

# Comparative methods for RNA structure analysis

Sebastian Will

Institute for Theoretical Chemistry, University of Vienna  
*will at tbi.univie.ac.at*

course material: <https://www.tbi.univie.ac.at/~will/AlgoSB19>

AlgoSB 2019—Day IV

# Comparative RNA Analysis—What?

- *compare* (potentially) homologous RNAs

fdhA	CGCCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAGGUGGCG
hdrA	GGCACCACUCGAAGGCUAAGCCAAAGUGGUGCU
vhuD	GUUCUCUCGGGAACCCGUCAAGGGACCGAGAGAAC
vhuU	AGCUCACAACCGAACCCAUUUGGGAGGUUGUGAGCU
fwdB	AUGUUGGAGGGGAACCCGUAAGGGACCCUCCAAGAU
selD	UUACGAUGUGCCGAACCCUUUAAGGGAGGCACAUCGAAA
fruA	CCUCGAGGGGAACCCGAAAGGGACCCGAGAGG

# Comparative RNA Analysis—What?

- *compare* (potentially) homologous RNAs

```
fdhA    CGCCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAGGUGGCCG
hdrA    GGCACCACUCGAAGGCUAAGCCAAAGUGGUGCU
vhuD    GUUCUCUCGGGAACCCGUCAAGGGACCGAGAGAAC
vhuU    AGCUCACAACCGAACCCAUUUGGGAGGUUGUGAGCU
fwdB    AUGUUGGAGGGGAACCCGUAAGGGACCCUCCAAGAU
selD    UUACGAUGUGCCGAACCCUUUAAGGGAGGCACAUCGAAA
fruA    CCUCGAGGGGAACCCGAAAGGGACCCGAGAGG
```

- *align*

```
fdhA    CGC-CACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-GUGG-CG
hdrA    GGC-ACC-ACUCGAAGGCU-----AAGCCAAAGU-GGUG-CU
vhuD    GUU-CUC-UCGGGAACCCGU-----CAAGGGACCGA-GAGA-AC
vhuU    AGC-UCACAACCGAACCCAU-----UUGGGAGGUUGUGAG-CU
fwdB    AUG-UUGGAGGGGAACCCGU-----AAGGGACCCUCCAAG-AU
selD    UUACGAUGUGCCGAACCCUU-----UAAGGGAGGCACAUCGAAA
fruA    CC-UCG--AGGGGAACCCGA-----AAGGGACCC--GAGA-GG
```

# Comparative RNA Analysis—What?

- *compare* (potentially) homologous RNAs

```
fdhA    CGCCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAGGUGGCCG
hdrA    GGCACCACUCGAAGGCUAAGCCAAAGUGGUGCU
vhuD    GUUCUCUCGGGAACCCGUCAAGGGACCGAGAGAAC
vhuU    AGCUCACAACCGAACCCAUUUGGGAGGUUGUGAGCU
fwdB    AUGUUGGAGGGGAACCCGUAAGGGACCCUCCAAGAU
selD    UUACGAUGUGCCGAACCCUUUAAGGGAGGCACAUCGAAA
fruA    CCUCGAGGGGAACCCGAAAGGGACCCGAGAGG
```

- *align*

```
fdhA    CGC-CACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-GUGG-CG
hdrA    GGC-ACC-ACUCGAAGGCU-----AAGCCAAAGU-GGUG-CU
vhuD    GUU-CUC-UCGGGAACCCGU-----CAAGGGACCGA-GAGA-AC
vhuU    AGC-UCACAACCGAACCCAU-----UUGGGAGGUUGUGAG-CU
fwdB    AUG-UUGGAGGGGAACCCGU-----AAGGGACCCUCCAAG-AU
selD    UUACGAUGUGCCGAACCCUU-----UAAGGGAGGCACAUCGAAA
fruA    CC-UCG--AGGGGAACCCGA-----AAGGGACCC--GAGA-GG
```

- consider and learn about *RNA structure*

```
AGC_CAC_AGGCGAACCCGU_____AAGGGACCCU_GAGG_AU
((..(((((((.....)))))))))..)) (-19.48)
```

## Comparative RNA Analysis—Why?

- *overcome limitations* of prediction from single sequences

Program	Sens	PPV	MCC	F-measure
RNAfold 2.1.9	0.742	0.795	0.767	0.765
UNAFold 3.8	0.693	0.767	0.727	0.725
RNAstructure 5.7	0.716	0.781	0.746	0.744

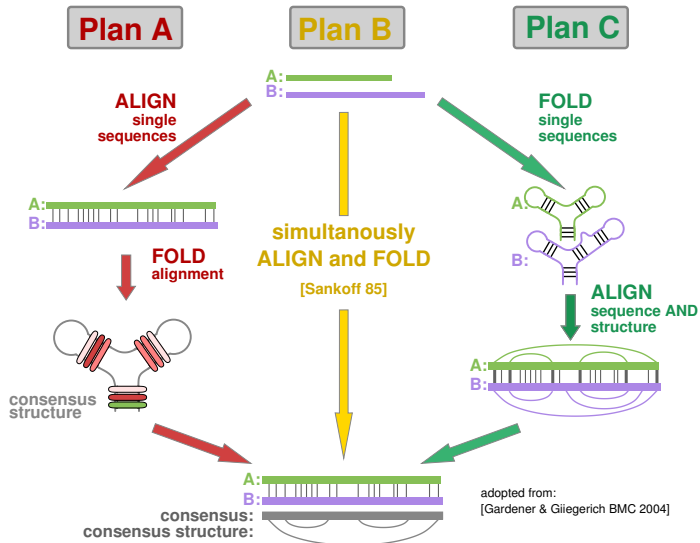
- single sequence stability does not help for *ncRNA gene finding*:  
“... *in general, the predicted stability of structural RNAs is not sufficiently distinguishable from the predicted stability of random sequences*”<sup>1</sup>
- pure sequence alignment cannot properly *compare remote RNAs*  
“... *sequence alignment alone, using the current algorithms, is generally inappropriate <50–60% sequence identity.*”<sup>2</sup>

---

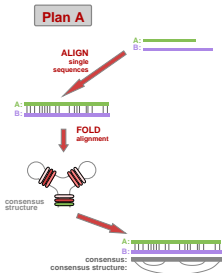
<sup>1</sup>Rivas, Eddy; 2001; [doi:10.1186/1471-2105-2-8](https://doi.org/10.1186/1471-2105-2-8)

<sup>2</sup>Gardner, Wilm, Washietl; 2005; [doi:10.1093/nar/gki541](https://doi.org/10.1093/nar/gki541)

# Comparative RNA Analysis—How?

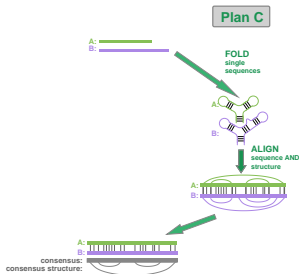


# ALIGN, then ANALYSE



- Covariation, R2R
- R-scape
- Pfold
- RNAalifold
- RNAz
- CMs, SCFGs, Infernal

# FOLD, then ALIGN

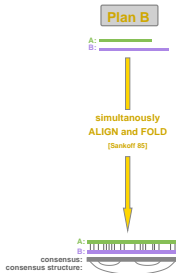


- RNAforester, MARNA
- noteworthy, algorithmically interesting (e.g. tree alignment vs. tree editing), ...

