Improving RNA secondary structure prediction

Ronny Lorenz
ronny@tbi.univie.ac.at

University of Vienna

Marseille, France, January 14, 2019

Secondary structure prediction is not perfect!

- can be done efficiently via DP (typically) in $\mathcal{O}(n^3)$
- very good accuracy for small RNAs
- accuracy drops to 40%-70% for longer sequences

How can we improve predictions?

- create better energy parameter set
- include ion concentrations
- guide the prediction with auxiliary data, e.g.
  1. comparative consensus structure prediction for homologous RNAs
  2. add constraints, e.g. experimental structure probing data
- extend the secondary structure model
  1. include pseudo-knots
  2. include additional (non-canonical) structure motifs
  3. include interaction with external factors
- folding dynamics, e.g. co-transcriptional folding

Guiding Structure Prediction with auxiliary data

1) Consensus structure prediction

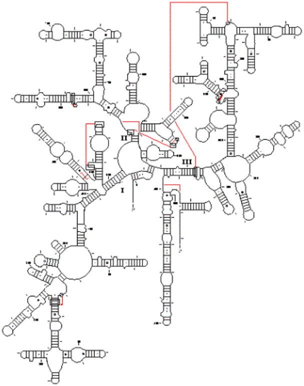
Consensus structures

Consensus structures are more **accurate**!

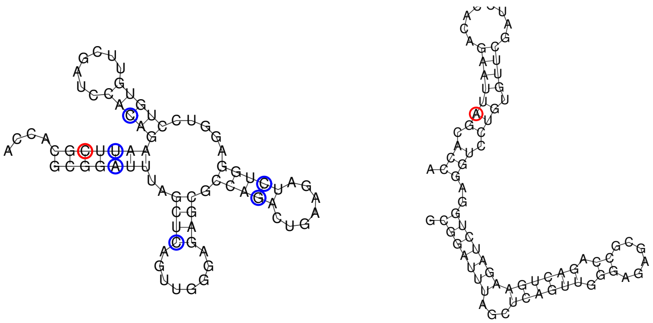
- Models of rRNA structures inferred from sequence comparison are highly accurate.
- Thermodynamic structure prediction often performs poorly

Comparative information may be included by:

- Considering the potential of structure conservation among homologous sequences
- Converting this information into a guiding potential



The Effect of Mutations



- Consistent and compensatory mutations often conserve the structure (blue)
- A single mutation (red) can radically change the structure
- Accumulating mutations quickly randomize any structure

Strategies for Predicting Consensus Structures

- Align Sequences, predict structure from alignment
  `RNAalifold`, `pfold`; `alidot`, `ConStruct`
  Sensitive to alignment errors
- Predict structures, then align structures
  `RNAforester`, `MARNA`
  Possibly sensitive to prediction errors
- Combine structure prediction and alignment
  The "Sankoff algorithm" `FoldAlign`, `DynAlign`, `stemloc`, `PMcomp`, `LocARNA`
- Alignment-free: Predict near-optimal coarse grained structures
  look for shapes common to all sequences `RNAcast`

Strategies for Predicting Consensus Structures

- Align Sequences, predict structure from alignment
  `RNAalifold`, `pfold`; `alidot`, `ConStruct`
  Sensitive to alignment errors
- Predict structures, then align structures
  `RNAforester`, `MARNA`
  Possibly sensitive to prediction errors
- Combine structure prediction and alignment
  The “Sankoff algorithm” `FoldAlign`, `DynAlign`, `stemloc`, `PMcomp`, `LocARNA`
- Alignment-free: Predict near-optimal coarse grained structures
  look for shapes common to all sequences `RNAcast`

Sebastian will talk about these things on Thursday...

2) Incorporate experimental structure probing data

Experimental structure probing

- Chemical or enzymatic probing experiments
- Some already used before first structure prediction approaches
- Specifically modify or cleave single stranded and/or double stranded regions
- Ribonucleases, lead(II), CMCT, DMS, SHAPE, inline probing, etc.

General protocol

- Prepare sample RNA and add probing reagent(s)
- Determine modification / cleavage sites with
  1. Gel electrophoresis
  2. Reverse transcription and (high throughput) sequencing
  3. Reverse transcription aborts at modified/cleaved site or yield a mutated nucleotide
- Convert reactivities into constraints (binary, probabilities, pseudo-energies)
- Manual or computational structure modeling

Probing signal is one-dimensional!

