Prediction of Conserved and Consensus RNA Structures

Christina Witwer

Institut für Theoretische Chemie und Molekulare Strukturbiologie

Universität Wien

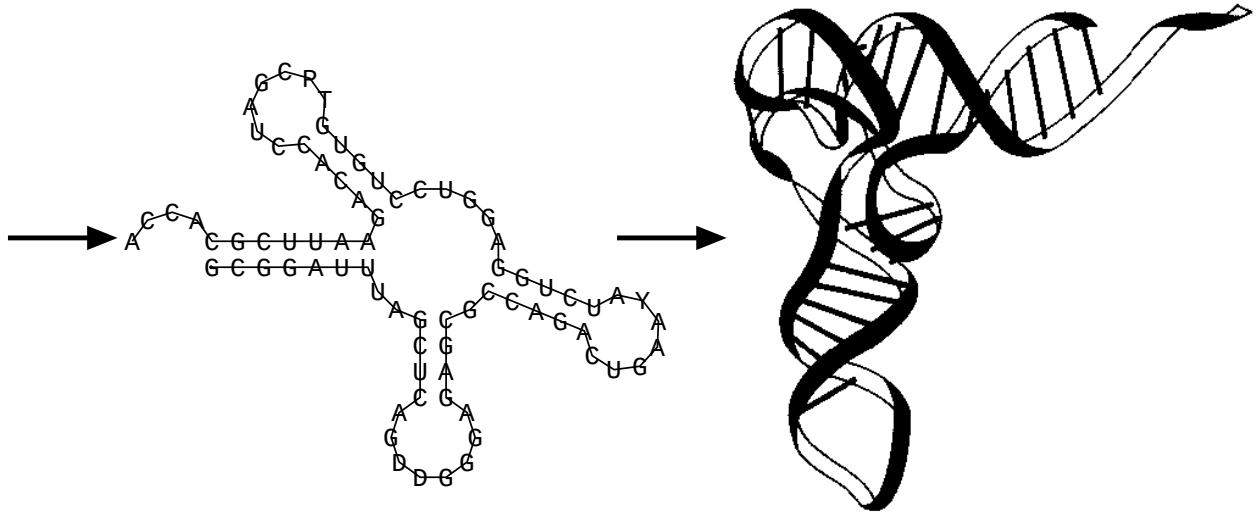
Wien, 2003

Outline

- About (conserved) RNA structures
- Prediction of conserved secondary structures without pseudoknots: alidot
  - Application to picornavirus genomes
- Prediction of consensus secondary structures with pseudoknots: hxmatch
  - Application to functional important RNAs in higher organisms

What are RNA Secondary Structures?

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA



A secondary structure is a list of base pairs that fulfills two constraints:

- A base may participate in at most one base pair.
- Base pairs must not cross, i.e., no two pairs (i, j) and (k, l) may have $i < k < j < l$. (no pseudo-knots)

Secondary structure forms before tertiary interactions.

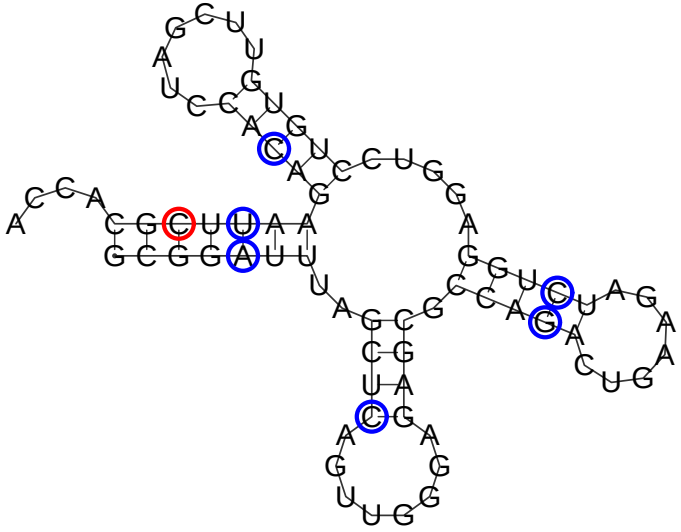
Why look for Conserved Secondary Structures?

- Better accuracy
  - Thermodynamic structure prediction is inaccurate
  - High accuracy structures infered from sequence comparison (requires many sequences)

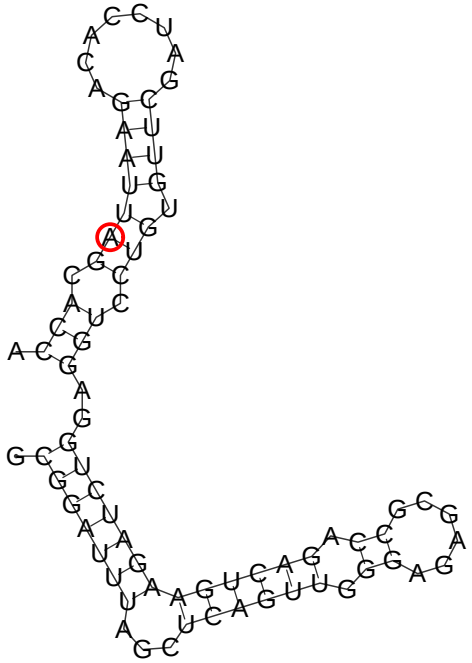
  $\Rightarrow$ combine to get the best of both worlds

- Search for *functional* structures
  - Almost all RNA molecules form secondary structure
  - Functional structures look just like any other
  - Functional structures are well conserved in evolution

