Comparative approach to discovering ncRNA
Project Question

• Akutsu’s algorithm gives a recursive formulation for pseudoknots.
• Sequence-structure alignments enable structure prediction through comparison to a homologous RNA (non-pseudoknotted) structure.
• Can sequence-structure alignments be extended to handle pseudoknots?
Comparative prediction of ncRNA

• In genome-genome comparisons, many sequences are found to be conserved.
• Can you use the pattern of conservation to detect if these are ncRNA sequences?
• QRNA is a software to do that.
QRNA: Approach

- Compute the 3 probabilities
  - $\Pr(\overline{XY}|\text{COD})$
  - $\Pr(\overline{XY}|\text{RNA})$
  - $\Pr(\overline{XY}|\text{OTH})$
\( \Pr(\overline{XY}|RNA) \)

- \( \Pr(\overline{XY}|RNA) = \sum_s \Pr(\overline{XY}|s,RNA) \Pr(s|RNA) \)
- While there are many structures, this expression can be computed efficiently.
- We start by describing a different formalism for computing RNA structure.
- We will show that the probabilistic, and energy frameworks are essentially equivalent.
Stochastic Context Free Grammars

• Consider the following CFG:
  - $S \rightarrow W$ (* Start *)
  - $W \rightarrow xWy \quad x,y \in \{A,C,G,U\}$ (* base-pairing *)
  - $W \rightarrow xW$ (* unpaired bases *)
  - $W \rightarrow WW$ (* branching *)
  - $W \rightarrow \square$ (* termination *)

• The CFG generates RNA sequences, with an associated structure.
Example
Computing RNA structures

• Consider the inverse problem: Given an RNA string, find the best parse (a sequence of Context Free rules that generate the sequence).

• This is equivalent to computing structure.
Computing the optimum parse
Stochastic Context Free Grammars

• Associate a probability with each rule. Hence, SCFG.
  - \( \text{Pr}(S \to W) \) (* Start *)
  - \( \text{Pr}(W \to x W y) \) \( x,y \in \{A,C,G,U\} \) (* base-pairing *)
  - \( \text{Pr}(W \to x W) \) (* unpaired bases *)
  - \( \text{Pr}(W \to WW) \) (* branching *)
  - \( \text{Pr}(W \to \square) \) (* termination *)

• Let \( a_{ij} \) be the probability that the RNA subsequence \( s[i..j] \) was generated by the SCFG.
  - \( a_{ij} = \text{Pr}(s[i..j] \mid \text{SCFG}) = \sum_\square \text{Pr}(s[i..j] \mid \square, \text{SCFG}) \text{Pr}(\square \mid \text{SCFG}) \)

• It is sufficient to compute \( a_{ij} \) for all \( i,j \).
Computing $\mathbf{v}_{i,j}$

- $\mathbf{v}_{i,j} = \Pr(W \to s[i] W s[j]) \mathbf{v}_{i+1,j-1}$
  + $\Pr(W \to s[i] W) \mathbf{v}_{i+1,j}$
  + $\Pr(W \to W s[j]) \mathbf{v}_{i,j-1}$
  + $\sum_k \Pr(W \to WW) \mathbf{v}_{i,k-1} \mathbf{v}_{k,j}$

- Computing the most likely parse:
- $v_{i,j} = \max \{ \Pr(W \to s[i] W s[j]) v_{i+1,j-1}, \Pr(W \to s[i] W) v_{i+1,j}, \Pr(W \to W s[j]) v_{i,j-1}, \max_k \Pr(W \to WW) v_{i,k-1} v_{k,j} \}$
SCFGs versus Energy minimization

• The two approaches (most likely parse and energy minimization) give equivalent answers.

• The full-likelihood function (\(\mathcal{L}\)) might sometimes be more meaningful from the max-likelihood parse. It helps answer the question: is the string s an RNA sequence?

