Day 3
5 slide synopsis of last lecture

Covariance Models (CMs) represent conserved RNA sequence/structure motifs

They allow accurate search

But

a) search is slow

b) model construction is laborious

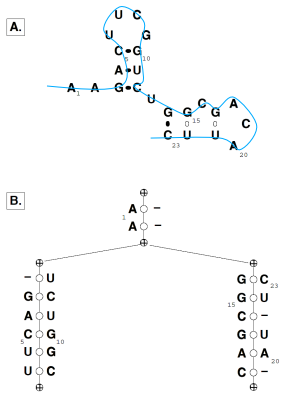
CM Structure

A: sequence + structure

B: the CM "guide tree"

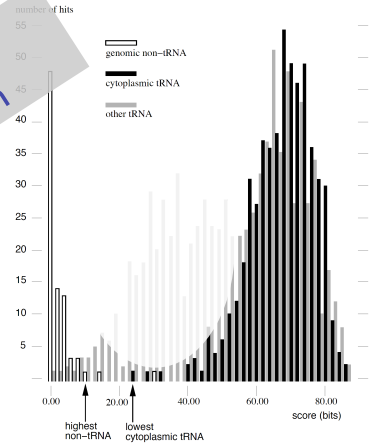
C: probabilities of letters/ pairs & of indels

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



Accurate Search

Example: searching for tRNAs



But Slow

CM Viterbi Alignment
(the "inside" algorithm)

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

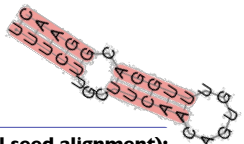
$$S_{ij}^{y} = \begin{cases} \max_z [S_{i+1,j-1}^{z} + \log T_{yz} + \log E_{x_i x_j}^{y}] & \text{match pair} \\ \max_z [S_{i+1,j}^{z} + \log T_{yz} + \log E_{x_i}^{y}] & \text{match/insert left} \\ \max_z [S_{i,j-1}^{z} + \log T_{yz} + \log E_{x_j}^{y}] & \text{match/insert right} \\ \max_z [S_{i,j}^{z}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