The Effect of Mutations



Consistent and compensatory mutations often conserve the structure



A single mutation (red) can radically change the structure

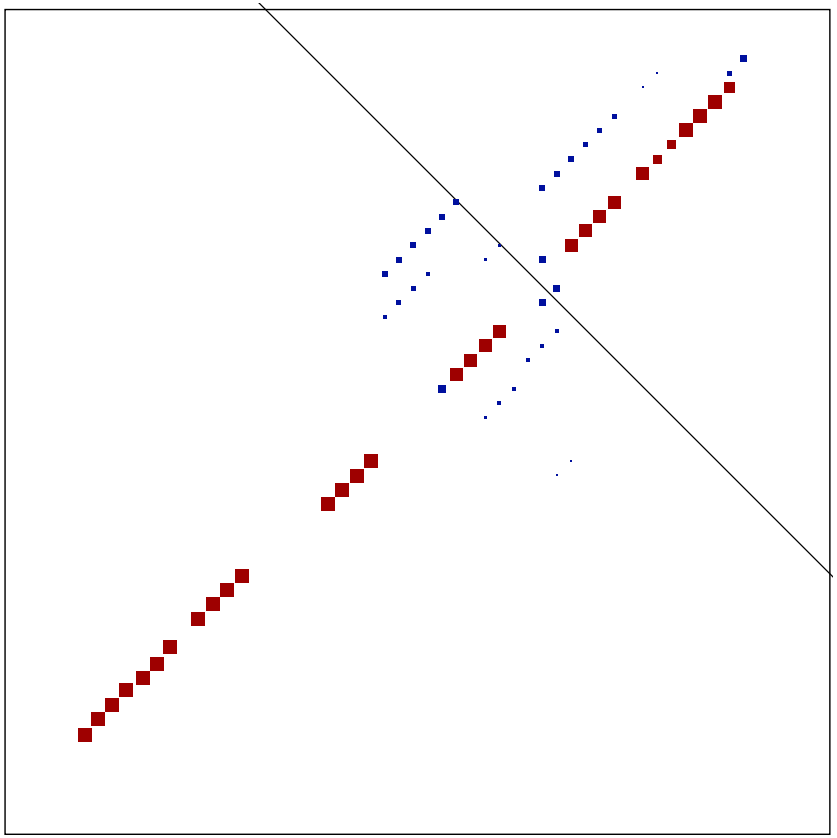
What does this tell us about conserved structures?

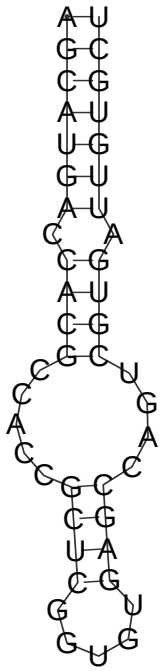
- 10% random mutations typically lead to unrelated structures.
- Common secondary structure motifs present in a group of sequences with less than 90 % average pairwise identity are most likely the result of stabilizing selection.
- If selection acts to preserve a structural element in spite of sequence variation then it must carry some function.

Thermodynamic structure prediction: RNAfold

Predicting base pairing probability matrix

based on McCaskill's partition function algorithm



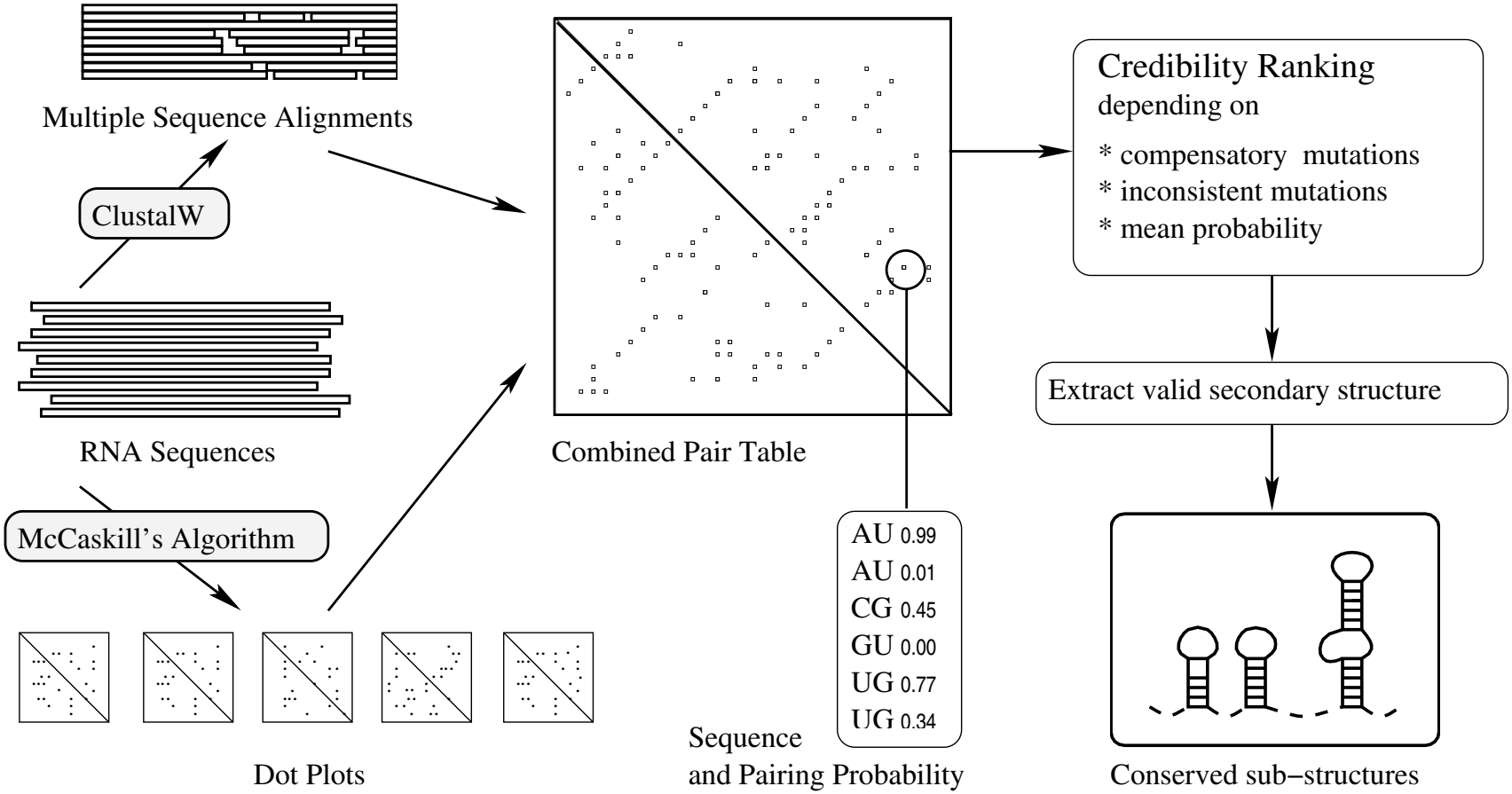


Thermodynamic equilibrium:

