RNA Search and Motif Discovery

Lecture 9
CSEP 590A
Autumn 2008

Outline

Whirlwind tour of ncRNA search & discovery
- RNA motif description (Covariance Model Review)
- Algorithms for searching
  - Rigorous & heuristic filtering
- Motif discovery
- Applications

Motif Description

RNA Motif Models

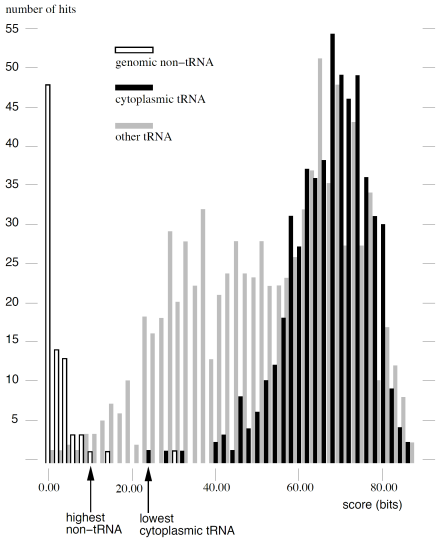
"Covariance Models" (Eddy & Durbin 1994)
- aka profile stochastic context-free grammars
- aka hidden Markov models on steroids

Model position-specific nucleotide preferences *and* base-pair preferences

Pro: accurate

Con: model building hard, search sloooow

Example: searching for tRNAs



Profile Hmm Structure

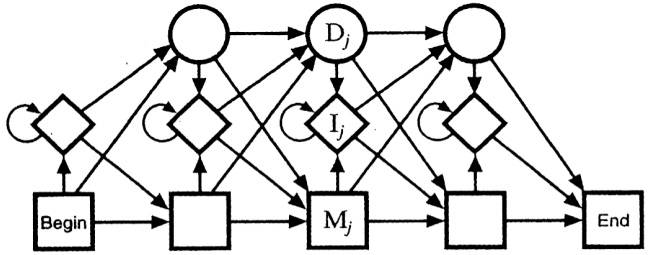


Figure 5.2 *The transition structure of a profile HMM.*

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

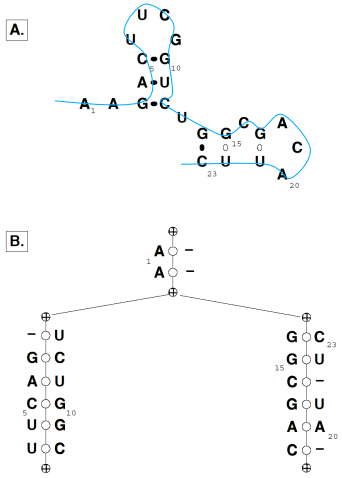
CM Structure

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)



CM Viterbi Alignment

x_i = i^{th} letter of input

x_{ij} = substring $i,...,j$ of input

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

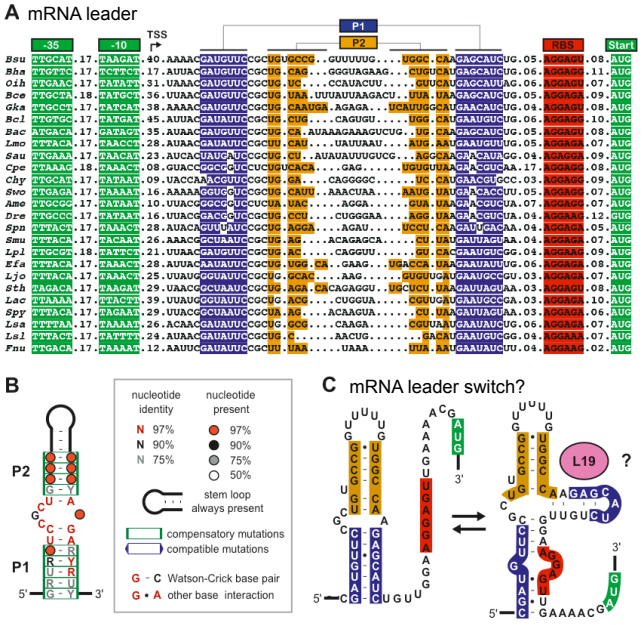
$E^y_{x_i,x_j}$ = $P(\text{emission of } x_i,x_j \text{ from state } y)$

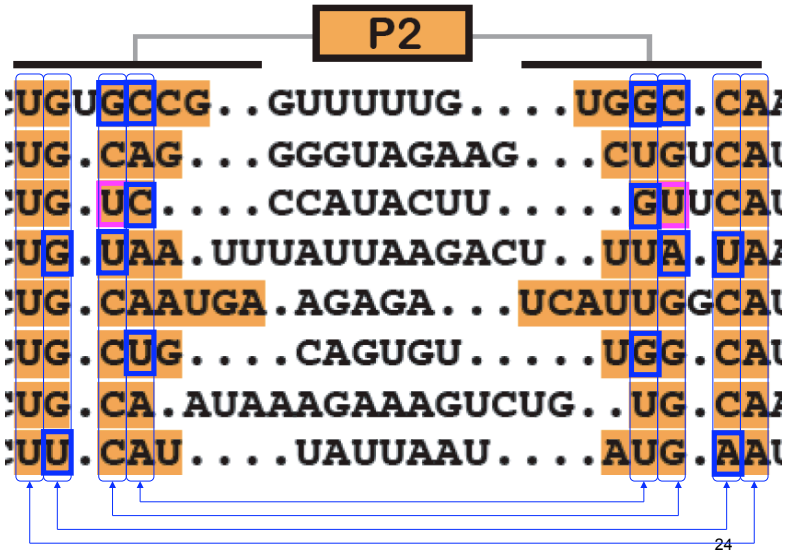
S^y_{ij} = $\max_\pi \log P(x_{ij} \text{ gen'd starting in state } y \text{ via path } \pi)$

