# The Tedious Task of Finding Common RNA Sequence Structure Properties AND Lattice Models and Energy Landscape

Rolf Backofen

Lehrstuhl für Bioinformatik
Albert-Ludwigs-Universität Freiburg

13. April 2011

Economist.com

RNA

# The rise and rise of RNA

Jun 14th 2007

From *The Economist* print edition

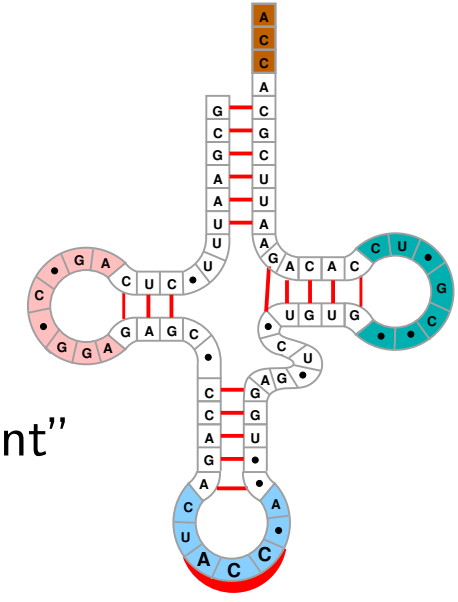
IT IS beginning to dawn on biologists that they may have got it wrong. Not completely wrong, but wrong enough to be embarrassing. For half a century their subject had been built around the relation between two sorts of chemical. Proteins, in the form of enzymes, hormones and so on, made things happen. DNA, in the form of genes, contained the instructions for making proteins. Other molecules were involved, of course. Sugars and fats were abundant (too abundant, in some people). And various vitamins and minerals made an appearance, as well. Oh, and there was also a curious chemical called RNA, which looked a bit like DNA but wasn't. It obediently carried genetic information from DNA in the nucleus to the places in the cell where proteins are made, rounded up the amino-acid units out of which those proteins are constructed, and was found in the protein factories themselves....

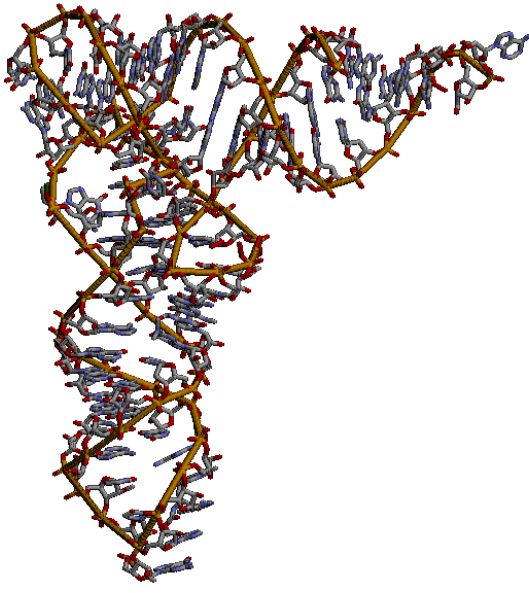
# Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models

# RNA-Molecules

- **RNA**:
  - bonds = secondary structure
  - hierarchical folding
    secondary structure first

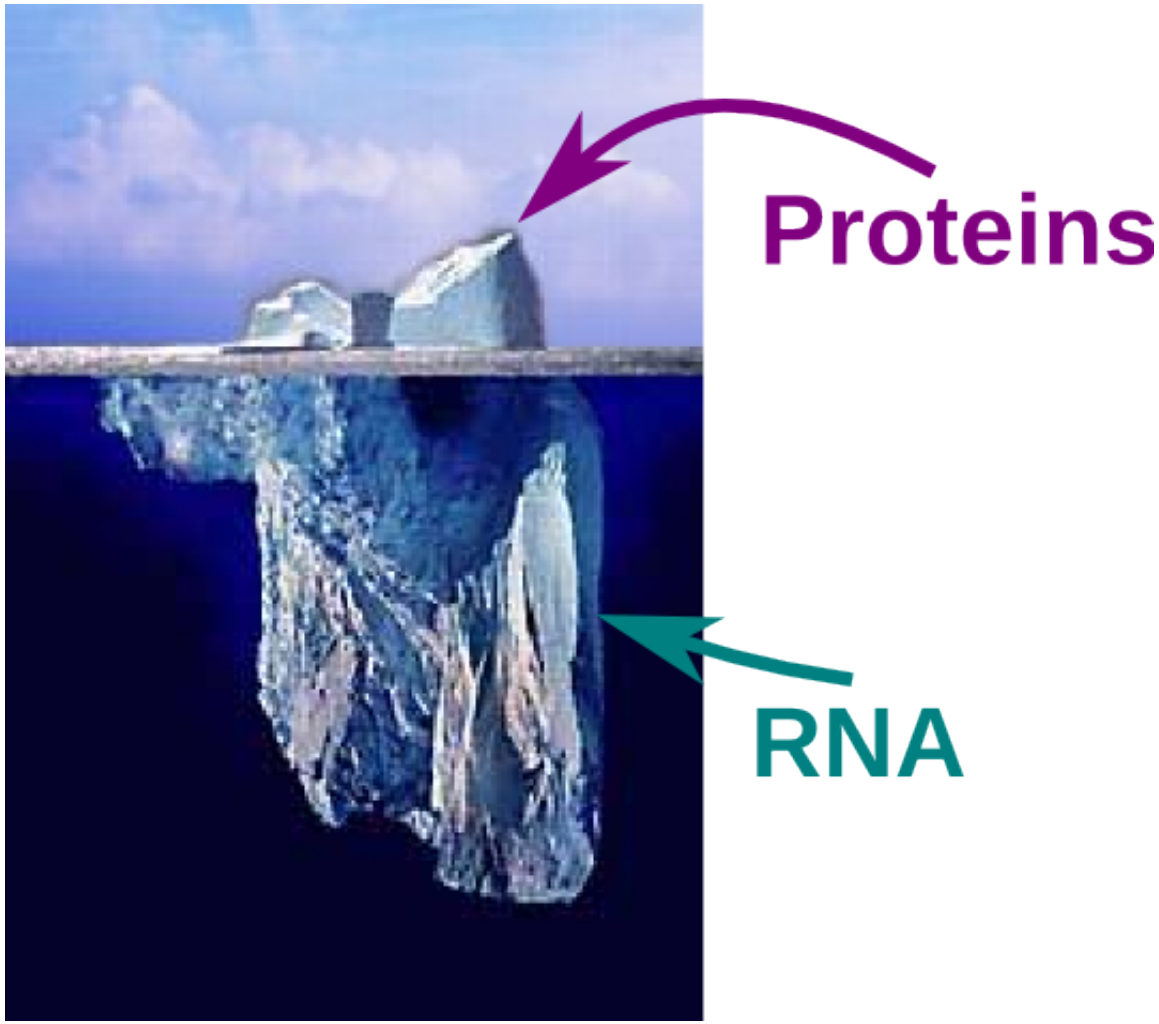




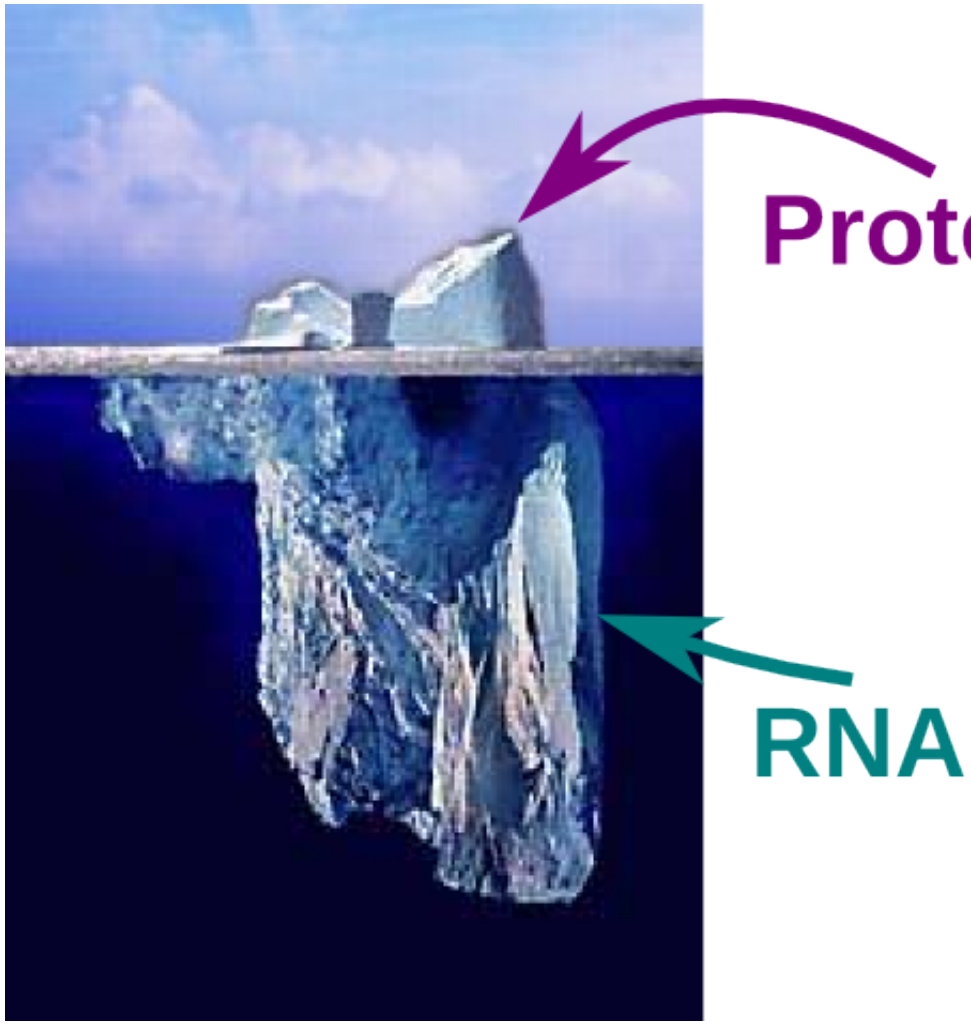
- properties
  - before: simple "transport element"
    DNA $\longrightarrow$ **RNA** $\longrightarrow$ protein
  - now: many functions
    - ribozyme: RNA-enzymes
    - non-coding RNAs, regulation etc.
    - RNA: scientific breakthrough 2002
    - Nobel prize for medicine 2006: RNAi

# How many possible ncRNA out there? (Cont)

# How many possible ncRNA out there? (Cont)



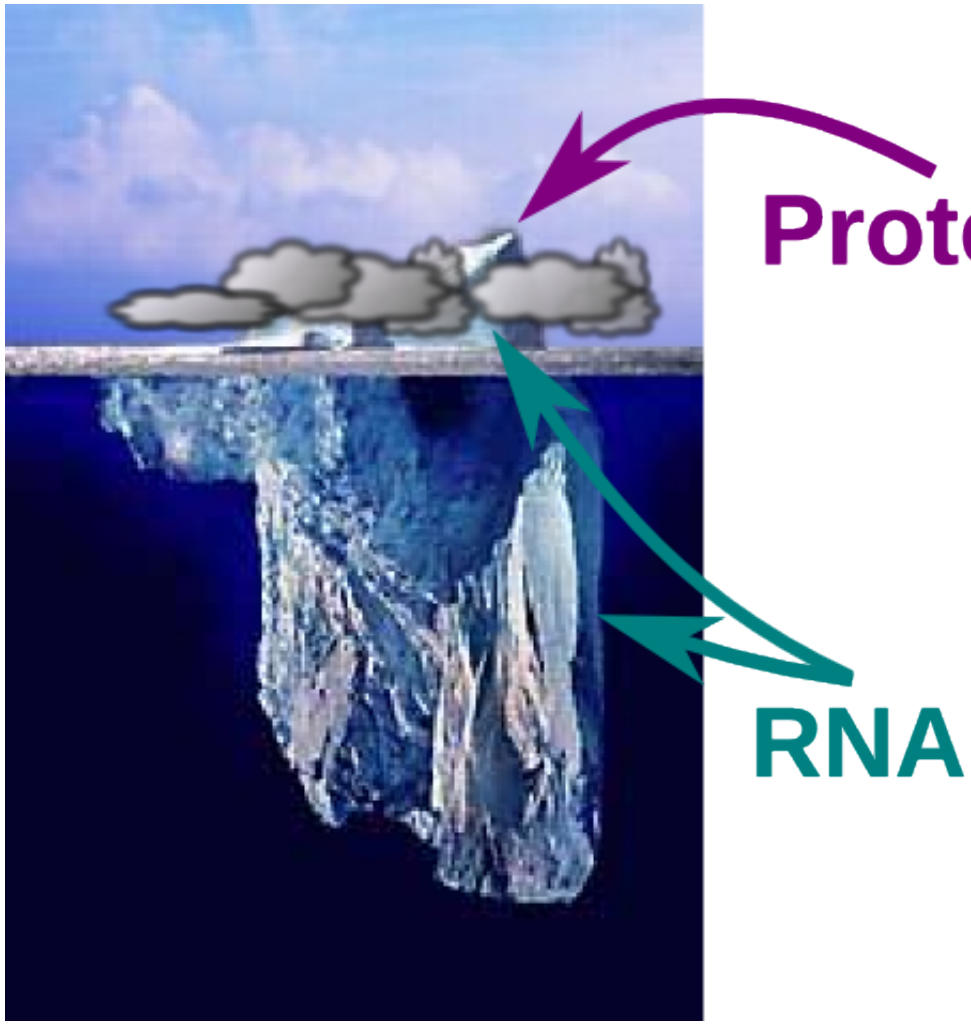
# How many possible ncRNA out there? (Cont)




In humans:

- approx. **1%** of genome encodes protein
- approx. 80-90% of genome **is** transcribed $\Rightarrow$ at least **98%** of transcribed RNA is **non-coding**

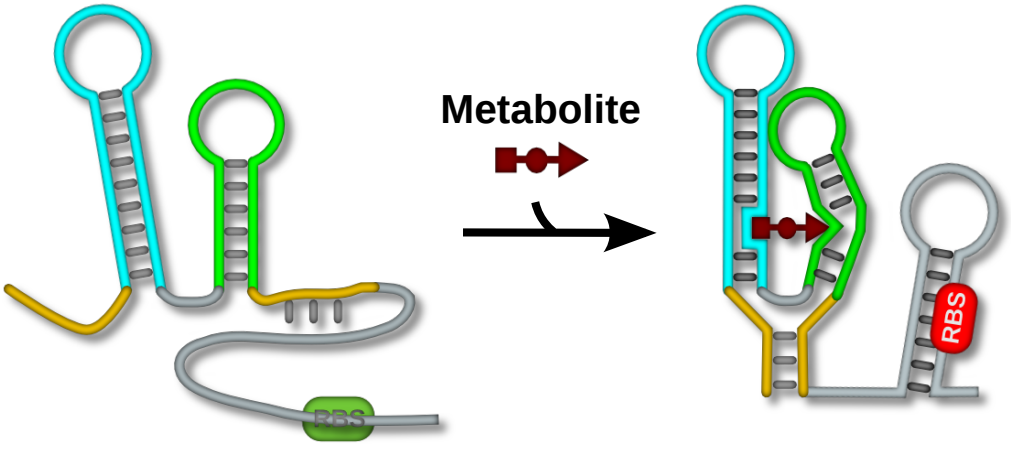
# How many possible ncRNA out there? (Cont)




In humans:

- approx. **1%** of genome encodes protein
- approx. 80-90% of genome **is** transcribed $\Rightarrow$ at least **98%** of transcribed RNA is **non-coding**

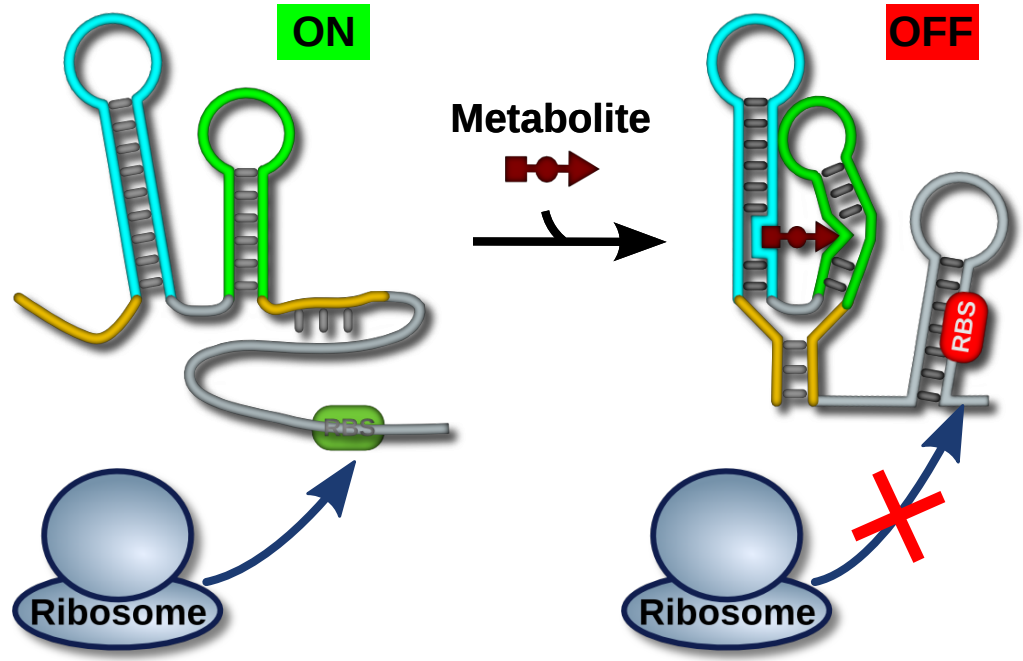
# Riboswitch




[Serganov et al. (mod.)]

- Riboswitches:
  - cis-acting RNA-elements included in the mRNA
  - can detect different metabolites.
  - direct regulation of associated mRNA

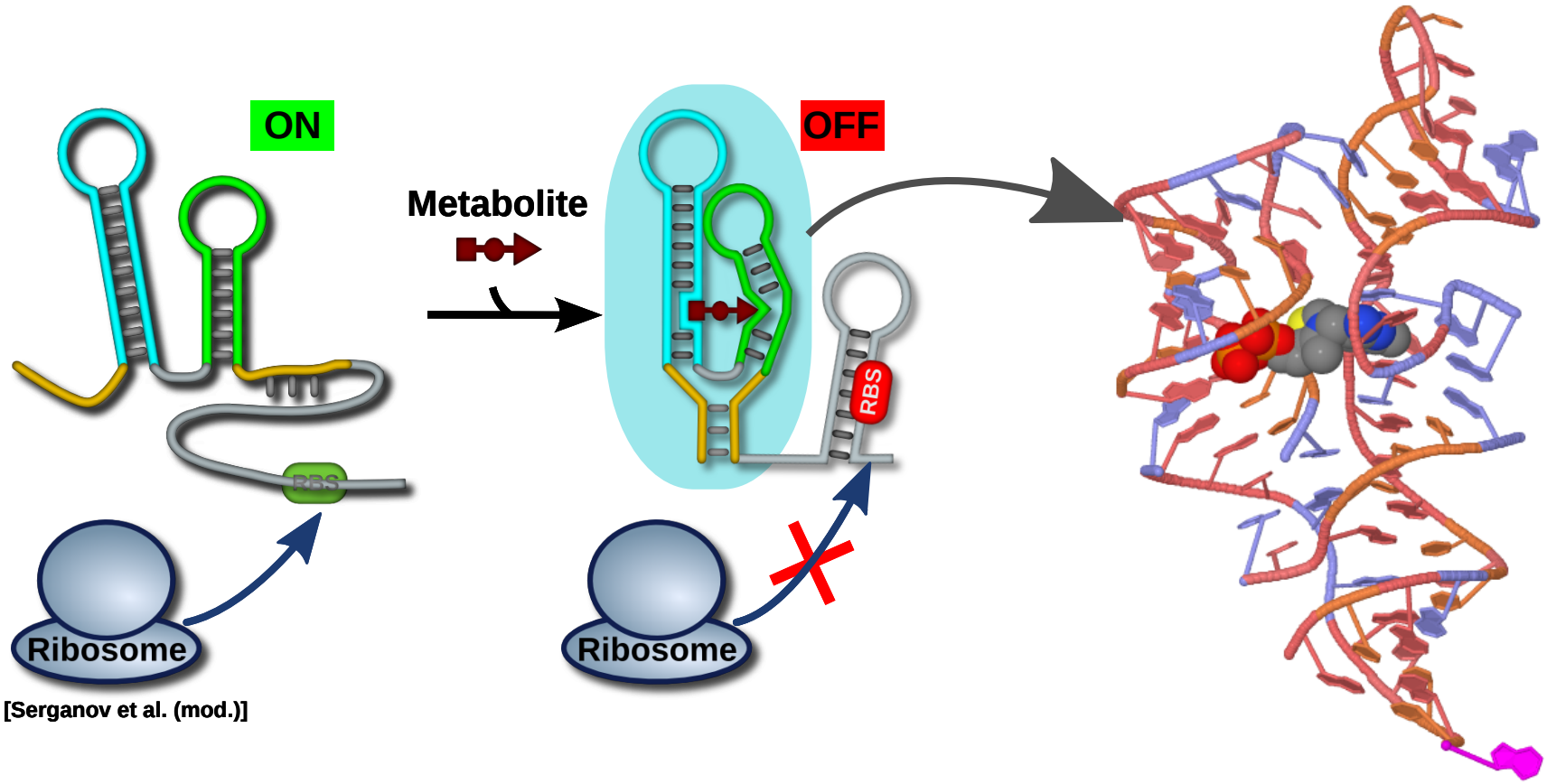
# Riboswitch




[Serganov et al. (mod.)]

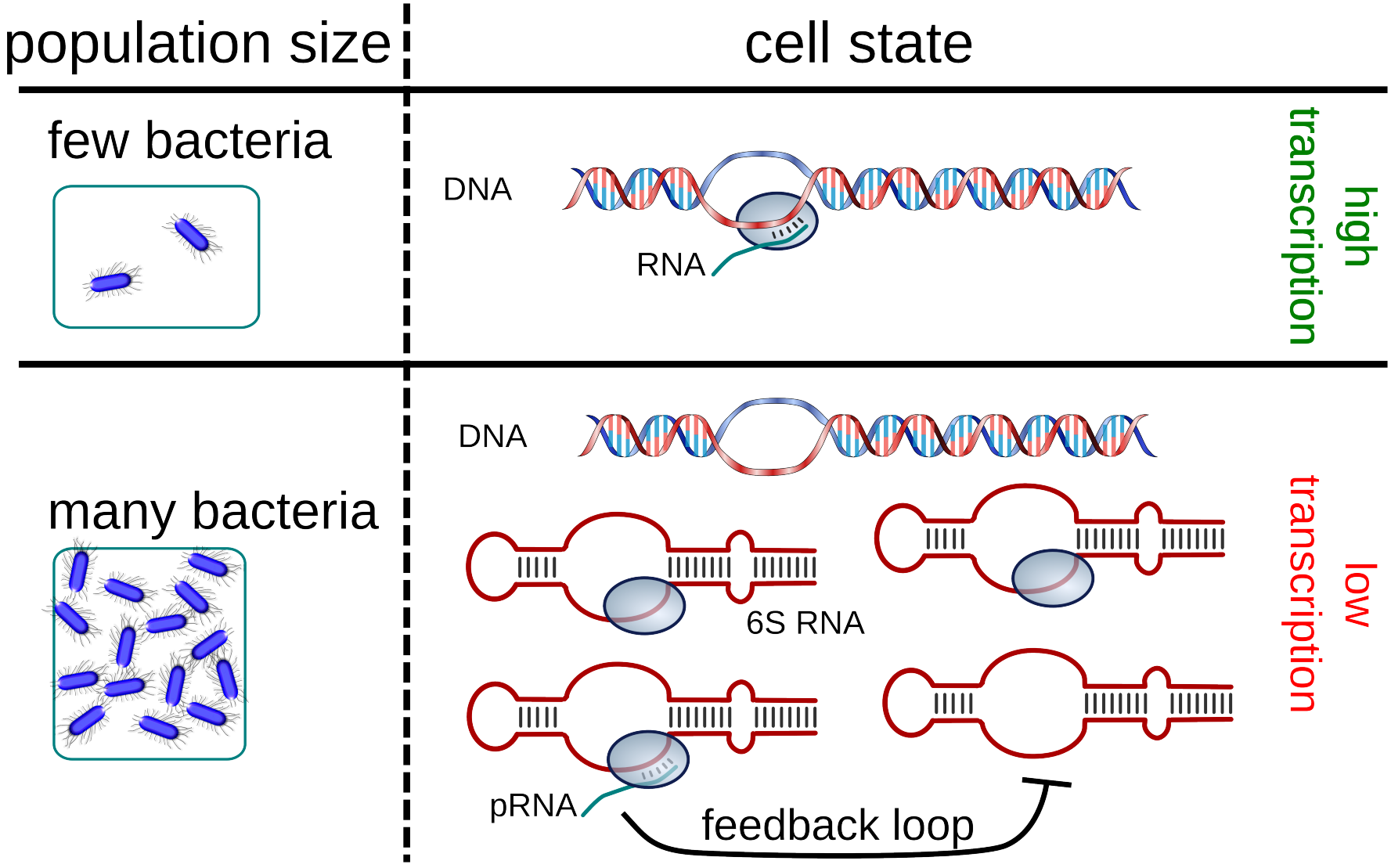
- Riboswitches:
  - cis-acting RNA-elements included in the mRNA
  - can detect different metabolites.
  - direct regulation of associated mRNA

