RNA Search and Motif Discovery

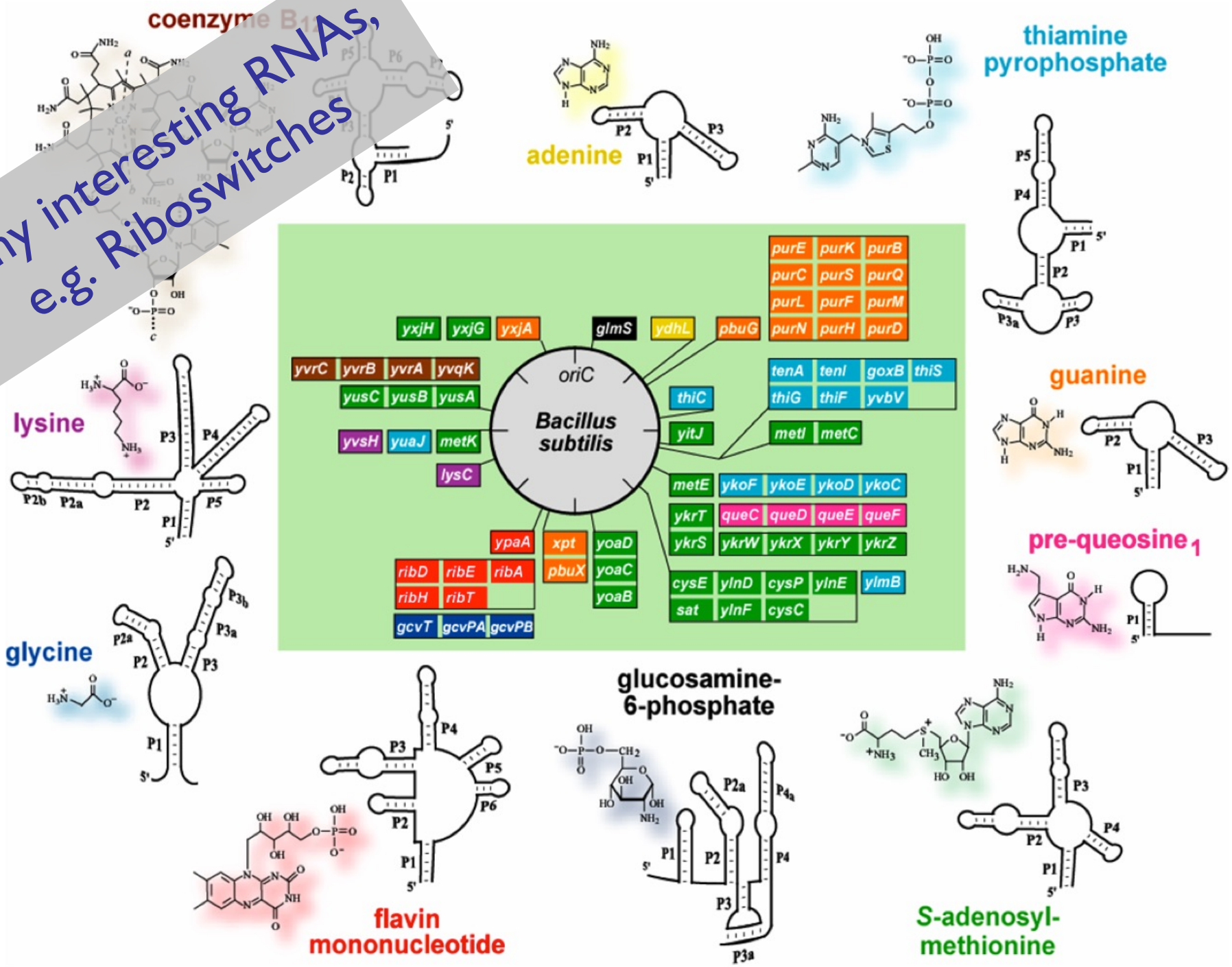
CSE 427
Computational Biology

Previous Lectures

Many biologically interesting roles for RNA

RNA secondary structure prediction

Many interesting RNAs, e.g. Riboswitches



Approaches to Structure Prediction

Maximum Pairing
- + works on single sequences
- + simple
- - too inaccurate

Minimum Energy
- + works on single sequences
- - ignores pseudoknots
- - only finds "optimal" fold

Partition Function
- + finds all folds
- - ignores pseudoknots

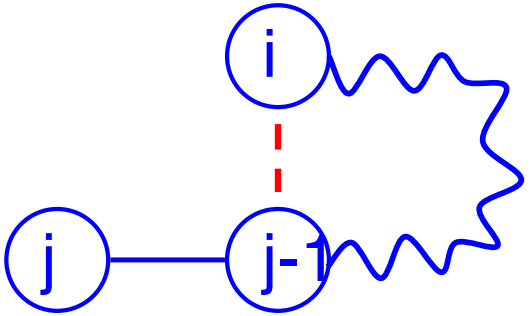
Structure Prediction

"Optimal pairing of r_i ... r_j"

Two possibilities

j Unpaired:

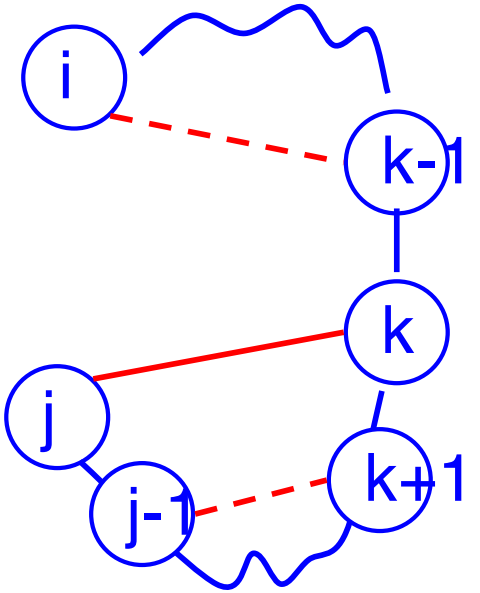
Find best pairing of r_i ... r_{j-1}



j Paired (with some k):

Find best r_i ... r_{k-1} +

best r_{k+1} ... r_{j-1} plus 1

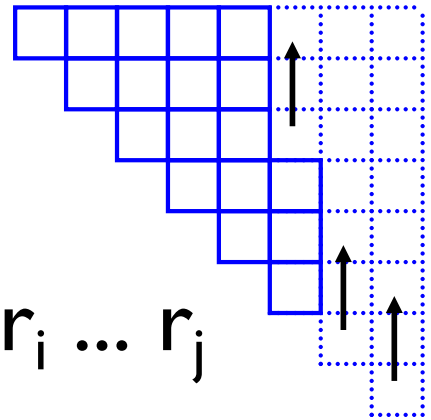


Why is it slow?

Why do pseudoknots matter?

Structure Prediction

Nussinov: Computation Order



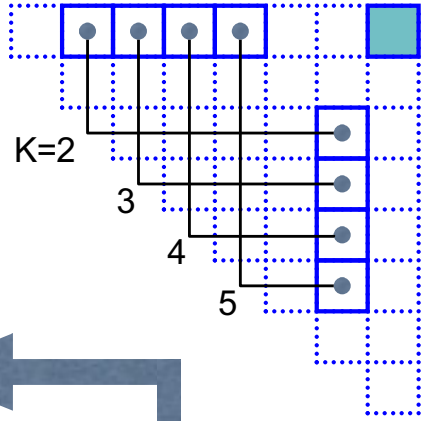
Or -energy

B(i,j) = # pairs in optimal pairing of r_i ... r_j

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

B(i,j) = max of:

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



Time: $O(n^3)$

Loop-based energy version is better; recurrences similar, slightly messier

Structure Prediction

Single Seq Prediction Accuracy

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

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

Definitely useful, but obviously imperfect

Structure Prediction

Approaches, II

Comparative sequence analysis

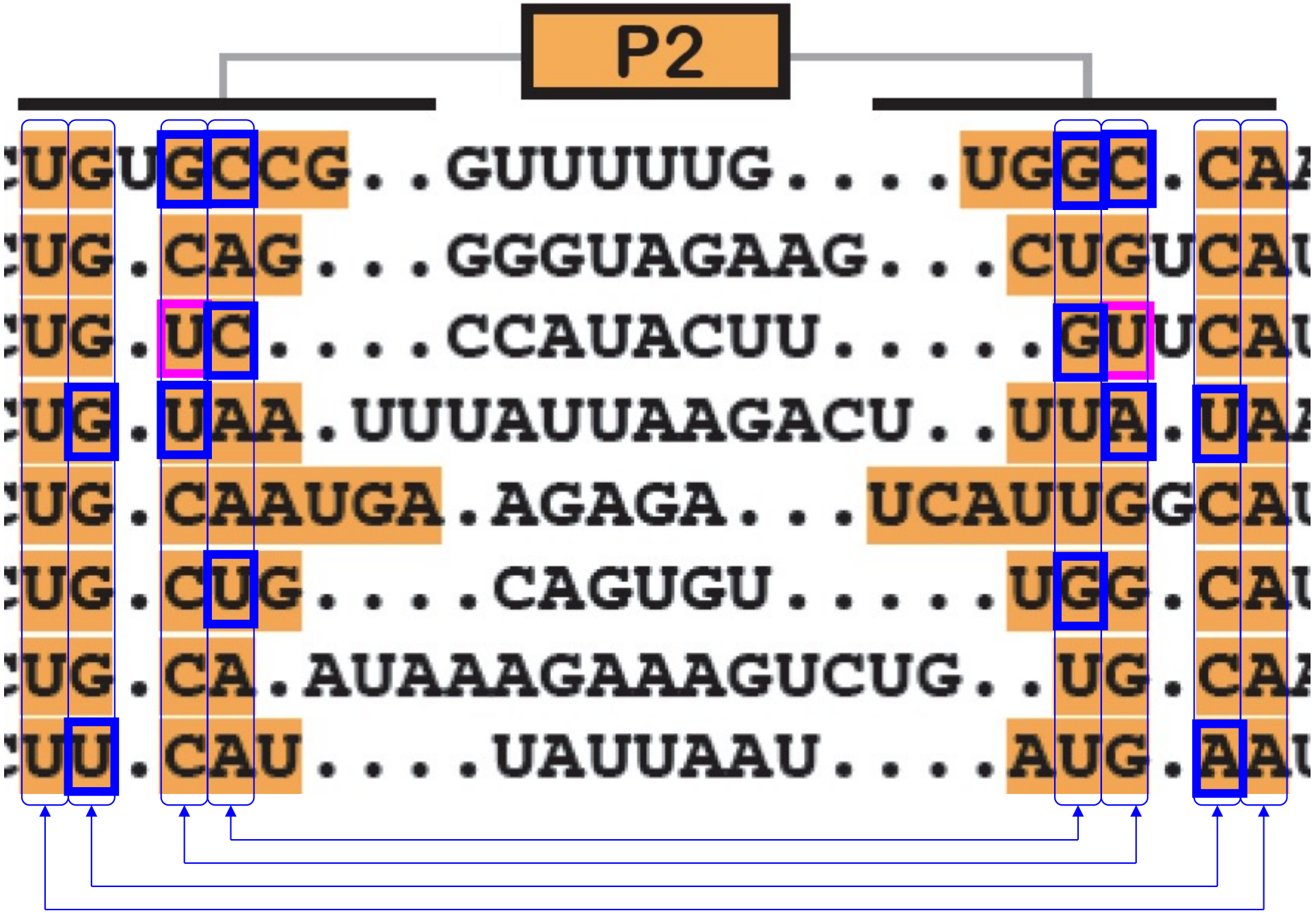
+ handles all pairings (potentially incl. pseudoknots)
- requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars

Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

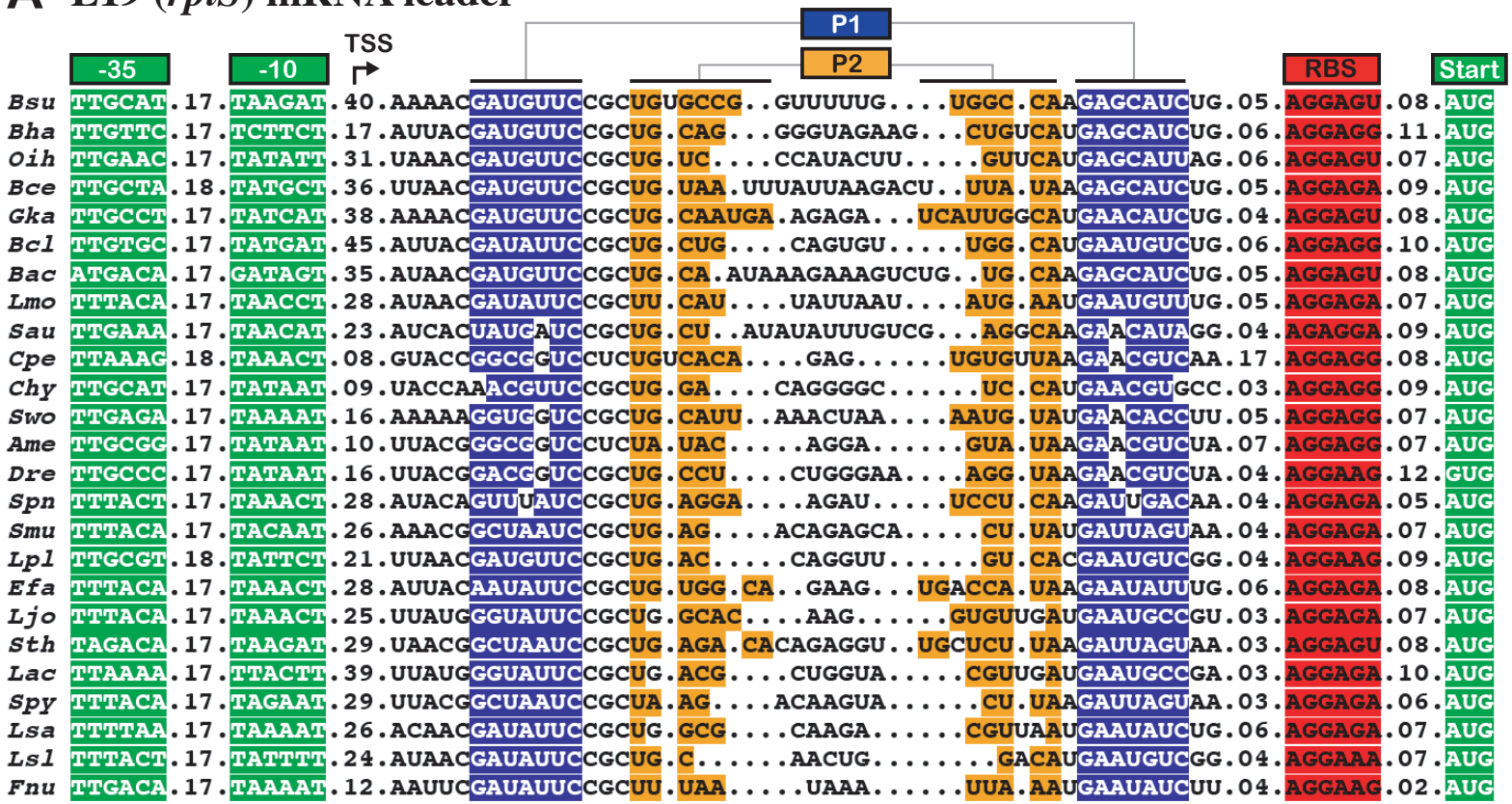
Today



Covariation is strong evidence for base pairing

Example: Ribosomal Autoregulation:

Excess L19 represses L19 (RF00556; 555-559 similar)



Mutual Information

x_k: letter from col k; f_{xk}: its freq in col k; $f_{xi,xj}$: pair freq

$$M_{ij} = \sum_{xi,xj} f_{xi,xj} \log_2 \frac{f_{xi,xj}}{f_{xi} f_{xj}}; \quad 0 \le M_{ij} \le 2$$ (4 letters ⇒ 2 bits)

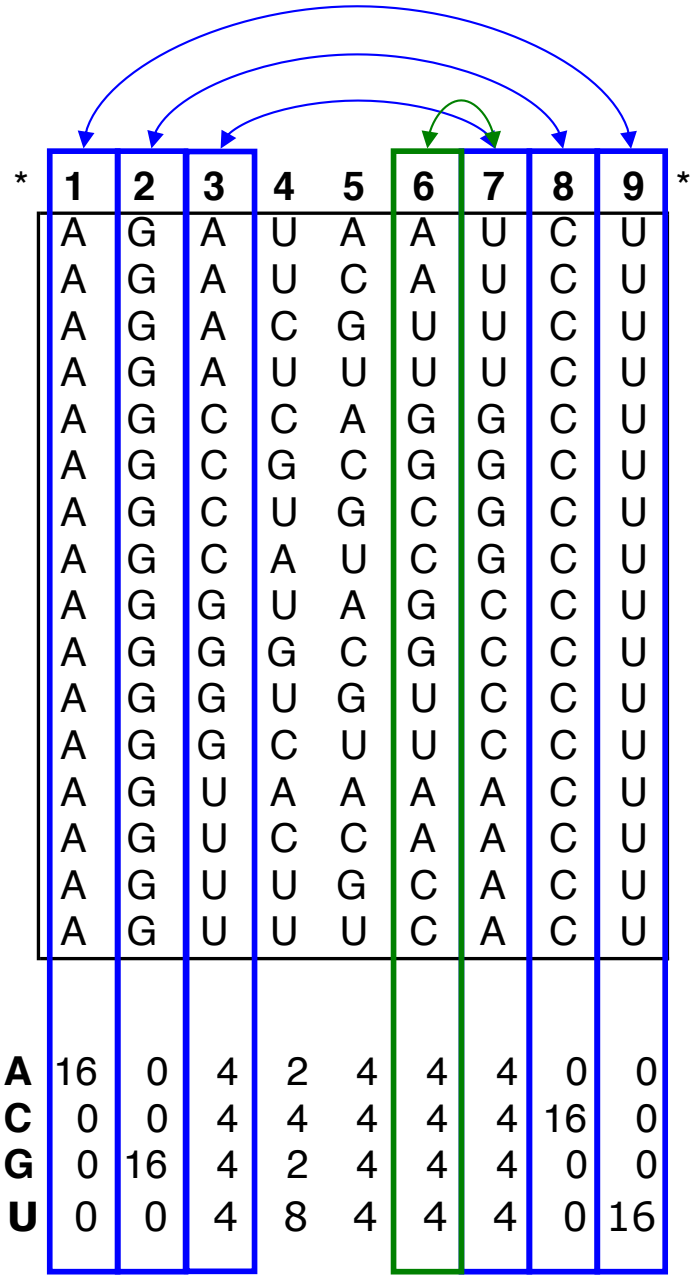
Max when *no* seq conservation but perfect pairing

MI={ given letter in col *i*, what is mate in col *j*?
expected score gain from using a pair state (below)

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)



