Outline

Task 1: RNA 2\textsuperscript{ary} Structure Prediction (last time)
Task 2: RNA Motif Models
  Covariance Models
  Training & “Mutual Information”
Task 3: Search
  Rigorous & heuristic filtering
Task 4: Motif discovery
Task 2: Motif Description
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

AACAAAGccggccaggcuuucAGUF
GAAUAUCUuuuggauuu......AGUF
GAAA..CA.................AGUF
GAAUAUCUuuaugauuu......AGUF

mostly G del ins all G
How to model an RNA “Motif”?

Add “column pairs” and pair emission probabilities for base-paired regions
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 sloooow
“RNA sequence analysis using covariance models”

Eddy & Durbin
Nucleic Acids Research, 1994
vol 22 #11, 2079-2088
(see also, Ch 10 of Durbin et al.)
What

A probabilistic model for RNA families

The “Covariance Model”
\[ \approx \text{A Stochastic Context-Free Grammar} \]

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 - current favorite)

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
Alignment Quality

**Trusted:**

DF6280 GCGGAUUAACUCAGUU GGG AGAGCCCAAGAGCUAGA AGUCUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DF6280G GCGGAUUAACUCAGUU GGG AGAGCCCAAGAGCUAGA AGUCUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DD6280 UCUGGAUUAAGUUUAU UUGCAGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DX1661 CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU GAGCCGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DS6280 GCGCAACUUUGCCGAGU GUUAAAGGGCGAAGAUAAGA AAGCUUUU GGCGUUUUGC CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU

**U100:**

DF6280 GCGGAUUAACUCAG UUGGAGAGCGCCAGACU GAGA AGACUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DF6280G GCGGAUUAACUCAG UUGGAGAGCGCCAGACU GAGA AGACUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DD6280 UCUGGAUUAAGUUUAU UUGCAGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DX1661 CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU GAGCCGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DS6280 GCGCAACUUUGCCGAGU GUUAAAGGGCGAAGAUAAGA AAGCUUUU GGCGUUUUGC CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU

**ClustalV:**

DF6280 GCGGAUUAACUCAG UUGGAGAGCGCCAGACU GAGA AGACUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DF6280G GCGGAUUAACUCAG UUGGAGAGCGCCAGACU GAGA AGACUGGA GUCCUGUGGUUCGAAUCACAGAAUUCGCCC
DD6280 UCUGGAUUAAGUUUAU UUGCAGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DX1661 CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU GAGCCGAAAGGGCGGCGCUUGGU GCUGCCAG AAUCGGGUUUCUUCCGUGCAGCAGCCC
DS6280 GCGCAACUUUGCCGAGU GUUAAAGGGCGAAGAUAAGA AAGCUUUU GGCGUUUUGC CGCAGGUGGAGGAGAGGCUAGAAGCUAGGCUAGGGAU
Comparison to TRNASCAN

Fichant & Burks - best heuristic then
- 97.5% true positive
- 0.37 false positives per MB

CM A1415 (trained on trusted alignment)
- > 99.98% true positives
- <0.2 false positives per MB

Current method-of-choice is “tRNAscanSE”, a CM-based scan with heuristic pre-filtering (including TRNASCAN?) for performance reasons.
**Profile Hmm Structure**

**Figure 5.2** *The transition structure of a profile HMM.*

- **M_j**: Match states (20 emission probabilities)
- **I_j**: Insert states *(Background emission probabilities)*
- **D_j**: 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)
Overall CM Architecture

One box (“node”) per node of guide tree

BEG/MATL/INS/DEL just like an HMM

MATP & BIF are the key additions: MATP emits pairs of symbols, modeling base-pairs; BIF allows multiple helices
CM Viterbi Alignment
(the “inside” algorithm)

\[ x_i = \text{ith letter of input} \]
\[ x_{ij} = \text{substring } i, \ldots, j \text{ of input} \]
\[ T_{yz} = P(\text{transition } y \rightarrow z) \]
\[ E_{x_i,x_j}^y = P(\text{emission of } x_i, x_j \text{ from state } y) \]
\[ S_{ij}^y = \max_\pi \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi) \]
CM Viterbi Alignment
(the “inside” algorithm)

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

\[ S_{ij}^y = \max_z \left[ S_{i+1, j-1}^z + \log T_{yz} + \log E_{x_i, x_j}^y \right] \quad \text{match pair} \]

\[ \max_z \left[ S_{i+1, j}^z + \log T_{yz} + \log E_{x_i}^y \right] \quad \text{match/insert left} \]

\[ \max_z \left[ S_{i, j-1}^z + \log T_{yz} + \log E_{x_j}^y \right] \quad \text{match/insert right} \]

\[ \max_z \left[ S_{i, j}^z + \log T_{yz} \right] \quad \text{delete} \]

\[ \max_{i < k \leq j} \left[ S_{i, k}^{y_{\text{left}}} + S_{k+1, j}^{y_{\text{right}}} \right] \quad \text{bifurcation} \]

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

compare: \(O(qn)\) for profile HMM
Nussinov: Max Pairing

\[ B(i,j) = \# \text{ pairs in optimal pairing of } r_i \ldots r_j \]

\[ B(i,j) = 0 \text{ for all } i, j \text{ with } i \geq j-4; \text{ otherwise} \]

\[ B(i,j) = \max \text{ of:} \]

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

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

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

Max when no seq conservation but perfect pairing

MI = expected score gain from using a pair state

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)

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.
Pseudoknots
disallowed  allowed  \( \left( \sum_{i=1}^{n} \max_j M_{i,j} \right)/2 \)