# Riboswitch

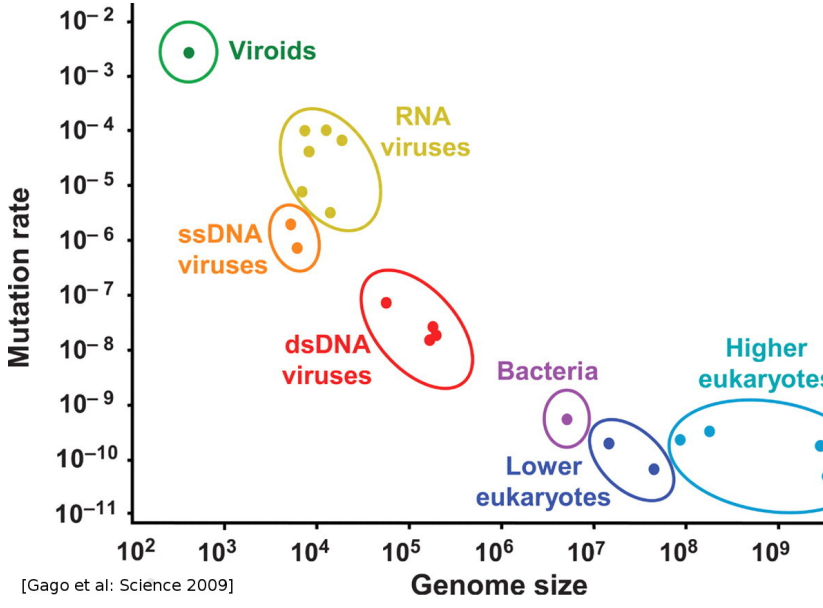



- Riboswitches:
  - cis-acting RNA-elements included in the mRNA
  - can detect different metabolites.
  - direct regulation of associated mRNA

# Examples I: 6S-RNA



# Viroids

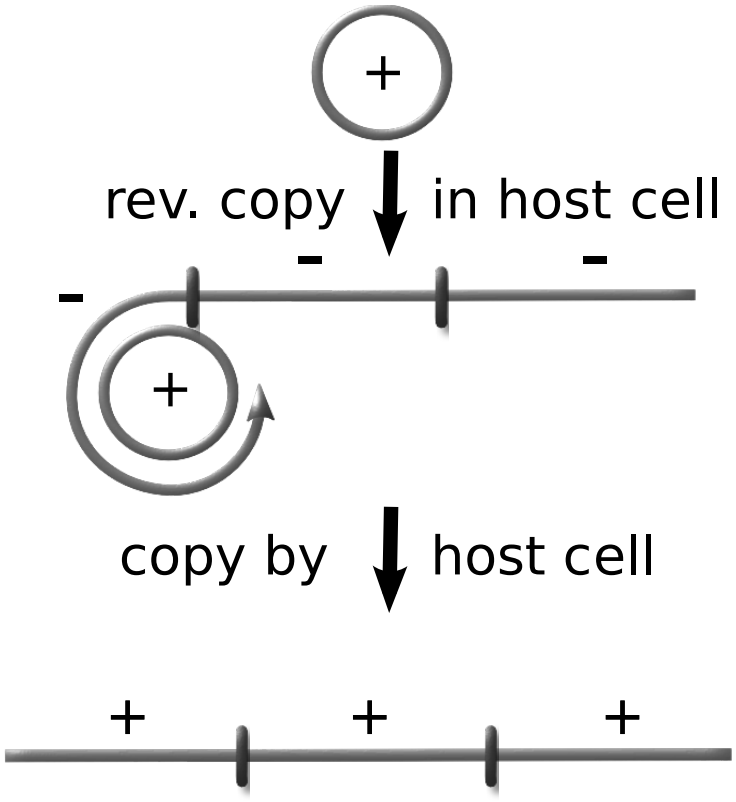




[Gago et al: Science 2009]

- small RNA pathogens infecting plants (240 – 400 nt)
- first identified:

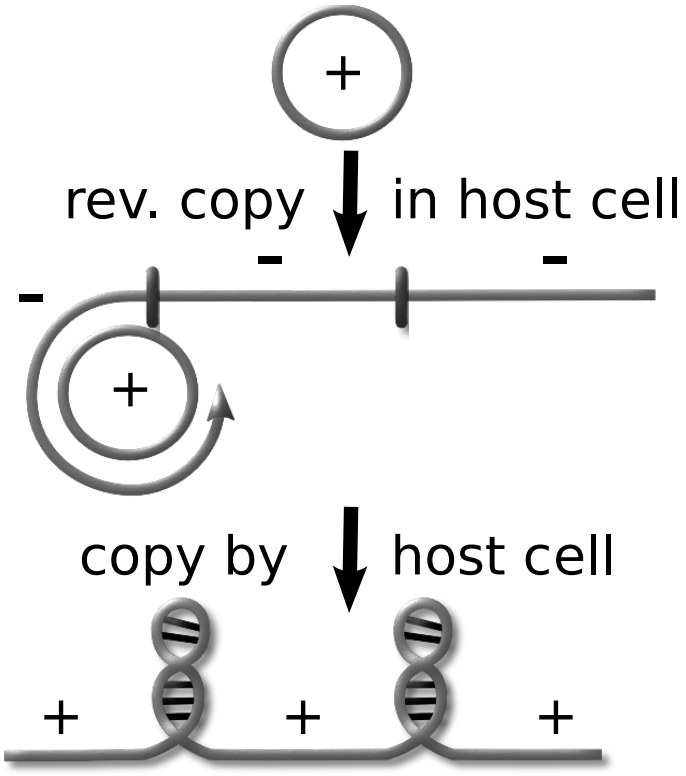
  Potato spindle tuber viroid (PSTVd)
- viroids are pure RNA, no protein, no capsule
- smallest known self-replicating unit

# Viroid Replication with Hammerhead



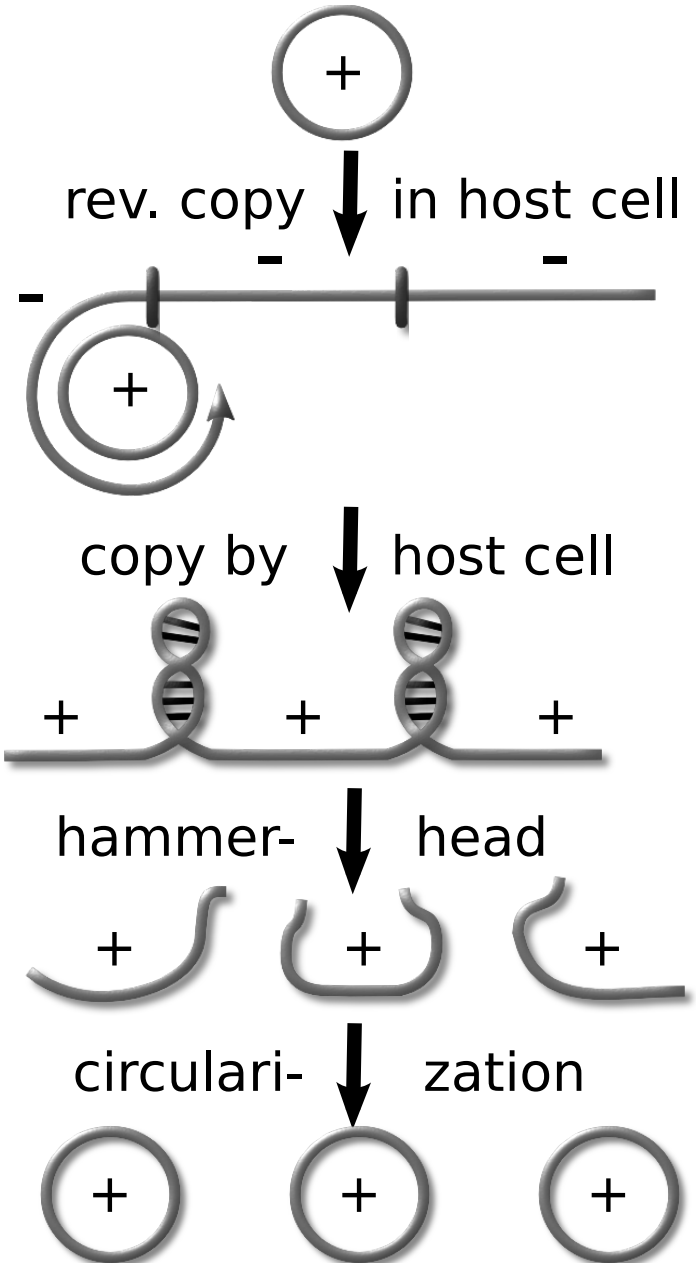

- viroid: circlular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by **self-cleavage through Hammerhead ribozyme**

# Viroid Replication with Hammerhead




- viroid: circlular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by **self-cleavage through Hammerhead ribozyme**

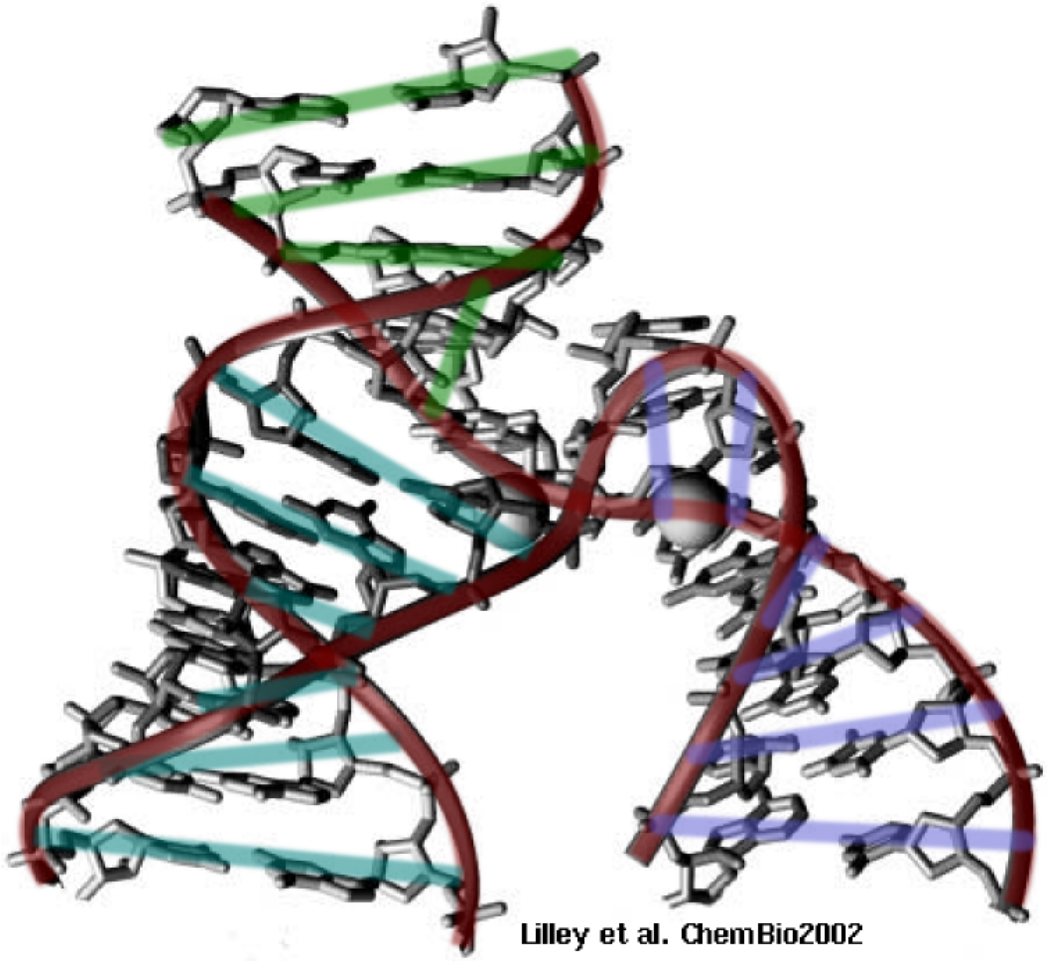
# Viroid Replication with Hammerhead




- viroid: circlular + strand
- host RNA-polymerase generates a longer − strand by going through the circular genome more than once
- new, longer plus strand is then synthesized by the host RNA polymerase
- split into viroid-units by **self-cleavage through Hammerhead ribozyme**

# Example: Ribozymes

- Ribozymes: RNA enzyme
- Hammerhead-Ribozyme:
  - detected as site-specific self-cleavage unit
  - many variants with different specifity generated
  - requires only two metal atom



# RNA Structure and Function

- Function is determined by sequence **and** structure
- Next generation sequencing technologies allow high-throughput data collection of sequence information
- ...but high-throughput structure determination is (still) mostly done algorithmically

IRES SECIS IRE


snRNA miRNA snoRNA tRNA

# Overview

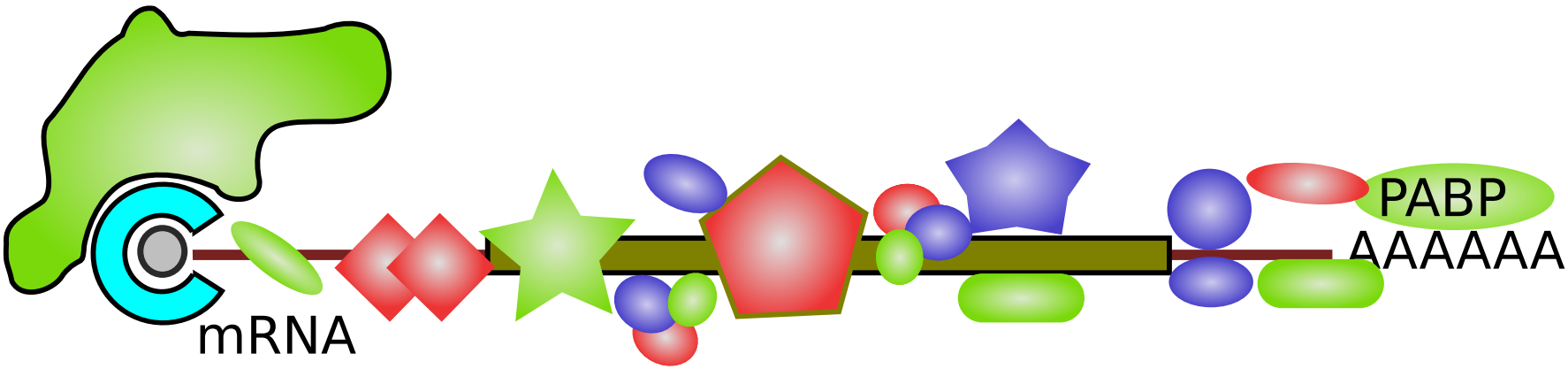
- Motivation

- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes

- barrier trees and HP-lattice models

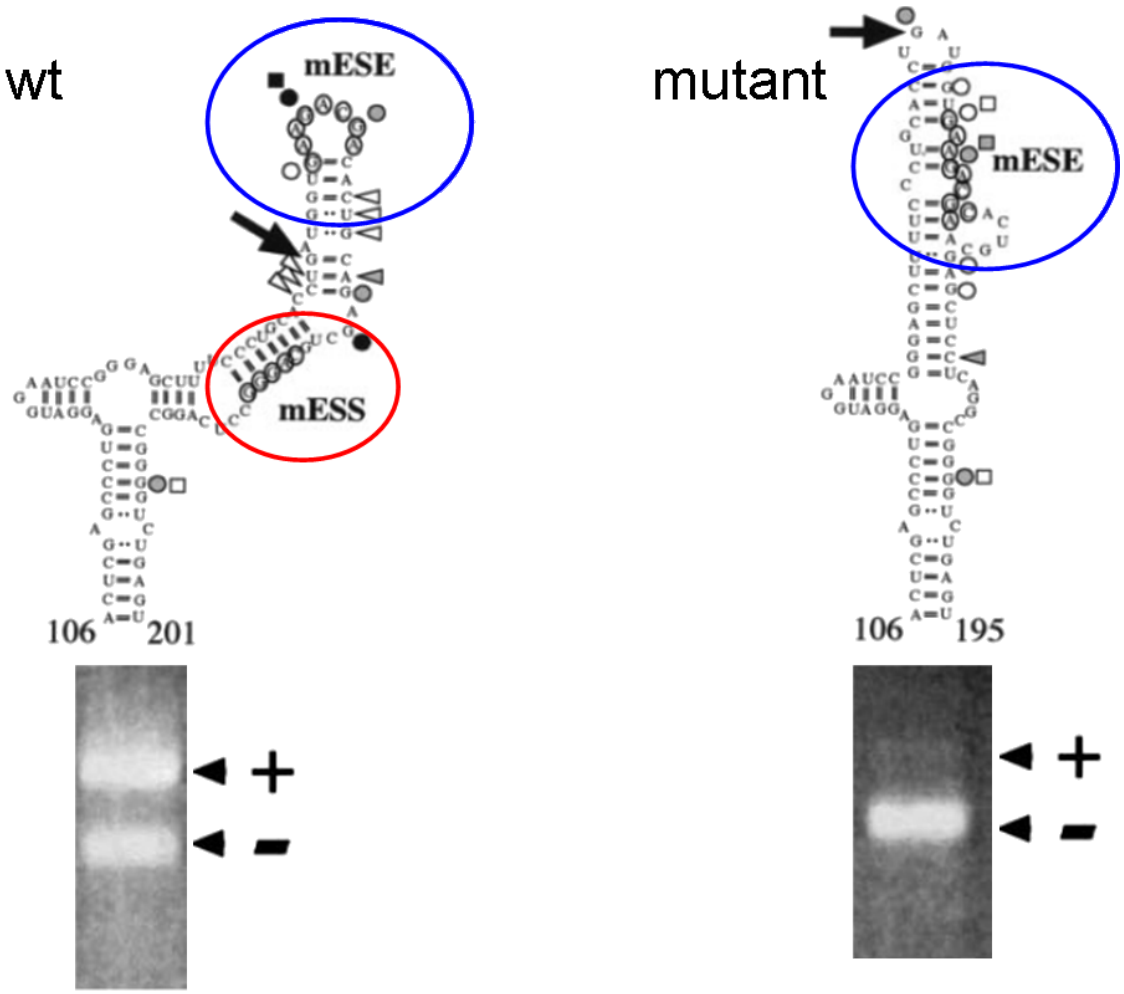
# mRNA is Loaded with Proteuns

# mRNA is Loaded with Proteuns



# Alternative Splicing and Secondary Structure

- one important feature missing: secondary structure
- example: fibronectin EDA exon



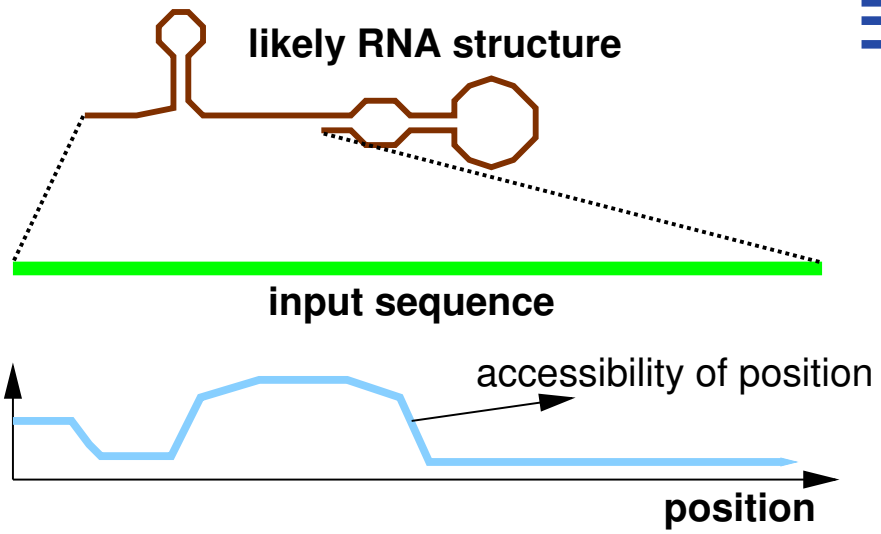

Buratti et al. Mol and Cell Bio. 24(3) 2004

# MEMERIS

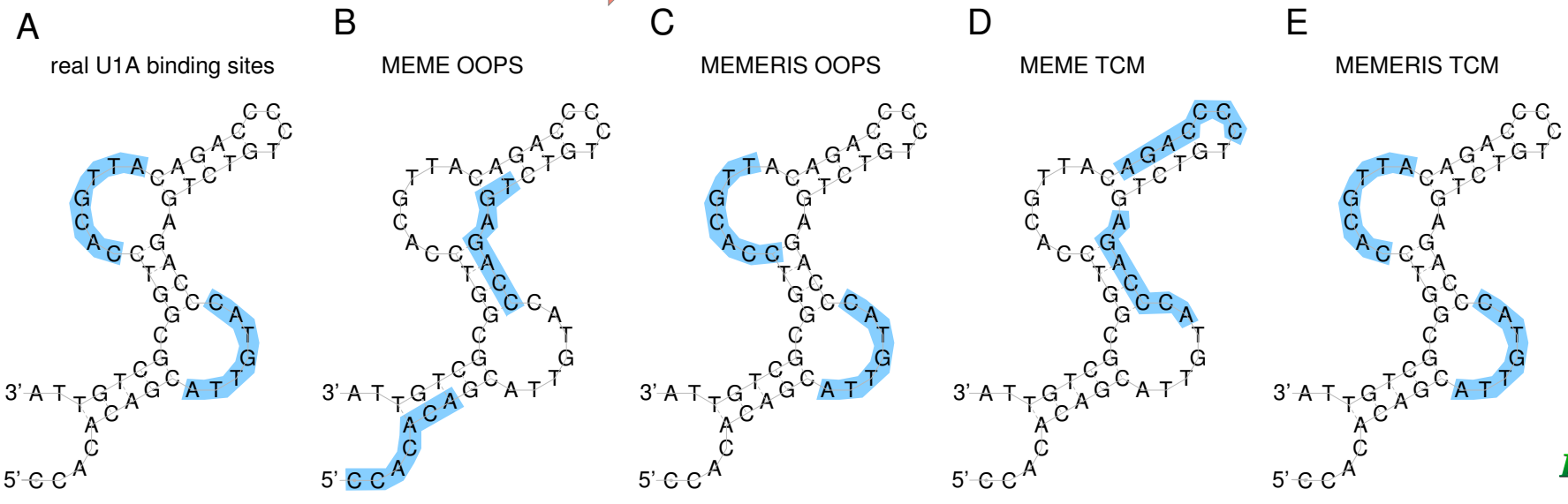
- **often: additional knowledge**

RNA: some splice factors prefer single-stranded sites

TFs: distances to TATA box, structural contexts, ...



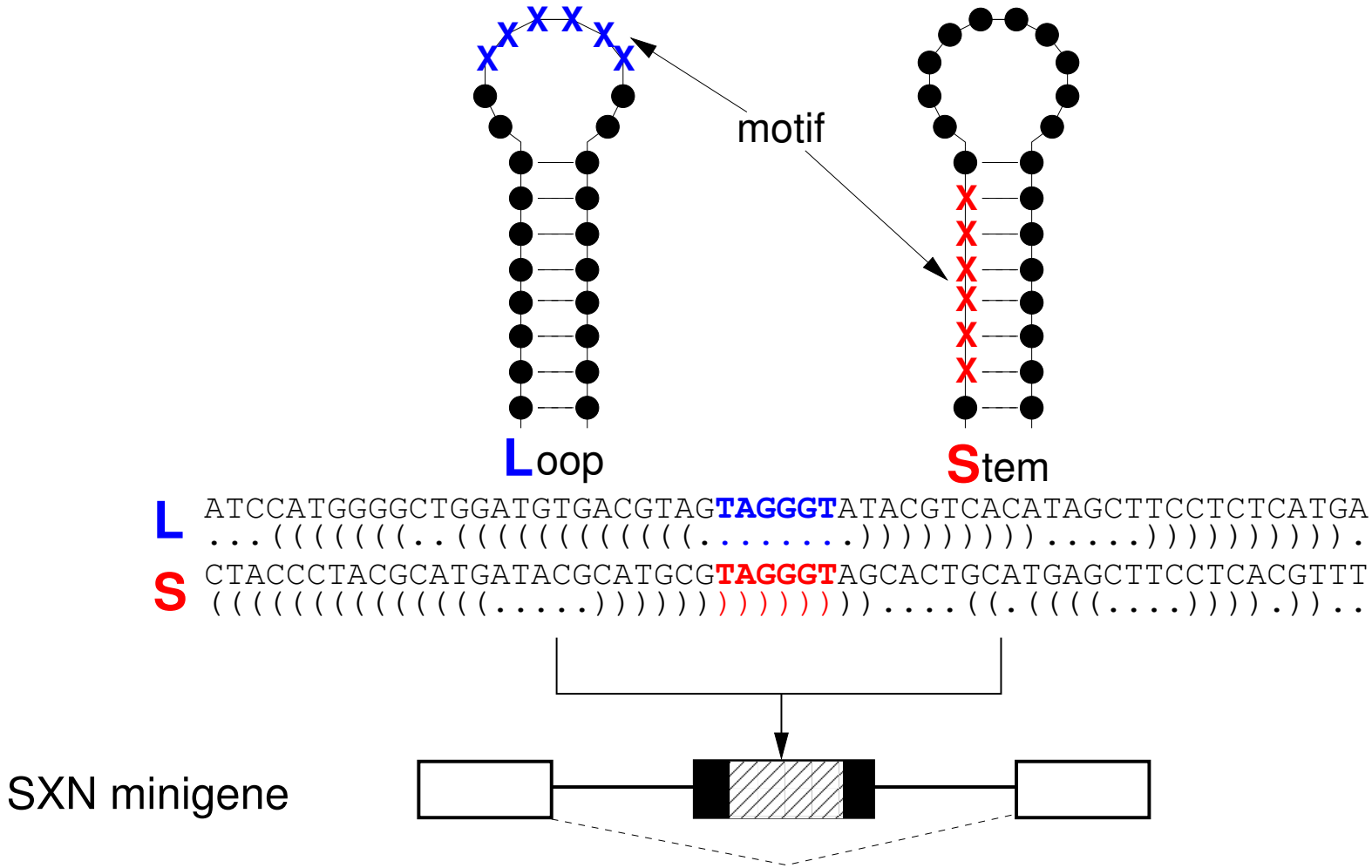

- integration into EM? ⇒ **prior probablities on start positions**




[Hiller et al. NAR 2006]