... but neglected here for time constraints :(

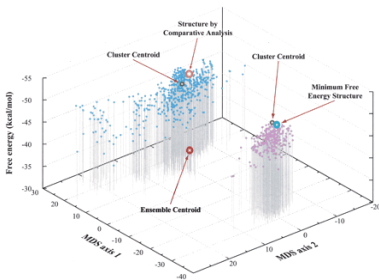


# Simultaneous ALIGN and FOLD



- The classic: Sankoff simultaneous alignment and folding (SA&F)
- “Gold standard” for RNA comparison
- Heuristic short cuts: STRAL, TurboFold II
- Sankoff-style: Dynalign, stemloc, Foldalign
- Fast SA&F (PMcomp-style): PMcomp, LocARNA, RAF, LocARNA-P, SPARSE

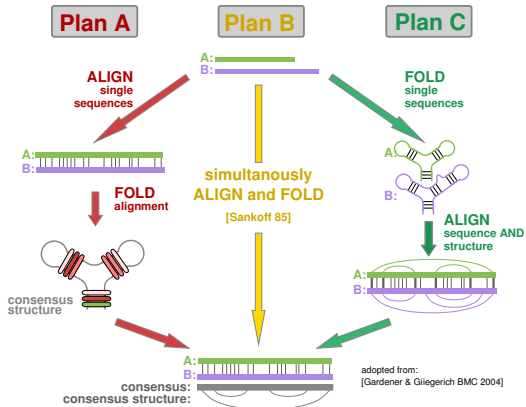
# Clustering



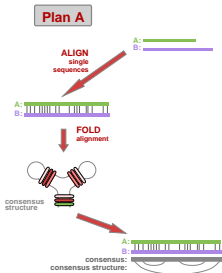
- clustering structures of one RNA<sup>3</sup>
- structure-based clustering of RNAs (RNAclust, GraphClust)

---

<sup>3</sup>e.g. Ding et al.; RNA 2005; [doi:10.1261/rna.2500605](https://doi.org/10.1261/rna.2500605)



# ALIGN, then ANALYSE



- Covariation, R2R
- R-scape
- Pfold
- RNAalifold
- RNAz
- CMs, SCFGs, Infernal

## Covariation hints at structure

- Functional RNAs are under selective pressure to preserve their secondary structure
- → Mutations must be compensated! (or wobble)

```
..((.....))..  
auGCaugaGCuc  
auCCaugaGGuc  
auCGaugaCGuc  
auUGaugaCGuc
```

- Inversely: compensatory mutations hint at functional structure

# Measuring Covariation: Mutual Information

123456789012  
..((.....))..  
auGCaugaGCuc  
auCCaugaGGuc  
auCGaugaCGuc

*Mutual Information* (of columns  $i$  and  $j$ ):

$$MI_{i,j} = \sum_{a,b \in \{A,C,G,U\}} f_{i,j}(ab) \log_2 \frac{f_{i,j}(ab)}{f_i(a)f_j(b)}$$

[aka *relative entropy*, *Kullback-Leibler divergence*]

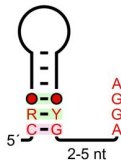
- $M_{1,12} = f_{1,12}(AC) \log \frac{f_{1,12}(AC)}{f_1(A)f_{12}C} = 1 \log 1 = 0$
- $M_{4,9} = f_{4,9}(CG) \log \frac{f_{4,9}(CG)}{f_4(C)f_9(G)} + f_{4,9}(GC) \log \frac{f_{4,9}(GC)}{f_4(G)f_9(C)}$   
 $\approx 0.66 \log 0.66/0.22 + 0.33 \log 0.33/0.22 \approx 0.86$

convention: "0 log 0 = 0"

# Covariation in Consensus Structure Visualization

```
# STOCKHOLM 1.0
martian      CAGGGAAACUGAUUUUAGGA
venusian    CGU.UUCG.ACGUA...AGGA
#=GC SS_cons <<<<.....>>>>.....
#=GC R2R_LABEL ...[.....]...1...2T...
#=GF R2R var_hairpin [ ]
#=GF R2R var_backbone_range 1 2
#=GF R2R turn_ss T -90
//
```

label & use

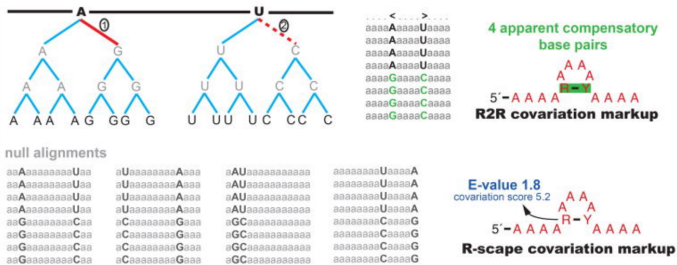


Visualizations created by the RNA drawing tool *R2R*<sup>4</sup>  
Covarying mutations are highlighted (green-ish)

<sup>4</sup>Weinberg, Breaker; 2011; [doi:10.1186/1471-2105-12-3](https://doi.org/10.1186/1471-2105-12-3)

# Significance of covariation in R-scape<sup>5</sup>

Independent positions show apparent covariation due to phylogeny



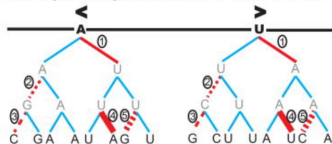
- to generate null model: estimate tree, then shuffle mutations
- in shuffled alignment make exactly the same mutations at same branches at random sequence positions
- preserves composition and substitutions, scrambles dependencies
- Overcomes problem of 'apparent' covariation, but destroys local conservation

<sup>5</sup>Rivas, Clements, Eddy. 2017. [doi:10.1038/nmeth.4066](https://doi.org/10.1038/nmeth.4066)



# Significance of covariation in R-scape<sup>5</sup>

Base paired positions show covariation due to structure



```

.....<.....>.....
aaaaCaaaaGaaaa
aaaaGaaaaCaaaa
aaaaAaaaaUaaaa
aaaaAaaaaUaaaa
aaaaUaaaaAaaaa
aaaaAaaaaUaaaa
aaaaGaaaaCaaaa
aaaaUaaaaAaaaa

```

5 apparent compensatory base pairs



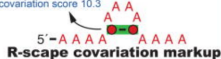
null alignments

```

aaaaaaaaCaaGa   aaaaGaaaaCaaaa   aCaaaaaaaaGaaaa   aaaaaaaaaGaaCaa
aaaaaaaaGaaCa   aaaaCaaaaGaaaa   aCaaaaaaaaGaaaa   aaaaaaaaaCaGaa
aaaAaaaUaaaaaa   aaUaaaaAaaaaaa   AaaaaUaaaaaaaa   aaaaUaAaaaaaa
aaAaaaUaaaaaa   aaUaaaaAaaaaaa   AaaaaUaaaaaaaa   aaaaUaAaaaaaa
aaaaaaaaAaaUaa   UaaaaaaAaaaaaa   aaaaaaaaaAaUaa   aaaaaaaaaAaUaa
aaaaaaaaUaaAaa   AaaaUaaaaaaaa   aaaaaaaaaUaAaa   aaaaaaaaaUaAaa
aaaaaaaaCaaGaa   GaaaaaaCaaaaaa   aaaaaaaaaCGAa   aaaaaaaaaCaaGaa
aaaaaaaaAaaUaa   UaaaaaaAaaaaaa   aaaaaaaaaAaUaa   aaaaaaaaaAaUaa

```

E-value  $9.4 \cdot 10^{-4}$   
covariation score 10.3



- to generate null model: estimate tree, then shuffle mutations
- in shuffled alignment make exactly the same mutations at same branches at random sequence positions
- preserves composition and substitutions, scrambles dependencies
- Overcomes problem of 'apparent' covariation, but destroys local conservation