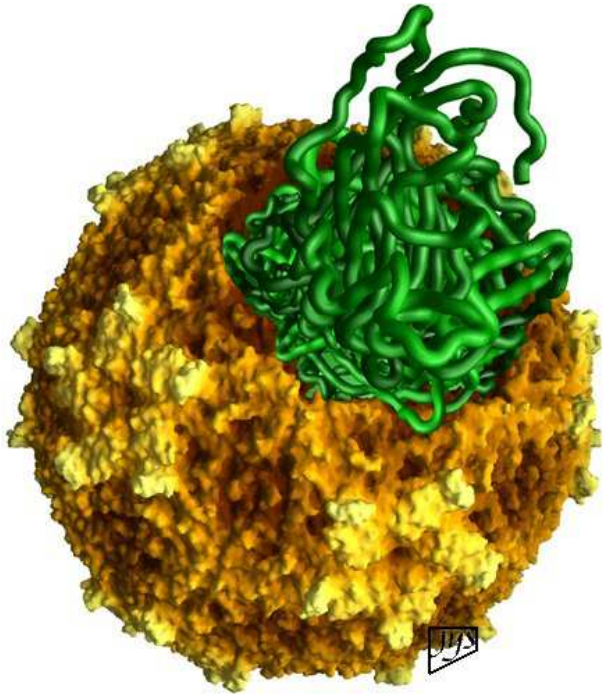
alternative low energy states are

occupied

Alidot


Multiple Sequence Alignments
ClustalW
RNA Sequences
McCaskill's Algorithm
Dot Plots
Combined Pair Table
AU 0.99
AU 0.01
CG 0.45
GU 0.00
UG 0.77
UG 0.34
Sequence
and Pairing Probability
Credibility Ranking
depending on
* compensatory mutations
* inconsistent mutations
* mean probability
Extract valid secondary structure
Conserved sub−structures

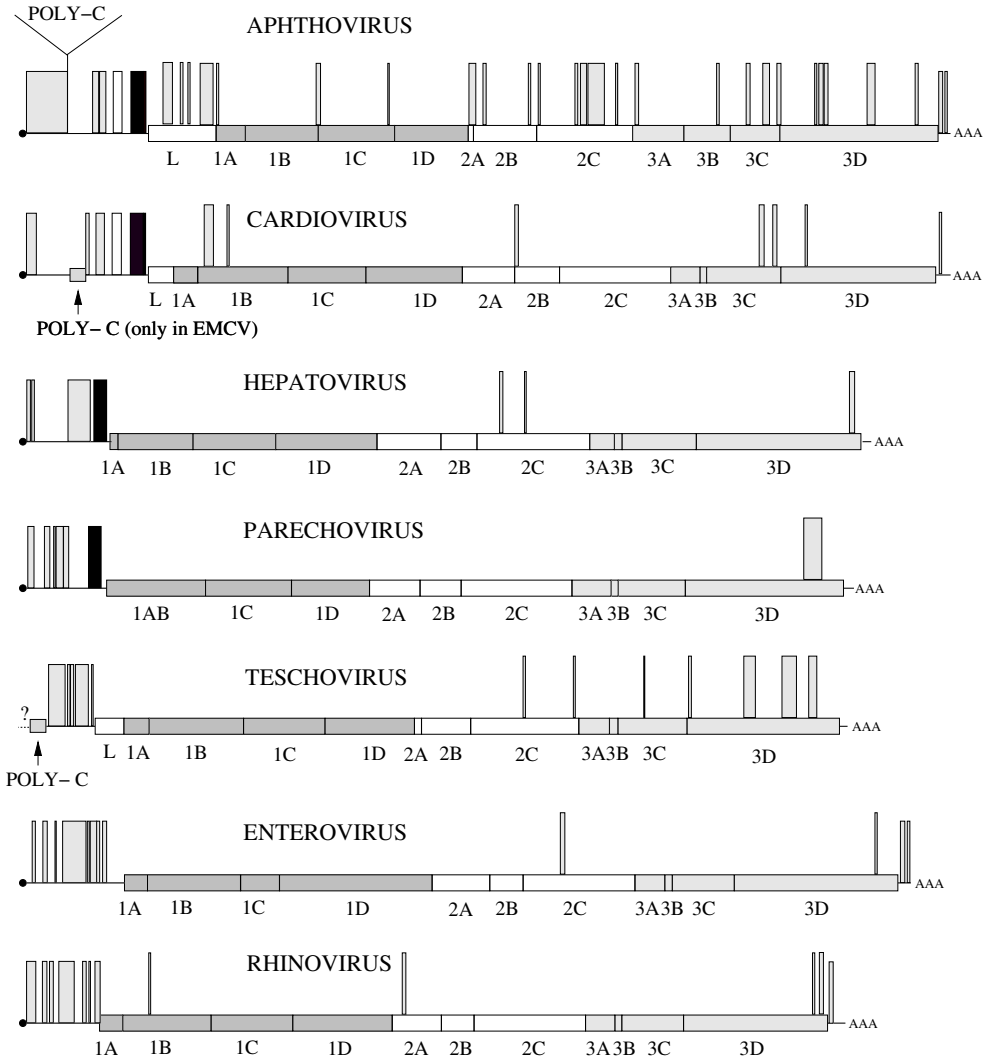
Family Picornaviridae



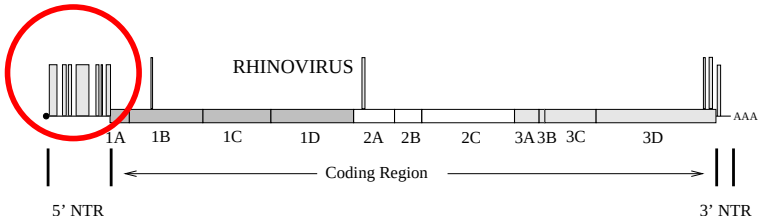
(currently) contains 9 genera:

Genus	Species (example)
Enterovirus	Poliovirus
Rhinovirus	Human Rhinovirus A
Cardiovirus	Encephalomyocarditis virus
Aphthovirus	Foot-and-mouth disease virus
Hepatovirus	Hepatitis A virus
Parechovirus	Human parechovirus
Teschovirus	Porcine teschovirus
Erbovirus	Equine rhinitis B virus
Kobuvirus	Aichi virus

