Jigsaw Puzzle
RNA structure alignment including pseudoknots

Christina Witwer

Institut für Theoretische Chemie und Molekulare Strukturbiologie

Universität Wien

Bled, 2004

Outline

- Introduction
- Carnac
  secondary structure alignment of two sequences
- Hxmatch
  consensus structure including pseudoknots on a set of aligned sequences
- Puzzle
  structure alignment including pseudoknots on a set of (poorly aligned) sequences

Prediction of Consensus Structure

- Secondary structure
  - fold an alignment
    `RNAalifold`, `qrna`; `alidot`, `ConStruct`
  - align structures
    `RNAforester`, `MARNA`
  - align and fold simultaneously (Sankoff Algorithm)
    `Foldalign`, `Dynalign`, `PMmatch` - expensive in CPU time
    `Carnac`

- Structure including pseudoknots
  - Structure prediction based on an alignment
    `ILM, Hxmatch`

CARNAC

Finding the common structure shared by two homologous RNAs

O. Perriquet, H. Touzet and M. Dauchet (2003). *Bioinformatics*, **19**, 108-116

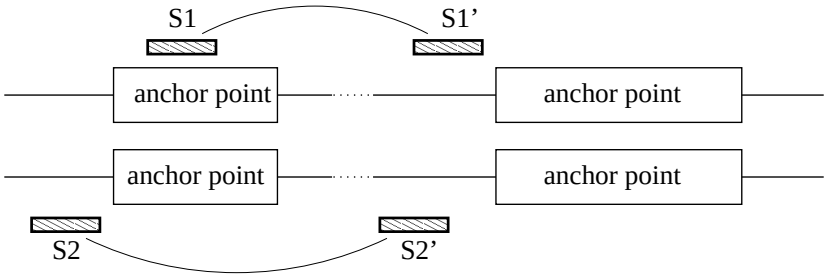
- scan each sequence for the best candidate stems
- search for regions of high sequence similarity

  $=>$ *anchor points*
- pairwise selection of matchable stems (several matchable partners are possible for a stem)
- construct common folding (variant of the Sankoff algorithm)

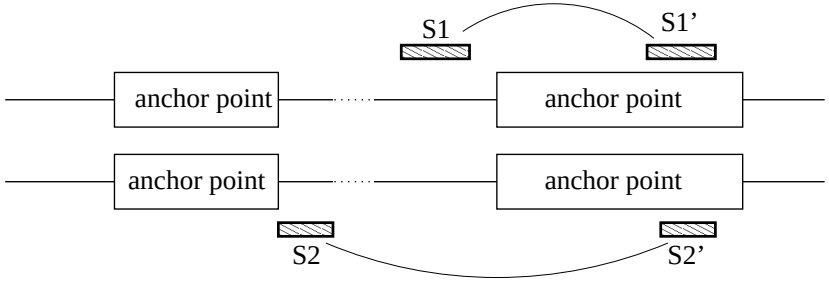
CPU time about $O(n^4)$

CARNAC (II)