<sup>5</sup>Rivas, Clements, Eddy. 2017. [doi:10.1038/nmeth.4066](https://doi.org/10.1038/nmeth.4066)

# Covariation and Thermodynamics: RNAalifold<sup>6</sup>

AF008220 GGAGGAUU-AGCUCAGCUGGGAGAGCAUCUGCCUACAAGC-----AGAGGGUCGGCGGUUCGAGCCCAGUACCCUCA  
M68929 GCGGAUUAU-AACUUAGGGGUAAAAGUUGCAGAUUGGGCUC-----UGAAAA-CACGGGUUCGAAUCCCGUUAUUCGCC  
X02172 GCCUUUAU-AGCUUAG-UGGUAAAAGCGAUAAACUGAAGAUU-----UAUUUACAUGUAGUUCGAUUCUCAUUAAGGGCA  
Z11880 GCCUUCU-AGCUCAG-UGGUAGAGCGCACGGCUUUUAACC-----GUGUGGUCGUGGGUUCGAUCCCCACGGAAGGCG  
D10744 GGAAAAUUGAUCAUCGGCAAGAUAAAGUUAUUUACUAAAAAUAGGAUUUAAUAAACCUGGUGAGUUCGAAUCUCAUUAUUCCG

---

<sup>6</sup>Bernhart et al. 2008. *doi:10.1186/1471-2105-9-474*

# Covariation and Thermodynamics: RNAalifold<sup>6</sup>

```
AF008220  GGAGGAUU-AGCUCAGCUGGGAGAGCAUCUGCCUACAAGC-----AGAGGGUCGGGGUUCGAGCCCAGUACUCCUCA
M68929   GCGGAUUAU-AACUUAGGGGUUAAAAGUUGCAGAUUGGGCUC-----UGAAAAA-CACGGGUUCGAAUCCCGUUAUUCGCC
X02172   GCCUUUAU-AGCUUAG-UGGUAAAAGCGAUAAACUGAAGAUU-----UAUUUACAUGUAGUUCGAUUCUCAUUAAGGGCA
Z11880   GCCUUCU-AGCUCAG-UGGUAGAGCGCACGGCUUUUAACC-----GUGUGGUCGUGGGUUCGAUCCCCACGGAAGGCG
D10744   GGAAAAUUGAUCAUCGGCAAGAUAAAGUUAUUUACUAAAAAUAGGAUUUAAUAAACCGGUGAGUUCGAAUCUCACAUUUUCCG
```

```
alifold  (((((((((...(((.....))))((((.....)).....))))....((((.....)))))))))).
```

(-49.58 = -17.46 + -32.12)

Predict consensus structure that is

- thermodynamically good
- ideally possible for all sequences (tolerate defects)
- supported by covariation

---

<sup>6</sup>Bernhart et al. 2008. [doi:10.1186/1471-2105-9-474](https://doi.org/10.1186/1471-2105-9-474)

# RNAalifold—or how to fold an alignment

*Given:* a multiple alignment

*Goal:* predict the (non-crossing) consensus structure of the alignment

# RNAalifold—or how to fold an alignment

*Given:* a multiple alignment

*Goal:* predict the (non-crossing) consensus structure of the alignment

*Trick:* alignment = *sequence* of columns

*Algorithmic ideas:*

- The optimal consensus structure minimizes a combination of
  - free energies for all the RNA sequences and
  - the conservation score (= evidence for base pairing).
- Since the consensus structure pairs columns and is non-crossing, its prediction works similar to the Zuker algorithm

## RNAalifold Recursions

$$F_{ij} = \min\{F_{ij-1}; \min_{i \leq k < j-m} F_{ik-1} + C_{kj}\}$$

$$C_{ij} = \beta\gamma(i, j)$$

$$+ \min \left\{ \begin{array}{l} \sum_{1 \leq \ell \leq K} \mathcal{H}_\ell(i, j) \\ \min_{i < i' < j' < j} \sum_{1 \leq \ell \leq K} C_{i'j'} + \mathcal{I}_\ell(i, j, i', j') \\ \min_{i < k < j} M_{i+1k} + M_{k+1j-1} + aK \end{array} \right.$$

$$M_{ij} = \min \left\{ \begin{array}{l} M_{ij-1} + cK; M_{i+1j} + cK; C_{ij} + bK \\ \min_{i < k < j} M_{ik} + M_{k+1j} \end{array} \right.$$

$\mathcal{H}_\ell(i, j)$  and  $\mathcal{I}_\ell(i, j, i', j')$ : energy contributions for  $\ell$ -th sequence.

Note: RNAalifold implements an unambiguous variant.

# RNAalifold Conservation Score

*conservation score*  $\gamma(i,j) = \text{covariation boni} + \text{penalties}$

*covariation boni:*

for each pair of sequences, where columns  $i$  and  $j$  could base pair:  
average hamming distances of left ends and right ends

*penalties:*

for each sequence:

if entries in columns  $i$  and  $j$

- are non-complementary bases:  $\delta$
- are one base and one gap:  $\delta$
- are both gaps:  $0.25\delta$





## Structure conservation

*Recall:* Given an alignment, RNAalifold computes the MFE (including conservation score) of any consensus structure

*Question:* Is there a truly-conserved consensus structure?

## Structure conservation

*Recall:* Given an alignment, RNAalifold computes the MFE (including conservation score) of any consensus structure

*Question:* Is there a truly-conserved consensus structure?

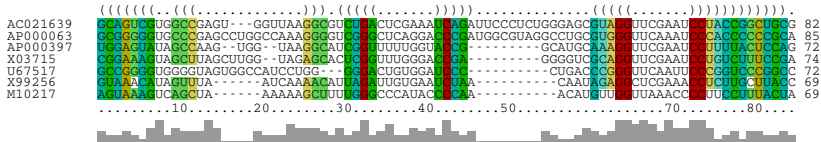
*Requires to put RNAalifold's MFE into relation! Is it as large as the average single sequence MFE's (from RNAfold)?*

*Structure Conservation Index (SCI)*

of alignment  $\mathcal{A}$  of  $K$  sequences  $S_i$

$$SCI(\mathcal{A}) := \frac{MFE_{\text{alifold}}(\mathcal{A})}{\text{mean}_i[MFE(S_i)]}$$

# SCI Example



*Single MFEs* (RNAfold): -31.20, -52.80, -22.00, -28.90, -35.60, -13.90, -13.90

*Consensus MFE* (RNAalifold): -25.67 (e -18.15, cons -7.52)

*Structure conservation index (SCI)*:

$$\frac{-25.6}{\text{mean}(-31.20, -52.80, -22.00, -28.90, -35.60, -13.90, -13.90)} = \frac{-25.6}{-28.33} \approx 0.91$$

## De novo ncRNA prediction—RNAz<sup>7</sup>

*Question:* Given alignment, is there an ncRNA?

- is there a truly conserved structure?
- can the single sequences form stable structures?

---

<sup>7</sup>Washietl, Hofacker, Stadler. 2005. [doi:10.1073/pnas.0409169102](https://doi.org/10.1073/pnas.0409169102)

## De novo ncRNA prediction—RNAz<sup>7</sup>

*Question:* Given alignment, is there an ncRNA?

- is there a truly conserved structure?
  - significance of structure conservation (SCI)
- can the single sequences form stable structures?
  - significance of stabilities (MFEs)

---

<sup>7</sup>Washietl, Hofacker, Stadler. 2005. [doi:10.1073/pnas.0409169102](https://doi.org/10.1073/pnas.0409169102)

## De novo ncRNA prediction—RNAz<sup>7</sup>

*Question:* Given alignment, is there an ncRNA?