Conserved Structures in Picornaviruses



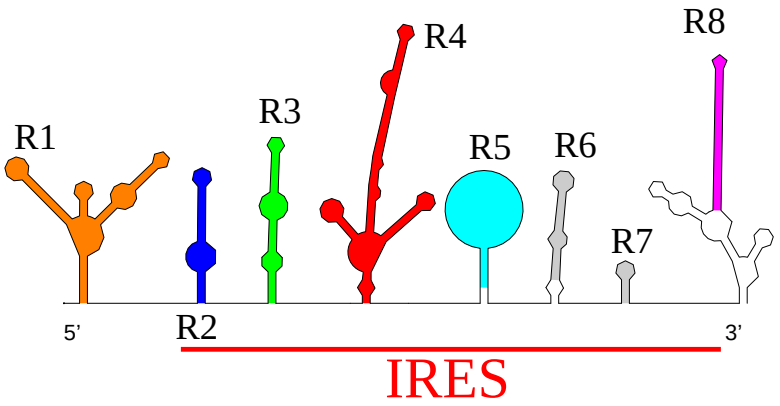
Internal Ribosome Entry Site (IRES)



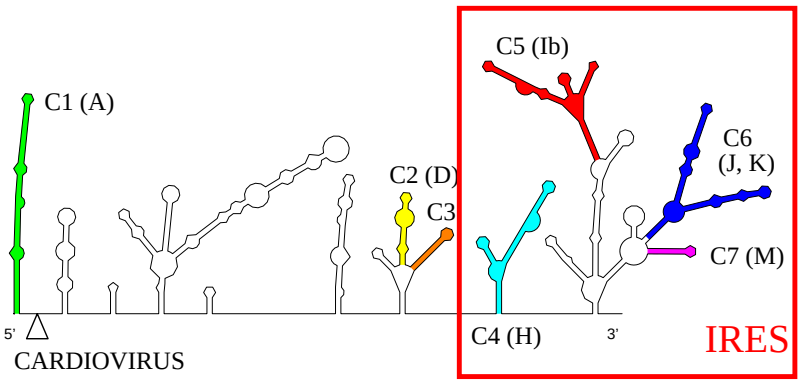
.

IRES	Genera
type I	Enterovirus & Rhinovirus
type II	Aphtho-, Cardio- & Parechovirus
type III	Hepatovirus

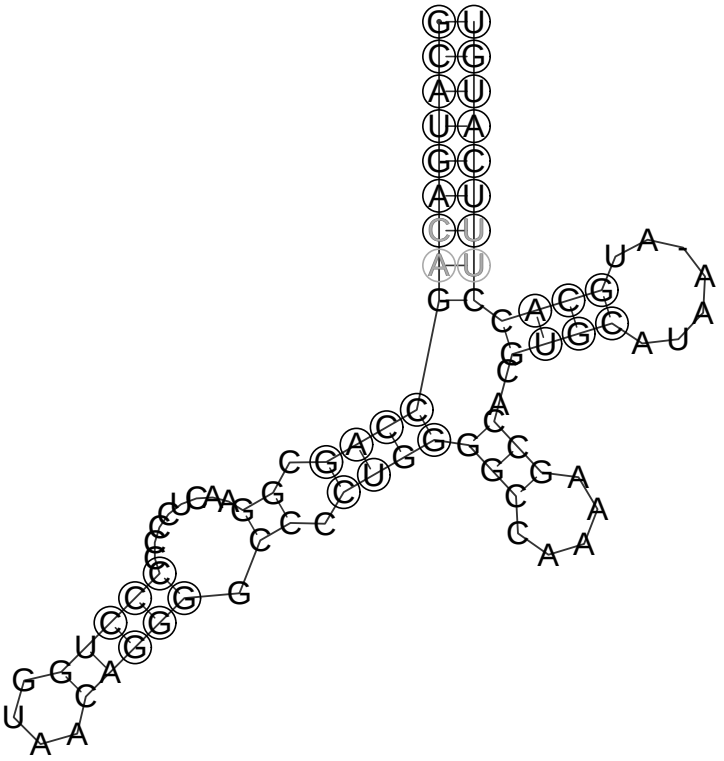
Rhinovirus 5'NTR: type I - IRES



Cardiovirus 5'NTR: type II - IRES



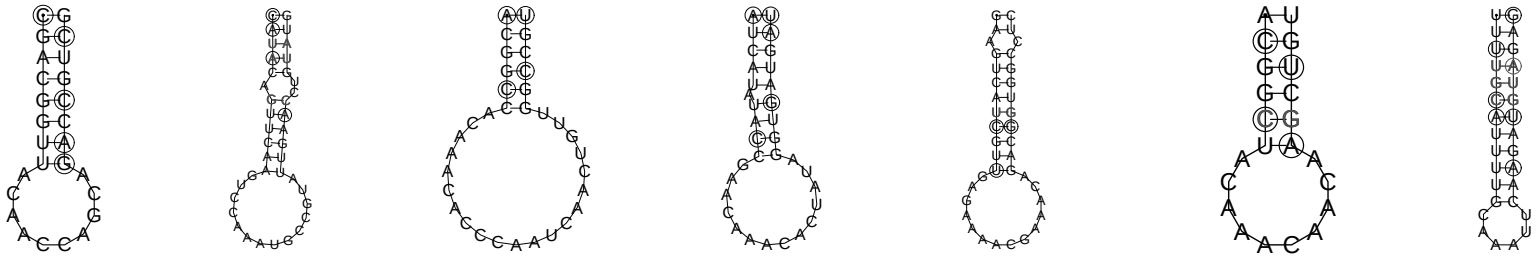
Type II IRES: element 'Ib'



Common structure of aphthovirus, cardiovirus and parechovirus.

Cis-acting-Replication Element (CRE)

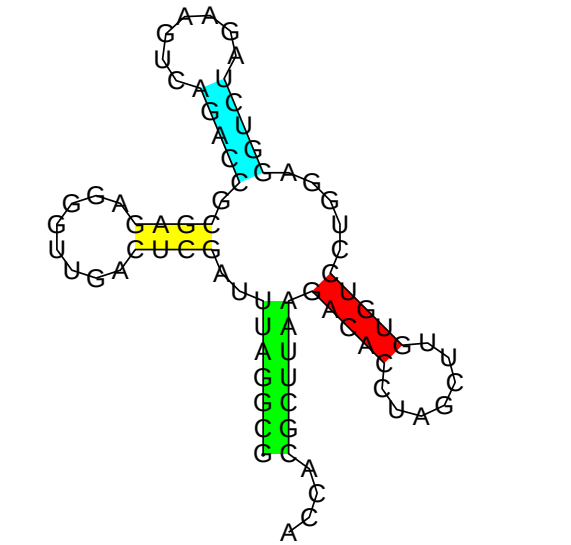
The function of the CRE probably involves the initiation of the synthesis of the negative-sense strand template RNA during virus replication.

