RNA Search and Motif Discovery

CSE 427
Winter 2008

Outline

Task 1: RNA 2^{ary} Structure Prediction (last time)

Task 2: RNA Motif Models

- Covariance Models
- Training & “Mutual Information”

Task 3: Search

- Rigorous & heuristic filtering

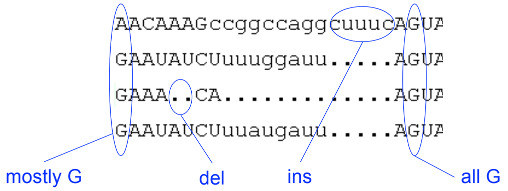
Task 4: Motif discovery

Task 2: Motif Description

How to model an RNA “Motif”?

Conceptually, start with a profile HMM:

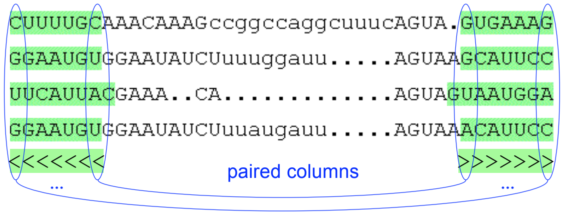
- from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position
- given a new seq, estimate likelihood that it could be generated by the model, & align it to the model



```
AACAAAGccggccaggcuuucAGUA
GAAUAUCUuuuggauu.....AGUA
GAAA..CA.............AGUA
GAAUAUCUuuaugauu.....AGUA
```

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
<<<<<<<                      >>>>>>>
```

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

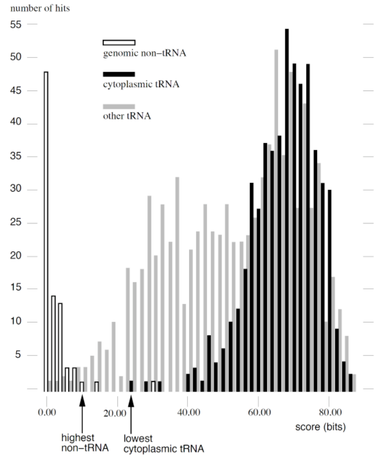
You set a score threshold

Anything above threshold → a “hit”

Scoring:

- “Forward” / “Inside” algorithm - sum over all paths
- Viterbi approximation - find single best path
  (Bonus: alignment & structure prediction)

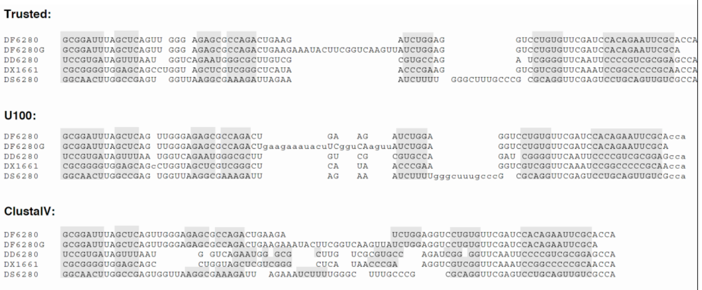
Example: searching for tRNAs



highest non–tRNA

lowest cytoplasmic tRNA

Alignment Quality



Trusted:

```
DF6280    GCGGAUUUAGCUCAGUU GGG AGAGCGCCAGACUGAAG                  AUCUGGAG            GUCCUGUGUUCGAUCCACAGAAUUCGCACCA
DF6280G   GCGGAUUUAGCUCAGUU GGG AGAGCGCCAGACUGAAGAAAUACUUCGGUCAAGUUAUCUGGAG            GUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAAU  GGUCAGAAUGGGCGCUUGUCG                  CGUGCCAG            A UCGGGGUUCAAUUCCCCGUCGCGGAGCCA
DX1661    CGCGGGUGGAGCAGCCUGGU AGCUCGUCGGGCUCAUA                   ACCCGAAG            GUCGUCGGUUCAAAUCCGGCCCCCGCAACCA
DS6280    GGCAACUUGGCCGAGU  GGUUAAGGCGAAAGAUUAGAA                  AUCUUUU  GGGCUUUGCCCG CGCAGGUUCGAGUCCUGCAGUUGUCGCCA
```

U100:

```
DF6280    GCGGAUUUAGCUCAG UUGGGAGAGCGCCAGACU           GA   AG     AUCUGGA            GGUCCUGUGUUCGAUCCACAGAAUUCGCAcca
DF6280G   GCGGAUUUAGCUCAG UUGGGAGAGCGCCAGACUgaagaaauacuUCgguCAaguuAUCUGGA            GGUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAA UGGUCAGAAUGGGCGCUU           GU   CG     CGUGCCA            GAU CGGGGUUCAAUUCCCCGUCGCGGAGcca
DX1661    CGCGGGUGGAGCAGcCUGGUAGCUCGUCGGGCU           CA   UA     ACCCGAA            GGUCGUCGGUUCAAAUCCGGCCCCCGCAAcca
DS6280    GGCAACUUGGCCGAG UGGUUAAGGCGAAAGAUU           AG   AA     AUCUUUUgggcuuugcccG  CGCAGGUUCGAGUCCUGCAGUUGUCGcca
```

ClustalV:

```
DF6280    GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGA                  UCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
DF6280G   GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAAAUACUUCGGUCAAGUUAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCA
DD6280    UCCGUGAUAGUUUAAU       G GUCAGAAUGG GCG    CUUG  UCGCGUGCC   AGAUCGG GGUUCAAUUCCCCGUCGCGGAGCCA
DX1661    CGCGGGUGGAGCAGC        CUGGUAGCUCGUCGGG    CUCA  UAACCCGA    AGGUCGUCGGUUCAAAUCCGGCCCCCGCAACCA
DS6280    GGCAACUUGGCCGAGUGGUUAAGGCGAAAGAUU  AGAAAUCUUUUGGGC  UUUGCCCG     CGCAGGUUCGAGUCCUGCAGUUGUCGCCA
```

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

Slightly different evaluation criteria

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

Profile Hmm Structure

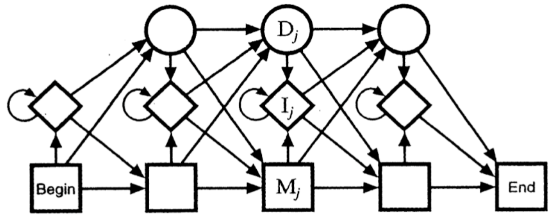


Figure 5.2 *The transition structure of a profile HMM.*

Mj: Match states (20 emission probabilities)

Ij: Insert states (Background emission probabilities)

Dj: Delete states (silent - no emission)

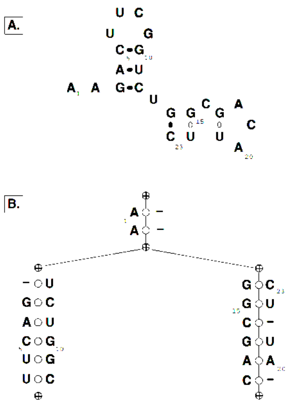
CM Structure

A: Sequence + structure

B: the CM “guide tree”

C: probabilities of letters/ pairs & of indels

Think of each branch being an HMM emitting both sides of a helix (but 3’ side emitted in reverse order)

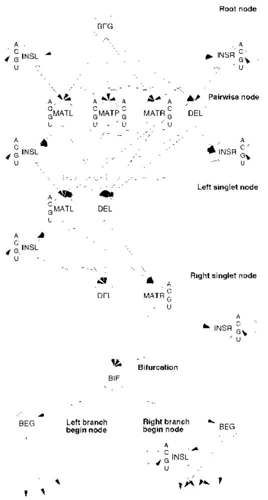


Overall CM Architecture

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)

$x_i \quad = i^{th}$ letter of input

$x_{ij} \quad =$ substring $i,\ldots,j$ of input

$T_{yz} \quad = P(\text{transition } y \to z)$

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

$S^{y}_{ij} \quad = \max_{\pi} \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi)$

CM Viterbi Alignment

(the “inside” algorithm)

$$S^{y}_{ij} = \max_{\pi} \log P(x_{ij} \text{ generated starting in state } y \text{ via path } \pi)$$

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

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

compare: O(qn) for profile HMM

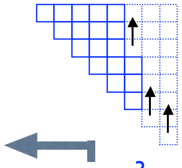
Nussinov: Max Pairing

B(i,j) = # pairs in optimal pairing of $r_i \ldots r_j$

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

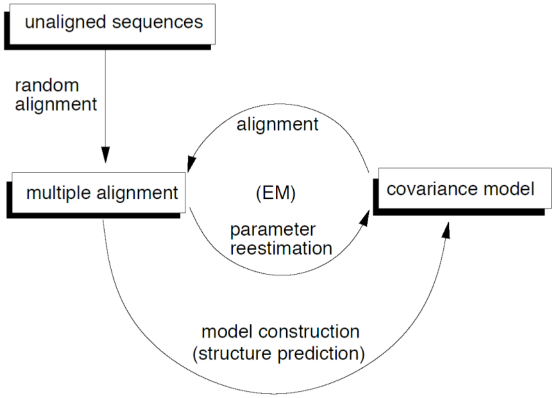
B(i,j) = max of:

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



Time: O(n^3)

Model Training



Mutual Information

$$M_{ij} = \sum_{xi,xj} f_{xi,xj} \log_2 \frac{f_{xi,xj}}{f_{xi} f_{xj}}; \quad 0 \le M_{ij} \le 2$$

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)

*	1	2	3	4	5	6	7	8	9	*
	A	G	A	U	A	A	U	C	U	
	A	G	A	U	C	A	U	C	U	
	A	G	A	C	G	U	U	C	U	
	A	G	A	U	U	U	U	C	U	
	A	G	C	C	A	G	G	C	U	
	A	G	C	G	C	G	G	C	U	
	A	G	C	U	G	C	G	C	U	
	A	G	C	A	U	C	G	C	U	
	A	G	G	U	A	G	C	C	U	
	A	G	G	G	C	G	C	C	U	
	A	G	G	U	G	U	C	C	U	
	A	G	G	C	U	U	C	C	U	
	A	G	U	A	A	A	A	C	U	
	A	G	U	C	C	A	A	C	U	
	A	G	U	U	G	C	A	C	U	
	A	G	U	U	U	C	A	C	U	
A	16	0	4	2	4	4	4	0	0	
C	0	0	4	4	4	4	4	16	0	
G	0	16	4	2	4	4	4	0	0	
U	0	0	4	8	4	4	4	0	16	

MI:	1	2	3	4	5	6	7	8	9
9	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0		
7	0	0	**2**	**0.30**	**0**	**1**			
6	0	0	**1**	**0.55**	**1**				
5	0	0	**0**	**0.42**					
4	0	0	**0.30**						
3	0	0							
2	0								
1									

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

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

Cols 7->6: unconserved, but each letter in 7 has only 2 possible mates in 6. M.I. = 1 bit.

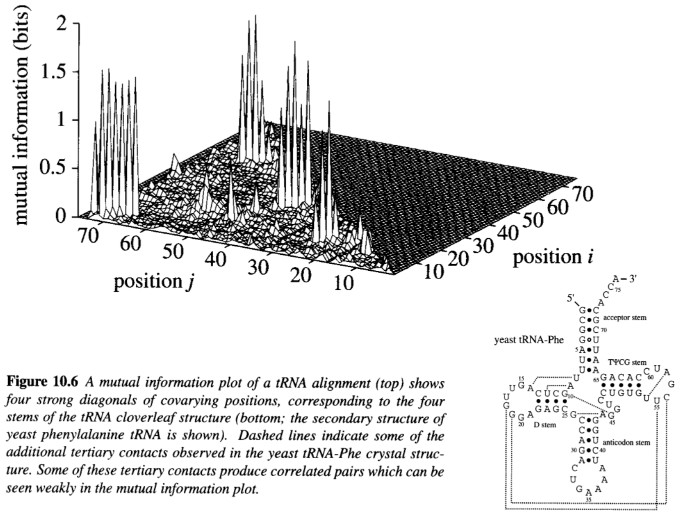
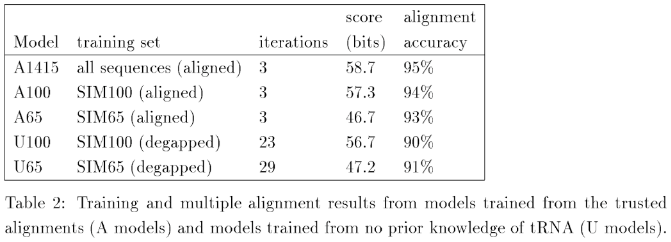


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary contacts observed in the yeast tRNA-Phe crystal structure. Some of these tertiary contacts produce correlated pairs which can be seen weakly in the mutual information plot.*

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

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

Table 1: Statistics of the training and test sets of 100 tRNA sequences each. The average identity in an alignment is the average pairwise identity of all aligned symbol pairs, with gap/symbol alignments counted as mismatches. Primary sequence information content is calculated according to [48]. Calculating pairwise mutual information content is an NP-complete problem of finding an optimum partition of columns into pairs. A lower bound is calculated by using the model construction procedure to find an optimal partition subject to a non-pseudoknotting restriction. An upper bound is calculated as sum of the single best pairwise covariation for each position, divided by two; this includes all pairwise tertiary interactions but overcounts because it does not guarantee a disjoint set of pairs. For the meaning of multiple alignment accuracy of ClustalV, see the text.



Model	training set	iterations	score (bits)	alignment accuracy
A1415	all sequences (aligned)	3	58.7	95%
A100	SIM100 (aligned)	3	57.3	94%
A65	SIM65 (aligned)	3	46.7	93%
U100	SIM100 (degapped)	23	56.7	90%
U65	SIM65 (degapped)	29	47.2	91%

Table 2: Training and multiple alignment results from models trained from the trusted alignments (A models) and models trained from no prior knowledge of tRNA (U models).

Rfam – an RNA family DB

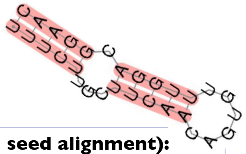
Griffiths-Jones, et al., NAR '03, '05

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

Rapidly growing:

- Rel 1.0, 1/03: 25 families, 55k instances
- Rel 7.0, 3/05: 503 families, >300k instances

Rfam



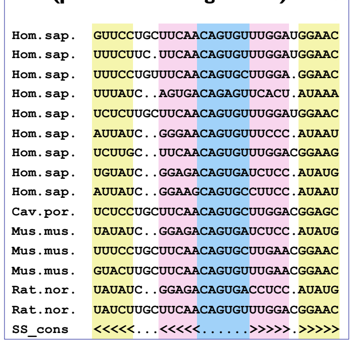
Input (hand-curated):

- MSA "seed alignment"
- SS_cons
- Score Thresh T
- Window Len W

Output:

- CM
- scan results & "full alignment"

IRE (partial seed alignment):



