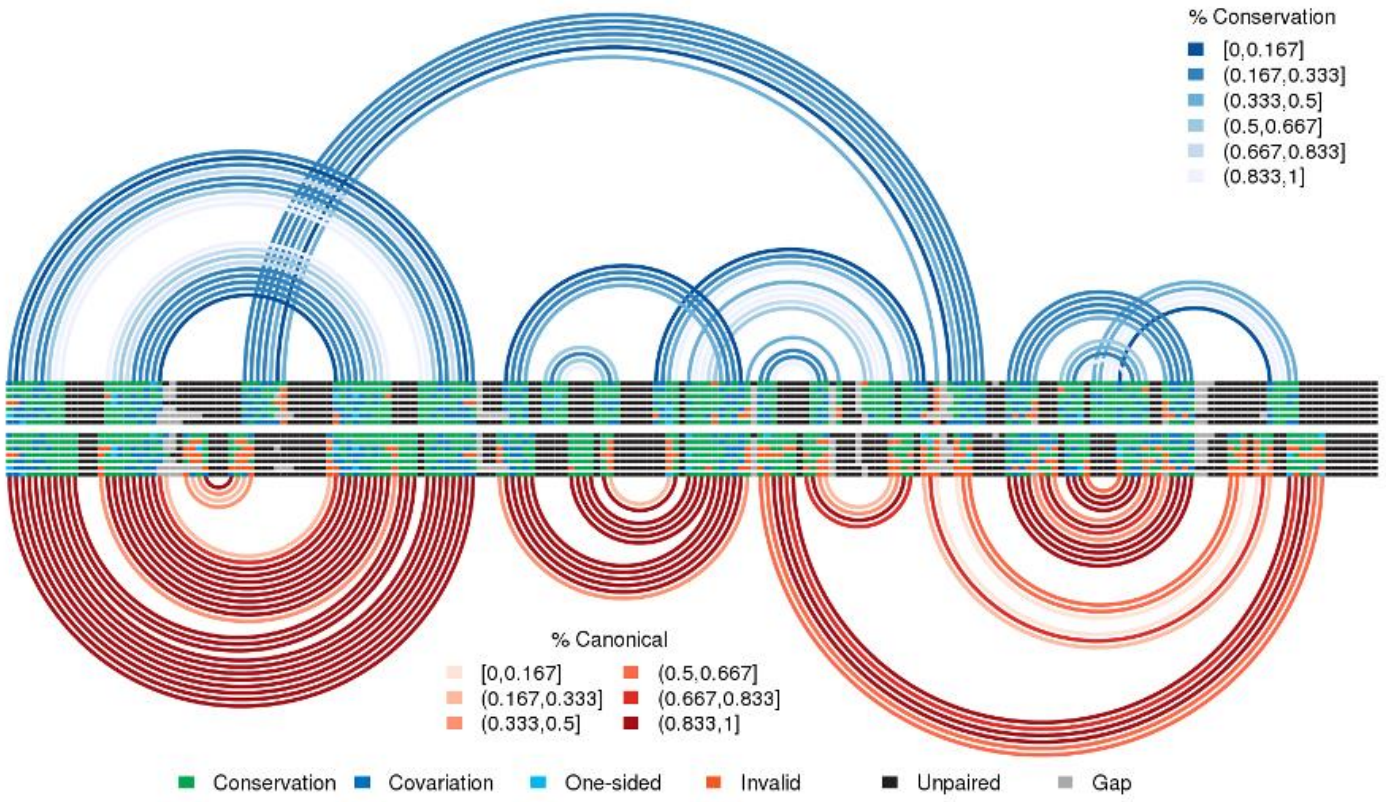
EECS730: Introduction to Bioinformatics

Lecture 10: Non-coding RNA secondary structure alignment

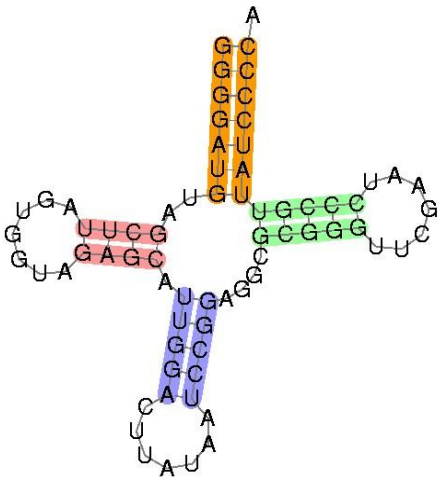


http://www.e-rna.org/r-chie/images/doublecov.png

Some slides were adapted from Dr. Shaojie Zhang (University of Central Florida)

