RNA Search and Motif Discovery

CSE 427
Computational Biology
Many biologically interesting roles for RNA
RNA secondary structure prediction
Many interesting RNAs, e.g. Riboswitches.
Approaches to Structure Prediction

Maximum Pairing
+ works on single sequences
+ simple
- too inaccurate

Minimum Energy
+ works on single sequences
- ignores pseudoknots
- only finds “optimal” fold

Partition Function
+ finds all folds
- ignores pseudoknots
“Optimal pairing of $r_i \ldots r_j$”

Two possibilities

j Unpaired:
Find best pairing of $r_i \ldots r_{j-1}$

j Paired (with some $k$):
Find best $r_i \ldots r_{k-1}$ + best $r_{k+1} \ldots r_{j-1}$ plus 1

Why is it slow?
Why do pseudoknots matter?
B(i,j) = # pairs in optimal pairing of \( r_i \ldots r_j \)

B(i,j) = 0 for all \( i, j \) with \( i \geq j-4 \); otherwise

B(i,j) = max of:

\[
\begin{cases}
    B(i,j-1) \\
    \max \{ B(i,k-1) + 1 + B(k+1,j-1) \mid i \leq k < j-4 \text{ and } r_k - r_j \text{ may pair} \}
\end{cases}
\]

Time: \( O(n^3) \)

Loop-based energy version is better; recurrences similar, slightly messier
Loop-based Energy Minimization

Detailed experiments show it’s more accurate to model based on loops, rather than just pairs.

Loop types

1. Hairpin loop
2. Stack
3. Bulge
4. Interior loop
5. Multiloop
Zuker: Loop-based Energy, 1

\( W(i,j) = \) energy of optimal pairing of \( r_i \ldots r_j \)

\( V(i,j) = \) as above, but forcing \( (i.e., \ subset \ with) \ pair \ i \cdot j \)

\( W(i,j) = V(i,j) = \infty \) for all \( i, j \) with \( i \geq j-4 \)

\[ W(i,j) = \min( W(i,j-1), \min \{ W(i,k-1)+V(k,j) \mid i \leq k < j-4 \} ) \]
Zuker: Loop-based Energy, II

\[ V(i,j) = \min(\text{eh}(i,j), \text{es}(i,j) + V(i+1,j-1), \text{VBI}(i,j), \text{VM}(i,j)) \]

\[ \text{VM}(i,j) = \min \{ W(i,k) + W(k+1,j) \mid i < k < j \} \]

\[ \text{VBI}(i,j) = \min \{ \text{ebi}(i,j,i',j') + V(i', j') \mid \]
\[ \quad i < i' < j' < j \land i'-i+j-j' > 2 \} \]

Time: \( O(n^4) \)

\( O(n^3) \) possible if \( \text{ebi}(.) \) is “nice”
Single Seq Prediction Accuracy

Mfold, Vienna,... [Nussinov, Zuker, Hofacker, McCaskill]

Estimates suggest ~50-75% of base pairs predicted correctly in sequences of up to ~300nt

Definitely useful, but obviously imperfect
Approaches, II

Comparative sequence analysis
  + handles all pairings (potentially incl. pseudoknots)
  - requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars
  Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystallography, NMR)
Covariation is strong evidence for base pairing
Example: Ribosomal Autoregulation

Excess L19 represses L19 (RF00556; 555-559 similar)
Mutual Information

\[ M_{ij} = \sum_{x_i, x_j} f_{x_i, x_j} \log_2 \frac{f_{x_i, x_j}}{f_{x_i} f_{x_j}}; \quad 0 \leq M_{ij} \leq 2 \]

Max when no seq conservation but perfect pairing

MI = expected score gain from using a pair state (below)

Finding optimal MI, (i.e. opt pairing of cols) is hard(?)

Finding optimal MI without pseudoknots can be done by dynamic programming
**M.I. Example (Artificial)**

<table>
<thead>
<tr>
<th>MI:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.55</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.42</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Cols 1 & 9, 2 & 8: perfect conservation & *might* be base-paired, but unclear whether they are. M.I. = 0

Cols 3 & 7: *No* conservation, but always W-C pairs, so seems likely they do base-pair. M.I. = 2 bits.

Cols 7->6: unconserved, but each letter in 7 has only 2 possible mates in 6. M.I. = 1 bit.
Figure 10.6 A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary contacts observed in the yeast tRNA-Phe crystal structure. Some of these tertiary contacts produce correlated pairs which can be seen weakly in the mutual information plot.
Problem: Find best (max total MI) pseudo-knot-free subset of column pairs among $i...j$.