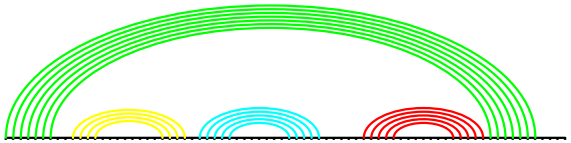


Aphthovirus	Enterovirus	Cardiovirus	HRV-A	HRV-B	Teschov.	Hepatov.
region:2C	2C	1B	2A	1B	2C	2C

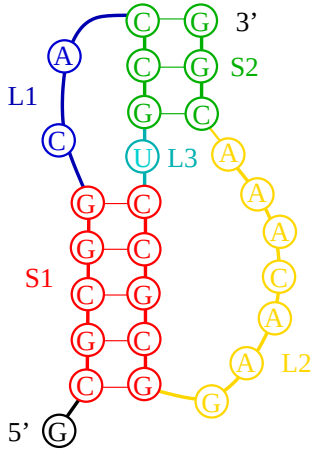
```
Aphto      ~~~~CGAC-GGUU------ACA-CCAAGCA------GACCGUCG~~~~~
Entero     CAUACAGU-UCAAG--------UCCAAAU-GCCGUAUUGAACCUGUAUG
Cardio     ~~~~~ACG-GCCA---CAAACACCCAAUCAACUGU-UGGCCGU~~~~~~
HRV-A      ~~~AUCAUAUACCGAACAAACA---------CUAUAGGUGAUGAU~~~~
HRV-B      GAAGUCAU-CGUUGAGAAAACG---AAACA------GACGGUGGCCUC~
Tescho     ~~~~~~AC-GGCU--ACAAACA-----ACA------AGCUGU~~~~~~~
Hepato     UUUUGCAU-UUUG---CAAA--------------UUCAAGAUGUAGAG~
           ~~~(((((-((((.....................)))))))))~~~~
           1.......10........20........30........40.........
```

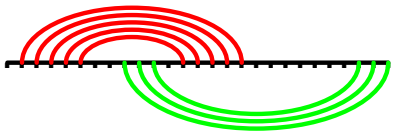
Secondary Structure Including Pseudoknots



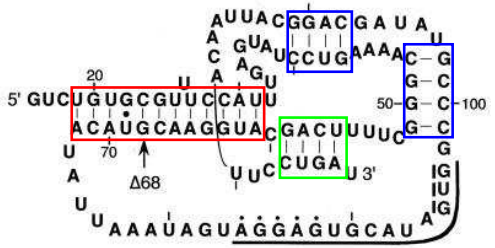


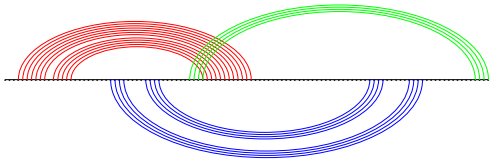
secondary structure





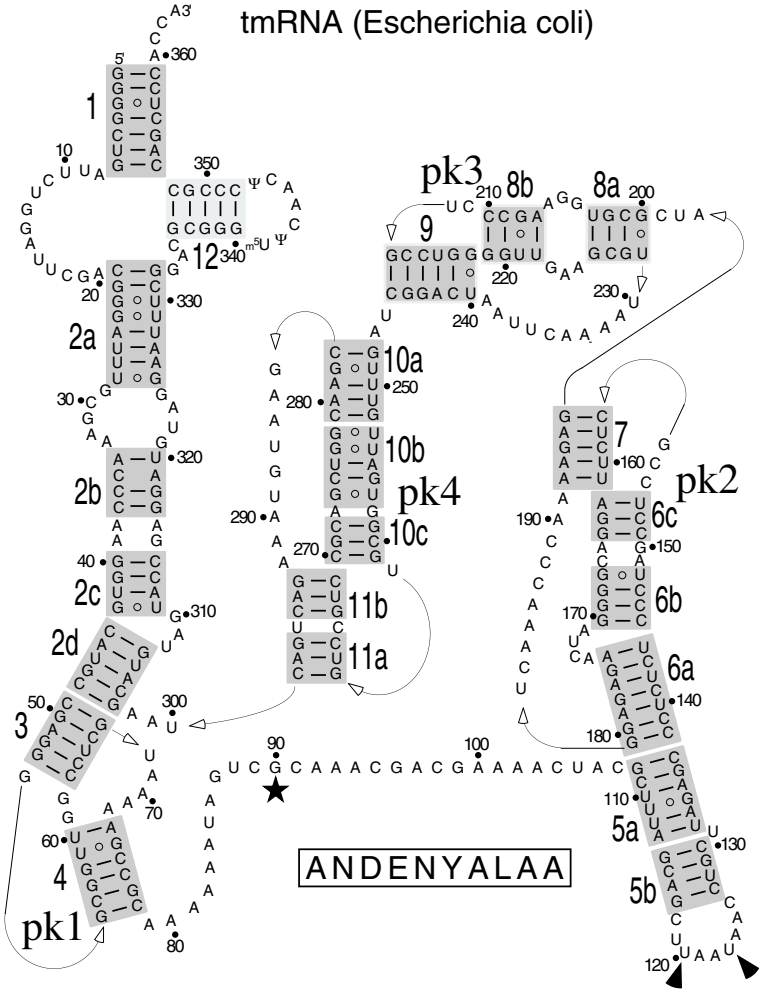
bi-secondary structure

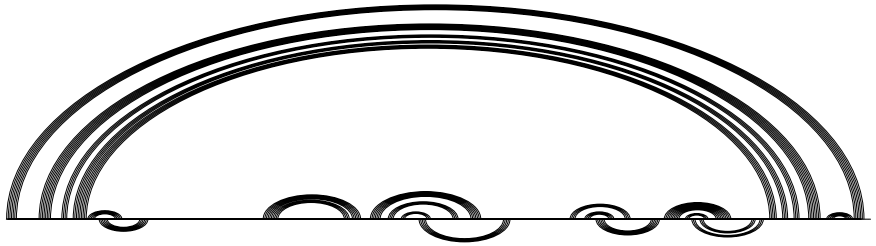




non-planar structure

tmRNA





Zwieb, C. et al, NAR (1999) 27(10):2063

Prediction of Secondary Structure Including Pseudoknots

Thermodynamic prediction of secondary structure without pseudoknots requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory.