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

compare: $O(qn)$ for profile HMM

Hand-made

Example Rfam Family



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   <<<<<...<<<<<.....>>>>>.>>>>
```

Today's Goals

Faster Search

- Infernal & RaveNnA

Automated Model-building

- CMfinder

Homology search

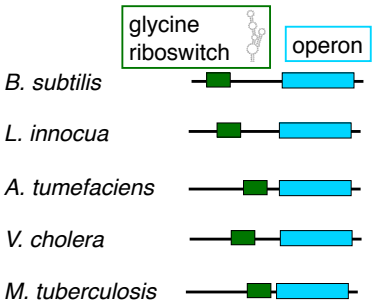
Sequence-based
- Smith-Waterman
- FASTA
- BLAST

Sharp decline in sensitivity at ~60-70% identity

So, use structure, too

Impact of RNA homology search

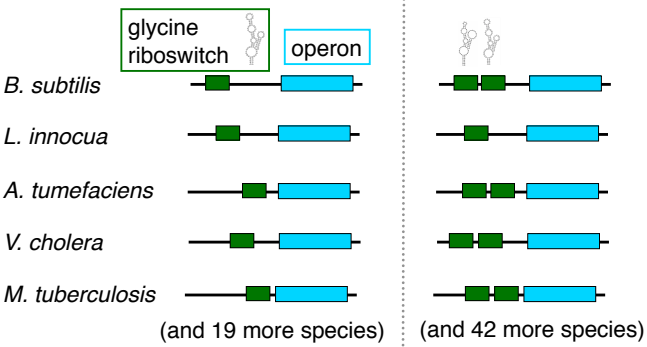
(Barrick, *et al.*, 2004)



Impact of RNA homology search

(Barrick, *et al.*, 2004)

(Mandal, *et al.*, 2004)



BLAST-based

CM-based

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

Zasha Weinberg

& W.L. Ruzzo

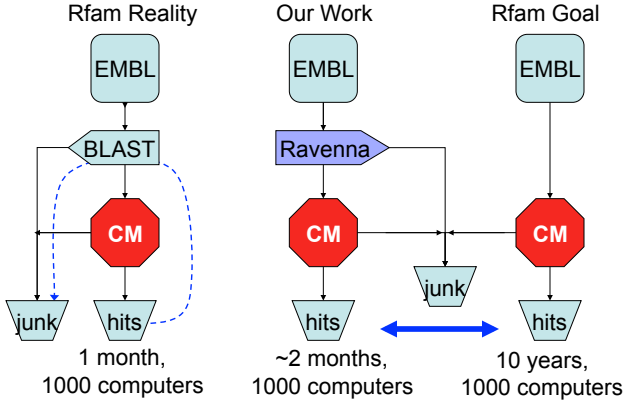
Recomb '04, ISMB '04, Bioinfo '06

RaveNnA: Genome Scale RNA Search

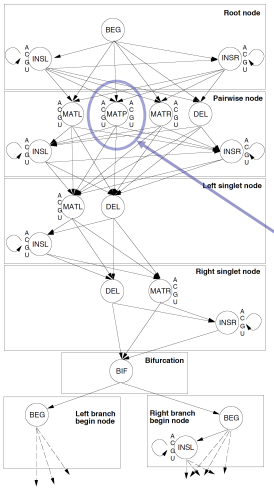
Typically 100x speedup over raw CM, w/ no loss in accuracy:
- Drop structure from CM to create a (faster) HMM
- Use that to pre-filter sequence;
- Discard parts where, provably, CM score < threshold;
- Actually run CM on the rest (the promising parts)
- Assignment of HMM transition/emission scores is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

Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

CM's are good, but slow

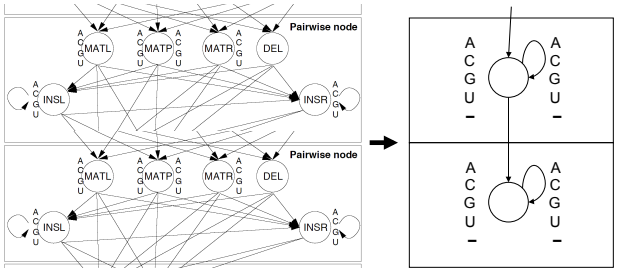


Covariance Model

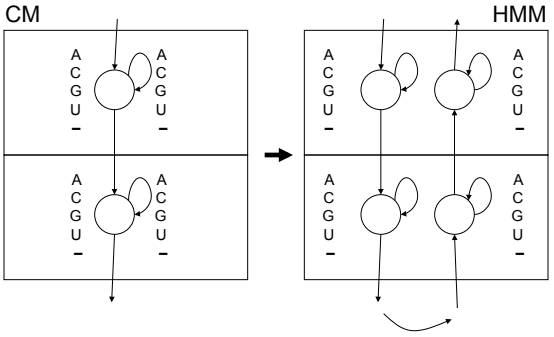


Key difference of CM vs HMM: Pair states emit paired symbols, corresponding to base-paired nucleotides; 16 emission probabilities here.

Oversimplified CM
(for pedagogical purposes only)



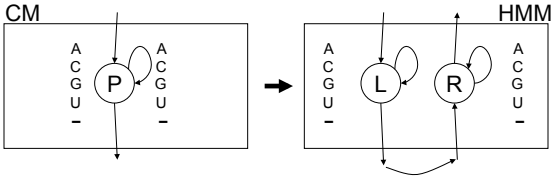
CM to HMM



25 emisions per state

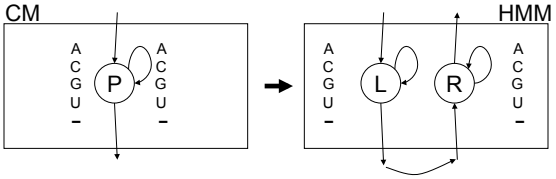
5 emissions per state, 2x states

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

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

NB: HMM not a prob. model

Rigorous Filtering

$P_{AA} \le L_A + R_A$
$P_{AC} \le L_A + R_C$
$P_{AG} \le L_A + R_G$
$P_{AU} \le L_A + R_U$
$P_{A-} \le L_A + R_-$
…

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

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

	Assuming ACGU ≈ 25%
Option 1: $L_U = R_A = R_G = 2$	Opt 1: $L_U + (R_A + R_G)/2 = 4$
Option 2: $L_U = 0, R_A = 1, R_G = 4$	Opt 2: $L_U + (R_A + R_G)/2 = 2.5$

Assignment of scores/ "probabilities"

Convex optimization problem

- Constraints: enforce rigorous property
- Objective function: filter as aggressively as possible

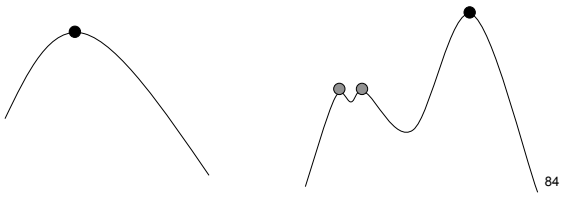
Problem sizes:

- 1000-10000 variables
- 10000-100000 inequality constraints

"Convex" Optimization

Convex:
local max = global max;
simple "hill climbing" works

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