RNA secondary structure alignment

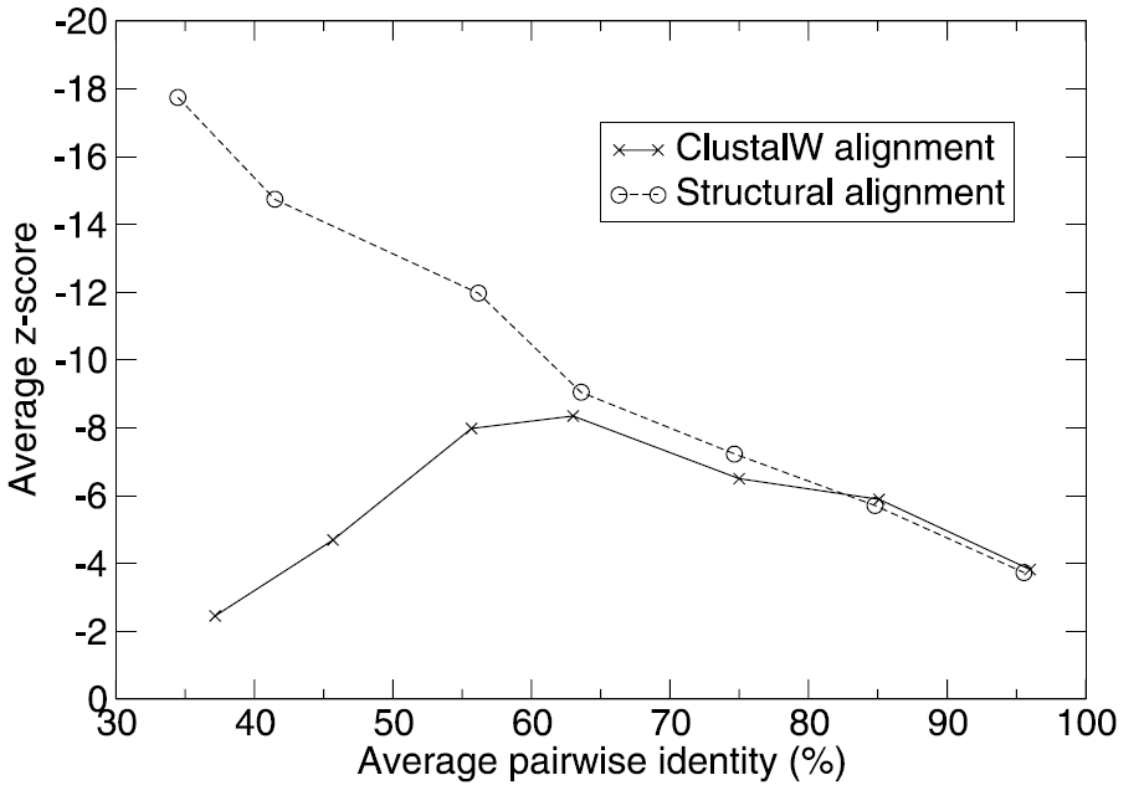
- Sequence alignment will not work for RNAs with low sequence similarity
- RNA conserved on secondary structure rather than primary sequence



Failure of sequence alignment on RNA

```
ACCCG-UUAAU
|  |||  ||*||
A-CCGUUUCAU
```

```
Stru1: >>>>    <<<<
Seq1:  GGGGCAACCCC
       ++++*||++++
Seq2:  AUCCGAAGGAU
```



(Based on pairwise alignments of SRP RNAs)

Need to consider secondary structure while comparing RNAs

- Sequence-Sequence alignment
  - Only works for high sequence similarity RNAs
- Structure-Structure alignment
  - Compute similarity of two RNA structures
- Sequence-Structure alignment
  - Can the sequence fold into a given secondary structure
- Simultaneous Alignment and folding
  - Compute the consensus structure two RNA sequences can fold into

Sequence-Sequence alignment

- Will only work for high-sequence similarity RNAs
- Can incorporate secondary structure information into sequence by designing a new alphabet
- “For each piece, RNAfold predicts whether each nucleotide is paired upstream, paired downstream or unpaired. To take advantage of fast primary sequence homology search programs, we map these sequences into a 12 letter alphabet representing nucleotide plus pairing direction.”

