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: CGCCACCCUGCGAACCCGAAUAUAAUAAUAAUACAGGGAGCAGGUGGCG
  - hdrA: GGCACCACUCGAAGGCUAAGCCAAAGUGGUGCU
  - vhuD: GUUCUCUCGGGAACCCGUCAAGGGACCGAGAGAAC
  - vhuU: AGCUCACAACCGAACCCAUUUGGGAGGUUGAGCU
  - fwdB: AUGUUGGAGGGGAACCCGUAAGGGACCCUCCAAGAU
  - selD: UUACGAUGUGCCGAACCCGUUUAAGGGAGGCACUAUCGAAA
  - fruA: CCUCGAGGGGAACCCGAAAAGGGACCCGAGAGG
Comparative RNA Analysis—What?

- **compare** (potentially) homologous RNAs
  
  - fdhA  \( \text{CGCACCACCGAACCCAAUUAAUUUUACAAAGGGAGCAGCUGGUGGCG} \)
  - hdrA  \( \text{GGCCACACUCGAAGGCUCUAAGCCAAAGUUGGUGGCU} \)
  - vhuD  \( \text{GUUCUCUCGGAACCCGUGCAAGGGACCGAGAGAC} \)
  - vhuU  \( \text{AGCUCACAACCGAACCCAUUUGGGAGGUUGUGACGCU} \)
  - fwdB  \( \text{AUGUUGGAGGGGAAACCCGUAAGGGACCGUCCAAAGAU} \)
  - selD  \( \text{UUACGAUGUGCCGAACCCUUUUAAGGGAGGCAACUCGAAA} \)
  - fruA  \( \text{CCUCGAGGGGAACCCGAAAGGGACCGAGAGG} \)

- **align**
  
  - fdhA  \( \text{CGCACCACCGAACCCAAUUAAUUUUACAAAGGGAGCAGCUGGCG} \)
  - hdrA  \( \text{GGCCACACUCGAAGGCUCUAAGCCAAAGUUGGUGGCU} \)
  - vhuD  \( \text{GUUCUCUCGGAACCCGUGCAAGGGACCGAGAGAC} \)
  - vhuU  \( \text{AGCUCACAACCGAACCCAUUUGGGAGGUUGUGACGCU} \)
  - fwdB  \( \text{AUGUUGGAGGGGAAACCCGUAAGGGACCGUCCAAAGAU} \)
  - selD  \( \text{UUACGAUGUGCCGAACCCUUUUAAGGGAGGCAACUCGAAA} \)
  - fruA  \( \text{CCUCGAGGGGAACCCGAAAGGGACCGAGAGG} \)

- consider and learn about RNA structure
  
  - AGC CAC AGGGGAACCCGUAACCCGAAAGGGACCGAGAGG
  
  - (-19.48)
Comparative RNA Analysis—What?

- **compare** (potentially) homologous RNAs
  
<table>
<thead>
<tr>
<th>RNA</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>fdhA</td>
<td>CGCCACCCUGCGAACCCAAAUUAAAUUAACAAAGGGAGCAGGUGGCG</td>
</tr>
<tr>
<td>hdrA</td>
<td>GGCACCACUCGAAAGGCUAAGCCAAGGGACCGAGAGAAC</td>
</tr>
<tr>
<td>vhuD</td>
<td>GUUCUCUCGGGAAACCCGUCAAGGGACCGAGACCAAC</td>
</tr>
<tr>
<td>vhuU</td>
<td>AGCUCACAACCGAACCCAUUUGGGAGGUUGUGAGCU</td>
</tr>
<tr>
<td>fwdB</td>
<td>AUGUUGGAGGGAACCCGUAAAGGACCGACCUCAAAGAU</td>
</tr>
<tr>
<td>selD</td>
<td>UUACGAUGUGCCGAACCCUUUUAAGGGAGGACACAUAGAAA</td>
</tr>
<tr>
<td>fruA</td>
<td>CCUCGAGGGGAAACCCGAAAGGGACCCGAGG</td>
</tr>
</tbody>
</table>