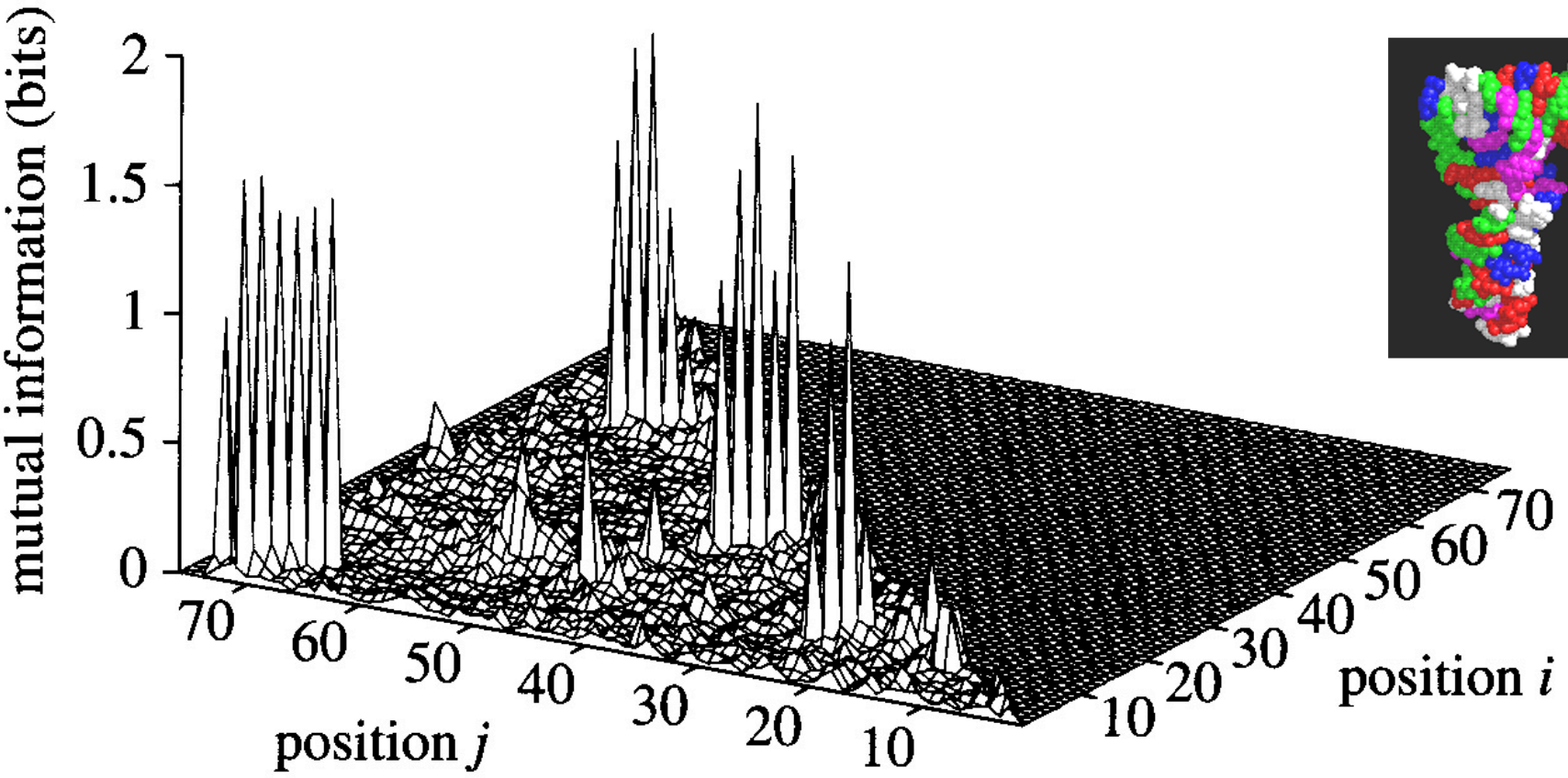
*	1	2	3	4	5	6	7	8	9	*
	A	G	A	U	A	A	U	C	U	
	A	G	A	U	C	A	U	C	U	
	A	G	A	C	G	U	U	C	U	
	A	G	A	U	U	U	U	C	U	
	A	G	C	C	A	G	G	C	U	
	A	G	C	G	C	G	G	C	U	
	A	G	C	U	G	C	G	C	U	
	A	G	C	A	U	C	G	C	U	
	A	G	G	U	A	G	C	C	U	
	A	G	G	G	C	G	C	C	U	
	A	G	G	U	G	U	C	C	U	
	A	G	G	C	U	U	C	C	U	
	A	G	U	A	A	A	A	C	U	
	A	G	U	C	C	A	A	C	U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

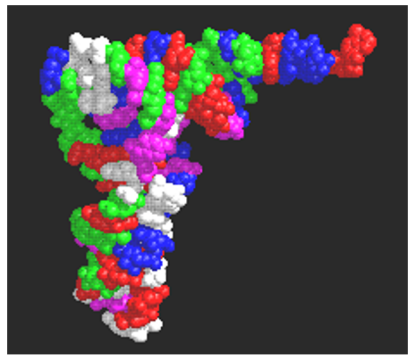
MI:	1	2	3	4	5	6	7	8	9
9	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0		
7	0	0	**2**	**0.30**	**0**	**1**			
6	0	0	**1**	**0.55**	**1**				
5	0	0	**0**	**0.42**					
4	0	0	**0.30**						
3	0	0							
2	0								
1									

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.





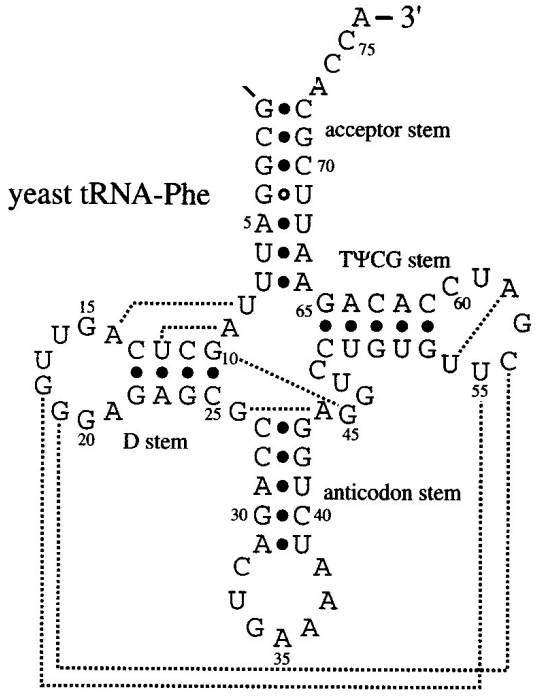


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary 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.*

MI-Based Structure-Learning

Problem: Find best (max total MI) pseudo-knot-free subset of column pairs among i…j.

Solution: "Just like Nussinov/Zucker folding"

$$S_{i,j} = \max \begin{cases} S_{i,j-1} & \text{j unpaired} \\ \max_{i \le k < j-4} S_{i,k-1} + M_{k,j} + S_{k+1,j-1} & \text{j paired} \end{cases}$$

BUT, need the right data—enough sequences at the right phylogenetic distance

Computational Problems

~~How to predict secondary structure~~

How to model an RNA "motif"
(I.e., sequence/structure pattern)

Given a motif, how to search for instances

Given (unaligned) sequences, find motifs

How to score discovered motifs

How to leverage prior knowledge

Motif Description

RNA Motif Models

"Covariance Models" (Eddy & Durbin 1994)

- aka profile stochastic context-free grammars (Sakakibara 94)
- aka hidden Markov models on steroids

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

Pro: accurate

Con: model building hard, search slow

Eddy & Durbin 1994: 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 – a very good tRNA-finder)

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

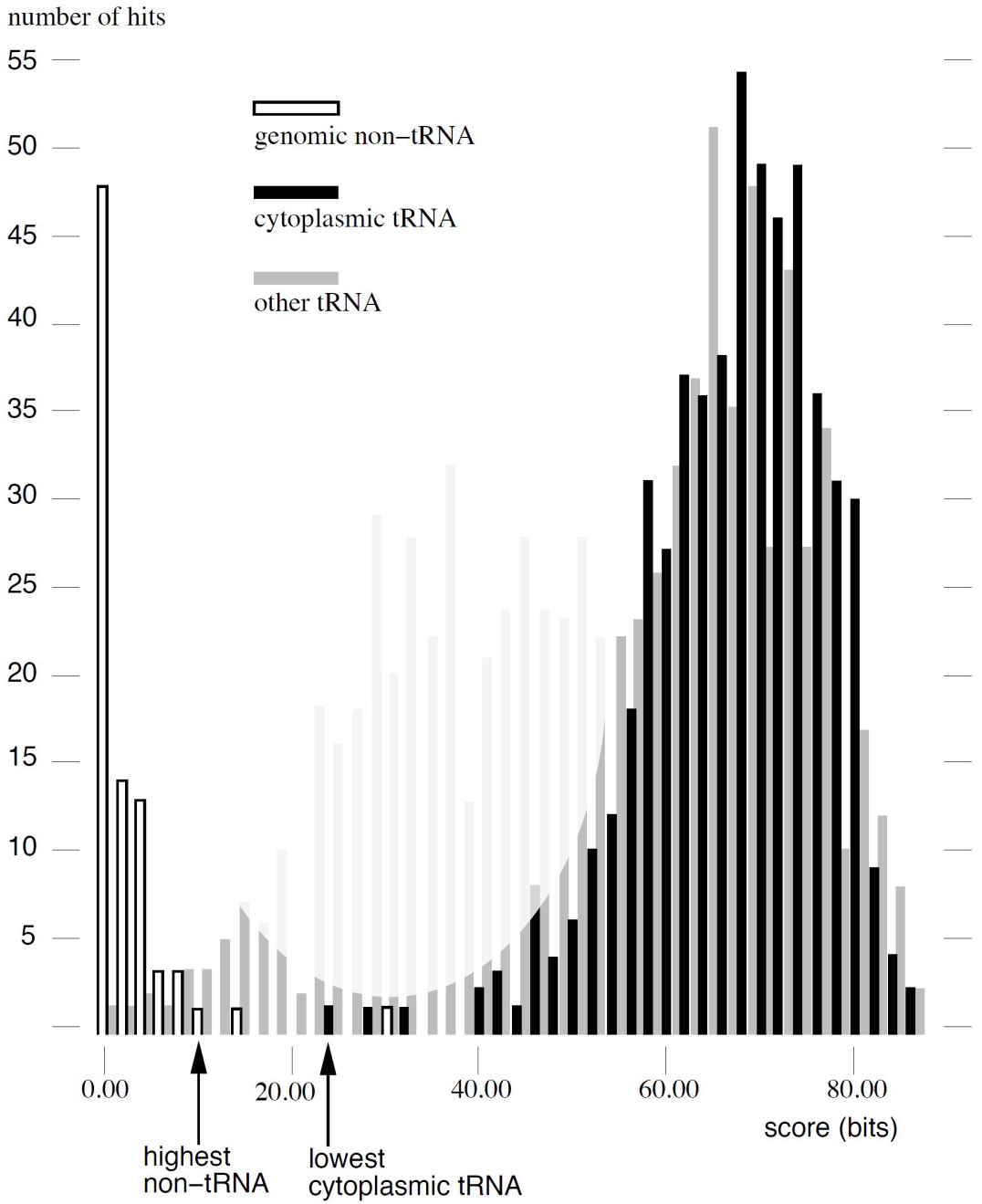
As with HMMs, given a sequence:

- 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



Recall

Profile Hmm Structure

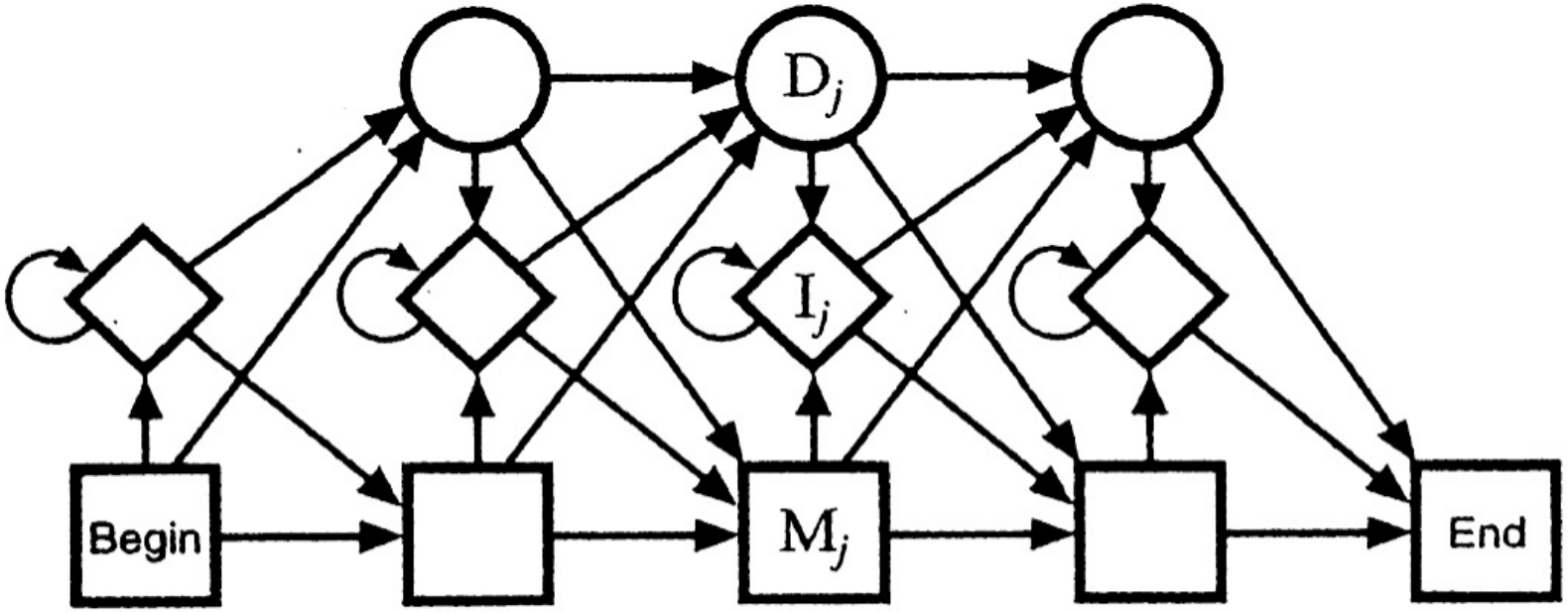


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

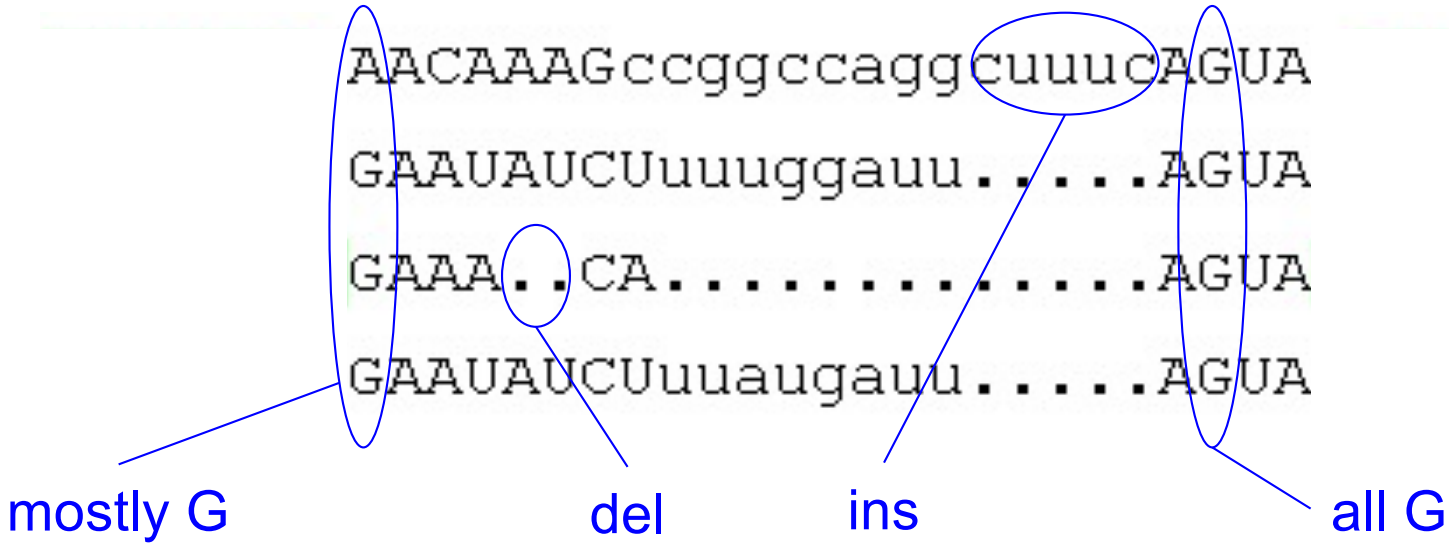
D_j: Delete states (silent - no emission)

How to model an RNA "Motif"?

Conceptually, start with a profile HMM:

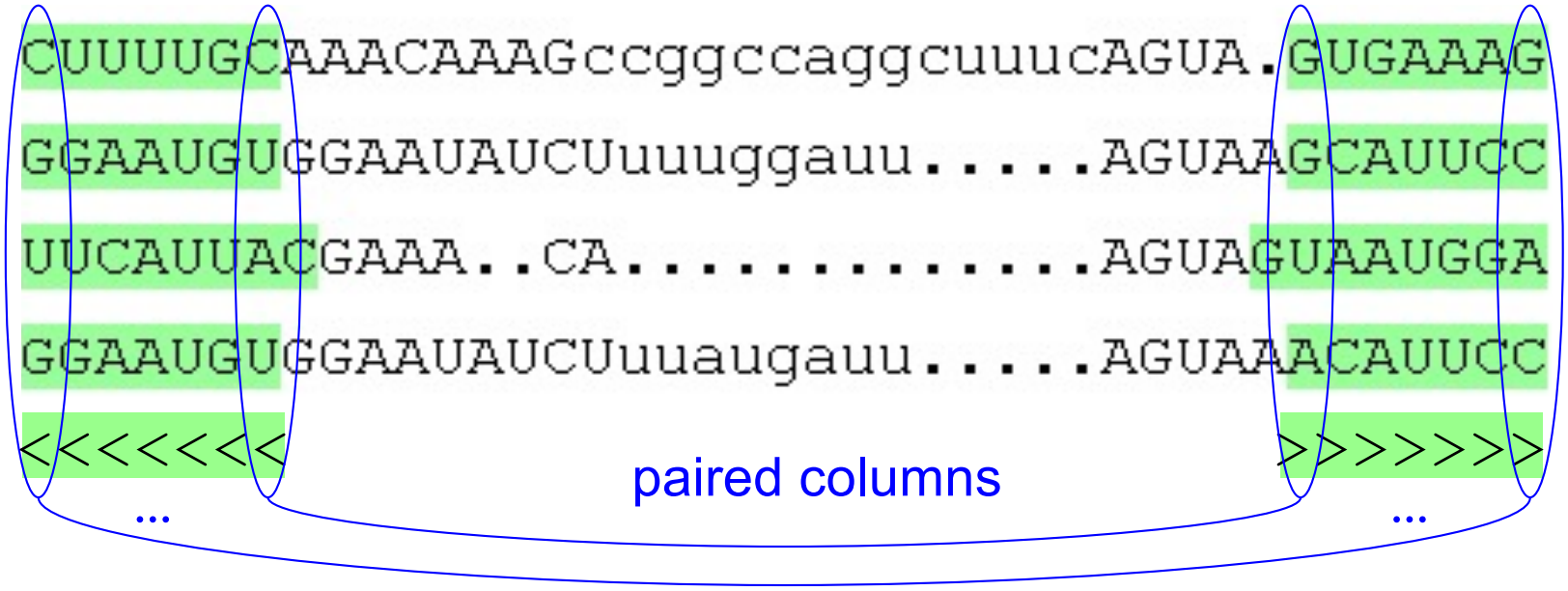
from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position

given a new seq, estimate likelihood that it could be generated by the model, & align it to the model



How to model an RNA "Motif"?