Tseng et al. 2007 *WSPC*

Sequence-Sequence alignment (folded-BLAST)

Table 2. CM scan best recovery motif comparison. For each ncRNA family and each homology search program used, the motif/CM that recovered the most instances of the particular family is listed. The actual motif identifications can be cross-referenced online. *sen.* is the recovery percentage (sensitivity), and *spe.* is the specificity of the CM scan; "None" indicates that no instances were recovered.

ncRNA	NCBI-BLAST		WU-BLAST		SSEARCH		*folded*-BLAST	
family	*sen.*	*spe.*	*sen.*	*spe.*	*sen.*	*spe.*	*sen.*	*spe.*
t-box	0.69	0.98	**0.71**	**0.99**	0.41	**0.99**	0.68	**0.99**
SAMI	0.94	**0.99**	**0.96**	**0.99**	0.84	**0.99**	0.94	**0.99**
TPP	0.84	**0.99**	0.95	**0.99**	0.54	**0.99**	**0.96**	**0.99**
purine	0.36	**0.99**	0.36	**0.99**	0.32	**0.99**	**0.37**	**0.99**
ylbH	0.01	0.5	0.01	**1.00**	**0.02**	**1.00**	0.01	0.33
cobalamin	None		0.84	0.82	0.72	**1.00**	**0.86**	0.99
lysine	0.79	**1.00**	**0.84**	0.82	0.72	**1.00**	0.74	**1.00**
SRP	0.1	0.99	0.1	**1.0**	**0.84**	0.98	0.77	0.98
RNaseP	**1.0**	**0.99**	**1.0**	**0.99**	**1.0**	**0.99**	**1.0**	**0.99**
FMN	**0.96**	**0.99**	**0.96**	**0.99**	**0.96**	**0.99**	**0.96**	0.98
glycine	None		0.08	0.98	None		**0.86**	**0.99**
preQ1	**0.01**	**0.04**	None		**0.01**	0.02	None	
ydaO	**0.97**	**1.00**	0.96	**1.00**	0.96	**1.00**	0.96	0.99
yybP	**0.26**	**1.00**	0.11	**1.00**	0.11	**1.00**	0.22	0.99
6S	0.09	**1.00**	**0.42**	0.92	0.09	**1.00**	0.29	**1.00**
ykoK	**0.96**	**1.00**	**0.96**	**1.00**	0.90	**1.00**	**0.96**	**1.00**
glmS	**0.95**	**1.00**	0.93	**1.00**	0.91	**1.00**	**0.95**	**1.00**
ykkC	None		None		**0.69**	**1.00**	**0.69**	**1.00**
moco	None		None		None		None	
SMK	**0.08**	**0.67**	None		None		None	
Median	0.31	**0.99**	0.56	**0.99**	0.61	**0.99**	**0.76**	**0.99**

Structure-Structure alignment

- Compare the distance of two known RNA secondary structures
- It is more useful nowadays as technologies for experimentally probing RNA secondary structures become mature
- Can be used to define distance for RNA structure clustering and defining RNA families based on conserved RNA secondary structures

Algorithm for structure-structure alignment

Procedure *AlignRNA*
begin
for intervals (i_1, j_1), $1 \leq i_1 < j_1 \leq m$
and (i_2, j_2), $1 \leq i_2 < j_2 \leq n$
(* Assume that the intervals are examined in lexicographically
increasing order of widths *)

$$Align[i_1, j_1, i_2, j_2] = \max \begin{cases} Align[i_1, j_1 - 1, i_2, j_2] + \gamma(s[j_1], `-') \\ Align[i_1, j_1, i_2, j_2 - 1] + \gamma(`-', t[j_2]) \\ Align[i_1, j_1 - 1, i_2, j_2 - 1] + \gamma(s[j_1], t[j_2]) \end{cases}$$