```
Hom.sap.   GUUCCUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCUUC.UUCAACAGUGUUUGGAUGGAAC
Hom.sap.   UUUCCUGUUUCAACAGUGCUUGGA.GGAAC
Hom.sap.   UUUAUC..AGUGACAGAGUUCACU.AUAAA
Hom.sap.   UCUCUUGCUUCAACAGUGUUUGGAUGGAAC
Hom.sap.   AUUAUC..GGGAACAGUGUUUCCC.AUAAU
Hom.sap.   UCUUGC..UUCAACAGUGUUUGGACGGAAG
Hom.sap.   UGUAUC..GGAGACAGUGAUCUCC.AUAUG
Hom.sap.   AUUAUC..GGAAGCAGUGCCUUCC.AUAAU
Cav.por.   UCUCCUGCUUCAACAGUGCUUGGACGGAGC
Mus.mus.   UAUAUC..GGAGACAGUGAUCUCC.AUAUG
Mus.mus.   UUUCCUGCUUCAACAGUGCUUGAACGGAAC
Mus.mus.   GUACUUGCUUCAACAGUGUUUGAACGGAAC
Rat.nor.   UAUAUC..GGAGACAGUGACCUCC.AUAUG
Rat.nor.   UAUCUUGCUUCAACAGUGUUUGGACGGAAC
SS_cons    <<<<<...<<<<<......>>>>>.>>>>>
```

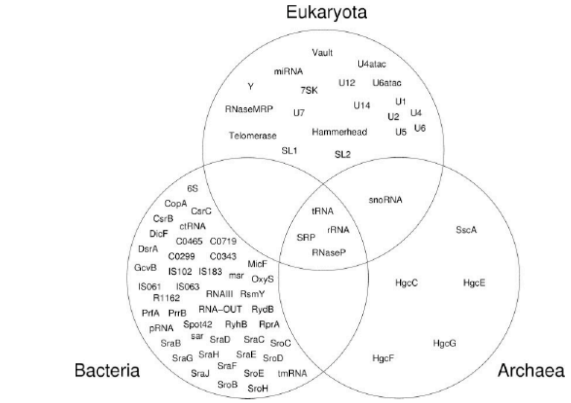


Figure 2. Taxonomic distribution of Rfam family members in the three kingdoms of life.

Task 3: Faster Search

Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg

& W.L. Ruzzo

Recomb ‘04, ISMB ‘04, Bioinfo ‘06

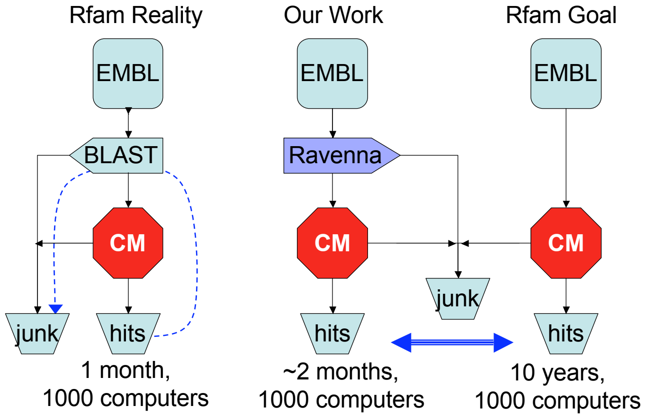
RaveNnA: Genome Scale RNA Search

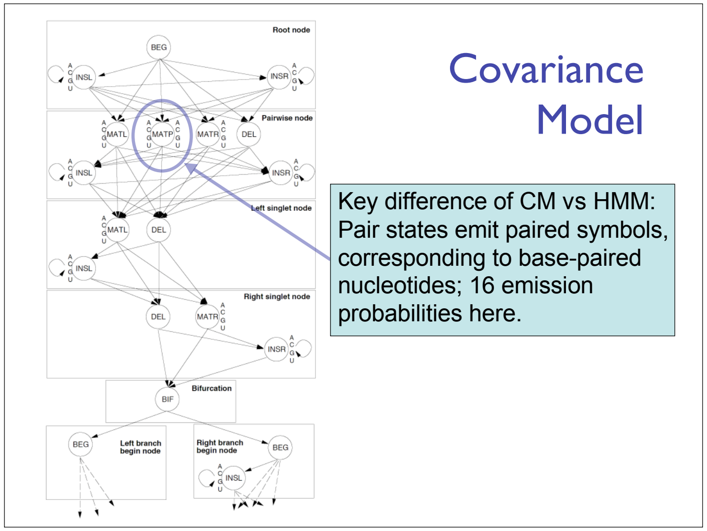
Typically 100x speedup over raw CM, w/ no loss in accuracy:

- drop structure from CM to create a (faster) HMM
- use that to pre-filter sequence;
- discard parts where, provably, CM will score < threshold;
- actually run CM on the rest (the promising parts)
- assignment of HMM transition/emission scores is key
  - (large convex optimization problem)

Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

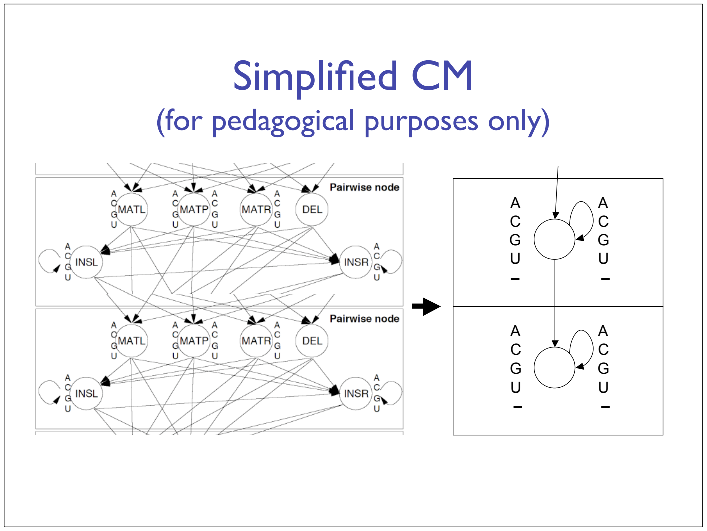
CM’s are good, but slow





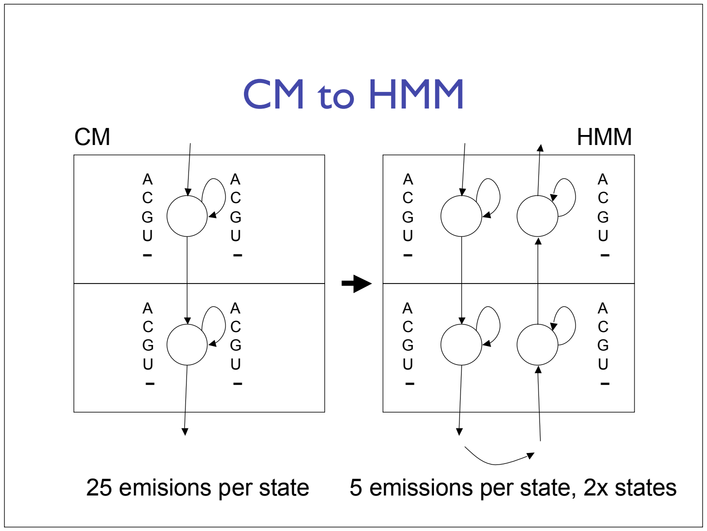
Covariance Model

Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.



Simplified CM

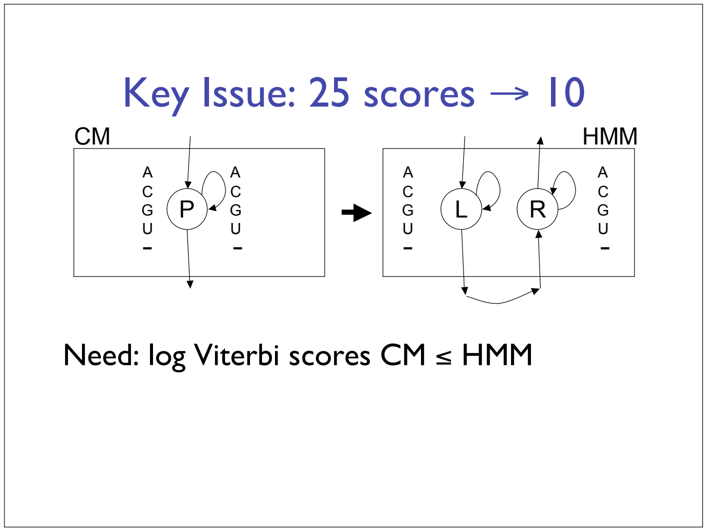
(for pedagogical purposes only)



CM to HMM

25 emisions per state

5 emissions per state, 2x states



Key Issue: 25 scores → 10

Need: log Viterbi scores CM ≤ HMM

Viterbi/Forward Scoring

Path π defines transitions/emissions

Score(π) = product of "probabilities" on π

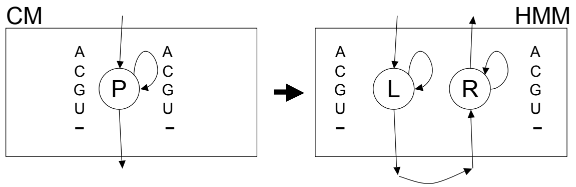
NB: ok if "probs" aren't, e.g. $\Sigma \neq 1$
(e.g. in CM, emissions are odds ratios vs 0th-order background)

For any nucleotide sequence x:

Viterbi-score(x) = max{ score(π) | π emits x }

Forward-score(x) = Σ{ score(π) | π emits x }

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

$$\begin{array}{lll}
P_{AA} \le L_A + R_A & P_{CA} \le L_C + R_A & \dots \\
P_{AC} \le L_A + R_C & P_{CC} \le L_C + R_C & \dots \\
P_{AG} \le L_A + R_G & P_{CG} \le L_C + R_G & \dots \\
P_{AU} \le L_A + R_U & P_{CU} \le L_C + R_U & \dots \\
P_{A-} \le L_A + R_- & P_{C-} \le L_C + R_- & \dots
\end{array}$$

NB: HMM not a prob. model

Rigorous Filtering

$$\begin{array}{l}
P_{AA} \le L_A + R_A \\
P_{AC} \le L_A + R_C \\
P_{AG} \le L_A + R_G \\
