Modeling and Searching for Non-Coding RNA

W.L. Ruzzo

http://www.cs.washington.edu/homes/ruzzo

http://www.cs.washington.edu/homes/ruzzo/courses/gs541/09sp
“... *protein* and *DNA* sequence analysis ... to determine the "periodic table of biology," i.e., the list of *proteins* ..., which can be regarded as the first stage in...”

*No mention of RNA...*
The Message

Cells make lots of RNA  *noncoding* RNA

Functionally important, functionally diverse

Structurally complex

New tools required
  alignment, discovery, search, scoring, etc.
The Outline

The problem: noncoding RNA

Why: it’s important

Some results

Some methods
RNA Secondary Structure: RNA makes helices too

Base pairs

A-U
C-G
U-G

Usually *single* stranded
RNA: Interest
Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.
“Classical” RNAs

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)
tRNA - transfer RNA (~61 kinds, ~ 75 nt)
snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)
RNaseP - tRNA processing (~300 nt)
a handful of others
Bacteria

Triumph of proteins

80% of genome is coding DNA

Functionally diverse

  receptors
  motors
  catalysts
  regulators  (Monod & Jakob, Nobel prize 1965)

  ...

16
Proteins Catalyze Biochemistry: Met Pathways

\[ \text{CH}_3\text{SCH}_2\text{CH}_2\text{C}\text{COO}^- \text{NH}_3^+ \]

\text{Methionine}

\[ \text{ATP} + \text{PP}_i + \text{H}_2\text{O} \rightarrow \text{SAM} \]

\[ \text{S-Adenosylmethionine} \]

\[ \text{CH}_3\text{SCH}_2\text{CH}_2\text{C}\text{COO}^- \text{NH}_3^+ \]

\text{S-Adenosylhomocysteine}

\[ \text{Homocysteine} \]

\[ \text{Serine} \]

\[ \text{H}_2\text{O} \]

\[ \text{THF} \]

\[ \text{N}^5\text{-methyl-THF} \]

\[ \text{biosynthetic methylation} \]

\[ \text{methyl acceptor} \]

\[ \text{methylated acceptor} \]
Proteins Regulate Biochemistry:
The MET Repressor

![Protein Diagram](A)

SAM

COOH

HOOC

![DNA Structure](B)

Protein

DNA

Alberts, et al. 3e.
Not the only way!

Protein way  Riboswitch alternative

Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003
Not the only way!

Protein way

Riboswitch alternatives

SAM-II


Corbino et al., Genome Biol. 2005

Alberts, et al., 3e.
Not the only way!

Protein way

Riboswitch alternatives

SAM-III

Fuchs et al., NSMB 2006


Corbino et al., Genome Biol. 2005
Not the only way!

Protein way  Riboswitch alternatives

Alberts, et al., 3e.

Corbino et al., Genome Biol. 2005

Weinberg et al., RNA 2008


Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006


Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006

Weinberg et al., RNA 2008
Tip of the iceberg?

Approximately 20 riboswitches known
Regulate ~20,000 operons in sequenced organisms
All found since 2003
More ribo-regulators are known than protein transcription factors in many species
Growing number of trans-acting elements, too
E.g., a recent RNA-seq study found >500 small RNA and 127 antisense RNA in V. cholerae
Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.
Vertebrates

Bigger, more complex genomes

<2% coding

But >5% conserved in sequence?

And 50-90% transcribed?

And *structural* conservation, if any, invisible
(without proper alignments, etc.)

What’s going on?
Vertebrate ncRNAs

mRNA, tRNA, rRNA, … of course

PLUS:

snRNA, spliceosome, snoRNA, teleomerase, microRNA, RNAi, SECIS, IRE, piwi-RNA, XIST (X-inactivation), ribozymes, …
Fastest Human Gene?
Human Predictions

Evofold


48,479 candidates (~70% FDR?)
RNAz


30,000 structured RNA elements

1,000 conserved across all vertebrates.

~1/3 in introns of known genes, ~1/6 in UTRs

~1/2 located far from any known gene
E Torarinsson, M Sawera, JH Havgaard, M Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure."


1800 candidates from 36970 (of 100,000) pairs
CMfinder


6500 candidates in ENCODE alone (better FDR, but still high)
Origin of Life?

Life needs

- information carrier: DNA
- molecular machines, like enzymes: Protein
- making proteins needs DNA + RNA + proteins
- making (duplicating) DNA needs proteins

Horrible circularities! How could it have arisen in an abiotic environment?
Origin of Life?