Finding the best structure including pseudoknots based on the standard energy model has been proved to be NP-complete.

For a significant class of pseudoknots polynomial time algorithms exist: Rivas and Eddy achieve $\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ memory.

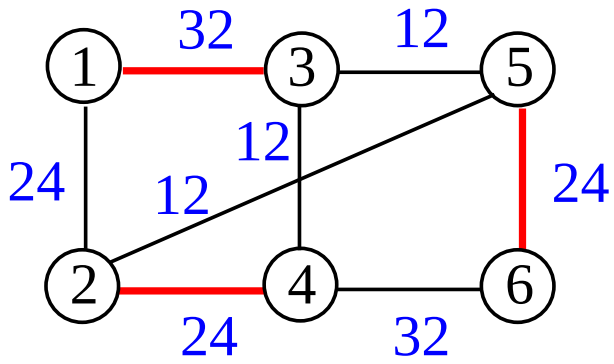
Different approach: pseudoknot prediction based on maximum weighted matching (Tabaska *et al.*, 1998)

Maximum Weighted Matching (MWM)

A *graph G* consists of a set V of vertices and a set E of edges.

$$G = (V, E)$$

A *weighted graph* has a weight w_{ij} assigned to each edge $[v_i, v_j] \in E$.



Matching: A subset $M \subseteq E$ of edges is called a matching, if each vertex is contained in no more than one edge.

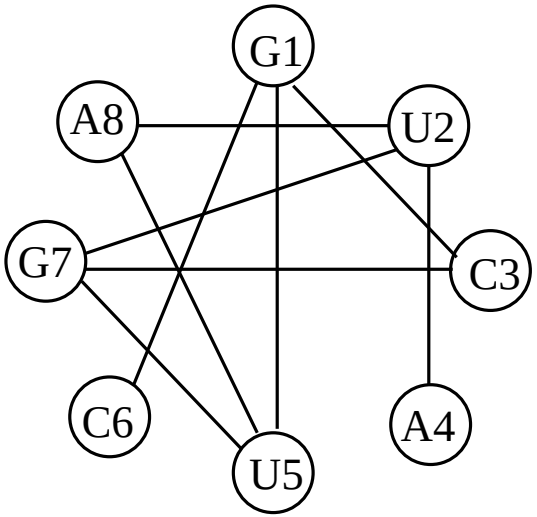
Maximum weighted matching: The sum of the weights of the edges forming the matching is maximal.

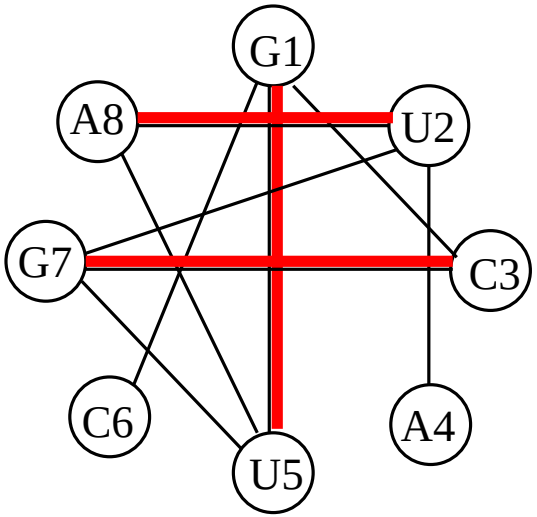
MWM and RNA Structure

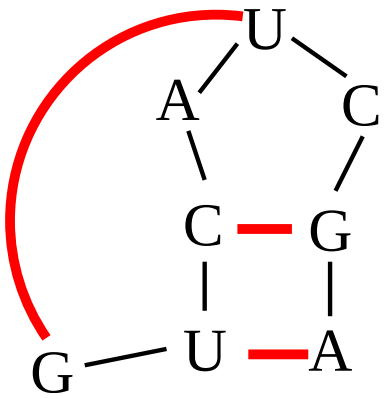
Sequence: GUCAUCGA

Scoring Matrix

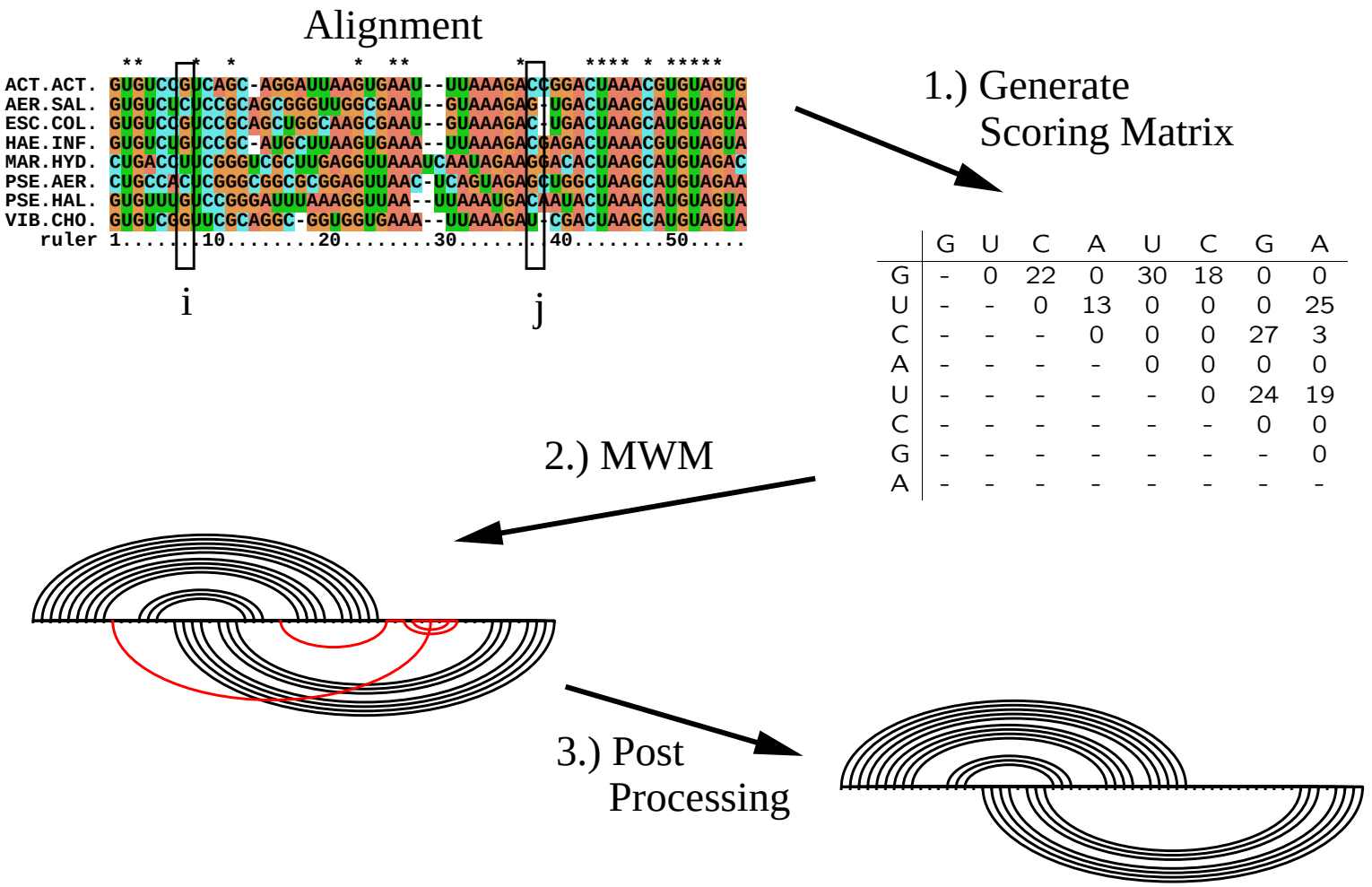
	G	U	C	A	U	C	G	A
G	-	0	8	0	30	18	0	0
U	-	-	0	13	0	0	32	25
C	-	-	-	0	0	0	27	3
A	-	-	-	-	0	0	0	0
U	-	-	-	-	-	0	24	9
C	-	-	-	-	-	-	0	0
G	-	-	-	-	-	-	-	0
A	-	-	-	-	-	-	-	-