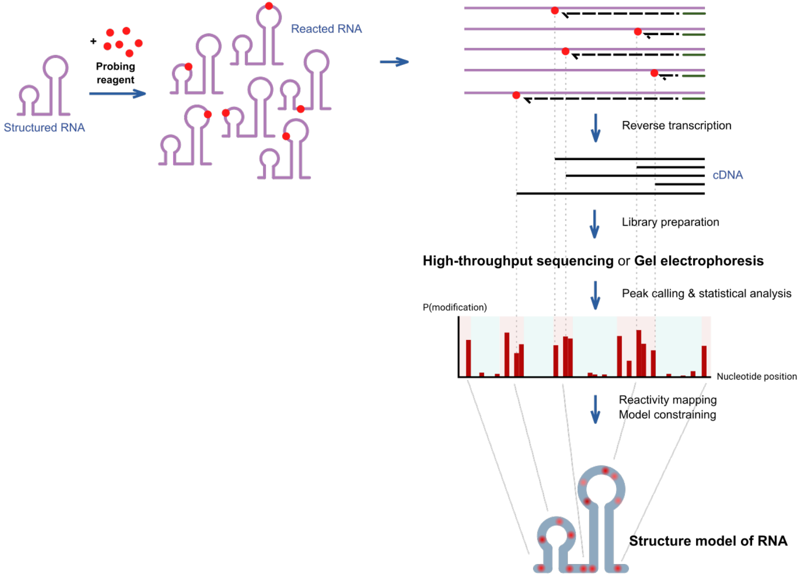
Experimental structure probing

- Chemical or enzymatic probing experiments
- Some already used before first structure prediction approaches
- Specifically modify or cleave single stranded and/or double stranded regions
- Ribonucleases, lead(II), CMCT, DMS, **SHAPE**, inline probing, etc.

General protocol

- Prepare sample RNA and add probing reagent(s)
- Determine modification / cleavage sites with
  1. Gel electrophoresis
  2. Reverse transcription and (high throughput) sequencing
  3. Reverse transcription aborts at modified/cleaved site or yield a mutated nucleotide
- Convert reactivities into constraints (binary, probabilities, pseudo-energies)
- Manual or computational structure modeling

Probing signal is one-dimensional!

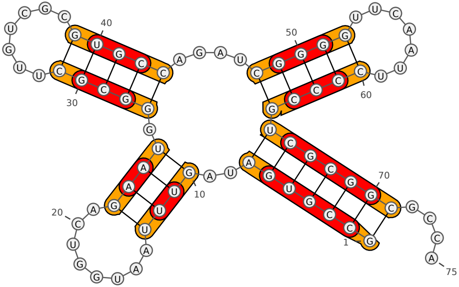


Adapted from *Ptrw08, A schematic figure explaining the steps in a typical chemical probing experiment to assay the structure of RNA molecules, CC BY-SA 4.0*

SHAPE reactivity in secondary structure prediction

Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE)

- Reactivity probes flexibility of backbone
- No nucleobase bias
- Assume flexible means unpaired
- Convert reactivity to pseudo-energy for prediction
  Deigan et al. [2009] (stacked pairs)

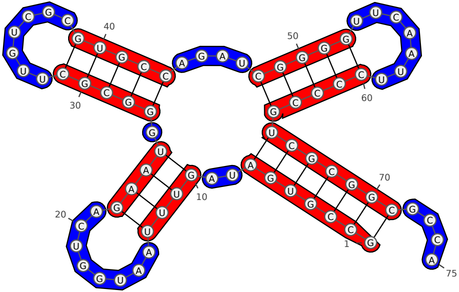


$$\Delta\ G(i) = m * ln(reactivity[i] + 1) + b$$

SHAPE reactivity in secondary structure prediction

Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE)

- Reactivity probes flexibility of backbone
- No nucleobase bias
- Assume flexible means unpaired
- Convert reactivity to pseudo-energy for prediction
  Zarringhalam et al. [2012] (unpaired bases and base pairs)



