Sequence-Structure Alignment — A General Formulation

"Unifying view on Edit Distance, SA&F, ..."

IN

- $S_1, \dots, S_k \in \Sigma$
- $P_1, \dots, P_k \in \{1, \dots, |S_i|\}$: sets of basepairs
- score on alignments

OUT

Alignment $A = (S_1^*, P_1^*, \dots, S_k^*, P_k^*)$ that maximizes score(A), where $S_i^*|_\Sigma = S_i$, "$P_i^*|_\Sigma$" $\subseteq P_i$, ...

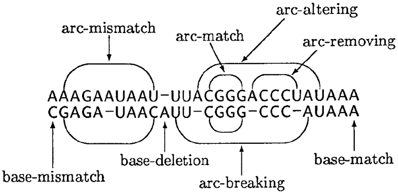
Exact conditions and score vary

problem classes: restrict input and output structures, score

Alignment with Fixed Input Structures

Jiang *et al.* A General Edit Distance between RNA Structures. *JCB*, 2002.

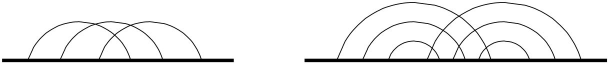
- “$P_i^*|_\Sigma$” $= P_i$, i.e. output structure = input structure
- score is rather general edit distance (breaking of basepairs)
- only pairwise, $k = 2$
- efficient only for NESTED/CROSSING with “not so general score”



Alignment with Fixed Input Structures – Pseudoknots

- CROSSING/CROSSING, i.e. pseudoknots allowed
- restricted pseudoknots:
  e.g., no crossing of 3 basepairs

  Patricia A. Evans. Finding common RNA pseudoknot structures in polynomial time. CPM 2006.



a) a three−knot

b) interleaved left−right endpoints

  Möhl, Will, Backofen. Lifting prediction to alignment of RNA pseudoknots. RECOMB 2009.

- general crossing:

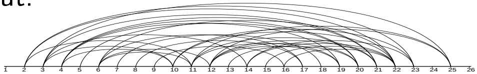
  Möhl, Will, Backofen. Fixed parameter tractable alignment of RNA structures including arbitrary pseudoknots. CPM 2008

Simultaneous Alignment and Folding (SA&F)

David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 1985.

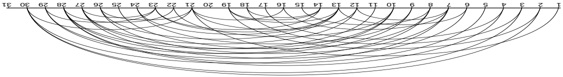
- “$P_i^*|_\Sigma$” $\subseteq P_i$
- input structures *crossing* (all potential basepairs)
- output structures *non-crossing*

Example Input:

$P_1 =$ 

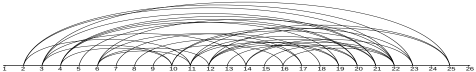
$S_1 = ACGGACUUACGGACUUGACUCGGACU$

$S_2 = CGGAACGUAUACGGACUCCAGACUACGUGCA$

$P_2 =$ 

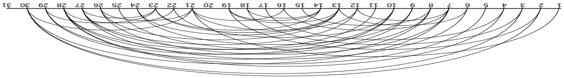
Example SA&F

IN:

$P_1 =$ 

$S_1 = ACGGACUUACGGACUUGACUCGGACU$

$S_2 = CGGAACGUAUACGGACUCCAGACUACGUGCA$

$P_2 =$ 

OUT:

$P_1^* \equiv$ `----.(.((..(........)..)).)...----`

$S_1^* =$ `----ACGGACUUACGGACUUGACUCGGACU----`

$S_2^* =$ `CGGAACGUAUACGGACUCCAGACUACG---UGCA`

$P_2^* \equiv$ `.....(.((..(........)..)).)---....`

Incomplete history of SA&F

- 1985 Sankoff. *Computationally heavy, no implementation*
- 1997 Foldalign (Gorodkin et *only stems, simpler energy*
- 2002 Dynalign (Mathews, Turner) *first "full" implementation*
- 2004 PMcomp (Hofacker et al.) *clever simplification*
- 2007 FoldalignM Mc (Torarinsson et al.), *PMcomp implementation*
- 2007 LocARNA (Will, et al.), *PMcomp-based, more time and space efficient, optionally local*
- 2008 RAF (Do, *et al.*), *PMcomp-based, sequence-sparsity, machine learning*
- 2011 LocARNA-P (Will, et al.), *efficient partition function*

