RNA Folding Algorithms with G-Quadruplexes

Ronny Lorenz[1], Stephan H. Bernhart[2], Fabian Externbrink[2], Jing Qin[4], Christian Höner zu Siederdissen[1], Fabian Amman[1], Andrea Tanzer[7], Ivo L. Hofacker[1,3], and Peter F. Stadler[2,1,4,3,5,6]

[1] Department of Theoretical Chemistry University of Vienna, Austria

[2] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Germany

[3] Center for non-coding RNA in Technology and Health, University of Copenhagen, Denmark

[4] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[5] Fraunhofer Institute for Cell Therapy and Immunology, Leipzig, Germany

[6] Santa Fe Institute, Santa Fe, USA

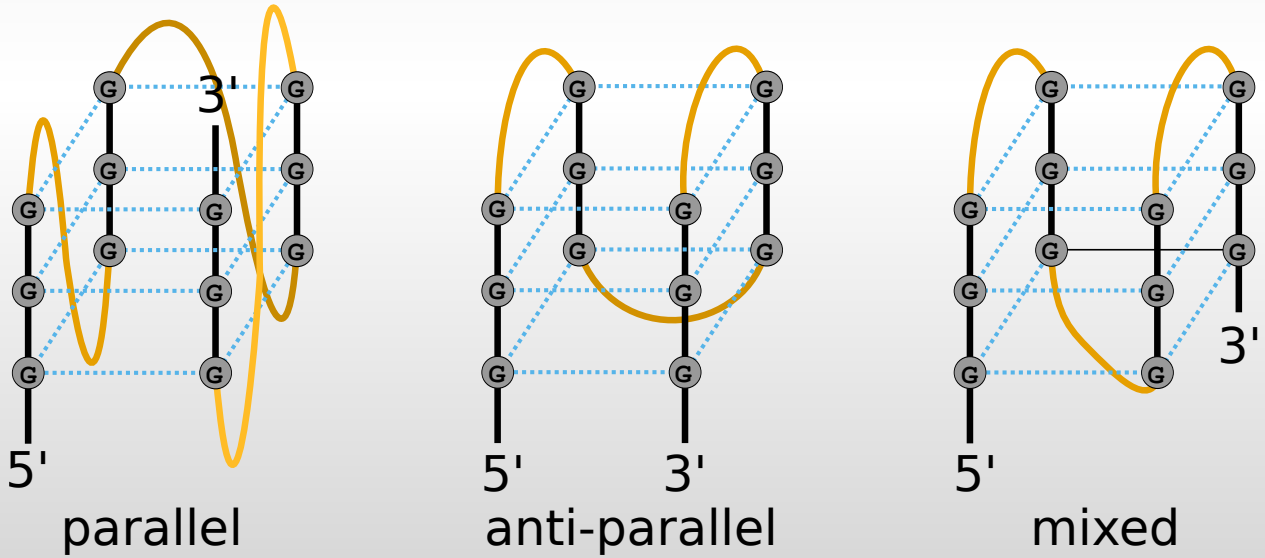
[7] Center for Genomic Regulation (CRG), Barcelona, Spain

Doubice, Czech Republic, October 5, 2012

universität wien

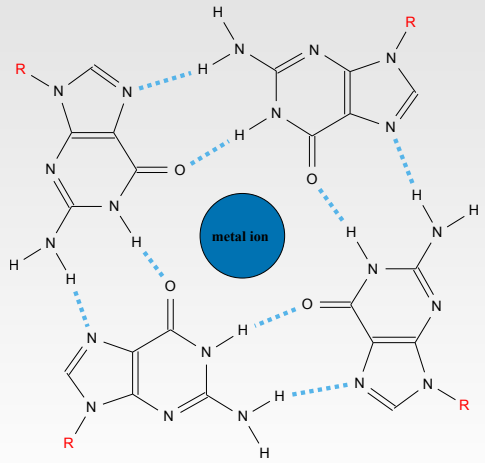
What are G-Quadruplexes

- G-rich nucleic acid sequences can form stacked arrangements of G-quartets
- Stable local structure of 4 interconnected strands
- 2-5 quartet layers connected by 3 short loops
- Sequence pattern follows $G_L N_{l_1} G_L N_{l_2} G_L N_{l_3} G_L$
- Several structure arrangements possible

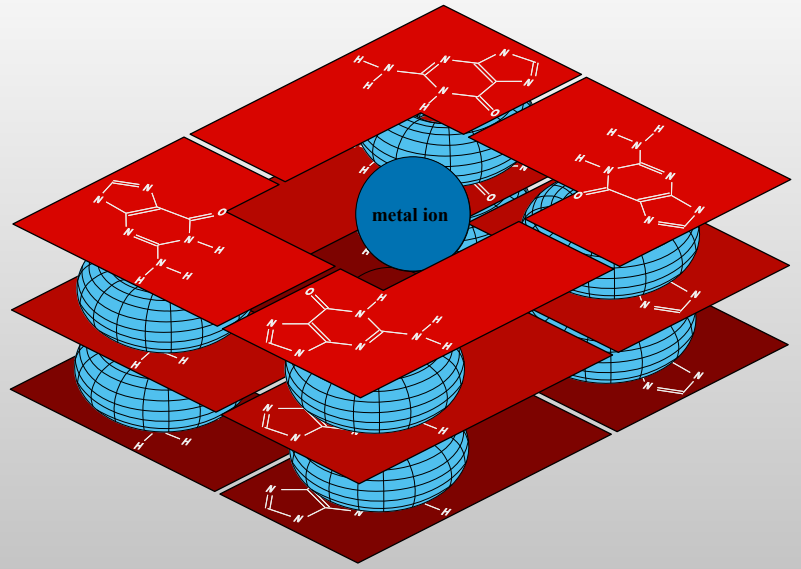


Why are G-Quadruplexes

8 Hogsteen-Watson Crick hydrogen bonds



$\pi-\mathrm{orbital}$ stacking between layers



Where are G-Quadruplexes

DNA:

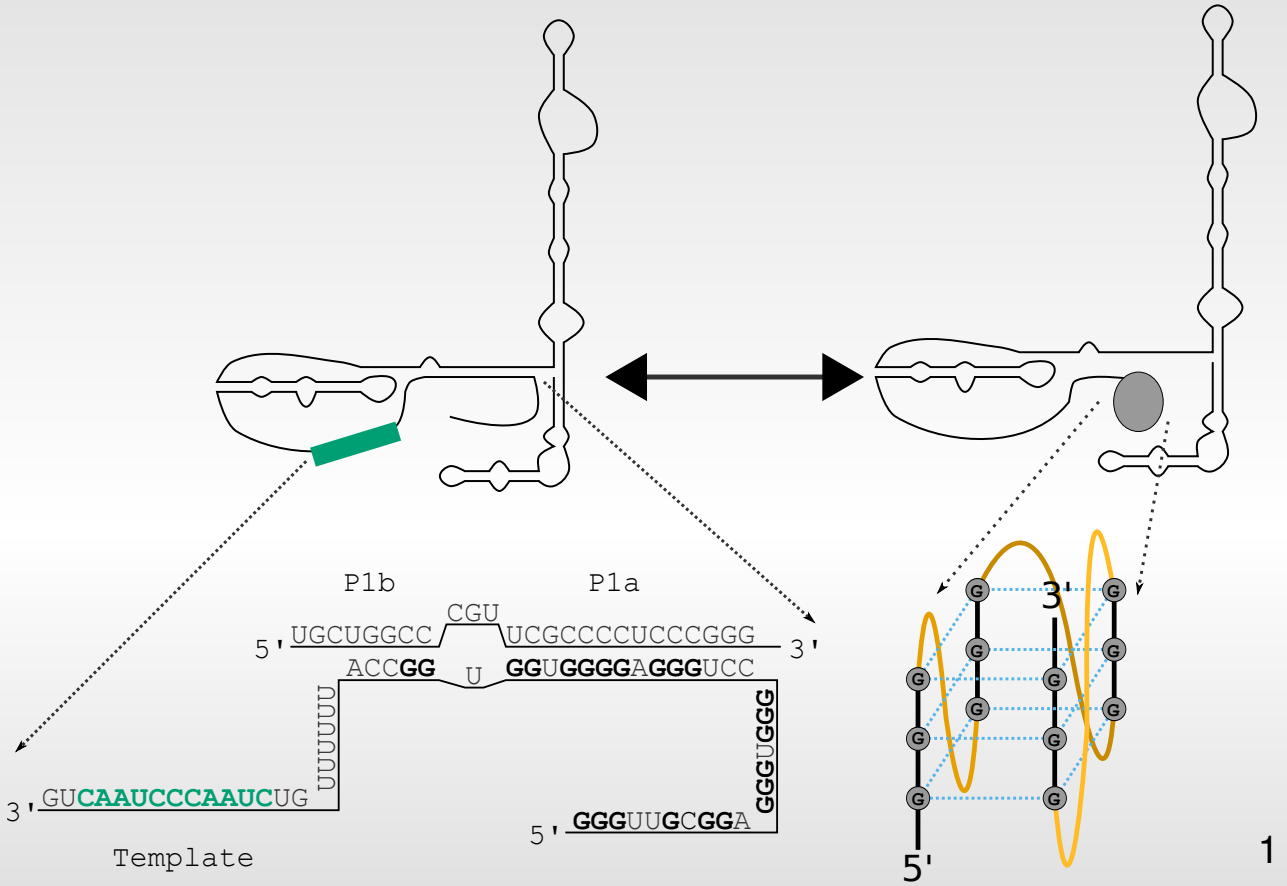
- Human Telomers: Telomerase inhibition
- Promotor Regions: Modulation of gene transcription
- Elsewhere: Interference with protein function

RNA:

- Eukaryote genomes: Translation modulation
  - 5′ and 3′ UTR of mRNAs: post-transcriptional control of gene expression
  - exonic regions of mRNAs: ligand for several G-quadruplex recognizing proteins
  - ncRNAs: function modulation (e.g. hTERC)
  - Elsewhere: Heterodimers in telomeric regions (TERRA)
- Viral RNA genomes: Dimerization (e.g. in HIV)
- Bacterial genomes: Control of slippage transcription

Where are RNA G-Quadruplexes

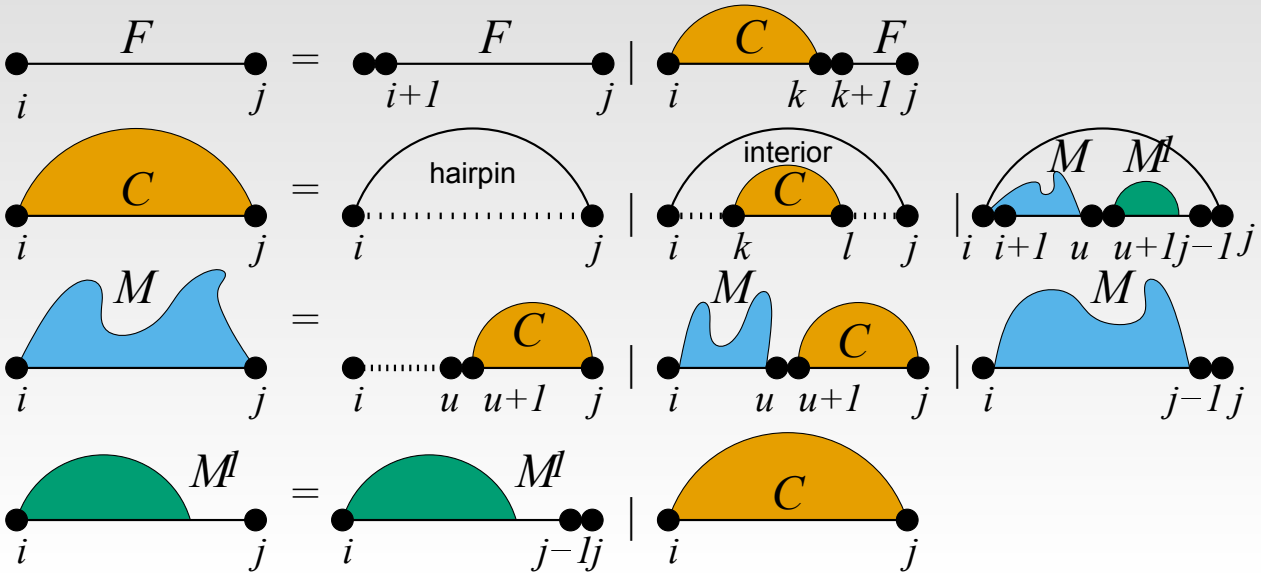
Human Telomerase RNA Component (hTERC):



[1] Adapted from Gros et al., ChemBioChem 2008

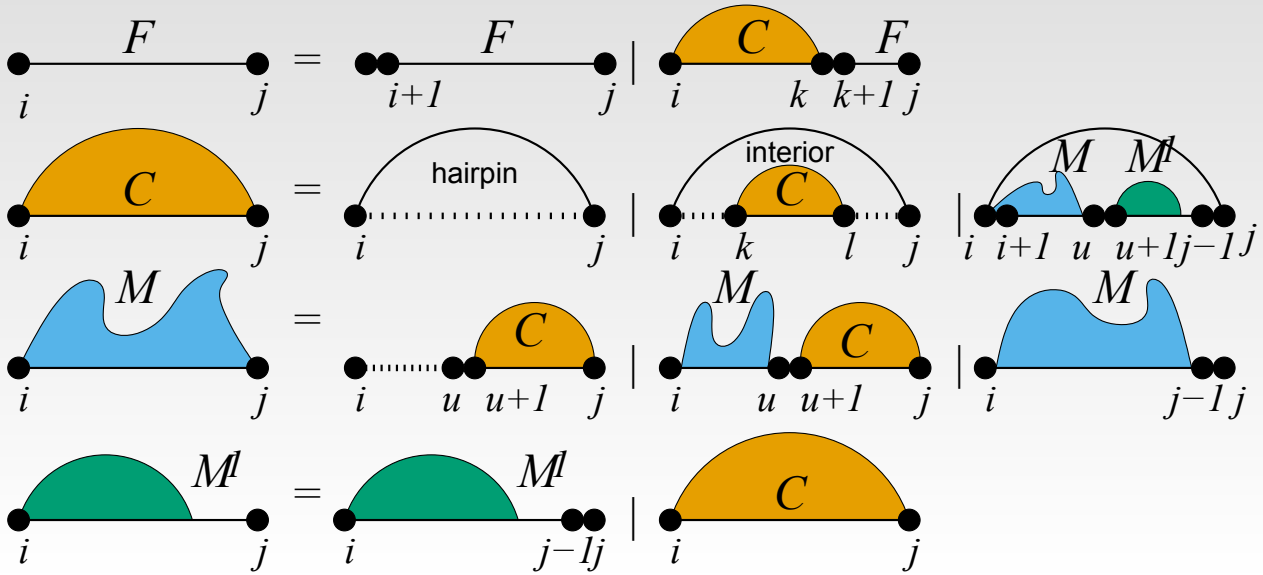
How to predict putativ stable G-quadruplexes from sequence data in silico?