- **align**

<table>
<thead>
<tr>
<th>RNA</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>fdhA</td>
<td>CGC-CACCCUGCGAACCCAAAUUAAAUUAACAAAGGGAGCAG-GUGG-CG</td>
</tr>
<tr>
<td>hdrA</td>
<td>GGC-ACC-ACUCGAAAGGCU------------------------AAGCCAAGAGU-GGGG-CU</td>
</tr>
<tr>
<td>vhuD</td>
<td>GUU-CUC-UCGGGAAACCCGU------------------------CAAGGGACCGA-GAGA-AC</td>
</tr>
<tr>
<td>vhuU</td>
<td>AGC-UCACAACCGAACCCAU------------------------UUGGAGGUUGUGAG-CU</td>
</tr>
<tr>
<td>fwdB</td>
<td>AUG-UUGGAGGGAACCCGU------------------------AAGGGACCCUCAAAG-AU</td>
</tr>
<tr>
<td>selD</td>
<td>UUACGAUGUGCCGAACCCUU------------------------UAAGGGAGGACACAUAGAAA</td>
</tr>
<tr>
<td>fruA</td>
<td>CC-UCG--AGGGGAAACCCCGA------------------------AAGGGACCC--GAGA-GG</td>
</tr>
</tbody>
</table>

- **consider and learn about RNA structure**

  AGC_CAC_AGGCGAACCAGGCU_____________AAGGGACCCU_GAGG_AU
  ((..((((((((((((((((((((((........................)))))))))))))))))))))))(-19.48)
**Comparative RNA Analysis—Why?**

- **overcome limitations** of prediction from single sequences

<table>
<thead>
<tr>
<th>Program</th>
<th>Sens</th>
<th>PPV</th>
<th>MCC</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAfold 2.1.9</td>
<td>0.742</td>
<td>0.795</td>
<td>0.767</td>
<td>0.765</td>
</tr>
<tr>
<td>UNAfold 3.8</td>
<td>0.693</td>
<td>0.767</td>
<td>0.727</td>
<td>0.725</td>
</tr>
<tr>
<td>RNAstructure 5.7</td>
<td>0.716</td>
<td>0.781</td>
<td>0.746</td>
<td>0.744</td>
</tr>
</tbody>
</table>

- 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”

- pure sequence alignment cannot properly **compare remote RNAs**
  
  “...sequence alignment alone, using the current algorithms, is generally inappropriate <50–60% sequence identity.”

---

1 Rivas, Eddy; 2001; [doi:10.1186/1471-2105-2-8]

2 Gardner, Wilm, Washietl; 2005; [doi:10.1093/nar/gki541]
Comparative RNA Analysis—How?

Plan A
- ALIGN single sequences
- FOLD alignment
- consensus structure

Plan B
- simultaneously ALIGN and FOLD
  - [Sankoff 85]
- consensus:
  - consensus structure

Plan C
- FOLD single sequences
- ALIGN sequence AND structure

adopted from:
- [Gardener & Giiegerich BMC 2004]
- [Sankoff 85]
ALIGN, then ANALYSE

- Covariation, R2R
- R-scape
- Pfold
- RNAalifold
- RNAz
- CMs, SCFGs, Infernal
• 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\(^3\)
- structure-based clustering of RNAs (RNAclust, GraphClust)

\(^3\)e.g. Ding et al.; RNA 2005; *doi:10.1261/rna.2500605*
Plan A: Align single sequences

Plan B: Align and Fold simultaneously

Plan C: Fold single sequences

Simultaneously Align and Fold

[Sankoff 85]

Consensus: Consensus structure

Adopted from:

[Gardener & Giiegerich BMC 2004]
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)
  \[ \ldots((\ldots))\ldots \]
  \[ \text{auGCaugaGCuc} \]
  \[ \text{auCCaugaGGuc} \]
  \[ \text{auCGaugaCGuc} \]
  \[ \text{auUGaugaCGuc} \]