• The probabilistic approach makes it easier to train parameters (using Bayesian methods).
Probability of RNA alignments

- How can we compute

\[ \Pr(\overline{XY}|\text{RNA}) \]

\[ = \sum \Pr(\overline{XY}|\overline{f}, \text{RNA}) \Pr(\overline{f}|\text{RNA}) \]
Probability of an RNA alignment

\[ P_{i,j} = \text{Pr}(W \rightarrow A[i] W A[j]) P_{i+1,j-1} + \text{Pr}(W \rightarrow A[i] W) P_{i+1,j} + \text{Pr}(W \rightarrow W A[j]) P_{i,j-1} + \sum_k \text{Pr}(W \rightarrow WW) P_{i,k-1} P_{k,j} \]
Computing RNA emission probabilities

\[ P^{RNA}(a_L a_R b_L b_R | t) = P(b_R | a_L a_R b_L t) P(a_L a_R b_L | t), \]
\[ = P(b_R | a_L a_R b_L t) P(a_R | a_L b_L t) P^{OTH}(a_L b_L | t), \]
\[ \approx p^{pair}(b_R | b_L t) p^{pair}(a_R | a_L t) P^{OTH}(a_L b_L | t), \]
\[ = \frac{p^{pair}(b_L b_R | t) p^{pair}(a_L a_R | t) P^{OTH}(a_L b_L | t)}{P(b_L | t) P(a_L | t)}. \]
Other models in QRNA

\[ P(\overline{XY}|\text{COD}) = \sum_f P(\overline{XY}|f, \text{COD})P(f|\text{COD}), \]

\[ P^{\text{COD}}(a_1a_2a_3, b_1b_2b_3|t) \approx \sum_{A,B} P(a_1a_2a_3|A)P(b_1b_2b_3|B)P(A,B|t), \]
Is the sequence RNA, coding or OTH?

• $\Pr(XY \mid \text{Model})$ can be computed for the 3 models (RNA, COD, OTH)
• $\Pr(\text{Model}_i \mid XY) = P(XY \mid \text{model}_i) \cdot P(\text{model}_i)/P(XY)$
• $P(XY) = \sum_j P(XY \mid \text{model}_j) \cdot P(\text{model}_j)$
QRNA results

- Multiple alignment of 63 Eukaryotic SRP-RNAs, and 52 RNaseP RNA
  - Use pair-wise alignments from the structural alignment
    - Alignments are classified according to sequence diversity
  - Use each sequence as query to Blast against other family members
- Sensitivity: fraction pairs predicted to be RNA
- Specificity: 1-(fraction predicted to be RNA after shuffling)
### Sensitivity and Specificity

<table>
<thead>
<tr>
<th>% ID</th>
<th># align</th>
<th>% sensitivity</th>
<th>% specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 &lt; 10</td>
<td>140</td>
<td>42.8 (60)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>10 &lt; 20</td>
<td>827</td>
<td>59.6 (493)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>20 &lt; 30</td>
<td>503</td>
<td>71.4 (359)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>30 &lt; 40</td>
<td>764</td>
<td>75.1 (574)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>40 &lt; 50</td>
<td>283</td>
<td>58.6 (166)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>50 &lt; 60</td>
<td>434</td>
<td>81.3 (353)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>60 &lt; 70</td>
<td>88</td>
<td>80.7 (71)</td>
<td>100.0 (0)</td>
</tr>
<tr>
<td>70 &lt; 80</td>
<td>70</td>
<td>91.4 (64)</td>
<td>97.1 (2)</td>
</tr>
<tr>
<td>80 &lt; 90</td>
<td>73</td>
<td>97.3 (71)</td>
<td>79.4 (15)</td>
</tr>
<tr>
<td>90 &lt; 100</td>
<td>61</td>
<td>93.4 (57)</td>
<td>27.9 (44)</td>
</tr>
<tr>
<td>100</td>
<td>99</td>
<td>93.9 (93)</td>
<td>29.3 (70)</td>
</tr>
</tbody>
</table>
QRNA results: experiment 2

• Each of the sequences was chosen in turn, and compared against members of its own family (WU-Blastn2).
  • Poor quality of alignments
  • Bias towards conserved sequences
  • 1003 (out of 3342 pairs) alignments were selected
Table 3: Similar analysis to the one presented in Table 2 for 586 BLASTN alignments of SRP RNAs and 417 BLASTN alignments of RNaseP RNAs.

<table>
<thead>
<tr>
<th>% ID</th>
<th># alignments</th>
<th>% sensitivity</th>
<th>% specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>% ID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 &lt; 70</td>
<td>419</td>
<td>15.3 (64)</td>
<td>99.5 (2)</td>
</tr>
<tr>
<td>70 &lt; 80</td>
<td>269</td>
<td>26.8 (72)</td>
<td>98.5 (4)</td>
</tr>
<tr>
<td>80 &lt; 90</td>
<td>131</td>
<td>61.1 (80)</td>
<td>89.5 (19)</td>
</tr>
<tr>
<td>90 &lt; 100</td>
<td>78</td>
<td>97.4 (76)</td>
<td>67.9 (53)</td>
</tr>
<tr>
<td>100</td>
<td>106</td>
<td>92.4 (98)</td>
<td>24.5 (80)</td>
</tr>
</tbody>
</table>