<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>

Table 1: Statistics of the training and test sets of 100 tRNA sequences each. The average identity in an alignment is the average pairwise identity of all aligned symbol pairs, with gap/symbol alignments counted as mismatches. Primary sequence information content is calculated according to [48]. Calculating pairwise mutual information content is an NP-complete problem of finding an optimum partition of columns into pairs. A lower bound is calculated by using the model construction procedure to find an optimal partition subject to a non-pseudoknotting restriction. An upper bound is calculated as sum of the single best pairwise covariation for each position, divided by two; this includes all pairwise tertiary interactions but overcounts because it does not guarantee a disjoint set of pairs. For the meaning of multiple alignment accuracy of ClustalV, see the text.
<table>
<thead>
<tr>
<th>Model</th>
<th>training set</th>
<th>iterations</th>
<th>score (bits)</th>
<th>alignment accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1415</td>
<td>all sequences (aligned)</td>
<td>3</td>
<td>58.7</td>
<td>95%</td>
</tr>
<tr>
<td>A100</td>
<td>SIM100 (aligned)</td>
<td>3</td>
<td>57.3</td>
<td>94%</td>
</tr>
<tr>
<td>A65</td>
<td>SIM65 (aligned)</td>
<td>3</td>
<td>46.7</td>
<td>93%</td>
</tr>
<tr>
<td>U100</td>
<td>SIM100 (degapped)</td>
<td>23</td>
<td>56.7</td>
<td>90%</td>
</tr>
<tr>
<td>U65</td>
<td>SIM65 (degapped)</td>
<td>29</td>
<td>47.2</td>
<td>91%</td>
</tr>
</tbody>
</table>

Table 2: Training and multiple alignment results from models trained from the trusted alignments (A models) and models trained from no prior knowledge of tRNA (U models).
Rfam – an RNA family DB
Griffiths-Jones, et al., NAR ‘03, ’05

Biggest scientific computing user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:
Rel 1.0, 1/03: 25 families, 55k instances
Rel 7.0, 3/05: 503 families, >300k instances
Input (hand-curated):  
MSA “seed alignment”  
SS_cons  
Score Thresh T  
Window Len W  

Output:  
CM  
scan results & “full alignment”

IRE (partial seed alignment):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>IRE (partial seed alignment):</td>
<td>GUUCCUGCUUCAACAGUGUUUGGAUGGAAC</td>
<td>UAUAUCGGGAACAGUGUUUCCC.AUAAU</td>
<td>UUUCUUC.UUCAACAGUGUUUGGAUGGAAC</td>
<td>UUUAUC.AGUGACAGAGUUCACU.AUAAA</td>
</tr>
<tr>
<td></td>
<td>UUUCUUC.UUCAACAGUGUUUGGAUGGAAC</td>
<td>UUUCUGUUUCAACAGUGCUUGGA.GGAAC</td>
<td>UUUAUC.AGUGACAGAGUUCACU.AUAAA</td>
<td>UCUUCUGCUUCAACAGUGUUUGGAUGGAAC</td>
</tr>
<tr>
<td></td>
<td>UUCCUGUUUCAACAGUGCUUGGA.GGAAC</td>
<td>UUUAUC.AGUGACAGAGUUCACU.AUAAA</td>
<td>UCUUCUGCUUCAACAGUGUUUGGAUGGAAC</td>
<td>UCUUCUGCUUCAACAGUGCUUGGA.GGAAC</td>
</tr>
<tr>
<td></td>
<td>UUUAUC.AGUGACAGAGUUCACU.AUAAA</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>UCUCUUGCUUCAACAGUGUUUGGAUGGAAC</td>
<td>UCUUGC.UUCAACAGUGUUUGGACGGAAG</td>
<td>UCUUCUGCUUCAACAGUGUUUCCC.AUAAU</td>
<td>UCUUGC.UUCAACAGUGUUUGGACGGAAG</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
</tr>
<tr>
<td></td>
<td>UCUUGC.UUCAACAGUGUUUGGACGGAAG</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUCUUGCUUCAACAGUGUUUGGACGGAAC</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUCUUGCUUCAACAGUGUUUGGACGGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
<tr>
<td></td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
<td>UGUAUC.GGAGACAGUGAUCUCC.AUAUG</td>
<td>UCUUCUGCUUCAACAGUGUUUGGA.GGAAC</td>
<td>AUUAUC.GGGAACAGUGUUUCCC.AUAAU</td>
</tr>
</tbody>
</table>

SS_cons <<<<<<...<<<<<<...>>>>>.>>>>>
Figure 2. Taxonomic distribution of Rfam family members in the three kingdoms of life.
Task 3: Faster Search
Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg & W.L. Ruzzo

Recomb ‘04, ISMB ‘04, Bioinfo ‘06
RaveNnA: Genome Scale RNA Search

Typically 100x speedup over raw CM, w/ no loss in accuracy:
- drop structure from CM to create a (faster) HMM
- use that to pre-filter sequence;
- discard parts where, provably, CM will score < threshold;
- actually run CM on the rest (the promising parts)
- assignment of HMM transition/emission scores is key
  (large convex optimization problem)

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
Covariance Model

Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.
Simplified CM
(for pedagogical purposes only)
CM to HMM

CM

HMM

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

Need: log Viterbi scores $CM \leq HMM$
Viterbi/Forward Scoring

Path \( \pi \) defines transitions/emissions

\[
\text{Score}(\pi) = \text{product of “probabilities” on } \pi
\]

NB: ok if “probs” aren’t, e.g. \( \Sigma \neq 1 \)
(e.g. in CM, emissions are odds ratios vs 0th-order background)