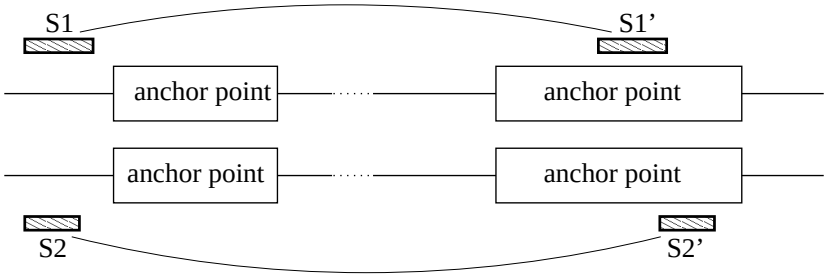
Pairwise selection of matchable stems



anchor point violation

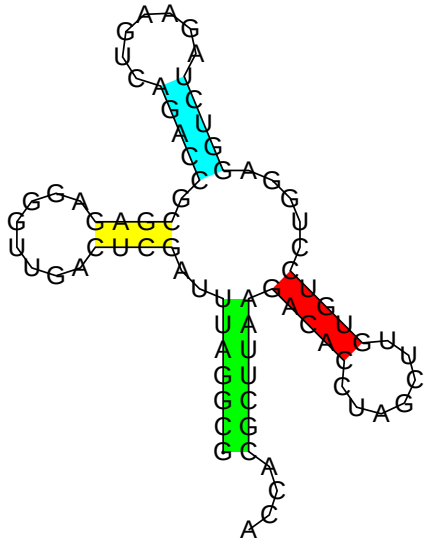


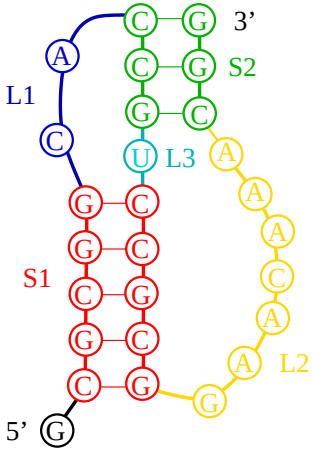
shift too large outside an anchor point

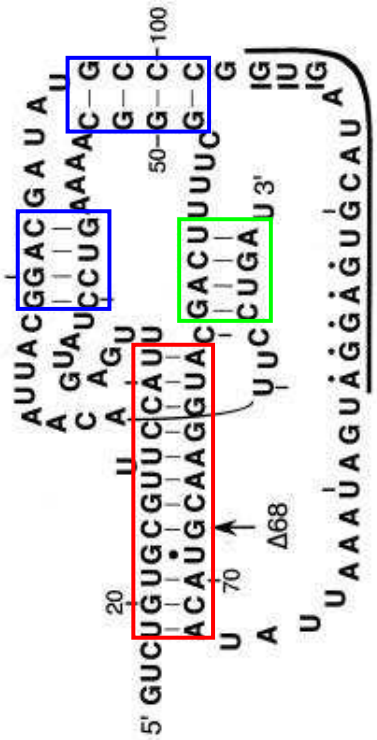


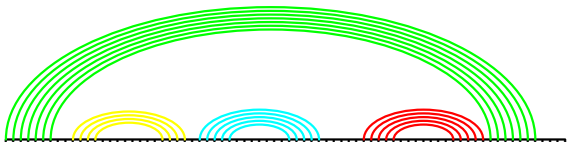
shift too large within an anchor point

Secondary Structure Including Pseudoknots

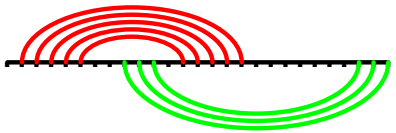




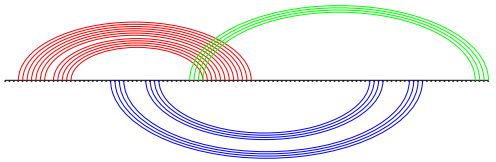




secondary structure
(outer-planar graph)



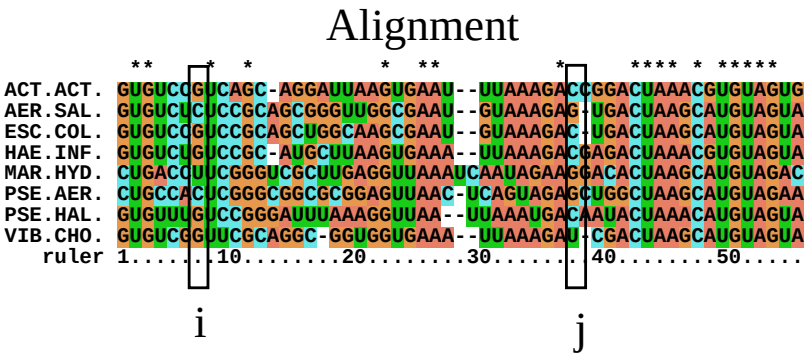
bi-secondary structure
(planar graph)



non-planar structure

Hxmatch

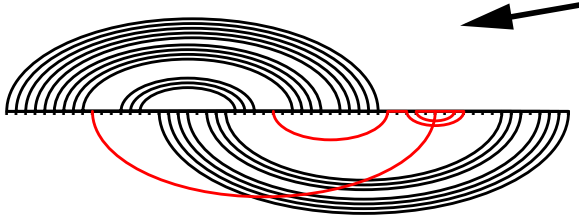
Prediction of RNA consensus bi-secondary structure based on a set of aligned sequences.