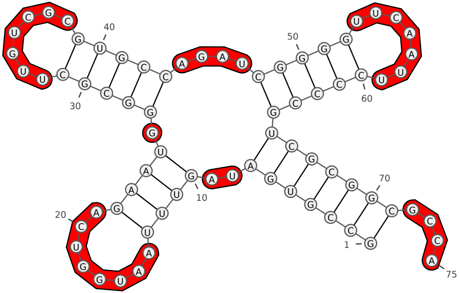
$$\Delta\ G(x, i) = \beta\ * |x - q_i|$$

$$x \in [0(unpaired), 1(paired)]$$

SHAPE reactivity in secondary structure prediction

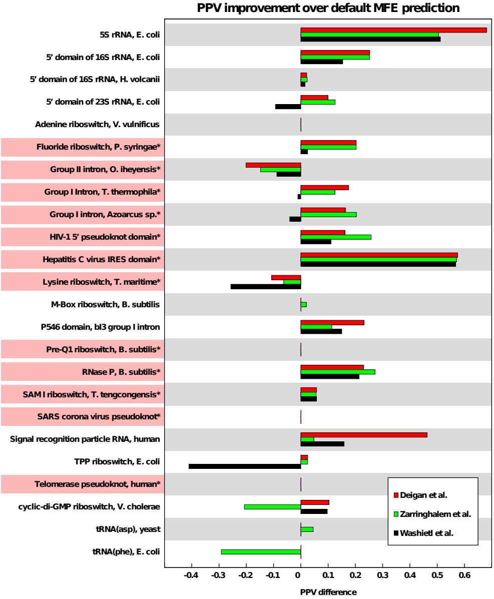
Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE)

- Reactivity probes flexibility of backbone
- No nucleobase bias
- Assume flexible means unpaired
- Convert reactivity to pseudo-energy for prediction
  Washietl et al. [2012] (unpaired bases)



$$F(\vec{\epsilon}) = \sum_{i=1}^{n} \frac{\epsilon_i^2}{\tau^2} + \sum_{i=1}^{n} \frac{(p_i(\vec{\epsilon}) - q_i)^2}{\sigma^2} \rightarrow min$$

SHAPE reactivity in secondary structure prediction



Conclusions

- Experimental data can substantially improve prediction
- High-throughput probing became quite popular in last decade
- Multiple predictions with different data for consensus modeling
- Methods such as `Shape-MaP` can even reveal multiple sites on a single RNA strand

Probing reactivities must be taken with great care! They ...

- tend to differ from one to the other experiment (even when performed in same lab)
- may have poor discriminative power
- usually reflect an ensemble of conformations
- include more than secondary structure (pseudoknots, tertiary interactions, etc)

So what?

- reactivity preparation must be robust
- tools need to be flexible with respect to inclusion of data
- deconvolution of probing data is still a problem

Outlook - Hands-on session

Secondary structure constraints:

- **Hard**: disallow certain parses of the decomposition scheme
  → add / remove particular (sub)structures from the candidates
- **Soft**: modify the energy contributions of the model
  → (de-)stabilize particular (sub)structures

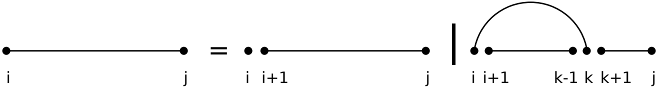
Mostly limited to particular use-cases

- suboptimal structures *sensu* M. Zuker
- mark modified bases (as unpaired)
- recompute optimal structure given a consensus
- simulations of translocating an RNA through a pore
- incorporate probing data (SHAPE, DMS, PARS)
- incorporate protein/ligand binding

The `ViennaRNA Package` provides a most generic implementation of hard and soft constraints!