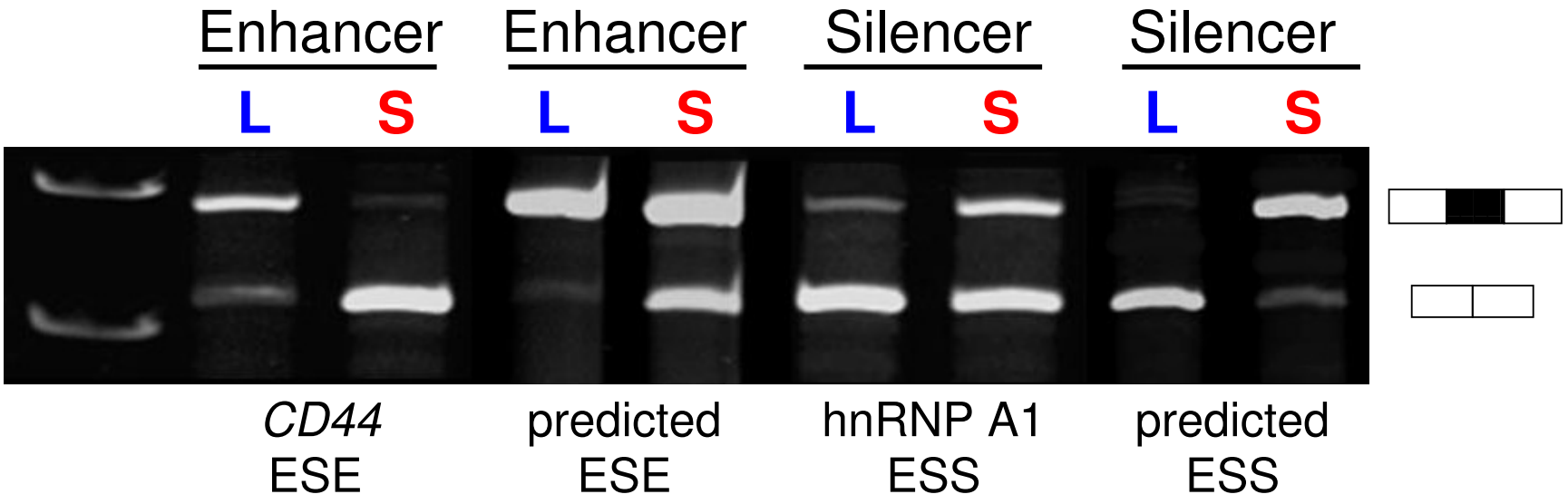
# Experimental Testing

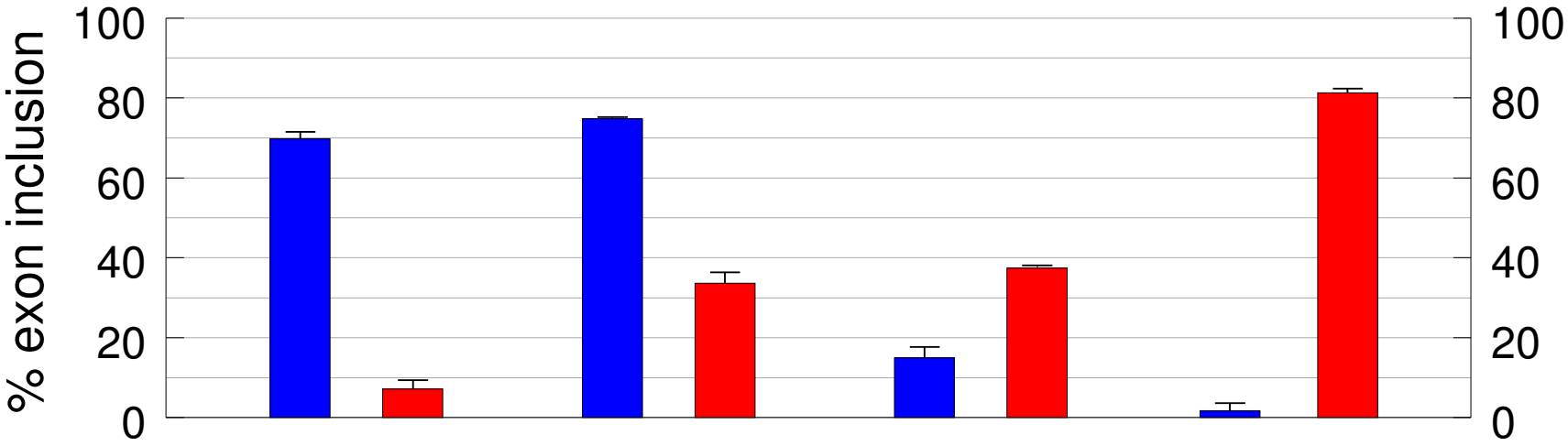
inserts with known splicing motifs (TAGGGT, hnRNP A1)




[Hiller et al. PLoS Genetics 2007]

# Experimental Testing




⇒ **secondary structure of ESE/ESS affects splicing**

[Hiller et al. PLoS Genetics 2007]

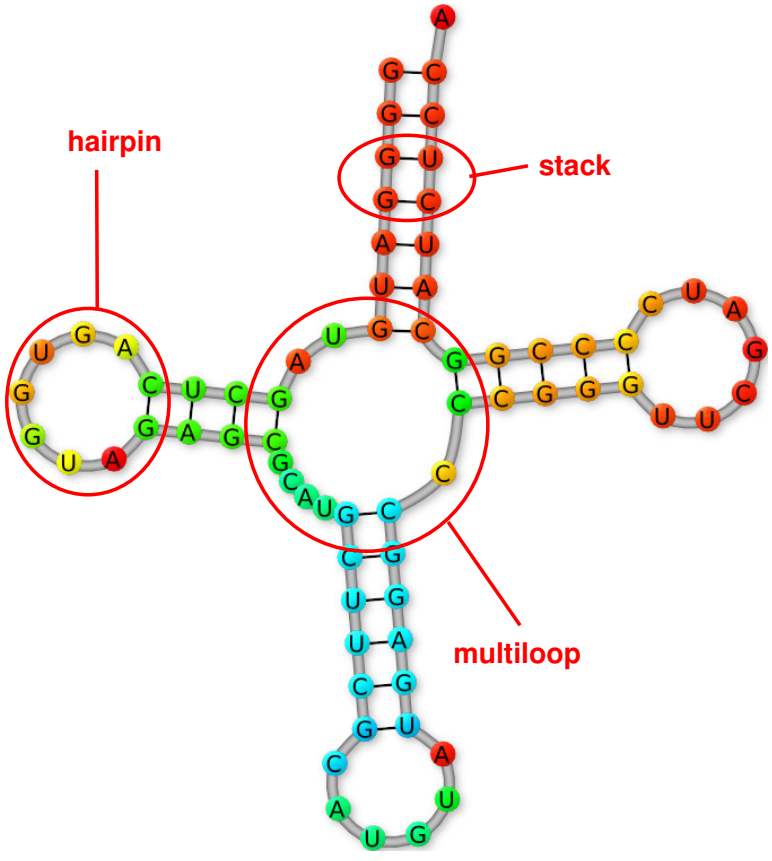
# Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models

# RNA Secondary Structure Prediction

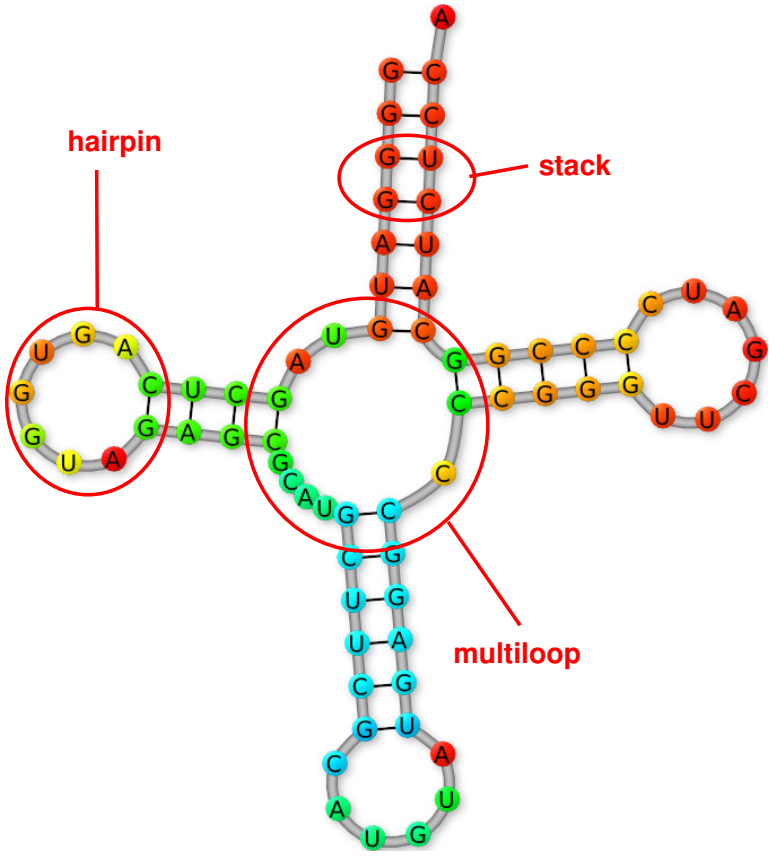
- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure

- Mouse tRNA-ALA:

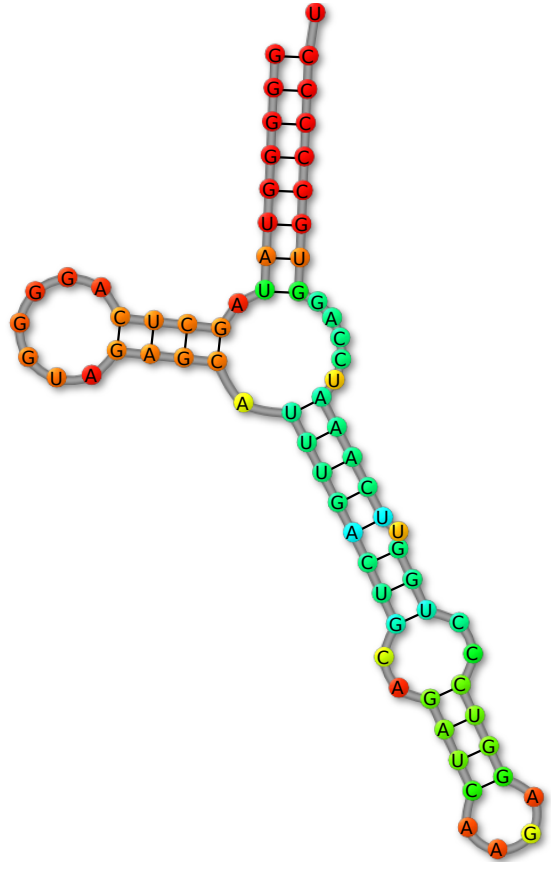


# RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:

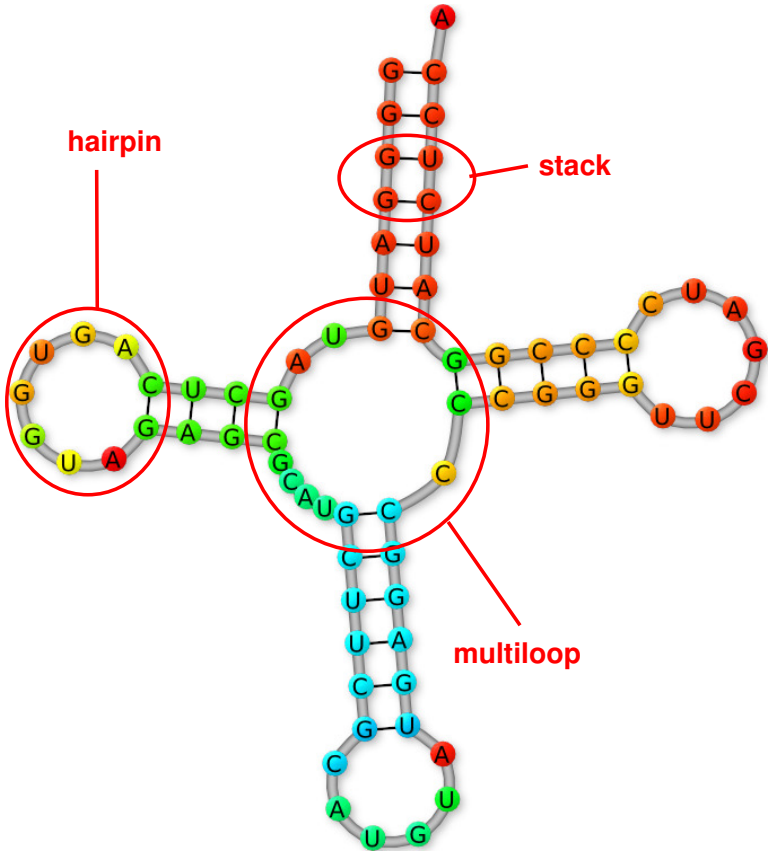



- Mouse tRNA-CYS:

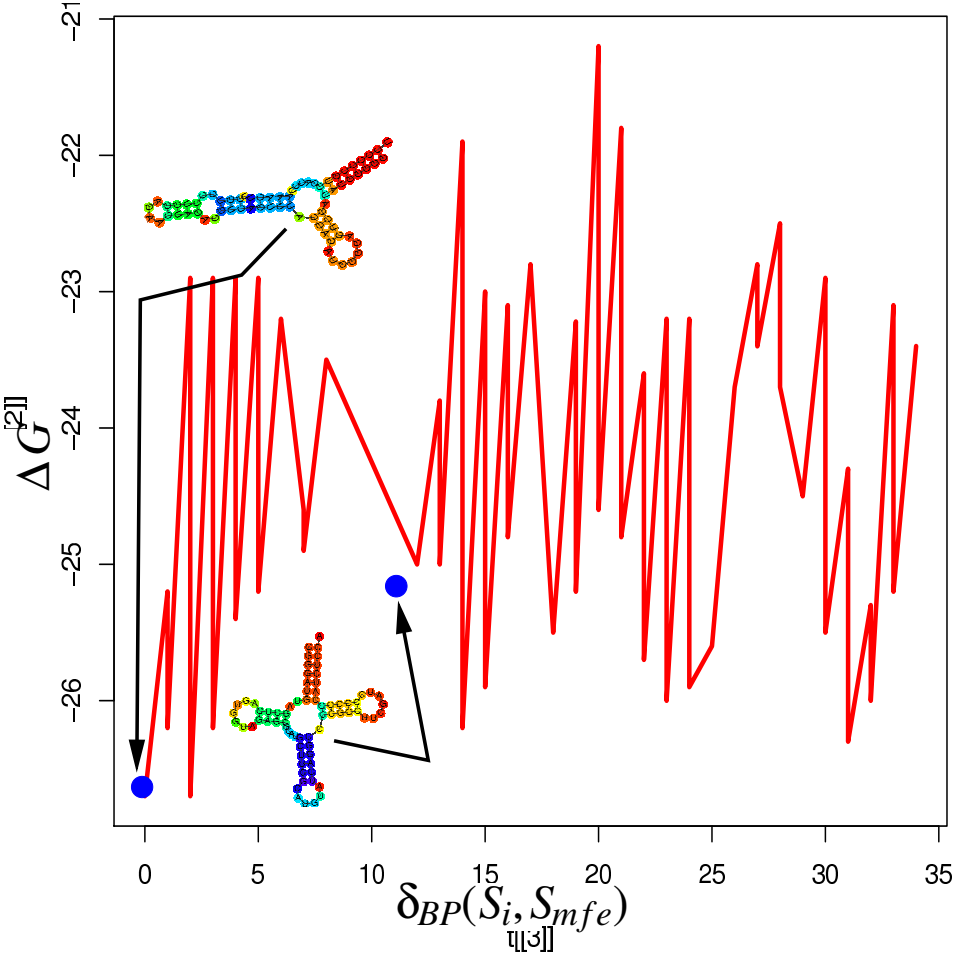


# RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:

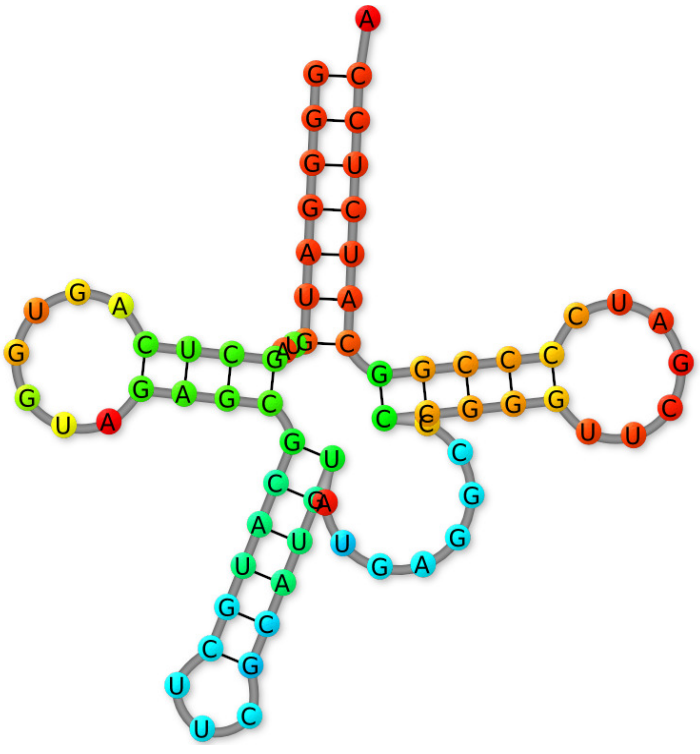



- Mouse tRNA-CYS:

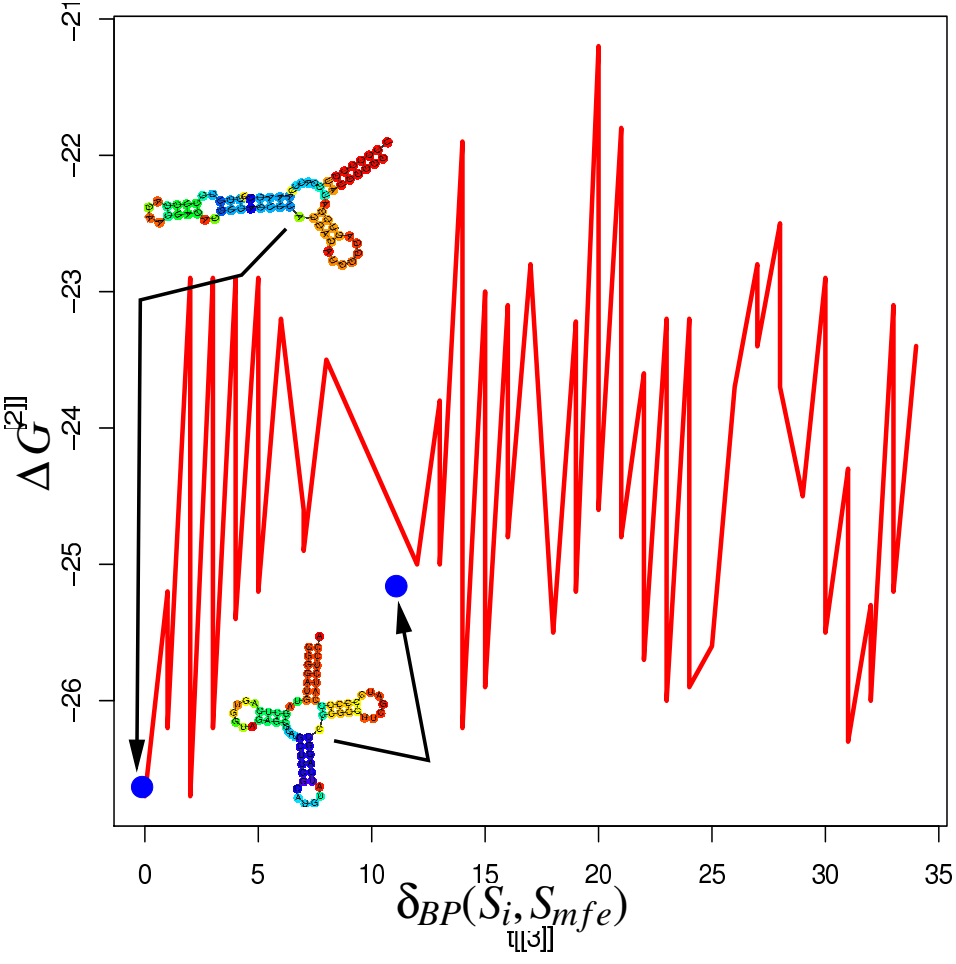


# RNA Secondary Structure Prediction

- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:

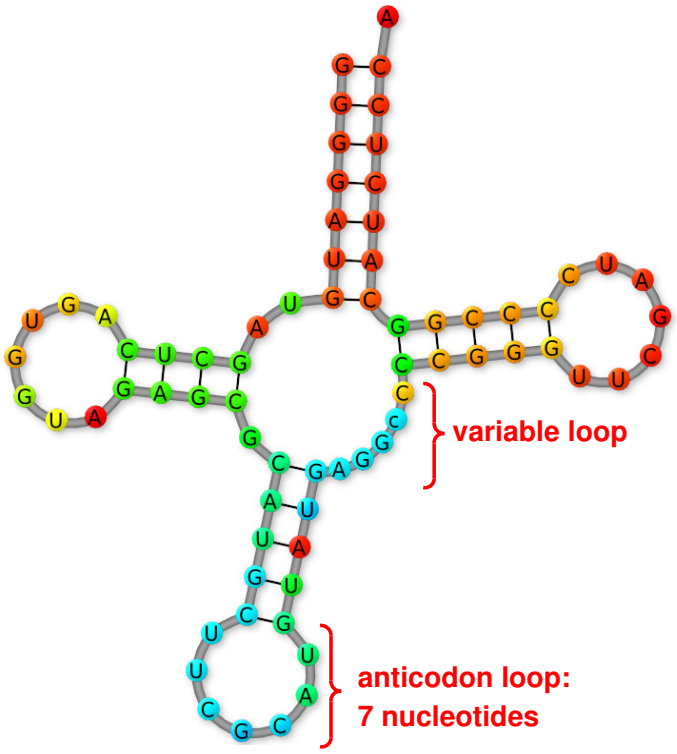


- Mouse tRNA-CYS:

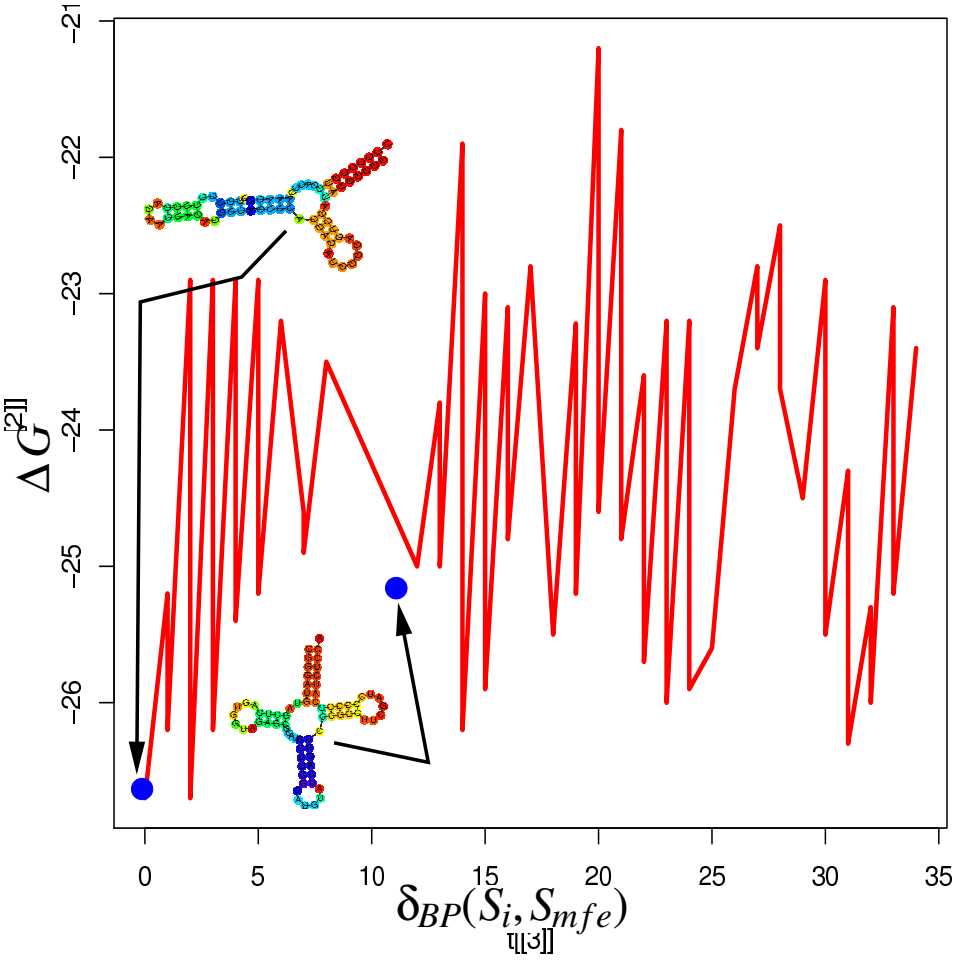


# RNA Secondary Structure Prediction

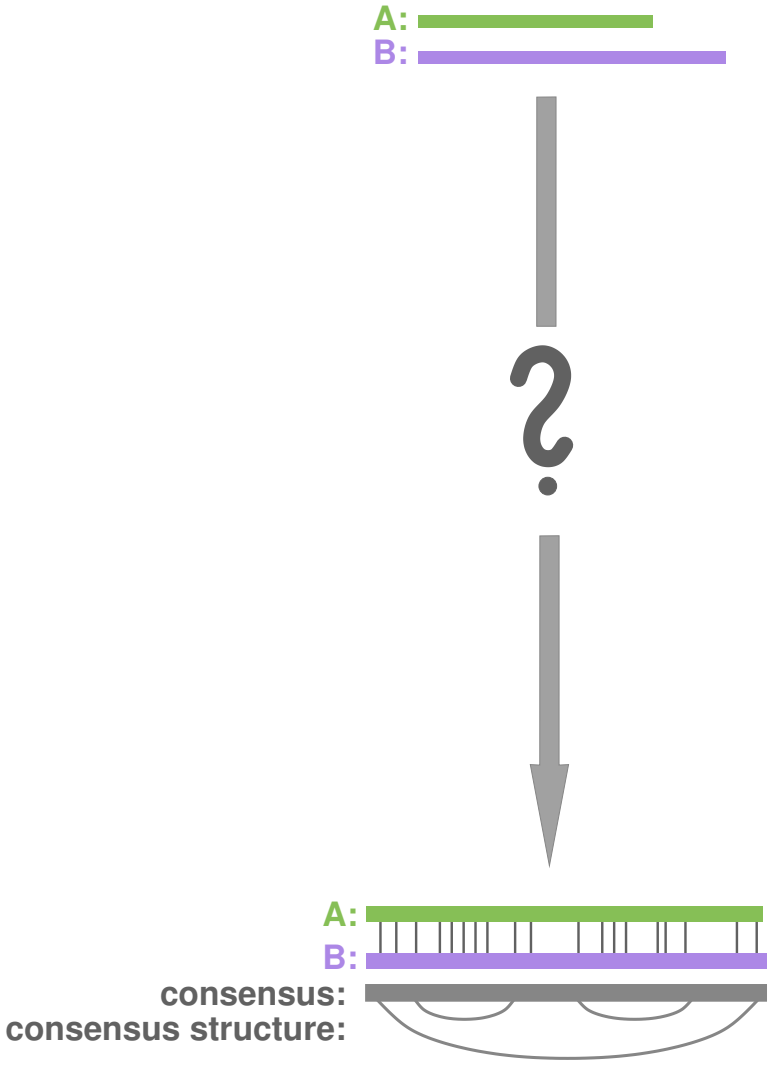
- Turner energy-model: free energies for *loops*
- efficient calculation of minimal free energy (MFE) structure
- problem: MFE is often wrong
- Mouse tRNA-ALA:




- Mouse tRNA-CYS:

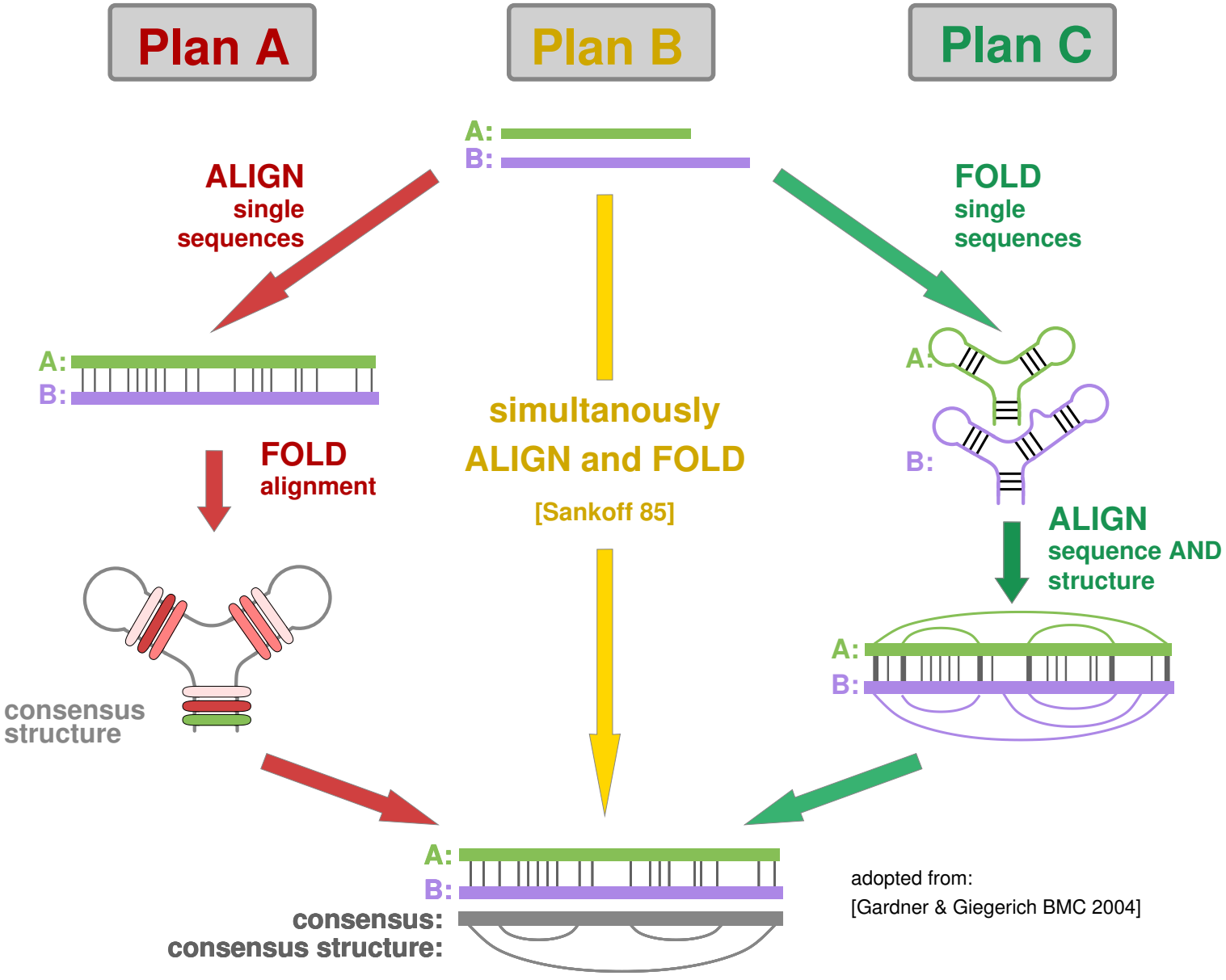


# Comparative RNA Analysis

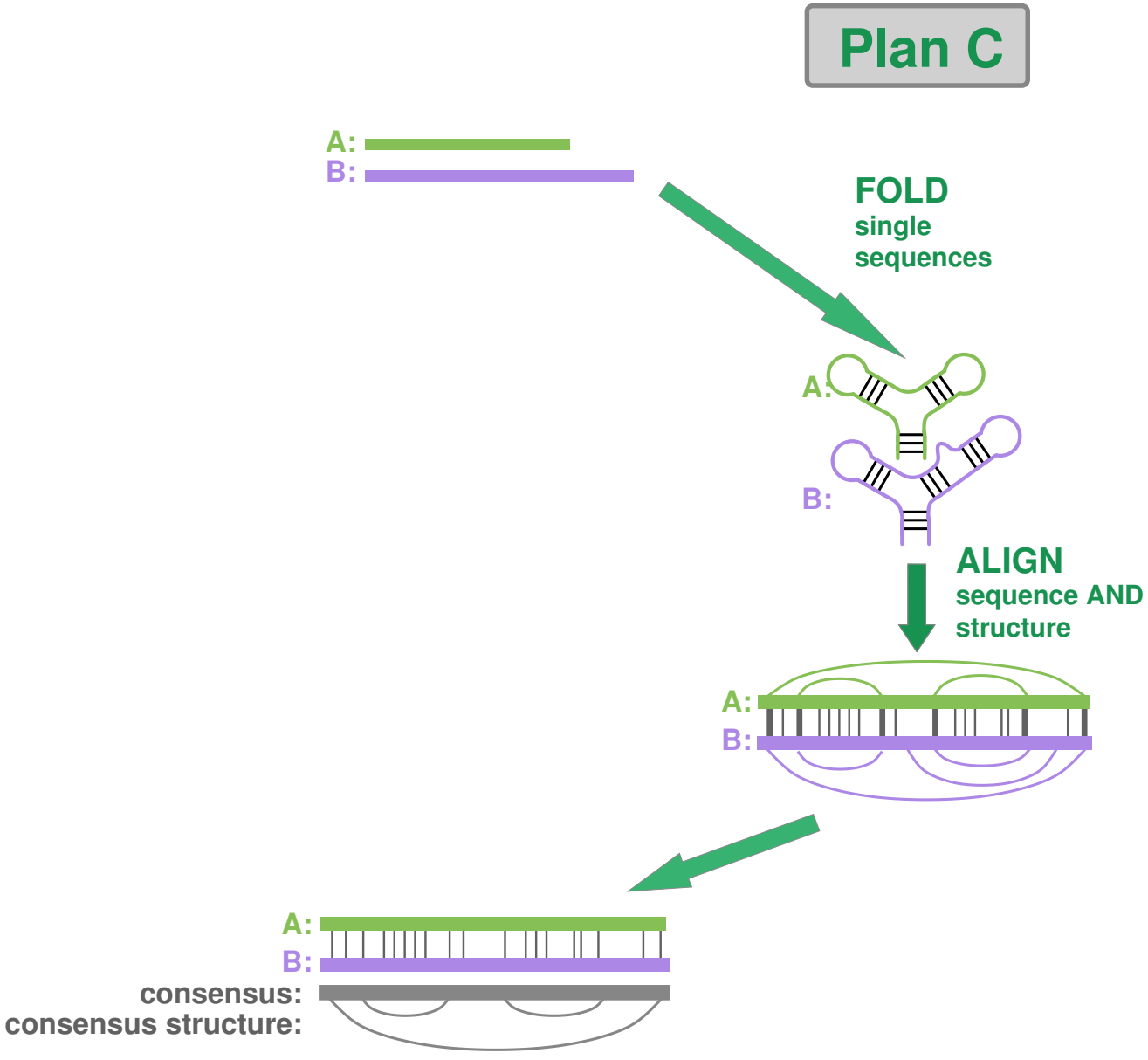



adopted from:
[Gardner & Giegerich BMC 2004]

# Comparative RNA Analysis

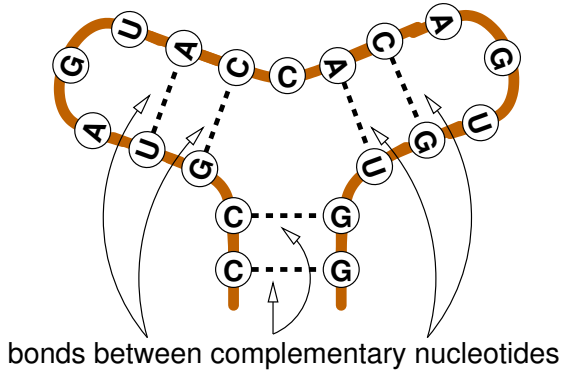


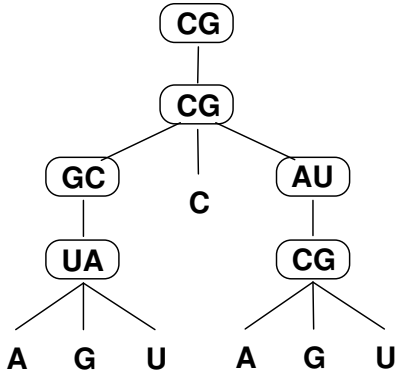
# Comparative RNA Analysis



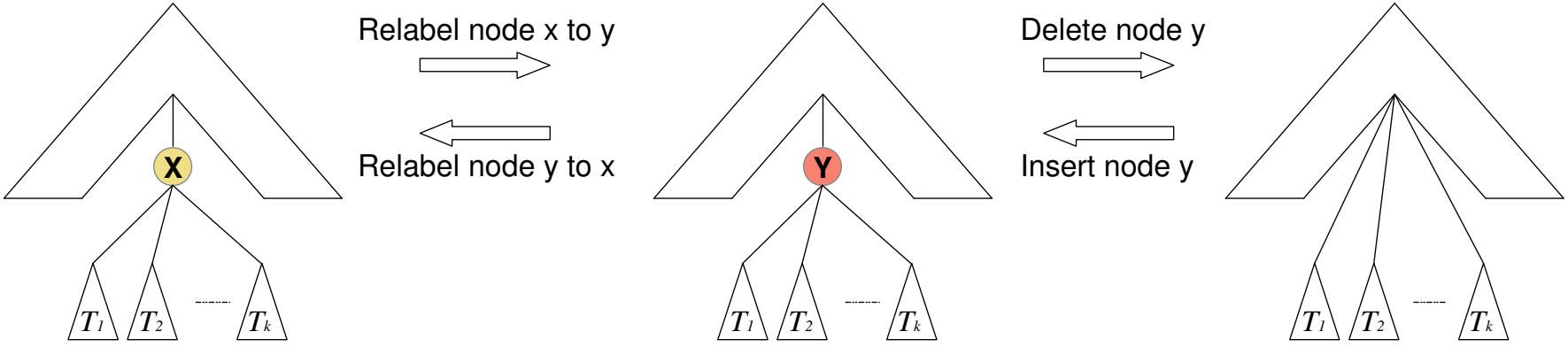
# Tree representation of RNA

- two representations of RNS secondary structure




- edit operation on trees



# Zhang and Sascha's Method

- Associated Recursion Equation
  - $\delta(\emptyset, \emptyset) = 0$
  - $\delta(F, \emptyset) = \delta(F - r_F, \emptyset) + c_{\text{del}}(r_F)$
  - $\delta(\emptyset, G) = \delta(\emptyset, G - r_G) + c_{\text{del}}(r_G)$
  - $$\delta(F, G) = \min \begin{cases} \delta(F - r_F, G) + c_{\text{del}}(r_F), \\ \delta(F, G - r_G) + c_{\text{del}}(r_G), \\ \delta(R_F^\circ, R_G^\circ) + \delta(F - R_F, G - R_G) + c_{\text{match}}(r_F, r_G) \end{cases}$$

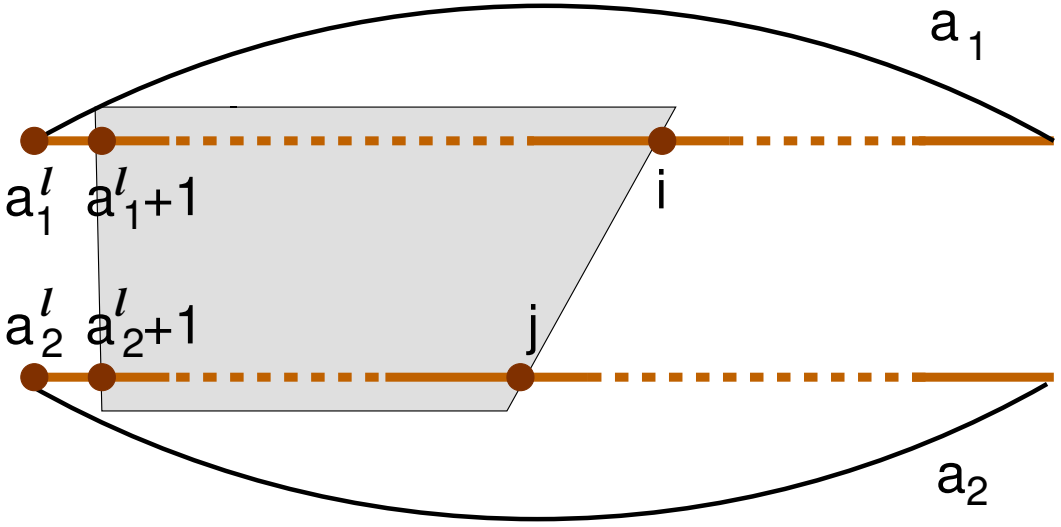
- $F^\circ$ is the special case of $F - r_F$ for $F$ rooted
- $R_F$ denotes the rightmost child in $F$
- $O(n^2 m^2)$ algorithm with $n = |F_1|$ and $m = |F_2|$

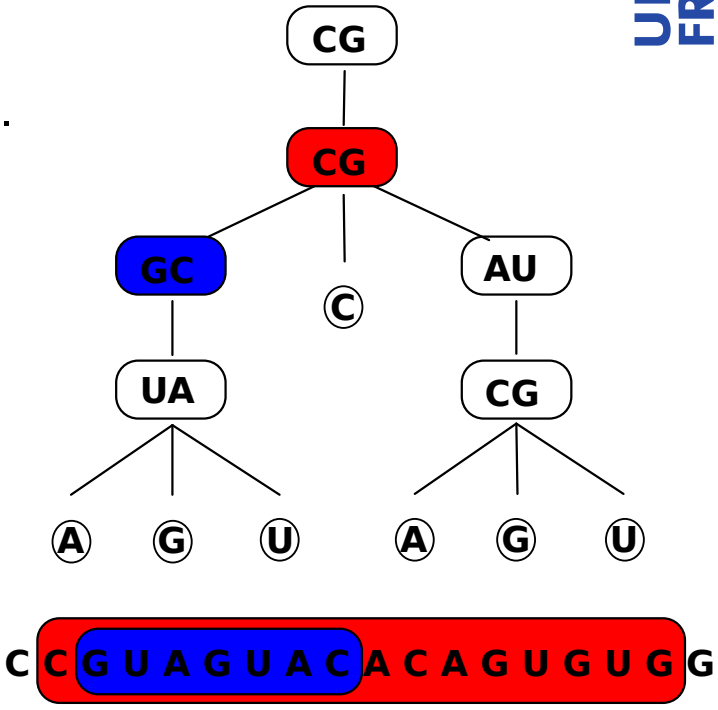
- Zhang and Sascha: fewer subproblems are needed
- *relevant* problems: prefix of $F^\circ$, where $F$ is a root having degree $> 2$.

# Same in our DP Notation

- forests $F$ and $G$: all regions $[i..i']$ and $[j..j']$.
  $\Rightarrow O(n^4)$ space and $O(n^6)$ time

- $\delta(F, G)$ then corresponds to $D(i, i', j, j')$ (alignment of subsequences)
- **But:** not all entries are considered

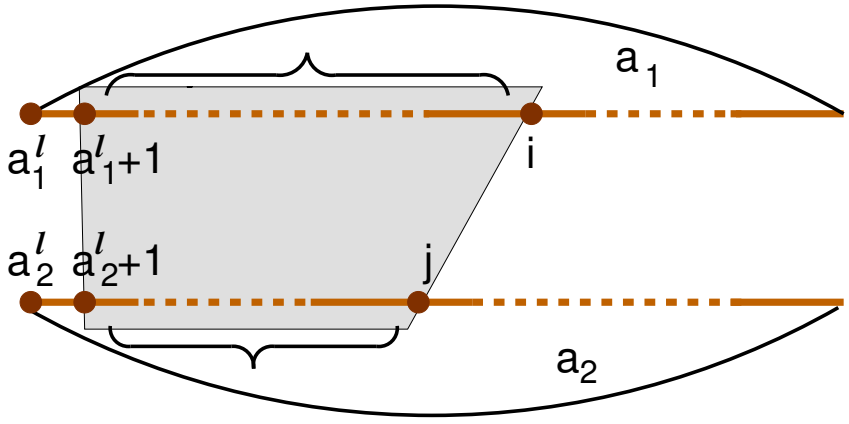



- **Hence:** $O(n^2)$-matrices $M^{a_1}_{a_2}(i, j)$ for all pairs of arcs $a_1$, $a_2$.

# Dynamic Programming for Sequence/Structure Alignmer

- matrices $M_{a_2}^{a_1}(i,j)$: subsequences/substructures under arcs $a_1$ and $a_2$

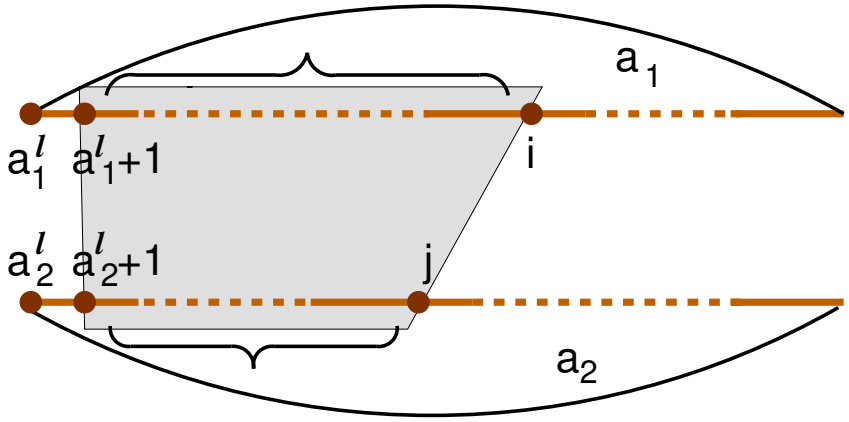



- recursion:

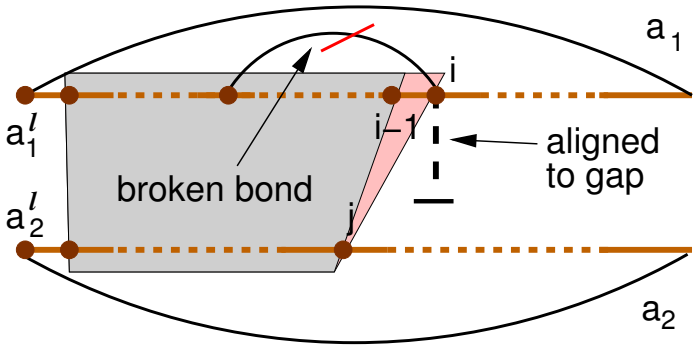
[Backofen&Will: JCBC 04]

# Dynamic Programming for Sequence/Structure Alignmen

- matrices $M_{a_2}^{a_1}(i,j)$: subsequences/substructures under arcs $a_1$ and $a_2$



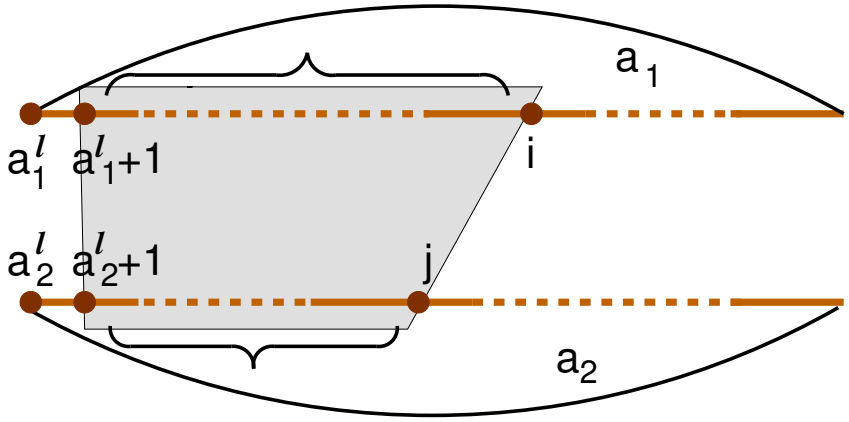

- recursion:
  - base (mis-)match, indel $\Rightarrow M_{a_2}^{a_1}(i-1,j)$ ...
    existing arcs broken in this case



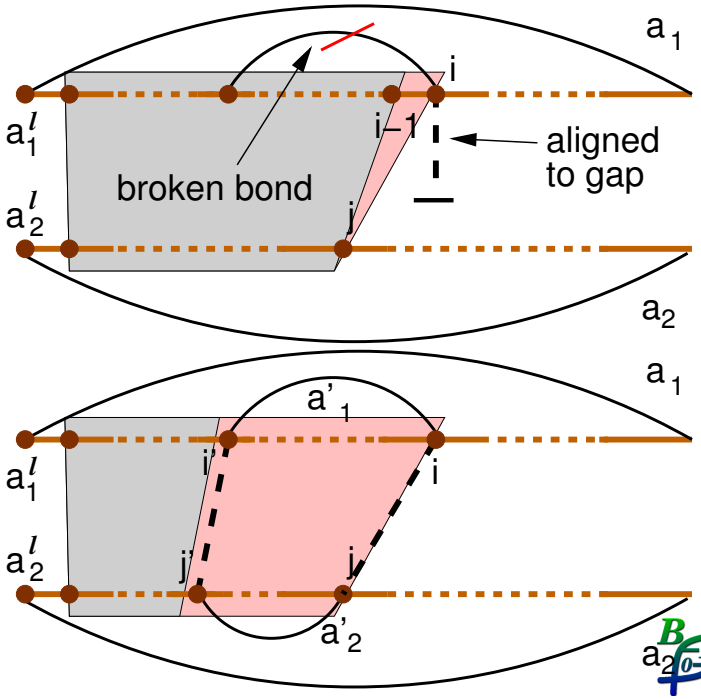

[Backofen&Will: JCBC 04]

# Dynamic Programming for Sequence/Structure Alignmen

- matrices $M_{a_2}^{a_1}(i,j)$: subsequences/substructures under arcs $a_1$ and $a_2$

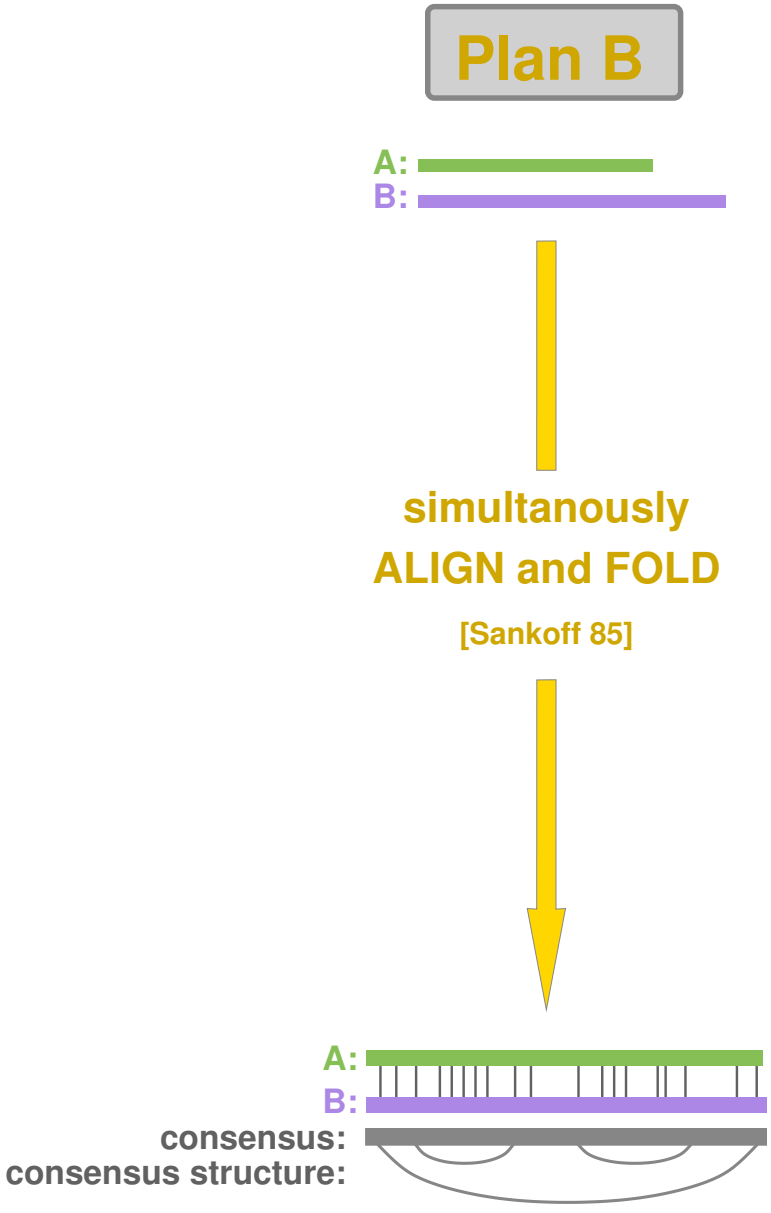



- recursion:
  - base (mis-)match, indel $\Rightarrow M_{a_2}^{a_1}(i-1,j)$ ...
    existing arcs broken in this case
  - arc match $\Rightarrow M_{a_2}^{a_1}(i'-1,j'-1) + M_{a_2'}^{a_1'}(i,j)$