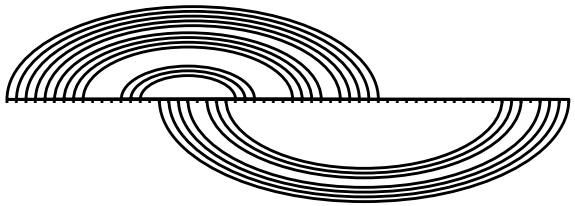
1.) Generate Scoring Matrix

	G	U	C	A	U	C	G	A
G	-	0	22	0	30	18	0	0
U	-	-	0	13	0	0	0	25
C	-	-	-	0	0	0	27	3
A	-	-	-	-	0	0	0	0
U	-	-	-	-	-	0	24	19
C	-	-	-	-	-	-	0	0
G	-	-	-	-	-	-	-	0
A	-	-	-	-	-	-	-	-

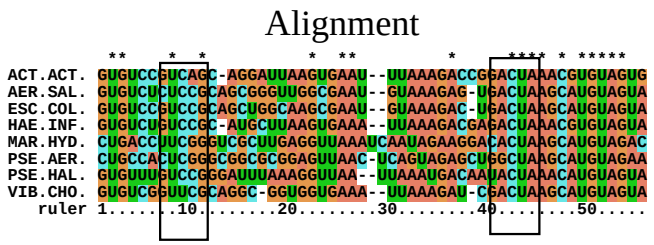
2.) MWM



3.) Post Processing



Hxmatch: Generating the Scoring Matrix



Thermodynamic score:

- based on the stacking energies of the helices

$$T_{ij} = \frac{1}{N} \sum_{\alpha \in \mathbb{A}} -\Delta G^{\alpha}_{\Psi} \text{ for all } ij \in \Psi$$

Covariation score:

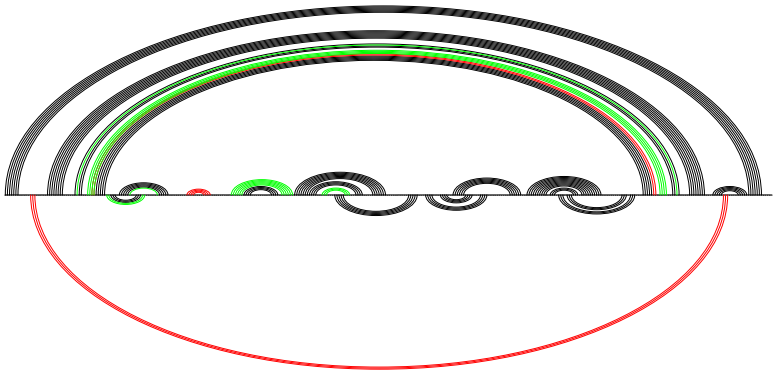
$$C_{ij} = \sum_{\mathrm{xy,x'y'}} f_{ij}(\mathrm{xy}) \mathbf{D}_{\mathrm{xy,x'y'}} f_{ij}(\mathrm{x'y'}) - \varphi_1 q_{ij}$$

where $\mathbf{D}_{\mathrm{xy,x'y'}}$ contains $d_H(xy, x'y')$ if xy and $x'y'$ are allowed pairs , else 0.

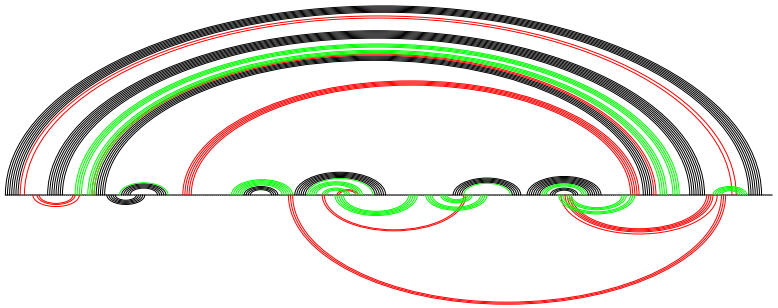
$d_{ij}^{\alpha,\beta}$ is the hamming distance of α and β at positions i and j (i.e. 0,1, or 2).

q_{ij} is the percentage of inconsistent sequences at positions i and j.