RNA secondary structure prediction



Efficient DP algorithm with asymptotic time complexity of $\mathcal{O}(n^3)$

RNA secondary structure prediction



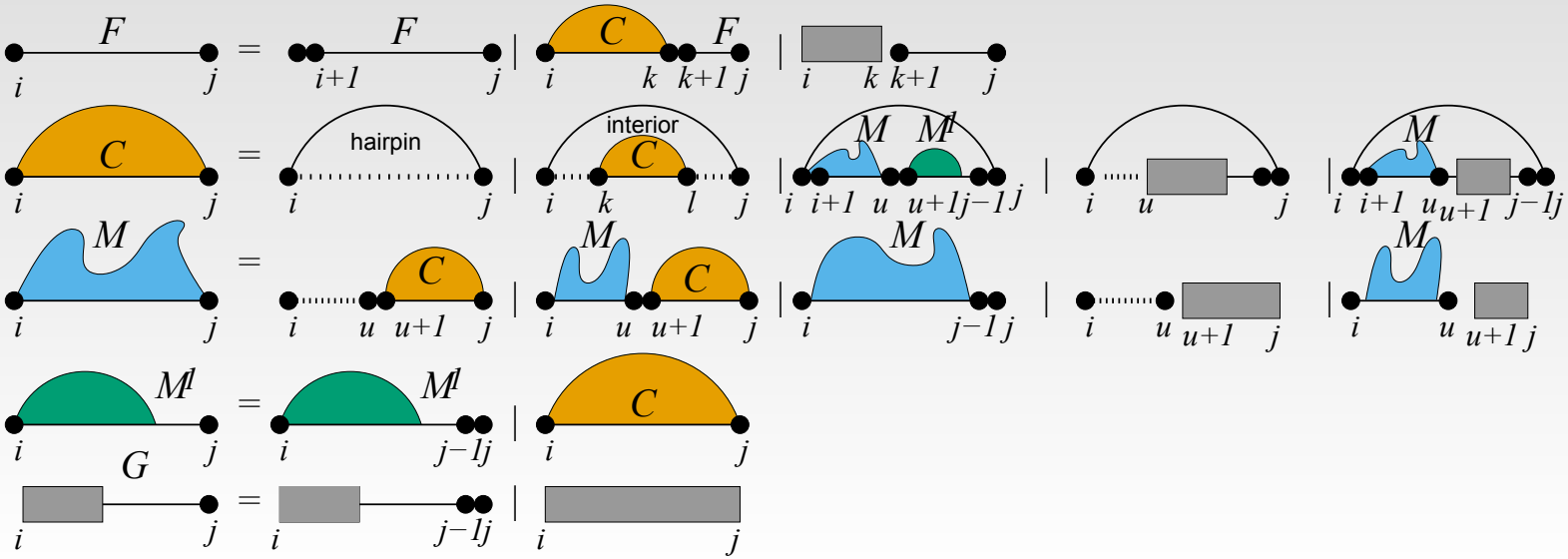
Efficient DP algorithm with asymptotic time complexity of $\mathcal{O}(n^3)$

Well parameterized tools available[2]

	Sensitivity	Specificity	MCC	F-measure
RNAfold 2.0	0.739	0.792	0.763	0.761
RNAfold 1.8.5	0.711	0.773	0.740	0.737
UNAFold	0.692	0.766	0.727	0.724
RNAStructure	0.715	0.781	0.745	0.742

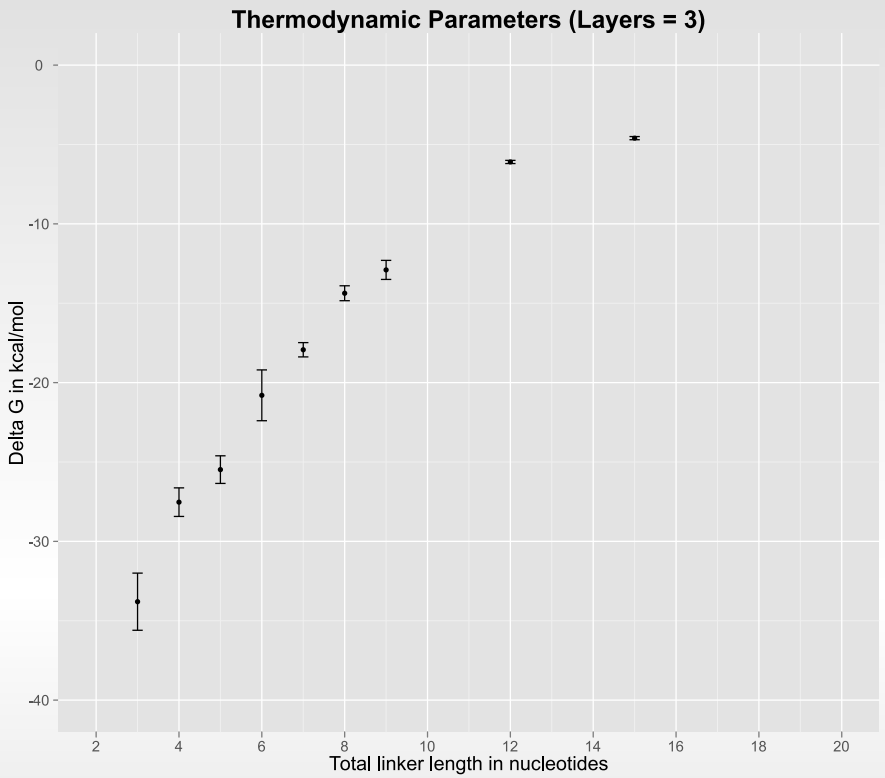
[2] `ViennaRNA Package 2.0`, Lorenz et al. 2011