For any nucleotide sequence \( x \):

\[
\text{Viterbi-score}(x) = \max \{ \text{score}(\pi) \mid \pi \text{ emits } x \}
\]

\[
\text{Forward-score}(x) = \sum \{ \text{score}(\pi) \mid \pi \text{ emits } x \}
\]
Key Issue: 25 scores → 10

Need: log Viterbi scores CM ≤ HMM

\[
\begin{align*}
P_{AA} & \leq L_A + R_A \\
P_{AC} & \leq L_A + R_C \\
P_{AG} & \leq L_A + R_G \\
P_{AU} & \leq L_A + R_U \\
P_{A-} & \leq L_A + R_-
\end{align*}
\]

\[
\begin{align*}
P_{CA} & \leq L_C + R_A \\
P_{CC} & \leq L_C + R_C \\
P_{CG} & \leq L_C + R_G \\
P_{CU} & \leq L_C + R_U \\
P_{C-} & \leq L_C + R_-
\end{align*}
\]

NB: HMM not a prob. model
Rigorous Filtering

Any scores satisfying the linear inequalities give rigorous filtering

Proof:
CM Viterbi path score
≤ “corresponding” HMM path score
≤ Viterbi HMM path score
(even if it does not correspond to any CM path)
Some scores filter better

\[ P_{UA} = 1 \leq L_U + R_A \]
\[ P_{UG} = 4 \leq L_U + R_G \]

Option 1:
\[ L_U = R_A = R_G = 2 \]

Option 2:
\[ L_U = 0, R_A = 1, R_G = 4 \]

Assuming ACGU ≈ 25%

Opt 1:
\[ L_U + (R_A + R_G)/2 = 4 \]

Opt 2:
\[ L_U + (R_A + R_G)/2 = 2.5 \]
Optimizing filtering

For any nucleotide sequence $x$:

\[
\text{Viterbi-score}(x) = \max \{ \text{score}(\pi) \mid \pi \text{ emits } x \} \\
\text{Forward-score}(x) = \sum \{ \text{score}(\pi) \mid \pi \text{ emits } x \}
\]

Expected Forward Score

\[
E(L_i, R_i) = \sum_{\text{all sequences } x} \text{Forward-score}(x) \cdot \text{Pr}(x)
\]

NB: $E$ is a function of $L_i, R_i$ only

Optimization:

Minimize $E(L_i, R_i)$ subject to score Lin.Ineq.s

This is heuristic (“forward↓ ⇒ Viterbi↓ ⇒ filter↓”)

But still rigorous because “subject to score Lin.Ineq.s”
Calculating $E(L_i, R_i)$

$$E(L_i, R_i) = \sum_x \text{Forward-score}(x) \times \text{Pr}(x)$$

Forward-like: for every state, calculate expected score for all paths ending there; easily calculated from expected scores of predecessors & transition/emission probabilities/scores
Minimizing $E(L_i, R_i)$

Calculate $E(L_i, R_i)$ symbolically, in terms of emission scores, so we can do partial derivatives for numerical convex optimization algorithm.

Forward:
\[
\begin{align*}
    f_k(i) &= P(x_1 \ldots x_i, \pi_i = k) \\
    f_i(i + 1) &= e_i(x_{i+1}) \sum_k f_k(i) a_{k,l}
\end{align*}
\]

Viterbi:
\[
    v_i(i + 1) = e_i(x_{i+1}) \cdot \max_k (v_k(i) a_{k,l})
\]
“Convex” Optimization

Convex:
local max = global max;
simple “hill climbing” works

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>

~100x speedup
## Results: New ncRNA’s?

<table>
<thead>
<tr>
<th>Name</th>
<th># found BLAST + CM</th>
<th># found rigorous filter + CM</th>
<th># new</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrococcus snoRNA</td>
<td>57</td>
<td>180</td>
<td>123</td>
</tr>
<tr>
<td>Iron response element</td>
<td>201</td>
<td>322</td>
<td>121</td>
</tr>
<tr>
<td>Histone 3’ element</td>
<td>1004</td>
<td>1106</td>
<td>102</td>
</tr>
<tr>
<td>Purine riboswitch</td>
<td>69</td>
<td>123</td>
<td>54</td>
</tr>
<tr>
<td>Retron msr</td>
<td>11</td>
<td>59</td>
<td>48</td>
</tr>
<tr>
<td>Hammerhead I</td>
<td>167</td>
<td>193</td>
<td>26</td>
</tr>
<tr>
<td>Hammerhead III</td>
<td>251</td>
<td>264</td>
<td>13</td>
</tr>
<tr>
<td>U4 snRNA</td>
<td>283</td>
<td>290</td>
<td>7</td>
</tr>
<tr>
<td>S-box</td>
<td>128</td>
<td>131</td>
<td>3</td>
</tr>
<tr>
<td>U6 snRNA</td>
<td>1462</td>
<td>1464</td>
<td>2</td>
</tr>
<tr>
<td>U5 snRNA</td>
<td>199</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>U7 snRNA</td>
<td>312</td>
<td>313</td>
<td>1</td>
</tr>
</tbody>
</table>
Task 4: Motif Discovery
RNA Motif Discovery

Typical problem: given a ~10-20 unaligned sequences of ~1kb, most of which contain instances of one RNA motif of, say, 150bp -- find it.

