RNA Search and 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

Outline

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

AGCAAGccggccccggauuuuAGUFG
GAUAUCUuuugauuu........AGUF
GAUAUCUuuugauuu........AGUF
GAUAUCUuuugauuu........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”
≈ 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
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.

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)

Profile Hmm Structure

Figure 5.2 The transition structure of a profile HMM.

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

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)

\[ x_i = i^{th} \text{ letter of input} \]
\[ x_{ij} = \text{substring } i,..., j \text{ of input} \]
\[ T_{yz} = P(\text{transition } y \rightarrow z) \]
\[ E_{x_i,x_j}^y = P(\text{emission of } x_i,x_j \text{ from state } y) \]
\[ S_{ij}^y = \max_z \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi) \]

Nussinov: Max Pairing

\[ B(i,j) = \# \text{ pairs in optimal pairing of } r_i \ldots r_j \]
\[ B(i,j) = 0 \text{ for all } i, j \text{ with } i \geq j-4; \text{ otherwise} \]
\[ B(i,j) = \max \{ B(i,j-1) \]
\[ \max \{ B(i,k-1)+1+B(k+1,j-1) \mid i \leq k < j-4 \text{ and } r_i-r_j \text{ may pair} \} \]

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

Model Training

- random alignment
- multiple alignment
- (EM) parameter reestimation
- model construction (structure prediction)
- covariance model
Mutual Information

\[ M_{ij} = \sum_{i,j} f_{i,j} \log \frac{f_{i,j}}{f_i f_j} ; \ 0 \leq M_{ij} \leq 2 \]

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.

<table>
<thead>
<tr>
<th>Position</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position</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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Pseudoknots
disallowed
allowed \( \left( \sum_{i,j} M_{ij} \right) / 2 \)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Avg.</th>
<th>Min</th>
<th>Max</th>
<th>ChustalV 1*</th>
<th>ChustalV 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>id</td>
<td>id</td>
<td>id</td>
<td>accuracy</td>
<td>accuracy</td>
</tr>
<tr>
<td>TEST</td>
<td>.402</td>
<td>.144</td>
<td>1.00</td>
<td>64%</td>
<td>44.7</td>
</tr>
<tr>
<td>SIM100</td>
<td>.306</td>
<td>.131</td>
<td>.996</td>
<td>54%</td>
<td>39.7</td>
</tr>
<tr>
<td>SIM65</td>
<td>.302</td>
<td>.111</td>
<td>.655</td>
<td>47%</td>
<td>31.8</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the training and test sets of 100 RNA 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 no-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 ChustalV, see text.
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

<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>91%</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 RNA (U models).
Task 3: 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 will 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

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
Our Work
Rfam Goal

EMBL
BLAST
CM
hits
junk
1 month, 1000 computers

EMBL
Ravenna
CM
hits
junk
~2 months, 1000 computers

EMBL
CM
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 → 10

Need: log Viterbi scores CM ≤ HMM
Viterbi/Forward Scoring

Path $\pi$ defines transitions/emissions
Score($\pi$) = 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 $x$:
  Viterbi-score($x$) = $\max \{ \text{score}(\pi) \mid \pi \text{ emits } x \}$
  Forward-score($x$) = $\Sigma \{ \text{score}(\pi) \mid \pi \text{ emits } x \}$

Rigorous Filtering

Any scores satisfying the linear inequalities give rigorous filtering
Proof:
CM Viterbi path score
  $\leq$ “corresponding” HMM path score
  $\leq$ Viterbi HMM path score
  (even if it does not correspond to any CM path)

Key Issue: 25 scores → 10

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

Some scores filter better

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 $x$:

- 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

$E(L, R) = \sum \text{all sequences } x \text{ Forward-score}(x)^{\times} \Pr(x)$

NB: $E$ is a function of $L$, $R$ only

Optimization:

- Minimize $E(L, R)$ subject to score Lin. Ineq.

This is heuristic ("forward $\downarrow$ Viterbi $\downarrow$ filter $\downarrow"$

But still rigorous because "subject to score Lin. Ineq.s"

Calculating $E(L_i, R_i)$

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

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

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

Convex:

- local max = global max;
- simple "hill climbing" works

Nonconvex:

- can be many local maxima, $\ll$ 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 mar</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>12</td>
</tr>
<tr>
<td>U4 snoRNA</td>
<td>203</td>
<td>290</td>
<td>7</td>
</tr>
<tr>
<td>G-box</td>
<td>128</td>
<td>331</td>
<td>3</td>
</tr>
<tr>
<td>U6 snRNA</td>
<td>1462</td>
<td>1464</td>
<td>2</td>
</tr>
<tr>
<td>US 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>

### RNA Motif Discovery

Task 4: 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

... 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 sequences, then look for common structure
Predict structures, then try to align them
Do both together

single-seq struct prediction only ~ 60% accurate;
exacerbated by flanking seq; no biologically-validated model for structural alignment

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 → CM → 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
- (EM) parameter reestimation
- Covariance model
- Model construction (structure prediction)

---

**CMfinder Outline**

- Folding predictions
- Heuristics
- Candidate alignment
- CM
- Search
- M step
- E step
- Realign