RNA secondary structure prediction with G-Quadruplexes



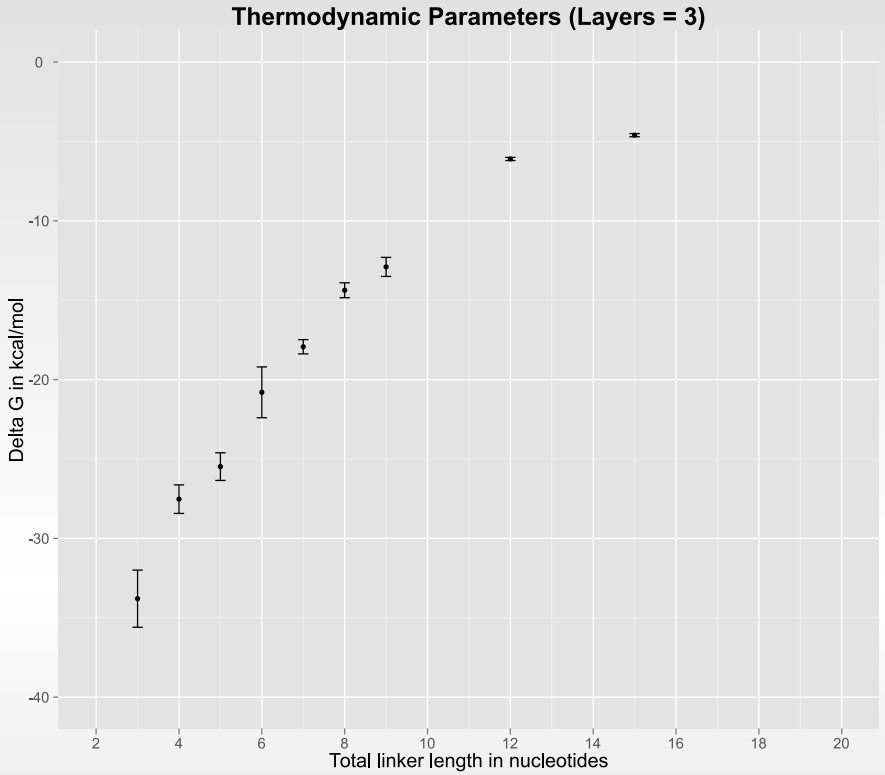
- G-quads are local closed structures
- can be treated like other substructures
- *potential* G-quads can be searched for in linear time
- energy contributions computed via pre-processing step

RNA secondary structure prediction with G-Quadruplexes



Data taken from Zhang et al., Biochemistry 2011

RNA secondary structure prediction with G-Quadruplexes

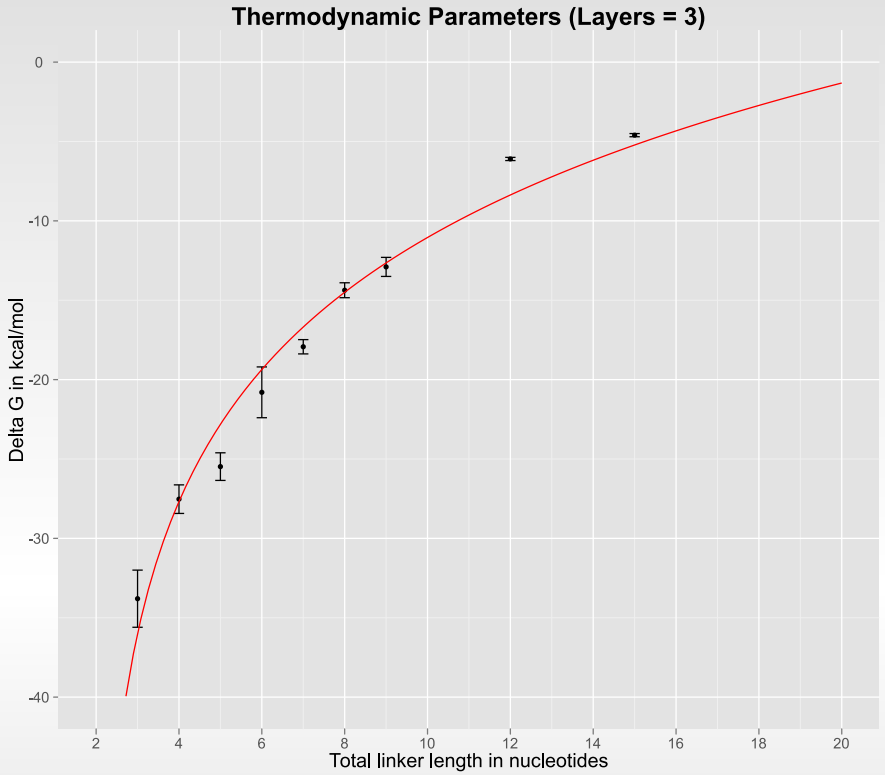


Data taken from Zhang et al., Biochemistry 2011

- Energy $\propto$ number of layers - 1
- Energy $\propto$ total linker length
- No effect of linker asymmetry or sequence composition

$$E(L, l) = a(L-1)g_0 + b\ln(l-2)$$

RNA secondary structure prediction with G-Quadruplexes



Data taken from Zhang et al., Biochemistry 2011

- Energy $\propto$ number of layers - 1
- Energy $\propto$ total linker length
- No effect of linker asymmetry or sequence composition

$$E(L, l) = a(L-1)g_0 + b\ln(l-2) \quad a = -18.00, b = 12.00$$

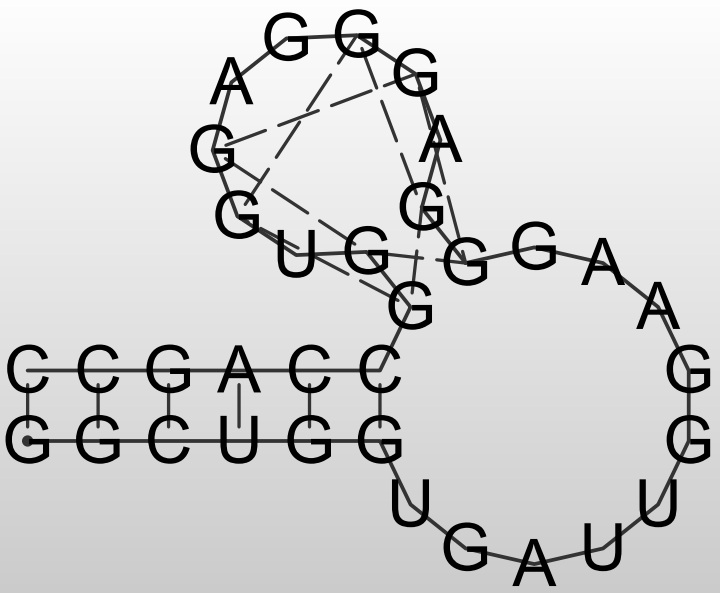
Integration into the `ViennaRNA Package`