Outlook - Hands-on session

generic hard and soft constraints (basic idea)



Outlook - Hands-on session

generic hard and soft constraints (basic idea)

$$N_{ij} = X_{ii} \cdot \{N_{i+1,j} + E^{u}(i)\} + \sum_{k=i+1}^{j} X_{ik} \cdot \{N_{i+1,k-1} + N_{k+1,j} + E^{bp}(i,k)\}$$

Outlook - Hands-on session

generic hard and soft constraints (basic idea)

$$N_{ij} = X_{ii} \cdot \{N_{i+1,j} + E^{u}(i)\} + \sum_{k=i+1}^{j} X_{ik} \cdot \{N_{i+1,k-1} + N_{k+1,j} + E^{bp}(i,k)\}$$

The `ViennaRNA Package` discriminates full Nearest Neighbor scheme

Hard constraints: X expressed in terms of a Boolean function

$$f : \mathbb{N}^m \times \mathbb{D} \rightarrow 0|1$$

Soft constraints: E expressed in terms of a Real-valued function

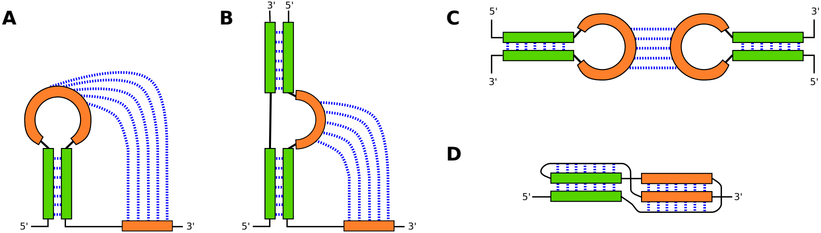
$$f : \mathbb{N}^m \times \mathbb{D} \rightarrow \mathbb{R}$$

with m nucleotide positions, and decomposition step $d \in \mathbb{D}$.

Extending the dynamic programming scheme

1) Pseudoknots

Pseudoknots



- quite common in natural RNA structures
- left out in most predictions due to algorithmic complexity (NP hard for arbitrarily complex pseudoknots)
- only a small number of energy models exist
- very sensitive to cation concentrations (Mg^{2+})

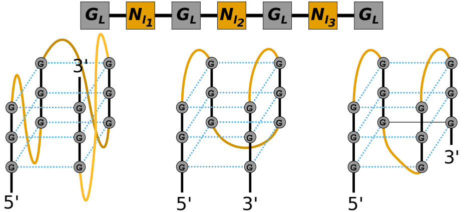
So what?

- limit predictive model to particular subclasses (*ab initio*)
- resort to heurists, e.g predict (suboptimal) secondary structures first and insert pseudoknots later (*a posteriori*)

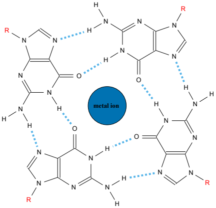
2) 2.5D motifs - The case of G-Quadruplexes

What are G-Quadruplexes

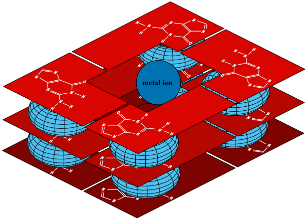
- G-rich nucleic acid sequences forming stacks of G-quartets
- Stable local structure of 4 interconnected strands
- 2-5 (**L**) quartet layers connected by 3 short loops (l_1, l_2, l_3)


G_L — N_{l_1} — G_L — N_{l_2} — G_L — N_{l_3} — G_L
G
3'
5'
3'
5'
3'
5'

parallel anti-parallel mixed



Hogsteen-Watson Crick bonds



$\pi-$orbital stacking

Where are G-Quadruplexes

DNA:

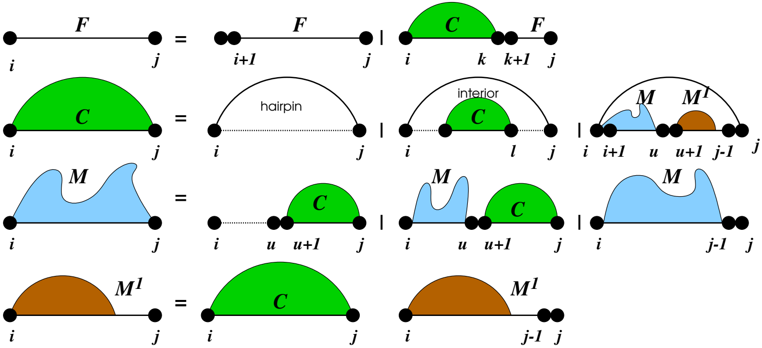
- Human Telomers: Telomerase inhibition
- Promotor Regions: Modulation of gene transcription
- Elsewhere: Interference with protein function