- is there a truly conserved structure?
  - significance of structure conservation (SCI)
- can the single sequences form stable structures?
  - significance of stabilities (MFEs)

RNAz evaluates alignment by

- computing SCI
- estimating Z-scores of MFEs (in relation to seq. composition)
- relating them to each other and alignment entropy

---

<sup>7</sup>Washietl, Hofacker, Stadler. 2005. [doi:10.1073/pnas.0409169102](https://doi.org/10.1073/pnas.0409169102)

## De novo ncRNA prediction—RNAz<sup>7</sup>

*Question:* Given alignment, is there an ncRNA?

- is there a truly conserved structure?
  - significance of structure conservation (SCI)
- can the single sequences form stable structures?
  - significance of stabilities (MFEs)

RNAz evaluates alignment by

- computing SCI
- estimating Z-scores of MFEs (in relation to seq. composition)
- relating them to each other and alignment entropy

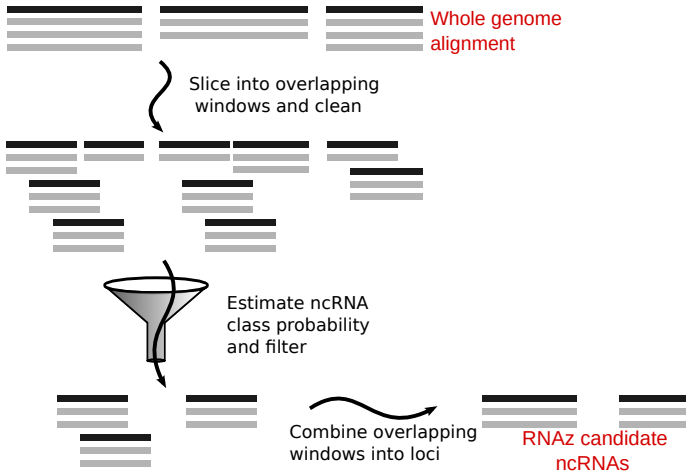
For high efficiency

- the MFE Z-scores are estimated after function learning from pre-computed distributions (SVM-based)
- combination via trained SVM

---


<sup>7</sup>Washietl, Hofacker, Stadler. 2005. [doi:10.1073/pnas.0409169102](https://doi.org/10.1073/pnas.0409169102)

# RNAz Screen





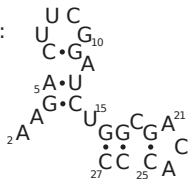
# RNA families: CMs—Infernal<sup>8</sup>—Rfam<sup>9</sup>

- *Infernal*: characterize RNA family and fast search for members  
 Inference of RNA alignments
- fundamental for *Rfam* (database of RNA families)  
 Rfam 14.0 (August 2018, 2791 families)  
 'hand-curated' seed alignments  $\Rightarrow$  *Infernal* full alignments
- models RNA families by *Stochastic Context Free Grammars (SCFGs)* as *Consensus Models (CMs)*

input multiple alignment:

[structure]	. . . <<<	>- >>:	<<- <.	. >>> .
human	. AAGACUUCGGAUCUGGCG .	ACA . CCC .		
mouse	aUACACUUCGGAUG - CACC .	AAA . GUGa		
orc	. AGGUCUUC - GCACGGGCAgCCA cUUC .			
	1            5            10            15            20            25            28			

example structure:



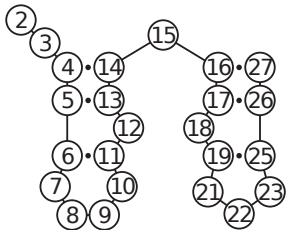
<sup>8</sup>Nawrocki, Eddy. 2013. [doi:10.1093/bioinformatics/btt509](https://doi.org/10.1093/bioinformatics/btt509)

<sup>9</sup><http://rfam.xfam.org/>

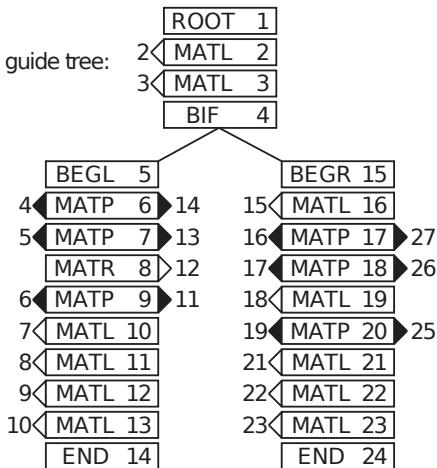
# Infernal Consensus Models (CMs)

- CMs are grammatical description of RNA families
- learn transition and output probabilities from alignment
- CMs extend profile HMMs (Pfam)

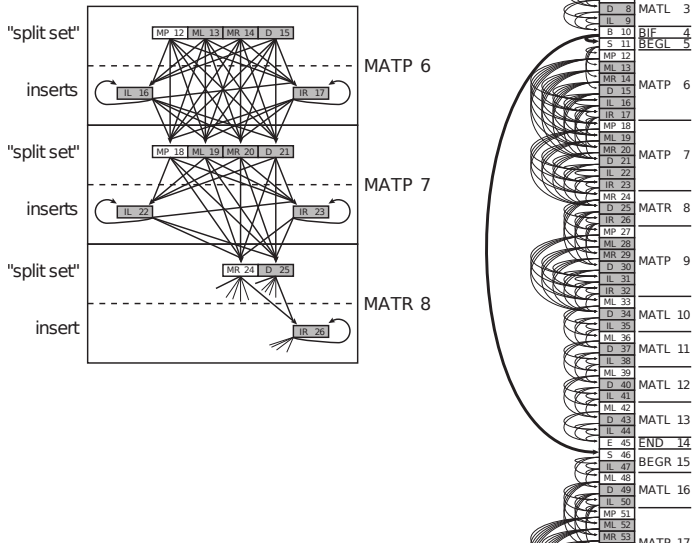
consensus structure:



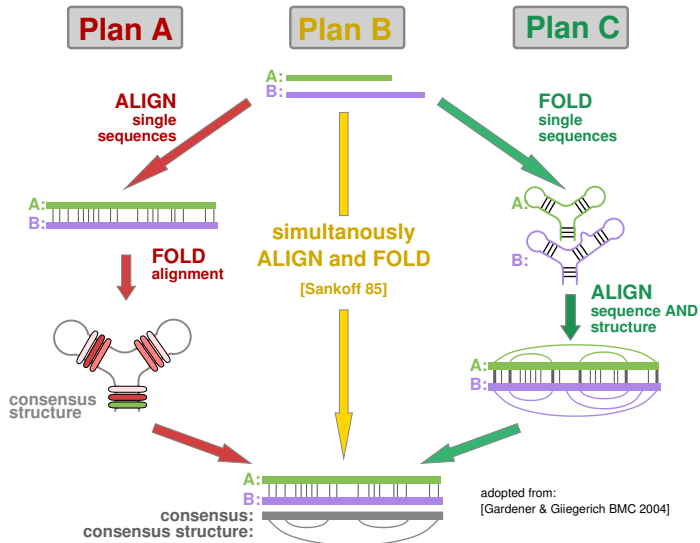
guide tree:



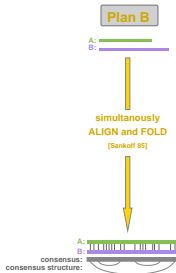
# Infernal Consensus Models



# Comparative RNA Analysis—How?



# Simultaneous ALIGN and FOLD

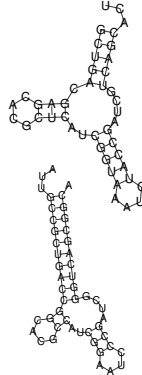
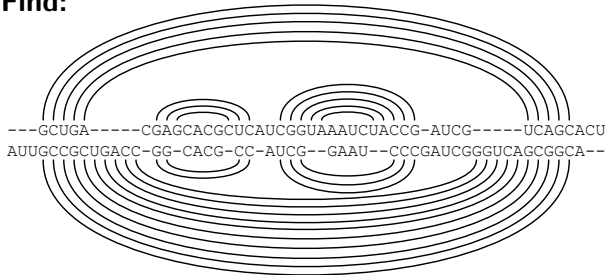


- The classic: Sankoff simultaneous alignment and folding (SA&F)
- “Gold standard” for RNA comparison
- Heuristic short cuts: STRAL, TurboFold II
- Sankoff-style: Dynalign, stemloc, Foldalign
- Fast SA&F (PMcomp-style): PMcomp, LocARNA, RAF, LocARNA-P, SPARSE

# Simultaneous Alignment and Folding<sup>10</sup>

**Given:** A = GCUGACGAGCACGCUCAUCGGUAAAUCUACCGAUCGUCAGCACU  
& B = AUUGCCGCUGACCGGCACGCCAUCGGAAUCCCGAUCGGGUCAGCGGCA

**Find:**



sequence similarity + energy A + energy B  $\rightarrow$  opt

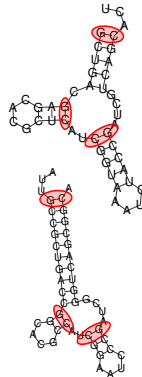
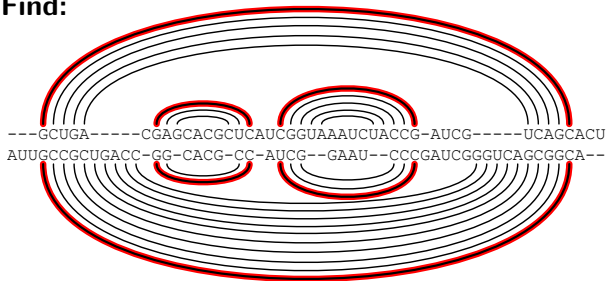
where alignment, structure A, & structure B are **compatible**

<sup>10</sup>Sankoff, 1985

# Simultaneous Alignment and Folding<sup>10</sup>

**Given:** A = GCUGACGAGCACGCUCAUCGGUAAAUCUACCGAUCGUCAGCACU  
& B = AUUGCCGCUGACCGGCACGCCAUCGGAAUCCCGAUCGGGUCAGCGGCA

**Find:**



sequence similarity + energy A + energy B  $\rightarrow$  opt

where alignment, structure A, & structure B are **compatible**

<sup>10</sup>Sankoff, 1985

# Sankoff's SA&F Algorithm

## Dynamic Programming



# Sankoff's SA&F Algorithm

## Dynamic Programming

RNA Energy Minimization [Zuker]

×

Sequence Alignment

# Sankoff's SA&F Algorithm

## Dynamic Programming

RNA Energy Minimization [Zuker]

×

Sequence Alignment

$O(n^6)$  = “extreme computational cost”

## PMcomp's Trick – Lightweight SA&F<sup>11</sup>

Sankoff: **sequence similarity**  
**+ energies of A and B** → **opt**

- **energies** composed of loop energies



<sup>11</sup>Hofacker et al., 2004. [doi:10.1093/bioinformatics/bth229](https://doi.org/10.1093/bioinformatics/bth229)

# PMcomp's Trick – Lightweight SA&F<sup>11</sup>

Sankoff: **sequence similarity**  
**+ energies of A and B** → **opt**

- **energies** composed of loop energies

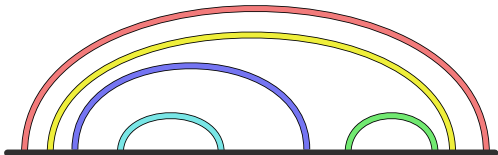


<sup>11</sup>Hofacker et al., 2004. [doi:10.1093/bioinformatics/bth229](https://doi.org/10.1093/bioinformatics/bth229)

## PMcomp's Trick – Lightweight SA&F<sup>11</sup>

PMcomp: **sequence similarity**  
+ **pseudo-energies of A and B** → **opt**

- **pseudo-energies** composed of “base pair energies”

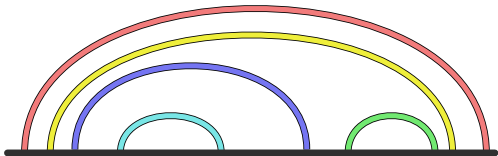


<sup>11</sup>Hofacker et al., 2004. [doi:10.1093/bioinformatics/bth229](https://doi.org/10.1093/bioinformatics/bth229)

## PMcomp's Trick – Lightweight SA&F<sup>11</sup>

PMcomp: **sequence similarity**  
+ **pseudo-energies of A and B** → **opt**

- **pseudo-energies** composed of “base pair energies”



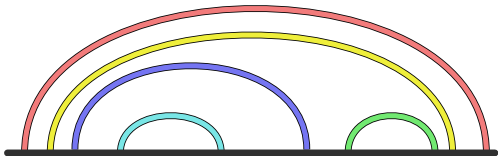
- Dynamic Programming  
Base Pair Maximization [Nussinov] × Sequence Alignment

<sup>11</sup>Hofacker et al., 2004. [doi:10.1093/bioinformatics/bth229](https://doi.org/10.1093/bioinformatics/bth229)

## PMcomp's Trick – Lightweight SA&F<sup>11</sup>

PMcomp: **sequence similarity**  
+ **pseudo-energies of A and B** → **opt**

- **pseudo-energies** composed of “base pair energies”



- Dynamic Programming  
Base Pair Maximization [Nussinov] × Sequence Alignment
- **cheaper computation (at same complexity)**

<sup>11</sup>Hofacker et al., 2004. [doi:10.1093/bioinformatics/bth229](https://doi.org/10.1093/bioinformatics/bth229)

## PMcomp: Nussinov-style Sankoff — Recursion

$$M_{ij;kl} = \max \begin{cases} M_{ij-1;kl-1} + \sigma(A_j, B_l) \\ M_{ij-1;kl} + \gamma \\ M_{ij;k l-1} + \gamma \\ \max_{j'l'} M_{ij'-1;k'l'-1} + D_{j'j;l'l} \end{cases}$$
$$D_{ij;kl} = M_{i+1j-1;k+1l-1} + \tau(i, j, k, l)$$



## PMcomp — Scoring

$$M_{ij;kl} = \max \begin{cases} M_{ij-1;kl-1} + \sigma(A_j, B_l) \\ M_{ij-1;kl} + \gamma \\ M_{ij;kl-1} + \gamma \\ \max_{j'l'} M_{ij'-1;kl'-1} + D_{j'j;l'l} \end{cases}$$

$$D_{ij;kl} = M_{i+1j-1;k+1l-1} + \tau(i, j, k, l)$$

Idea:

- $\tau(i, j, k, l) = \Psi_{ij}^A + \Psi_{kl}^B$
- $\Psi_{ij}^A, \Psi_{kl}^B$ : log odds scores for base-pairs
- “McCaskill”-basepair probabilities vs. background



Hofacker *et al.* Alignment of RNA base pairing probability matrices. *Bioinformatics*, 2004.