[Backofen&Will: JCBC 04]

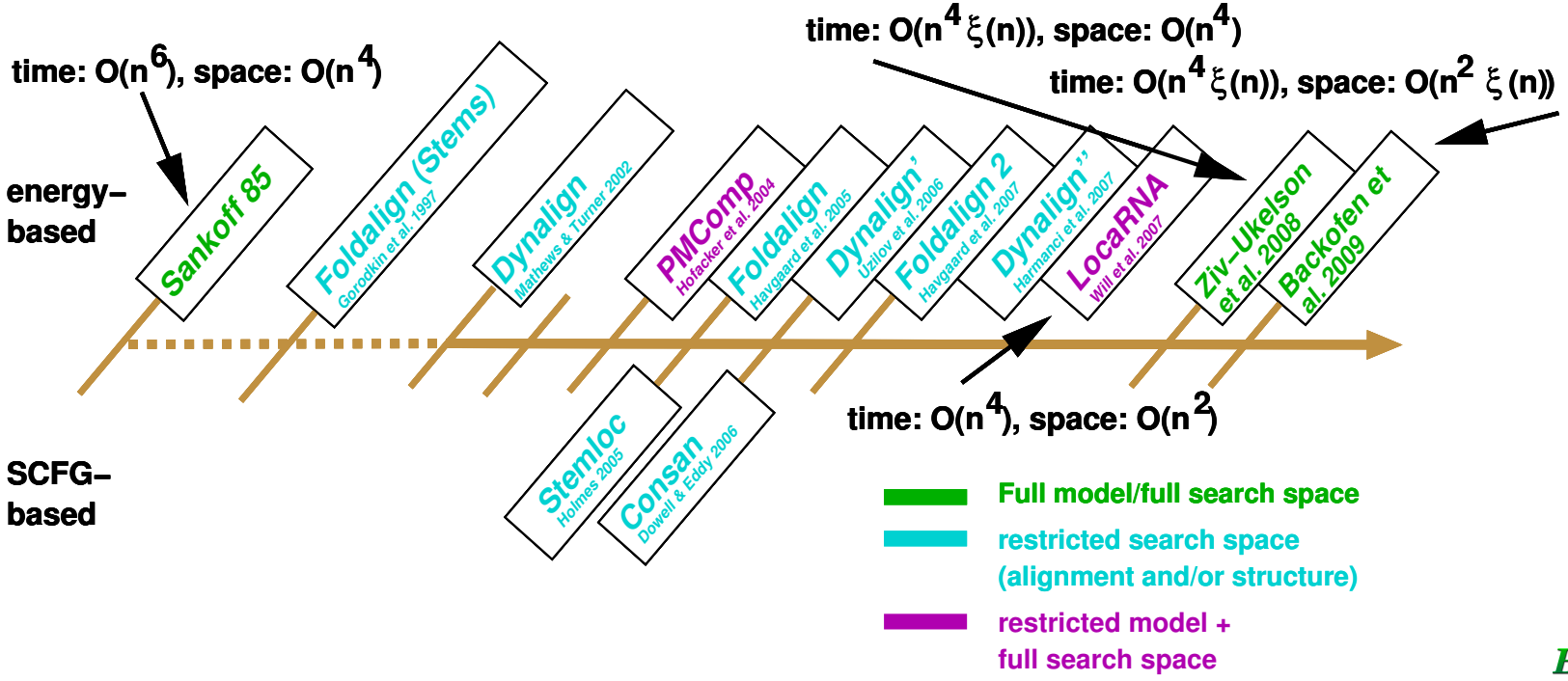
# Comparative RNA Analysis



# Sankoff-like approaches

- Sankoff is the gold standard *BUT requires extreme amount of space and time* *[Gardner & Giegerich 2004]*

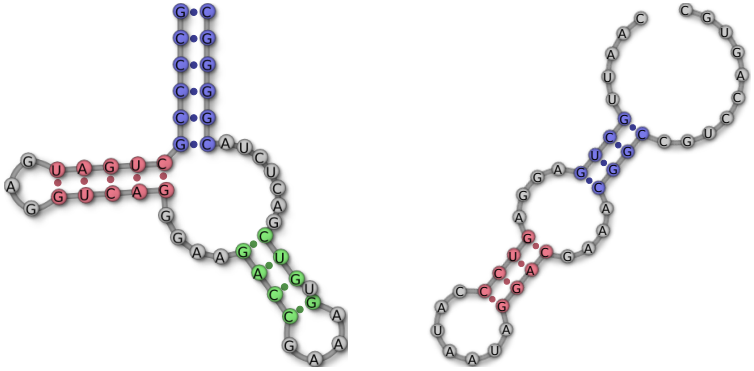
  (time: $O(n^6)$, space $O(n^4)$)
- hence: Sankoff-like approaches are restricted versions




[Will et al: PLoS Comp. Bio 07],[Backofen et al: CPM 09], [Möhl et al: RECOMB 09]...

# Problem: Suboptimal Structures

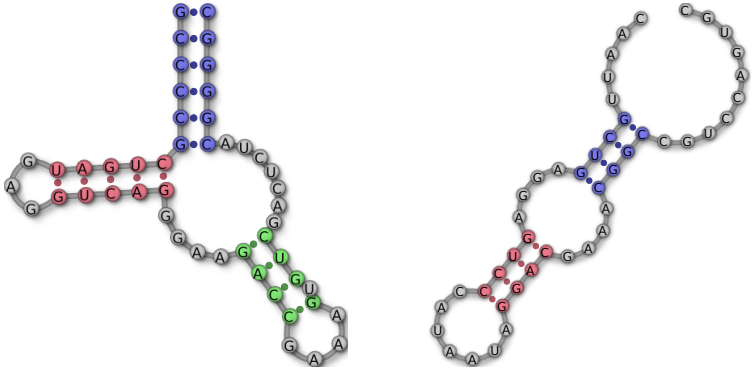
- example: two hammerhead ribozymes



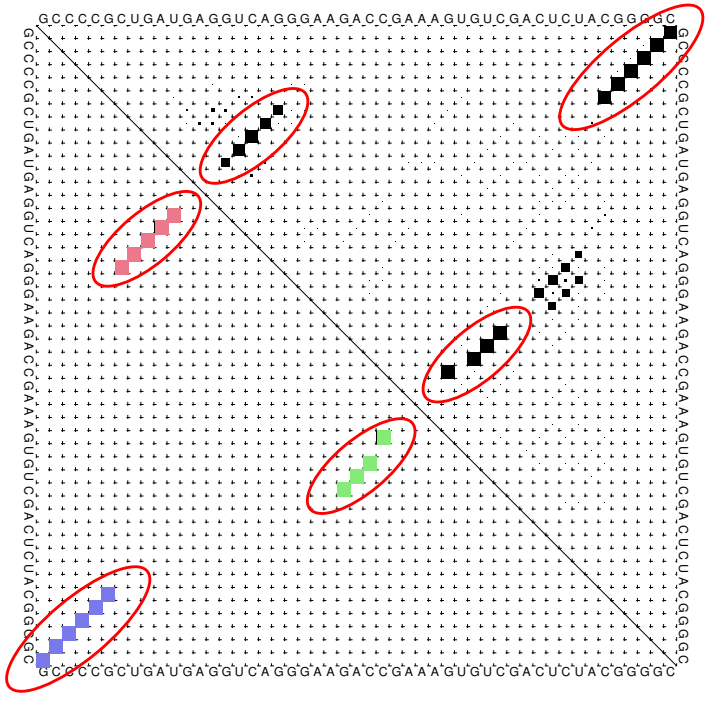
[Will et al: PLoS Comp. Bio 07]

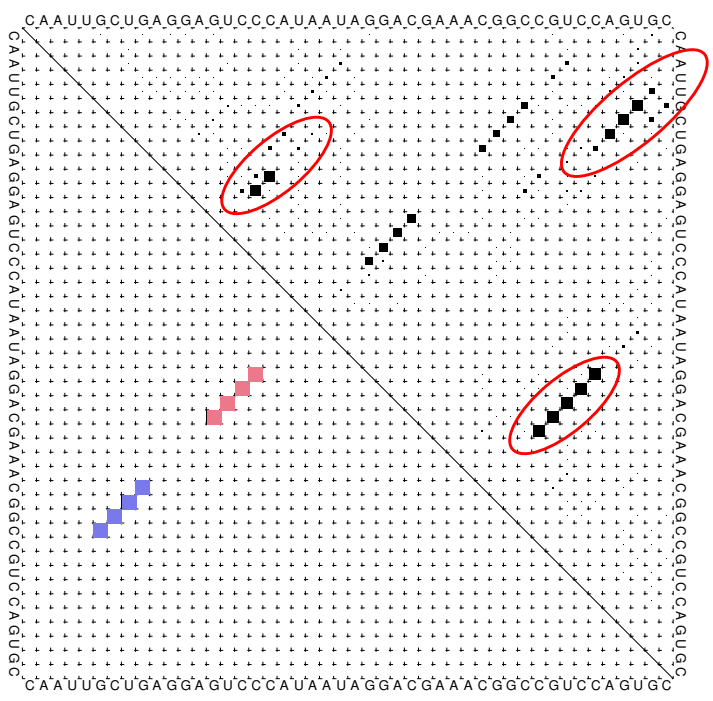
# Problem: Suboptimal Structures

- example: two hammerhead ribozymes



- corresponding dotplots

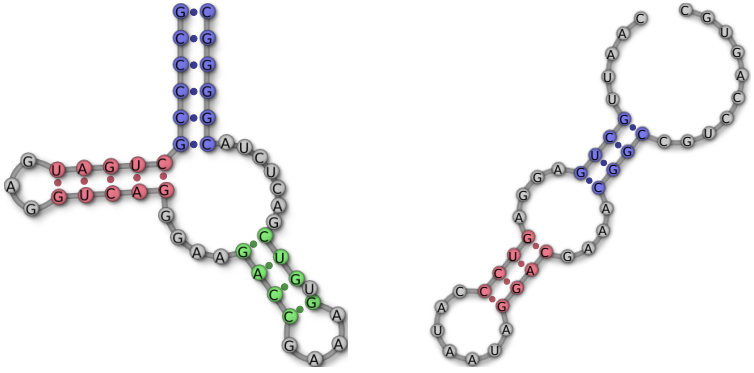




[Will et al: PLoS Comp. Bio 07]

# Problem: Suboptimal Structures

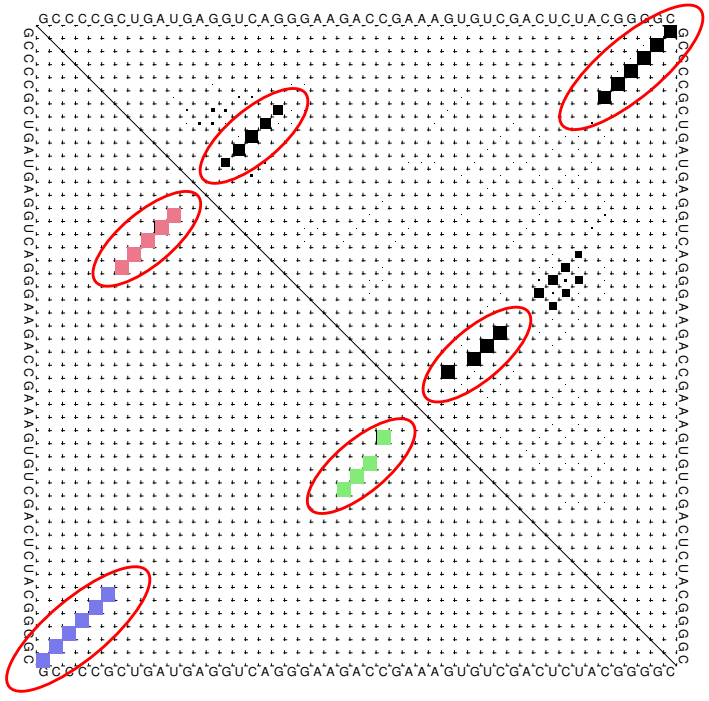
- example: two hammerhead ribozymes

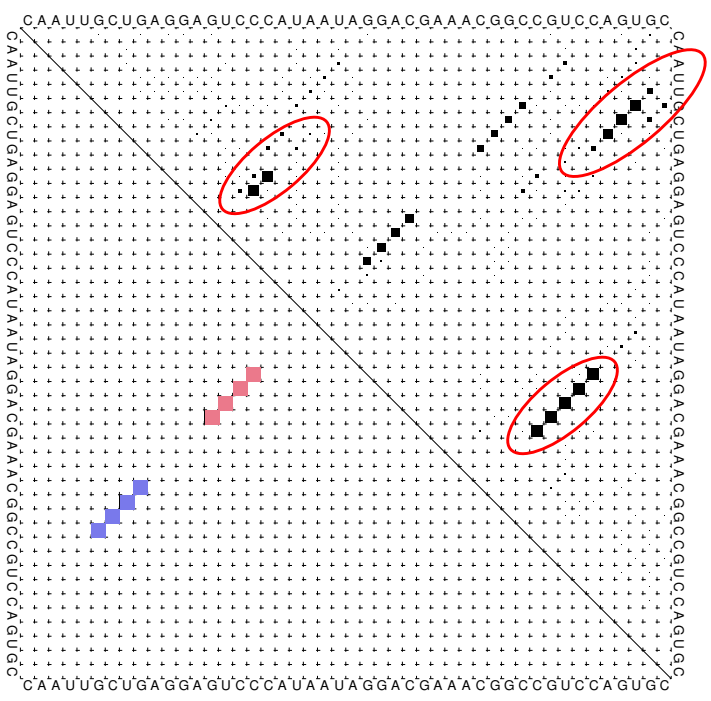


**LocARNA**

- alignment of dotplots
- efficient version of Sankoff

- corresponding dotplots





[Will et al: PLoS Comp. Bio 07]

# Probabilistic Consistency Transformation

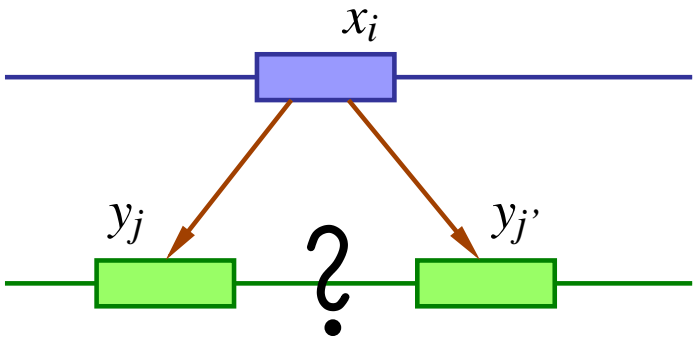
- **remaining problem:** progressive alignment

  *use probabilistic consistency transformation ala ProbCons*

- Idea:
  - Given set of sequences $\mathcal{S}$
  - for all pairs $x, y \in \mathcal{S}$ of sequences calculated:

    *match probabilities* $P(x_i \sim y_j | x, y)$

Then:



# Probabilistic Consistency Transformation

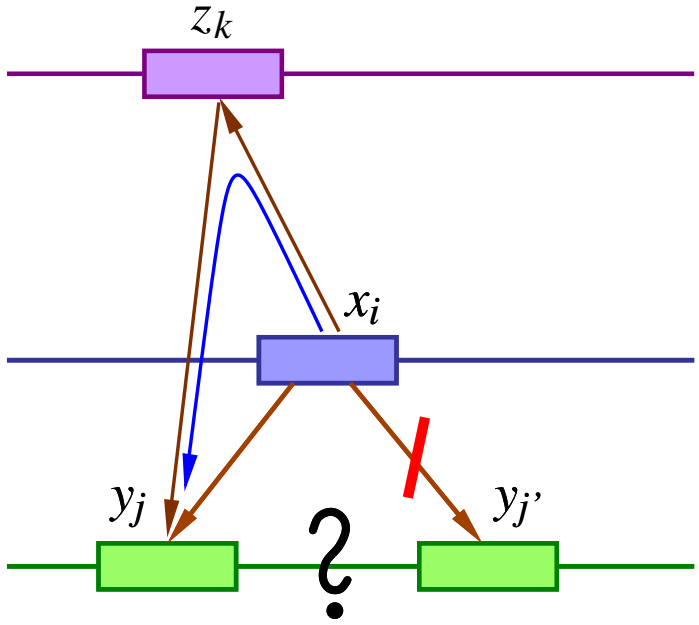
- **remaining problem:** progressive alignment

  *use probabilistic consistency transformation ala ProbCons*

- Idea:
  - Given set of sequences $\mathcal{S}$
  - for all pairs $x, y \in \mathcal{S}$ of sequences calculated:

    *match probabilities* $P(x_i \sim y_j | x, y)$

Then:



# Probabilistic Consistency Transformation

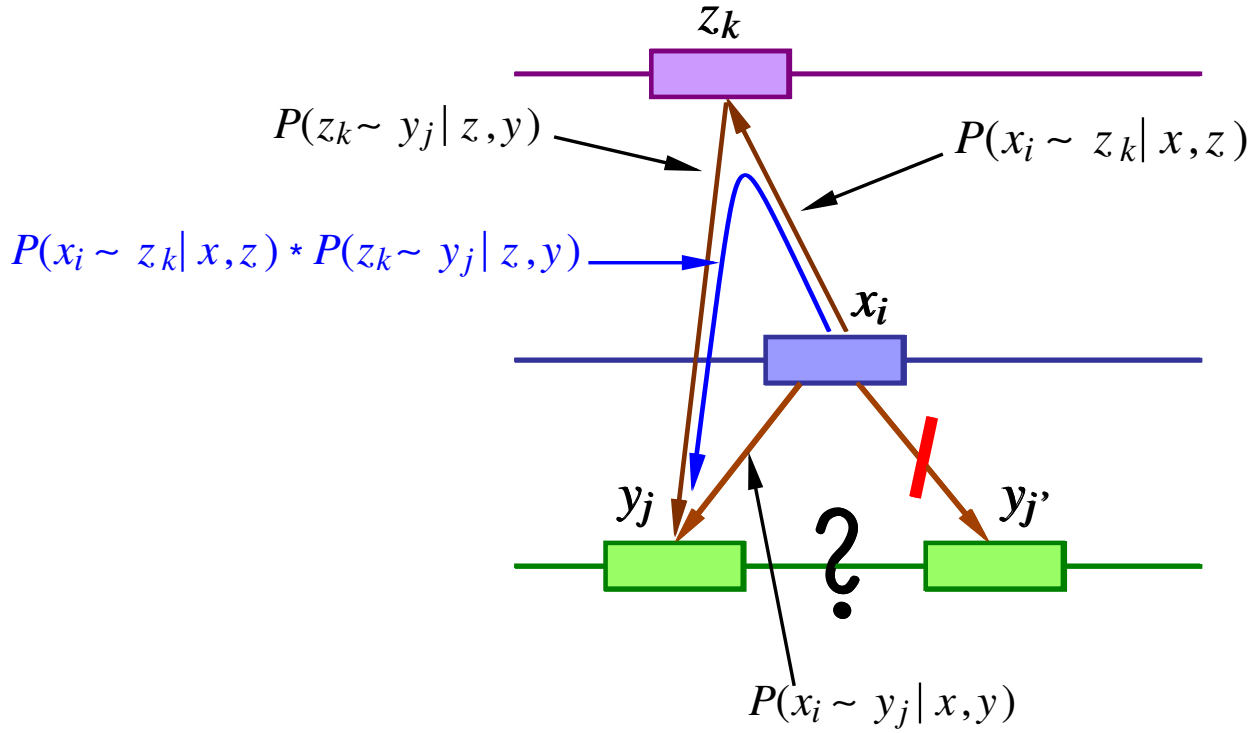
- **remaining problem:** progressive alignment

  *use probabilistic consistency transformation ala ProbCons*

- Idea:
  - Given set of sequences $\mathcal{S}$
  - for all pairs $x, y \in \mathcal{S}$ of sequences calculated:

    *match probabilities* $P(x_i \sim y_j | x, y)$

Then:



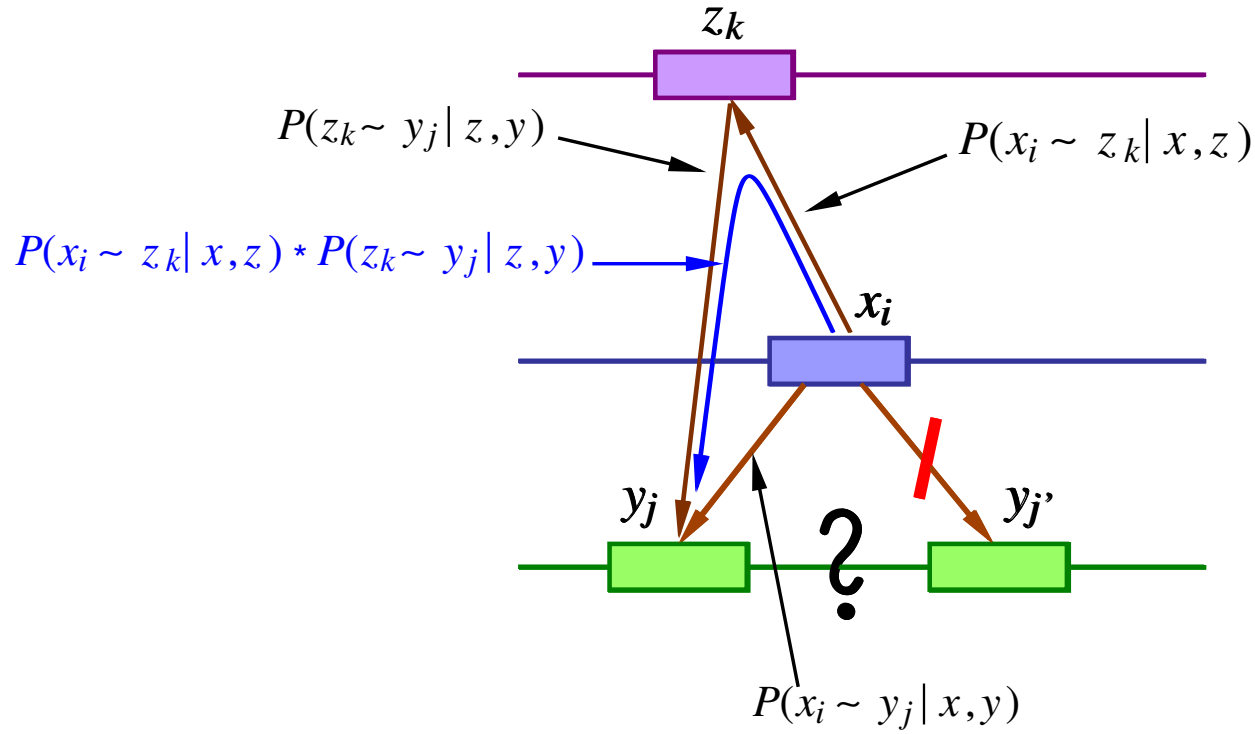
# Probabilistic Consistency Transformation

- **remaining problem:** progressive alignment

  *use probabilistic consistency transformation ala ProbCons*

- Idea:
  - Given set of sequences $\mathcal{S}$
  - for all pairs $x, y \in \mathcal{S}$ of sequences calculated:

    *match probabilities* $P(x_i \sim y_j | x, y)$



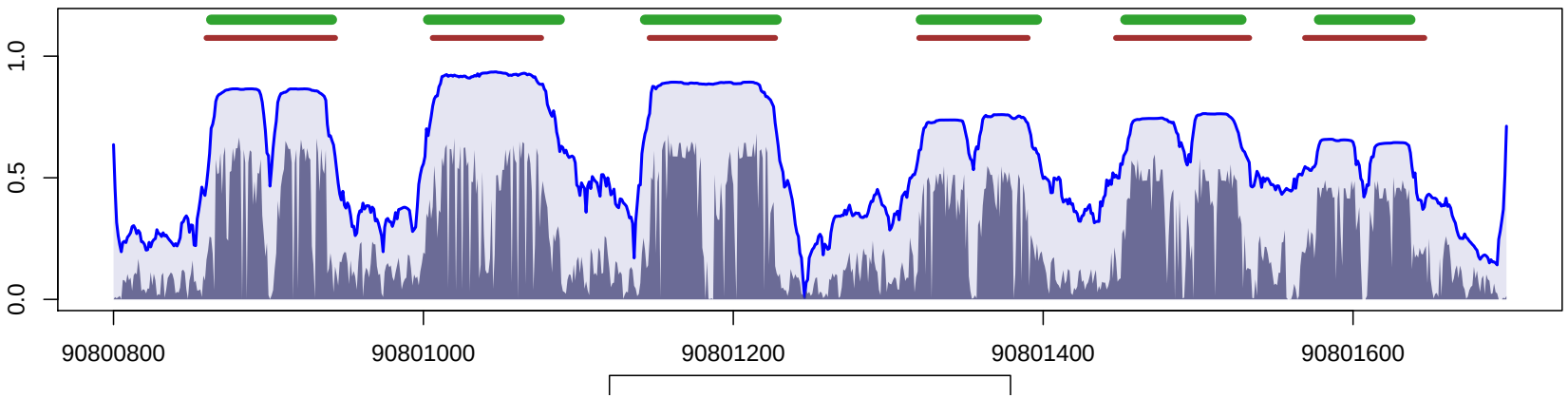

Then:

- do this for all intermediate sequences $z \in \mathcal{S}$

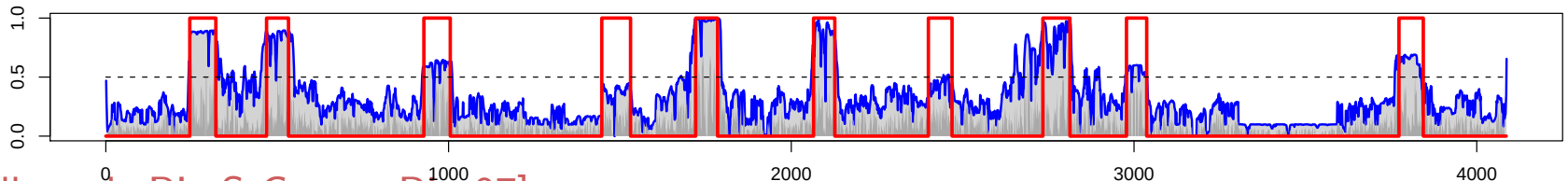
# Reliability Profiles

- genomic cluster with known ncRNAs
- align corresponding regions in 10/5 vertebrates
- show reliability profile for human DNA

cluster of 6 micro RNAs, length ≈900



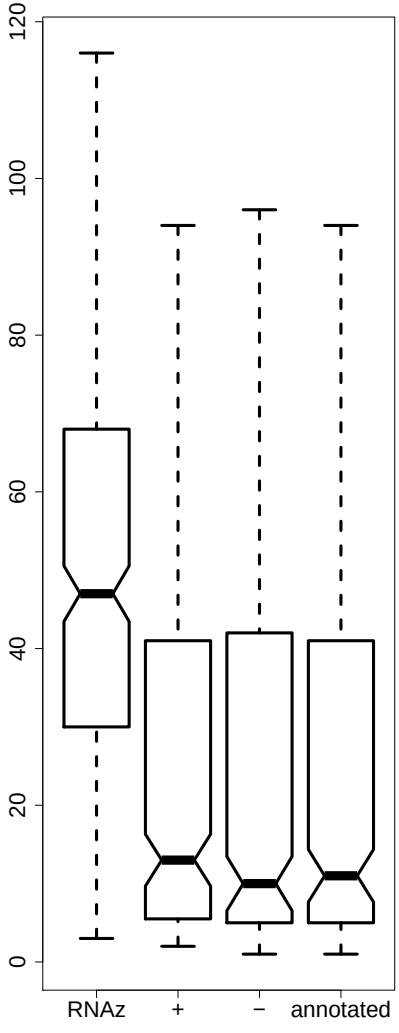

cluster of 10 CD-Box snoRNAs 'GAS5', length ≈4000

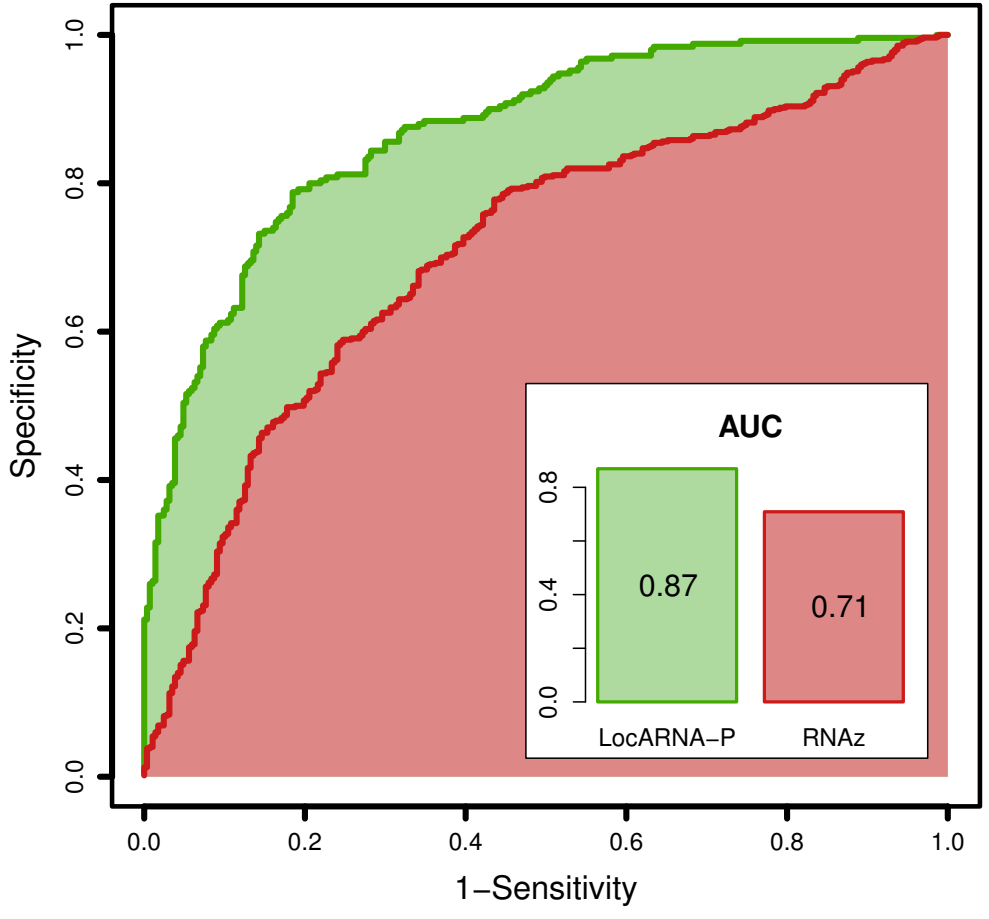



[Will et al: PLoS Comp. Bio 07]

# Boundary Prediction

- median
- use LocaRNA/ boundary prediction / reliability as post-processing filter



# Overview

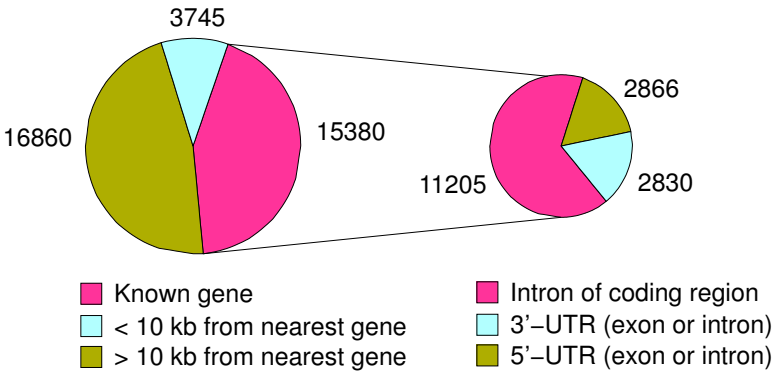
- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes
- barrier trees and HP-lattice models

# Classification of putative ncRNAs

- RNAz: detects functional RNA secondary structures in **multiple sequence alignments**

- results of human RNAz-scan

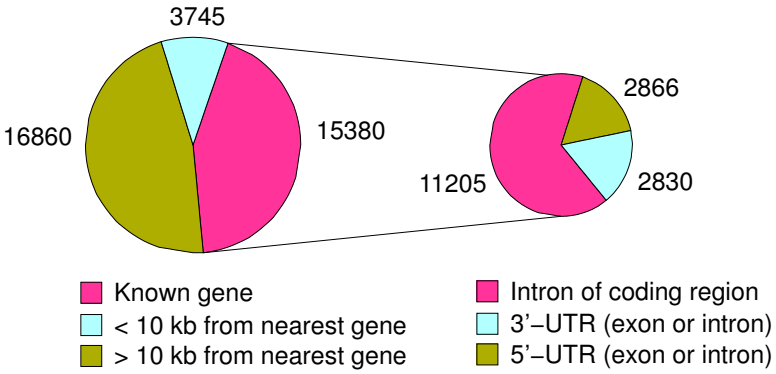
  [Washietl et al. Nature Biotech. 2005]

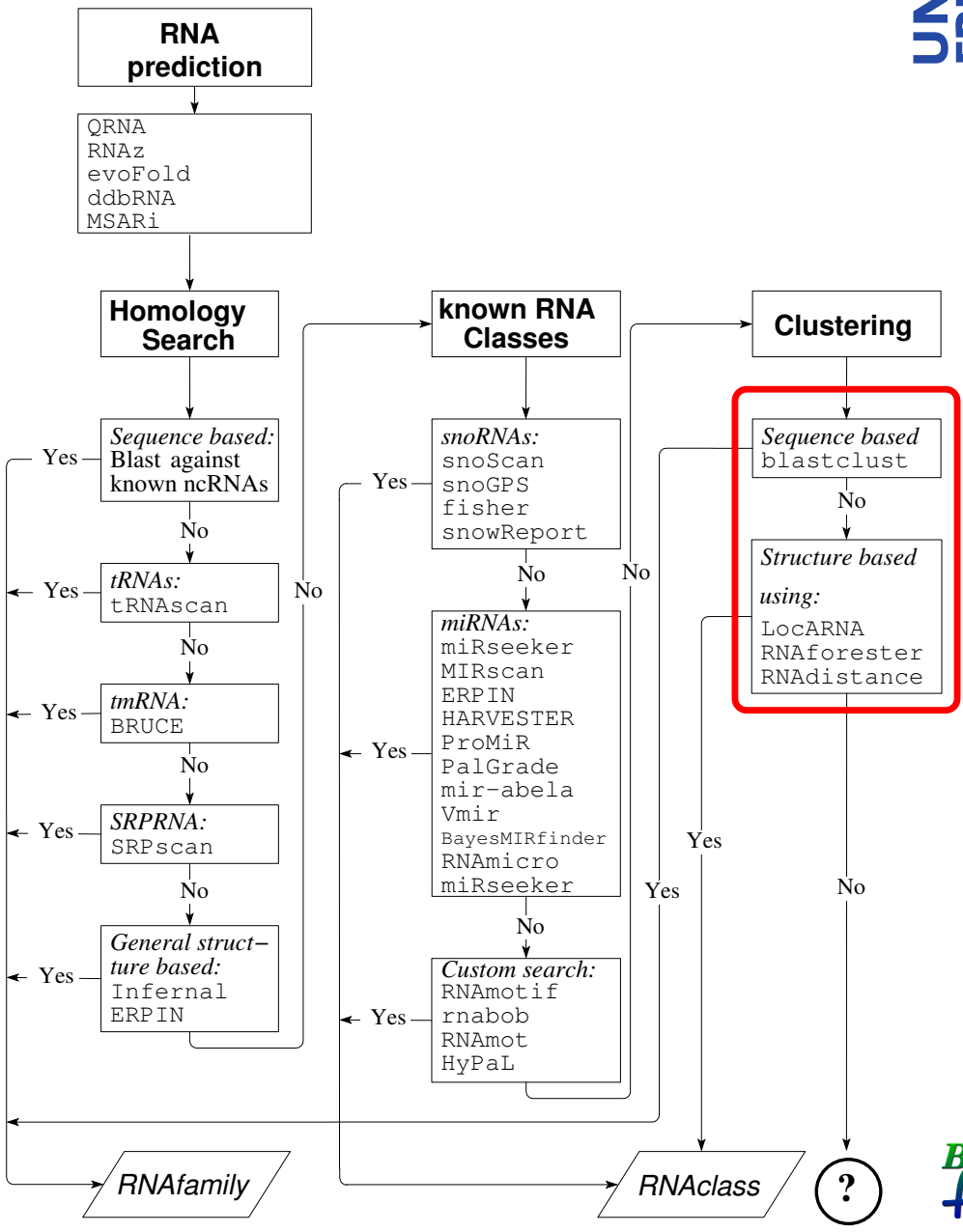


# Classification of putative ncRNAs

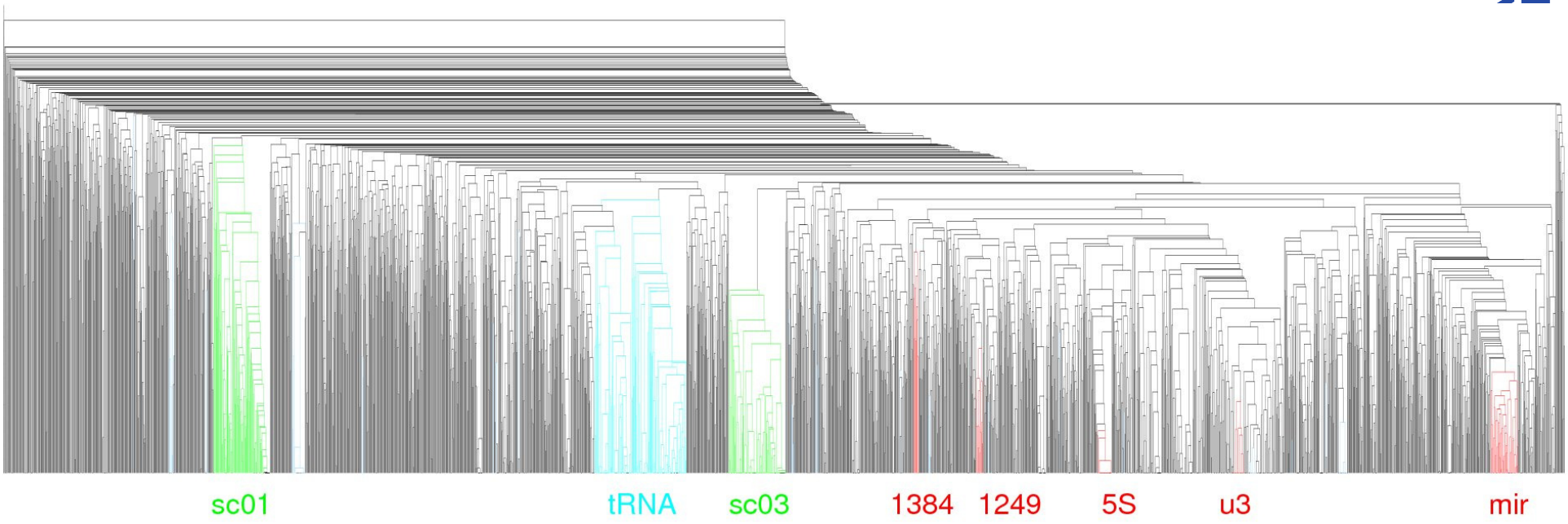
- RNAz: detects functional RNA secondary structures in **multiple sequence alignments**
- results of human RNAz-scan

  [Washietl et al. Nature Biotech. 2005]



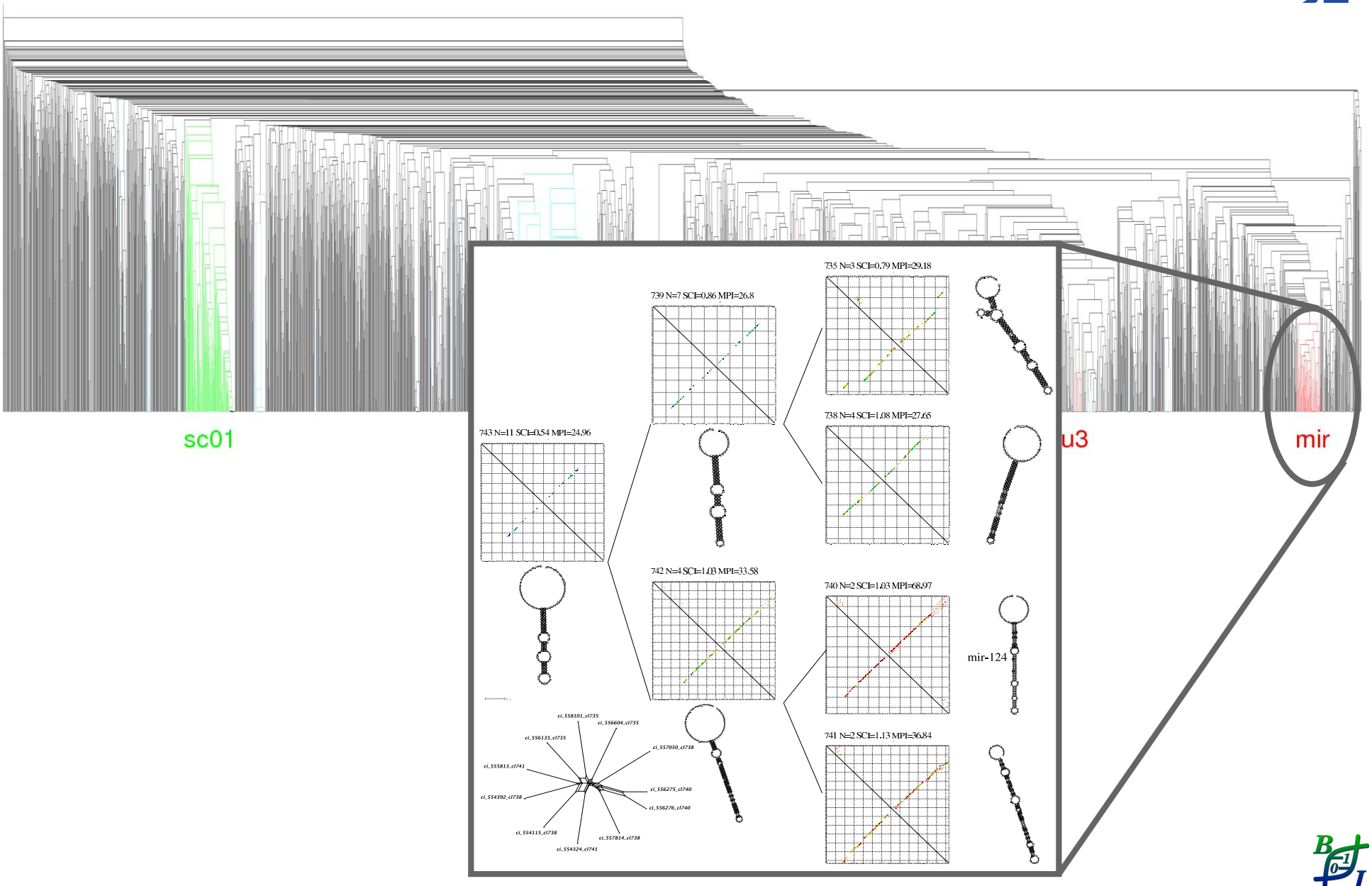



# LocARNA: Clustering of RNAz ncRNA Predictions




[Will et al: PLoS Comp. Bio 07]

# LocARNA: Clustering of RNAz ncRNA Predictions

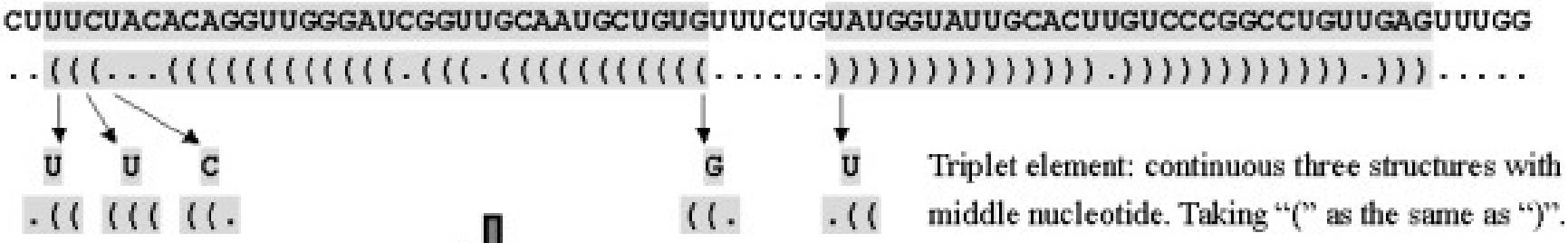



[Will et al: PLoS Comp. Bio 07]

# Classification of ncRNA: MicroRNA Example

- problem: how to classify ncRNAs from properties of RNA 2D structure
  $\Rightarrow$ *learn graph properties*

A sequence and predicted hairpin secondary structure: only stem portions (shadow regions) of the hairpin are computed.




Triplet element: continuous three structures with middle nucleotide. Taking "(" as the same as ")".

32 triplet element features --- 32-dimension vector:

( U(((, U((., U(.(, U(.., U.((, U.(., U..(, U..., G(((, G((., … )

Counting the appearances of the triplet elements:

( 12, 4, 3, 1, 2, 0, 0, 0, 10, 1, … )

Normalizing the triplet element count vector:

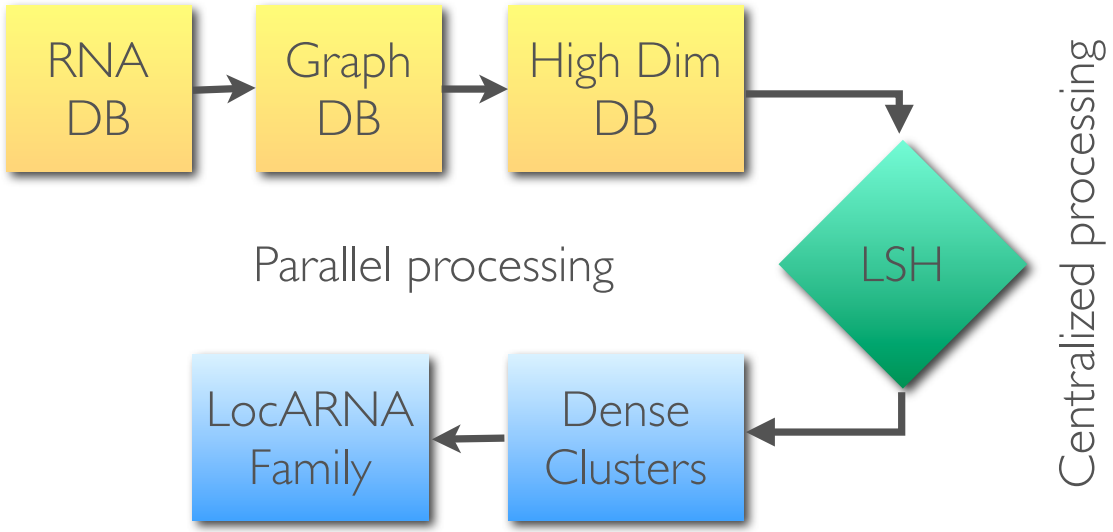
( 0.1846, 0.0615, 0.0462, 0.0154, 0.0308, 0, 0, 0, 0.1538, 0.0154, … )

# Clustering: How to cluster RNAz predictions?

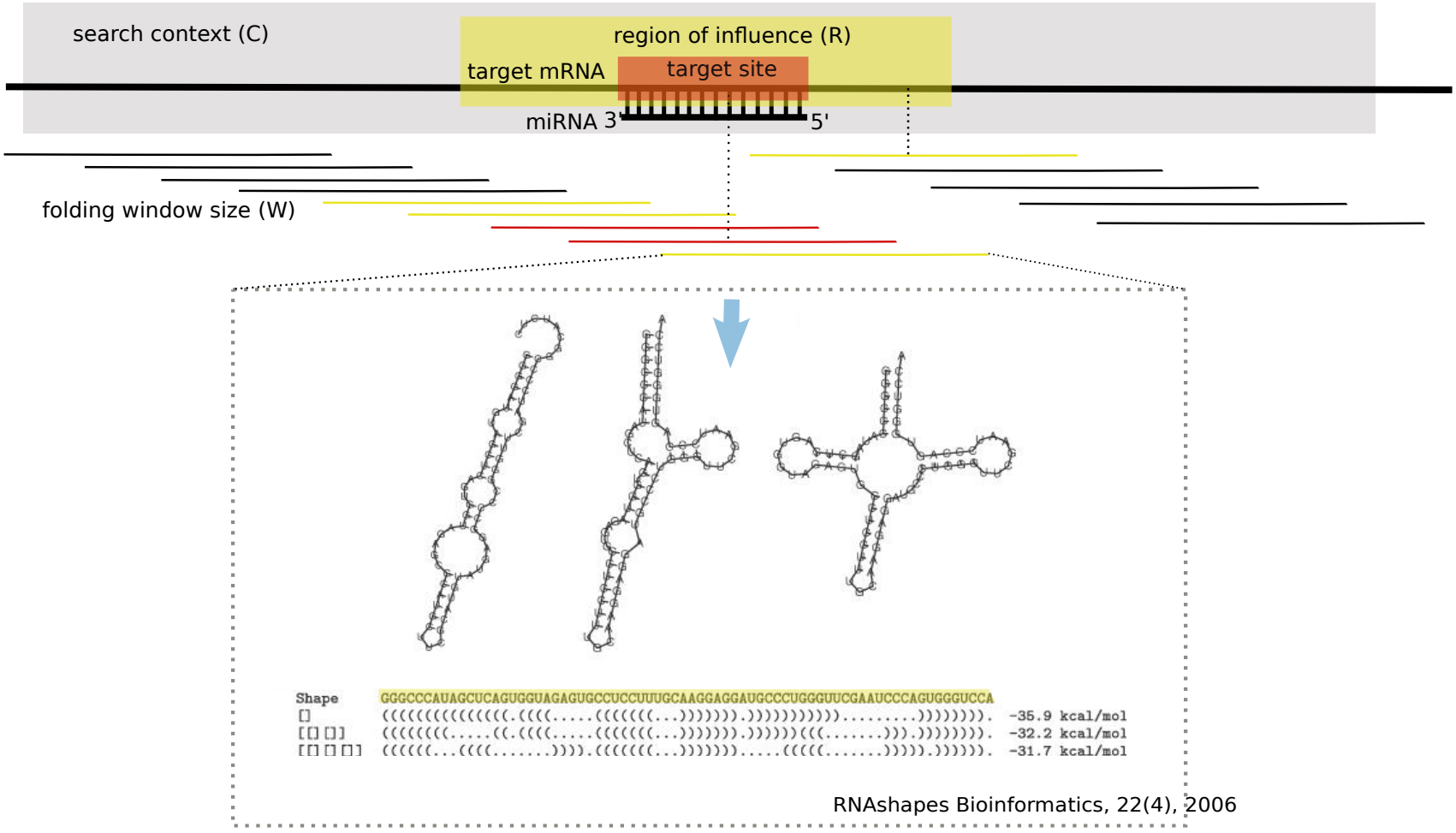
- Problem: still too much data for LocaRNA

  *16.000 Drosophila or 36.000 human RNAz hits*

- solution: modified cluster pipeline (Fabrizio Costa)
  - built bbuild graphs (using RNAShapes) from RNA sequences
  - convert them to high dimensional sparse vectors (graph kernel)
  - Use LSH to efficiently retrieve neighbors and density
  - Return highly dense clusters
  - Refine RNA family models in clusters by LocARNA