- Inversely: compensatory mutations hint at functional structure
Mutual Information (of columns $i$ and $j$):

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

[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

Visualizations created by the RNA drawing tool R2R⁴
Covarying mutations are highlighted (green-ish)

⁴Weinberg, Breaker; 2011; doi:10.1186/1471-2105-12-3
Significance of covariation in R-scape

- 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

---

Rivas, Clements, Eddy. 2017. doi:10.1038/nmeth.4066
Significance of covariation in R-scape

- 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

---

Rivas, Clements, Eddy. 2017. doi:10.1038/nmeth.4066
Covariation and Thermodynamics: RNAalifold

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF008220</td>
<td>GGAGGAUUA-AGCUCAGCUGGGAGAGCAUCUGCUGCUAACAGC-AGAGGGUCGGCGGUUCAGCUGCCUCCA</td>
</tr>
<tr>
<td>M68929</td>
<td>GCGGAUAU-AACUUAGGGGUUAAAGUUGCAGAUUGUGGCUC-UGAAAA-CACGGGUUCGAUCCCCGUUAUUCGCC</td>
</tr>
<tr>
<td>X02172</td>
<td>GCCUUUAU-AGCUAGUGGUAAAGCAAAACUGAAGAU-UUUUACAUGUAGUUCGAUUCUCAUAAAGGCA</td>
</tr>
<tr>
<td>Z11880</td>
<td>GCCUUCCU-AGCUACAG-UGGUAGAGCGCACGGCUUUAAACC-GUGUGUGCGUGGGUUCGAAUCCCCACGGAAGGCG</td>
</tr>
<tr>
<td>D10744</td>
<td>GGAAAAUUGAUCAUCGGCAAGAAUGUUAAUACUAUAAUAUAGGAUUAAUAAACCUGGAGAUGUCGAUUCUCAUAAUUCCG</td>
</tr>
</tbody>
</table>

(-49.58 = -17.46 + -32.12)

Predict consensus structure that is
• thermodynamically good
• ideally possible for all sequences (tolerate defects)
• supported by covariation

Covariation and Thermodynamics: RNAalifold\(^6\)

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>AF008220</td>
<td>GGAGGAUU-AGCUAAGAAGGAGAGCAUCUGCCUUAACAAGC-...</td>
</tr>
<tr>
<td>M68929</td>
<td>GCGGGAUAU-ACUUAGGGGUAAGAUUGCAGAUUGGCCUC-----UGAAAA-CACGGGUUCGAUCCGUAAUUUCCG</td>
</tr>
<tr>
<td>X02172</td>
<td>GCCUUUAU-AGCUUAG-UGGUAAGCGAUAAACUGAAGAUU-----UAUUUACAUUGAUUCGAUUCUCAUAAAGGCA</td>
</tr>
<tr>
<td>Z11880</td>
<td>GCCUUCCU-AGCUACG-UGGUAGACGCACGCUUUUAAACC-----GUGUGGUCGUGGUUCGAUCCACCACGGAAGCG</td>
</tr>
<tr>
<td>D10744</td>
<td>GGAAGGAUCAUCGCAAGAAGUAAUUCUAAUAAUGGAGAUAUAAUACUACCUUGAGAUGCAGAUCACAUUUCCG</td>
</tr>
</tbody>
</table>

```
RNAalifold (((((((...(((........))))((((((.......)).........))))....(((((.......))))))))))))).
```

\((-49.58 = -17.46 + -32.12\)

Predict consensus structure that is

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

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

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

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

\[ M_{ij} = \min \left\{ M_{ij-1} + cK; M_{i+1j} + cK; C_{ij} + bK \right\} \]

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

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: $0.25\delta$
### RNAalifold Example