RNA can carry information, too
  RNA double helix; RNA-directed RNA polymerase
RNA can form complex structures
RNA enzymes exist (ribozymes)
RNA can control, do logic (riboswitches)

The “RNA world” hypothesis:
  1st life was RNA-based

Some extant RNAs are relics of that origin, some are “modern” inventions
“Classical” RNAs

tRNA - transfer RNA (~61 kinds, ~75 nt)
rRNA - ribosomal RNA (~4 kinds, 120-5k nt)
snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)
RNaseP - tRNA processing (~300 nt)
RNase MRP - rRNA processing; mito. rep. (~225 nt)
SRP - signal recognition particle; membrane targeting (~100-300 nt)
SECIS - selenocysteine insertion element (~65nt)
6S - ? (~175 nt)
Semi-classical RNAs
(discovery in mid 90’s)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)
Recent discoveries

siRNA (Nobel prize 2006: Fire & Mello)

microRNAs (Lasker prize 2008: Ambros, Baulcombe & Ruvkun)

riboswitches

many ribozymes

regulatory elements

...

Hundreds of families

Rfam release 1, 1/2003: 25 families, 55k instances
Rfam release 9, 7/2008: 603 families, 896k instances
Rfam release 9.1, 1/2009: 1372 families, ??? instances
Why?

RNA’s fold, and function

Nature uses what works
Example: Glycine Regulation

How is glycine level regulated?

Plausible answer:

transcription factors (proteins) bind to DNA to turn nearby genes on or off
The Glycine Riboswitch

Actual answer (in many bacteria):

Mandal et al. Science 2004
Bacillus subtilis
gcvT RNA

(Mandal, Lee, Barrick, Weinberg, Emilsson, Ruzzo, Breaker, Science 2004)
Fig. 3. Cooperative binding of two glycine molecules by the VC I-II RNA. Plot depicts the fraction of VC II (open) and VC I-II (solid) bound to ligand versus the concentration of glycine. The constant, n, is the Hill coefficient for the lines as indicated that best fit the aggregate data from four different regions (fig. S3). Shaded boxes demark the dynamic range (DR) of glycine concentrations needed by the RNAs to progress from 10%- to 90%-bound states.
Riboswitches

~ 20 ligands known; multiple nonhomologous solutions for some
dozens to hundreds of instances of each
TPP known in archaea & eukaryotes
one known in bacteriophage
on/off; transcription/translation; splicing;
combinatorial control
In some bacteria, more riboregulators identified than protein TFs
all found since ~2003
Bacillus subtilis

oriC

purE purK purB
purC purS purQ
purL purF purM
purN purH purD

yxrC yxrB yxrA yxrK
yusC yusB yusA
yvaH yvaJ metK
lysC

yjaA

thiC

yjiA

glmS

ydhL

gluS

yjaB

ribD ribE ribA
ribbonA
ribH ribT

riboT

gcvT
gcvP

gcvB

S-adenosylmethionine

glucosamine-6-phosphate

flavin mononucleotide

thiamine pyrophosphate

adenine

pre-queosine

lysine

glycine
The protein way

Riboswitch alternatives

SAM-I


SAM-II

Corbino et al., Genome Biol. 2005

SAM-III

Fuchs et al., NSMB 2006

SAM-IV

Weinberg et al., RNA 2008
59

Methionine

\[ \text{CH}_3\text{SCH}_2\text{CH}_2\text{C}\text{OO}^- + \text{H}_2\text{O} \rightarrow \text{ATP} + \text{PP}_i + \text{P}_i \]

\[ \text{S-Adenosylmethionine (SAM)} \]

\[ \text{CH}_3\text{SCH}_2\text{CH}_2\text{C}\text{OO}^- + \text{Adenosine} \rightarrow \text{S-Adenosylhomocysteine} \]

\[ \text{S-Adenosylhomocysteine} \]

\[ \text{Homocysteine} \]

\[ \text{Serine} \rightarrow \text{H}_2\text{O} \]

\[ \text{THF} \rightarrow \text{N}^5\text{-methyl-THF} \]

\[ \text{biosynthetic methylation} \]

\[ \text{methyl acceptor} \rightarrow \text{methylated acceptor} \]
Wanted

Good structure prediction tools
Good motif descriptions/models
Good, fast search tools
   (“RNA BLAST”, etc.)
Good, fast motif discovery tools
   (“RNA MEME”, etc.)

Importance of structure makes last 3 hard
Why is RNA hard to deal with?