## Complexity PMcomp

$$M_{ij;kl} = \max \begin{cases} M_{ij-1;kl-1} + \sigma(A_j, B_l) \\ M_{ij-1;kl} + \gamma \\ M_{ij;k l-1} + \gamma \\ \max_{j'l'} M_{ij'-1;k'l'-1} + D_{j'j;l'l} \end{cases}$$
$$D_{ij;kl} = M_{i+1j-1;k+1l-1} + \tau(i, j, k, l)$$

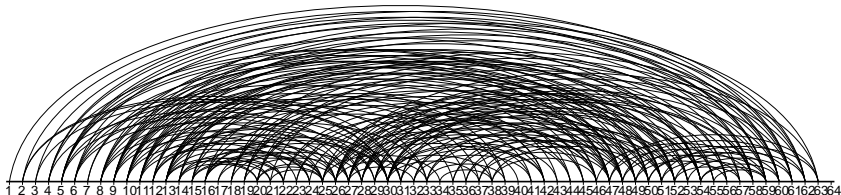
- $O(n^2 \cdot m^2)$  entries in  $M$
- per entry:  $O(nm)$  time

Total Complexity:  $O(n^3 m^3)$  time,  $O(n^2 m^2)$  space

# LocARNA<sup>12</sup>: Fast and Accurate Sankoff

Ideas:

- follow PMcomp idea for scoring
- only consider significant base pairs: “cut-off probability”



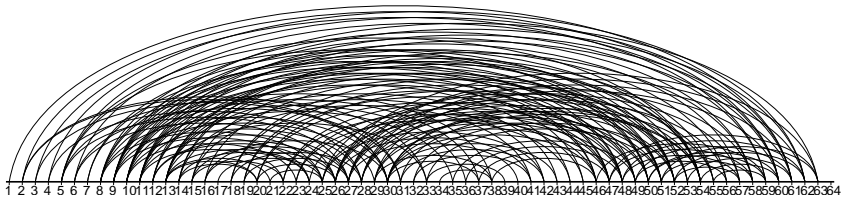
- reformulate recursion
- profit in time and space complexity

---

<sup>12</sup>Will et al., 2007. [doi:10.1371/journal.pcbi.0030065](https://doi.org/10.1371/journal.pcbi.0030065)

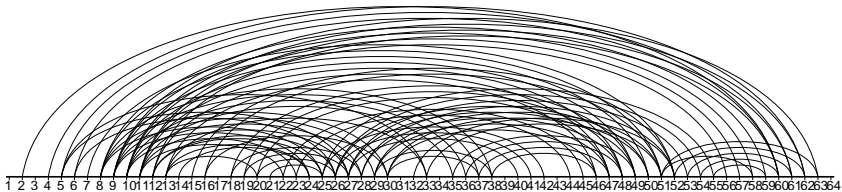
# Effect of Base-Pair Filtering

$$p_{\text{cutoff}} = 0.01$$



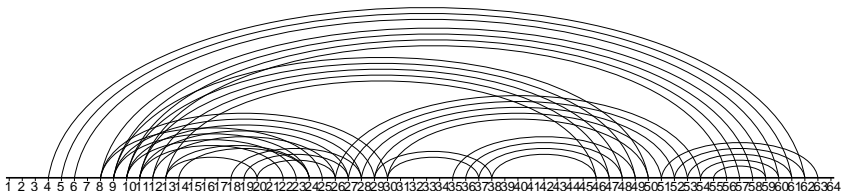
# Effect of Base-Pair Filtering

$$p_{\text{cutoff}} = 0.05$$

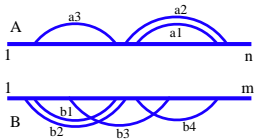


# Effect of Base-Pair Filtering

$$p_{\text{cutoff}} = 0.1$$



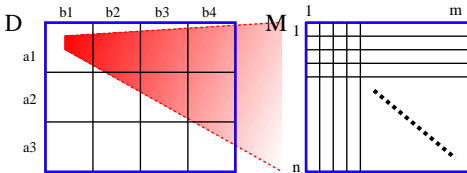
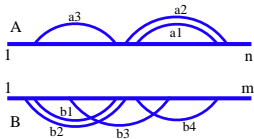
# LocARNA Basic Algorithm: Matrices



D

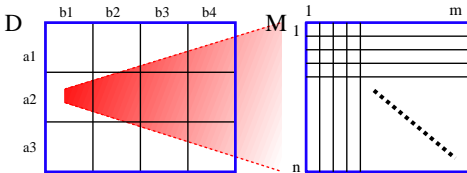
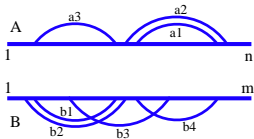
	b1	b2	b3	b4
a1				
a2				
a3				

# LocARNA Basic Algorithm: Matrices

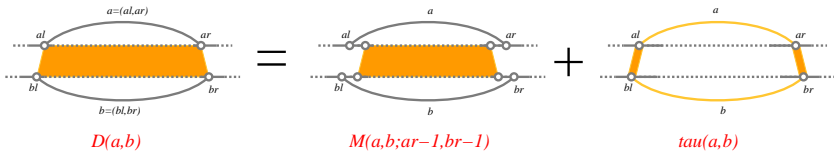




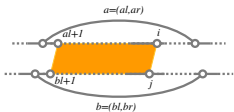
# LocARNA Basic Algorithm: Matrices



# LocARNA Basic Algorithm: Recursion

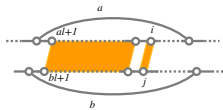


# LocARNA Basic Algorithm: Recursion

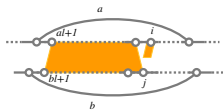


$M(a,b;i,j)$

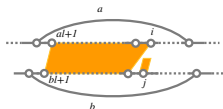
= max



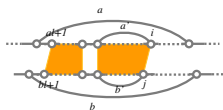
$M(a,b;i-1,j-1) + \sigma(A_i, B_j)$



$M(a,b;i,j-1) + \gamma$



$M(a,b;i-1,j) + \gamma$



$\max_{a'b'} M(a,b;a'-1,b'-1) + D(a',b')$   
 where  $a'r=i, b'r=j$

## Complexity LocARNA

$$M^{a,b}(i,j) = \max \begin{cases} M^{a,b}(i-1, j-1) + \sigma(A_i, B_j) \\ M^{a,b}(i-1, j) + \gamma \\ M^{a,b}(i, j-1) + \gamma \\ \max_{a', b'} M^{a,b}(a'-1, b'-1) + D(a', b') \\ \quad \text{where } a'_r = i, b'_r = j \end{cases}$$
$$D(a, b) = M^{a,b}(a_r - 1, b_r - 1) + \tau(a, b)$$

Probability threshold  $p_{\text{cutoff}} \Rightarrow \text{deg} \leq 1/p_{\text{cutoff}} \in O(1)$

- compute  $D(a, b)$  for all base pair edges:  
 $\Rightarrow O(|P_1||P_2|) =_{(!)} O(nm)$  pairs of base pairs  $(a, b)$
- $O(nm \cdot \text{rdeg}_1 \text{rdeg}_2) =_{(!)} O(nm)$  time per  $(a, b)$