PMcomp: A Realistic Nussinov-style Sankoff-Algorithm

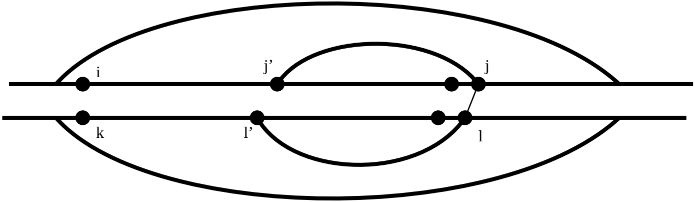
Idea:

- Simplify Energy Model of SA&F:
  Loop-based (Zuker-style) ⇒ Base-pair-based (Nussinov-style)
- Advantage?
- Problem?
- Add realistic energy scoring again!: McCaskill pair probabilities

PMcomp: Nussinov-style Sankoff — Recursion

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

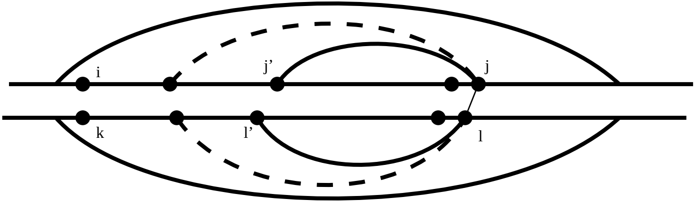
$$D_{i\,j;k\,l} = M_{i+1\,j-1;k+1\,l-1} + \tau(i,j,k,l)$$



PMcomp: Nussinov-style Sankoff — Recursion

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

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



PMcomp — Scoring

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

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

Idea:

- $\tau(i, j, k, l) = \Psi^A_{ij} + \Psi^B_{kl}$
- Ψ^A_{ij}, Ψ^B_{kl}: 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

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

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

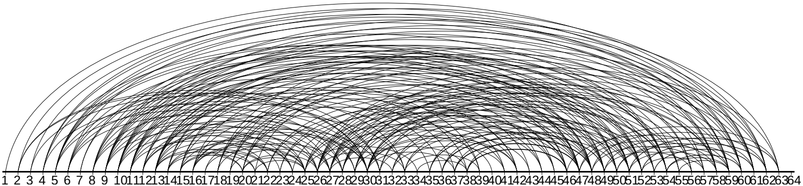
- $O(n^2 \cdot m^2)$ entries in M
- per entry: $O(nm)$ time

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

LocARNA: Making PMcomp/Sankoff practical

Ideas:

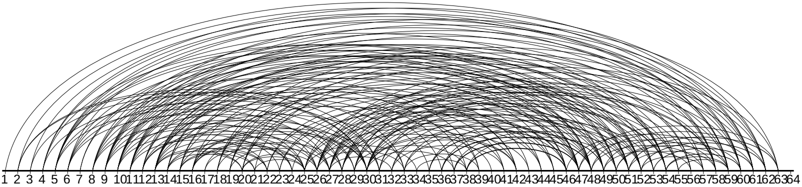
- follow PMcomp idea for scoring
- only consider significant base pairs: “cut-off probability”