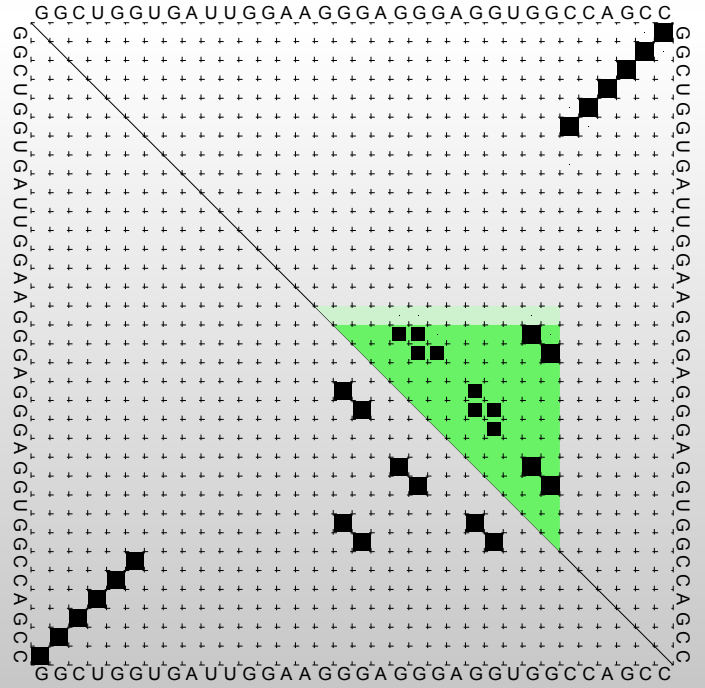
`RNAfold`	MFE-, Centroid- and MEA-Structure, Base Pair Probabilities, Partition Function for Single Sequences
`RNAalifold`	MFE-, Centroid- and MEA-Structure, Base Pair Probabilities, Partition Function for Sequence Alignment
`RNAcofold`	MFE-Structure, Concentration Dependent Base Pair Probabilities, Partition Function for Dimers
`RNALfold`	Locally Stable Structure Prediction
`RNAplfold`	Locally Stable Structure Base Pair Probabilities, Probability for beeing unpaired *(in progress)*
`RNAsubopt`	Suboptimal Structure Prediction for Single Sequences and Sequence Dimers *(in progress)*

RNA secondary structure prediction with G-Quadruplexes

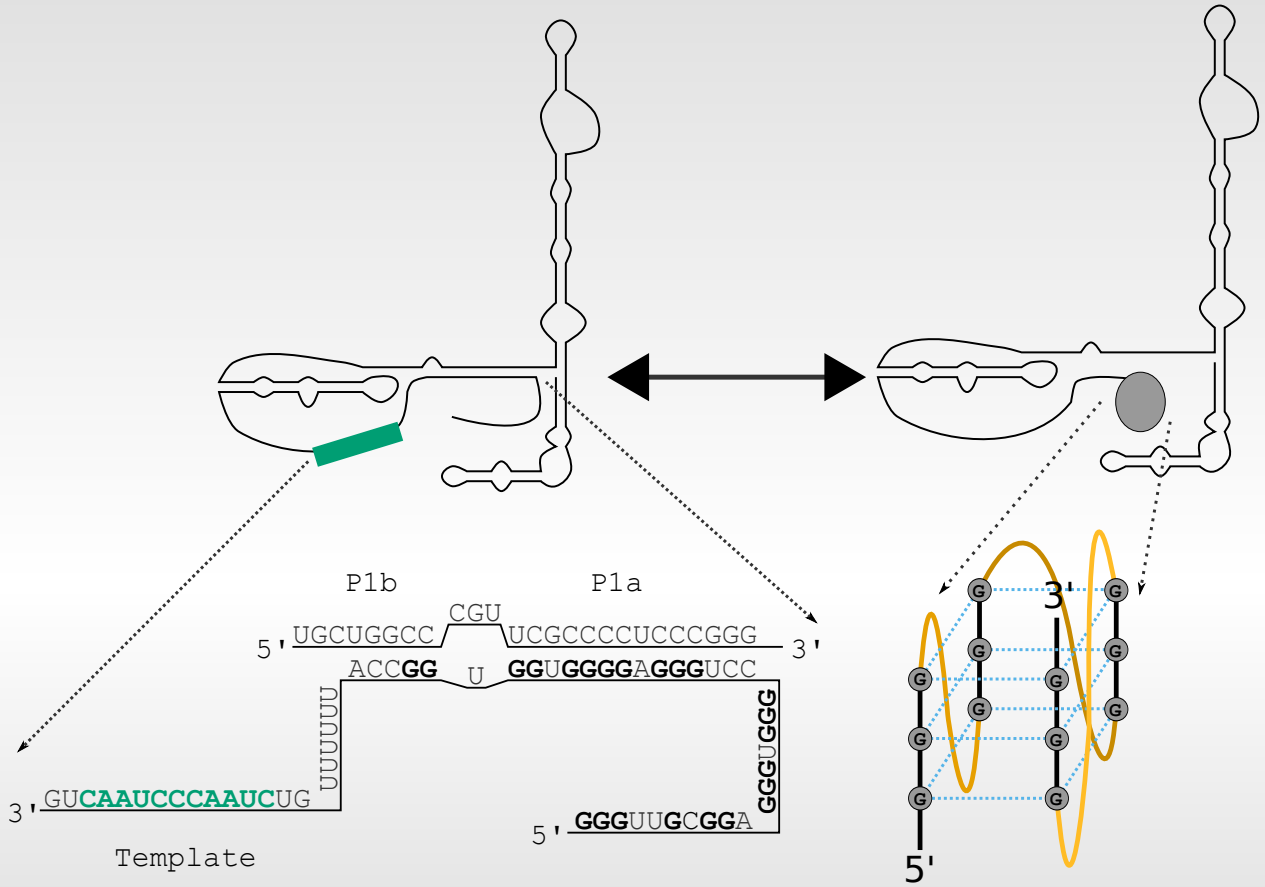
```
$ RNAfold -p

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
```