Example: 5’ UTRs of orthologous glycine cleavage genes from γ-proteobacteria
Approaches

Align sequences, then look for common structure

Predict structures, then try to align them

Do both together
“Obvious” Approach I: Align First, Predict from Multiple Sequence Alignment

Compensatory mutations reveal structure, (core of “comparative sequence analysis”)

but usual alignment algorithms penalize them (twice)
Pitfall for sequence-alignment-first approach

Structural conservation ≠ Sequence conservation
Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions

same-colored boxes *should* be aligned
Approaches

Align sequences, then look for common structure

Predict structures, then try to align them
  single-seq struct prediction only ~ 60% accurate;
  exacerbated by flanking seq; no biologically-validated
  model for structural alignment

Do both together
  Sankoff – good but slow
  Various heuristics – still tend to be slow
Our Approach: CMfinder

Simultaneous alignment, folding and CM-based motif description using an EM-style learning procedure

Yao, Weinberg & Ruzzo, Bioinformatics, 2006
Alignment $\rightarrow$ CM $\rightarrow$ Alignment

Similar to HMM, but slower
Builds on Eddy & Durbin, ‘94
But new way to infer which columns to pair, via a principled combination of mutual information and predicted folding energy
And, it’s local, not global, alignment (harder)
Model Training (Eddy-Durbin)

unaligned sequences

random alignment

multiple alignment

alignment

(EM)

parameter reestimation

covariance model

model construction (structure prediction)
CMfinder Outline

- Folding predictions
- Heuristics
- Candidate alignment
- Realign
- CM
- Search

M-step uses M.I. + folding energy for structure prediction
Structure Inference

Part of M-step is to pick a structure that maximizes data likelihood

We combine:

- mutual information
- position-specific priors for paired/unpaired

intuition: for similar seqs, little MI; fall back on single-sequence folding predictions

data-dependent, so not strictly Bayesian
CMfinder Accuracy
(on Rfam families with flanking sequence)
# Summary of Rfam test families and results

<table>
<thead>
<tr>
<th>ID</th>
<th>Family</th>
<th>Rfam ID</th>
<th>#seqs</th>
<th>%id</th>
<th>length</th>
<th>#hp</th>
<th>CMfinder</th>
<th>CW/Pfold</th>
<th>CW/RNAalifold</th>
<th>Carnac</th>
<th>Foldalign</th>
<th>ComRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cobalamin</td>
<td>RF00174</td>
<td>71</td>
<td>49</td>
<td>216</td>
<td>4</td>
<td><strong>0.59</strong></td>
<td>0.05</td>
<td>0</td>
<td>X</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ctRNA_pGAl</td>
<td>RF00236</td>
<td>17</td>
<td>74</td>
<td>83</td>
<td>2</td>
<td><strong>0.91</strong></td>
<td>0.70</td>
<td>0.72</td>
<td>0</td>
<td>0.86</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Entero_CRE</td>
<td>RF00048</td>
<td>56</td>
<td>81</td>
<td>61</td>
<td>1</td>
<td><strong>0.89</strong></td>
<td>0.74</td>
<td>0.22</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Entero_OriR</td>
<td>RF00041</td>
<td>35</td>
<td>77</td>
<td>73</td>
<td>2</td>
<td><strong>0.94</strong></td>
<td>0.75</td>
<td>0.76</td>
<td>0.80</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>5</td>
<td>glmS</td>
<td>RF00234</td>
<td>14</td>
<td>58</td>
<td>188</td>
<td>4</td>
<td><strong>0.83</strong></td>
<td>0.12</td>
<td>0.18</td>
<td>0</td>
<td>-</td>
<td>0.13</td>
</tr>
<tr>
<td>6</td>
<td>Histone3</td>
<td>RF00032</td>
<td>63</td>
<td>77</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>Intron_gpII</td>
<td>RF00029</td>
<td>75</td>
<td>55</td>
<td>92</td>
<td>2</td>
<td><strong>0.80</strong></td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>IRE</td>
<td>RF00037</td>
<td>30</td>
<td>68</td>
<td>30</td>
<td>1</td>
<td><strong>0.77</strong></td>
<td>0.22</td>
<td>0</td>
<td>0</td>
<td>0.38</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>let-7</td>
<td>RF00027</td>
<td>9</td>
<td>69</td>
<td>84</td>
<td>1</td>
<td><strong>0.87</strong></td>
<td>0.08</td>
<td>0.42</td>
<td>0</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>10</td>
<td>lin-4</td>
<td>RF00052</td>
<td>9</td>
<td>69</td>
<td>72</td>
<td>1</td>
<td><strong>0.78</strong></td>
<td>0.51</td>
<td>0.75</td>
<td>0.41</td>
<td>0.65</td>
<td>0.24</td>
</tr>
<tr>
<td>11</td>
<td>Lysine</td>
<td>RF00168</td>
<td>48</td>
<td>48</td>
<td>183</td>
<td>4</td>
<td><strong>0.77</strong></td>
<td>0.24</td>
<td>0</td>
<td>X</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>mir-10</td>
<td>RF00104</td>
<td>11</td>
<td>66</td>
<td>75</td>
<td>1</td>
<td><strong>0.66</strong></td>
<td>0.59</td>
<td>0.60</td>
<td>0</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>13</td>
<td>Purine</td>
<td>RF00167</td>
<td>29</td>
<td>55</td>
<td>103</td>
<td>2</td>
<td><strong>0.91</strong></td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0.27</td>
</tr>
<tr>
<td>14</td>
<td>RFN</td>
<td>RF00050</td>
<td>47</td>
<td>66</td>
<td>139</td>
<td>4</td>
<td>0.39</td>
<td><strong>0.68</strong></td>
<td>0.26</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>Rhino_CRE</td>
<td>RF00220</td>
<td>12</td>
<td>71</td>
<td>86</td>
<td>1</td>
<td><strong>0.88</strong></td>
<td>0.52</td>
<td>0.52</td>
<td>0.69</td>
<td>0.41</td>
<td>0.61</td>
</tr>
<tr>
<td>16</td>
<td>s2m</td>
<td>RF00164</td>
<td>23</td>
<td>80</td>
<td>43</td>
<td>1</td>
<td>0.67</td>
<td><strong>0.80</strong></td>
<td>0.45</td>
<td>0.64</td>
<td>0.63</td>
<td>0.29</td>
</tr>
<tr>
<td>17</td>
<td>S_box</td>
<td>RF00162</td>
<td>64</td>
<td>66</td>
<td>112</td>
<td>3</td>
<td><strong>0.72</strong></td>
<td>0.11</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>SECIS</td>
<td>RF00031</td>
<td>43</td>
<td>43</td>
<td>68</td>
<td>1</td>
<td><strong>0.73</strong></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>Tymo_tRNA-like</td>
<td>RF00233</td>
<td>22</td>
<td>72</td>
<td>86</td>
<td>4</td>
<td><strong>0.81</strong></td>
<td>0.33</td>
<td>0.36</td>
<td>0.30</td>
<td>0.80</td>
<td>0.48</td>
</tr>
</tbody>
</table>