#### RNA Sequences

<table>
<thead>
<tr>
<th>Accession</th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC021639</td>
<td>CGAGUGGCCGAGU---GGUUAAGGCGUGCCAAUCCUACCGGCUGCG</td>
<td>82</td>
</tr>
<tr>
<td>AP000063</td>
<td>GCGGGGGUGCCCGAGCCUGGCCAAAGGGGUCGGGCUCAGGACCCGAUGGCGUAGGCCUGC</td>
<td>85</td>
</tr>
<tr>
<td>AP000397</td>
<td>UGGAGUAUAGCCAAG--UGG--UAAGGCAUCGGUUUUUGGUACCG---------GCAUGC</td>
<td>72</td>
</tr>
<tr>
<td>X03715</td>
<td>CGGAAAGUAGCUUAGCUUGG--UAGAGCACUCGGUUUGGGACCGA---------GGGGUC</td>
<td>74</td>
</tr>
<tr>
<td>U67517</td>
<td>GCCGGGGUGGGGUAGUGGCCAUCCUG---GGGACUGUGGAUCCC----------CUGAC</td>
<td>72</td>
</tr>
<tr>
<td>X99256</td>
<td>GUAAACAUAGUUUA------AUCAAAACAUUAGAUUGAAUCUAA----------CAAU</td>
<td>69</td>
</tr>
<tr>
<td>M10217</td>
<td>AGUAAAGUCAGCUA------AAAAAGCUUUUGGGCCCAUACCCCAA----------ACAU</td>
<td>69</td>
</tr>
</tbody>
</table>

#### RNA Structure

(-49.58 = -17.46 + -32.12)
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?
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 $S_i$

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

\[
\text{mean}(-31.20, -52.80, -22.00, -28.90, -35.60, -13.90, -13.90) = \frac{-25.6}{-28.33} \approx 0.91
\]
De novo ncRNA prediction—RNAz\textsuperscript{7}

Question: Given alignment, is there an ncRNA?

- is there a truly conserved structure?

- can the single sequences form stable structures?

\textsuperscript{7}Washietl, Hofacker, Stadler. 2005. doi:10.1073/pnas.0409169102
De novo ncRNA prediction—RNAz\textsuperscript{7}

*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)

\textsuperscript{7}Washietl, Hofacker, Stadler. 2005. \textit{doi:10.1073/pnas.0409169102}
De novo ncRNA prediction—RNAz

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

\[\text{\cite{Washietl, Hofacker, Stadler. 2005.} doi:10.1073/pnas.0409169102}\]
De novo ncRNA prediction—RNAlz

**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)

RNAlz 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

---

RNAz Screen

Whole genome alignment

Slice into overlapping windows and clean

Estimate ncRNA class probability and filter

Combine overlapping windows into loci

RNAz candidate ncRNAs
RNA families: CMs—Infernal\textsuperscript{8}—Rfam\textsuperscript{9}

- *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 ⇒ Infernal full alignments

- models RNA families by *Stochastic Context Free Grammars (SCFGs)* as *Consensus Models (CMs)*

\textsuperscript{8}Nawrocki, Eddy. 2013. \textit{doi:10.1093/bioinformatics/btt509}

\textsuperscript{9}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)
Infernal Consensus Models

"split set" inserts

"split set" inserts

"split set" insert

MATP 6

MATP 7

MATR 8

MATL 2

MATL 3

MATL 4

MATL 5

MATL 6

MATL 7

MATL 8

MATL 9

MATL 10

MATL 11

MATL 12

MATL 13

MATL 14

MATL 15

MATL 16

MATL 17

MATL 18

MATL 19

MATL 20

MATL 21

MATL 22

MATL 23

MATL 24

BEGL 4

BIF 5

BEGR 15

END 14
Comparative RNA Analysis—How?