≈ break even

~100x speedup

Averages 283 times faster than CM

Results: new ncRNAs (?)

Name	# Known (BLAST + CM)	# New (rigorous filter + CM)
Pyrococcus snoRNA	57	123
Iron response element	201	121
Histone 3' element	1004	102*
Retron msr	11	48
Hammerhead I	167	26
Hammerhead III	251	13
U6 snRNA	1462	2
U7 snRNA	312	1
cobalamin riboswitch	170	7
13 other families	5-1107	0

Heuristic Filters

Rigorous filters optimized for worst case

Possible to trade improved speed for small loss in sensitivity?

Yes – profile HMMs as before, but optimized for average case

Often 10x faster, modest loss in sensitivity

Software

Ravenna implements both rigorous and heuristic filters

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

- E.g., dynamic "banding" of dynamic programming matrix based on the insight that large deviations from consensus length must have low scores.

CM Search Summary

Still slower than we might like, but dramatic speedup over raw CM is possible with:

- No loss in sensitivity (provably), or
- Even faster with modest (and estimable) loss in sensitivity

RNA Motif Discovery

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.

RNA Motif Discovery

Typical problem: given a 10-20 unaligned sequences of 1-10kb, most of which contain instances of one RNA motif of 100-200bp -- find it.

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

Example: corresponding introns of orthogolous vertebrate genes

Approaches

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

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

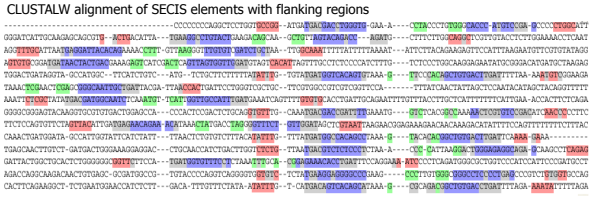
Joint: Do both together

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-first: align sequences, then look for common structure

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

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

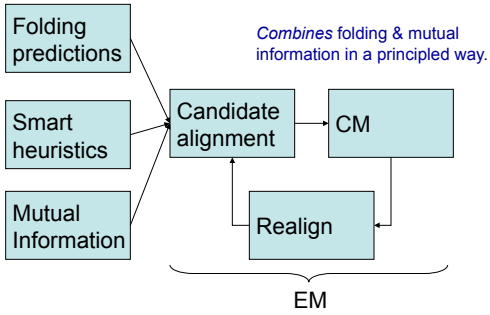
Joint: Do both together

- Sankoff – good but slow
- Heuristic

CMFinder

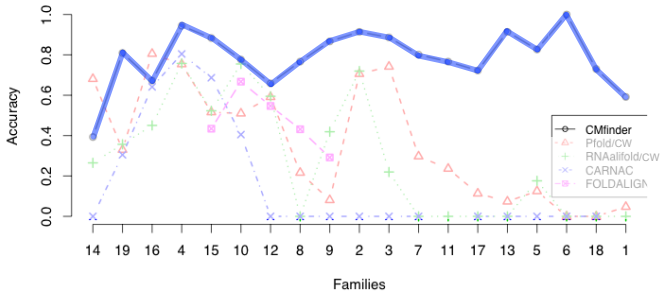
Simultaneous alignment, folding & motif description

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



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 I:

Prokaryotes

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.

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.

Predicting New *cis*-Regulatory RNA Elements

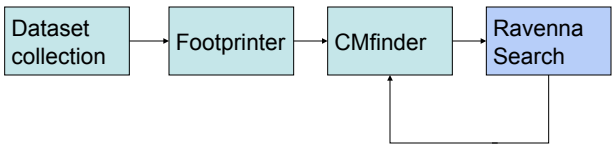
Goal:

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

Difficulties:

- Low sequence similarity: alignment difficult
- Varying flanking sequence
- Motif missing from some input genes

Use the Right Data; Do Genome Scale Search



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)

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…

Approach

Get bacterial genomes