A: Only 29% identity! *Structure* often trumps *sequence*
Structure Prediction
RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape
Single Seq Secondary Structure Prediction

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

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

Definitely useful, but obviously imperfect
Motif Description
“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”
≈ 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
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
Profile Hmm Structure

Figure 5.2 The transition structure of a profile HMM.

Mj: Match states (4 emission probabilities)
Ij: Insert states (Background emission probabilities)
Dj: Delete states (silent - no emission)
How to model an RNA “Motif”?

Covariance Models (aka “profile SCFG”)

Probabilistic models, like profile HMMs, but adding “column pairs” and pair emission probabilities for base-paired regions
Covariation is strong evidence for base pairing
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

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

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

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

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

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

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

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

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

compare: \( O(qn) \) for profile HMM
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

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></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.
Modeling and Searching for Non-Coding RNA

W.L. Ruzzo

http://www.cs.washington.edu/homes/ruzzo

http://www.cs.washington.edu/homes/ruzzo/courses/gs541/09sp
Outline

Whirlwind tour of ncRNA search & discovery
  RNA motif description (Covariance Model Review)
  Motif search
    Rigorous & heuristic filtering
  Motif discovery
Applications
  Prokaryotes
  Vertebrates
Open problems
An Important Application: Rfam
Rfam – an RNA family DB
Griffiths-Jones, et al., NAR ’03, ’05, ’08

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
Rel 9.0, 7/08: 603 families, 896k instances
Rel 9.1, 1/09: 1372 families, ??? instances
Input (hand-curated):
  MSA “seed alignment”
  SS_cons
Score Threshold
Window Len W
Output:
  CM
  scan results &
  “full alignment”
An Important Need:
Faster Search
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
  (large convex optimization problem)

Weinberg & Ruzzo, Bioinformatics, 2004, 2006
CM’s are good, but slow

Rfam Reality

EMBL → BLAST → CM → junk
1 month, 1000 computers

Our Work

EMBL → Ravenna → CM → hits
~2 months, 1000 computers

Rfam Goal

EMBL → CM → junk → hits
10 years, 1000 computers
Oversimplified CM
(for pedagogical purposes only)
CM to HMM

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

Need: log Viterbi scores CM $\leq$ HMM
Key Issue: 25 scores $\rightarrow$ 10

Need: log Viterbi scores $CM \leq HMM$

\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 \approx 25%

<table>
<thead>
<tr>
<th>Opt 1</th>
<th>Opt 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ L_U + (R_A + R_G)/2 = 4 ]</td>
<td>[ L_U + (R_A + R_G)/2 = 2.5 ]</td>
</tr>
</tbody>
</table>
What should the scores be?

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

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

$\approx$ break even

$\sim 100x$ 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><em>Pyrococcus</em> 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>
Motif Discovery
RNA Motif Discovery

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

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 (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
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
Finding optimal MI, (i.e. opt pairing of cols) is hard(?)
Finding optimal MI without pseudoknots can be done by dynamic programming
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.
CMfinder Accuracy
(on Rfam families with flanking sequence)
Applications:
ncRNA discovery in prokaryotes and vertebrates

Key issue in both cases is exploiting prior knowledge to focus on promising data
Application 1

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

Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo.  
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, find similar genes (“homologs”), look at adjacent DNA

(Not strategy used in vertebrates - 1000x larger genomes)
A pipeline for RNA motif genome scans

Genome Scale Search: Why

Many riboswitches, e.g., are present in ~5 copies per genome
In most close relatives
More examples give better model, hence even more examples, fewer errors
More examples give more clues to function - critical for wet lab verification

