RNA Search and Motif Discovery

CSEP 527
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?
Nussinov:

Computation Order

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

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

\[ B(i,j) = \max \{ B(i,j-1) \]

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

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

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

Structure prediction via comparative analysis

Covariance Models (CMs) represent RNA sequence/structure motifs

Fast CM search

Motif Discovery

Applications in prokaryotes & vertebrates
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
A. L19 (rplS) mRNA leader

Example: Ribosomal Autoregulation
Excess L19 represses L19 (RF00556; 555-559 similar)
Mutual Information

\( x_k \): letter from col \( k \); \( f_{x_k} \): its freq in col \( k \); \( f_{x_i, x_j} \): pair freq

\[
M_{ij} = \sum_{x_i, x_j} f_{x_i, x_j} \log_2 \frac{f_{x_i, x_j}}{f_{x_i} f_{x_j}} ; \quad 0 \leq M_{ij} \leq 2 \quad (4 \text{ letters } \Rightarrow 2 \text{ bits})
\]

Max when no seq conservation but perfect pairing

\( \text{MI} \left\{ \begin{array}{l}
given \text{ letter in col } i, \text{ what is mate in col } j? \\
\text{expected score gain from using a pair state} \quad (\text{below})
\end{array} \right. \)

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>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.55</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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.
**MI-Based Structure-Learning**

**Problem:** Find best (max total MI) pseudo-knot-free subset of column pairs among $i\ldots j$.

**Solution:** “Just like Nussinov/Zucker folding”

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

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” (i.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 (Sakakibara 94)
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”
  ≈ 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.*

$M_j$: Match states (20 emission probabilities)

$I_j$: Insert states (Background emission probabilities)

$D_j$: 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

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

Add “column pairs” and pair emission probabilities for base-paired regions

CUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAAAG
GGAAUGUGGAAUAUCUuuuggauu......AGUAAGCAUUCC
UUCAUUACGAAA..CA.................AGUAGUAAUGGA
GGAAUGUGGAAUAUCUuuaugauu......AGUAAACAUUCC
<<<<<><<<<<<<<<<

... paired columns ...
**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)

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

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

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>

3 test sets from ED 94

Disallowing / allowing pseudoknots

\[
\left( \sum_{i=1}^{n} \max_{j} M_{ij} \right) / 2
\]
Model Training

unaligned sequences

random alignment

multiple alignment

alignment

(EM)

parameter reestimation

covariance model

model construction

(structure prediction)
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.
tRNASEcanSE

Uses 3 older heuristic tRNA finders as prefilter
Uses CM built as described for final scoring
Actually 3(?) different CMs
  eukaryotic nuclear
  prokaryotic
  organellar
Used in “all” genome annotation projects
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):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|          | GUUCCUGCUUCAACAGUGUUUUGGAUGGAAC | UUUCUUC.UUCAACAGUGUUUUGGAUGGAAC | UUUCUGUUUCAACAGUGCUUGGA.GGAAC | UUUAUC..AGUGACAGAGUUCACU.AUAAA | UCUCUUGCUUCAACAGUGUUUUGGAUGGAAC | AUUAUC..GGGAACAGUGUUUUCCC.AUAUU | UCUUGC..UUCAACAGUGUUUUGGACGGAAG | UGUAUC..GGAGACAGUGAUCUC.AUAUG | AUUAUC..GGAGACAGUGCUCUCC.AUAUU | UCUCUGCUUCAACAGUGCUUGGACGGAAC | UUUAUC..GGAGACAGUGAUCUC.AUAUG | UUCCUGCUUCAACAGUGCUUGGAACGGAAC | GUACUUGCUUCAACAGUGUUUUGGAACGGAAC | AUUAUC..GGAGACAGUGAACCUC.C.37AUG | AUUAUC..GGAGACAGUGAACCUC.C.37AUG | <<<<<...<<<......>>>>.>>>>>.
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 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>
<tr>
<td>Rel 13.0</td>
<td>9/17</td>
<td>2686</td>
<td></td>
</tr>
<tr>
<td>Rel 14.3</td>
<td>9/20</td>
<td>3446</td>
<td></td>
</tr>
</tbody>
</table>

DB size:
- Rel 1.0, 1/03: 55k instances (~8GB)
- Rel 7.0, 3/05: 363k instances (~160GB)
- Rel 9.0, 7/08: 636k instances (~320GB)
From: Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families
Rfam – key issues