- reformulate recursion
- profit in time and space complexity

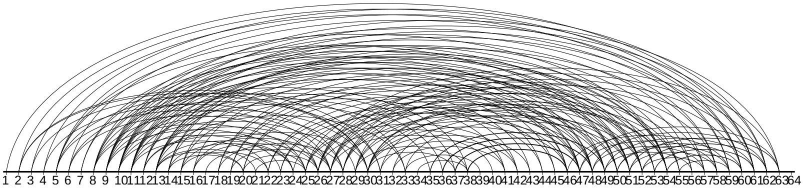
Effect of Base-Pair Filtering

$p_{\text{cutoff}} = 0.005$



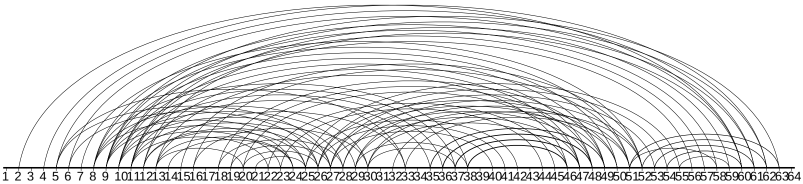
Effect of Base-Pair Filtering

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



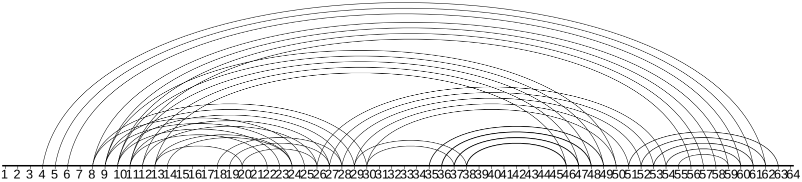
Effect of Base-Pair Filtering

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

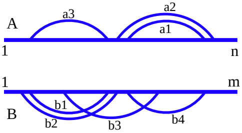


Effect of Base-Pair Filtering

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

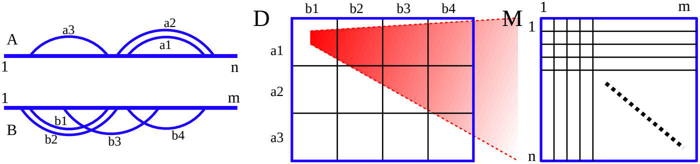


Locarna Basic Algorithm: Matrices

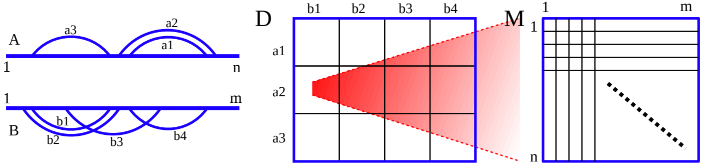


D	b1	b2	b3	b4
a1				
a2				
a3				

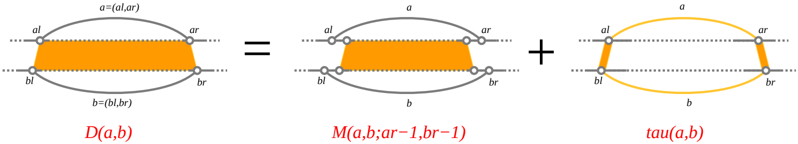
Locarna Basic Algorithm: Matrices



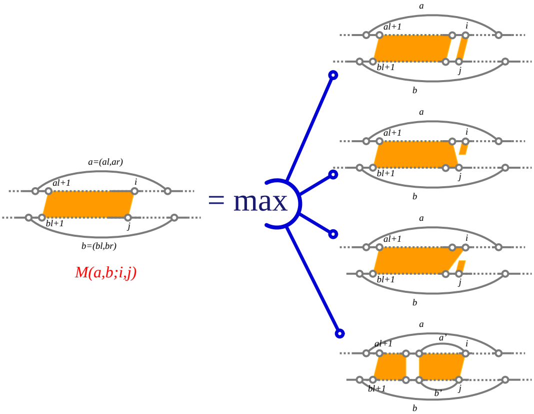
Locarna Basic Algorithm: Matrices



Locarna Basic Algorithm: Recursion



Locarna Basic Algorithm: Recursion



M(a,b;i,j)

M(a,b;i−1,j−1) + sigma(Ai,Bj)

M(a,b;i,j−1) + gamma

M(a,b;i−1,j) + gamma

max a'b': M(a,b;a'l−1,b'l−1) + D(a',b')
where a'r=i, b'r=j

Locarna Basic Algorithm: Recursion

$$M^{a\,b}(i,j) = \max \begin{cases} M^{a\,b}(i-1,j-1) + \sigma(A_i,B_j) \\ M^{a\,b}(i-1,j) + \gamma \\ M^{a\,b}(i,j-1) + \gamma \\ \max\limits_{a'\,b'} M^{a\,b}(a'_l - 1, b'_l - 1) + D(a',b') \\ \qquad \text{where } a'_r = i, b'_r = j \end{cases}$$

$$D(a,b) = M^{a\,b}(a_r - 1, b_r - 1) + \tau(a,b)$$

Complexity LocARNA

$$M^{a\,b}(i,j) = \max \begin{cases} M^{a\,b}(i-1,j-1) + \sigma(A_i,B_j) \\ M^{a\,b}(i-1,j) + \gamma \\ M^{a\,b}(i,j-1) + \gamma \\ \max\limits_{a'\,b'} M^{a\,b}(a'_l - 1, b'_l - 1) + D(a',b') \\ \qquad \text{where } a'_r = i, b'_r = j \end{cases}$$