Solution: “Just like Nussinov/Zucker folding”

$$S_{i,j} = \max \begin{cases} S_{i,j-1} \\ \max_{i \leq k < j-4} S_{i,k-1} + M_{k,j} + S_{k+1,j-1} \end{cases} \quad \begin{cases} j \text{ unpaired} \\ j \text{ paired} \end{cases}$$

BUT, need the right data—enough sequences at the right phylogenetic distance
Computational Problems

How to predict secondary structure
How to model an RNA “motif” (l.e., sequence/structure pattern)
Given a motif, how to search for instances
Given (unaligned) sequences, find motifs
How to score discovered motifs
How to leverage prior knowledge
Motif Description
RNA Motif Models

“Covariance Models” (Eddy & Durbin 1994)
aka profile stochastic context-free grammars
aka hidden Markov models on steroids
Model position-specific nucleotide preferences and base-pair preferences

Pro: accurate
Con: model building hard, search slow
Eddy & Durbin 1994: What

A probabilistic model for RNA families
  The “Covariance Model”
  \approx A Stochastic Context-Free Grammar
  A generalization of a profile HMM

Algorithms for Training
  From aligned or unaligned sequences
  Automates “comparative analysis”
  Complements Nusinov/Zucker RNA folding

Algorithms for searching
Main Results

Very accurate search for tRNA

(Precursor to tRNAscanSE – a very good tRNA-finder)

Given sufficient data, model construction comparable to, but not quite as good as, human experts

Some quantitative info on importance of pseudoknots and other tertiary features
Probabilistic Model Search

As with HMMs, given a sequence:
   You calculate likelihood ratio that the model could generate the sequence, vs a background model
   You set a score threshold
   Anything above threshold → a “hit”

Scoring:
   “Forward” / “Inside” algorithm - sum over all paths
   Viterbi approximation - find single best path
   (Bonus: alignment & structure prediction)
Example: searching for tRNAs
Profile HMM Structure

Figure 5.2 The transition structure of a profile HMM.

Mj: Match states (20 emission probabilities)
Ij: Insert states (Background emission probabilities)
Dj: Delete states (silent - no emission)
How to model an RNA “Motif”?  

Conceptually, start with a profile HMM:   
from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position   
given a new seq, estimate likelihood that it could be generated by the model, & align it to the model
How to model an RNA “Motif”? 

Add “column pairs” and pair emission probabilities for base-paired regions
Figure 5.2  The transition structure of a profile HMM.

Mj:  Match states (20 emission probabilities)
Ij:  Insert states (Background emission probabilities)
Dj:  Delete states (silent - no emission)
CM Structure

A: Sequence + structure

B: the CM “guide tree”

C: probabilities of letters/ pairs & of indels

Think of each branch being an HMM emitting both sides of a helix (but 3’ side emitted in reverse order)
CM Viterbi Alignment
(the “inside” algorithm)

\( x_i \) = \( i^{th} \) letter of input

\( x_{ij} \) = substring \( i, ..., j \) of input

\( T_{yz} \) = \( P(\text{transition } y \rightarrow z) \)

\( E_{x_i, x_j}^y \) = \( P(\text{emission of } x_i, x_j \text{ from state } y) \)

\( S_{ij}^y \) = \( \max_{\pi} \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi) \)
## CM Viterbi Alignment

*(the “inside” algorithm)*

\[ S_{ij}^y = \max_\pi \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi) \]

\[
S_{ij}^y = \begin{cases} 
\max_z [S_{i+1, j-1}^z + \log T_{yz} + \log E_{x_i, x_j}^y] & \text{match pair} \\
\max_z [S_{i+1, j}^z + \log T_{yz} + \log E_{x_i}^y] & \text{match/insert left} \\
\max_z [S_{i, j-1}^z + \log T_{yz} + \log E_{x_j}^y] & \text{match/insert right} \\
\max_z [S_{i, j}^z + \log T_{yz}] & \text{delete} \\
\max_{i < k \leq j} [S_{i, k}^{y_{left}} + S_{k+1, j}^{y_{right}}] & \text{bifurcation}
\end{cases}
\]