Add "column pairs" and pair emission probabilities for base-paired regions



Does not handle "paired columns" above

Profile Hmm Structure

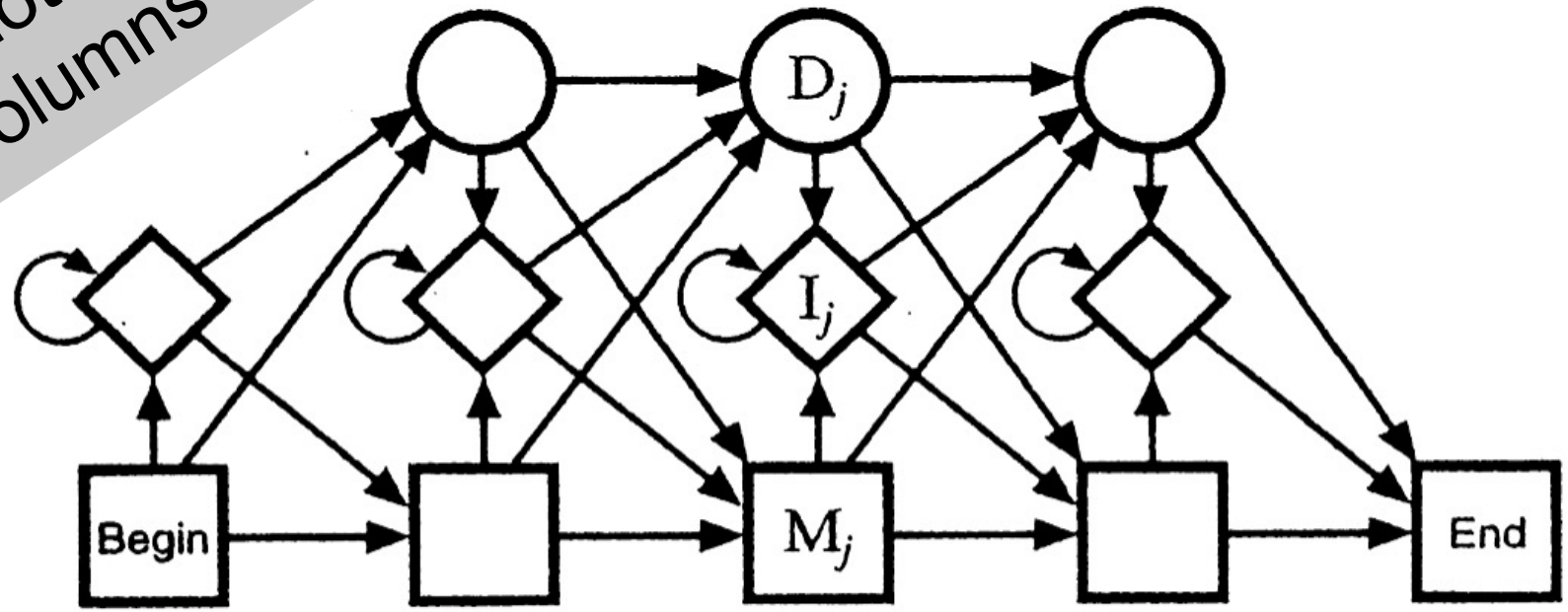


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

D_j: Delete states (silent - no emission)

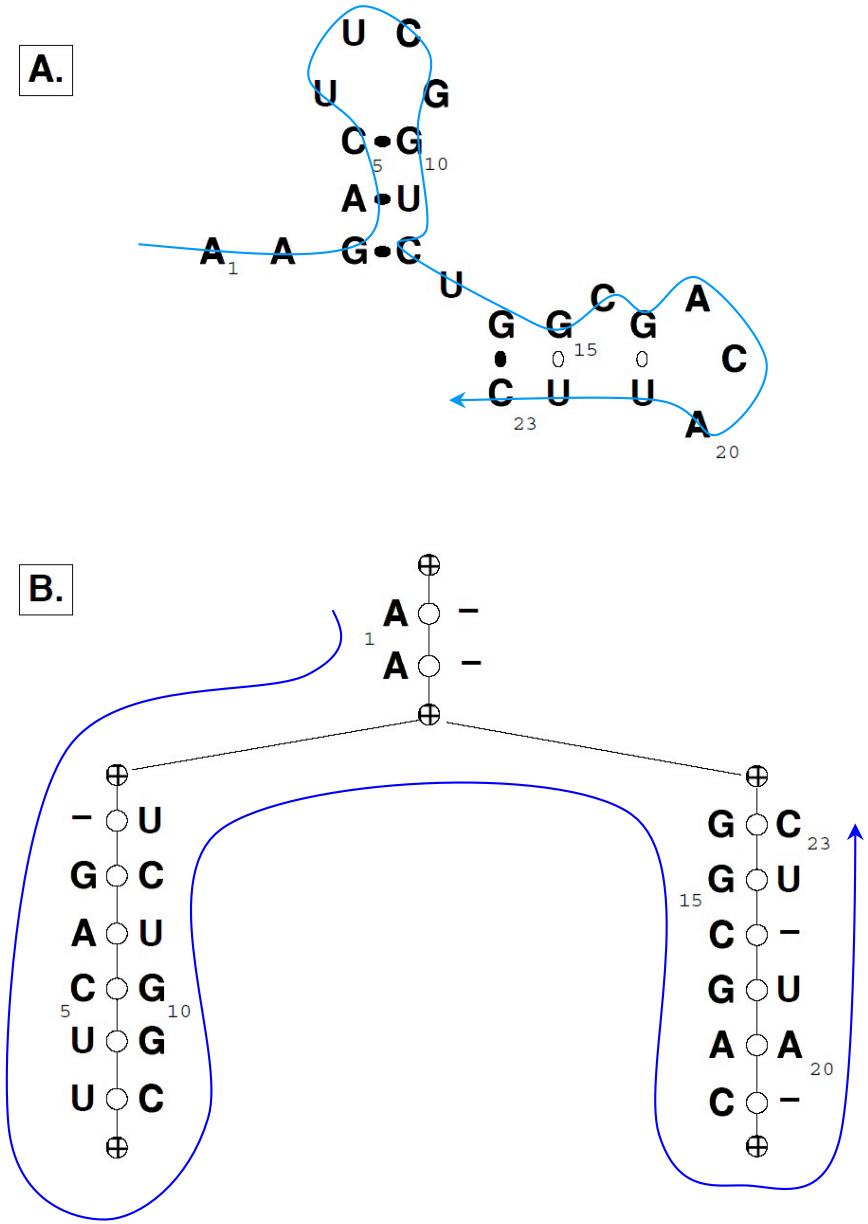
CM Structure

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)

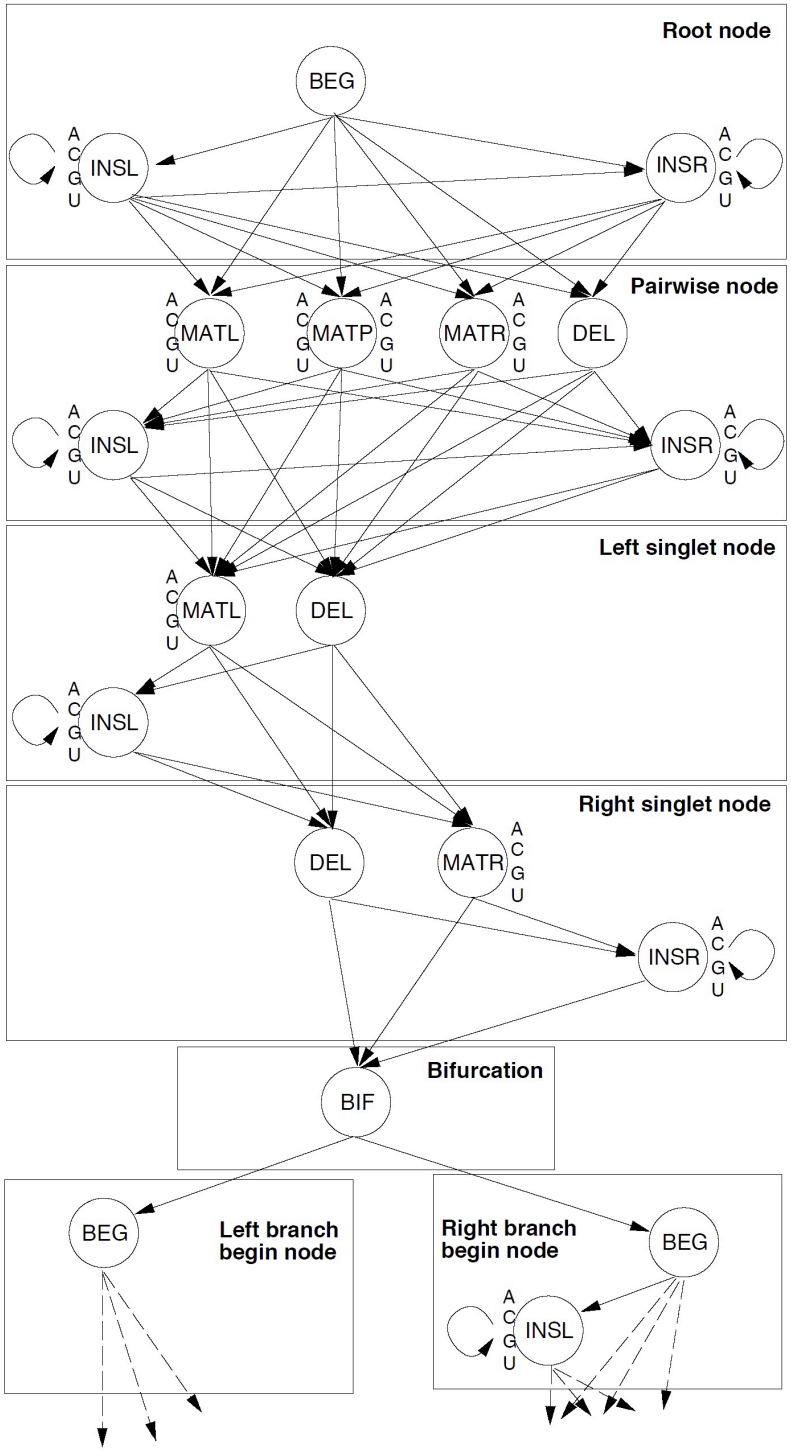


Overall CM Architecture

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

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

Primary vs Secondary Info

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

3 test sets from ED 94

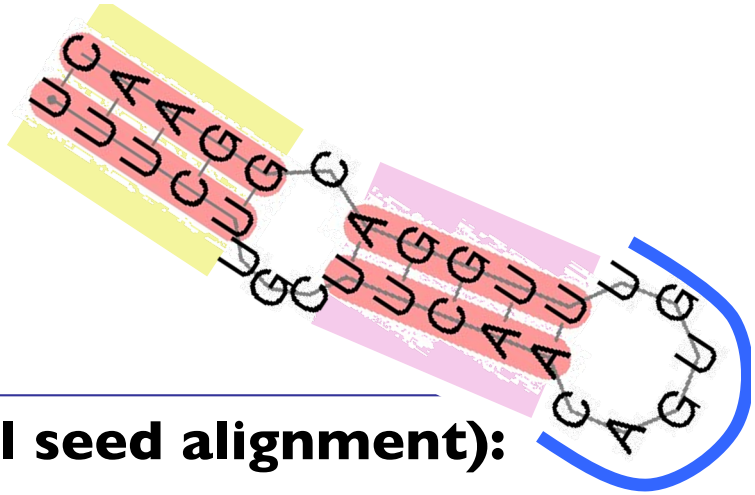
disallowing / allowing pseudoknots

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

An Important Application: Rfam

A Database of RNA Families

RF00037: Example Rfam Family



Input (hand-curated):

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

Output:

- CM
- scan results & "full alignment"
- phylogeny, etc.

