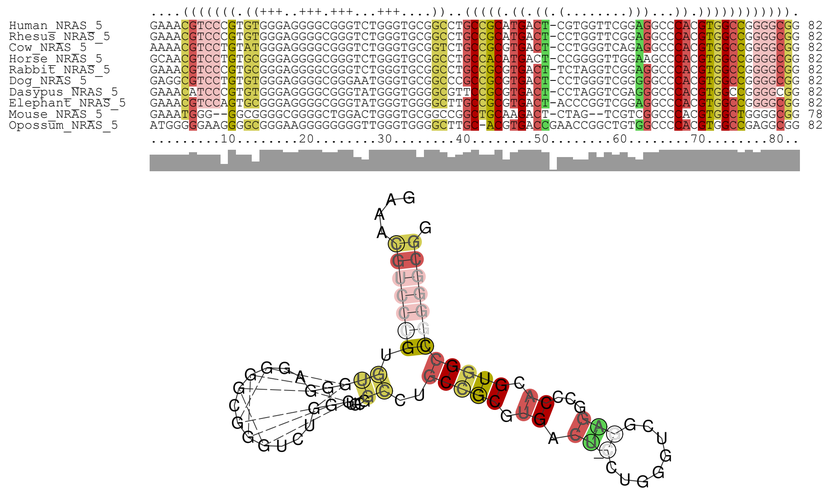
The modified recursion scheme was implemented in *C* as part of the ***ViennaRNA Package*** [3]. As a result, this now empowers the program `RNAfold` to compute MFE, partition function, base pair- and quadruplex- probabilities, centroid structure, and maximum expected accuracy (*MEA*) structure for the extended structure space including G-quadruplexes.

```
$ RNAfold -p

Input string (upper or lower case); @ to quit
....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,....8
GGCUGGUGAUUGGAAGGGAGGGAGGUGGCCAGCC
length = 34
GGCUGGUGAUUGGAAGGGAGGGAGGUGGCCAGCC
((((((..........++.++..++.++))))))
 minimum free energy = -21.39 kcal/mol
((((((..........(..........))))))) 
 free energy of ensemble = -28.59 kcal/mol
((((((..........++.++..++.++)))))) {-22.29 d=0.09}
 frequency of mfe structure in ensemble 8.38749e-06; ensemble diversity 0.17
```

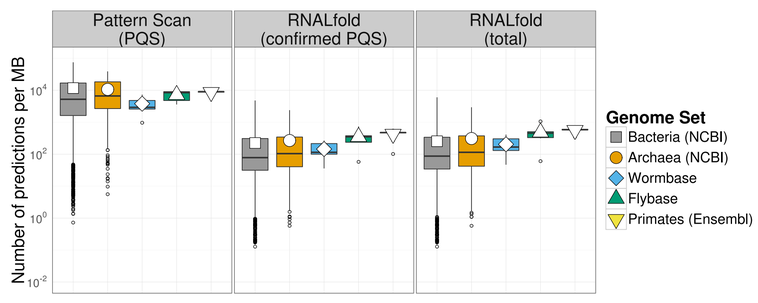


Similarly, the programs `RNAalifold`, `RNALfold` and `RNAcofold` were adapted to constitute a consensus structure prediction algorithm, a scanning variant for locally stable secondary structures, and a RNA dimer variant to predict secondary structures of two RNAs upon hybridization, respectively.

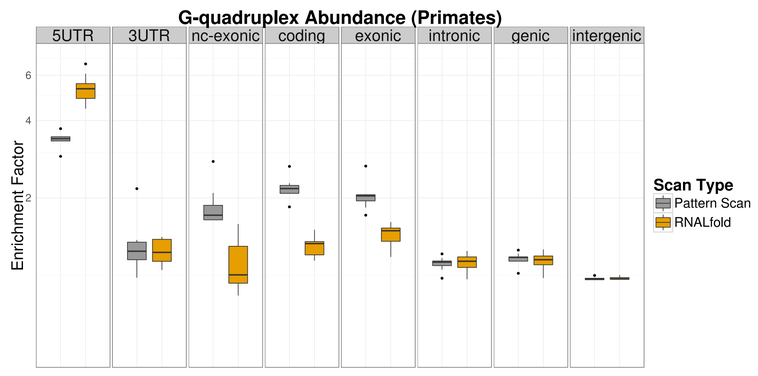


4. Genome Screening

G-quadruplex forming sequence motifs, so called putative quadruplex sites (*PQS*), are very abundant in most genomes. To test whether a PQS exhibits a thermodynamically stable G-quadruplex we conducted a large genome wide survey with `RNALfold`. Each PQS where `RNALfold` predicted at least one stable G-quadruplex was regarded *'confirmed'*. The average confirmation rates of *1.96%* (*bacteria*), *2.38%* (*archaea*), *3.92%* (*nematodes*), *4.79%* (*drosophilidae*) and *4.91%* (*primates*) revealed that the majority of PQS are **dominated by canonical secondary structures** rather then G-quadruplexes.



We then analyzed the enrichment of PQS and thermodynamically stable G-quadruplexes within subsets of the genome. Therefore, the genome was partitioned into *'genic'* and *'intergenic'* regions, according to whether there is an overlap with annotated genes or not. The *'genic'* part of the primate genomes was further split into *'exonic'* and *'intronic'* regions to distinguish between *'coding'*, *'5UTR'* and *'3UTR'* within the *'exonic'* part. The remaining *'exonic'* regions are labelled *'nc-exonic'*. To address the ambiguity of overlapping transcripts, we applied a hierarchy to the projection of the annotation to the genome: *'coding'*, *'5UTR'*, *'3UTR'*, *'nc-exonic'*, *'intronic'*.



We find that G-quadruplexes are **enriched** in genic regions, most notably **in the 5'-UTR** of primates (*median: 5.12fold*). This suggests that G-quadruplexes are abundant functional features of mRNA. In contrast to that, intergenic regions seem to be even slightly depleted. The analysis also showed that relying solely on PQS data tends to overestimate the enrichment. For coding regions, this can be interpreted as a selective pressure against stable G-quadruplexes which are likely to interfer with the process of translation.

5. Discussion

Self enclosed structural elements like RNA G-quadruplexes can be included into the standard DP secondary structure prediction algorithms. The `ViennaRNA Package` now provides a set of programs that implement these elements for energy minimization and partition function computation, local folding algorithms and consensus structures for sequence alignments. Even a simple approach to predict intramolecular G-quadruplexes in RNA-RNA dimers is provided. A large scale enrichment analysis shows that G-quadruplexes are abundant in exonic regions, particularly in 5'-UTRs, whereas their stability is reduced in protein coding sequences. This strongly suggests a function in translational control. However, limited knowledge of the energy function especially for cases not covered by the experimental data remains an uncertainty. Still, it becomes clear that competing canonical secondary structures dominate an overwhelming majority of putative genomic G-quadruplex forming sequences.

References

[1] **"2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction"**, Lorenz, R. et al., *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP:99 (2013)

[2] **"A sequence independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology"**, A. Y. Zhang et al., *Biochemistry*, vol. 50, pp. 7251–7258 (2011)

[3] **"ViennaRNA Package 2.0"**, R. Lorenz et al., *Algorithms in Molecular Biology*, vol. 6, p. 26 (2011)

Acknowledgements

This work was supported in part by the German Research Foundation (*STA 850/7-2, under the auspices 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"*.