Time $O(qn^3)$, $q$ states, seq len $n$

compare: $O(qn)$ for profile HMM
Primary vs Secondary Info

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Avg id</th>
<th>Min id</th>
<th>Max id</th>
<th>ClustalV accuracy</th>
<th>1° info (bits)</th>
<th>2° info (bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEST</td>
<td>.402</td>
<td>.144</td>
<td>1.00</td>
<td>64%</td>
<td>43.7</td>
<td>30.0-32.3</td>
</tr>
<tr>
<td>SIM100</td>
<td>.396</td>
<td>.131</td>
<td>.986</td>
<td>54%</td>
<td>39.7</td>
<td>30.5-32.7</td>
</tr>
<tr>
<td>SIM65</td>
<td>.362</td>
<td>.111</td>
<td>.685</td>
<td>37%</td>
<td>31.8</td>
<td>28.6-30.7</td>
</tr>
</tbody>
</table>

Disallowing / allowing pseudoknots

\[
\left( \sum_{i=1}^{n} \max_j M_{ij} \right) / 2
\]
An Important Application: Rfam

A Database of RNA Families
RF00037: Example Rfam Family

Input (hand-curated):
- MSA “seed alignment”
- SS_cons
- Score Thresh T
- Window Len W

Output:
- CM
- scan results & “full alignment”
- phylogeny, etc.

**IRE (partial seed alignment):**

| Hom. sap. | GUUCUGCUUCACAGUGUUUGGAUGGAAC |
| Hom. sap. | UUUCUUCUCAACAGUGUUUGGAUGGAAC |
| Hom. sap. | UUUCUGUUUCAACAGUGCUUGGA.GGAAC |
| Hom. sap. | UUUAUC..AGUGACAGAGUUCACU.AUAAA |
| Hom. sap. | UCUCUUUGCUUCACAGUGUUUGGAUGGAAC |
| Hom. sap. | AUUAUC..GGGAACAGUGUUUCCC.AUAAU |
| Hom. sap. | UCUUGC..UCCAACAGUGUUUGGACGGAAG |
| Hom. sap. | UGUAUC..GGAGACAGUGAUCUCC.AUAAUG |
| Hom. sap. | AUUAUC..GGAGACAGUGCUCCUCCC.AUAAU |
| Cav. por. | UCUCCUCUUCACAGUGCUUGGACGGAAC |
| Mus. mus. | UAUAGACAGUGAUCUCC.AUAAUG |
| Mus. mus. | UUCCCUGCUUCACAGUGCUUGGAACGGAAC |
| Mus. mus. | GUACUUGCUUCACAGUGUUUGGACGGAAC |
| Rat. nor. | UAUAGACAGUGAUCUCC.AUAAUG |
| Rat. nor. | UAUCUUGCUUCACAGUGUUUGGACGGAAC |

SS_cons <<<<<<...<<<<<<......>>>>>.>>>>>
Rfam – an RNA family DB
Griffiths-Jones, et al., NAR ’03, ’05, ’08, ’11, ’12

Was biggest scientific comp user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:

<table>
<thead>
<tr>
<th>Release</th>
<th>Date</th>
<th>Families</th>
<th>Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel 1.0</td>
<td>1/03</td>
<td>25</td>
<td>55k</td>
</tr>
<tr>
<td>Rel 7.0</td>
<td>3/05</td>
<td>503</td>
<td>363k</td>
</tr>
<tr>
<td>Rel 9.0</td>
<td>7/08</td>
<td>603</td>
<td>636k</td>
</tr>
<tr>
<td>Rel 9.1</td>
<td>1/09</td>
<td>1372</td>
<td>1148k</td>
</tr>
<tr>
<td>Rel 10.0</td>
<td>1/10</td>
<td>1446</td>
<td>3193k</td>
</tr>
<tr>
<td>Rel 11.0</td>
<td>8/12</td>
<td>2208</td>
<td>6125k</td>
</tr>
<tr>
<td>Rel 12.0</td>
<td>9/14</td>
<td>2450</td>
<td>19623k</td>
</tr>
<tr>
<td>Rel 12.1</td>
<td>4/16</td>
<td>2474</td>
<td>9m</td>
</tr>
</tbody>
</table>

DB size:

~8GB
~160GB
~320GB
Covariance Models (CMs) represent conserved RNA sequence/structure motifs.

They allow accurate search.