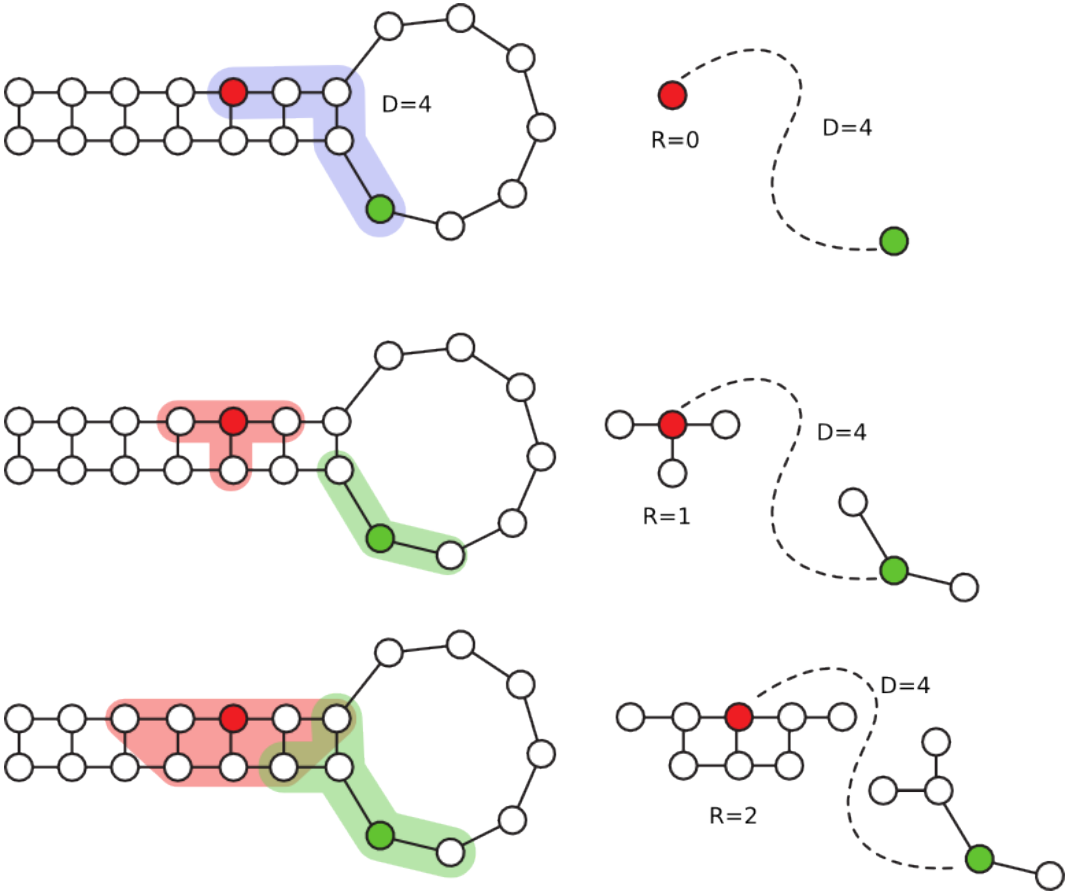


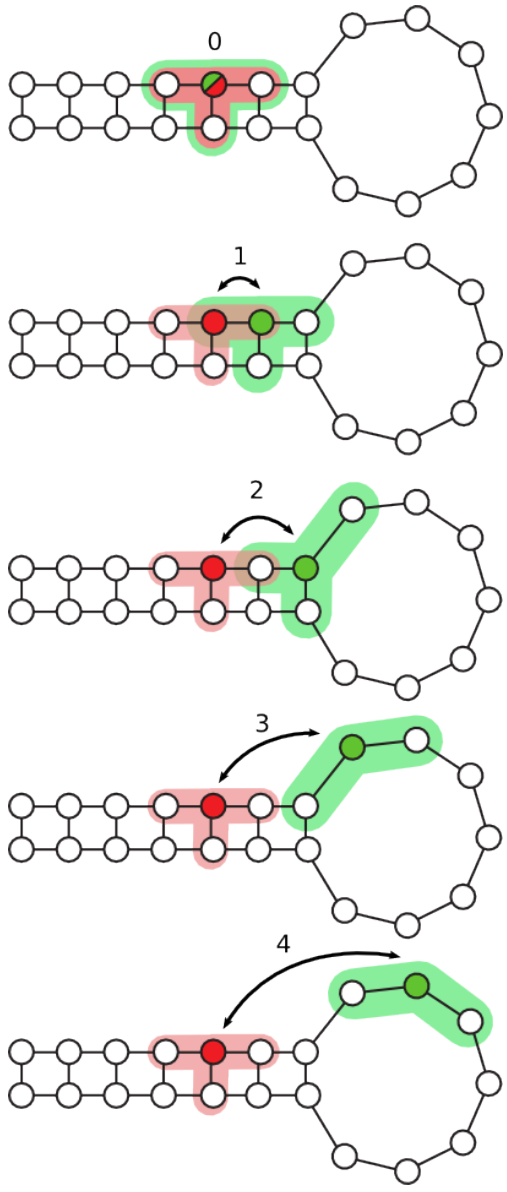
# Input graphs with RNAshapes




RNAshapes Bioinformatics, 22(4), 2006

# Features: all pairs of near small subgraphs




**Interpretation:** consider the occurrence of each subgraph in the context provided by the other subgraphs

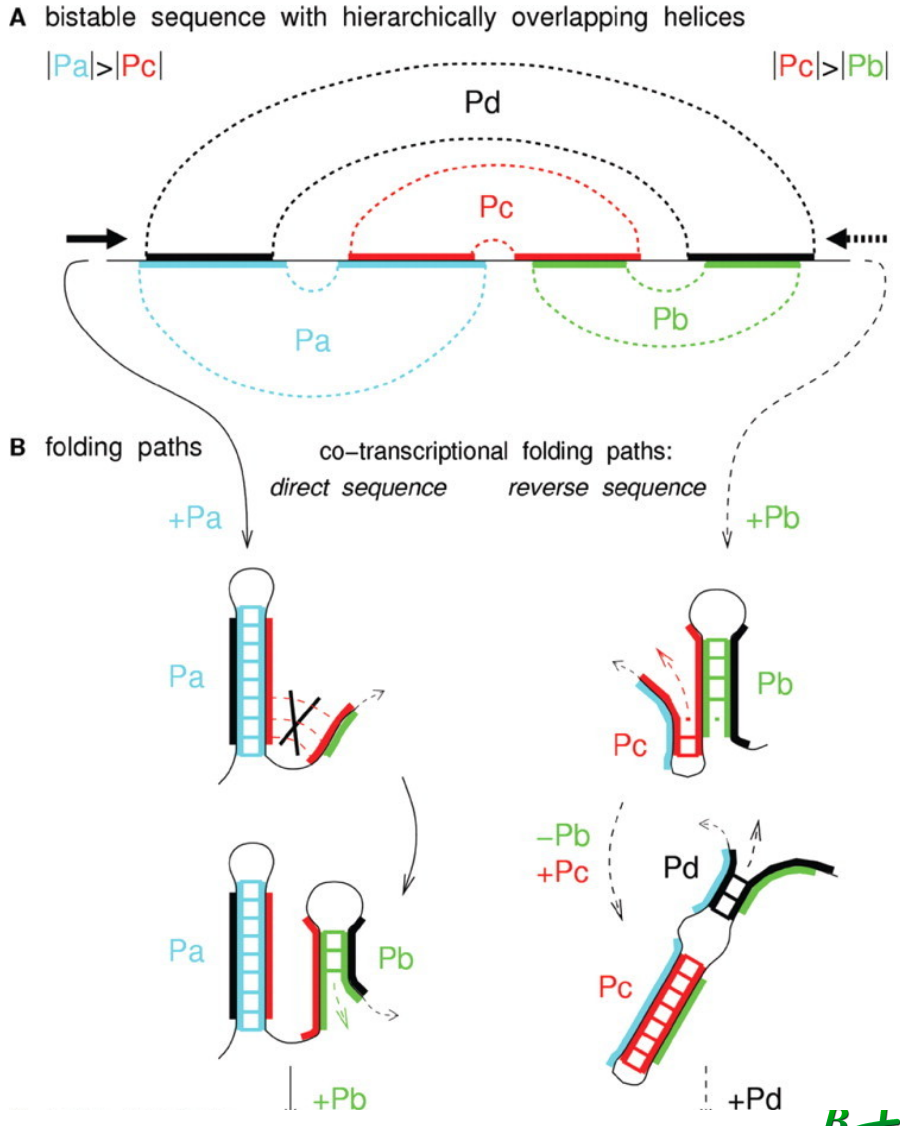
# Overview

- Motivation
- sequence motif with secondary structure properties.
- finding the structure: RNA sequence/structure alignment
- RNA classification: clustering RNA into structural classes

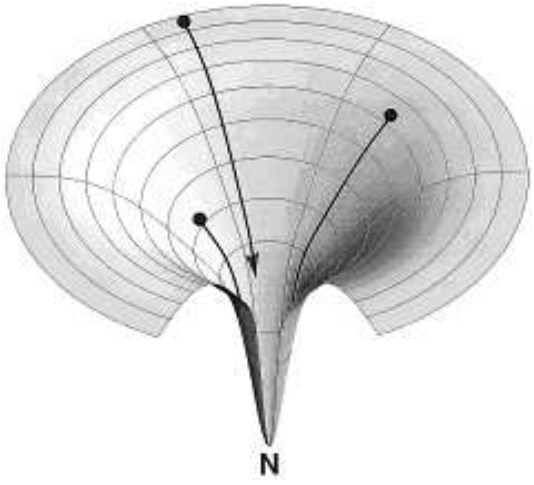
- barrier trees and HP-lattice models

# Kinetic versus Thermodynamic Folding

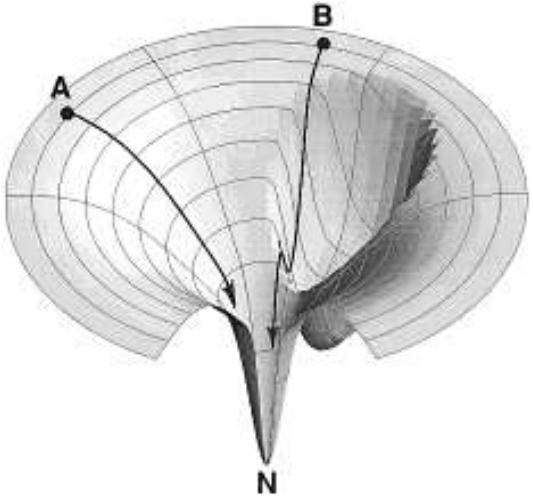
- so far: consideration of thermodynamic stable folding
  *minimum free energy*
- however: folding is a kinetic process
  *suboptimal structure favourable*
- example: co-transcriptional folding of RNA
- technique: investigation of energy landscapes



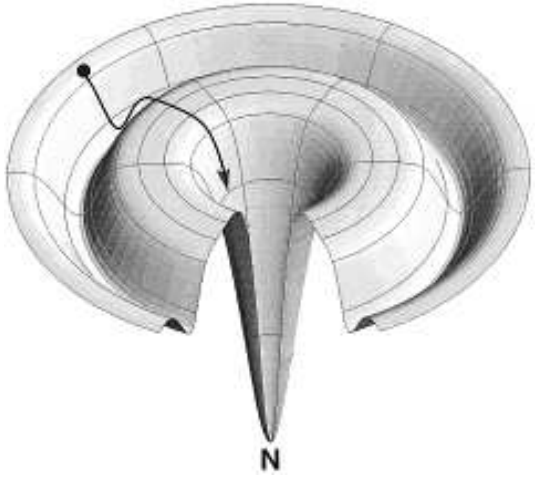
# Landscape Schemes



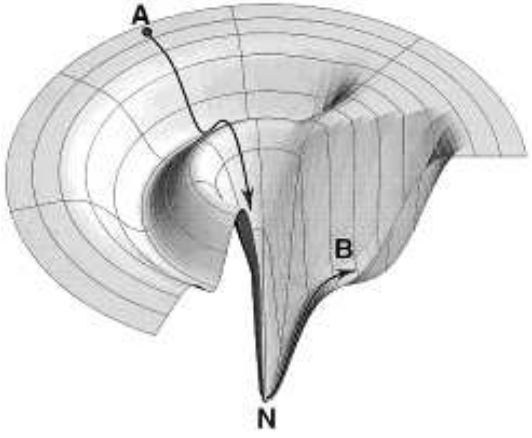

smooth



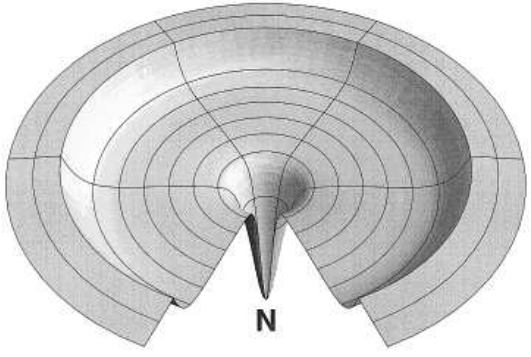

energetic trap



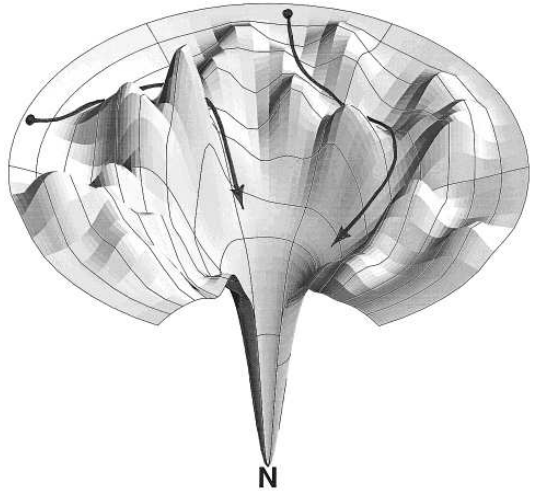

folding intermediate




unfolding trap



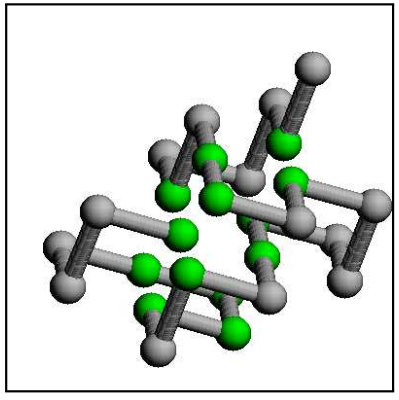

entropic trap



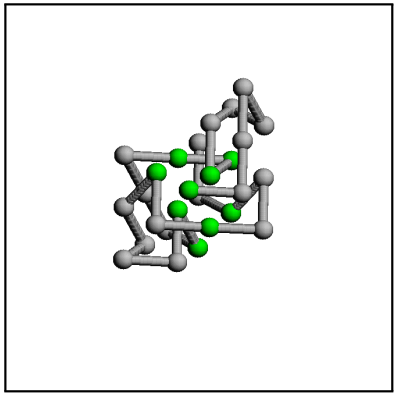

superfunnel like

# HP Lattice Model of Proteins

- only backbone structure
  positions $\hat{=}$ lattice positions

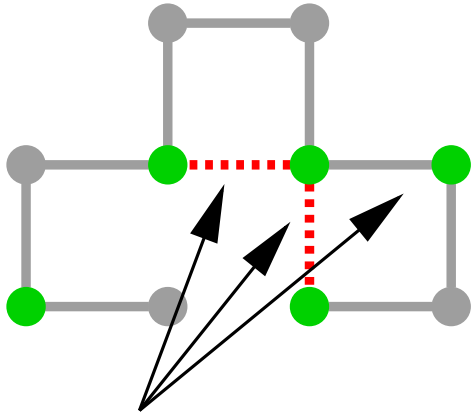


cubic



FCC

- simplified energy function:
  e.g. only hydrophobic force
  native=maximal number of HH-contacts



HH–contact

# General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

[Backofen&Will: CP2001]

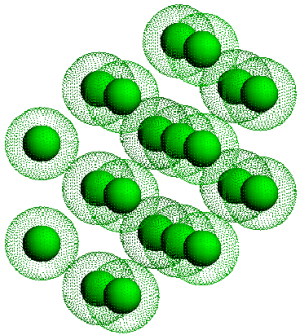
# General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

  Step 1: compute lower energy bounds

  estimate contacts (within layers, between layers)

$\Rightarrow$ Step 1



[Backofen&Will: CP2001]

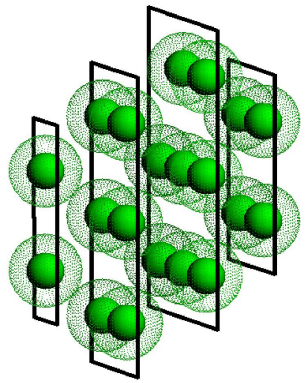
# General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

  Step 1: compute lower energy bounds

  estimate contacts (within layers, between layers)

$\Rightarrow$ Step 1




[Backofen&Will: CP2001]

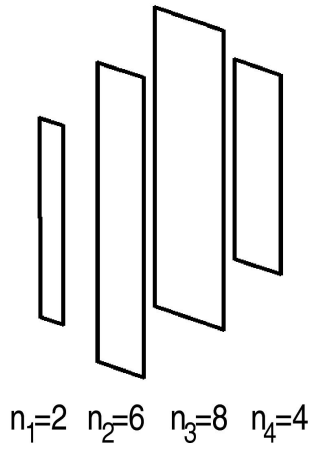
# General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

  Step 1: compute lower energy bounds

  estimate contacts (within layers, between layers)

$\overrightarrow{\text{Step 1}}$




[Backofen&Will: CP2001]

# General Approach To Lattice Folding

- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

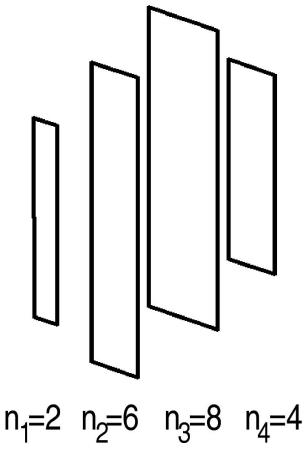
  Step 1: compute lower energy bounds

  estimate contacts (within layers, between layers)

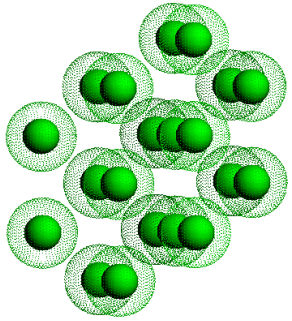
  Step 2: construct hydrophobic cores

  use bounds from last step, precomputed

$\Rightarrow$ Step 1




$\Rightarrow$ Step 2



[Backofen&Will: CP2001]

# General Approach To Lattice Folding

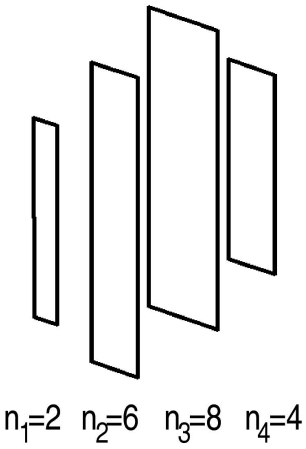
- Algorithm consist of three steps:
- Step 1 and 2 are precomputation steps

  Step 1: compute lower energy bounds
  estimate contacts (within layers, between layers)

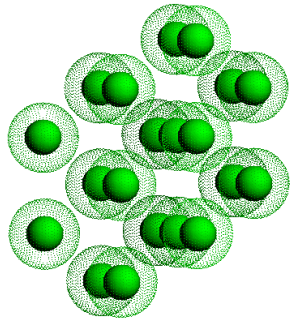
  Step 2: construct hydrophobic cores
  use bounds from last step, precomputed

  Step 3: thread sequence to hydrophobic cores of size $n$.
  using constraint propagation

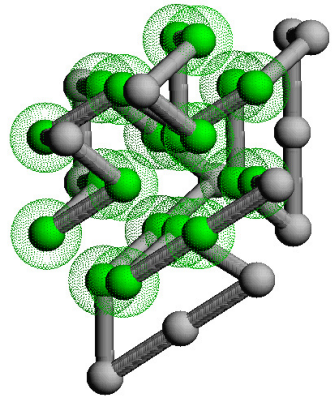
$\Rightarrow$ Step 1




$\Rightarrow$ Step 2



$\Rightarrow$ Step 3



[Backofen&Will: CP2001]

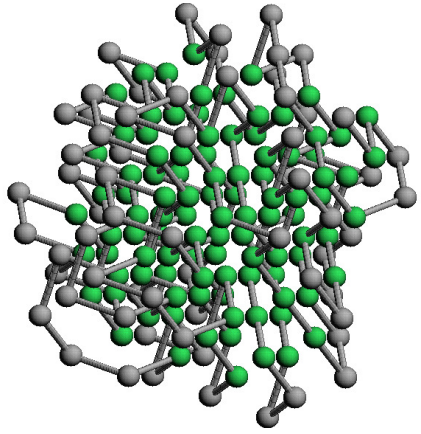
# Comparison of Results

- small selection of previous approaches:

| authors | model | dim. | maxlen | algorithm | comment |
|---|---|---|---|---|---|
| [Yue& Dill PhysRevE93] | cubic HP | 3 | 36 | branch-and-bound | optimality proven |
| [Yue&Dill PNAS95] | cubic HP | 3 | 88 | branch-and-bound | optimality proven |
| [Sazhin et al. 01] | cubic HP, FCC | 3 | 34 | branch-and-bound | not always optimal |
| [Cui et al. PNAS02] | square HP | 2 | 18 | compl. enum | |
| [Hart&Istrail JCB97] | FCC side chain | 3 | — | approximation | 86% of optimum |
| [Agarwala et al. JMB97] | FCC HP | 3 | — | approximation | $\frac{3}{5}$ of optimum |

- our results:
  - native conformation up to length **300**
  - proof of optimality
  - number of conformations of length n: $\approx 4.5^n$

    $\Rightarrow$ **search space handled $\approx 4.5^{190}$ bigger**
  - only existing non-heuristic algorithm for **FCC**



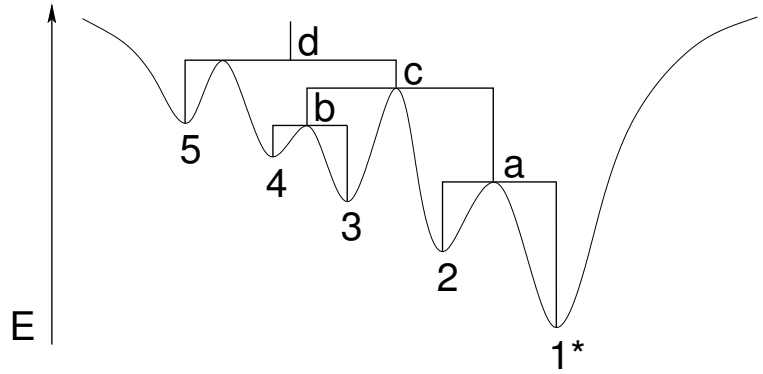
threading on 100-Hs core

| seq. | length | runtime |
|---|---|---|
| S1 | 135 | 9 s |
| S2 | 151 | 15 s |
| S3 | 161 | 18 s |
| S4 | 164 | 11 s |

[Backofen&Will: Constraint 2006]

# Investigation of Landscape

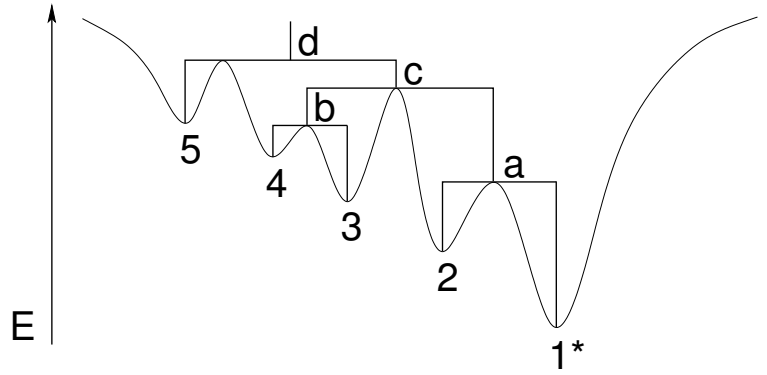
- **Goal: quantification of complexity in self-organising biomolecular systems**
  - determination of ensemble of low energy conformations
  - calculation of *barrier-trees*
  - determination of kinetic parameters



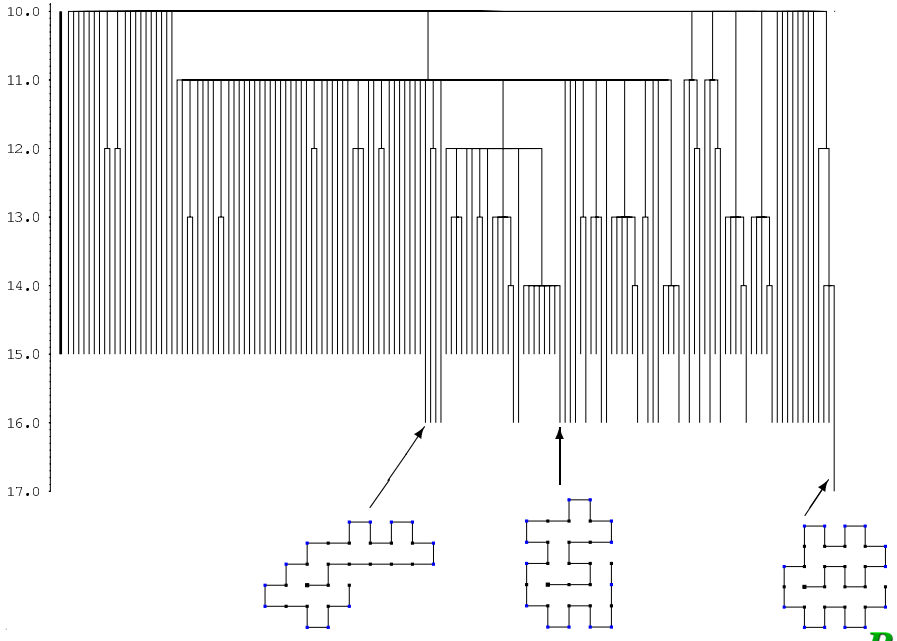

[Wolfinger et al.: Europhysics Letters 2006]