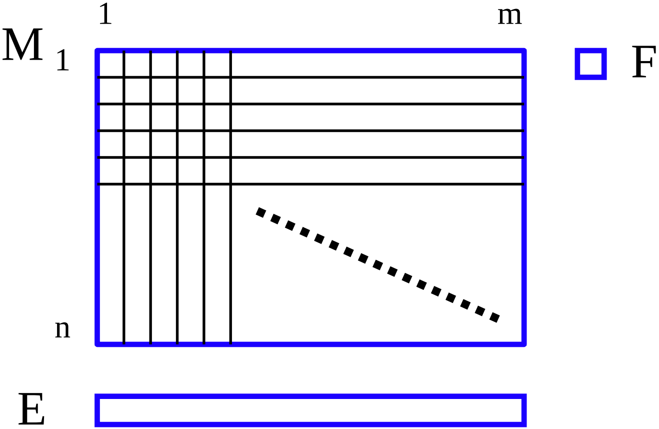
$$D(a,b) = M^{a\,b}(a_r - 1, b_r - 1) + \tau(a,b)$$

- compute $D(a,b)$ for all base-pairs edges: $a \in P_1, b \in P_2$ [and a,b compatible] $\quad \Longrightarrow O(|P_1||P_2|)$
- combine $D(a,b)$-computation for common $(a_l,b_l) \Rightarrow O(nm)$
- per (a_l,b_l): $O(nm \cdot \text{rdeg}_1\,\text{rdeg}_2)$

Total Complexity: $O(nm|P_1||P_2|)$ time, $O(|P_1||P_2| + nm)$ space

Affine Gap Cost

- Basic algorithm: linear gap cost
- Affine gap cost $g(k) = \alpha + \beta \cdot k$: ala Gotoh



Affine Gap Cost

$$M^{a\,b}(i,j) = \max \begin{cases} M^{a\,b}(i-1,j-1) + \sigma(A_i, B_j) \\ E_i^{a\,b}(j) \\ F_{i\,j}^{a\,b} \\ \max\limits_{a'\,b'} M^{a\,b}(a'_l - 1, b'_l - 1) + D(a', b') \\ \qquad \text{where } a'_r = i, b'_r = j \end{cases}$$

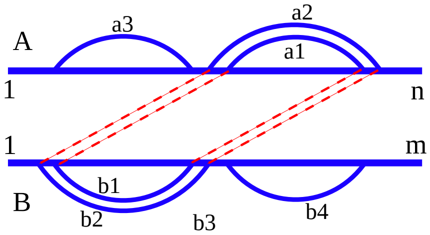
$$D(a,b) = M^{a\,b}(a_r - 1, b_r - 1) + \tau(a,b)$$

$$E_i^{a\,b}(j) = \max\{E_{i-1}^{a\,b}(j) + \beta, M^{a\,b}(i-1,j) + \alpha + \beta\}$$

$$F_{i\,j}^{a\,b} = \max\{F_{i\,j-1}^{a\,b} + \beta, M^{a\,b}(i,j-1) + \alpha + \beta\}$$

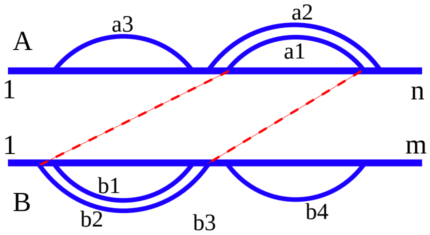
Stacking

- Distinguish stacked and un-stacked base pair matches
- Implementation without change of recursion structure
- No additional computational cost



Stacking

- Distinguish stacked and un-stacked base pair matches
- Implementation without change of recursion structure
- No additional computational cost



Stacking Recursion

$$M^{a\,b}(i,j) = \max \begin{cases} M^{a\,b}(i-1,j-1) + \sigma(A_i, B_j) \\ M^{a\,b}(i-1,j) + \gamma \\ M^{a\,b}(i,j-1) + \gamma \\ \max\limits_{a'\,b'} M^{a\,b}(a'_l - 1, b'_l - 1) + D(a', b') \\ \qquad \text{where } a'_r = i, b'_r = j \end{cases}$$

$$D(a,b) = \max \begin{cases} M^{a\,b}(a_r - 1, b_r - 1) + \tau(a,b) \\ D(a', b') + \tau'(a,b) \\ \quad \text{where } (a,b) \text{ stacked to } (a', b') \end{cases}$$

LocARNA: sequence local alignment

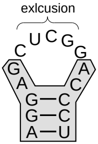
- find best alignment of subsequences
- special "last" recursion for pseudo-arcs a_0, b_0

$$M^{a_0\,b_0}(i,j) = \max \begin{cases} 0 \\ M^{a_0\,b_0}(i-1,j-1) + \sigma(A_j, B_l) \\ M^{a_0\,b_0}(i-1,j) + \gamma \\ M^{a_0\,b_0}(i,j-1) + \gamma \\ \max\limits_{a'\,b'} M^{a_0\,b_0}(a'_l - 1, b'_l - 1) + D(a', b') \\ \qquad \text{where } a'_r = i, b'_r = j \end{cases}$$

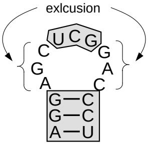
- back-trace from maximal entry to 0-entry (cf. local sequence alignment).