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

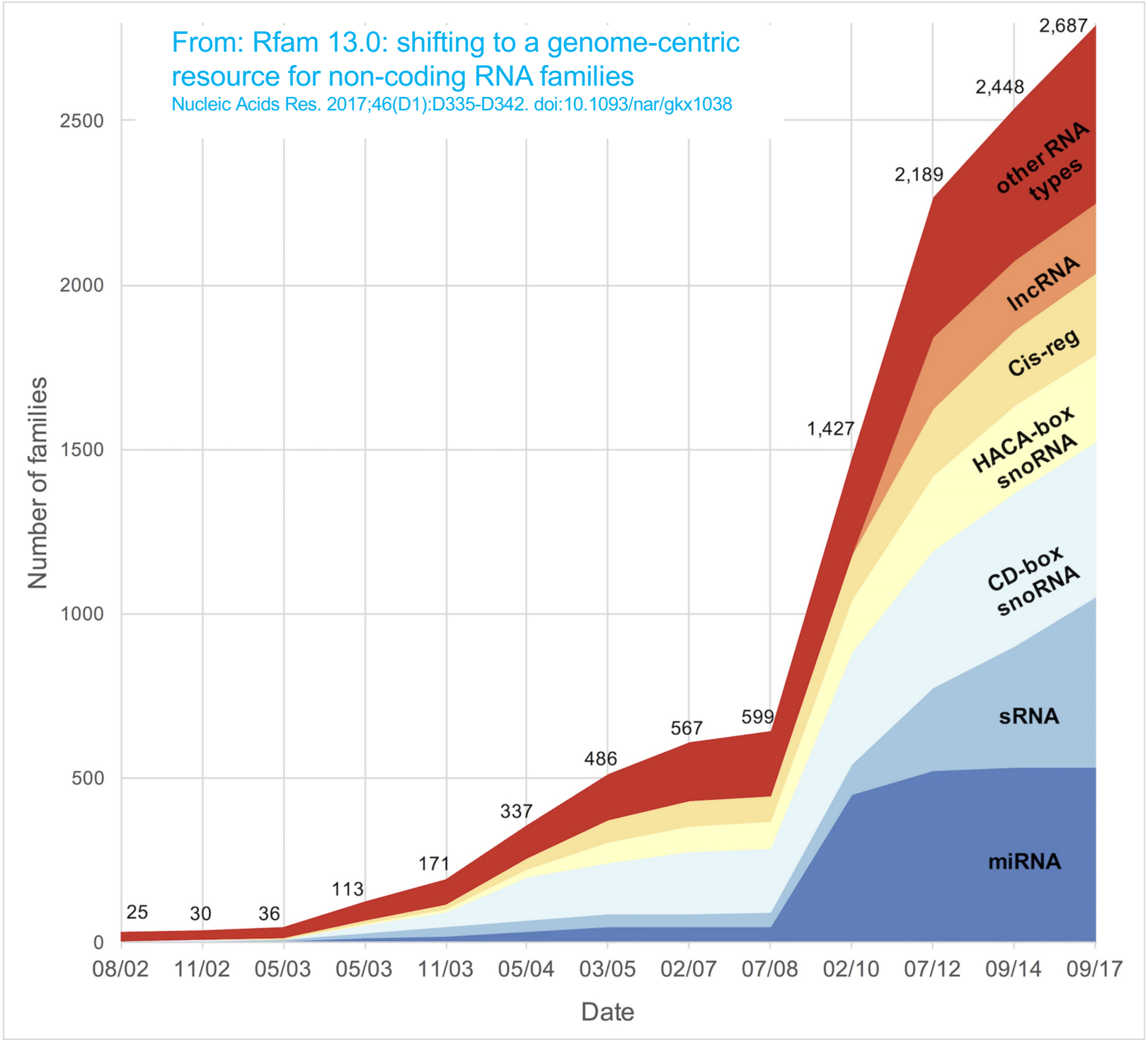
Rfam – an RNA family DB

Griffiths-Jones, et al., NAR '03, '05, '08, '11, '12

Was biggest scientific comp user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:

		DB size:
Rel 1.0, 1/03:	25 families, 55k instances	
Rel 7.0, 3/05:	503 families, 363k instances	~8GB
Rel 9.0, 7/08:	603 families, 636k instances	
Rel 10.0, 1/10:	1446 families, 3193k instances	~160GB
Rel 11.0, 8/12:	2208 families, 6125k instances	~320GB
Rel 12.0, 9/14:	2450 families, 19623k instances	
Rel 12.1, 4/16:	2474 families, 9m instances	
Rel 13.0, 9/17:	2686 families	
Rel 14.3, 9/20:	3446 families	


From: Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families
Nucleic Acids Res. 2017;46(D1):D335-D342. doi:10.1093/nar/gkx1038
Number of families
Date
0
500
1000
1500
2000
2500
08/02
11/02
05/03
05/03
11/03
05/04
03/05
02/07
07/08
02/10
07/12
09/14
09/17
25
30
36
113
171
337
486
567
599
1,427
2,189
2,448
2,687
other RNA types
lncRNA
Cis-reg
HACA-box snoRNA
CD-box snoRNA
sRNA
miRNA

CM Summary

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

They allow accurate search

But

a) search is slow

b) model construction is laborious

An Important Need: Faster Search

Homology search

"Homolog" – similar by descent from common ancestor

Sequence-based

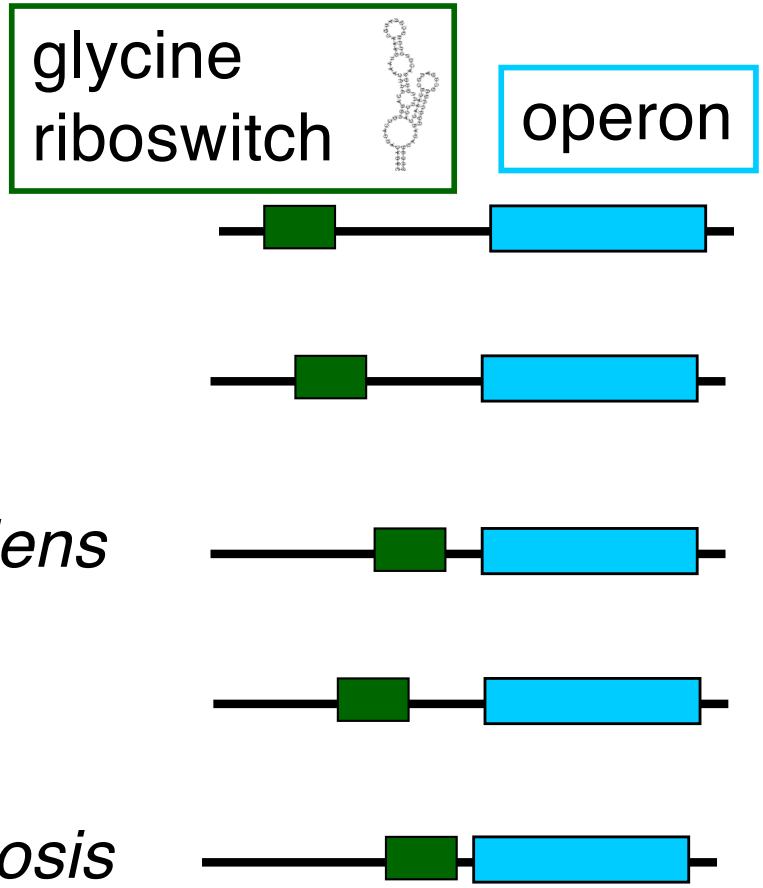
- Smith-Waterman
- FASTA
- BLAST

For RNA, sharp decline in sensitivity at ~60-70% identity

So, use structure, too

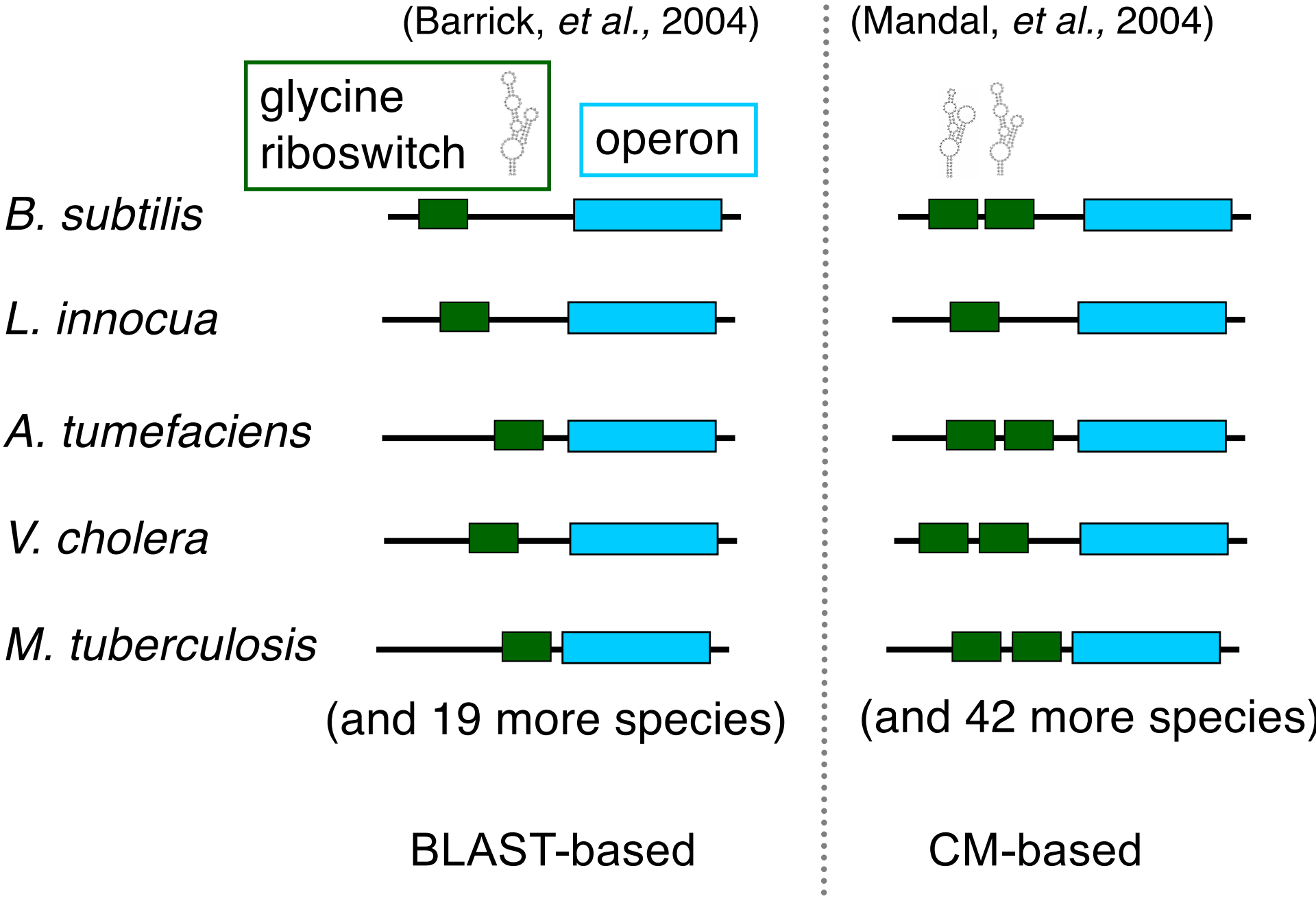
Impact of RNA homology search

(Barrick, *et al.,* 2004)