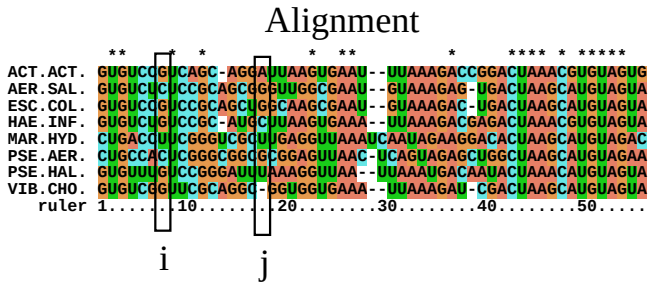


Prediction of RNA Secondary Structures Including Pseudoknots



	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	-	-	-	-	-	-	-	-

Imatch (Tabaska *et al.*, 1998)



Thermodynamic score:

- Positive score for allowed base pairs
- Negative score otherwise

Mutual Information (MI) Score:

$$M_{ij} = \sum_{\mathrm{X,Y}} f_{ij}(\mathrm{XY}) \log \frac{f_{ij}(\mathrm{XY})}{f_i(\mathrm{X}) f_j(\mathrm{Y})}$$

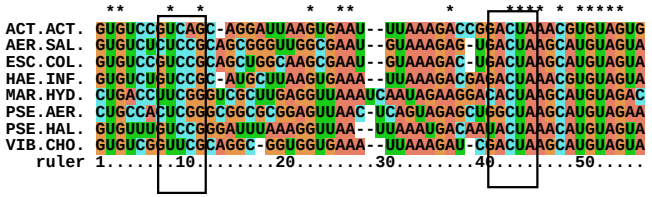
$\pm$ Uses no prior knowledge about secondary structures

- $+$ can detect tertiary contacts and functional constraints
- $-$ poor signal to noise for small data sets
- $-$ only compensatory mutations contribute, consistent mutations (GC $\rightarrow$ GU) are neglected

$\Rightarrow$ not ideal for small number of sequences

Hxmatch

Alignment



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:

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

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).

Hxmatch

Combination of thermodynamic and covariation score:

$$\Pi_{ij} = T_{ij} + \varphi C_{ij}$$

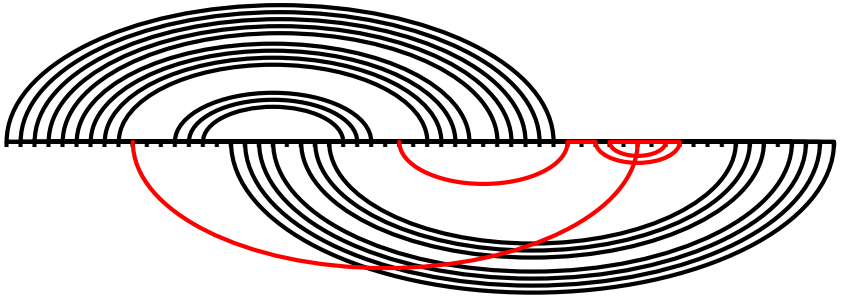
Optional `hxmatch` uses a threshold Π^*, then

$$\Pi_{ij} = \begin{cases} 0 & \text{if } \Pi_{ij} < \Pi^* \\ \Pi_{ij} & \text{if } \Pi_{ij} \geq \Pi^* \end{cases}$$

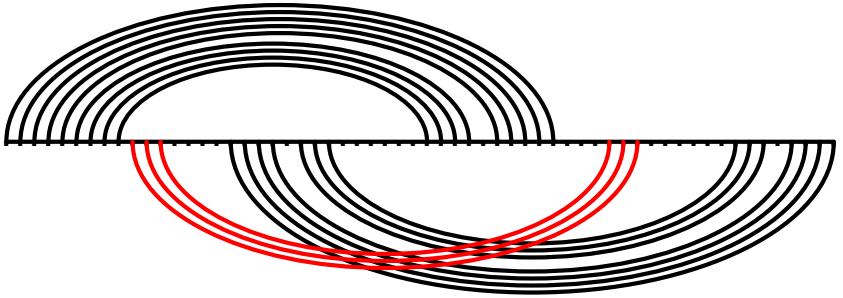
The input graph for the MWM algorithm is given by the scoring matrix Π_{ij}.