**Average Accuracy:** 0.79 0.36 0.28 0.17 0.60 0.19
**Average Specificity:** 0.81 0.42 0.57 **0.83** 0.60 0.65
**Average Sensitivity:** 0.77 0.36 0.23 0.13 0.61 0.17
Task 5: Application

A Computational Pipeline for High Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes.

Searching for noncoding RNAs

CM’s are great, but where do they come from?

An approach: comparative genomics

Search for motifs with common secondary structure in a set of functionally related sequences.

Challenges

Three related tasks

- Locate the motif regions.
- Align the motif instances.
- Predict the consensus secondary structure.

Motif search space is huge!

- Motif location space, alignment space, structure space.
Predicting New cis-Regulatory RNA Elements

Goal:
- Given unaligned UTRs of coexpressed or orthologous genes, find common structural motifs

Difficulties:
- Low sequence similarity: alignment difficult
- Varying flanking sequence
- Motif missing from some input genes
A pipeline for RNA motif genome scans

Genome Scale Search: Why

Most riboswitches, e.g., are present in ~5 copies per genome
Throughout (most of) clade
More examples give better model, hence even more examples, fewer errors
More examples give more clues to function - critical for wet lab verification
Genome Scale Search

CMfinder is directly usable for/with search
Results

Analyzed most sequenced bacteria (~2005)
- bacillus/clostridia
- gamma proteobacteria
- cyanobacteria
- actinobacteria
- firmicutes
Identify CDD group members