LocARNA: structure local alignment

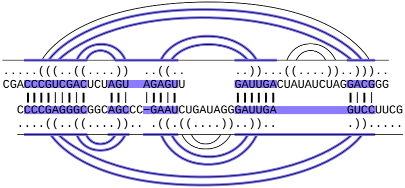
What is structure local?



allowed



disallowed



Find best alignment of "connected" sub-structures.

Idea

- exclusions, allow only one per basepair-match per sequence
- counting: 0/1 exclusions in seq 1, 0/1 exclusions in seq 2
  $\Longrightarrow$ 4 states/matrices
- Gotoh's trick: exclusion opening + exclusion extension
  $\Longrightarrow$ 8 states/matrices

Reward: Structure locality without increasing complexity

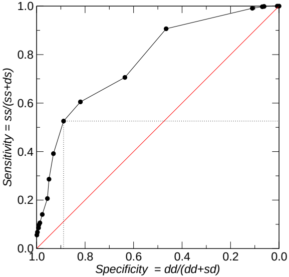
Application of LocARNA: Clustering of RNAs

- 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 hiererchical clustering (WPGMA)
  - identify meaningful clusters
- Application: cluster RNAs from RNAz screen
  RNAz can identify potential non-coding RNAs in genomes

more about RNAz and prediction of ncRNA in genomes:
Guest Lecture: Thursday, Oct 27: Stefan Washietl

Evaluation: Reproducing RNA families of Rfam

Rfam = collection of RNA families and their alignments
(= known classification)

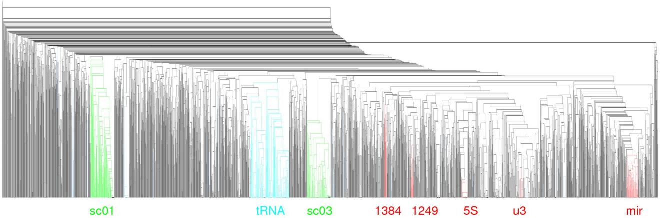


:

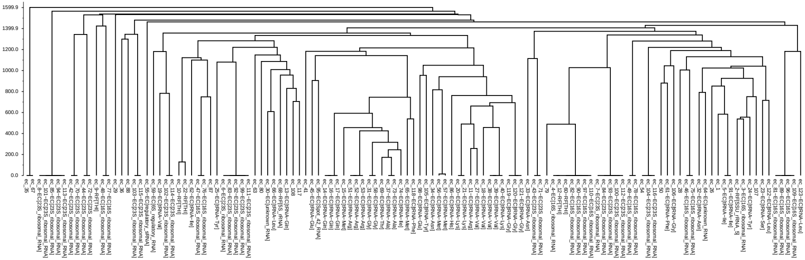
Minimum recall level	Average recall	Average precision
0.50	0.5818	0.8280
0.55	0.6996	0.7819
0.60	0.7277	0.7530
0.65	0.7596	0.7117
0.70	0.8092	0.6831
0.75	0.8519	0.5949
0.80	0.8763	0.5701
0.85	0.9381	0.4794
0.90	0.9599	0.4419
0.95	0.9766	0.3907