Overly narrow families

Variant structures/unstructured RNAs

Spliced RNAs

RNA pseudogenes

- Human ALU is SRP-related w/ 1.1m copies
- Mouse B2 repeat (350k copies) tRNA related

Speed & sensitivity

Motif discovery/hand-made models
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
(glycine riboswitch)
(operon)
(and 19 more species)

(Mandal, et al., 2004)

BLAST-based
CM-based
(and 42 more species)
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

BLAST → CM

CM → junk, hits

1 month, 1000 computers

Our Work

EMBL → Ravenna

Ravenna → CM

CM → hits

~2 months, 1000 computers

Rfam Goal

EMBL → CM

CM → junk

10 years, 1000 computers
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 score < threshold;
- Actually run CM on the rest (the promising parts)
- Assignment of HMM transition/emission scores is key
  (a large convex optimization problem)

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

CM

25 emissions per state

HMM

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

Need: log Viterbi scores $\text{CM} \leq \text{HMM}$
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. $\sum \neq 1$

(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 $CM \leq HMM$

- $P_{AA} \leq L_A + R_A$
- $P_{AC} \leq L_A + R_C$
- $P_{AG} \leq L_A + R_G$
- $P_{AU} \leq L_A + R_U$
- $P_{A-} \leq L_A + R_-$
- $P_{CA} \leq L_C + R_A$
- $P_{CC} \leq L_C + R_C$
- $P_{CG} \leq L_C + R_G$
- $P_{CU} \leq L_C + R_U$
- $P_{C-} \leq L_C + R_-

NB: HMM not a prob. model
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:
\[ L_U = 0, R_A = 1, R_G = 4 \]

<table>
<thead>
<tr>
<th>Assuming ACGU (\approx 25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Opt 1:</strong></td>
</tr>
<tr>
<td>( L_U + (R_A + R_G)/2 = 4 )</td>
</tr>
</tbody>
</table>

| **Opt 2:**                  |
| \( L_U + (R_A + R_G)/2 = 2.5 \) |
Optimizing filtering

For any nucleotide sequence $x$:
- Viterbi-score($x$) = $\max\{\text{score}(\pi) \mid \pi \text{ emits } x\}$
- Forward-score($x$) = $\sum\{\text{score}(\pi) \mid \pi \text{ emits } x\}$

Expected Forward Score
- $E(L_i, R_i) = \sum_{\text{all sequences } x} \text{Forward-score}(x) \cdot \Pr(x)$
- 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$\downarrow \Rightarrow$ Viterbi$\downarrow \Rightarrow$ filter$\downarrow"$")
  - But still rigorous because "subject to score Lin.Ineq.s"

Under 0th-order background model
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 $E(L_i, R_i)$
(subject to linear constraints)

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

Forward:

$\begin{align*}
  f_k(i) &= P(x_1 \ldots x_i, \pi_i = k) \\
  f_l(i + 1) &= e_l(x_{i+1}) \sum_k f_k(i) a_{k,l}
\end{align*}$

Viterbi:

$vl(i + 1) = e_l(x_{i+1}) \cdot \max_k (v_k(i) a_{k,i})$

$\frac{\partial E(L_1, L_2, \ldots)}{\partial L_i}$
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>

Approximately break even, ~100x speedup

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>Pyrococcus 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>
Results: With additional work

<table>
<thead>
<tr>
<th></th>
<th># with BLAST+CM</th>
<th># with rigorous filter series + CM</th>
<th># new</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rfam tRNA</td>
<td>58609</td>
<td>63767</td>
<td>5158</td>
</tr>
<tr>
<td>Group II intron</td>
<td>5708</td>
<td>6039</td>
<td>331</td>
</tr>
<tr>
<td>tRNAscan-SE (human)</td>
<td>608</td>
<td>729</td>
<td>121</td>
</tr>
<tr>
<td>tmRNA</td>
<td>226</td>
<td>247</td>
<td>21</td>
</tr>
<tr>
<td>Lysine riboswitch</td>
<td>60</td>
<td>71</td>
<td>11</td>
</tr>
</tbody>
</table>

And more…
Ravenna implements both rigorous and heuristic filters

Infernal (engine behind Rfam) implements heuristic filters and some other (important) accelerations

E.g., dynamic “banding” of dynamic programming matrix based on the insight that large deviations from consensus length must have low scores.
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
Motif Discovery
RNA Motif Discovery

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

Key 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.
RNA Motif Discovery

Would be great if: given 100 complete genomes from diverse species, we could automatically find all the RNAs.