*Total Complexity:*  $O(n^2 m^2)$  time,  $O(nm)$  space

## LocARNA implements various extensions

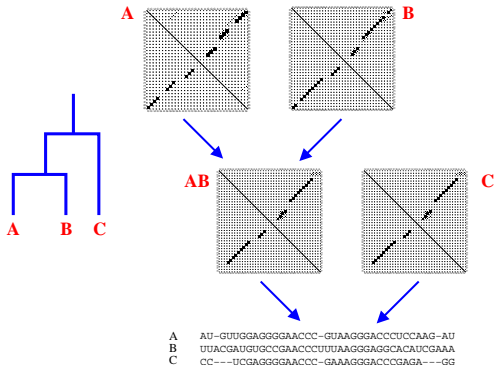
- more realistic “affine” gap cost
- sequence and structure locality
- anchor and structure constraints
- multiple alignment
- scoring of stacks
- normalized local alignment
- partition functions (LocARNA-P<sup>13</sup>)
- stronger sparsification and added structural flexibility (SPARSE<sup>14</sup>)

---

<sup>13</sup>Will et al., 2012. [doi:10.1261/rna.029041.111](https://doi.org/10.1261/rna.029041.111)

<sup>14</sup>Will et al., 2015. [doi:10.1093/bioinformatics/btv185](https://doi.org/10.1093/bioinformatics/btv185)

# Multiple LocARNA (mlocarna): Progressive Alignment



- pairwise comparison all-2-all
- guide tree
- aligning alignments along guide tree
- heuristic (does not guarantee global optimum)

# LocARNA Example Input

*Unaligned sequences, unknown structures:*

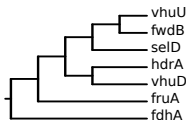
```
>fruA
CCUCGAGGGGAACCCGAAAGGGACCCGAGAGG
>fdhA
CGCCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAGGUGGCG
>vhuU
AGCUCACAACCGAACCCAUUUGGGAGGUUGUGAGCU
>hdrA
GGCACCACUCGAAGGCUAAGCCAAAGUGGUGCU
>vhuD
GUUCUCUCGGGAACCCGUCAAGGGACCGAGAGAAC
>selD
UUACGAUGGCCGAACCCUUUAAGGGAGGCACAUCGAAA
>fwdB
AUGUUGGAGGGGAACCCGUAAGGGACCCUCCAAGAU
```

# LocARNA Example Output

## Similarities:

-	-123	1433	1842	2319	848	2906
-123	-	2158	1406	2361	249	1224
1433	2158	-	2555	3250	3069	5410
1842	1406	2555	-	3766	1750	2084
2319	2361	3250	3766	-	3449	3679
848	249	3069	1750	3449	-	2977
2906	1224	5410	2084	3679	2977	-

## Guide tree:



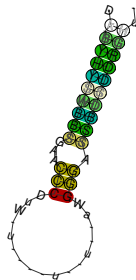
((((vhuU,fwdB),selD),(hdrA,vhuD)),fruA),fdhA);

## Alignment and consensus structure:

```

.(((.(((((((.....((((.....))))))))))))))..
vhuU AG-CUCACAAACGAACC AUU-----U  GGAGGUUGUGAGCU- 36
fwdB AU-GUUGGAGGGGAACCGUA-----A  GGGACCUCUCAAAGAU- 36
selD UUAACGAUGUGCGGAACCCUUU-----AA  GGGAGGCACAUCGAAA 39
hdrA G--GCACCACTCGAAGGC--U-----AA  CCAAAGUGGUGCU-- 33
vhuD G--UUCUCUCGGGAACCGUC-----AA  GGGACCAGAGAGAAC-- 35
fruA ---CCUCGAGGGGAACCGG-A-----AA  GGGACCAGAGAGG-- 32
fdhA CG-CCACCCUGGAAACC AAUAUAAAAUAUACAA  GGGAGAG-GUGGCG- 48
.....10.....20.....30.....40.....50

```

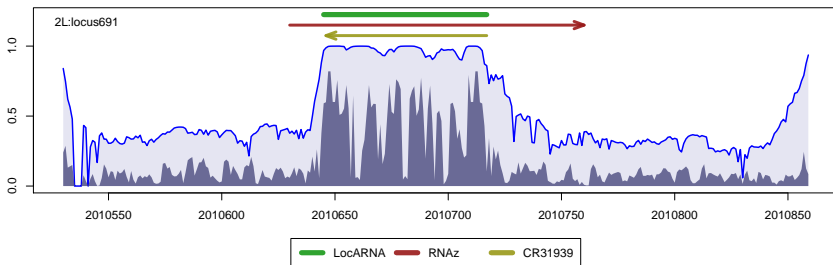




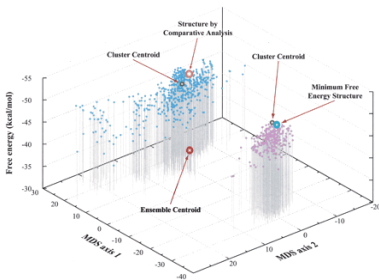
# Probabilities of RNA alignments

- LocARNA-P extends LocARNA to compute structure alignment *probabilities*  
(using a statistical mechanics approach; 'partition functions')
- distinguishes sequence match and structure match probabilities
- calculates local, column-wise quality of multiple alignments:  
reliability profiles
- predicts ncRNA boundaries

## Structure Alignment Reliability (STAR) Profile:



# Clustering



- clustering structures of one RNA<sup>3</sup>
- structure-based clustering of RNAs (RNAclust, GraphClust)

---

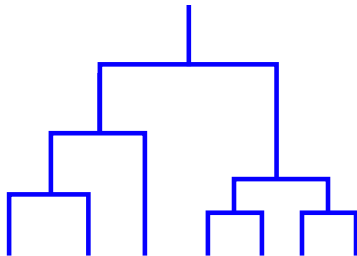
<sup>3</sup>e.g. Ding et al.; RNA 2005; [doi:10.1261/rna.2500605](https://doi.org/10.1261/rna.2500605)

## General ideas about RNA clustering

- cluster a set of RNAs (e.g. predicted ncRNA candidates from a genome)  
    [different problem: cluster set of structures of one RNA]
- structure-based, unknown structure; ideally: plan B
- naive:  $O(n^2)$  comparisons  $\Rightarrow$  Distance matrix
- first idea: hierarchical clustering (UPGMA, NJ)
- how to identify sub-groups that form distinguished clusters?

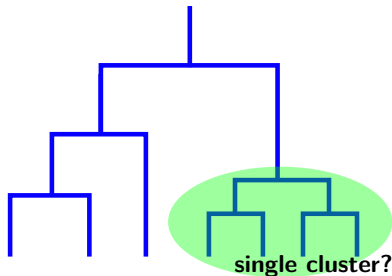
# General ideas about RNA clustering

- cluster a set of RNAs (e.g. predicted ncRNA candidates from a genome)  
[different problem: cluster set of structures of one RNA]
- structure-based, unknown structure; ideally: plan B
- naive:  $O(n^2)$  comparisons  $\Rightarrow$  Distance matrix
- first idea: hierarchical clustering (UPGMA, NJ)
- how to identify sub-groups that form distinguished clusters?



# General ideas about RNA clustering

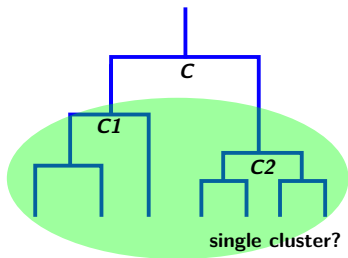
- cluster a set of RNAs (e.g. predicted ncRNA candidates from a genome)  
[different problem: cluster set of structures of one RNA]
- structure-based, unknown structure; ideally: plan B
- naive:  $O(n^2)$  comparisons  $\Rightarrow$  Distance matrix
- first idea: hierarchical clustering (UPGMA, NJ)
- how to identify sub-groups that form distinguished clusters?