Hxmatch Results: tmRNA, 8 sequences



databank alignment
$SS = 84\%$, $SP = 91\%$



clustalw alignment
$SS = 51\%$, $SP = 69\%$

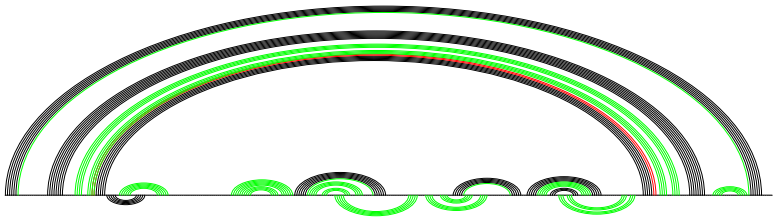
$SS = TP/RP * 100$

$SP = TP/(TP + FP) * 100$

TP . . . #true positives

RP . . . #pairs in the reference structure

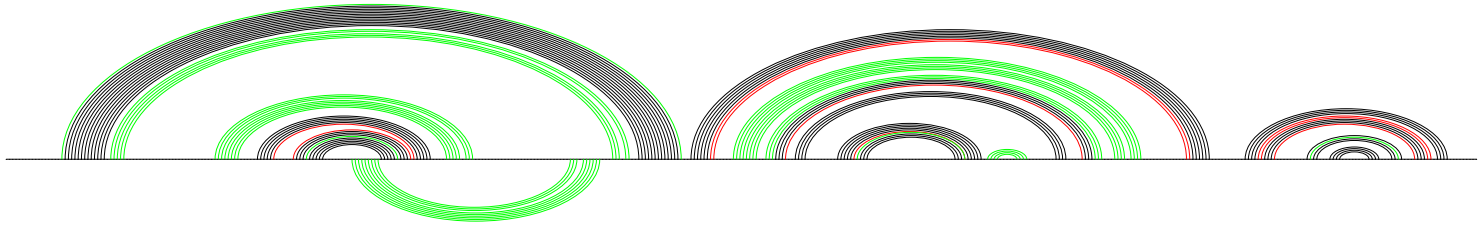
FP . . . #false positives



clustalw alignment (only the best)
$SS = 41\%$, $SP = 96\%$

Hxmatch: Results

Telomerase RNA: 8 sequences
clustalw alignment (only the best)



$SS = 60\%,\ SP = 84\%$

For all datasets investigated:

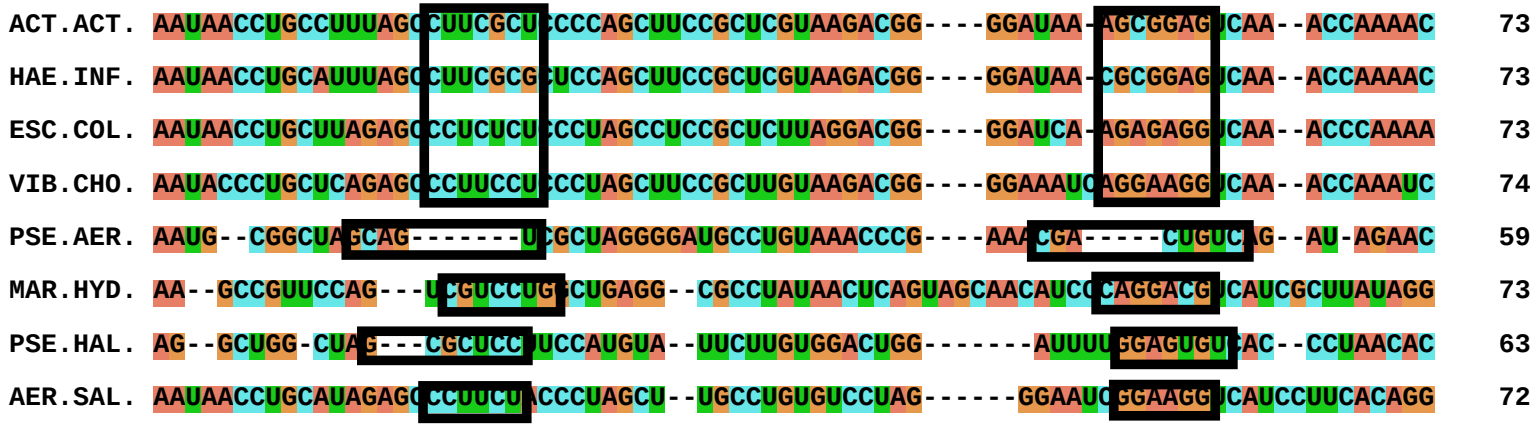
$SS : 40 - 60\%$

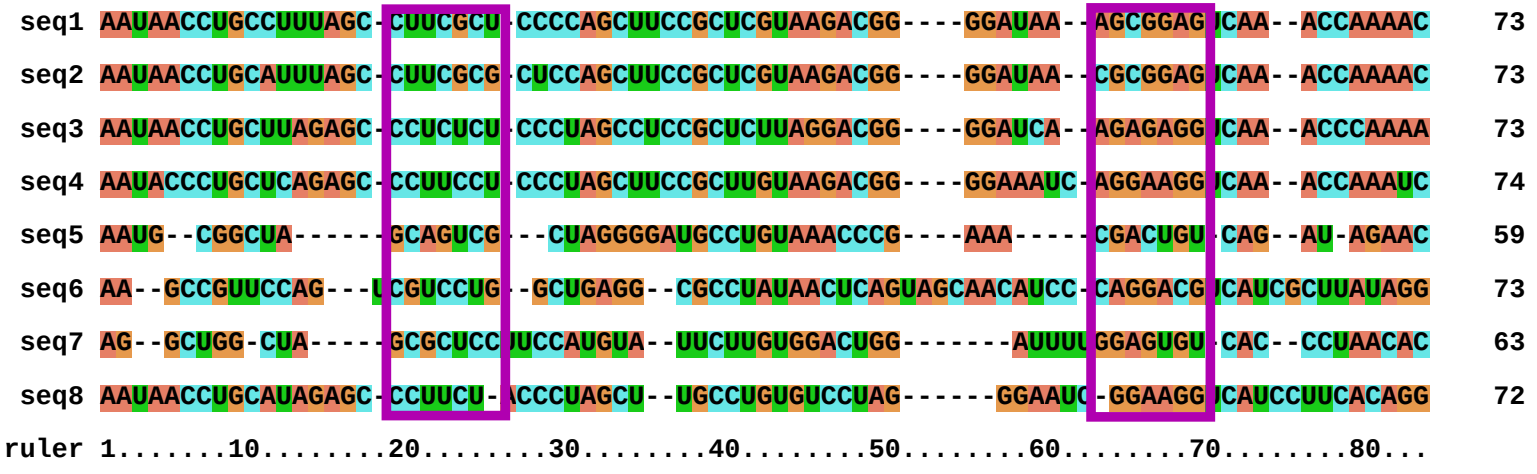
$SP : 85 - 95\%$

Puzzle(I)

- believe `hx`match prediction => 'true' helices
- 'repair' the true helices
  - search for good helices *close* to the true helix
  - take the thermodynamically most stable helix as true
  - adapt the alignment accordingly
- split each sequence into intervals given by the true helices

Puzzle(II): Repair true helices





Puzzle(III): preselect helices

for each pair of intervals:

- select the thermodynamically most stable helices of each sequence and put them into the list of 'excellent' helices
- for each excellent helix:
  - search for good helices in the other sequences *close* to the excellent helix
  - take the thermodynamically most stable helix as a match
- select helices which have a match in at least 80% of the sequences

Puzzle(IV)

- calculate hxmatch score for each selected helix
- choose base pairs with the help of the Maximum Weighted Matching algorithm
- possibly start again with relaxed constraints

Puzzle - Summary

- Hxmatch prediction $\mathcal{O}(n^3)$
- 'repair' predicted helices $\mathcal{O}(N * l^2 * n)$
- select candidate matching helices $\mathcal{O}(N^2 * l^2 * n^2)$
- calculate score (combine thermodynamics and covariation)
- MWM $\mathcal{O}(n^3)$

n . . . length of the alignment; N . . . number of sequences;

l . . . length of the interval defining 'close'

Thanks