LocARNA: Clustering of RNAz ncRNA Predictions

- Clustering of 3332 putative ncRNAs in *Ciona intestinalis*

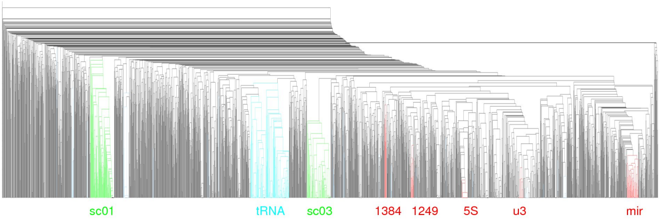


- Clustering of bacterial RNAz predictions

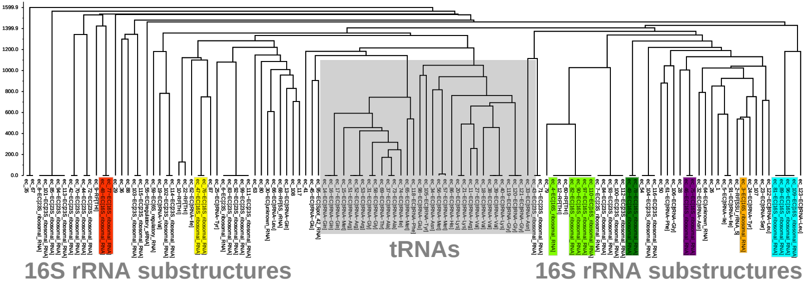


LocARNA: Clustering of RNAz ncRNA Predictions

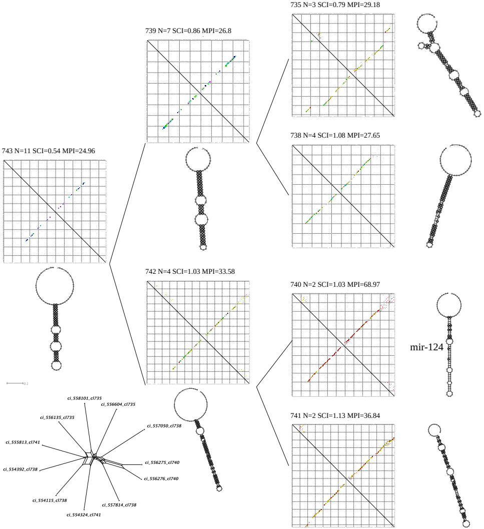
- Clustering of 3332 putative ncRNAs in *Ciona intestinalis*



- Clustering of bacterial RNAz predictions



LocARNA Cluster: Known and Predicted microRNAs


735 N=3 SCI=0.79 MPI=29.18
739 N=7 SCI=0.86 MPI=26.8
738 N=4 SCI=1.08 MPI=27.65
743 N=11 SCI=0.54 MPI=24.96
742 N=4 SCI=1.03 MPI=33.58
740 N=2 SCI=1.03 MPI=68.97
mir-124
741 N=2 SCI=1.13 MPI=36.84
0.1
ci_558101_cl735
ci_556604_cl735
ci_556135_cl735
ci_557050_cl738
ci_555813_cl741
ci_556275_cl740
ci_554392_cl738
ci_556276_cl740
ci_554115_cl738
ci_557814_cl738
ci_554324_cl741

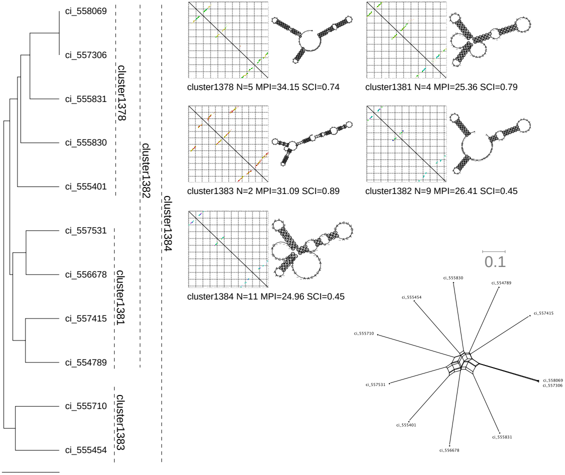
- taken from clustering of 3332 predicted ncRNAs in *C. intestinalis*