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

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

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





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

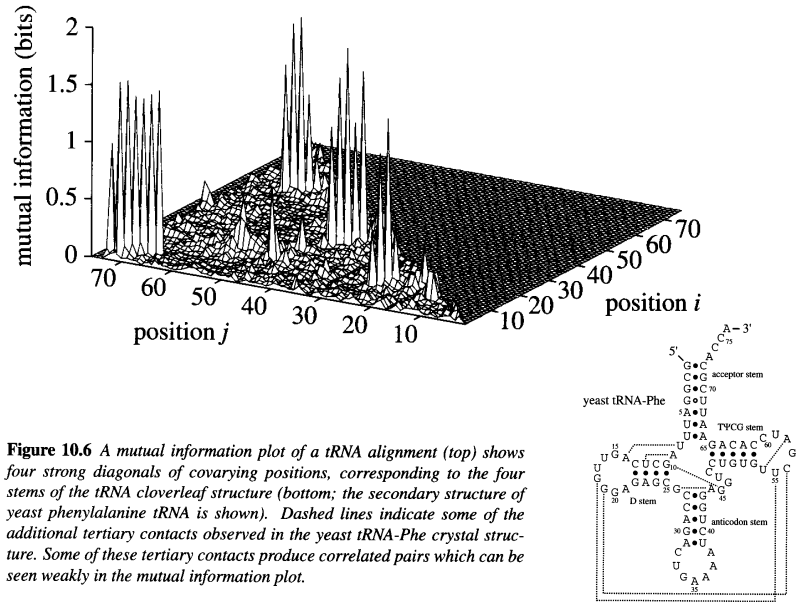


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary 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 $\left(\sum_{i=1}^{n} \max_j M_{i,j}\right)/2$

Dataset	Avg. id	Min id	Max id	ClustalV accuracy	1° info (bits)	2° info (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.

Rfam – an RNA family DB

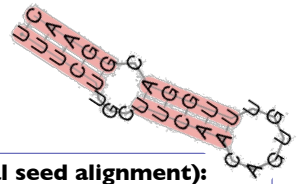
Griffiths-Jones, et al., NAR '03,'05

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

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 score < threshold;

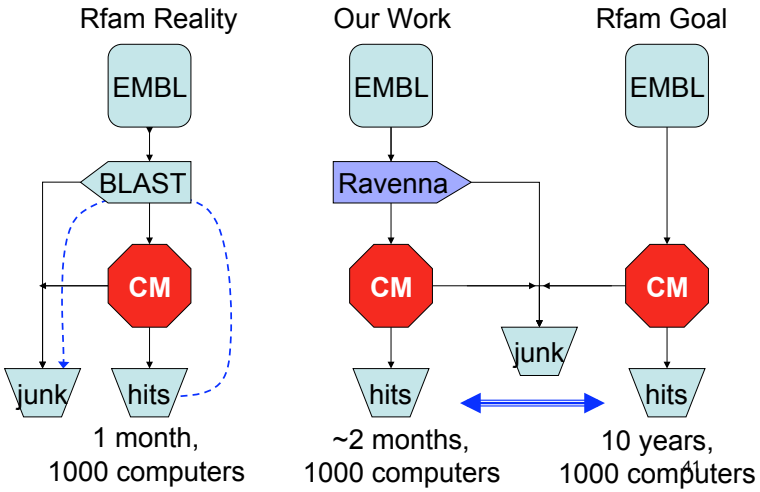
actually run CM on the rest (the promising parts)

assignment of HMM transition/emission scores is key

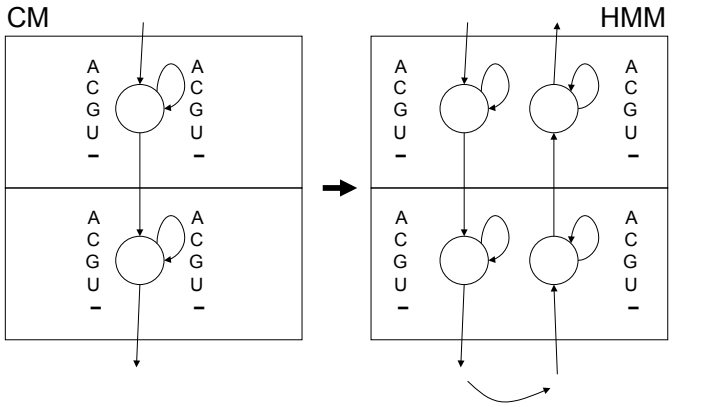
(large convex optimization problem)

Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

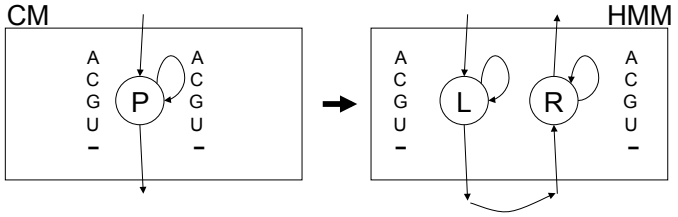
CM's are good, but slow



CM to HMM



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

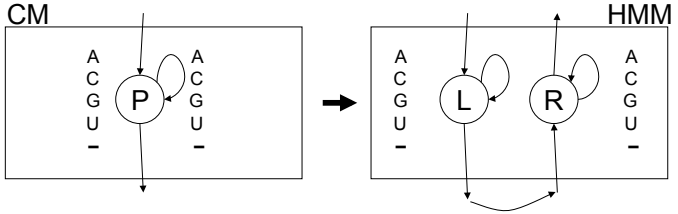
NB: ok if "probs" aren't, e.g. $\Sigma \neq 1$
(e.g. in CM, emissions are odds ratios vs 0th-order background)