if there exist $i_1 \leq k_1 < j_1, i_2 \leq k_2 < j_2$
s.t. $(k_1, j_1) \in S_1, (k_2, j_2) \in S_2$

$$Align[i_1, j_1, i_2, j_2] = \max \begin{cases} Align[i_1, j_1, i_2, j_2], \\ Align[i_1, k_1 - 1, i_2, k_2 - 1] + \\ Align[k_1 + 1, j_1 - 1, k_2 + 1, j_2 - 1] \\ +\delta(k_1, j_1, k_2, j_2) + \gamma(s[k_1], t[k_2]) + \gamma(s[j_1], t[j_2]) \end{cases}$$

end

Bafna et al. 1995, DIMACS report

Complexity of Structure-Structure alignment algorithm

- O(n^4), because k1 and k2 can only take on constant values
- Zhong and Zhang further reduced the complexity to O(n^3)

Table 1 Comparison on running time of ERA, LocARNA, and RNAforester

RNA family	length (bp)	num. pairs	ERA (sec)	LocARNA (sec)	ERA vs. LocARNA (fold)	RNAforester (sec)	ERA vs. RNAforester (fold)
tRNA	78	21	0.017	0.100	5.882	0.047	2.765
Gly riboswitch	105	22	0.015	0.277	18.46	0.162	10.80
U12 spliceosome	160	42	0.035	0.311	8.886	0.657	18.77
Phage_pRNA	244	43	0.124	0.647	5.218	6.935	55.93
tmRNA	367	64	0.929	22.45	24.16	225.4	242.6
biocoid_3UTR	549	155	4.898	170.3	34.77	13.99	2.856
snR86	1004	333	53.15	4862	91.48	5.579	-9.527*
Sacc_telomerase	1162	181	23.93	522.3	21.82	3697	154.5

*ERA is slower than RNAforester when aligning snR86 RNA structures.

Zhong and Zhang, 2013, BMC Bioinformatics

Sequence-structure alignment

- Compared with structure-structure alignment, we loss the information regarding the secondary structure of one of the sequences
- Naively, we need to try all possible, say k2, for branching in the recursive function
- This naïve way gives rise to an O(n^5) algorithm

Sequence-structure alignment

- However, we notice that we know the secondary structure of one RNA sequence.
- We also observe that the branching case is only used for considering multi-branch loop
- We only compute the expensive recursive function when we know that we are handling a multi-branch loop case (by looking at the know secondary structure)
- This will reduce the complexity to O(n^4)
- Zhang et al. further reduced the complexity to O(n^3), See Zhang et al. 2005 IEEE/ACM TCBB

Sequence-structure alignment algorithm

Procedure *InferStructure*
begin
for intervals (i_1, j_1), $1 \leq i_1 < j_1 \leq n$
and intervals (i_2, j_2), $1 \leq i_2 < j_2 \leq m$
(* Assume that the intervals are examined in lexicographically increasing order of widths*)

$$Align[i_1, j_1, i_2, j_2] = \max \begin{cases} Align[i_1+1, j_1, i_2, j_2] + \gamma(s[i_1], `-') \\ Align[i_1, j_1, i_2+1, j_2] + \gamma(`-', t[i_2]) \\ Align[i_1+1, j_1, i_2+1, j_2] + \gamma(s[i_1], t[i_2]) \\ Align[i_1, j_1-1, i_2, j_2] + \gamma(s[j_1], `-') \\ Align[i_1, j_1, i_2, j_2-1] + \gamma(`-', t[j_2]) \\ Align[i_1, j_1-1, i_2, j_2-1] + \gamma(s[j_1], t[j_2]) \end{cases}$$

if $(i_1, j_1) \in S$ and
$t[i_2]$ and $t[j_2]$ are complementary base-pairs

$$Align[i_1, j_1, i_2, j_2] = \max \begin{cases} Align[i_1, j_1, i_2, j_2], \\ \delta(i_1, j_1, i_2, j_2) + \gamma(s[i_1], t[i_2]) \\ +\gamma(s[j_1], t[j_2]) + Align[i_1+1, j_1-1, i_2+1, j_2-1] \end{cases}$$