But inclusion of non-examples can degrade motif…
Genome Scale Search

CMfinder is directly usable for/with search

Folding predictions

Smart heuristics

Candidate alignment

CM

Search

Realign
Identify CDD group members

2946 CDD groups

Retrieve upstream sequences

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</th>
<th>Score</th>
<th>#</th>
<th>ID</th>
<th>Gene</th>
<th>Description</th>
<th>CDD</th>
<th>Rfam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RAV</td>
<td>CMF</td>
<td>FP</td>
<td>RAV</td>
<td>CMF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>43</td>
<td>107</td>
<td>3400</td>
<td>367</td>
<td>11</td>
<td>9904</td>
<td>IlvB</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>344</td>
<td>3115</td>
<td>96</td>
<td>22</td>
<td>13174</td>
<td>COG3859 Predicted membrane protein</td>
</tr>
<tr>
<td>2</td>
<td>77</td>
<td>1284</td>
<td>2376</td>
<td>112</td>
<td>6</td>
<td>11125</td>
<td>MetH</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>5</td>
<td>2327</td>
<td>30</td>
<td>26</td>
<td>9991</td>
<td>COG0116 Predicted N6-adenine-specific DNA methylase</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>66</td>
<td>2228</td>
<td>49</td>
<td>18</td>
<td>4383</td>
<td>DHBP</td>
</tr>
<tr>
<td>7</td>
<td>145</td>
<td>952</td>
<td>1429</td>
<td>51</td>
<td>7</td>
<td>10390</td>
<td>GuaA</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>108</td>
<td>1322</td>
<td>29</td>
<td>13</td>
<td>10732</td>
<td>GcvP</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>749</td>
<td>1235</td>
<td>28</td>
<td>7</td>
<td>24631</td>
<td>DUF149 Uncharacterised BCR, YbaB family COG0718</td>
</tr>
<tr>
<td>10</td>
<td>123</td>
<td>1358</td>
<td>1222</td>
<td>36</td>
<td>6</td>
<td>10986</td>
<td>CbiB</td>
</tr>
<tr>
<td>20</td>
<td>137</td>
<td>1133</td>
<td>899</td>
<td>32</td>
<td>7</td>
<td>9895</td>
<td>LysA</td>
</tr>
<tr>
<td>21</td>
<td>36</td>
<td>141</td>
<td>896</td>
<td>22</td>
<td>10</td>
<td>10727</td>
<td>TerC</td>
</tr>
<tr>
<td>39</td>
<td>202</td>
<td>684</td>
<td>664</td>
<td>25</td>
<td>5</td>
<td>11945</td>
<td>MgtE</td>
</tr>
<tr>
<td>40</td>
<td>26</td>
<td>74</td>
<td>645</td>
<td>19</td>
<td>18</td>
<td>10323</td>
<td>GlmS</td>
</tr>
<tr>
<td>53</td>
<td>208</td>
<td>192</td>
<td>561</td>
<td>21</td>
<td>5</td>
<td>10892</td>
<td>OpuBB</td>
</tr>
<tr>
<td>122</td>
<td>99</td>
<td>239</td>
<td>413</td>
<td>10</td>
<td>7</td>
<td>11784</td>
<td>EmrE</td>
</tr>
<tr>
<td>255</td>
<td>392</td>
<td>281</td>
<td>268</td>
<td>8</td>
<td>6</td>
<td>10272</td>
<td>COG0398 Uncharacterized conserved protein</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 Cobalamin</td>
<td>183</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>RF00504 Glycine</td>
<td>92</td>
<td>0.56</td>
<td>0.96</td>
</tr>
<tr>
<td>RF00234 glmS</td>
<td>34</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00168 Lysine</td>
<td>80</td>
<td>0.82</td>
<td>0.98</td>
</tr>
<tr>
<td>RF00167 Purine</td>
<td>86</td>
<td>0.86</td>
<td>0.93</td>
</tr>
<tr>
<td>RF00050 RFN</td>
<td>133</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>RF00011 RNaseP_bact_b</td>
<td>144</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>RF00162 S_box</td>
<td>208</td>
<td>0.95</td>
<td>0.97</td>
</tr>
<tr>
<td>RF00169 SRP_bact</td>
<td>177</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>RF00230 T-box</td>
<td>453</td>
<td>0.96</td>
<td>0.61</td>
</tr>
<tr>
<td>RF00059 THI</td>
<td>326</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00442 ykkC-yxkD</td>
<td>19</td>
<td>0.90</td>
<td>0.53</td>
</tr>
<tr>
<td>RF00380 ykoK</td>
<td>49</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>RF00080 yybP-ykoY</td>
<td>41</td>
<td>0.32</td>
<td>0.89</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td>145</td>
<td>0.84</td>
</tr>
<tr>
<td>median</td>
<td></td>
<td>113</td>
<td>0.91</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. 1After 2nd RaveNnA scan, membership Sn of Glycine, Cobalamin increased to 76% and 98% resp., Glycine Sp unchanged, but Cobalamin Sp dropped to 84%.
Application II

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

boxed = confirmed riboswitch (+2 more)

New Riboswitches
(all lab-verified)

SAM – IV (S-adenosyl methionine)
SAH (S-adenosyl homocystein)
MOCO (Molybdenum cofactor)
PreQ1 – II (queuosine precursor)
GEMM (cyclic di-GMP)
Application III

ncRNAs in Vertebrates
ncRNA discovery in Vertebrates

Natural approach: Align, Fold, Score

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, where alignments aren’t trustworthy

Thousands of candidates

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

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin

*Genome Research, Feb 2008, 18(2):242-251*  
PMID: 18096747
Search in Vertebrates

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.
  High false positive rate, but still suggests 1000’s of RNAs.