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

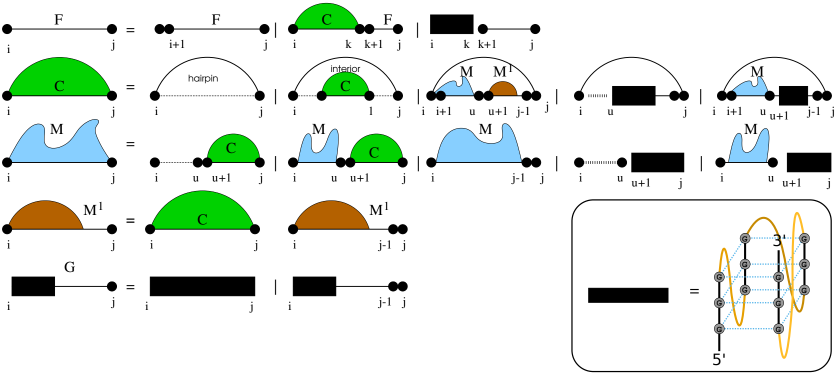
RNA secondary structure prediction with G-Quadruplexes

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



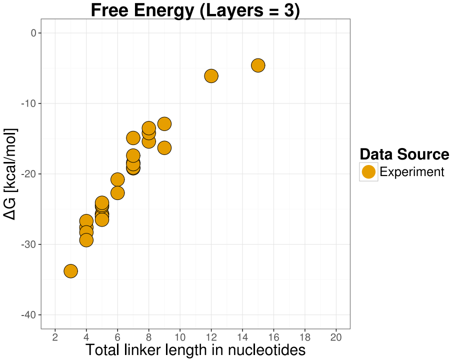
RNA secondary structure prediction with G-Quadruplexes

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



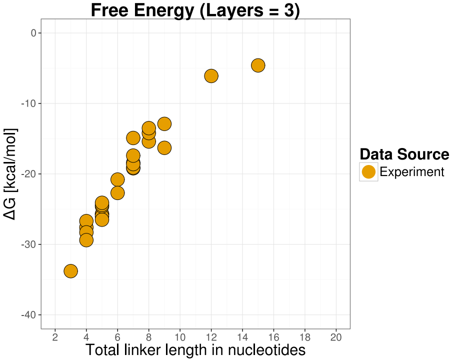
RNA secondary structure prediction with G-Quadruplexes

UV melting data from Zhang et al., Biochemistry 2011



RNA secondary structure prediction with G-Quadruplexes

UV melting data from Zhang et al., Biochemistry 2011

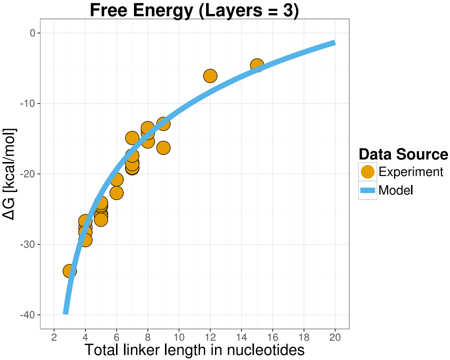


- Energy $\propto$ number of layers - 1
- Energy $\propto$ total linker length
- No effect of linker asymmetry or sequence composition

$$\begin{aligned} E(L,l,T) &= a(T)(L-1) + b(T)\ln(l-2) \quad (1)\\ a(T) &= H_a + TS_a \quad (2)\\ b(T) &= H_b + TS_b \quad (3) \end{aligned}$$

RNA secondary structure prediction with G-Quadruplexes

UV melting data from Zhang et al., Biochemistry 2011

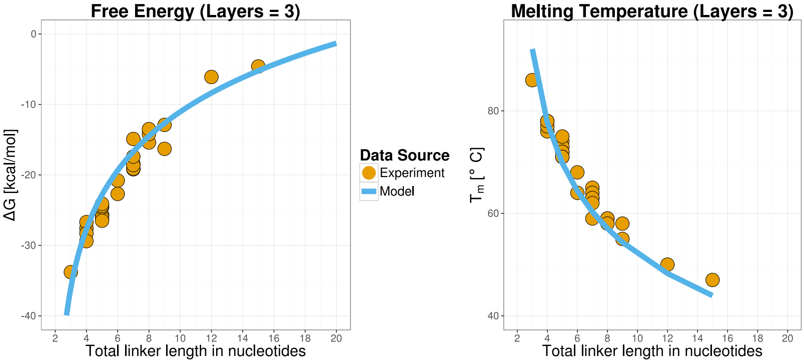


- Energy $\propto$ number of layers - 1
- Energy $\propto$ total linker length
- No effect of linker asymmetry or sequence composition