Plan A

ALIGN single sequences

Plan B

ALIGN and FOLD simultanously [Sankoff 85]

Plan C

FOLD single sequences

ALIGN sequence AND structure

consensus structure

consensus:

adopted from: [Gardener & Giiegerich BMC 2004]
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$^{10}$

Given: $A = \text{GCUGACGACGCACGCUCAUCGGUAAAUUCUACCAGAUCGUCAACGACU}$
$\& B = \text{AUUGCCGUGACCGGCAACGCAUCGGAAUCCCGAUCGGGUCAGCGGCA}$

Find:

sequence similarity + energy $A$ + energy $B \rightarrow \text{opt}$

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

$^{10}$Sankoff, 1985
Simultaneous Alignment and Folding

Given: \[ A = \text{GCUGACGAGCAGCUCUCAUCGGUAAAUACUACCGAUUCGUCAGCACU} \]
& \[ B = \text{AUUGCCGCUUGACCGCAUCGCAUCGAAUCCCGAUCCCAGGUUCAGCGGCA} \]

Find:

sequence similarity + energy A + energy B → opt

where alignment, structure A, & structure B are compatible

\[ ^{10} \text{Sankoff, 1985} \]
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

\[ O(n^6) = \text{“extreme computational cost”} \]
PMcomp’s Trick – Lightweight SA&F

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

- energies composed of loop energies

---

11 Hofacker et al., 2004. doi:10.1093/bioinformatics/bth229
PMcomp’s Trick – Lightweight SA&F

Sankoff: \( \text{sequence similarity} + \text{energies of A and B} \rightarrow \text{opt} \)

- \textbf{energies} composed of loop energies

\(^{11}\)Hofacker et al., 2004. doi:10.1093/bioinformatics/bth229
PMcomp’s Trick – Lightweight SA&F\(^\text{11}\)

PMcomp: \textbf{sequence similarity} \\
\begin{itemize}
  \item \textbf{pseudo-energies} composed of “base pair energies”
\end{itemize}

---

\(^{11}\)Hofacker et al., 2004. \textit{doi:10.1093/bioinformatics/bth229}
PMcomp’s Trick – Lightweight SA&F

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

---

11 Hofacker et al., 2004. doi:10.1093/bioinformatics/bth229
PMcomp’s Trick – Lightweight SA&F\textsuperscript{11}

PMcomp: \hspace{1em} \textbf{sequence similarity} \hspace{1em} + \hspace{1em} \textbf{pseudo-energies of A and B} \rightarrow \textbf{opt}

- \textbf{pseudo-energies} composed of “base pair energies”

- Dynamic Programming
  - Base Pair Maximization [Nussinov] \times \text{Sequence Alignment}

- cheaper computation (at same complexity)

\textsuperscript{11}Hofacker et al., 2004. \textit{doi:10.1093/bioinformatics/bth229}
PMcomp: Nussinov-style Sankoff — Recursion

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

\[ D_{ij;kl} = M_{i+1,j-1;k+1,l-1} + \tau(i,j,k,l) \]
PMcomp — Scoring

\[
M_{i\,j;\,k\,l} = \max \left\{ \begin{array}{l}
M_{i\,j-1;\,k\,l-1} + \sigma(A_j, B_l) \\
M_{i\,j-1;\,k\,l} + \gamma \\
M_{i\,j;\,k\,l-1} + \gamma \\
\max_{j',l'} M_{i\,j'-1;\,k\,l'-1} + D_{j'\,j;\,l'\,l} \end{array} \right. \\
\]

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

Idea:

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

Complexity PMcomp

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

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

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

Total Complexity: \( O(n^3m^3) \) time, \( O(n^2m^2) \) space
LocARNA\textsuperscript{12}: 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

\textsuperscript{12}Will et al., 2007. \textit{doi: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
LocARNA Basic Algorithm: Matrices
LocARNA Basic Algorithm: Matrices
LocARNA Basic Algorithm: Recursion

\[ D(a,b) = M(a,b;ar-1,br-1) + \tau(a,b) \]
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

\[ M^a_b(i, j) = \max \begin{cases} 
M^a_b(i - 1, j - 1) + \sigma(A_i, B_j) \\
M^a_b(i - 1, j) + \gamma \\
M^a_b(i, j - 1) + \gamma \\
\max_{a', b'} M^a_b(a'_l - 1, b'_l - 1) + D(a', b') 
\end{cases} \]

where \( a'_r = i, b'_r = j \)

\[ D(a, b) = M^a_b(a_r - 1, b_r - 1) + \tau(a, b) \]