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

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

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

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

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 $E(L_i, R_i)$

$E(L_i, R_i) = \Sigma_x$ 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:

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

Viterbi:

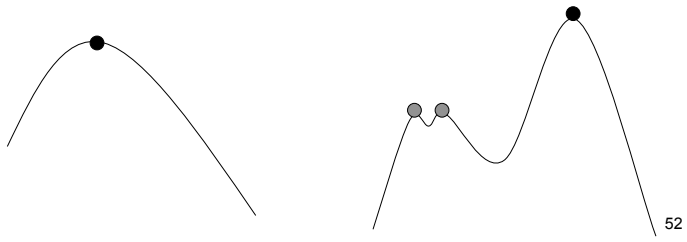
$$v_l(i+1) = e_l(x_{i+1}) \cdot \max_k (v_k(i)\, a_{k,l})$$

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

"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

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

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.

Cmfinder--A Covariance Model Based RNA Motif Finding Algorithm

Bioinformatics, 2006, 22(4): 445-452

Zizhen Yao
Zasha Weinberg
Walter L. Ruzzo
University of Washington, Seattle

Approaches

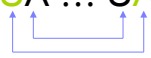
Align sequences, then look for common structure

Predict structures, then try to align them

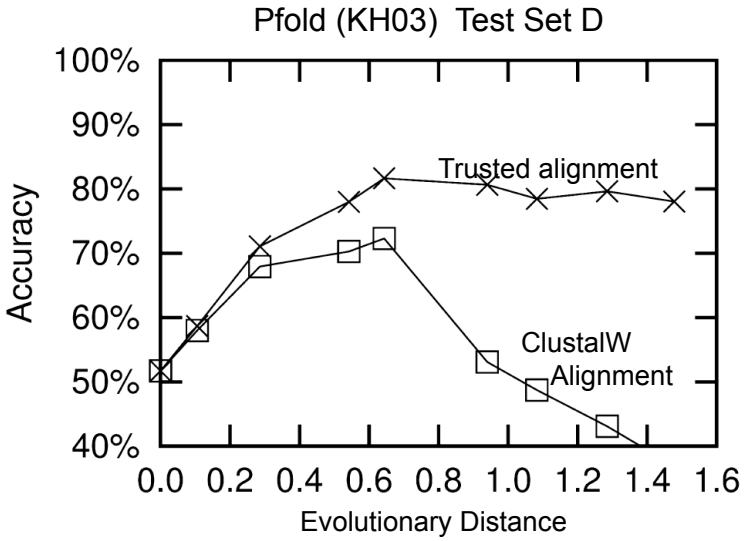
Do both together

"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)



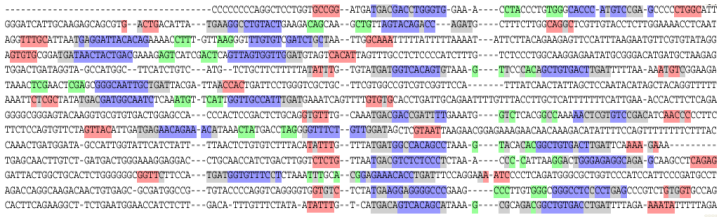
Knudsen & Hein, Pfold: RNA secondary structure prediction using stochastic context-free grammars, Nucleic Acids Research, 2003, v 31,3423–3428

Pitfall for sequence-alignment-first approach

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions



same-colored boxes *should* be aligned

Approaches

Align sequences, then look for common structure

Predict structures, then try to align them

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

Do both together

- Sankoff – good but slow
- Heuristic

Our Approach: CMfinder

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

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

Design Goals

Find RNA motifs in unaligned sequences

Seq conservation exploited, but not required

Robust to inclusion of unrelated sequences

Robust to inclusion of flanking sequence

Reasonably fast and scalable

Produce a probabilistic model of the motif that can be directly used for homolog search

Alignment → CM → Alignment

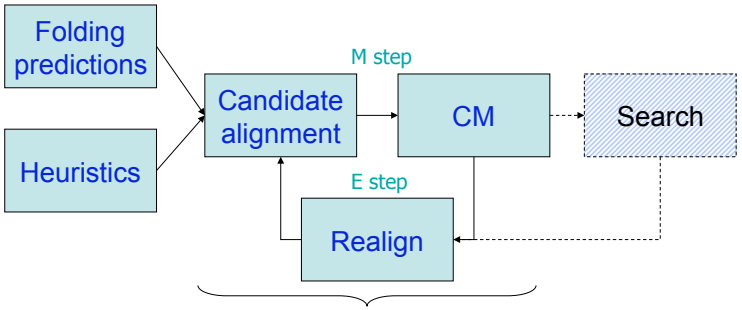
Similar to HMM, but slower

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)