else if $(i_1, j_1) \in S` - S$ and
$(k, j_1)) = rightchild(i_1, j_1)$

$$Align[i_1, j_1, i_2, j_2] = \max \begin{cases} Align[i_1, j_1, i_2, j_2], \\ \max_{i_2 < l < j_2} \{Align[i_1, k-1, i_2, l-1] + Align[k, j_1, l, j_2]\} \end{cases}$$

end

Bafna et al. 1995, DIMACS report

Simultaneous alignment and folding

- What if we know none of the two RNA sequences' secondary structures, and we want to compute the consensus structure that can be formed by them?
- Scoring: a combination of sequence & structural similarity and thermodynamic stability of the consensus structure
- Algorithm with only similarity: Simple! Just try all k1 and k2!
- O(n^6) time complexity; unfortunately, no easy way to reduce it

Simultaneous alignment and folding

- Due to David Sankoff; also called Sankoff's algorithm
- Incorporating similarity and stability
- Under Zuker-Sankoff energy model

Sankoff's algorithm

$$
\begin{aligned}
&C(i_1, j_1; i_2, j_2) \\
&= \min \begin{cases}
e(s_1) + e(s_2) + D(i_1+1, j_1-1; i_2+1, j_2-1), & s_1, s_2 \textit{ hairpins closed by } (i_1, j_1), (i_2, j_2) \textit{ respectively,} \\
\min \{e(s_1) + e(s_2) + C(p_1, q_1; p_2, q_2) \\
\qquad + D(i_1+1, p_1; i_2+1, p_2) + D(q_1, j_1-1; q_2, j_2-1)\}, & s_1, s_2 \textit{ are 2-loops closed by } (i_1, j_1), (i_2, j_2) \textit{ with } (p_1, q_1), (p_2, q_2) \textit{ accessible,} \\
& \qquad p_1 - i_1 + j_1 - q_1 - 2 \leqq U,\ p_2 - i_2 + j_2 - q_2 - 2 \leqq U, \\
& \qquad \textit{or one of } \begin{cases} s_1 = \phi & \textit{and } (p_1, q_1) = (i_1, j_1) \\ s_2 = \phi & \textit{and } (p_2, q_2) = (i_2, j_2), \end{cases} \\
\min\limits_{\substack{i_1 < h_1 < j_1 - 1 \\ i_2 < h_2 < j_2 - 1}} \{G(i_1+1, h_1; i_2+1, h_2) \\
\qquad + G(h_1+1, j_1-1; h_2+1, j_2-1) + 2A\},
\end{cases}
\end{aligned}
$$

Sankoff, 1985, SIAM

Sankoff's algorithm cont.

$$G(i_1, j_1; i_2, j_2)$$

$$= \min \begin{cases} C(i_1, j_1; i_2, j_2) + 2P + D(i_1, i_1; i_2, i_2) + D(j_1, j_1; j_2, j_2), \\ \min\limits_{\substack{i_1 < h_1 < j_1 \\ i_2 < h_2 < j_2}} \min \begin{cases} G(i_1, h_1; i_2, h_2) + (j_1 - h_1 + j_2 - h_2)Q \\ \qquad + D(h_1 + 1, j_1; h_2 + 1, j_2), \\ G(i_1, h_1; i_2, h_2) + G(h_1 + 1, j_1; h_2 + 1, j_2), \\ (h_1 - i_1 + 1 + h_2 - i_2 + 1)Q \\ \qquad + G(h_1 + 1, j_1; h_2 + 1, j_2) + D(i_1, h_1; i_2, h_2), \end{cases} \end{cases}$$

$$F(i_1, j_1; i_2, j_2) = \min \begin{cases} C(i_1, j_1; i_2, j_2) + D(i_1, i_1; i_2, i_2) + D(j_1, j_1; j_2, j_2), \\ \min\limits_{\substack{i_1 \leqq h_1 < j_1 \\ i_2 \leqq h_2 < j_2}} \{F(i_1, h_1; i_2, h_2) + F(h_1 + 1, j_1; h_2 + 1, j_2)\}, \\ D(i_1, j_1; i_2, j_2), \end{cases}$$

Sankoff, 1985, SIAM