State of the art: that’s hopeless

Hope: can we exploit biological knowledge to narrow the search space?
RNA Motif Discovery

More promising problem: given a 10-20 unaligned sequences of a few kb, most of which contain instances of one RNA motif of 100-200bp -- find it.

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

Example: corresponding introns of orthologous vertebrate genes

Orthologs = counterparts in different species
Approaches

Align-First: Align sequences, then look for common structure

Fold-First: Predict structures, then try to align them

Joint: Do both together
“Align First” Approach: Predict Struct from Multiple Alignment

… GA … UC …
… GA … UC …
… GA … UC …
… CA … UG …
… CC … GG …
… UA … UA …

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-first: align sequences, then look for common structure

Fold-first: 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

Joint: Do both together

  Sankoff – good but slow

Heuristic
Our Approach: CMfinder
RNA motifs from unaligned sequences

Simultaneous *local* alignment, folding and CM-based motif description via an EM-style learning procedure

- Sequence conservation exploited, but not required
- Robust to inclusion of unrelated and/or flanking sequence
- Reasonably fast and scalable
- Produces a probabilistic model of the motif that can be directly used for homolog search

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006
CMFinder

Simultaneous alignment, folding & motif description
Yao, Weinberg & Ruzzo, Bioinformatics, 2006

- Folding predictions
- Smart heuristics
- Mutual Information
- Candidate alignment
- CM
- Realign

Combines folding & mutual information in a principled way.
Initial Alignment Heuristics

fold sequences separately
candidates: regions with low folding energy
compare candidates via “tree edit” algorithm
find best “central” candidates & align to them
BLAST anchors
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
  (based on single sequence thermodynamic folding predictions)
- intuition: for similar seqs, little MI; fall back on single-sequence folding predictions
- data-dependent, so not strictly Bayesian

Details: see paper
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>0.59</td>
<td>0.05</td>
<td>0</td>
<td>X</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ctRNA_pGA1</td>
<td>RF00236</td>
<td>17</td>
<td>74</td>
<td>83</td>
<td>2</td>
<td>0.91</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>0.89</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>0.94</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>0.83</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>0.0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Intron_gpI</td>
<td>RF00029</td>
<td>75</td>
<td>55</td>
<td>92</td>
<td>2</td>
<td>0.80</td>
<td>0.30</td>
<td>0</td>
<td>0</td>
<td>0.38</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>0.77</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>0.87</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>0.78</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>0.77</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>0.66</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>0.91</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>0.68</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>0.88</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>0.80</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>0.72</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>0.73</td>
<td>0.0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>0</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>0.81</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

*Min/Max in col  Bold = best in row*
Discovery in Bacteria

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

Zizhen Yao¹*, Jeffrey Barrick²*, Zasha Weinberg³, Shane Neph¹,⁴, Ronald Breaker²,³,⁵, Martin Tompa¹,⁴, Walter L. Ruzzo¹,⁴

Published online 9 July 2007

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

Zasha Weinberg¹*, Jeffrey E. Barrick²,³, Zizhen Yao⁴, Adam Roth², Jane N. Kim¹, Jeremy Gore¹, Joy Xin Wang¹,², Elaine R. Lee¹, Kirsten F. Block¹, Narasimhan Sudarsan¹, Shane Neph⁵, Martin Tompa⁴,⁵, Walter L. Ruzzo⁴,⁵ and Ronald R. Breaker¹,²,³
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
Use the Right Data; Do Genome Scale Search

Dataset collection → Footprinter → CMfinder → Ravenna Search
Right Data: Why/How

We can recognize, say, 5-10 good examples amidst 20 extraneous ones (but not 5 in 200 or 2000) of length 1k or 10k (but not 100k)
Regulators often near regulatees (protein coding genes), which are usually recognizable cross-species
So, look near similar genes (“homologs”)
Many riboswitches, e.g., are present in ~5 copies per genome
(Not strategy used in vertebrates - 1000x larger genomes)
A pipeline for RNA motif genome scans

Overall Pipeline & Processing Times

Input from ~70 complete Firmicute genomes available in late 2005-early 2006, totaling ~200 megabases