CMfinder Outline



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

Initial Alignment Heuristics

fold sequences separately

candidates: regions with low folding energy

compare candidates via "tree edit" algorithm

find best "central" candidates & align to them

BLAST anchors

Structure Inference

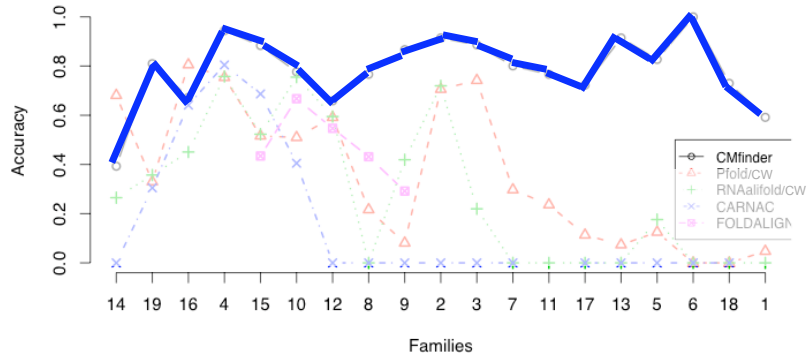
Part of M-step is to pick a structure that maximizes data likelihood

We combine:

- mutual information
- position-specific priors for paired/unpaired
  - (based on single sequence thermodynamic folding predictions)
- intuition: for similar seqs, little MI; fall back on single-sequence folding predictions
- data-dependent, so not strictly Bayesian

CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Application I

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

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)

Approach

Get bacterial genomes

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

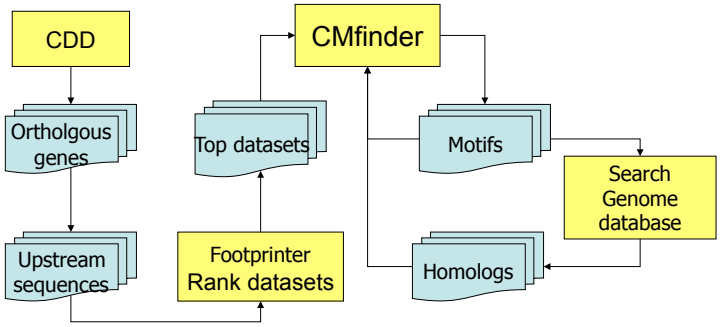
Find most promising genes, based on conserved sequence motifs (Footprinter)

From those, find structural motifs (CMfinder)

Genome-wide search for more instances (Ravenna)

Expert analyses (Breaker Lab, Yale)

A pipeline for RNA motif genome scans



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

Many riboswitches, e.g., are present in ~5 copies per genome

In most close relatives

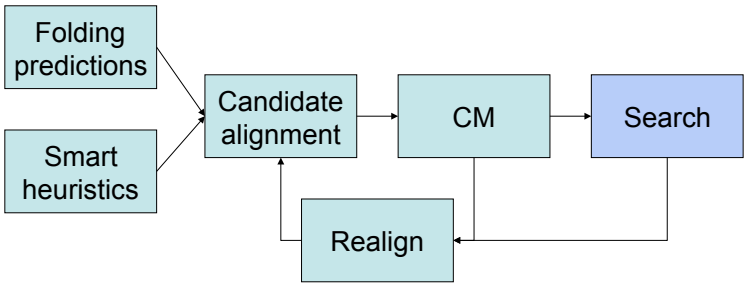
More examples give better model, hence even more examples, fewer errors

More examples give more clues to function - critical for wet lab verification

But inclusion of non-examples can degrade motif…

Genome Scale Search

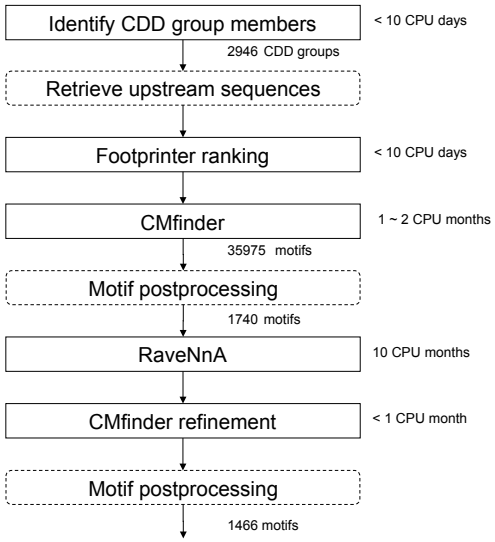
CMfinder is directly usable for/with search



Results

Have analyzed most sequenced bacteria (~2005)

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



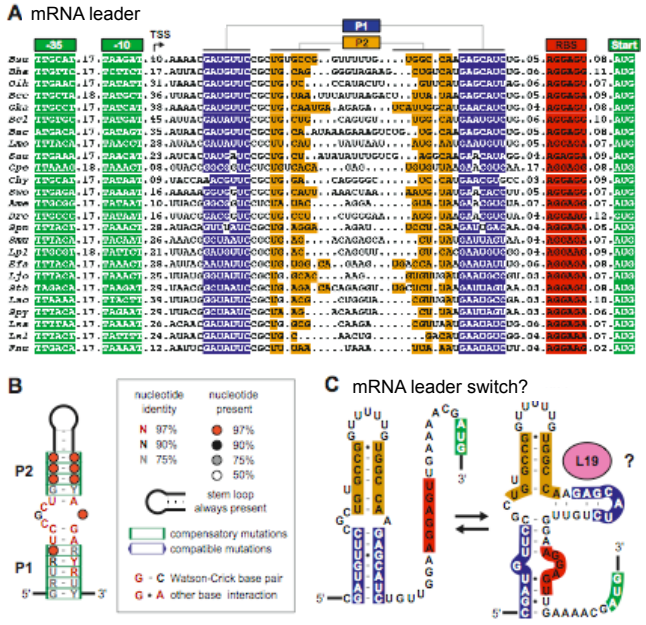
Rank			Score	#		CDD			Rfam
RAV	CMF	FP		RAV	CMF	ID	Gene	Description	
0	43	107	3400	367	11	9904	IlvB	Thiamine pyrophosphate-requiring enzymes	RF00230 T-box
1	10	344	3115	96	22	13174	COG3859	Predicted membrane protein	RF00059 THI
2	77	1284	2376	112	6	11125	MetH	Methionine synthase I specific DNA methylase	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%.