P_{AU} \le L_A + R_U \\
P_{A-} \le L_A + R_- \\
\dots
\end{array}$$

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

$P_{UA} = 1 \le L_U + R_A$

$P_{UG} = 4 \le L_U + R_G$

Option 1:
$L_U = R_A = R_G = 2$

Option 2:
$L_U = 0, R_A = 1, R_G = 4$

Assuming ACGU ≈ 25%

Opt 1:
$L_U + (R_A + R_G)/2 = 4$

Opt 2:
$L_U + (R_A + R_G)/2 = 2.5$

Optimizing filtering

For any nucleotide sequence x:

- Viterbi-score(x) = max{ score(π) | π emits x }
- Forward-score(x) = Σ{ score(π) | π emits x }

Expected Forward Score

- E(L_i, R_i) = $\Sigma_{\text{all sequences } x}$ Forward-score(x)*Pr(x)
- NB: E is a function of L_i, R_i only

Under 0th-order background model

Optimization:

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

- This is heuristic (“forward↓ ⇒ Viterbi↓ ⇒ filter↓”)
- But still rigorous because “subject to score Lin.Ineq.s”

Calculating E(L_i, R_i)

E(L_i, R_i) = Σ_x Forward-score(x)*Pr(x)

Forward-like: for every state, calculate expected score for all paths ending there; easily calculated from expected scores of predecessors & transition/emission probabilities/scores

Minimizing E(L_i, R_i)

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

Forward:

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

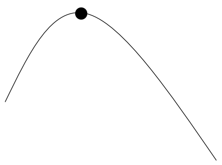
Viterbi:

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

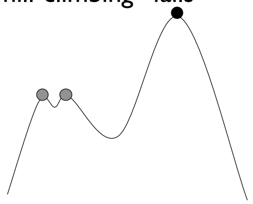
$$\frac{\partial E(L_1, L_2, \ldots)}{\partial L_i}$$

“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)

Filtering fraction	# families (compact)	# families (expanded)
< 10^{-4}	105	110
10^{-4} - 10^{-2}	8	17
.01 - .10	11	3
.10 - .25	2	2
.25 - .99	6	4
.99 - 1.0	7	3

~100x speedup

Results: New ncRNA's?

Name	# found BLAST + CM	# found rigorous filter + CM	# new
Pyrococcus snoRNA	57	180	123
Iron response element	201	322	121
Histone 3' element	1004	1106	102
Purine riboswitch	69	123	54
Retron msr	11	59	48
Hammerhead I	167	193	26
Hammerhead III	251	264	13
U4 snRNA	283	290	7
S-box	128	131	3
U6 snRNA	1462	1464	2
U5 snRNA	199	200	1
U7 snRNA	312	313	1

Task 4: Motif Discovery

RNA Motif Discovery

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

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

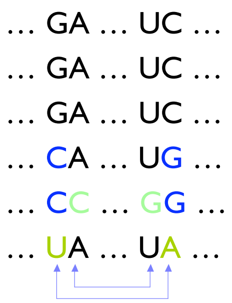
Approaches

Align sequences, then look for common structure

Predict structures, then try to align them

Do both together

"Obvious" Approach I: Align First, Predict from Multiple Sequence Alignment



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

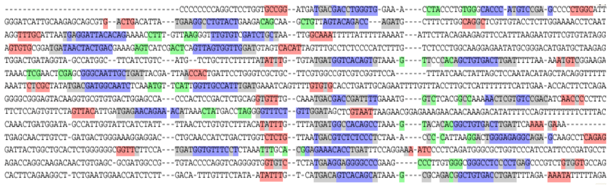
Compensatory mutations reveal structure, (core of "comparative sequence analysis") *but* usual alignment algorithms penalize them (twice)

Pitfall for sequence-alignment-first approach

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions



```
---------------------------------------CCCCCCCCAGGCTCCTGGTGCCGG--ATGATGACGACCTGGGTG-GAA-A----CCTACCCTGTGGGCACCC-ATGTCCGA-GCCCCCTGGCATT
GGGATCATTGCAAGAGCAGCGTG--ACTGACATTA---TGAAGGCCTGTACTGAAGACAGCAA--GCTGTTAGTACAGACC---AGATG----CTTTCTTGGCAGGCTCGTTGTACCTCTTGGAAAACCTCAAT
AGGTTTGCATTAATGAGGATTACACAGAAAACCTTT-GTTAAGGGTTTGTGTCGATCTGCTAA--TTGGCAAATTTTTATTTTTAAAAT---ATTCTTACAGAAGAGTTCCATTTAAGAATGTTCGTGTATAGG
AGTGTGCGGATGATAACTACTGACGAAAGAGTCATCGACTCAGTTAGTGGTTGGATGTAGTCACATTAGTTTGCCTCTCCCCATCTTTG----TCTCCCTGGCAAGGAGAATATGCGGGACATGATGCTAAGAG
TGGACTGATAGGTA-GCCATGGC--TTCATCTGTC---ATG--TCTGCTTCTTTTTATATTTG--TGTATGATGGTCACAGTGTAAA-G----TTCCCACAGCTGTGACTTGATTTTTAA-AAATGTCGGAAGA
TAAACTCGAACTCGAGCGGGCAATTGCTGATTACGA-TTAACCACTGATTCCTGGGTCGCTGC--TTCGTGGCCGTCGTCGGTTCCA-------TTTATCAACTATTAGCTCCAATACATAGCTACAGGTTTTT
AAATTCTCGCTATATGACGATGGCAATCTCAAATGT-TCATTGGTTGCCATTTGATGAAATCAGTTTTGTGTGCACCTGATTGCAGAATTTTGTTTACCTTGCTCATTTTTTTCATTGAA-ACCACTTCTCAGA
GGGGCGGGAGTACAAGGTGCGTGTGACTGGAGCCA---CCCACTCCGACTCTGCAGGTGTTTG--CAAATGACGACCGATTTTGAAATG----GTCTCACGGCCAAAAACTCGTGTCCGACATCAACCCCCTTC
TTCTCCAGTGTTCTAGTTACATTGATGAGAACAGAA-ACATAAACTATGACCTAGGGTTTCT--GTTGGATAGCTCGTAATTAAGAACGGAGAAAGAACAACAAAGACATATTTTCCAGTTTTTTTCTTTAC
CAAACTGATGGATA-GCCATTGGTATTCATCTATT---TTAACTCTGTGTCTTTACATATTTG--TTTATGATGGCCACAGCCTAAA-G----TACACACGGCTGTGACTTGATTCAAAA-GAAA--------
TGAGCAACTTGTCT-GATGACTGGGAAAGGAGGAC---CTGCAACCATCTGACTTGGTCTCTG--TTAATGACGTCTCTCCCTCTAA-A----CCC-CATTAAGGACTGGGAGAGGCAGA-GCAAGCCTCAGAG
GATTACTGGCTGCACTCTGGGGGGCGGTTCTTCCA---TGATGGTGTTTCCTCTAAATTTGCA--CGGAGAAACACCTGATTTCCAGGAAA-ATCCCCTCAGATGGGCGCTGGTCCCATCCATTCCCGATGCCT
AGACCAGGCAAGACAACTGTGAGC-GCGATGGCCG---TGTACCCCAGGTCAGGGGTGGTGTC--TCTATGAAGGAGGGGCCCGAAG-----CCCTTGTGGGCGGGCCTCCCCTGAGCCCGTCTGTGGTGCCAG
CACTTCAGAAGGCT-TCTGAATGGAACCATCTCTT---GACA-TTTGTTTCTATA-ATATTTG--T-CATGACAGTCACAGCATAAA-G----CGCAGACGGCTGTGACCTGATTTTAGA-AAATATTTTTAGA
```

same-colored boxes *should* be aligned

Approaches

Align sequences, then look for common structure

Predict structures, then try to align them

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

Do both together

- Sankoff – good but slow
- Various heuristics – still tend to be slow

Our Approach: CMfinder

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

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

Alignment → CM → Alignment

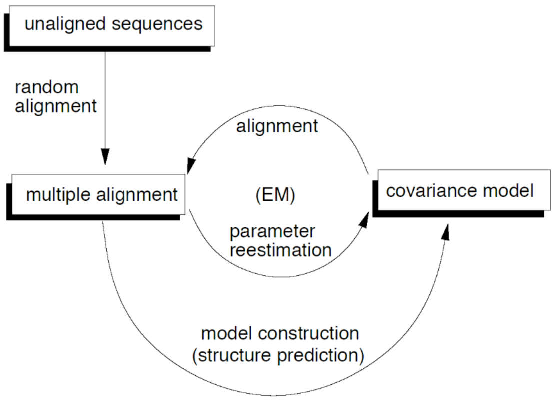
Similar to HMM, but slower

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

alignment

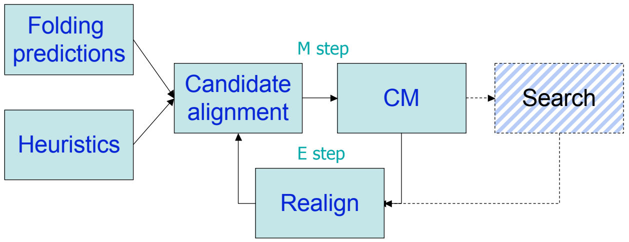
(EM)

parameter reestimation

covariance model

model construction (structure prediction)

CMfinder Outline



Folding predictions

Heuristics

Candidate alignment

M step

CM

Search

E step

Realign

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

Structure Inference

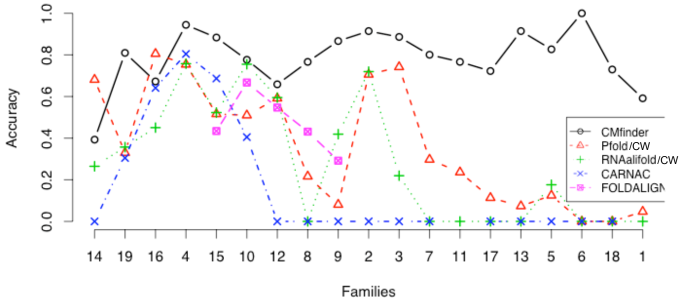
Part of M-step is to pick a structure that maximizes data likelihood

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

CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Summary of Rfam test families and results

ID	Family	Rfam ID	#seqs	%id	length	#hp	CMfinder	CW/Pfold	CW/RNAalifold	Carnac	Foldalign	ComRNA
1	Cobalamin	RF00174	71	49	216	4	**0.59**	0.05	0	X	-	0
2	ctRNA_pGA1	RF00236	17	74	83	2	**0.91**	0.70	0.72	0	0.86	0
3	Entero_CRE	RF00048	56	81	61	1	**0.89**	0.74	0.22	0	-	0