# Investigation of Landscape

- **Goal: quantification of complexity in self-organising biomolecular systems**
  - determination of ensemble of low energy conformations
  - calculation of *barrier-trees*
  - determination of kinetic parameters




- application to lattice proteins



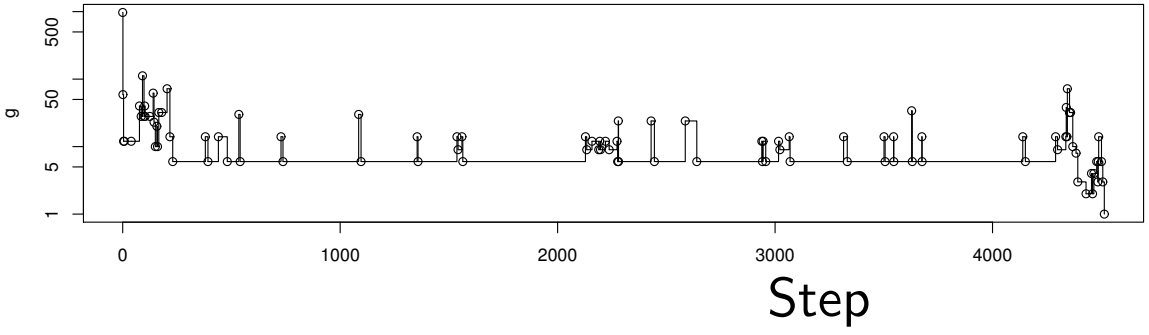

[Wolfinger et al.: Europhysics Letters 2006]

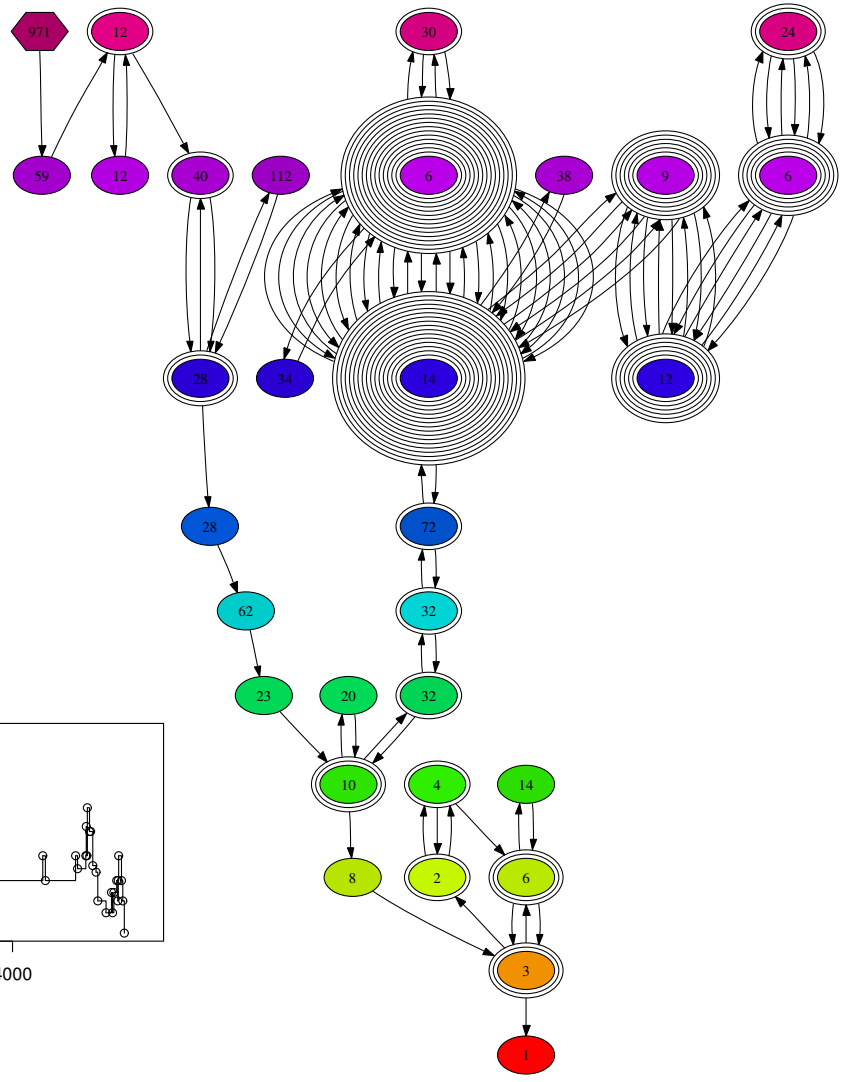
# Application: Design of protein-like Sequences

- find sequences with *exactly* *one* optimal structure
- stochastic local search

*node*: accepted sequences
*edges*: simulation step/mutation

Degeneracy



# Take-Home Messages

3 major problems in RNA

- sequence/structure motifs $\Rightarrow$ RNA-binding proteins
  `Memeris`: sequence motifs with structural properties.
- RNA comparative structure prediction
  `LocaRNA`: currently most efficient Sankoff-like approach
- RNA classification: clustering
  graph-kernel based approach

lattice model: NP-hard problems are solvable

- NP-hard doesn't mean that you cannot do it
- here: folding HP-models to optimality
- message: don't be afraid, ask your local computer scientist

# Acknowledgements


**Our group**

- Anke Busch
- Fabrizio Costa
- Steffen Heyne
- Michael Hiller
- Sita Lange
- Rileen Sinha
- Robert Kleinkauf
- Kousik Kundu
- Dominic Rose
- Andreas Richter
- Martin Mann
- Daniel Maticzka
- Mathias Möhl
- Sebastian Will

**Freiburg**

- Klaus Palme
- Claude Becker
- Wolfgang Hess
- Claudia Steglich
- Anke Becker
- ...

**Wien/Leipzig**

- Peter Stadler
- Ivo Hofacker
- Kristin Reiche

**Erlangen/U. Kentucky**

- Stefan Stamm

**MIT**

- Bonnie Berger

**SFU Vancouver**

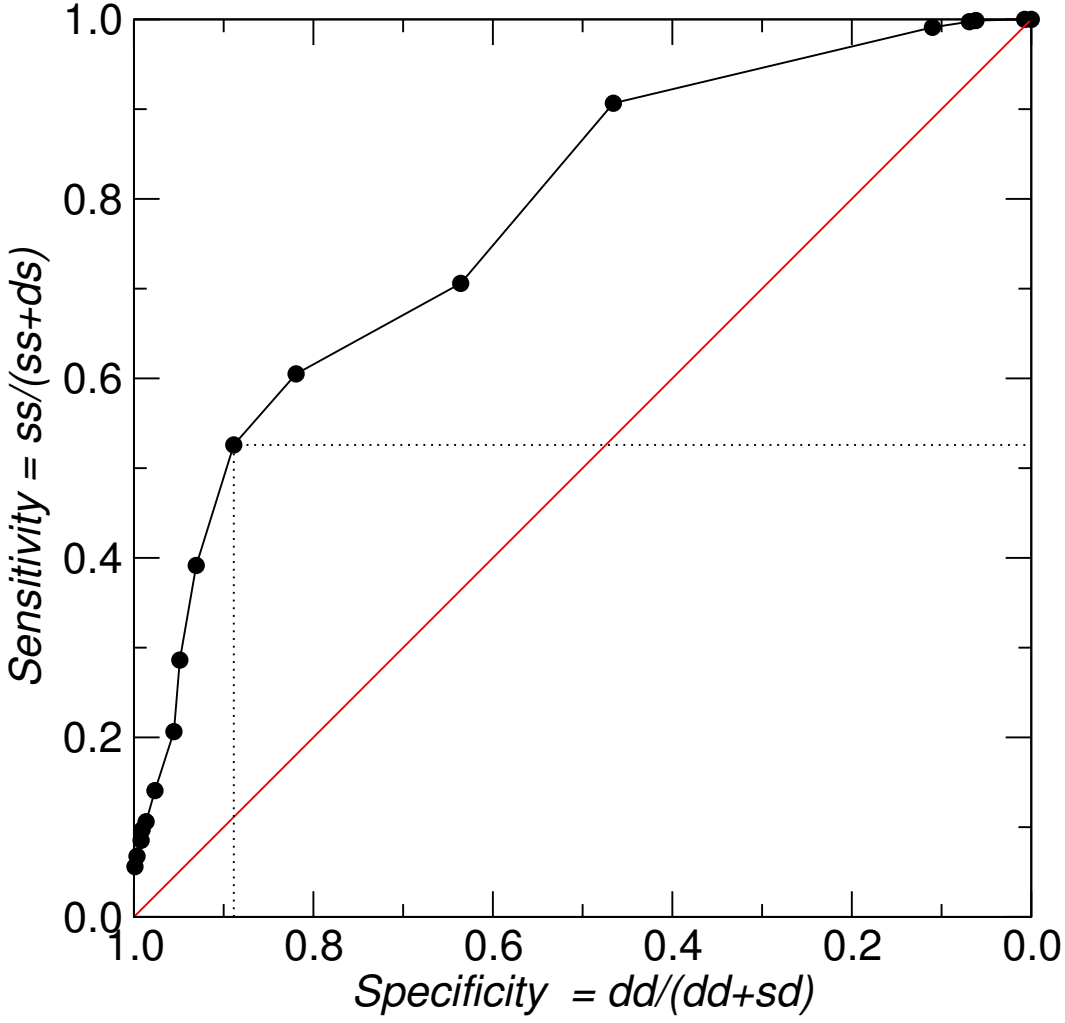
- S. Cenk Sahinalp
- Hamidreza Chitsaz
- Raheleh Salari

**DFG-Excellence Cluster "BIOSS"**
**BMBF "FRISYS"**
**DFG-SPP "regulatory bacterial RNAs"**
**DFG-SPP "InKomBio"**

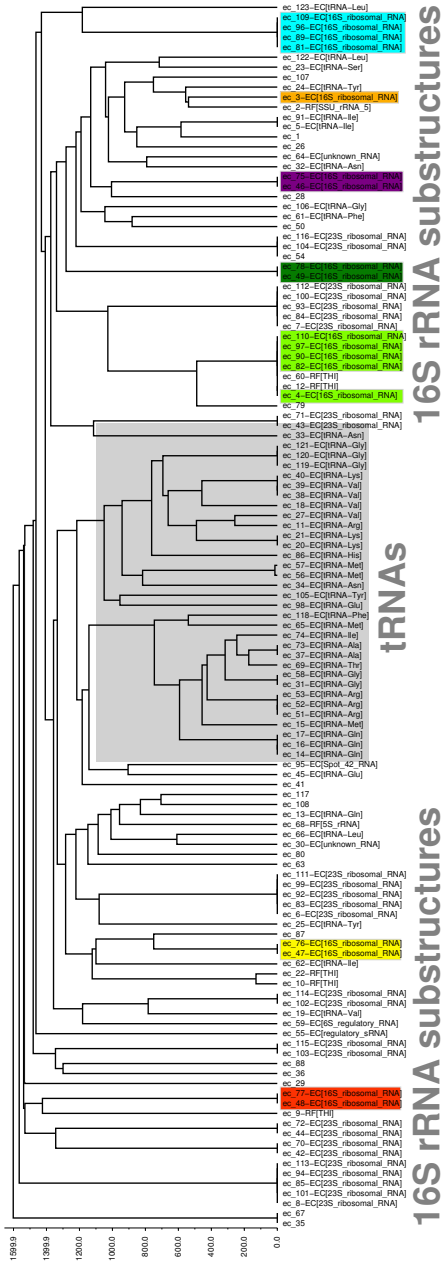
# Thank You for Your Attention

# LocARNA: Test on RFAM seed alignments

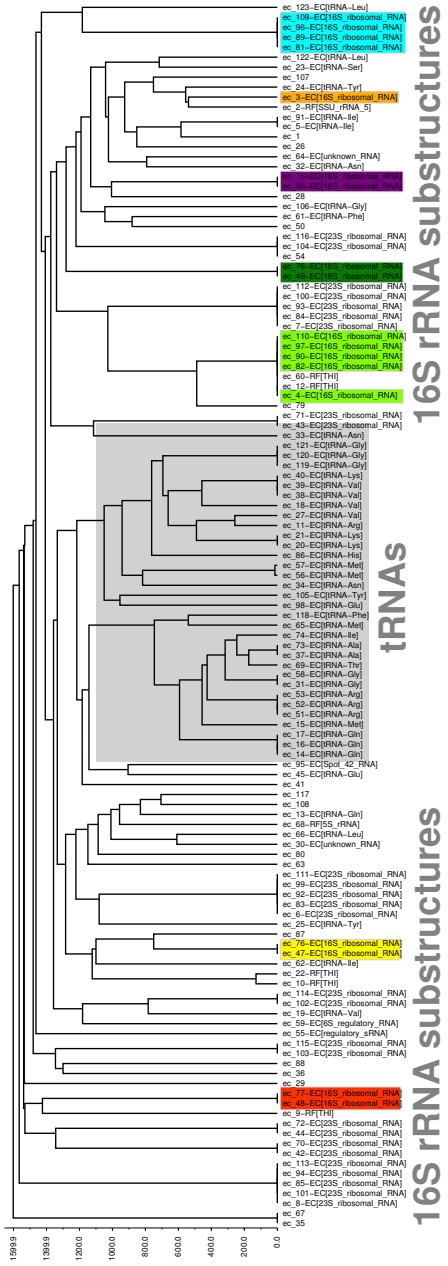
- ROC curve the global comparison of clustering and RFAM families

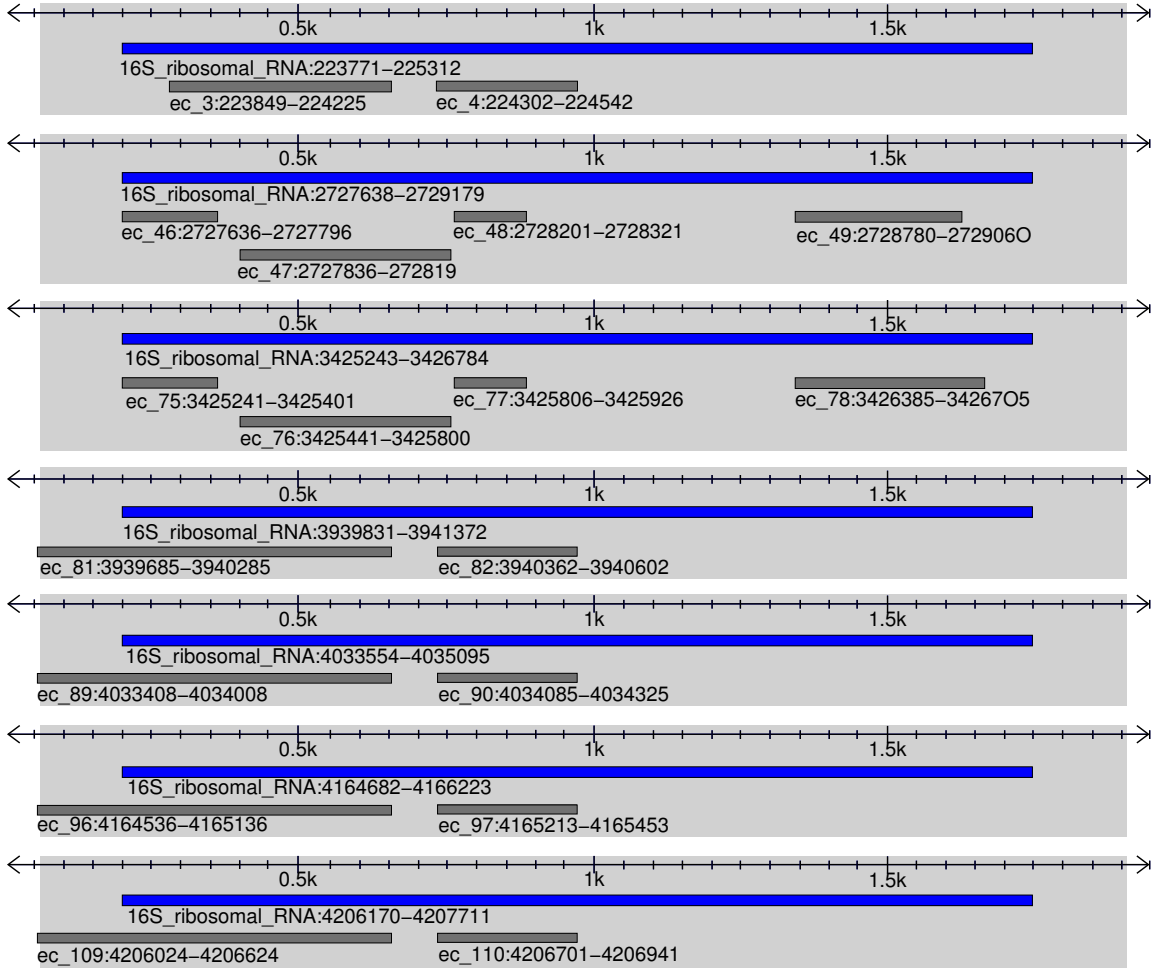


# Clustering of bacterial ncRNA predictions

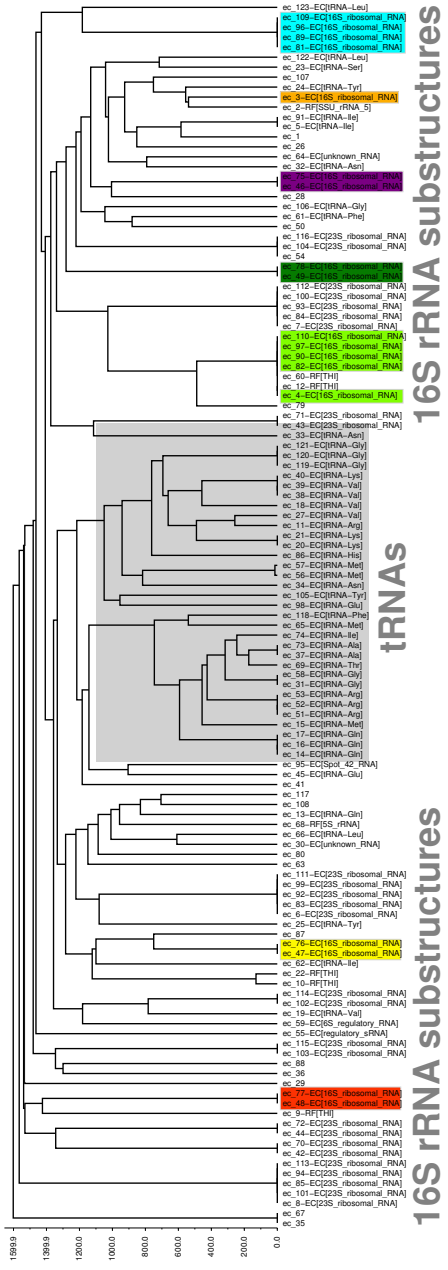


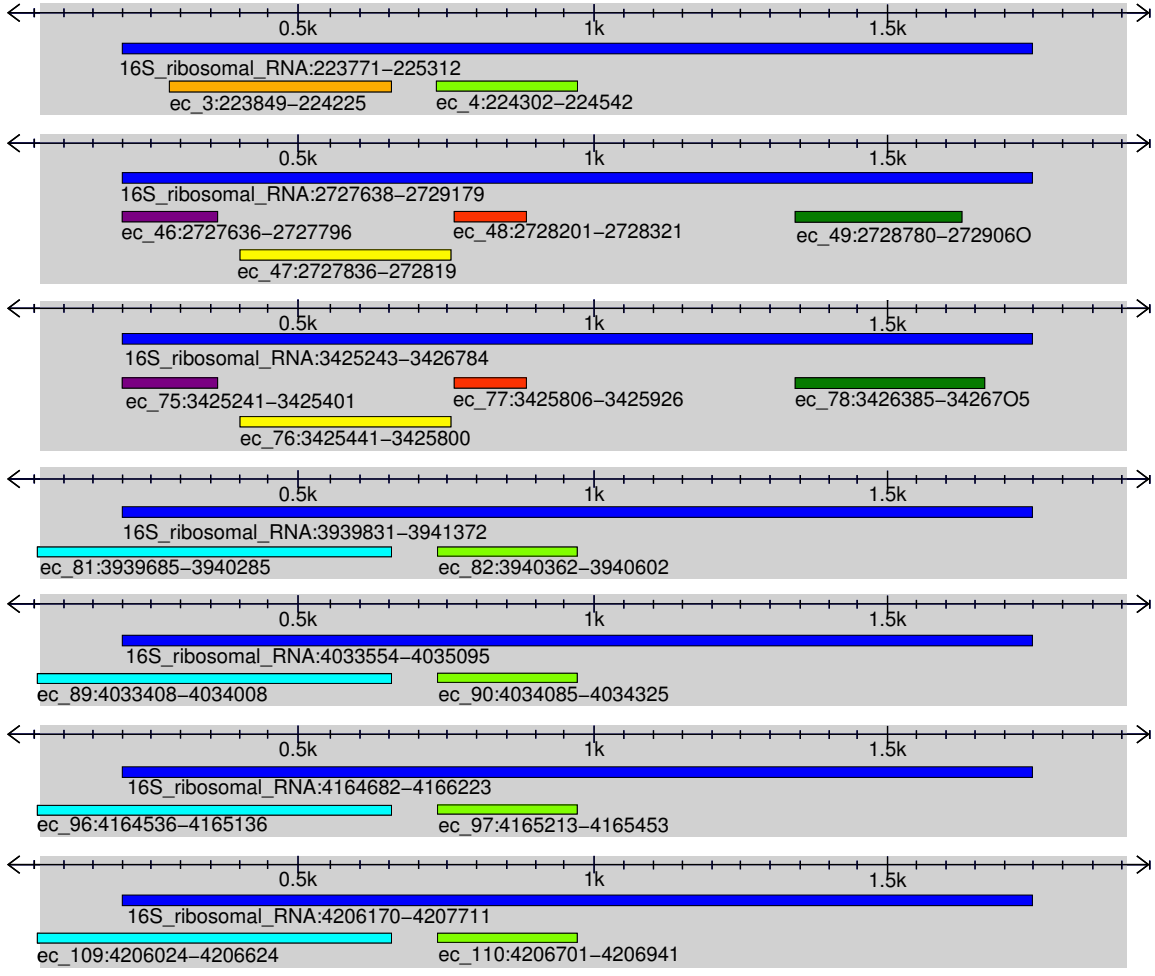
# Clustering of bacterial ncRNA predictions



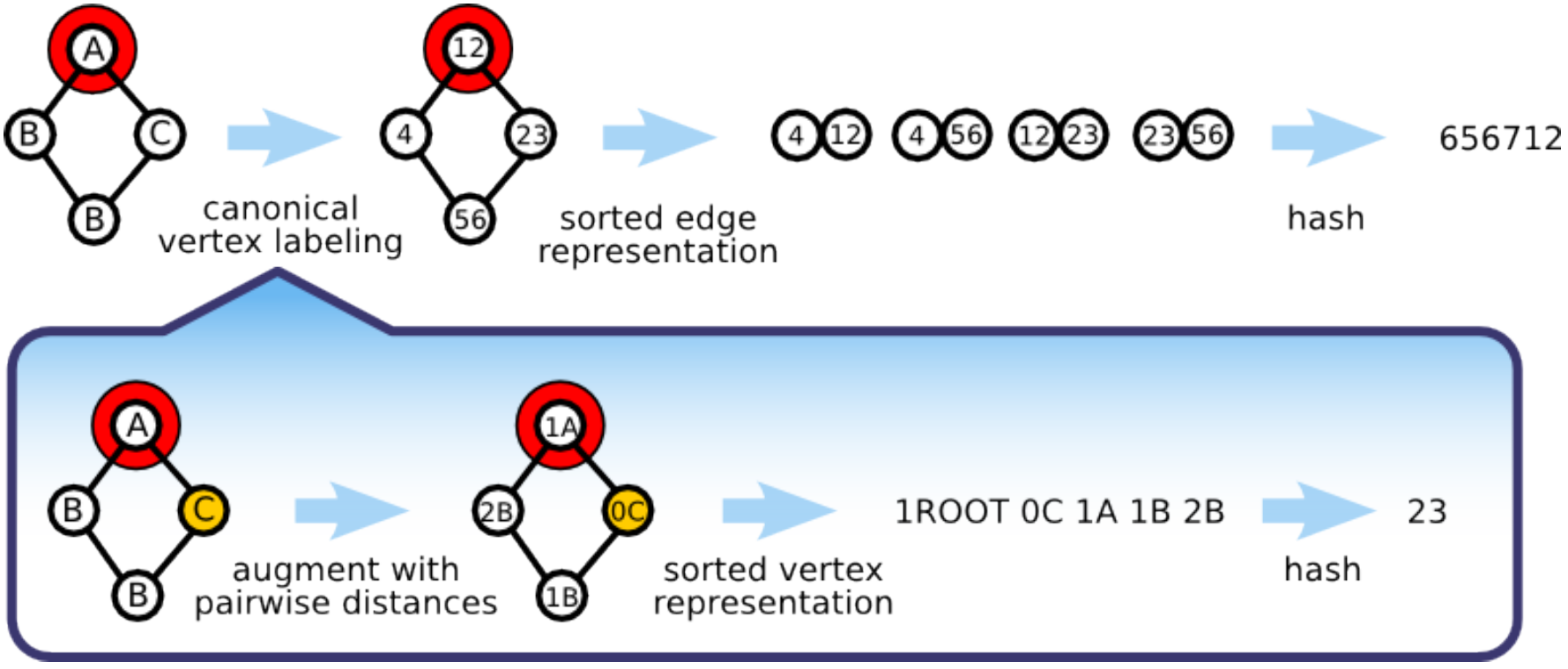
# Clustering of bacterial ncRNA predictions



# Mapping graphs into vector spaces

Given a feature (a pair of near small subgraphs) compute an integer encoding via a hashing technique




Complexity dominated by edge sorting or all-pairwise-distance computation in small subgraphs $\mapsto$ efficient (linear) in practice