% GC
QRNA: Results

- Comparison of E. coli and S. typhii
- E. coli was partitioned into 115 RNA, 4290 ORFs, and 2367 intergenic
- Each region blasted against S. typhii, and QRNA was used on “quality” alignments
  - 354 alignments to RNA
  - 4946 to ORFs
  - 11509 alignments to intergenic regions (Repeats?)
## Genomic comparison Results

<table>
<thead>
<tr>
<th></th>
<th>ncRNA</th>
<th>RNA regions annotated</th>
<th>ORF</th>
<th>RNA regions annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>148</td>
<td>33</td>
<td>7422</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>115 OTH</td>
<td>3937 COD</td>
<td></td>
<td>3397 COD</td>
</tr>
<tr>
<td>intergenic</td>
<td>1974</td>
<td>351 COD</td>
<td>1562</td>
<td>OTH</td>
</tr>
<tr>
<td></td>
<td>61 OTH</td>
<td>1562 OTH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Blastn does not produce good alignments from a structural viewpoint.
- Can we use paired SCFGs to redo the alignment and the structure? In principle, yes, but it is expensive.
- Rivas and Eddy did not use a true comparison of orthologs. Would that help?
Computing Structural alignments

For all intervals \((i,j)\) in \(s_{1..n}\), and \((k,l)\) in \(t_{1..m}\)

\[
S(i, j, k, l) = \max \left\{ \begin{array}{l}
S(i + 1, j, k + 1, l) + g(i, j, k, l) \\
S(i + 1, j, k, l) + g(s[i], t[j]) \\
S(i + 1, j, k + 1, l) + g(s[i], t[j]) \\
\vdots \\
\max_{j', l'} \{ S(i, j', k, l') + S(j', k, k', l') \} 
\end{array} \right. 
\]
Project Question

• Can you improve upon QRNA with the following:
  - Structural alignments to obtain better results
  - Filtering to make search efficient. Most pairs should be discarded without computing a structural alignment.
RNA multiple alignments

• Why should we compute multiple (structural) alignments for RNA?
Structural Alignment

Conserved sequences, and conserved structure are more apparent in multiple alignments.
Computing Structural Alignments

• Analogy: In sequence alignment, the score for aligning a column is position independent.
• In profiles, or HMMs, position specific scoring is used to distinguish conserved positions from non-conserved positions
• Similar ideas can be used for RNA.
Covariance models=RNA profiles

A

A

A

A

U

U

A

U

S

\downarrow

W_1

\downarrow

\downarrow

a W_2

a W_4 b

\downarrow

W_3 b

\downarrow

\downarrow

a W_4 b

\vdots

\vdots

\vdots

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
<th>A</th>
<th>U</th>
<th>U</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>U</td>
<td>U</td>
<td>-</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>U</td>
<td>-</td>
<td>A</td>
</tr>
</tbody>
</table>
Aligning a sequence to a covariance model

- We align each node of the covariance model (it is tree like, but may be a graph).
- The alignment score follows the same recurrence as in Lecture 7, but with position specific probabilities.
- Example:
  \[ A[W_i,(i,j)] = -\log \left( \Pr[W_i \rightarrow s[i] W_j s[j]] + A[W_j,(i+1,j-1)] \right) \]
- If we wish to compute the probability that a sequence belongs to a family, we compute the total likelihood (sum over all probabilities)
- If we wish to compute the structure of an unknown sequence by comparison to a covariance model, we compute the max likelihood parse in this graph.
Covariance models and ncRNA discovery

- Given a family of ncRNA sequences, scan a genomic sequence with a covariance model and retrieve all high scoring sub-sequences.
- This is the most common method, but it is expensive.
- Assume covariance model has $m$ states, and the substring has at most $n$ symbols, and the database has $L$ symbols.
- Alignment cost = $O(n^2m_1+n^3m_2)$
- Total time =?
Computing covariance models

• If we are given a CM, a multiple structural alignment is 'easy'.
  - In turn, align each sequence to the CM.
• If we are given a multiple alignment, computing the covariance model is easy
• For simultaneous prediction, a Bayesian iterative approach is used
  - Compute a seed alignment
  - Use the alignment to compute a CM
  - Use the CM to compute a new alignment
  - Iterate
Project

• **Compute an RNA multiple alignment.**
• **Existing methods do not work well without good seed alignment, and require excessive hand curation.**