4	Entero_OriR	RF00041	35	77	73	2	**0.94**	0.75	0.76	0.80	0.52	0.52
5	glmS	RF00234	14	58	188	4	**0.83**	0.12	0.18	0	-	0.13
6	Histone3	RF00032	63	77	26	1	**1**	0	0	0	-	0
7	Intron_gpII	RF00029	75	55	92	2	**0.80**	0.30	0	0	-	0
8	IRE	RF00037	30	68	30	1	**0.77**	0.22	0	0	0.38	0
9	let-7	RF00027	9	69	84	1	**0.87**	0.08	0.42	0	0.71	0.78
10	lin-4	RF00052	9	69	72	1	**0.78**	0.51	0.75	0.41	0.65	0.24
11	Lysine	RF00168	48	48	183	4	**0.77**	0.24	0	X	-	0
12	mir-10	RF00104	11	66	75	1	**0.66**	0.59	0.60	0	0.48	0.33
13	Purine	RF00167	29	55	103	2	**0.91**	0.07	0	0	-	0.27
14	RFN	RF00050	47	66	139	4	0.39	**0.68**	0.26	0	-	0
15	Rhino_CRE	RF00220	12	71	86	1	**0.88**	0.52	0.52	0.69	0.41	0.61
16	s2m	RF00164	23	80	43	1	0.67	**0.80**	0.45	0.64	0.63	0.29
17	S_box	RF00162	64	66	112	3	**0.72**	0.11	0	0	-	0
18	SECIS	RF00031	43	43	68	1	**0.73**	0	0	0	-	0
19	Tymo_tRNA-like	RF00233	22	72	86	4	**0.81**	0.33	0.36	0.30	0.80	0.48
				Average Accuracy:			**0.79**	0.36	0.28	0.17	0.60	0.19
				Average Specificity:			0.81	0.42	0.57	**0.83**	0.60	0.65
				Average Sensitivity:			**0.77**	0.36	0.23	0.13	0.61	0.17

Task 5: Application

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

Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo.
PLoS Computational Biology. 3(7): e126, July 6, 2007.

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

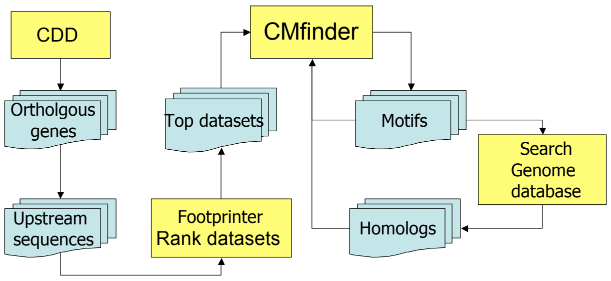
Goal:

- 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



CDD

Ortholgous genes

Upstream sequences

Footprinter Rank datasets

Top datasets

CMfinder

Motifs

Search Genome database

Homologs

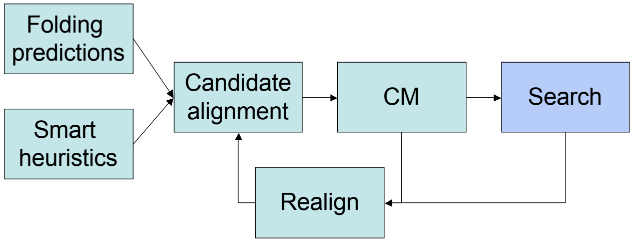
Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo. A Computational Pipeline for High Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. PLoS Computational Biology. 3(7): e126, July 6, 2007.

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



Folding predictions

Smart heuristics

Candidate alignment

CM

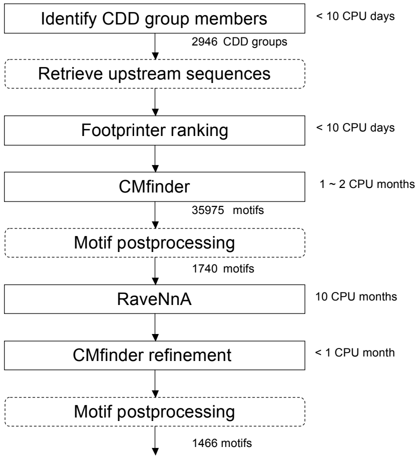
Search

Realign

Results

Analyzed most sequenced bacteria (~2005)

- bacillus/clostridia
- gamma proteobacteria
- cyanobacteria
- actinobacteria
- firmicutes



Identify CDD group members

< 10 CPU days

2946 CDD groups

Retrieve upstream sequences

Footprinter ranking

< 10 CPU days

CMfinder

1 ~ 2 CPU months

35975 motifs

Motif postprocessing

1740 motifs

RaveNnA

10 CPU months

CMfinder refinement

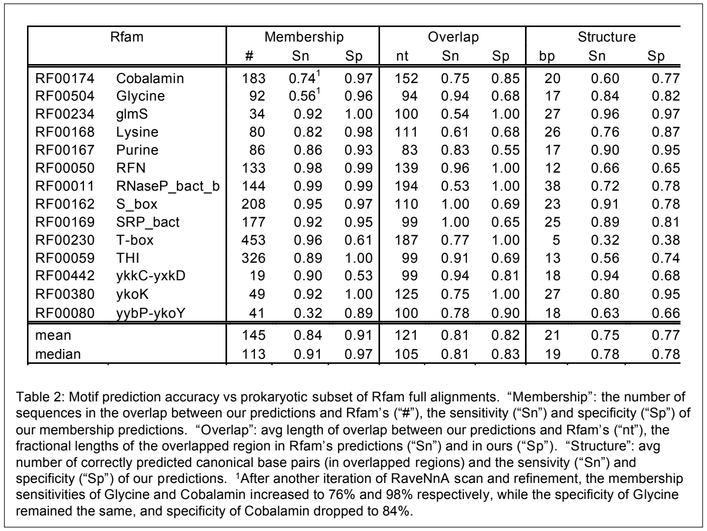
< 1 CPU month

Motif postprocessing

1466 motifs

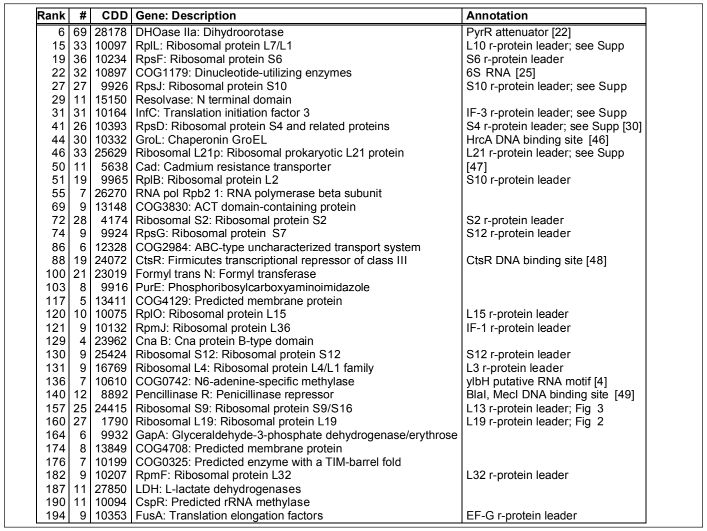
Rank			Score	#		CDD			Rfam
RAV	CMF	FP		RAV	CMF	ID	Gene	Descriptio n	
0	43	107	3400	367	11	9904	IlvB	Thiamine pyrophosphate-requiring enzymes	RF00230 T-box
1	10	344	3115	96	22	13174	COG3859	Predicted membrane protein	RF00059 THI
2	77	1284	2376	112	6	11125	MetH	Methionine synthase I specific DNA methylase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

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"

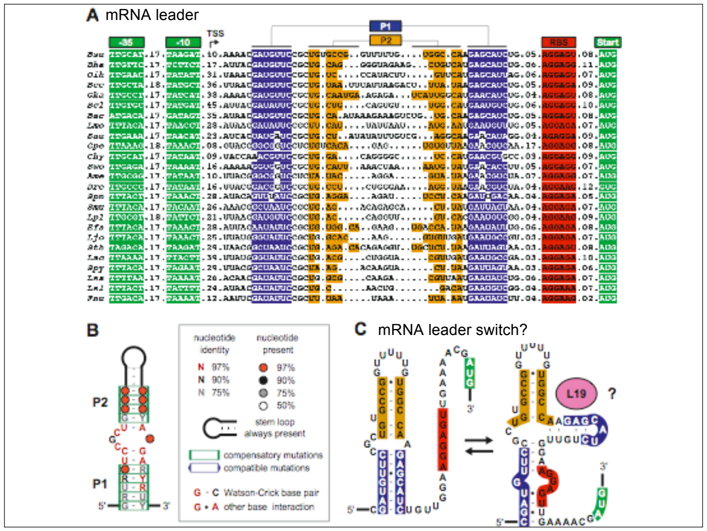


Rfam		Membership			Overlap			Structure		
		#	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174	Cobalamin	183	0.74[1]	0.97	152	0.75	0.85	20	0.60	0.77
RF00504	Glycine	92	0.56[1]	0.96	94	0.94	0.68	17	0.84	0.82
RF00234	glmS	34	0.92	1.00	100	0.54	1.00	27	0.96	0.97
RF00168	Lysine	80	0.82	0.98	111	0.61	0.68	26	0.76	0.87