- Identify CDD group members: 2946 CDD groups, < 10 CPU days
- Retrieve upstream sequences
- Footprinter ranking: < 10 CPU days
- CMfinder: 35975 motifs, 1 ~ 2 CPU months
  - Motif postprocessing: 1740 motifs
  - RaveNnA: 10 CPU months
  - CMfinder refinement: < 1 CPU month
  - Motif postprocessing: 1466 motifs
### Table 1: Motifs that correspond to Rfam families

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score RAV</th>
<th>Score CMF</th>
<th>Score FP</th>
<th># RAV</th>
<th># CMF</th>
<th># FP</th>
<th>ID</th>
<th>Gene</th>
<th>Description</th>
<th>CDD</th>
<th>Rfam</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
<td>107</td>
<td></td>
<td>3400</td>
<td>367</td>
<td>11</td>
<td>9904</td>
<td>IlvB</td>
<td>Thiamine pyrophosphate-requiring enzymes</td>
<td></td>
<td>RF00230 T-box</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>344</td>
<td></td>
<td>3115</td>
<td>96</td>
<td>22</td>
<td>13174</td>
<td>COG3859</td>
<td>Predicted membrane protein</td>
<td></td>
<td>RF00059 THI</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>1284</td>
<td></td>
<td>2376</td>
<td>112</td>
<td>6</td>
<td>11125</td>
<td>MetH</td>
<td>Methionine synthase I specific DNA methylase</td>
<td></td>
<td>RF00162 S_box</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>5</td>
<td></td>
<td>2327</td>
<td>30</td>
<td>26</td>
<td>9991</td>
<td>COG0116</td>
<td>Predicted N6-adenine-specific DNA methylase</td>
<td></td>
<td>RF00011</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>66</td>
<td></td>
<td>2228</td>
<td>49</td>
<td>18</td>
<td>4383</td>
<td>DHBP</td>
<td>3,4-dihydroxy-2-butanol 4-phosphate synthase</td>
<td></td>
<td>RF00050 RFN</td>
</tr>
<tr>
<td>7</td>
<td>145</td>
<td>952</td>
<td></td>
<td>1429</td>
<td>51</td>
<td>7</td>
<td>10390</td>
<td>GuaA</td>
<td>GMP synthase</td>
<td></td>
<td>RF00167 Purine</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>108</td>
<td></td>
<td>1322</td>
<td>29</td>
<td>13</td>
<td>10732</td>
<td>GcvP</td>
<td>Glycine cleavage system protein P</td>
<td></td>
<td>RF00504 Glycine</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>749</td>
<td></td>
<td>1235</td>
<td>28</td>
<td>7</td>
<td>24631</td>
<td>DUF149</td>
<td>Uncharacterised BCR, YbaB family COG0718</td>
<td></td>
<td>RF00169 SRP_bact</td>
</tr>
<tr>
<td>10</td>
<td>123</td>
<td>1358</td>
<td></td>
<td>1222</td>
<td>36</td>
<td>6</td>
<td>10986</td>
<td>CbiB</td>
<td>Cobalamin biosynthesis protein CobD/CbiB</td>
<td></td>
<td>RF00174 Cobalamin</td>
</tr>
<tr>
<td>20</td>
<td>137</td>
<td>1133</td>
<td></td>
<td>899</td>
<td>32</td>
<td>7</td>
<td>9895</td>
<td>LysA</td>
<td>Diaminopimelate decarboxylase</td>
<td></td>
<td>RF00168 Lysine</td>
</tr>
<tr>
<td>21</td>
<td>36</td>
<td>141</td>
<td></td>
<td>896</td>
<td>22</td>
<td>10</td>
<td>10727</td>
<td>TerC</td>
<td>Membrane protein TerC</td>
<td></td>
<td>RF00080 yybP-ykoY</td>
</tr>
<tr>
<td>39</td>
<td>202</td>
<td>684</td>
<td></td>
<td>664</td>
<td>25</td>
<td>5</td>
<td>11945</td>
<td>MgtE</td>
<td>Mg/Co/Ni transporter MgtE</td>
<td></td>
<td>RF00380 ykoK</td>
</tr>
<tr>
<td>40</td>
<td>26</td>
<td>74</td>
<td></td>
<td>645</td>
<td>19</td>
<td>18</td>
<td>10323</td>
<td>GlmS</td>
<td>Glucosamine 6-phosphate synthetase</td>
<td></td>
<td>RF00234 glmS</td>
</tr>
<tr>
<td>53</td>
<td>208</td>
<td>192</td>
<td></td>
<td>561</td>
<td>21</td>
<td>5</td>
<td>10892</td>
<td>OpuBB</td>
<td>ABC-type proline/glycine betaine transport systems</td>
<td></td>
<td>RF00005 tRNA1</td>
</tr>
<tr>
<td>122</td>
<td>99</td>
<td>239</td>
<td></td>
<td>413</td>
<td>10</td>
<td>7</td>
<td>11784</td>
<td>EmrE</td>
<td>Membrane transporters of cations and cationic drug</td>
<td></td>
<td>RF00442 ykkC-yxkD</td>
</tr>
<tr>
<td>255</td>
<td>392</td>
<td>281</td>
<td></td>
<td>268</td>
<td>8</td>
<td>6</td>
<td>10272</td>
<td>COG0398</td>
<td>Uncharacterized conserved protein</td>
<td></td>
<td>RF00023 tmRNA</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 FootPrinter 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>#</td>
<td>Sn</td>
<td>Sp</td>
</tr>
<tr>
<td>RF00174</td>
<td>183</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>RF00504</td>
<td>92</td>
<td>0.56</td>
<td>0.96</td>
</tr>
<tr>
<td>RF00234</td>
<td>34</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00168</td>
<td>80</td>
<td>0.82</td>
<td>0.98</td>
</tr>
<tr>
<td>RF00167</td>
<td>86</td>
<td>0.86</td>
<td>0.93</td>
</tr>
<tr>
<td>RF00050</td>
<td>133</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>RF00011</td>
<td>144</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>RF00162</td>
<td>208</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>RF00169</td>
<td>177</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>RF00230</td>
<td>453</td>
<td>0.96</td>
<td>0.61</td>
</tr>
<tr>
<td>RF00059</td>
<td>326</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00442</td>
<td>19</td>
<td>0.90</td>
<td>0.53</td>
</tr>
<tr>
<td>RF00380</td>
<td>49</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00080</td>
<td>41</td>
<td>0.32</td>
<td>0.89</td>
</tr>
<tr>
<td>mean</td>
<td>145</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>median</td>
<td>113</td>
<td>0.91</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Tbl 2: Prediction accuracy compared to prokaryotic subset of Rfam full alignments.
Membership: # of seqs in overlap between our predictions and Rfam’s, the sensitivity (Sn) and specificity (Sp) of our membership predictions. Overlap: the avg len 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: the avg # of correctly predicted canonical base pairs (in overlapped regions) in the secondary structure (bp), and sensitivity and specificity of our predictions. †After 2nd RaveNnA scan, membership Sn of Glycine, Cobalamin increased to 76% and 98% resp., Glycine Sp unchanged, but Cobalamin Sp dropped to 84%.
Table 3: High ranking motifs not found in Rfam