Retrieve upstream sequences

Footprinter ranking

CMfinder

Motif postprocessing

RaveNnA

CMfinder refinement

Motif postprocessing

Identify CDD group members

< 10 CPU days

Retrieve upstream sequences

< 10 CPU days

Footprinter ranking

< 10 CPU days

CMfinder

35975 motifs

Motif postprocessing

1740 motifs

RaveNnA

10 CPU months

CMfinder refinement

< 1 CPU month

Motif postprocessing

1466 motifs
<table>
<thead>
<tr>
<th>Rank RAV</th>
<th>Score RAV</th>
<th># ID Gene</th>
<th>Description</th>
<th>CDD</th>
<th>Rfam</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 43 107</td>
<td>3400 367 11</td>
<td>9904 IlvB</td>
<td>Thiamine pyrophosphate-requiring enzymes</td>
<td>RF00230 T-box</td>
<td></td>
</tr>
<tr>
<td>1 10 344</td>
<td>3115 96 22</td>
<td>13174 COG3859</td>
<td>Predicted membrane protein</td>
<td>RF00059 THI</td>
<td></td>
</tr>
<tr>
<td>2 77 1284</td>
<td>2376 112 6</td>
<td>11125 MetH</td>
<td>Methionine synthase I specific DNA methylase</td>
<td>RF00162 S_box</td>
<td></td>
</tr>
<tr>
<td>3 0 5</td>
<td>2327 30 26</td>
<td>9991 COG0116</td>
<td>Predicted N6-adenine-specific DNA methylase</td>
<td>RF00011 RNaseP_bact_b</td>
<td></td>
</tr>
<tr>
<td>4 6 66</td>
<td>2228 49 18</td>
<td>4383 DHBP</td>
<td>3,4-dihydroxy-2-butanoic 4-phosphate synthase</td>
<td>RF00050 RFN</td>
<td></td>
</tr>
<tr>
<td>7 145 952</td>
<td>1429 51 7</td>
<td>10390 GuaA</td>
<td>GMP synthase</td>
<td>RF00167 Purine</td>
<td></td>
</tr>
<tr>
<td>8 17 108</td>
<td>1322 29 13</td>
<td>10732 GcvP</td>
<td>Glycine cleavage system protein P</td>
<td>RF00504 Glycine</td>
<td></td>
</tr>
<tr>
<td>9 37 749</td>
<td>1235 28 7</td>
<td>24631 DUF149</td>
<td>Uncharacterised BCR, YbaB family COG0718</td>
<td>RF00169 SRP_bact</td>
<td></td>
</tr>
<tr>
<td>10 123 1358</td>
<td>1222 36 6</td>
<td>10986 CbiB</td>
<td>Cobalamin biosynthesis protein CobD/CbiB</td>
<td>RF00174 Cobalamin</td>
<td></td>
</tr>
<tr>
<td>20 137 1133</td>
<td>899 32 7</td>
<td>9895 LysA</td>
<td>Diaminopimelate decarboxylase</td>
<td>RF00168 Lysine</td>
<td></td>
</tr>
<tr>
<td>21 36 141</td>
<td>896 22 10</td>
<td>10727 TerC</td>
<td>Membrane protein TerC</td>
<td>RF00080 yybP-ykoY</td>
<td></td>
</tr>
<tr>
<td>39 202 684</td>
<td>664 25 5</td>
<td>11945 MgtE</td>
<td>Mg/Co/Ni transporter MgtE</td>
<td>RF00380 ykoK</td>
<td></td>
</tr>
<tr>
<td>40 26 74</td>
<td>645 19 18</td>
<td>10323 GlmS</td>
<td>Glucosamine 6-phosphate synthetase</td>
<td>RF00234 glmS</td>
<td></td>
</tr>
<tr>
<td>53 208 192</td>
<td>561 21 5</td>
<td>10892 OpuBB</td>
<td>ABC-type proline/glycine betaine transport systems</td>
<td>RF00005 tRNA^1</td>
<td></td>
</tr>
<tr>
<td>122 99 239</td>
<td>413 10 7</td>
<td>11784 EmrE</td>
<td>Membrane transporters of cations and cationic drug</td>
<td>RF00442 ykkC-ykkD</td>
<td></td>
</tr>
<tr>
<td>255 392 281</td>
<td>268 8 6</td>
<td>10272 COG0398</td>
<td>Uncharacterized conserved protein</td>
<td>RF00023 tmRNA</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Motifs that correspond to Rfam families. “Rank”: the three columns show ranks for refined motif clusters after genome scans (“RAV”), CMfinder motifs before genome scans (“CMF”), and FootPrint results (“FP”). We used the same ranking scheme for RAV and CMF. “Score”
<table>
<thead>
<tr>
<th>Rfam</th>
<th>Membership</th>
<th>Overlap</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#   Sn</td>
<td>nt</td>
<td>bp   Sn</td>
</tr>
<tr>
<td>RF00174 Cobalamin</td>
<td>183 0.74(^1)</td>
<td>152 0.75</td>
<td>20 0.60</td>
</tr>
<tr>
<td>RF00504 Glycine</td>
<td>91 0.56(^1)</td>
<td>94 0.94</td>
<td>17 0.84</td>
</tr>
<tr>
<td>RF00234 glmS</td>
<td>34 0.92</td>
<td>100 0.54</td>
<td>27 0.96</td>
</tr>
<tr>
<td>RF00168 Lysine</td>
<td>80 0.82</td>
<td>111 0.61</td>
<td>26 0.76</td>
</tr>
<tr>
<td>RF00167 Purine</td>
<td>86 0.86</td>
<td>83 0.83</td>
<td>17 0.90</td>
</tr>
<tr>
<td>RF00050 RFN</td>
<td>133 0.98</td>
<td>139 0.96</td>
<td>12 0.66</td>
</tr>
<tr>
<td>RF00011 RNaseP_bact_b</td>
<td>144 0.99</td>
<td>194 0.53</td>
<td>38 0.72</td>
</tr>
<tr>
<td>RF00162 S_box</td>
<td>208 0.95</td>
<td>110 1.00</td>
<td>23 0.91</td>
</tr>
<tr>
<td>RF00169 SRP_bact</td>
<td>177 0.92</td>
<td>99 1.00</td>
<td>25 0.89</td>
</tr>
<tr>
<td>RF00230 T-box</td>
<td>453 0.96</td>
<td>187 0.77</td>
<td>5 0.32</td>
</tr>
<tr>
<td>RF00059 THI</td>
<td>326 0.89</td>
<td>99 0.91</td>
<td>13 0.56</td>
</tr>
<tr>
<td>RF00442 ykkC-yxkD</td>
<td>19 0.90</td>
<td>99 0.94</td>
<td>18 0.94</td>
</tr>
<tr>
<td>RF00380 ykoK</td>
<td>49 0.92</td>
<td>125 0.75</td>
<td>27 0.80</td>
</tr>
<tr>
<td>RF00080 yybP-ykoY</td>
<td>41 0.32</td>
<td>100 0.78</td>
<td>18 0.63</td>
</tr>
</tbody>
</table>

| mean | 145 0.84 | 121 0.81 | 21 0.75 | 0.77 |
| median | 113 0.91 | 105 0.81 | 19 0.78 | 0.78 |