RF00167	Purine	86	0.86	0.93	83	0.83	0.55	17	0.90	0.95
RF00050	RFN	133	0.98	0.99	139	0.96	1.00	12	0.66	0.65
RF00011	RNaseP_bact_b	144	0.99	0.99	194	0.53	1.00	38	0.72	0.78
RF00162	S_box	208	0.95	0.97	110	1.00	0.69	23	0.91	0.78
RF00169	SRP_bact	177	0.92	0.95	99	1.00	0.65	25	0.89	0.81
RF00230	T-box	453	0.96	0.61	187	0.77	1.00	5	0.32	0.38
RF00059	THI	326	0.89	1.00	99	0.91	0.69	13	0.56	0.74
RF00442	ykkC-yxkD	19	0.90	0.53	99	0.94	0.81	18	0.94	0.68
RF00380	ykoK	49	0.92	1.00	125	0.75	1.00	27	0.80	0.95
RF00080	yybP-ykoY	41	0.32	0.89	100	0.78	0.90	18	0.63	0.66
mean		145	0.84	0.91	121	0.81	0.82	21	0.75	0.77
median		113	0.91	0.97	105	0.81	0.83	19	0.78	0.78

Table 2: Motif prediction accuracy vs prokaryotic subset of Rfam full alignments. "Membership": the number of sequences in the overlap between our predictions and Rfam's ("#"), the sensitivity ("Sn") and specificity ("Sp") of our membership predictions. "Overlap": avg length of overlap between our predictions and Rfam's ("nt"), the fractional lengths of the overlapped region in Rfam's predictions ("Sn") and in ours ("Sp"). "Structure": avg number of correctly predicted canonical base pairs (in overlapped regions) and the sensivity ("Sn") and specificity ("Sp") of our predictions. [1]After another iteration of RaveNnA scan and refinement, the membership sensitivities of Glycine and Cobalamin increased to 76% and 98% respectively, while the specificity of Glycine remained the same, and specificity of Cobalamin dropped to 84%.



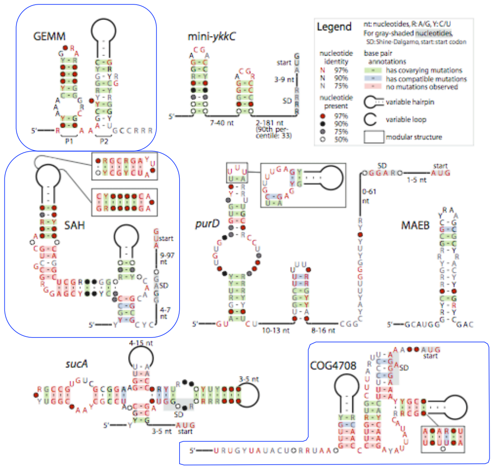
Rank	#	CDD	Gene: Description	Annotation
6	69	28178	DHOase IIa: Dihydroorotase	PyrR attenuator [22]
15	33	10097	RplL: Ribosomal protein L7/L1	L10 r-protein leader; see Supp
19	36	10234	RpsF: Ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: Dinucleotide-utilizing enzymes	6S RNA [25]
27	27	9926	RpsJ: Ribosomal protein S10	S10 r-protein leader; see Supp
29	11	15150	Resolvase: N terminal domain	
31	31	10164	InfC: Translation initiation factor 3	IF-3 r-protein leader; see Supp
41	26	10393	RpsD: Ribosomal protein S4 and related proteins	S4 r-protein leader; see Supp [30]
44	30	10332	GroL: Chaperonin GroEL	HrcA DNA binding site [46]
46	33	25629	Ribosomal L21p: Ribosomal prokaryotic L21 protein	L21 r-protein leader; see Supp
50	11	5638	Cad: Cadmium resistance transporter	[47]
51	19	9965	RplB: Ribosomal protein L2	S10 r-protein leader
55	7	26270	RNA pol Rpb2 1: RNA polymerase beta subunit	
69	9	13148	COG3830: ACT domain-containing protein	
72	28	4174	Ribosomal S2: Ribosomal protein S2	S2 r-protein leader
74	9	9924	RpsG: Ribosomal protein S7	S12 r-protein leader
86	6	12328	COG2984: ABC-type uncharacterized transport system	
88	19	24072	CtsR: Firmicutes transcriptional repressor of class III	CtsR DNA binding site [48]
100	21	23019	Formyl trans N: Formyl transferase	
103	8	9916	PurE: Phosphoribosylcarboxyaminoimidazole	
117	5	13411	COG4129: Predicted membrane protein	
120	10	10075	RplO: Ribosomal protein L15	L15 r-protein leader
121	9	10132	RpmJ: Ribosomal protein L36	IF-1 r-protein leader
129	4	23962	Cna B: Cna protein B-type domain	
130	9	25424	Ribosomal S12: Ribosomal protein S12	S12 r-protein leader
131	9	16769	Ribosomal L4: Ribosomal protein L4/L1 family	L3 r-protein leader
136	7	10610	COG0742: N6-adenine-specific methylase	ylbH putative RNA motif [4]