- local and global alignment of base pairing probability matrices

- detection of conserved structural RNAs by clustering

- successfully tested on RFAM

Case Study 1



cluster1378 N=5 MPI=34.15 SCI=0.74

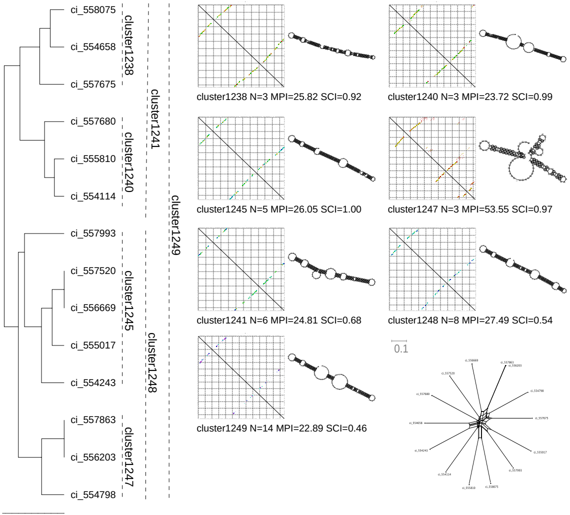
cluster1381 N=4 MPI=25.36 SCI=0.79

cluster1383 N=2 MPI=31.09 SCI=0.89

cluster1382 N=9 MPI=26.41 SCI=0.45

cluster1384 N=11 MPI=24.96 SCI=0.45

Case Study 2



cluster1238 N=3 MPI=25.82 SCI=0.92

cluster1240 N=3 MPI=23.72 SCI=0.99

cluster1245 N=5 MPI=26.05 SCI=1.00

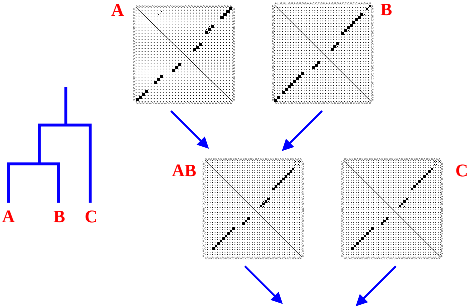
cluster1247 N=3 MPI=53.55 SCI=0.97

cluster1241 N=6 MPI=24.81 SCI=0.68

cluster1248 N=8 MPI=27.49 SCI=0.54

cluster1249 N=14 MPI=22.89 SCI=0.46

Multiple LocARNA: Progressive Alignment



```
A     AU-GUUGGAGGGGAACCC-GUAAGGGACCCUCCAAG-AU
B     UUACGAUGUGCCGAACCCUUUAAGGGAGGCACAUCGAAA
C     CC---UCGAGGGGAACCC-GAAAGGGACCCGAGA---GG
```

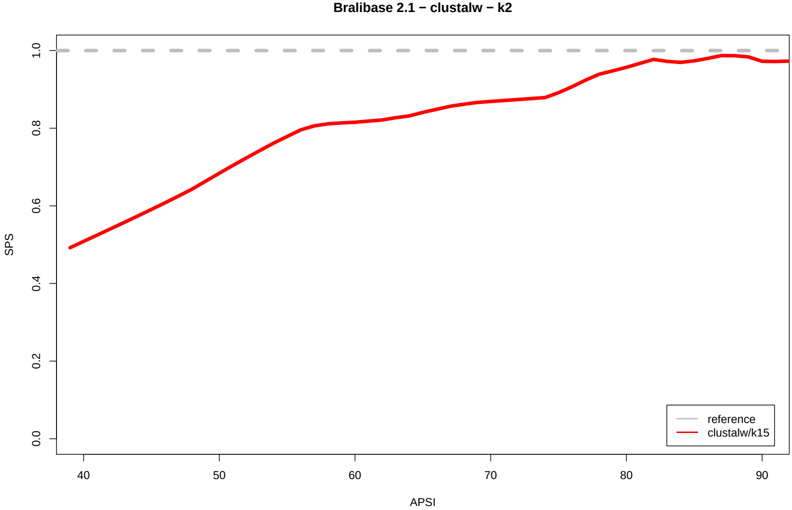
- pairwise comparison all-2-all
- guide tree
- aligning alignments along guide tree
- heuristic: can make mistakes

BRALIBASE 2.1

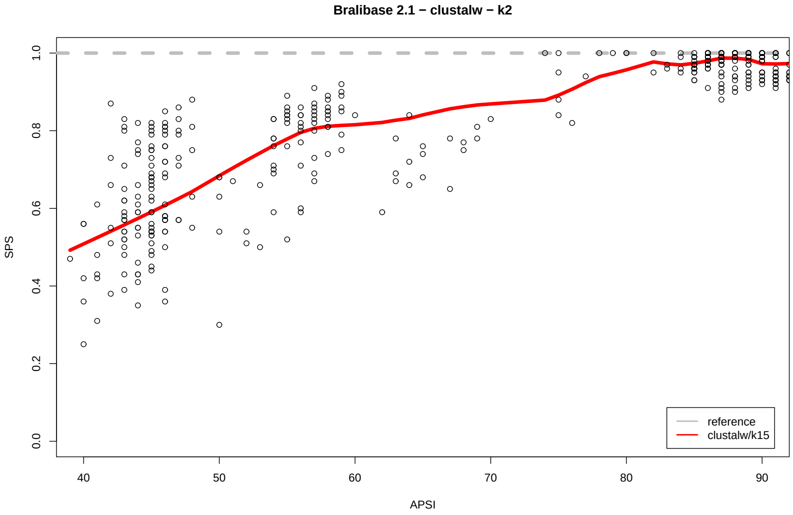
Compilation of "true" RNA alignments from Rfam
Benchmark set for multiple RNA alignment