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

Summary of Rfam test families and results

<table>
<thead>
<tr>
<th>ID</th>
<th>Family</th>
<th>Knt.ID</th>
<th>Gene</th>
<th>Seq.</th>
<th>Length</th>
<th>Pyl</th>
<th>Chimeric</th>
<th>CM-Prf</th>
<th>CWM-RNAfolded</th>
<th>Cymac</th>
<th>Foldalign</th>
<th>CD-RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>coli</td>
<td>KO00074</td>
<td>12</td>
<td>40</td>
<td>216</td>
<td>0.80</td>
<td>0.05</td>
<td>X</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>mRNAG1A</td>
<td>KO00036</td>
<td>17</td>
<td>74</td>
<td>85</td>
<td>0.80</td>
<td>0.75</td>
<td>0.01</td>
<td>0.72</td>
<td>0</td>
<td>0.80</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>BloodCRE</td>
<td>KO00048</td>
<td>56</td>
<td>81</td>
<td>61</td>
<td>0.80</td>
<td>0.74</td>
<td>0.01</td>
<td>0.22</td>
<td>0</td>
<td>0.80</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Sf1aAHR</td>
<td>KO00041</td>
<td>32</td>
<td>77</td>
<td>78</td>
<td>0.81</td>
<td>0.65</td>
<td>0.01</td>
<td>0.78</td>
<td>0.80</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>5</td>
<td>Gadd32</td>
<td>KO00034</td>
<td>63</td>
<td>50</td>
<td>46</td>
<td>0.80</td>
<td>0.65</td>
<td>0.01</td>
<td>0.78</td>
<td>0.80</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>MDA5</td>
<td>KO00002</td>
<td>63</td>
<td>50</td>
<td>46</td>
<td>0.80</td>
<td>0.65</td>
<td>0.01</td>
<td>0.78</td>
<td>0.80</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>7</td>
<td>SirtuinEl</td>
<td>KO00029</td>
<td>75</td>
<td>55</td>
<td>92</td>
<td>0.97</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.90</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>lps-3</td>
<td>KO00037</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>0.87</td>
<td>0.42</td>
<td>0</td>
<td>0.71</td>
<td>0.78</td>
<td>0.65</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>lps-4</td>
<td>KO00038</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>0.78</td>
<td>0.51</td>
<td>0.01</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>10</td>
<td>Lspv</td>
<td>KO00048</td>
<td>48</td>
<td>48</td>
<td>45</td>
<td>0.97</td>
<td>0.24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>mm-10</td>
<td>KO00044</td>
<td>11</td>
<td>90</td>
<td>75</td>
<td>0.86</td>
<td>0.59</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0.46</td>
<td>0.35</td>
</tr>
<tr>
<td>12</td>
<td>ParB</td>
<td>KO00007</td>
<td>20</td>
<td>55</td>
<td>103</td>
<td>0.90</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.37</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>Rsf1</td>
<td>KO00050</td>
<td>47</td>
<td>69</td>
<td>159</td>
<td>0.68</td>
<td>0.26</td>
<td>0</td>
<td>0</td>
<td>0.27</td>
<td>0.17</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>RmsfCRE</td>
<td>KO00020</td>
<td>12</td>
<td>71</td>
<td>86</td>
<td>0.88</td>
<td>0.22</td>
<td>0.02</td>
<td>0.69</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>15</td>
<td>col-10</td>
<td>KO00044</td>
<td>23</td>
<td>90</td>
<td>45</td>
<td>0.88</td>
<td>0.48</td>
<td>0.04</td>
<td>0.43</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>16</td>
<td>Lys</td>
<td>KO00012</td>
<td>64</td>
<td>55</td>
<td>113</td>
<td>0.72</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>Selins</td>
<td>KO00003</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>0.78</td>
<td>0.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>TransRNA5</td>
<td>KO00053</td>
<td>22</td>
<td>72</td>
<td>86</td>
<td>0.65</td>
<td>0.23</td>
<td>0.02</td>
<td>0.30</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Average Accuracy: 0.79, Average Specificity: 0.81, Average Sensitivity: 0.77

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

Retrieve upstream sequences

Folding predictions

Candidate alignment

Smart heuristics

CM

Realign

Search

Results

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score</th>
<th># CDD</th>
<th>Genes</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>43</td>
<td>107</td>
<td>3400</td>
<td>367</td>
<td>10</td>
</tr>
<tr>
<td>0</td>
<td>43</td>
<td>107</td>
<td>3400</td>
<td>367</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1: Most hit correspond to Rfam families. “Rank” is the column dev ranks for ranked motif clusters after genome scale (RAV), CMfinder refinement before genome scans (CMF), and FootPrinter results (FP). We used the same ranking scheme for RAV and CMF. “Score” is the description of the motif, and “Value” is the score.
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 (‘n’), the sensitivity (‘Sn’) and specificity (‘Sp’) of our membership predictions. "Overlap": avg length of overlap between our predictions and Rfam’s (‘n’); 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 (on overlapped regions) and the sensitivity (‘Sn’) and specificity (‘Sp’) of our predictions. After another iteration of RaveRna scan and refinement, the membership sensitivities of Glycine and Cobalamin increased to 76% and 87% respectively, while the specificity of Glycine remained the same, and specificity of Cobalamin dropped to 84%.

Task 5: Application

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

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


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

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.

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
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 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>
</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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>

Comparison with Evofold, RNAz

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

Realignment

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

Floods 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

Frontiers & Opportunities

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

...