140	12	8892	Pencillinase R: Penicillinase repressor	BlaI, MecI DNA binding site [49]
157	25	24415	Ribosomal S9: Ribosomal protein S9/S16	L13 r-protein leader; Fig 3
160	27	1790	Ribosomal L19: Ribosomal protein L19	L19 r-protein leader; Fig 2
164	6	9932	GapA: Glyceraldehyde-3-phosphate dehydrogenase/erythrose	
174	8	13849	COG4708: Predicted membrane protein	
176	7	10199	COG0325: Predicted enzyme with a TIM-barrel fold	
182	9	10207	RpmF: Ribosomal protein L32	L32 r-protein leader
187	11	27850	LDH: L-lactate dehydrogenases	
190	11	10094	CspR: Predicted rRNA methylase	
194	9	10353	FusA: Translation elongation factors	EF-G r-protein leader



Task 5: Application

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

Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker.
Nucl. Acids Res., July 2007 35: 4809-4819.



boxed = confirmed riboswitch (+2 more)

ncRNA discovery in Vertebrates

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

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

Genome Research, Jan 2008

ncRNA discovery in Vertebrates

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

- Evofold (Pedersen *et al.* 2006)
- RNAz (Washietl *et al.* 2005)

We explore regions with weak sequence conservation

Approach

Extract ENCODE Multiz alignments

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

Apply CMfinder to both strands.

10,106 predictions, 6,587 clusters.

- False positive rate: 50% based on a heuristic ranking function.

Overlap w/ Indel Purified Segments

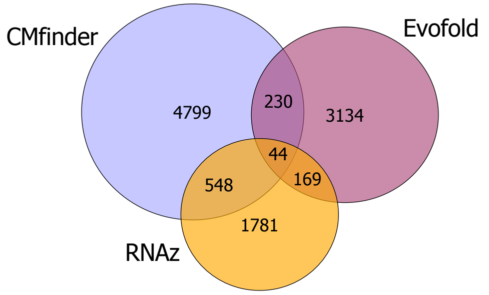
IPS presumed to signal purifying selection

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

Strong P-value for their overlap w/ IPS

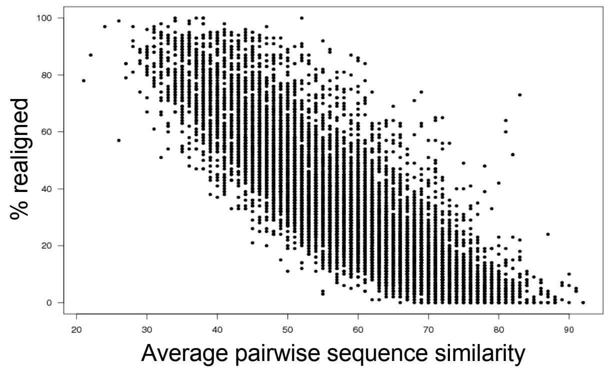
G+C	data	P	N	Expected	Observed	P-value	%
0-35	igs	0.062	380	23	24.5	0.430	5.8%
35-40	igs	0.082	742	61	70.5	0.103	11.3%
40-45	igs	0.082	1216	99	129.5	0.00079	18.5%
45-50	igs	0.079	1377	109	162.5	5.16E-08	20.9%
50-100	igs	0.070	2866	200	358.5	2.70E-31	43.5%
all	igs	0.075	6581	491	747.5	1.54E-33	100.0%

Comparison with Evofold, RNAz

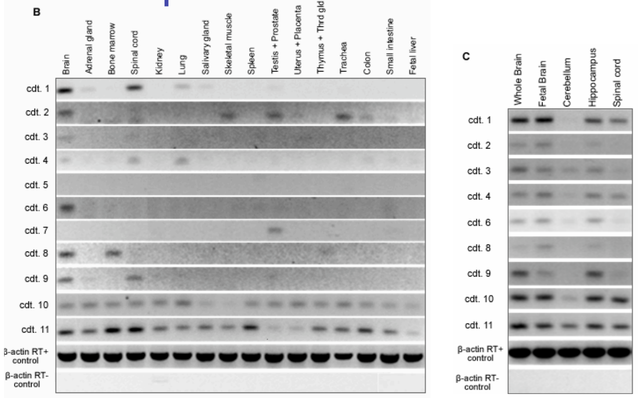


Small overlap (w/ highly significant p-values) emphasizes complementarity

Realignment



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.

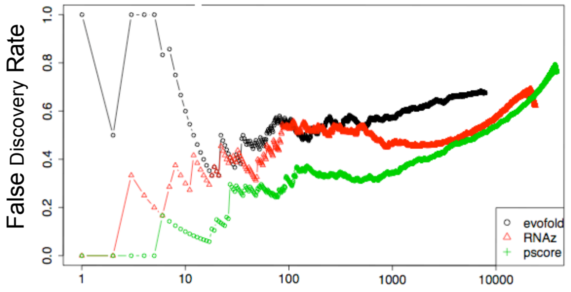
Better model for gaps

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

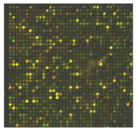
Covariance Models - powerful but expensive tool for ncRNA motif representation, search, discovery

Rigorous/Heuristic filtering - typically 100x speedup in search with no/little loss in accuracy

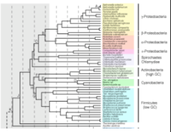
CMfinder - good CM-based motif discovery in unaligned sequences

- Pipeline integrating comp and bio for ribowitch discovery
- Potentially many ncRNAs with weak sequence conservation in vertebrates.

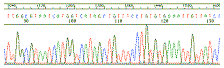
Course Wrap Up



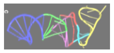
“High-Throughput BioTech”



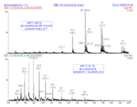
Sensors
- DNA sequencing
- Microarrays/Gene expression
- Mass Spectrometry/Proteomics
- Protein/protein & DNA/protein interaction



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 trafficing
- ncRNA discovery
- chromatin dynamics
- …

Exciting Times

Lots to do

Various skills needed

I hope I've given you a taste of it

Thanks!