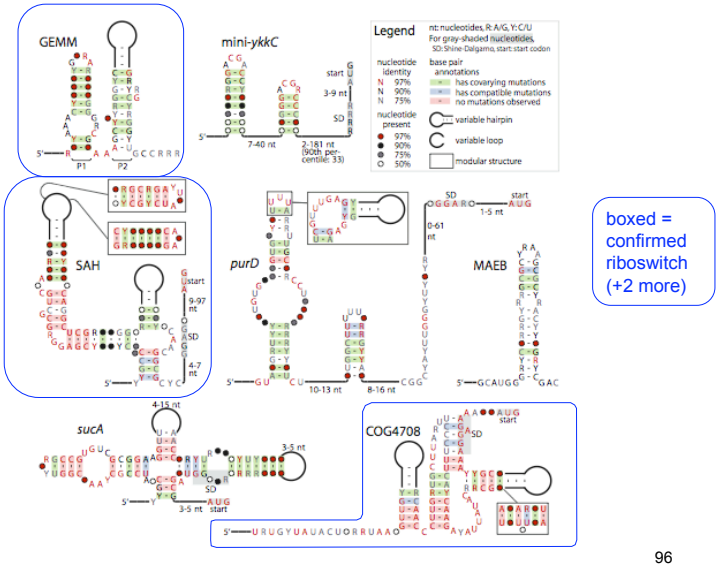
Rank	#	CDD	Gene: Description	Annotation
6	69	28178	DHOase IIa: Dihydroorotase	PyrR attenuator [22]
15	33	10097	RplL: Ribosomal protein L7/L1	L10 r-protein leader; see Supp
19	36	10234	RpsF: Ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: Dinucleotide-utilizing enzymes	6S RNA [25]
27	27	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



Application II

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.



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

New Riboswitches

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

GEMM regulated genes

Pili and flagella	Chitin
Secretion	Membrane Peptide
Chemotaxis	Other - *tfoX*, cytochrome c
Signal transduction	

GEMM sense a metabolite (cyclic di-GMP) produced for signal transduction or for cell-cell communication.

Utility?

Unknown

BUT

E.g., there are no known human riboswitches, so potentially fewer side effects from drugs that might target them

Some such drugs (w/ previously unknown targets) have been known for decades!

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.

Search in Vertebrates

Extract ENCODE Multiz alignments

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

Apply CMfinder to both strands.

10,106 predictions, 6,587 clusters.

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

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

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

Assoc w/ coding genes

Many known human ncRNAs lie in introns

Several of our candidates do, too, including some of the tested ones

- #6: *SYN3* (Synapsin 3)
- #10: *TIMP3*, antisense within *SYN3* intron
- #9: *GRM8* (glutamate receptor metabotropic 8)

Overlap with known transcripts

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

40% intergenetic, 60% overlap with protein coding gene

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

Overlap w/ Indel Purified Segments

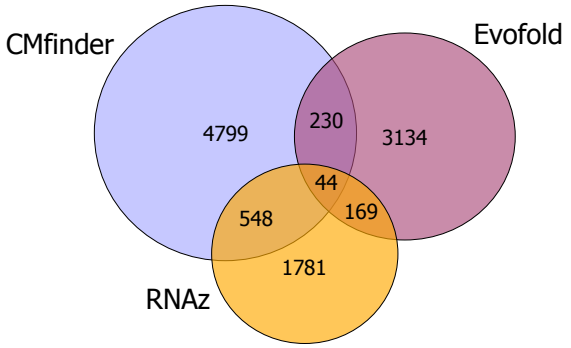
IPS presumed to signal purifying selection

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

Strong P-value for their overlap w/ IPS

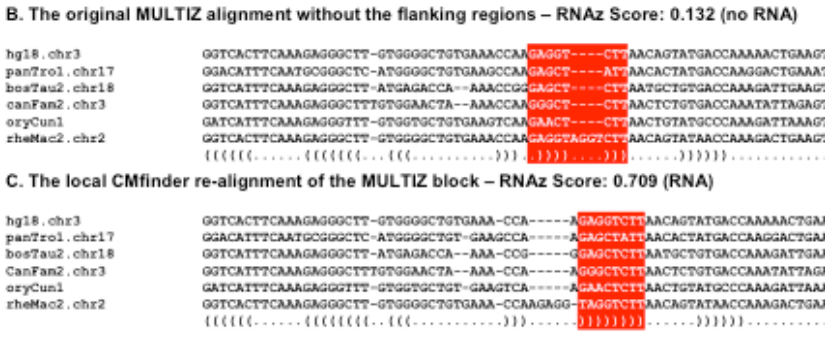
G+C	data	P	N	Expected	Observed	P-value	%
0-35	igs	0.062	380	23	24.5	0.430	5.8%
35-40	igs	0.082	742	61	70.5	0.103	11.3%
40-45	igs	0.082	1216	99	129.5	0.00079	18.5%
45-50	igs	0.079	1377	109	162.5	5.16E-08	20.9%
50-100	igs	0.070	2866	200	358.5	2.70E-31	43.5%
all	igs	0.075	6581	491	747.5	1.54E-33	100.0%