(We’ve applied CMfinder to whole human genome:
  $O(1000)$ CPU years. Analysis in progress.)

Trust 17-way alignment for orthology, not for detailed alignment
### 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>G+C</th>
<th>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</td>
<td>igs</td>
<td>0.062</td>
<td>380</td>
<td>23</td>
<td>24.5</td>
<td>0.430</td>
<td>5.8%</td>
</tr>
<tr>
<td>35-40</td>
<td>igs</td>
<td>0.082</td>
<td>742</td>
<td>61</td>
<td>70.5</td>
<td>0.103</td>
<td>11.3%</td>
</tr>
<tr>
<td>40-45</td>
<td>igs</td>
<td>0.082</td>
<td>1216</td>
<td>99</td>
<td>129.5</td>
<td>0.00079</td>
<td>18.5%</td>
</tr>
<tr>
<td>45-50</td>
<td>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>
</tr>
<tr>
<td>50-100</td>
<td>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>
</tr>
<tr>
<td>all</td>
<td>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>
</tr>
</tbody>
</table>
### Alignment Matters

#### B. The original MULTIZ alignment without the flanking regions – RNAz Score: 0.132 (no RNA)

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>hg18.chr3</td>
<td>GGTCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----CTGGCAAGTGATATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>panTro1.chr17</td>
<td>GGACATTCTGATCGGGCTC-ATGGGCTGTAACAGAGGAGCT-----ATAAACACTATGACCAAAAGCTGAAA!</td>
</tr>
<tr>
<td>bosTau2.chr18</td>
<td>GGTCAATGGCAAGAGGGCTTT-ATGAGACA--AAACCAGGGAGCT-----CTGGCAAGTGATATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>canFam2.chr3</td>
<td>GGTCAATGGCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----CTGGCAAGTGATATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>oryCun1</td>
<td>GATCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----CTGGCAAGTGATATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>rheMac2.chr2</td>
<td>GGTCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----CTGGCAAGTGATATGACCAAAACTGAAGG!</td>
</tr>
</tbody>
</table>

#### C. The local CMfinder re-alignment of the MULTIZ block – RNAz Score: 0.709 (RNA)

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>hg18.chr3</td>
<td>GGTCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----AGAGGTCTTAAACAGTGATATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>panTro1.chr17</td>
<td>GGACATTCTGATCGGGCTC-ATGGGCTGTAACAGAGGAGCT-----AGAGCTATLAAACACTATGACCAAAAGCTGAAA!</td>
</tr>
<tr>
<td>bosTau2.chr18</td>
<td>GGTCAATGGCAAGAGGGCTTT-ATGAGACA--AAACCAGGGAGCT-----GGAGCTCTDAATGCTGATGACCAAAAGATTGAAGJ</td>
</tr>
<tr>
<td>CanFam2.chr3</td>
<td>GGTCAATGGCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----AGAGCTATLAAACACTATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>oryCun1</td>
<td>GATCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----AGAGCTATLAAACACTATGACCAAAACTGAAGG!</td>
</tr>
<tr>
<td>rheMac2.chr2</td>
<td>GGTCACTTCAAGAGGGCTTT-GTGGGGCTGTTAACCAGTTGAGGT-----AGAGCTATLAAACACTATGACCAAAAGCTGAAA!</td>
</tr>
</tbody>
</table>
Realignment

Average pairwise sequence similarity

% realigned

Average pairwise sequence similarity
10 of 11 top (differentially) expressed
Open Problems - Better CM’s

Optional- and variable-length stems
Riboswitches & other regulatory RNAs often switch between conformations; better search & alignment exploiting both alternatives?

“Augmented” CM handling pseudoknots probably too slow for scan, but plausibly could be used for alignment

Better use of prior knowledge? (GNRA tetraloops, single-stranded A’s, structure motifs, …)
Open Problems - Better algorithms & scoring

incorporating phylogeny in model construction & scoring
  e.g. “mutual information” ignores it
improve scoring by “shuffling”
other ideas for scan filtering
comparing & clustering RNA structures
search/alignment/inference with splicing
Open Problems - Applications & Biology

clustering intergenic sequences, esp prokaryotic

systematic look at eukaryotic UTRs

how to cluster? how to score?

“swiss-cheese phylogenies”

evidence for selection (no dN/dS)
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
Plenty of room for more!