$$\begin{aligned} E(L, l, T) &= a(T)(L-1) + b(T)\ln(l-2) && (1) \\ a(T) &= H_a + TS_a && (2) \\ b(T) &= H_b + TS_b && (3) \end{aligned}$$

RNA secondary structure prediction with G-Quadruplexes

UV melting data 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

$$
\begin{aligned}
E(L, l, T) &= a(T)(L-1) + b(T)\ln(l-2) \qquad (1)\\
a(T) &= H_a + TS_a \qquad (2)\\
b(T) &= H_b + TS_b \qquad (3)
\end{aligned}
$$

RNA secondary structure prediction with G-Quadruplexes

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

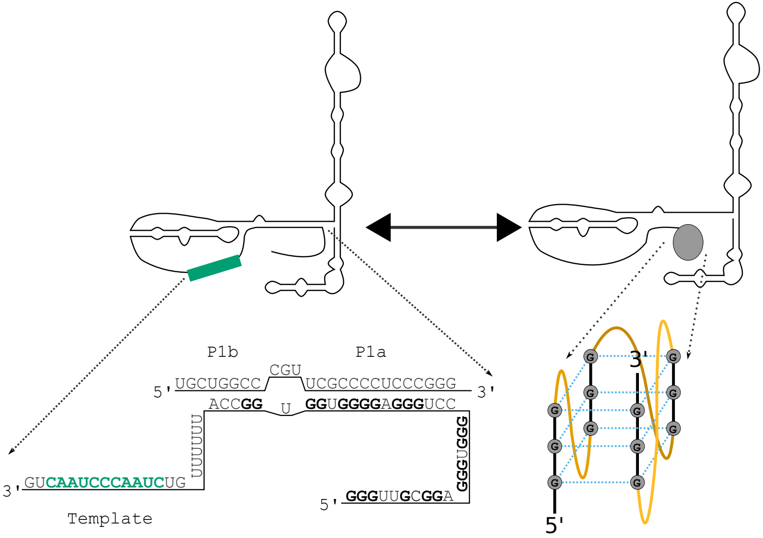
`RNAcofold` MFE-Structure, Concentration Dependent Base Pair Probabilities, Partition Function for Dimers

`RNALfold` Locally Stable Structure Prediction for Single Sequences

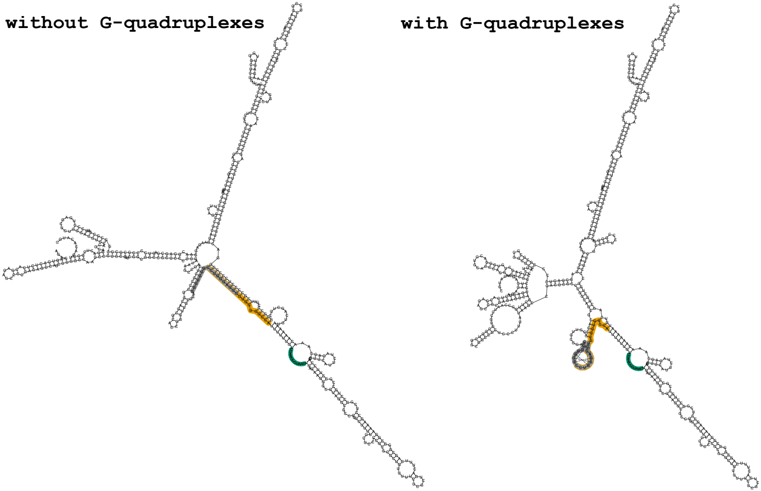
`RNALalifold` Locally Stable Structure Prediction for Sequence Alignment

`RNAsubopt` Suboptimal Structure Prediction for Single Sequences and Sequence Dimers

human Telomerase RNA Component (hTERC)