Comparison with Evofold, RNAz

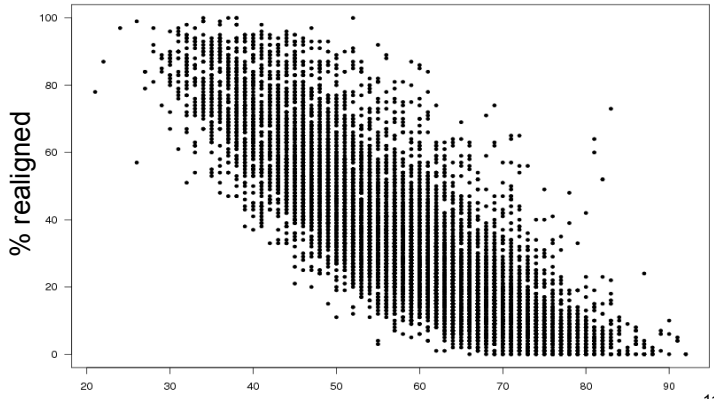


Small overlap (w/ highly significant p-values) emphasizes complementarity

Alignment Matters



Realignment

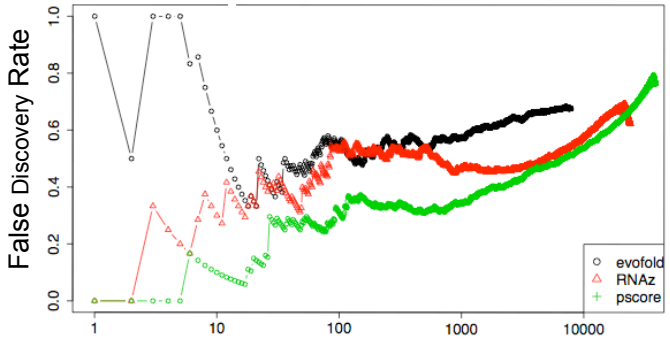


New scoring scheme

Goal: improve false discovery rate for top ranking motifs

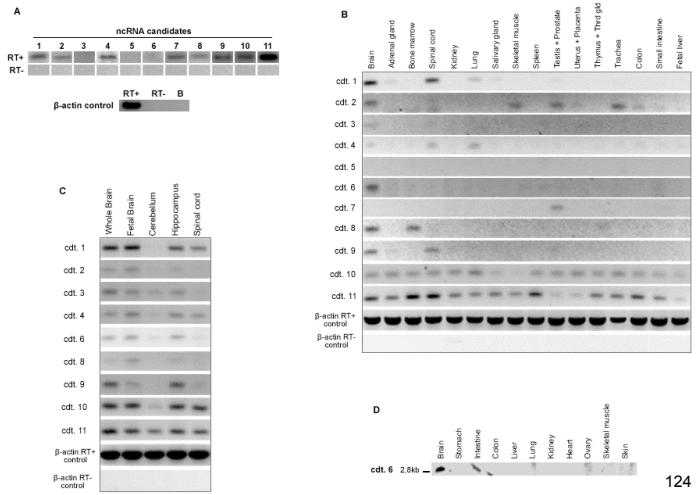
- Current methods can not improve beyond 50% FDR by using higher score threshold.
- Neither RNAz nor Evofold are robust on poorly conserved and gappy regions. (Of course, they weren't designed to be.)

Test on CMfinder motifs in ENCODE regions



FDR vs score ranks in the original alignments

10 of 11 top (differentially) expressed



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.

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

Thanks!

Discovering ncRNAs in prokaryotes through genome-wide clustering

Elizabeth Tseng
UW CSE

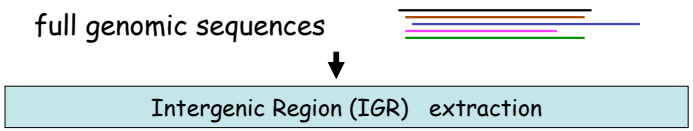
Our work

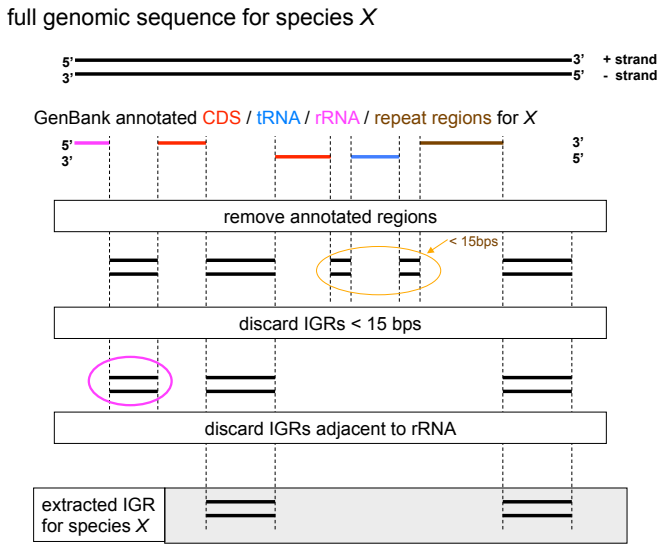
- Goal
  - Clustering for homologous ncRNA prediction
- Our Approach
  - Cluster genomic sequences by homology
  - Incorporate secondary structure information
- Challenges
  - Input: large search space
  - Homology inference: what tools to use?
  - How to evaluate?