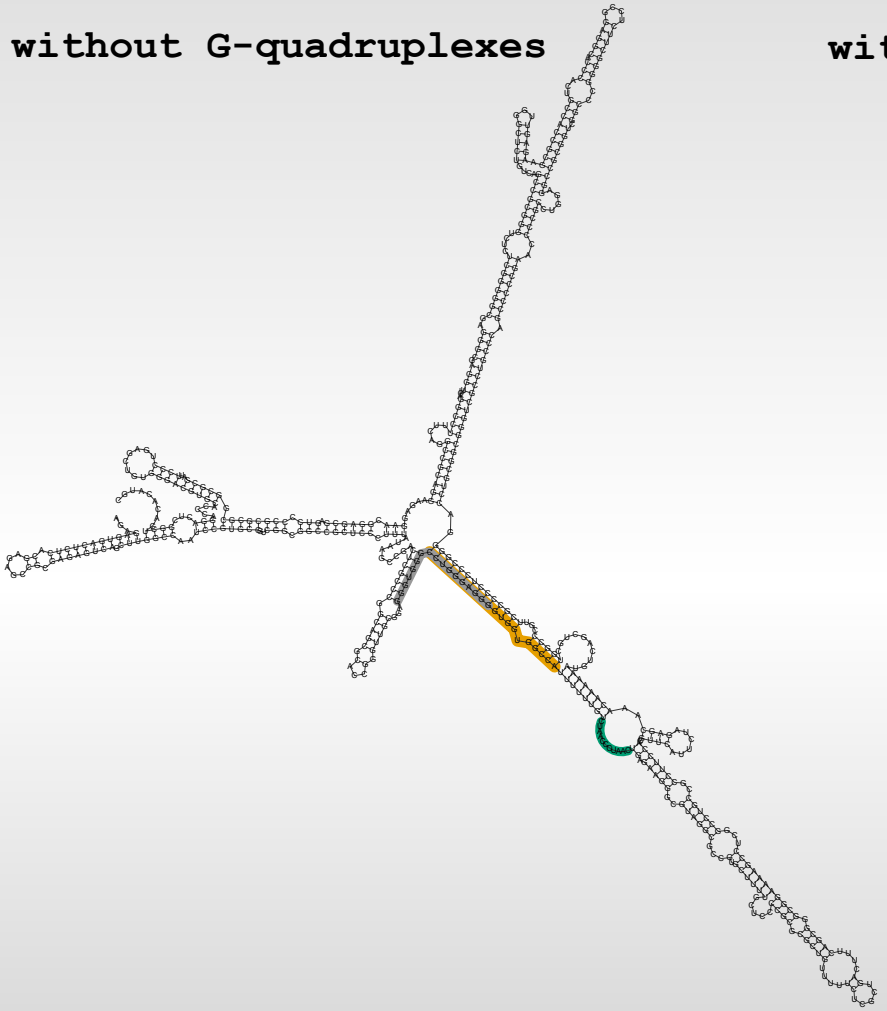


human Telomerase RNA Component (hTERC)

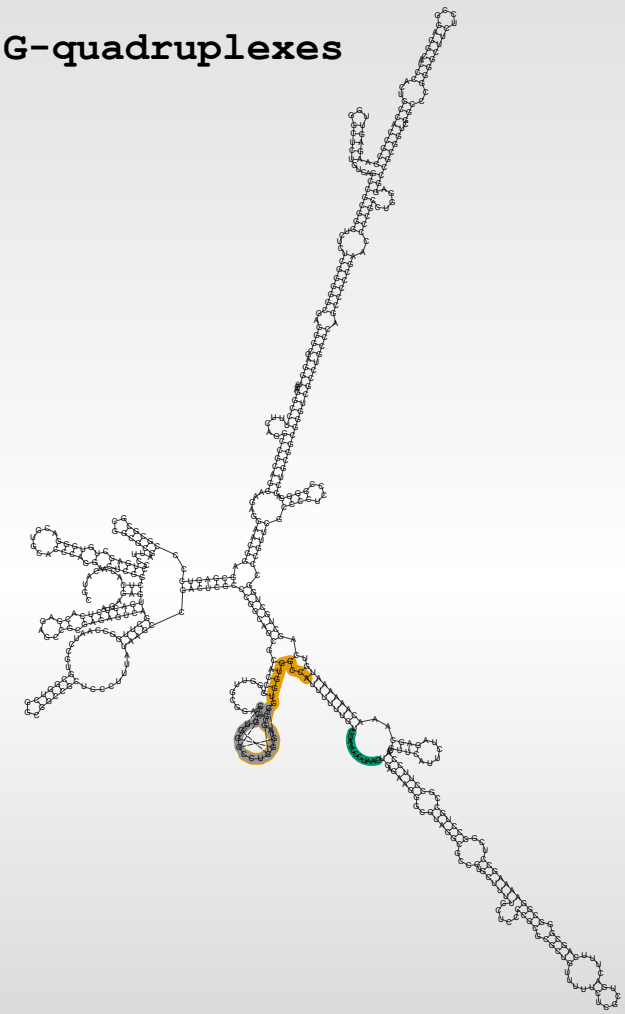


human Telomerase RNA Component (hTERC)

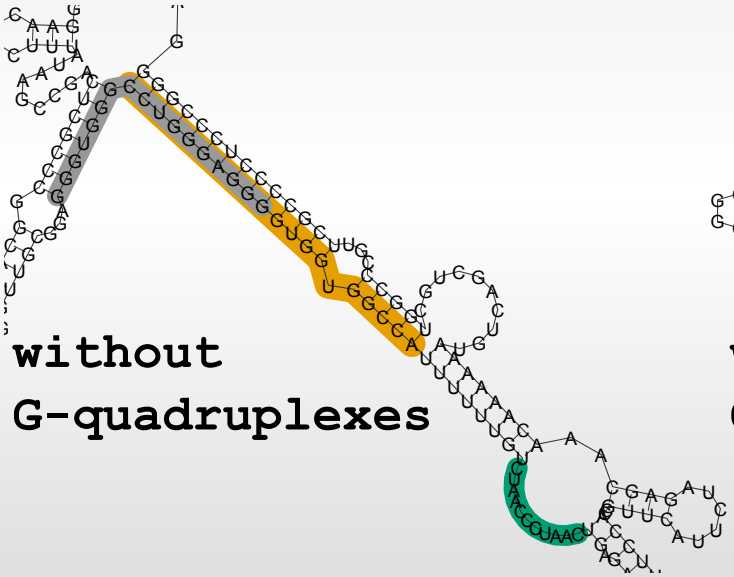
without G-quadruplexes



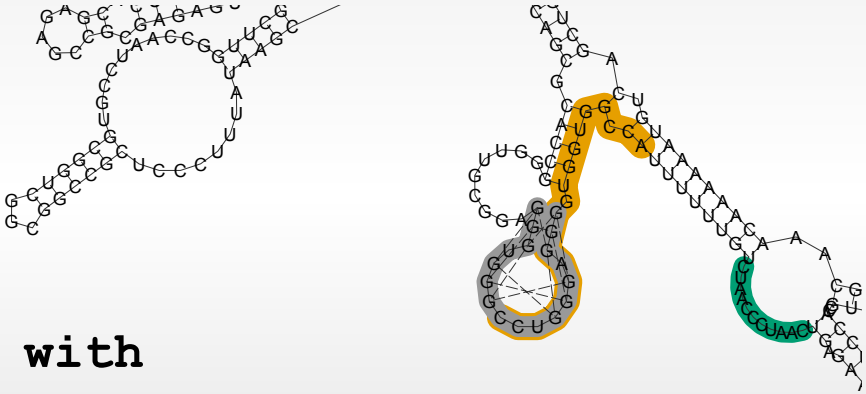
with G-quadruplexes



human Telomerase RNA Component (hTERC)



without
G-quadruplexes

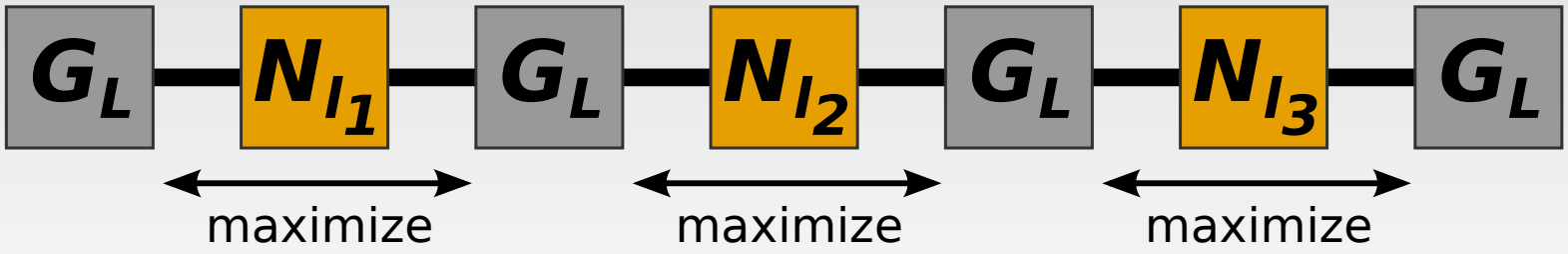


with
G-quadruplexes

human Telomerase RNA Component (hTERC)