<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></td>
</tr>
<tr>
<td>31</td>
<td>31</td>
<td>10164</td>
<td>lnfC: Translation initiation factor 3</td>
<td>IF-3 r-protein leader; see Supp</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</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: Penicillinase repressor</td>
<td>Blal, Mcel 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></td>
</tr>
<tr>
<td>194</td>
<td>9</td>
<td>10353</td>
<td>FusA: Translation elongation factors</td>
<td>EF-G r-protein leader</td>
</tr>
</tbody>
</table>
A  L19 (rplS) mRNA leader

Example: Ribosomal Autoregulation
Excess L19 represses L19 (RF00556; 555-559 similar)
Examples: 6 (of 22) Representative motifs

GEMM
Sudarsan, et al
Science, 2008

SAM-IV
Weinberg, et al
RNA ’08

SAH
Wang, et al
Mol Cell, 2008

MoCo
Regulski et al
Mol Microbiol ’08

sucA

Legend
nt: nucleotides, SD: Shine-Dalgarno
start: start codon, R: A/G, Y: C/U

nucleotide identity
N 97%
N 90%
N 75%

base pair annotations
has covarying mutations
has compatible mutations
no mutations observed

nucleotide present
variable hairpin
variable loop
modular loop
modular structure

Weinberg, et al.

ncRNA Summary

ncRNA is a “hot” topic
For family homology modeling: CMs
Training & search like HMM (but slower)
Dramatic acceleration possible
Automated model construction possible
New computational methods yield new discoveries
Many open problems