Table 2: Motif prediction accuracy vs prokaryotic subset of Rfam full alignments. “Membership”: the number of sequences in the overlap between our predictions and Rfam’s (“#”), the sensitivity (“Sn”) and specificity (“Sp”) of our membership predictions. “Overlap”: avg length of overlap between our predictions and Rfam’s (“nt”), the fractional lengths of the overlapped region in Rfam’s predictions (“Sn”) and in ours (“Sp”). “Structure”: avg number of correctly predicted canonical base pairs (in overlapped regions) and the sensitivity (“Sn”) and specificity (“Sp”) of our predictions. \(^1\)After another iteration of RaveNnA scan and refinement, the membership sensitivities of Glycine and Cobalamin increased to 76% and 98% respectively, while the specificity of Glycine remained the same, and specificity of Cobalamin dropped to 84%.
<table>
<thead>
<tr>
<th>Rank</th>
<th>#</th>
<th>CDD</th>
<th>Gene: Description</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>69</td>
<td>28178</td>
<td>DHOase IIa: Dihydroorotase</td>
<td>PyrR attenuator [22]</td>
</tr>
<tr>
<td>15</td>
<td>33</td>
<td>10097</td>
<td>RplL: Ribosomal protein L7/L1</td>
<td>L10 r-protein leader; see Supp</td>
</tr>
<tr>
<td>19</td>
<td>36</td>
<td>10234</td>
<td>RpsF: Ribosomal protein S6</td>
<td>S6 r-protein leader</td>
</tr>
<tr>
<td>22</td>
<td>32</td>
<td>10897</td>
<td>COG1179: Dinucleotide-utilizing enzymes</td>
<td>6S RNA [25]</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>9926</td>
<td>RpsJ: Ribosomal protein S10</td>
<td>S10 r-protein leader; see Supp</td>
</tr>
<tr>
<td>29</td>
<td>11</td>
<td>15150</td>
<td>Resolvase: N terminal domain</td>
<td>IF-3 r-protein leader; see Supp</td>
</tr>
<tr>
<td>31</td>
<td>31</td>
<td>10164</td>
<td>InfC: Translation initiation factor 3</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>26</td>
<td>10393</td>
<td>RpsD: Ribosomal protein S4 and related proteins</td>
<td>S4 r-protein leader; see Supp [30]</td>
</tr>
<tr>
<td>44</td>
<td>30</td>
<td>10332</td>
<td>GroL: Chaperonin GroEL</td>
<td>HrcA DNA binding site [46]</td>
</tr>
<tr>
<td>46</td>
<td>33</td>
<td>25629</td>
<td>Ribosomal L21p: Ribosomal prokaryotic L21 protein</td>
<td>L21 r-protein leader; see Supp [47]</td>
</tr>
<tr>
<td>50</td>
<td>11</td>
<td>5638</td>
<td>Cad: Cadmium resistance transporter</td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>19</td>
<td>9965</td>
<td>RplB: Ribosomal protein L2</td>
<td>S10 r-protein leader</td>
</tr>
<tr>
<td>55</td>
<td>7</td>
<td>26270</td>
<td>RNA pol Rpb2 1: RNA polymerase beta subunit</td>
<td></td>
</tr>
<tr>
<td>69</td>
<td>9</td>
<td>13148</td>
<td>COG3830: ACT domain-containing protein</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>28</td>
<td>4174</td>
<td>Ribosomal S2: Ribosomal protein S2</td>
<td>S2 r-protein leader</td>
</tr>
<tr>
<td>74</td>
<td>9</td>
<td>9924</td>
<td>RpsG: Ribosomal protein S7</td>
<td>S12 r-protein leader</td>
</tr>
<tr>
<td>86</td>
<td>6</td>
<td>12328</td>
<td>COG2984: ABC-type uncharacterized transport system</td>
<td></td>
</tr>
<tr>
<td>88</td>
<td>19</td>
<td>24072</td>
<td>CtsR: Firmicutes transcriptional repressor of class III</td>
<td>CtsR DNA binding site [48]</td>
</tr>
<tr>
<td>100</td>
<td>21</td>
<td>23019</td>
<td>Formyl trans N: Formyl transferase</td>
<td></td>
</tr>
<tr>
<td>103</td>
<td>8</td>
<td>9916</td>
<td>PurE: Phosphoribosylcarboxyaminomimidazole</td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>5</td>
<td>13411</td>
<td>COG4129: Predicted membrane protein</td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>10</td>
<td>10075</td>
<td>RplO: Ribosomal protein L15</td>
<td>L15 r-protein leader</td>
</tr>
<tr>
<td>121</td>
<td>9</td>
<td>10132</td>
<td>RpmJ: Ribosomal protein L36</td>
<td>IF-1 r-protein leader</td>
</tr>
<tr>
<td>129</td>
<td>4</td>
<td>23962</td>
<td>Cna B: Cna protein B-type domain</td>
<td></td>
</tr>
<tr>
<td>130</td>
<td>9</td>
<td>25424</td>
<td>Ribosomal S12: Ribosomal protein S12</td>
<td>S12 r-protein leader</td>
</tr>
<tr>
<td>131</td>
<td>9</td>
<td>16769</td>
<td>Ribosomal L4: Ribosomal protein L4/L1 family</td>
<td>L3 r-protein leader</td>
</tr>
<tr>
<td>140</td>
<td>12</td>
<td>8892</td>
<td>Pencillinase R: Pencillinase repressor</td>
<td>Blal, Mecl DNA binding site [49]</td>
</tr>
<tr>
<td>157</td>
<td>25</td>
<td>24415</td>
<td>Ribosomal S9: Ribosomal protein S9/S16</td>
<td>L13 r-protein leader; Fig 3</td>
</tr>
<tr>
<td>160</td>
<td>27</td>
<td>1790</td>
<td>Ribosomal L19: Ribosomal protein L19</td>
<td>L19 r-protein leader; Fig 2</td>
</tr>
<tr>
<td>164</td>
<td>6</td>
<td>9932</td>
<td>GapA: Glyceraldehyde-3-phosphate dehydrogenase/erythrose</td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>8</td>
<td>13849</td>
<td>COG4708: Predicted membrane protein</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>7</td>
<td>10199</td>
<td>COG0325: Predicted enzyme with a TIM-barrel fold</td>
<td></td>
</tr>
<tr>
<td>182</td>
<td>9</td>
<td>10207</td>
<td>RpmF: Ribosomal protein L32</td>
<td>L32 r-protein leader</td>
</tr>
<tr>
<td>187</td>
<td>11</td>
<td>27850</td>
<td>LDH: L-lactate dehydrogenases</td>
<td></td>
</tr>
<tr>
<td>190</td>
<td>11</td>
<td>10094</td>
<td>CspR: Predicted rRNA methylase</td>
<td>EF-G r-protein leader</td>
</tr>
<tr>
<td>194</td>
<td>9</td>
<td>10353</td>
<td>FusA: Translation elongation factors</td>
<td></td>
</tr>
</tbody>
</table>
mRNA leader

mRNA leader switch?
Task 5: Application

Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline.