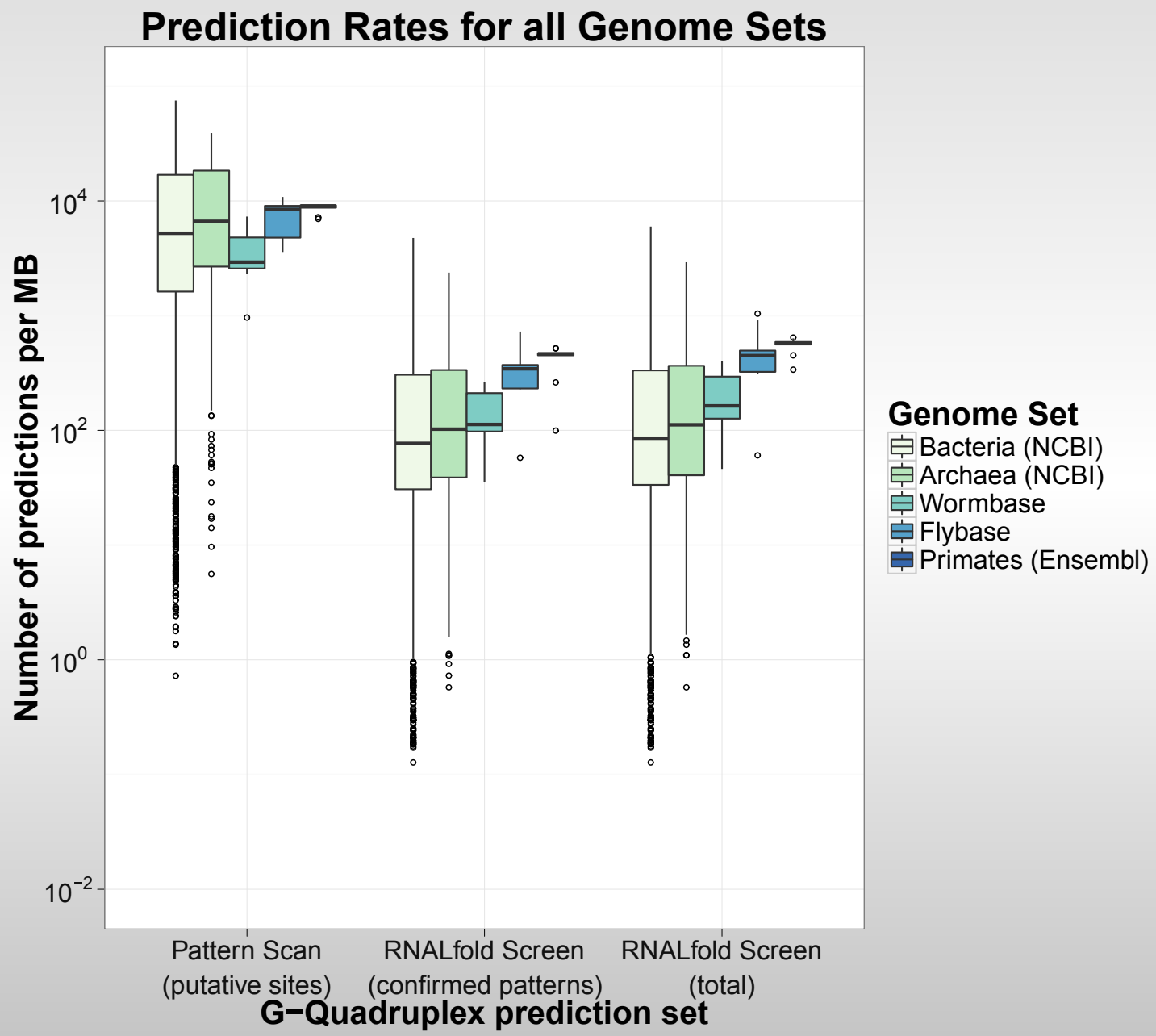
Genome wide G-Quadruplex Analysis

- Get putative G-Quadruplex sites (PGS) by scanning for

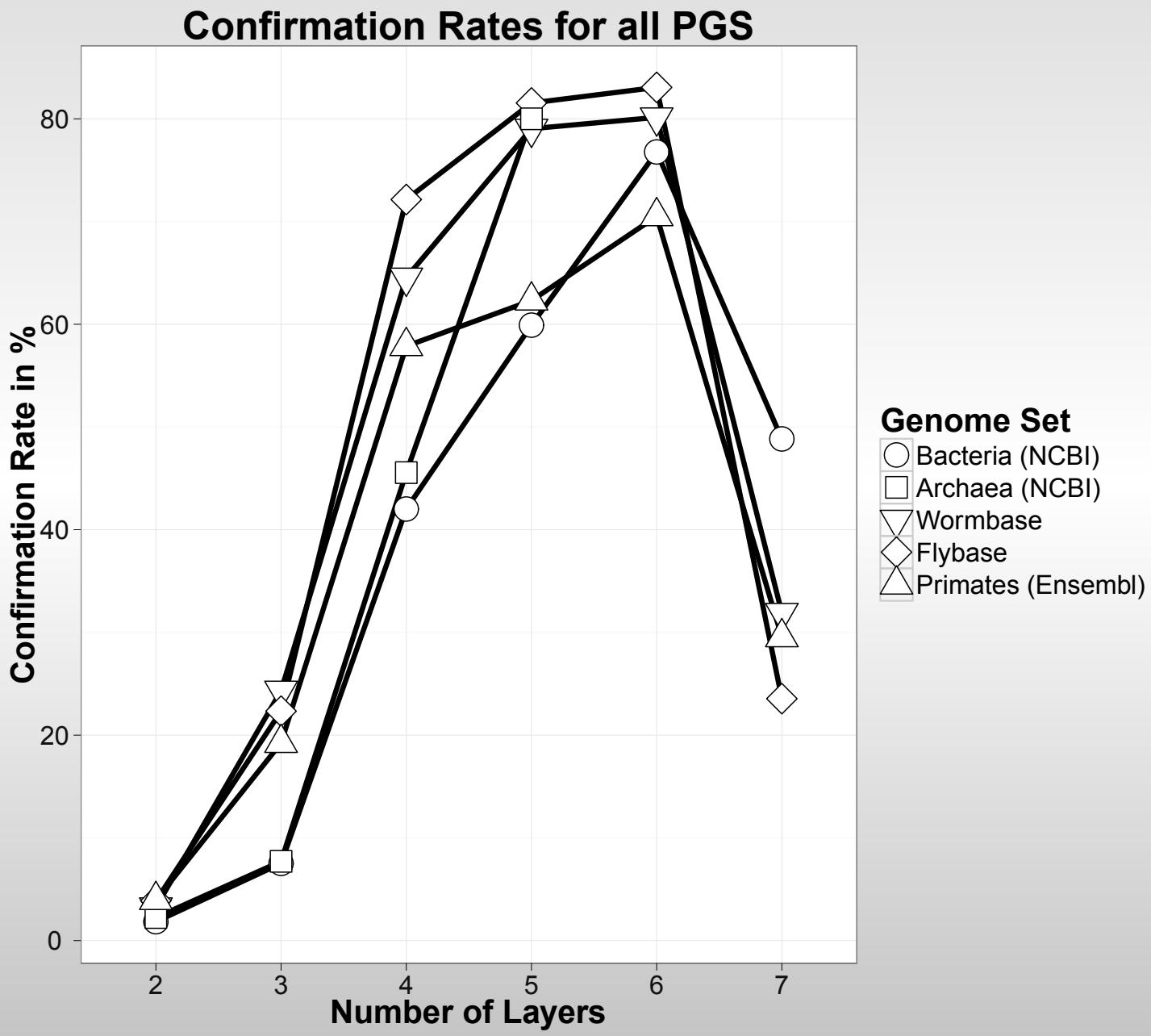


- Add 5′ and 3′ flanking region to each PGS
- Use these sequences to predict locally stable structures (`RNALfold`)
- Count how many putative sites are confirmed
- Count all unique stable G-Quadruplexes

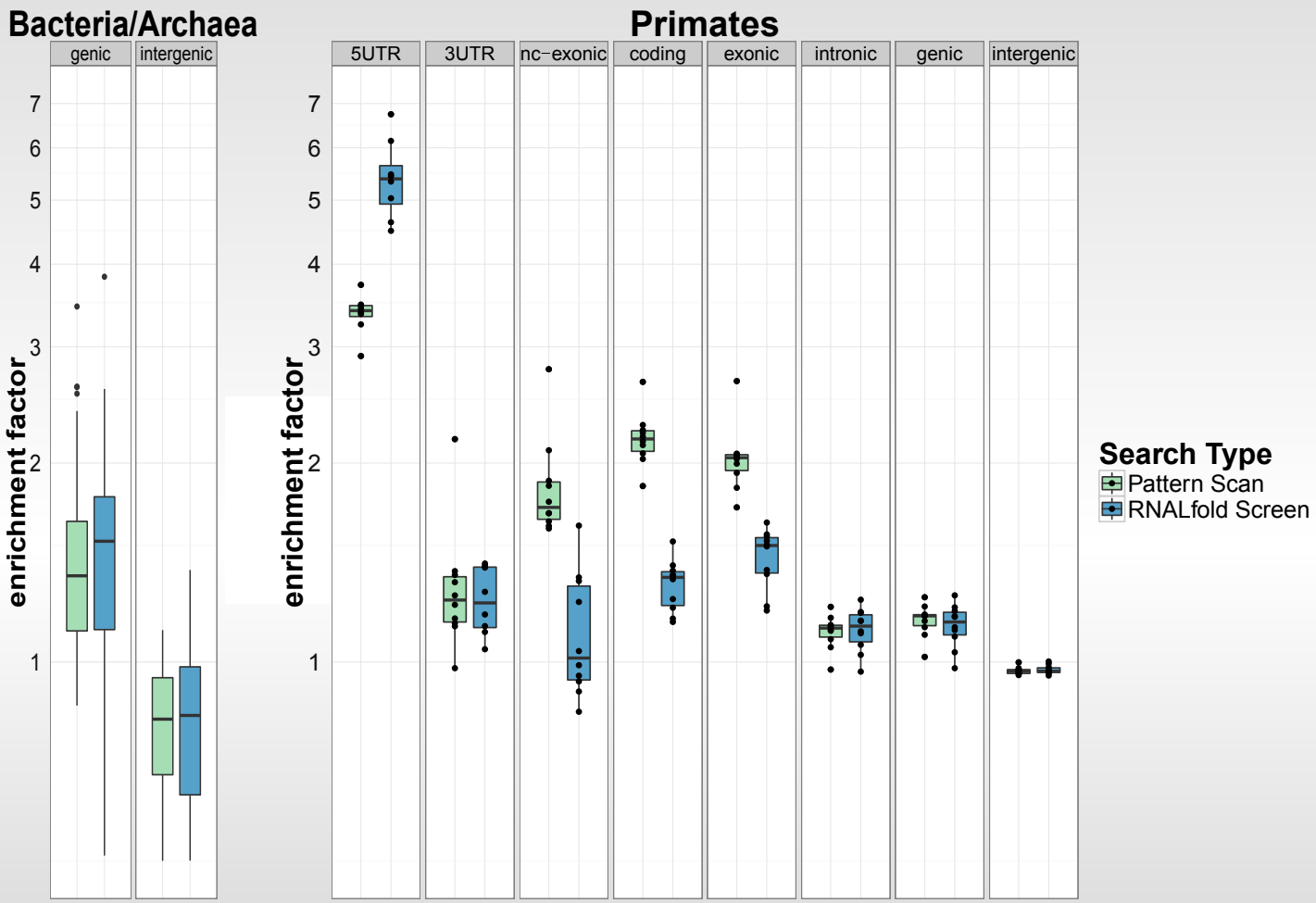
Genome wide G-Quadruplex Analysis



Genome wide G-Quadruplex Analysis



Enrichment Analysis



Conclusion and Outlook

- G-quadruplexes are important elements in gene regulation and cell life cycle
- Straight forward integration of G-tetrads into RNA folding DP recursions
- Implementation readily available and soon in main release of `ViennaRNA Package` (`http://www.tbi.univie.ac.at/RNA`)
- Genome wide scans for putative stable G-quadruplexes
- Only a very small amount ($\approx 2\%$) of PGS lead to thermodynamically stable G-quadruplexes
- Intersection with annotation data and enrichment analysis
- Cation (Na^+, K^+, Mg^{2+}) concentration dependancy
- RNA/RNA G-quadruplex Duplex structure prediction
- DNA G-quadruplex prediction
- RNA/DNA heterodimer G-quadruplexes

Conclusion and Outlook

- G-quadruplexes are important elements in gene regulation and cell life cycle
- Straight forward integration of G-tetrads into RNA folding DP recursions
- Implementation readily available and soon in main release of `ViennaRNA Package` (`http://www.tbi.univie.ac.at/RNA`)
- Genome wide scans for putative stable G-quadruplexes
- Only a very small amount ($\approx 2\%$) of PGS lead to thermodynamically stable G-quadruplexes
- Intersection with annotation data and enrichment analysis
- Cation (Na^+, K^+, Mg^{2+}) concentration dependancy
- RNA/RNA G-quadruplex Duplex structure prediction
- DNA G-quadruplex prediction
- RNA/DNA heterodimer G-quadruplexes

Need for more (better) energy parameters

Thanks to