But

a) search is slow

b) model construction is laborious
An Important Need: Faster Search
Homology search

“Homolog” – similar by descent from common ancestor

Sequence-based
  - Smith-Waterman
  - FASTA
  - BLAST

For RNA, sharp decline in sensitivity at ~60-70% identity

So, use structure, too
Impact of RNA homology search

(Barrick, et al., 2004)

- **B. subtilis**
- **L. innocua**
- **A. tumefaciens**
- **V. cholera**
- **M. tuberculosis**

(and 19 more species)
Impact of RNA homology search

(Barrick, et al., 2004)

B. subtilis
L. innocua
A. tumefaciens
V. cholera
M. tuberculosis
(and 19 more species)

BLAST-based

(Mandal, et al., 2004)

glycine riboswitch
operon

B. subtilis
L. innocua
A. tumefaciens
V. cholera
M. tuberculosis
(and 42 more species)

CM-based
6S mimics an open promoter

Barrick et al. RNA 2005
Trotochaud et al. NSMB 2005
Willkomm et al. NAR 2005
Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg

& W.L. Ruzzo

Recomb ‘04, ISMB ‘04, Bioinfo ‘06
CM’s are good, but slow

Rfam Reality

EMBL

BLAST

CM

junk

hits

1 month,
1000 computers

Our Work

EMBL

Ravenna

CM

hits

~2 months,
1000 computers

Rfam Goal

EMBL

CM

junk

hits

10 years,
1000 computers
CM to HMM

25 emissions per state  
5 emissions per state, 2x states
Key Issue: 25 scores $\rightarrow$ 10

Need: log Viterbi scores $\text{CM} \leq \text{HMM}$

$P_{AA} \leq L_A + R_A$
$P_{AC} \leq L_A + R_C$
$P_{AG} \leq L_A + R_G$
$P_{AU} \leq L_A + R_U$
$P_{A-} \leq L_A + R_-$

$P_{CA} \leq L_C + R_A$
$P_{CC} \leq L_C + R_C$
$P_{CG} \leq L_C + R_G$
$P_{CU} \leq L_C + R_U$
$P_{C-} \leq L_C + R_-$

NB: HMM not a prob. model
Assignment of scores/ “probabilities”

Convex optimization problem

- **Constraints**: enforce rigorous property
- **Objective function**: filter as aggressively as possible

Problem sizes:

- 1000-10000 variables
- 10000-100000 inequality constraints
“Convex” Optimization

Convex:
local max = global max;
simple “hill climbing” works
(but better ways, often)

Nonconvex:
can be many local maxima,
≪ global max;
“hill-climbing” fails
### Estimated Filtering Efficiency
(139 Rfam 4.0 families)

<table>
<thead>
<tr>
<th>Filtering fraction</th>
<th># families (compact)</th>
<th># families (expanded)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; $10^{-4}$</td>
<td>105</td>
<td>110</td>
</tr>
<tr>
<td>$10^{-4}$ - $10^{-2}$</td>
<td>8</td>
<td>17</td>
</tr>
<tr>
<td>.01 - .10</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>.10 - .25</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>.25 - .99</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>.99 - 1.0</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

Averages 283 times faster than CM
Results: new ncRNAs (?)

<table>
<thead>
<tr>
<th>Name</th>
<th># Known (BLAST + CM)</th>
<th># New (rigorous filter + CM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pyrococcus</em> snoRNA</td>
<td>57</td>
<td>123</td>
</tr>
<tr>
<td>Iron response element</td>
<td>201</td>
<td>121</td>
</tr>
<tr>
<td>Histone 3’ element</td>
<td>1004</td>
<td>102*</td>
</tr>
<tr>
<td>Retron msr</td>
<td>11</td>
<td>48</td>
</tr>
<tr>
<td>Hammerhead I</td>
<td>167</td>
<td>26</td>
</tr>
<tr>
<td>Hammerhead III</td>
<td>251</td>
<td>13</td>
</tr>
<tr>
<td>U6 snRNA</td>
<td>1462</td>
<td>2</td>
</tr>
<tr>
<td>U7 snRNA</td>
<td>312</td>
<td>1</td>
</tr>
<tr>
<td>Cobalamin riboswitch</td>
<td>170</td>
<td>7</td>
</tr>
<tr>
<td>13 other families</td>
<td>5-1107</td>
<td>0</td>
</tr>
</tbody>
</table>
CM Search Summary

Still slower than we might like, but dramatic speedup over raw CM is possible with:

No loss in sensitivity (provably), or
Even faster with modest (and estimable) loss in sensitivity