For each gene, get 10-30 close orthologs (CDD)

Find most promising genes, based on conserved sequence motifs (Footprinter)

From those, find structural motifs (CMfinder)

Genome-wide search for more instances (Ravenna)

Expert analyses (Breaker Lab, Yale)

Processing Times

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

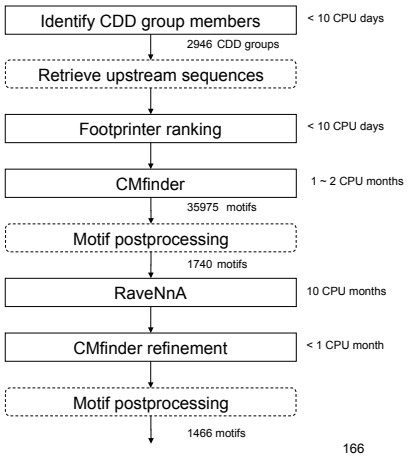


Table 1: Motifs that correspond to Rfam families

Rank RAV	Rank CMF	Rank FP	Score	# RAV	# CMF	CDD ID	CDD Gene	CDD Description	Rfam
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"

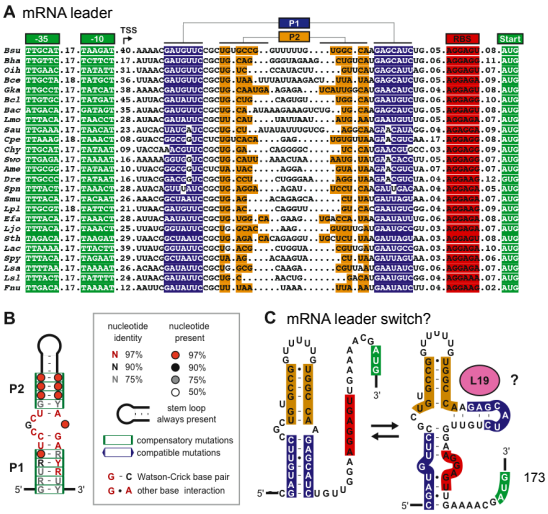
Rfam		Membership			Overlap			Structure		
		#	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174	Cobalamin	183	0.74[1]	0.97	152	0.75	0.85	20	0.60	0.77
RF00504	Glycine	92	0.56[1]	0.96	94	0.94	0.68	17	0.84	0.82
RF00234	glmS	34	0.92	1.00	100	0.54	1.00	27	0.96	0.97
RF00168	Lysine	80	0.82	0.98	111	0.61	0.68	26	0.76	0.87
RF00167	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

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. [1]After 2nd RaveNnA scan, membership Sn of Glycine, Cobalamin increased to 76% and 98% resp., Glycine Sp unchanged, but Cobalamin Sp dropped to 84%.

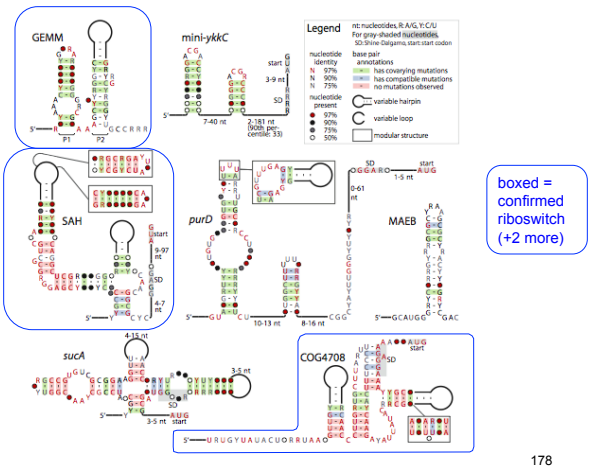
Table 3: High ranking motifs not found in Rfam

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

172



173



178

Weinberg, et al. Nucl. Acids Res., July 2007 35: 4809-4819.

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)

179

ncRNA discovery in Vertebrates

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

199

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)

} Thousands of candidates

We explore regions with weak sequence conservation, where alignments aren't trustworthy

} Thousands more

CMfinder Search in Vertebrates

Extract ENCODE Multiz alignments
- Remove exons, most conserved elements.
- 56017 blocks, 8.7M bps.

} Trust 17-way alignment for orthology, not for detailed alignment

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: many 100's of CPU years. Analysis in progress.)

Overlap with known transcripts

Input regions include only one known ncRNA
- hsa-mir-483, and we found it.

40% intergenetic, 60% overlap with protein coding gene

Sense	Antisense	Both	Intron	5'UTR	3'UTR
1332 (33.8%)	1721 (43.7%)	884 (22.5%)	3274 (83.1%)	551 (14%)	89 (2.3%)

Overlap w/ Indel Purified Segments

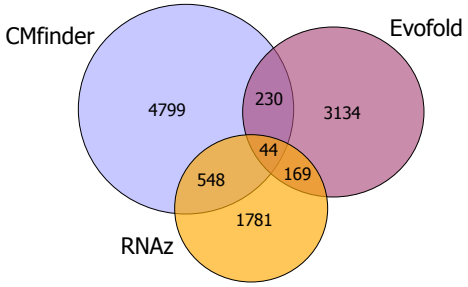
IPS presumed to signal purifying selection

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

Strong association with "Indel purified segments" - I.e., apparently under selection

Strong association with known genes

Alignment Matters

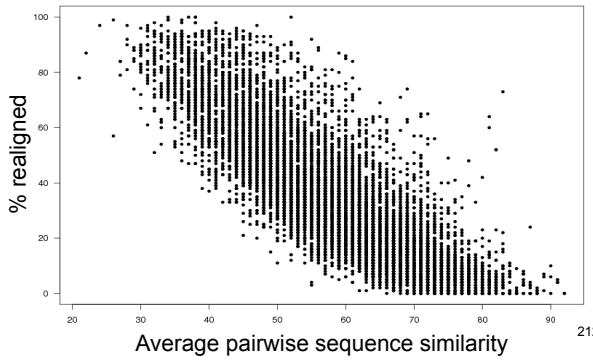
The original MULTIZ alignment without flanking regions. **RNAz Score: 0.132 (no RNA)**