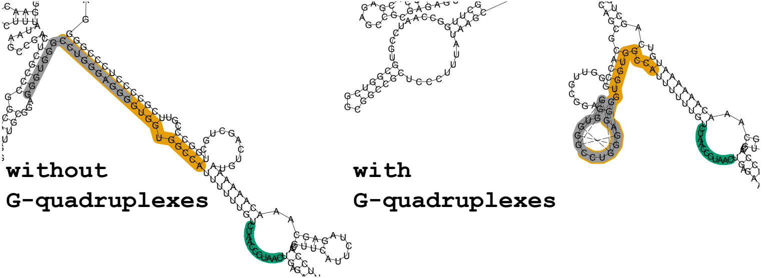
human Telomerase RNA Component (hTERC)



without G-quadruplexes

with G-quadruplexes

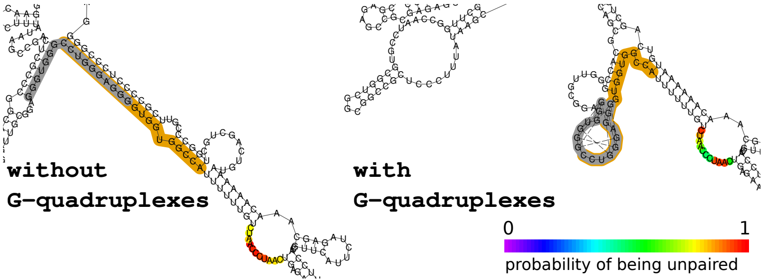
human Telomerase RNA Component (hTERC)



without G-quadruplexes

with G-quadruplexes

human Telomerase RNA Component (hTERC)



without G-quadruplexes

with G-quadruplexes

0 1

probability of being unpaired

RNA secondary structure prediction with G-Quadruplexes

G-quadruplexes are . . .

- important elements in gene regulation and cell life cycle
- in competition with *regular* structure formation
- straight forward to integrate into RNA folding DP recursions

Answers:

- Genome wide scans reveal only a very small amount ($\approx$ 2%) of PGS lead to thermodynamically stable G-quadruplexes [1]
- sometimes conserved across species
- same scheme may be applicable to other 2.5D motifs

What's missing:

- cation (Na^{+}, K^{+}, Mg^{2+}) concentration dependancy
- interstrand G-quadruplex structure prediction
- DNA G-quadruplex prediction
- RNA/DNA heterodimer G-quadruplexes

[1] Lorenz et al. 2013, ”2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction”

3) Ligand binding

Ligand binding and Constraints

Recall the partition function

$$Q = \sum_{s\in\Omega} e^{-E(s)/RT}$$

Including a ligand L with dissociation constant K_d and concentration c for an RNA with a single binding site (aptamer motif) leads to

$$Q_L = Q + Q^A \cdot \frac{K_d}{c}, \quad \text{with} \quad Q^A = \sum_{s|A\in s} e^{-E(s)/RT}$$

For more than one binding site $A_1, A_2, \ldots$ this quickly becomes infeasible to compute

$$Q_L = Q + (Q^{A_1} + Q^{A_2}) \cdot \frac{K_d}{c} + Q^{A_1 A_2} \cdot (\frac{K_d}{c})^2 + \ldots$$

Ligand binding and Constraints

Recall the partition function

$$Q \quad = \quad \sum_{s \in \Omega} e^{-E(s)/RT}$$

Including a ligand L with dissociation constant K_d and concentration c for an RNA with a single binding site (aptamer motif) leads to

$$Q_L \quad = \quad Q + Q^A \cdot \frac{K_d}{c}, \quad \text{with} \quad Q^A = \sum_{s | A \in s} e^{-E(s)/RT}$$

For more than one binding site $A_1, A_2, \ldots$ this quickly becomes infeasible to compute

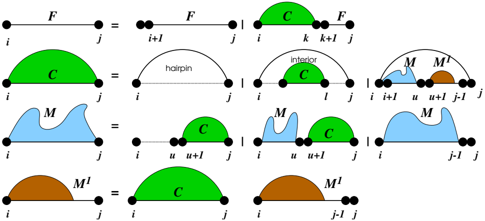
$$Q_L \quad = \quad Q + (Q^{A_1} + Q^{A_2}) \cdot \frac{K_d}{c} + Q^{A_1 A_2} \cdot (\frac{K_d}{c})^2 + \ldots$$