Overview

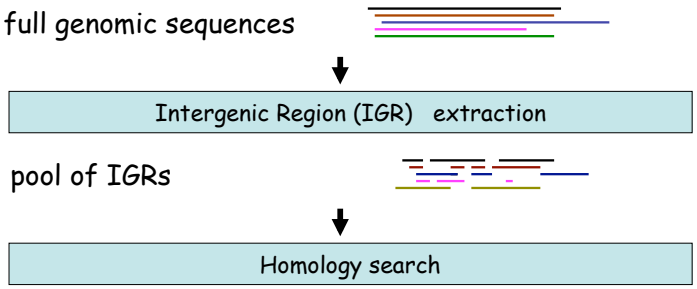
- Motivation
- Approach
  - Clustering based on homology
  - Incorporating secondary structure information
- Evaluation
- Conclusion

Overview of approach

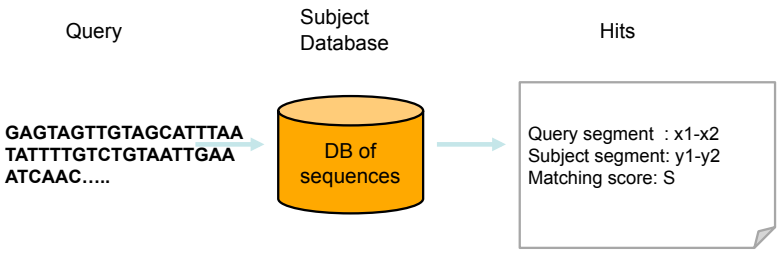




Overview of approach



Homology search programs



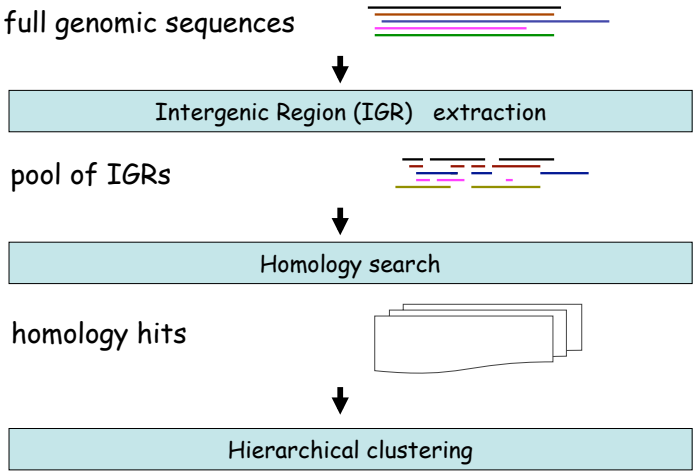
Homology search programs

- Popular homology search programs:
  - NCBI-BLAST
  - WU-BLAST
  - FASTA
  - SSEARCH

Uses dynamic programming to find matching regions between two sequences

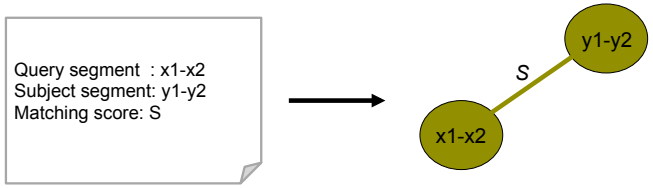
SSEARCH 10 times slower than the rest

Overview of approach



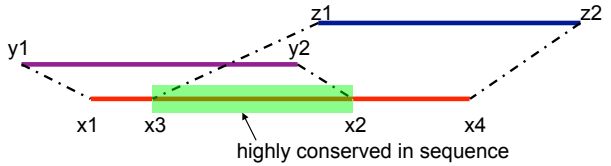
Hierarchical clustering

- Homology program produces a list of hits between IGR *segments*
  - IGR segments ➔ *nodes*
  - A hit between two segments ➔ connecting *edge*
  - Similarity score ➔ *edge weight*

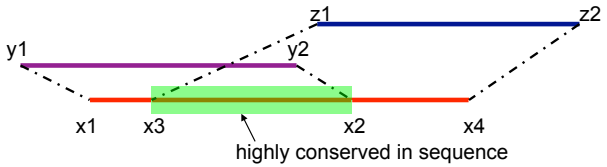


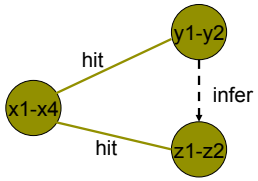
What if segments overlap?

What if (x1,x2) and (x3,x4) overlap by a significant portion?



What if segments overlap?

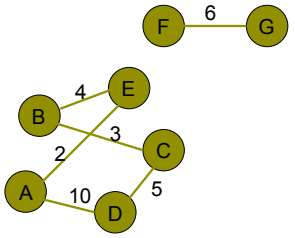


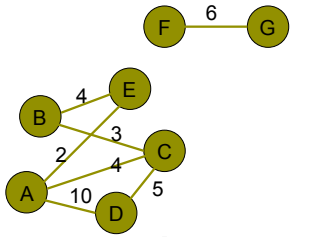


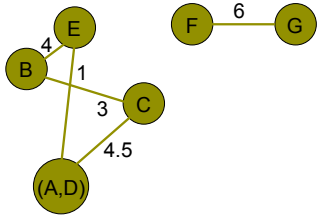
WPGMA

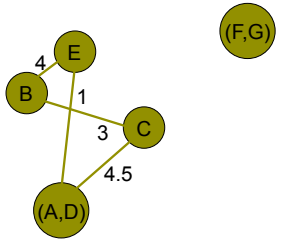
(Weighted Pair Group Method using arithmetic Averaging)

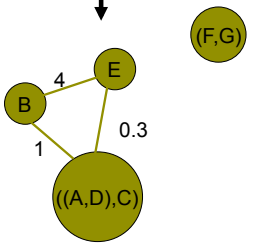
- While exists a connecting edge
  - Select the edge with highest weight
  - Replace the connected two nodes with an new *internal* node
  - Update edge associations



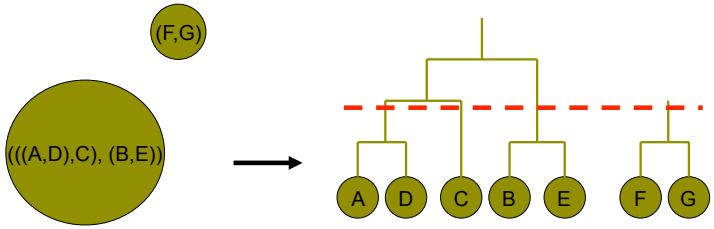








Use size threshold to cut down tree size

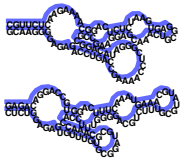


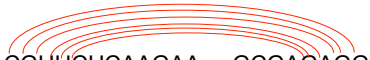
Cluster too big?