# Clustering using LocARNA

- GOAL: identify groups of related RNAs
- IN: set of RNAs
- OUT: hierarchical clustering of RNAs
- Steps
  - compare RNAs all-2-all using LocARNA
  - cluster-tree by hierarchical clustering (UPGMA)
  - identify meaningful clusters
- Application: cluster RNAs from RNAz screen

## The Duda rule<sup>15</sup> in RNAclust<sup>16</sup>



Combine C1 and C2?

Test hypothesis:

“C is single cluster”

- evaluate minimum free energies of sequences  $E_i$  (RNAfold)
- evaluate MFE of consensus structures  $E_{cons}(C)$  (RNAalifold)
- consider squared error

$$\Delta(C) = \sum_{i \in C} (E_i - E_{cons}(C))^2$$

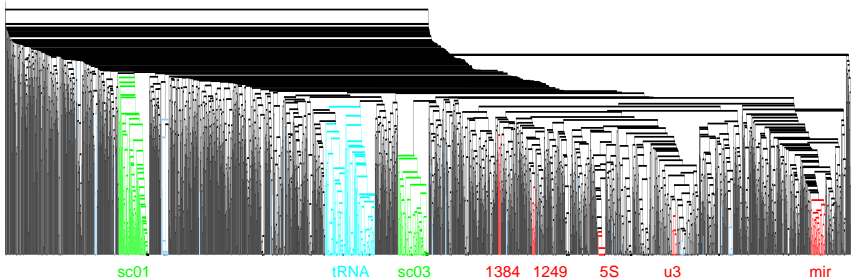
- $\frac{\Delta(C1) + \Delta(C2)}{\Delta(C)} < \theta$ , then reject

e.g. we could achieve MCC 0.8 in an evaluation on Rfam

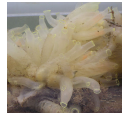
<sup>15</sup>Duda et al. Pattern Classification, 2001

<sup>16</sup><http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAclust/>

# Clustering of RNAz ncRNA Predictions

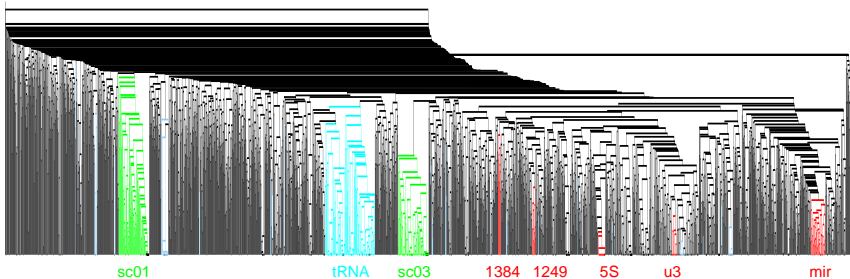


Clustering of 3332 putative ncRNAs in *Ciona intestinalis*

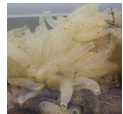




# Clustering of RNAz ncRNA Predictions

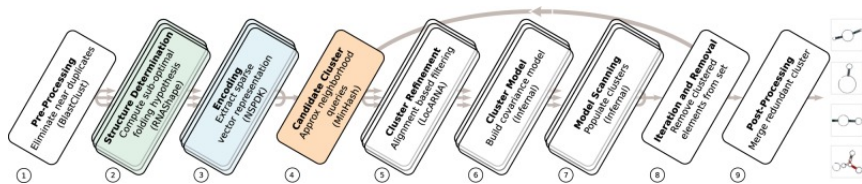


Clustering of 3332 putative ncRNAs in *Ciona intestinalis*



- putative ncRNAs from RNAz screen
- requires  $3332 \cdot 3331/2 \approx 5.5 \times 10^6$  LocARNA alignments
- e.g. 16,000 predicted ncRNAs in *Drosophila*; 37,000 in Human

# GraphClust<sup>17</sup>: Workflow and Results

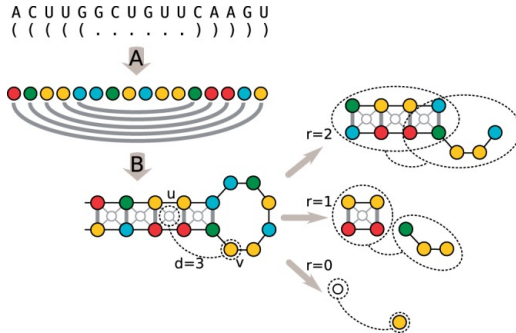


Species	Type	Method	Input	Size (Mb)	Time <sup>a</sup>	Cluster	MPI <sub>avg</sub>	SCI <sub>&gt;0.5</sub>
<i>Benchmark</i>								
Bacteria	Small ncRNAs	Misc	363	0.06	6.8 h	39	0.75	29
Human	Predicted RNA elements	EVOFAM	699	0.03	0.3 h	37	0.52	36
Misc	Small ncRNAs	Rfam	3900	0.51	36 h	130	0.64	98
<i>De-novo discovery</i>								
Fugu	LincRNAs	RNA-seq	5877	0.09	10.3 h	99	0.39	16
Fugu	Predicted RNA elements	RNAZ	11 287	1.36	13.3 h	97	0.39	22
Fruit fly	Predicted RNA elements	RNAZ	17 765	2.15	20.4 h	95	0.34	23
Human	LincRNAs	RNA-seq	31 418	5.40	3.6 d	95	0.34	3
Human	Predicted RNA elements	EVOFOLD	37 258	1.37	5.7 d	117	0.75	109
Human	3'UTRs	RefSeq	118 514	21.91	12.8 d	106	0.34	13
Σ			227 081	32.88	25.7 d	815	–	349

<sup>17</sup>Heyne et al., 2012. [doi:10.1093/bioinformatics/bts224](https://doi.org/10.1093/bioinformatics/bts224)

# GraphClust's Efficiency: Graph Features

The RNAs are represented as sets of structural *graph features*



## GraphClust's Efficiency

Main idea: Find clusters by “*Approximative neighborhood queries*”

- Use *Locality Sensitive Hashing (LSH)*. Let  $x, y$  be sets of features (representing two RNAs).

Define 400 independent *LSH functions*  $h_1, \dots, h_{400}$ , such that

$$h_i(x) = h_i(y) \text{ with probability } J(x, y) = \frac{x \cap y}{x \cup y}.$$

*MinHashing*: Choose  $h(x)$  as index of the minimal feature in  $x$  given some permutation of all features.

- build 400 *reverse* indices  $Z_i$  to find the  $x$  where  $h_i(x) = c$
- now:  $y \in Z_i(h_i(x))$  with probability  $J(x, y)$ !

⇒ find potential neighbors  $y$  of any  $x$  in constant time by searching through the most frequent elements in the multiset  $\bigcup_i Z_i(h_i(x))$ .

## Many remaining special issues

- using sparsity for further speed up
- pseudoknots
- non-canonical base pairs
- window-less de-novo prediction
- improved multiple alignment
- local (multiple) structure alignment
- local clustering
- multiple conserved structures
- ...

## Outlook to hands-on tutorial: From A to B and back again

- Analyzing alignments
- How (not) to use LocARNA
- Finding ncRNA candidates: RNAz screens and clustering

# Outlook to hands-on tutorial: From A to B and back again

- Analyzing alignments
- How (not) to use LocARNA
- Finding ncRNA candidates: RNAz screens and clustering

Please prepare for the hand on session: perform installations before class this afternoon

Detailed installation instructions are provided at the start of <https://www.tbi.univie.ac.at/~will/AlgoSB19/NOTES.txt>

Course Material: <https://www.tbi.univie.ac.at/~will/AlgoSB19/>