Solution: Explicitly include aptamer into decomposition scheme

Ligand binding in unstructured regions

1. What about using generic soft-constraints?

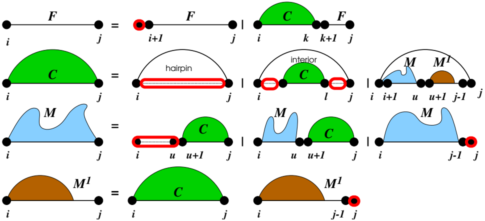
Nearest Neighbor Model



Ligand binding in unstructured regions

1. What about using generic soft-constraints?

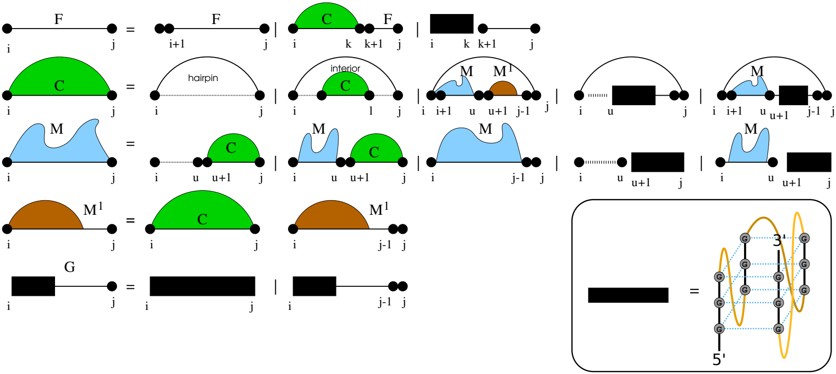
Nearest Neighbor Model



Ligand binding in unstructured regions

2. Extending the decomposition scheme

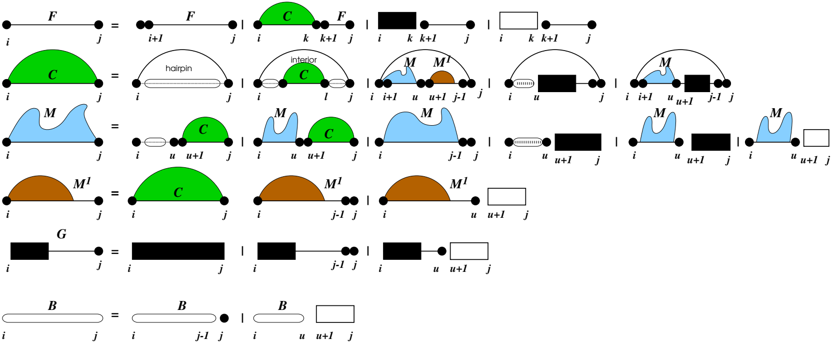
Nearest Neighbor Model with G-Quadruplexes



Ligand binding in unstructured regions

2. Extending the decomposition scheme

Nearest Neighbor Model with G-Quadruplexes and Ligands


F = F | C F | |
i j i+1 j i k k+1 j i k k+1 j i k k+1 j
C = hairpin | interior C | M M1 | | M
i j i j i l j i i+1 u u+1 j-1 j i u j i i+1 u u+1 j-1 j
M = C | M C | M | | M | M
i j i u u+1 j i u u+1 j i j-1 j i u u+1 j i u u+1 j i u u+1 j
M1 = C M1 | M1
i j i j i j-1 j i u u+1 j
G = | |
i j i j i j-1 j i u u+1 j
B = B | B
i j i j-1 j i u u+1 j

Conclusion

- ligand binding may be dealt with using constraints
- generally this leads to combinatorial explosion of constrained computations
- specific aptamer motifs may be included by extending the recursion scheme

The `ViennaRNA Package` implements ligand binding in $O(n^3)$

- to hairpin- or interior loop-like motifs (through soft constraints)
- to unstructured domains (through extension of decomposition scheme)

Drawbacks:

- still, cooperate effects of ligand binding is neglected
- changes in concentration requires re-computation of partition function

Let’s get our hands dirty trying out what we’ve learned so far in the afternoon!