Set	#Sequences	#Alignments
k2	2	8976
k3	3	4835
k5	5	2405
k7	7	1426
k10	10	845
k15	15	503

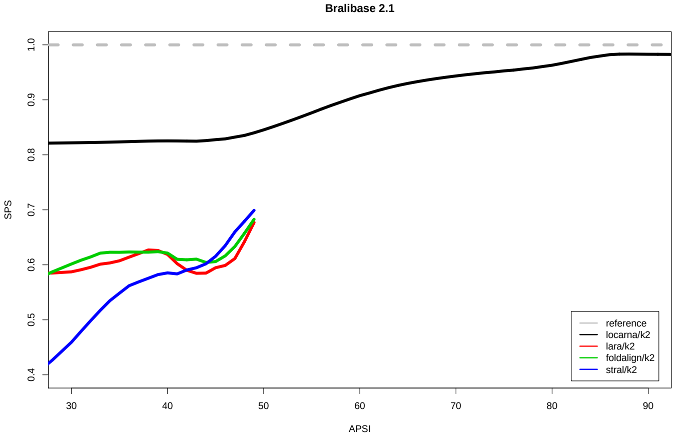
Bralibase SPS plots



Bralibase SPS plots

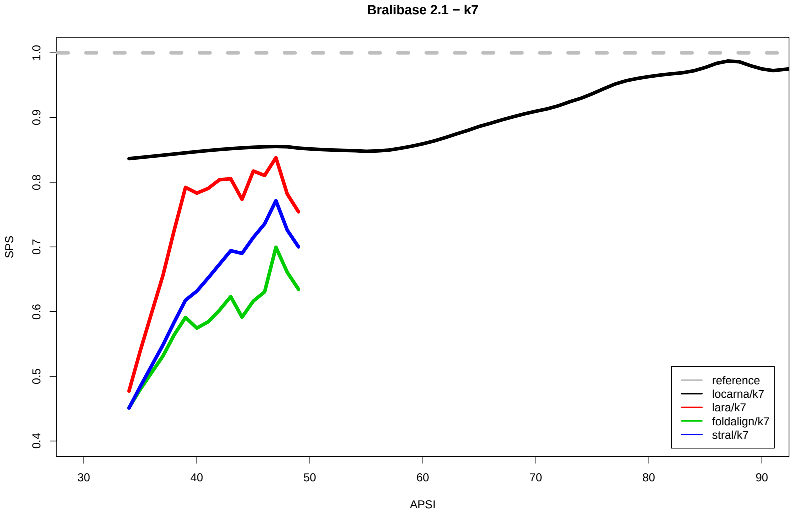


Pairwise LocARNA vs. Others

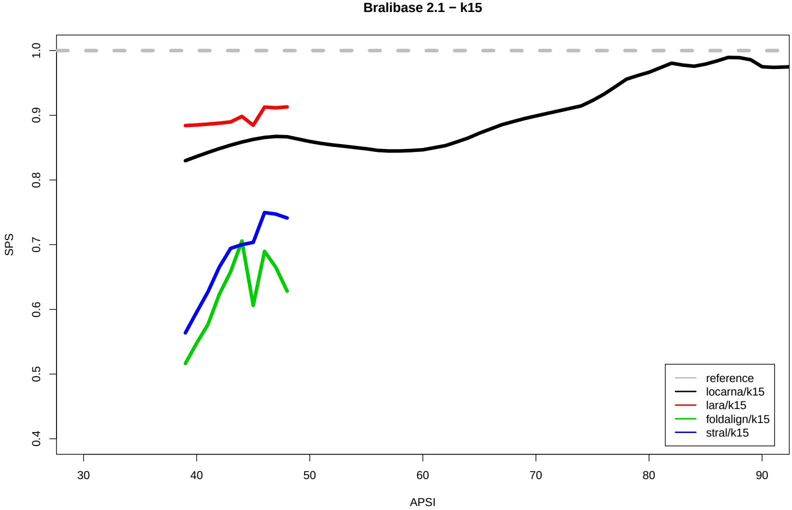


Data for Lara, Foldalign, Stral: Bauer, Klau, Reinert. BMC 2007.
Only ≤ 50% available.

Multiple LocARNA vs. Others - 7 sequences



Multiple LocARNA vs. Others - 15 sequences


Bralibase 2.1 − k15
SPS
APSI
reference
locarna/k15
lara/k15
foldalign/k15
stral/k15