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

- compute \( D(a, b) \) for all base pair edges:
  \[ \Rightarrow O(|P_1||P_2|) = O(nm) \text{ pairs of base pairs } (a,b) \]
- \( O(nm \cdot rdeg_1 rdeg_2) = O(nm) \text{ time per } (a, b) \)

**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\textsuperscript{13})
- stronger sparsification and added structural flexibility (SPARSE\textsuperscript{14})

\textsuperscript{13}Will et al., 2012. doi:10.1261/rna.029041.111
\textsuperscript{14}Will et al., 2015. doi: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)
Unaligned sequences, unknown structures:

>fruA
CCUCGAGGGGAAACCGAAAGGGACCCGAGAGG
>fdhA
CGCCACCCUGCAACCCCAAAUUAUAAAUAUAAUCAAGGGAGCAGGUGGCG
>vhuU
AGCUCAACAACCGAACCCAUUUGGGAGGUUGUGAGCU
>hdrA
GGCACCACUGCAAGGCUAACCCAAUGGUGGUGCU
>vhuD
GUUCUCUGGGAACCGUCAAGGGACCGAGAAC
>selD
UUACGAUGUGCACACCCUCUUUAAGGGAGGCGACACUACGAAA
>fwdB
AUGUUGGAGGGGAAACCGUAAAGGGACCCUCCAAGAU
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,fwdB),selD),(hdrA,vhuD)),fruA),fdhA));
```

```
 vhuU AG-CUCACAACGAACCCAUU-------------UGGGAGGUUGUGAGCU- 36
 fwdB AU-GUUGGAGGGGAACCCGUA-------------AGGGACCCUCCAAGAU- 36
 selD UUACGAUGUGCCGAACCCUUU------------AAGGGAGGCACAUCGAAA 39
 hdrA G--GCACCACUCGAAGGC--U------------AAGCCAAAGUGGUGCU-- 33
 vhuD G--UUCUCUCGGGAACCCGUC------------AAGGGACCGAGAGAAC-- 35
 fruA ---CCUCGAGGGGAACCCG-A------------AAGGGACCCGAGAGG--- 32
 fdhA CG-CCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-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:

![Graph showing STAR profile](image-url)
• clustering structures of one RNA\textsuperscript{3}
• structure-based clustering of RNAs (RNAclust, GraphClust)

\textsuperscript{3}e.g. Ding et al.; RNA 2005; \textit{doi: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?

![Diagram of a dendrogram with a single cluster highlighted.](image-url)
**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\textsuperscript{15} in RNAclust\textsuperscript{16}

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
  \[
  \Delta(C) = \sum_{i \in C} (E_i - E_{cons}(C))^2
  \]
- \[\frac{\Delta(C_1) + \Delta(C_2)}{\Delta(C)} < \theta\], then reject
  e.g. we could achieve MCC 0.8 in an evaluation on Rfam

\textsuperscript{15}Duda et al. Pattern Classification, 2001
\textsuperscript{16}http://www.bioinf.uni-leipzig.de/~kristin/Software/RNAclust/
Clustering of 3332 putative ncRNAs in *Ciona intestinalis*
Clustering of 3332 putative ncRNAs in *Ciona intestinalis*

- putative ncRNAs from RNAz screen
- requires $3332 \cdot \frac{3331}{2} \approx 5.5 \times 10^6$ LocARNA alignments
- e.g. 16,000 predicted ncRNAs in Drosophila; 37,000 in Human
GraphClust\textsuperscript{17}: Workflow and Results

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

\textsuperscript{17}Heyne et al., 2012. doi:10.1093/bioinformatics/bts224
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* $h_1, \ldots, h_{400}$, 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 $J(x, y)!$

$\Rightarrow$ 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

Please prepare for the hands-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/
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/