Overview

- Motivation
- Approach
  - Clustering based on homology
  - Incorporating secondary structure information
- Evaluation
- Conclusion

Secondary structure info





CGUUCUCAAGAA.....GCGAGAGGGGAACG
GAGACAGGACCG.....AC….AGAGGUCUC

- More conserved in structure than sequence
- Can we include secondary structure when searching for homologs?

Secondary structure info

CGUUCUCAAGAA.....GCGAGAGGGGAACG
GAGACAGGACCG.....AC….AGAGGUCUC

<<<<<<_____....._____>>>>>>
<<<<_____.....__….____>>>>

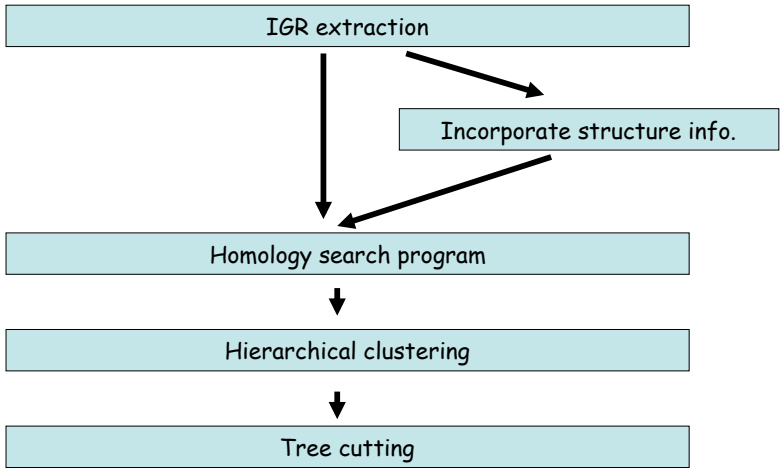
- convert 4-alphabet (A,U,C,G) to 12-alphabet {<, >, _} x {A,U,C,G}
- allow for mismatches between alphabets that are from different nucleotides, but the same structure
- DIY scoring matrix
  - (C< , C<) ➔ great
  - (C< , G<) ➔ good
  - (C< , U_) ➔ bad

Predicting structures on IGR

Given an IGR:

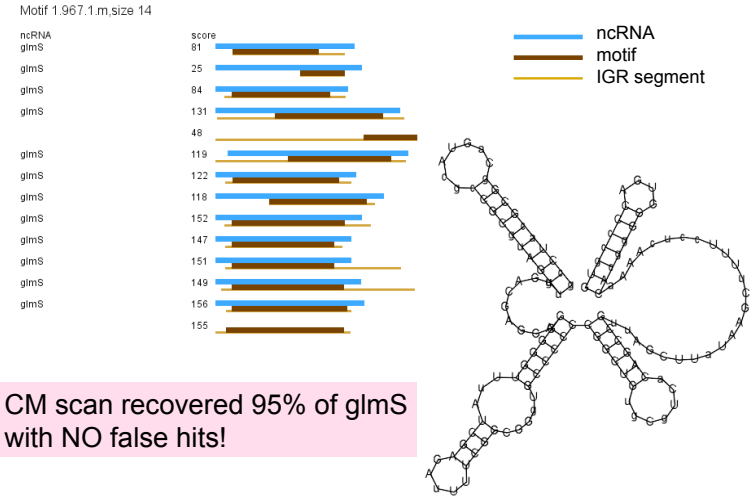
1. Break into overlapping pieces (prev. slide)
2. Feed each piece to RNAfold ➔ obtain structure
3. Convert pieces from 4-alphabet to 12-alphabet
4. Use homology program with DIY scoring matrix
5. Same clustering process…

INPUT: Genomic sequences



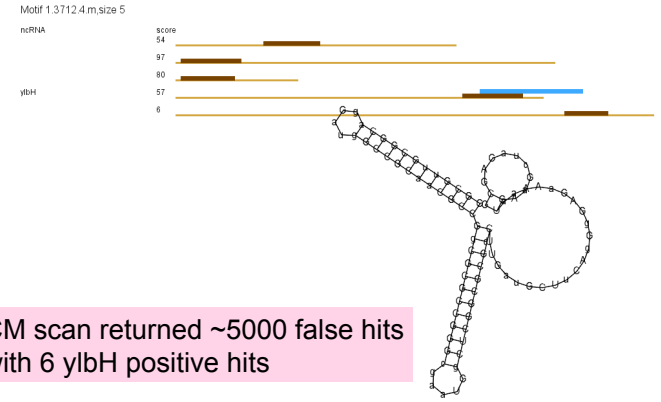
OUTPUT: Clusters of IGR segments

Example of a good scan



CM scan recovered 95% of glmS with NO false hits!

Example of a bad scan



CM scan returned ~5000 false hits with 6 ylbH positive hits