```
Human   GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGT----CTTAACAGTATGACCAAAAACTGAAGTT
Chimp   GGACATTTCAATGCGGGCTC-ATGGGGCTGTGAAGCCAAGAGCT----ATTAACACTATGACCAAGGACTGAAAT
Cow     GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAACCGGGAGCT----CTTAATGCTGTGACCAAAGATTGAAGTT
Dog     GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAACCAAGGGCT----CTTAACTCTGTGACCAAATATTAGAGTT
Rabbit  GATCATTTCAAAGAGGGTTT-GTGGTGCTGTGAAGTCAAGAACT----CTTAACTGTATGCCCAAAGATTAAAGTT
Rhesus  GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGTAGGTCTTAACAGTATAACCAAAGACTGAAGTT
Str     ((((((.....(((((((...(((..........)))..)))).....))).....))))))............
```

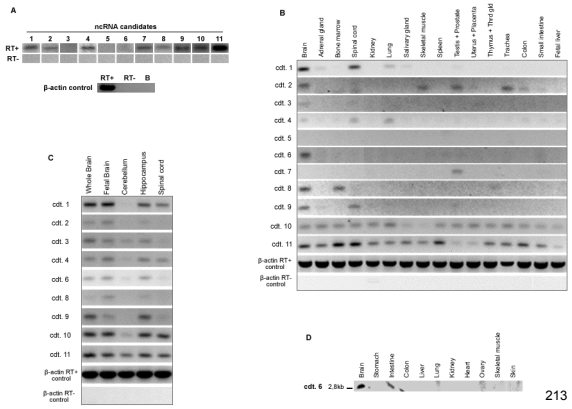
The local CMfinder re-alignment of the MULTIZ block. **RNAz Score: 0.709 (RNA)**

```
Human   GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCA-----AGAGGTCTTAACAGTATGACCAAAAACTGAAG
Chimp   GGACATTTCAATGCGGGCTC-ATGGGGCTGT-GAAGCCA-----AGAGCTATTAACACTATGACCAAGGACTGAAA
Cow     GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAA-CCG-----GGAGCTCTTAATGCTGTGACCAAAGATTGAAG
Dog     GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAA-CCA-----AGGGCTCTTAACTCTGTGACCAAATATTAGAG
Rabbit  GATCATTTCAAAGAGGGTTT-GTGGTGCTGT-GAAGTCA-----AGAACTCTTAACTGTATGCCCAAAGATTAAAG
Rhesus  GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCAAGAGG-TAGGTCTTAACAGTATAACCAAAGACTGAAG
Str     ((((((.....(((((((..(((..........)))).....)))))))).....))))))............
```

Realignment



10 of 11 top (differentially) expressed



Vertebrate Summary

- Lots of *structurally* conserved ncRNA
- Functional significance often unclear
- But high rate of confirmed tissue-specific expression in (small) set of top candidates in humans
- BIG CPU demands…
- Still need for further methods development & application

ncRNA Summary

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

Thanks!