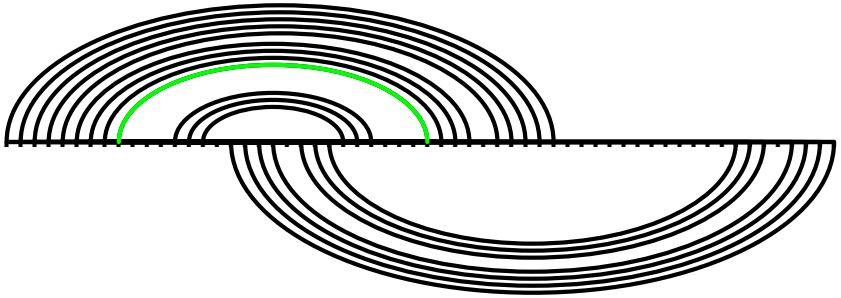
Postprocessing the Outcome of MWM

- remove helices with length < 3



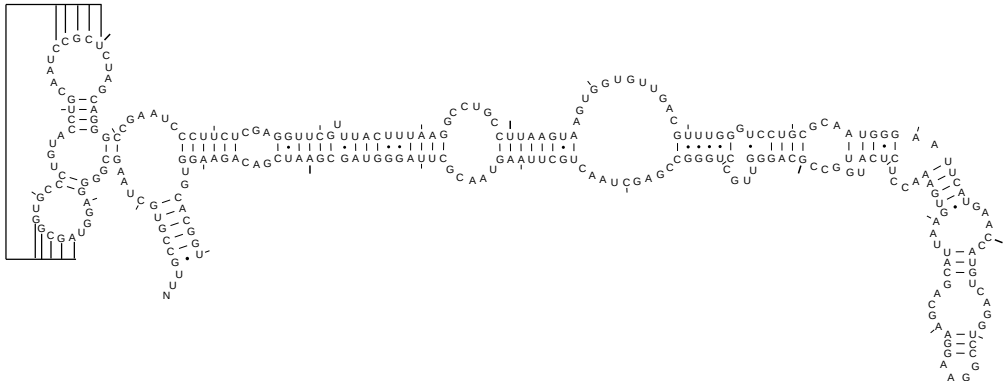
- remove helices not consistent with a bi-secondary structure



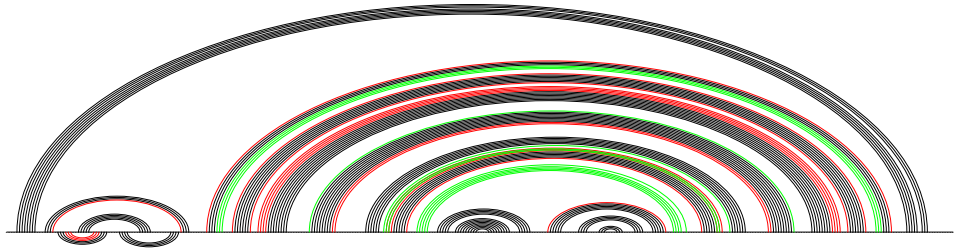


- extend existing helices

Results: SRP RNA



alignment of 6 SRP RNA sequences



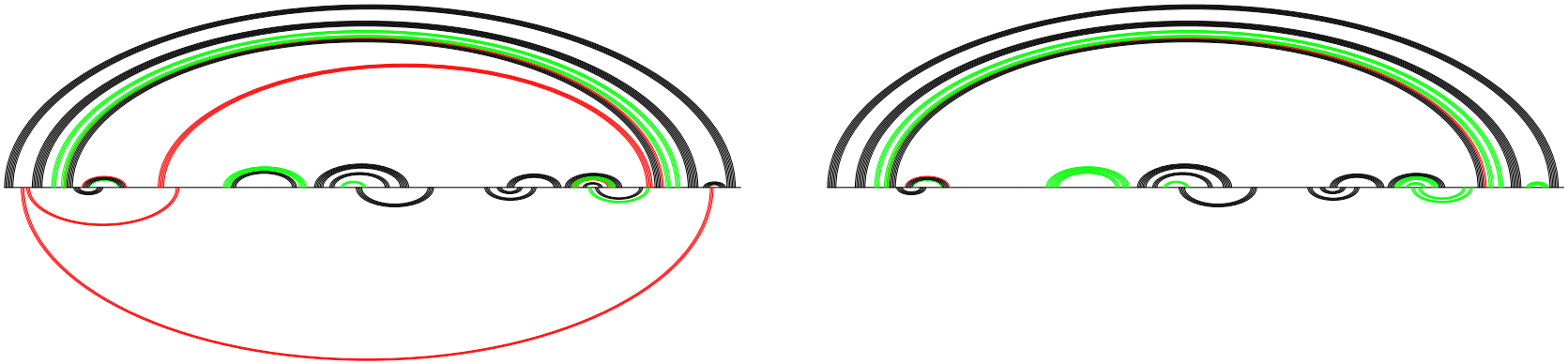
no threshold used

88 % base pairs correctly identified

Structure compared to the phylogenetically derived structure of *Halobacterium halobium* SRP RNA

Results: tmRNA

Alignment of 8 tmRNA sequences



no threshold used

76 % base pairs correctly identified

threshold applied

61 % base pairs correctly identified

Structure compared to the phylogenetically derived structure of *Escherichia coli* tmRNA.

Results

	N	correct basepairs	correct helices	false positive helices	correct pseudoknots
SRP RNA	6	88 %	17/18	0	2/2
RNase P RNA	9	67 %	15/20	2	2/2
tmRNA	8	76 %	16/19	3	4/4

Comparison to Other Methods

Alignment of 8 tmRNA sequences				
Method	correct basepairs	correct helices	false positive helices	correct pseudoknots
`hxmatch`	76 %	16/19	3	4/4
`imatch` (Tabaska *et al.*, 1998)	61 %	12/19	2	3/4
`x2s` (Juan & Wilson, 1999)	-	12/19	5	1/4

Conclusion

- Structural elements which have a biochemical function are conserved by evolution
- In contrast to phylogenetic methods only few sequences are necessary for the prediction
- Database of conserved structure elements of viruses available at `http://www.tbi.univie.ac.at`. Currently contains Picornaviridae, Hepadnaviridae, and Flaviviridae.
- Application to mRNAs and non-coding RNAs in higher organisms
- Knowledge of secondary structure elements can be used to guide deletion/mutation studies

Thanks

- Peter Stadler (Bioinformatik U. Leipzig)
- Ivo Hofacker (TBI)
- Christoph Flamm (TBI)
- Christian Mandl (Virologie U. Wien)
- Caroline Thurner, Roman Stocsits,
  Martin Fekete, Susanne Rauscher (TBI)