boxed = confirmed riboswitch (+2 more)
ncRNA discovery in Vertebrates

Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions

E. Torarinsson, Z. Yao, E. D. Wiklund, J. B. Bramsen, C. Hansen, J. Kjems, N. Tommerup, W. L. Ruzzo and J. Gorodkin

Genome Research, Jan 2008
ncRNA discovery in Vertebrates

Previous studies focus on highly conserved regions (Washietl, Pedersen et al. 2007)

Evofold (Pedersen et al. 2006)

RNAz (Washietl et al. 2005)

We explore regions with weak sequence conservation
Approach

Extract ENCODE Multiz alignments

  Remove exons, most conserved elements.
  56017 blocks, 8.7M bps.

Apply CMfinder to both strands.

10,106 predictions, 6,587 clusters.

  False positive rate: 50% based on a heuristic ranking function.
Overlap w/ Indel Purified Segments

IPS presumed to signal purifying selection

Majority (64%) of candidates have >45% G+C

Strong P-value for their overlap w/ IPS

<table>
<thead>
<tr>
<th></th>
<th>G+C data</th>
<th>P</th>
<th>N</th>
<th>Expected</th>
<th>Observed</th>
<th>P-value</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-35 igs</td>
<td>0.062</td>
<td>380</td>
<td>23</td>
<td>24.5</td>
<td>0.430</td>
<td>5.8%</td>
<td></td>
</tr>
<tr>
<td>35-40 igs</td>
<td>0.082</td>
<td>742</td>
<td>61</td>
<td>70.5</td>
<td>0.103</td>
<td>11.3%</td>
<td></td>
</tr>
<tr>
<td>40-45 igs</td>
<td>0.082</td>
<td>1216</td>
<td>99</td>
<td>129.5</td>
<td>0.00079</td>
<td>18.5%</td>
<td></td>
</tr>
<tr>
<td>45-50 igs</td>
<td>0.079</td>
<td>1377</td>
<td>109</td>
<td>162.5</td>
<td>5.16E-08</td>
<td>20.9%</td>
<td></td>
</tr>
<tr>
<td>50-100 igs</td>
<td>0.070</td>
<td>2866</td>
<td>200</td>
<td>358.5</td>
<td>2.70E-31</td>
<td>43.5%</td>
<td></td>
</tr>
<tr>
<td>all igs</td>
<td>0.075</td>
<td>6581</td>
<td>491</td>
<td>747.5</td>
<td>1.54E-33</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>
Comparison with Evofold, RNAz

Small overlap (w/ highly significant p-values) emphasizes complementarity
Realignment

% realigned

Average pairwise sequence similarity
Experimental Validation
New scoring scheme

Goal: improve false discovery rate for top ranking motifs

Current methods can not improve beyond 50% FDR by using higher score threshold.
Neither RNAz nor Evofold are robust on poorly conserved and gappy regions.
Method

Goal: given a structural alignments, determine its significance.

Phylo-SCFG as in EvoFold

- SCFG to capture consensus secondary structure
- Evolution models to capture mutations among species
Improvement over Evofold

Model single stranded regions as mixture of conserved and non conserved components.

Better model for gaps

Consider secondary structure folding energy

For each base pair, assign a score based on its posterior likelihood.

Take the sum of all such pairs.
Test on CMfinder motifs in ENCODE regions

FDR vs score ranks in the original alignments
Summary

ncRNA - apparently widespread, much interest
Covariance Models - powerful but expensive tool for ncRNA motif representation, search, discovery
Rigorous/Heuristic filtering - typically 100x speedup in search with no/little loss in accuracy
CMfinder - good CM-based motif discovery in unaligned sequences
  Pipeline integrating comp and bio for ribowitch discovery
Potentially many ncRNAs with weak sequence conservation in vertebrates.
Course Wrap Up
“High-Throughput BioTech”

Sensors
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction

Controls
- Cloning
- Gene knock out/knock in
- RNAi

Flooding of data

“Grand Challenge” problems
CS/Math/Stats Points of Contact

Scientific visualization
  Gene expression patterns

Databases
  Integration of disparate, overlapping data sources
  Distributed genome annotation in face of shifting underlying coordinates

AI/NLP/Text Mining
  Information extraction from journal texts with inconsistent nomenclature,
  indirect interactions, incomplete/inaccurate models,…

Machine learning
  System level synthesis of cell behavior from low-level heterogeneous data (DNA
  sequence, gene expression, protein interaction, mass spec,

Algorithms
  …
Frontiers & Opportunities

New data:
- Proteomics, SNPs, association studies, array CGH,
- comparative sequence information, methylation,
- chromatin structure, ChIP-seq, ncRNA, interactome

New methods:
- graphical models? rigorous filtering?

Data integration
- many, complex, noisy sources

Systems Biology
Open Problems:
splicing, alternative splicing
multiple sequence alignment (genome scale, w/ RNA etc.)
protein & RNA structure
interaction modeling
network models
RNA trafficking
ncRNA discovery
chromatin dynamics
...
Exciting Times

Lots to do
Various skills needed
I hope I’ve given you a taste of it
Thanks!