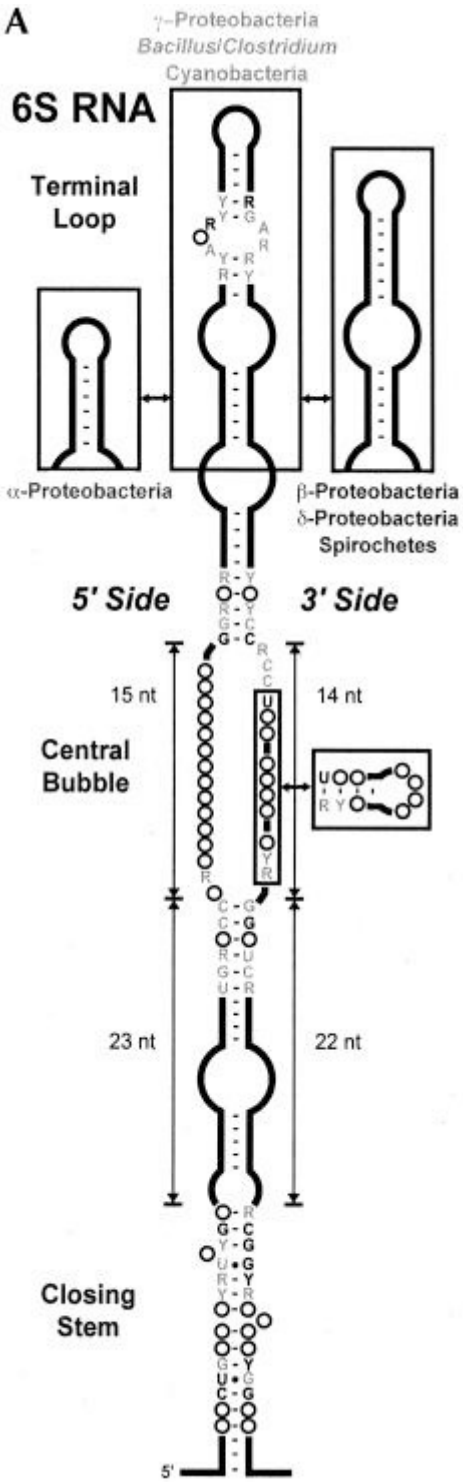


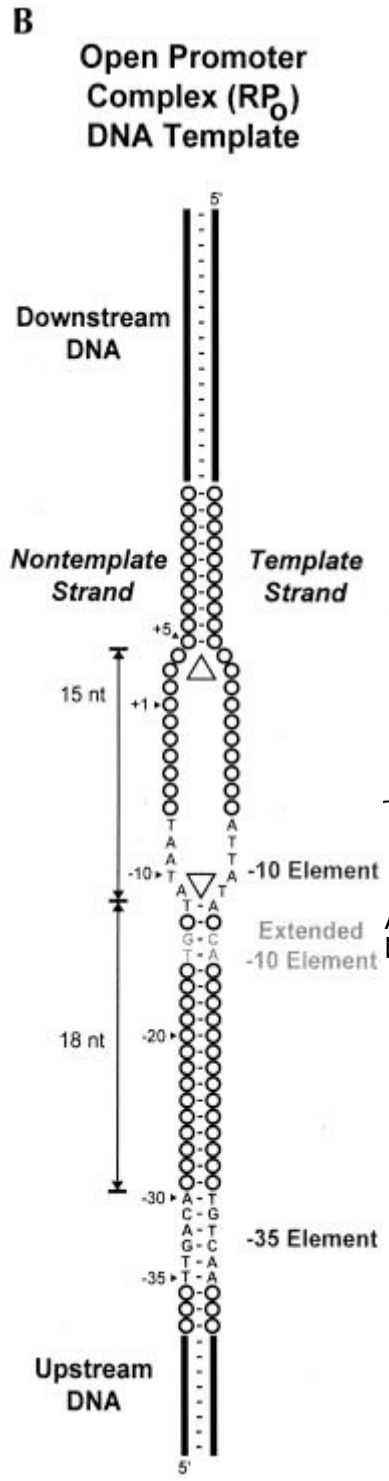
(and 19 more species)

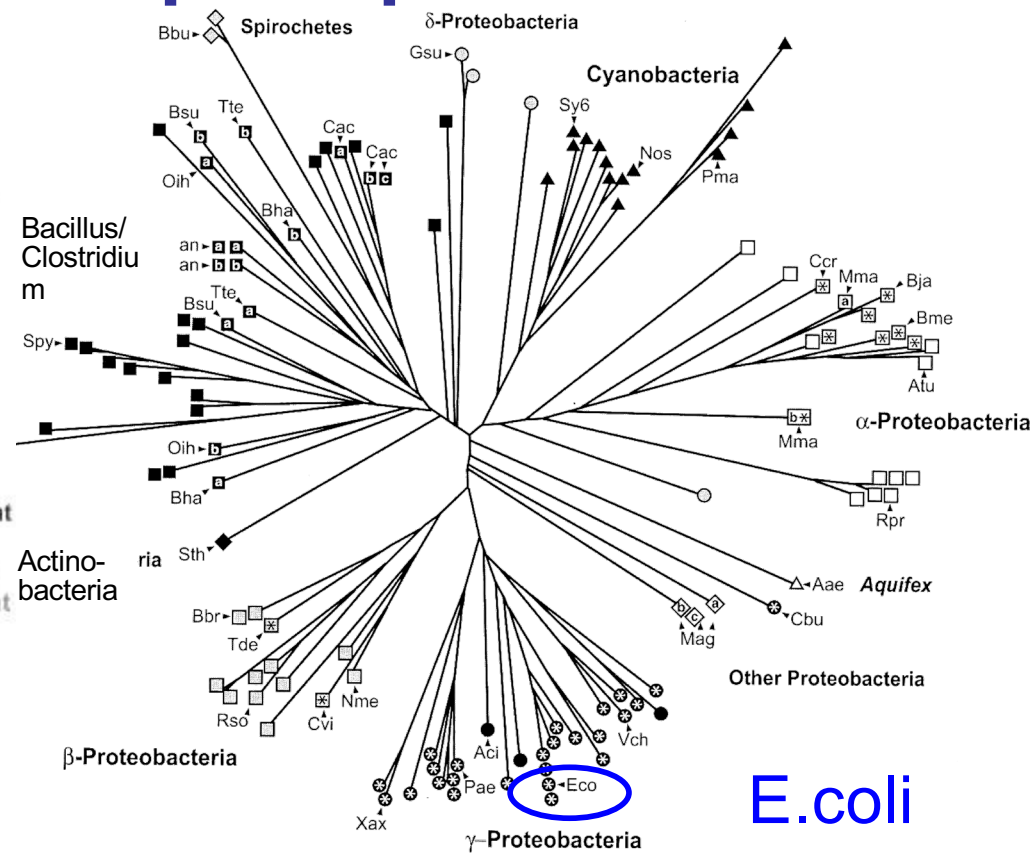
Impact of RNA homology search



6S mimics an open promoter







E.coli

Barrick et al. *RNA* 2005
Trotochaud et al. *NSMB* 2005
Willkomm et al. NAR 2005

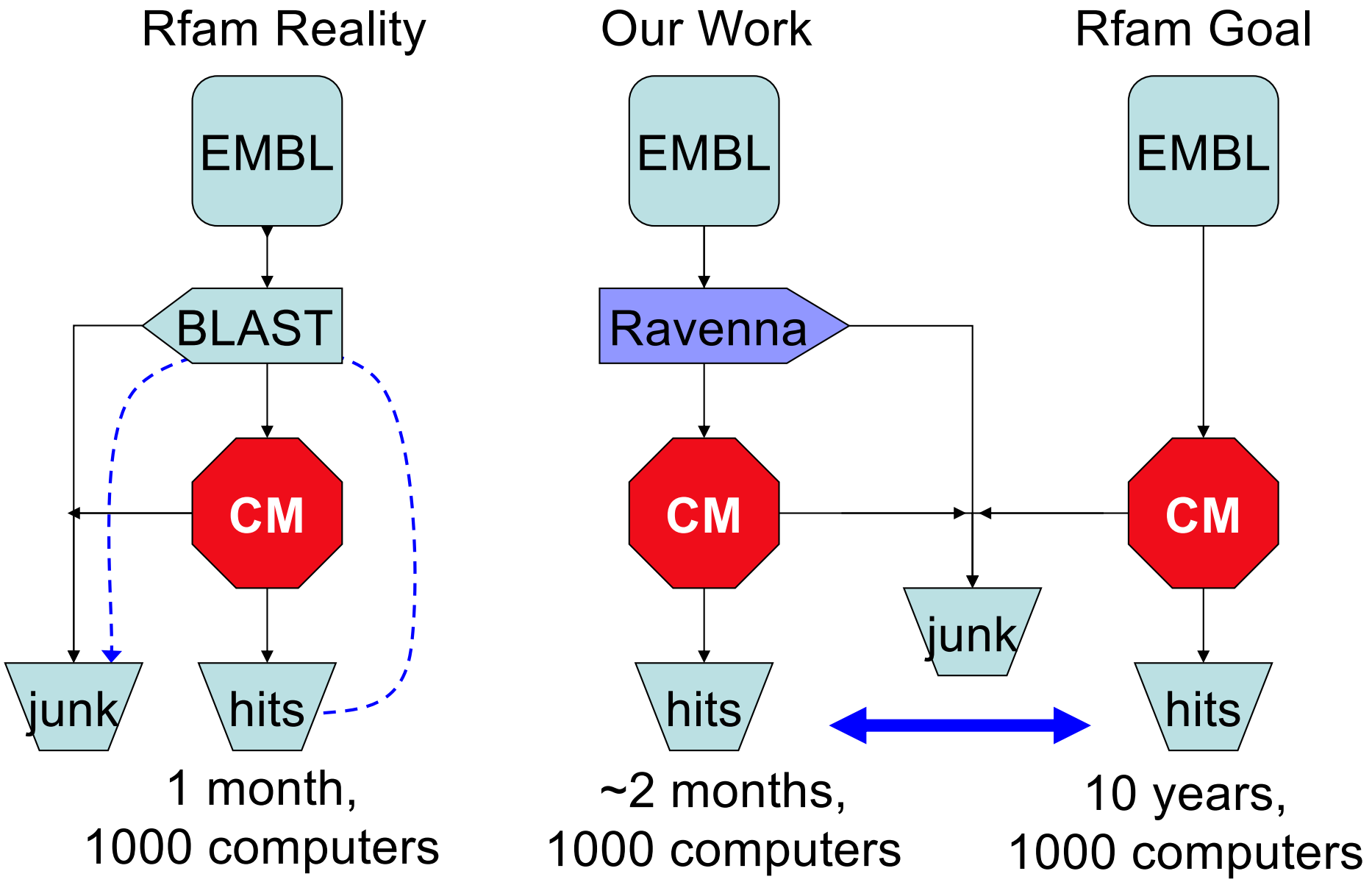
Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg

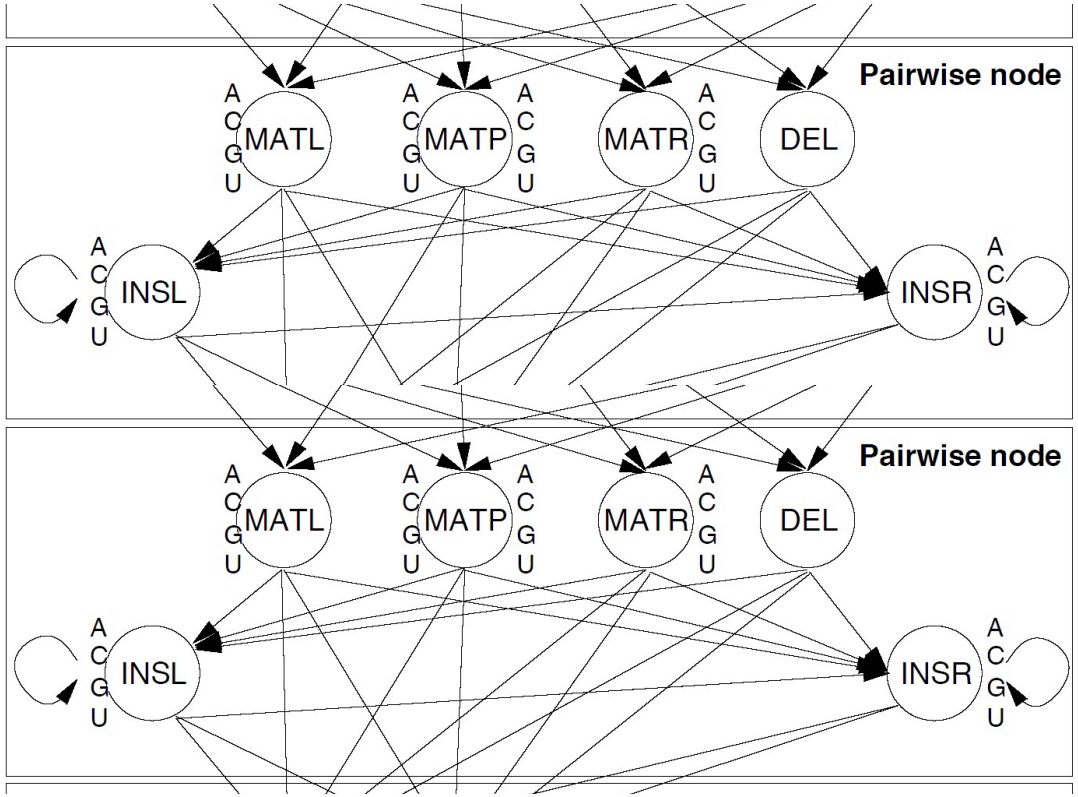
& W.L. Ruzzo

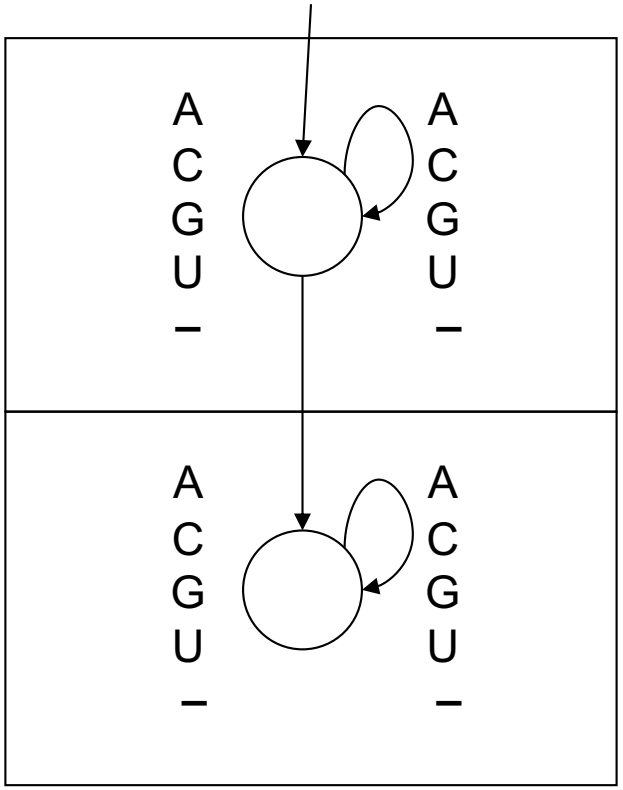
Recomb ‘04, ISMB ‘04, Bioinfo ‘06

CM's are good, but slow

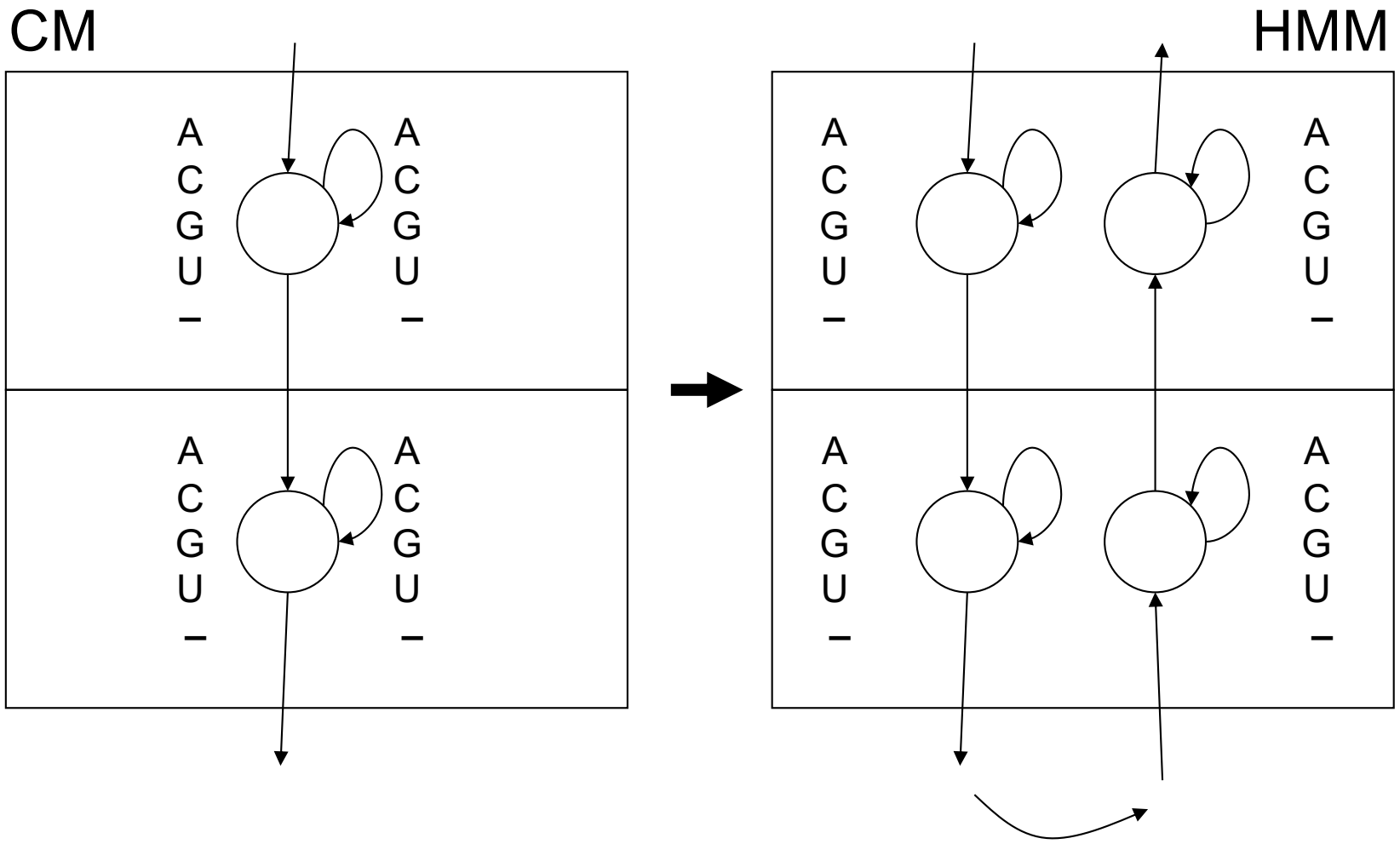


Oversimplified CM
(for pedagogical purposes only)





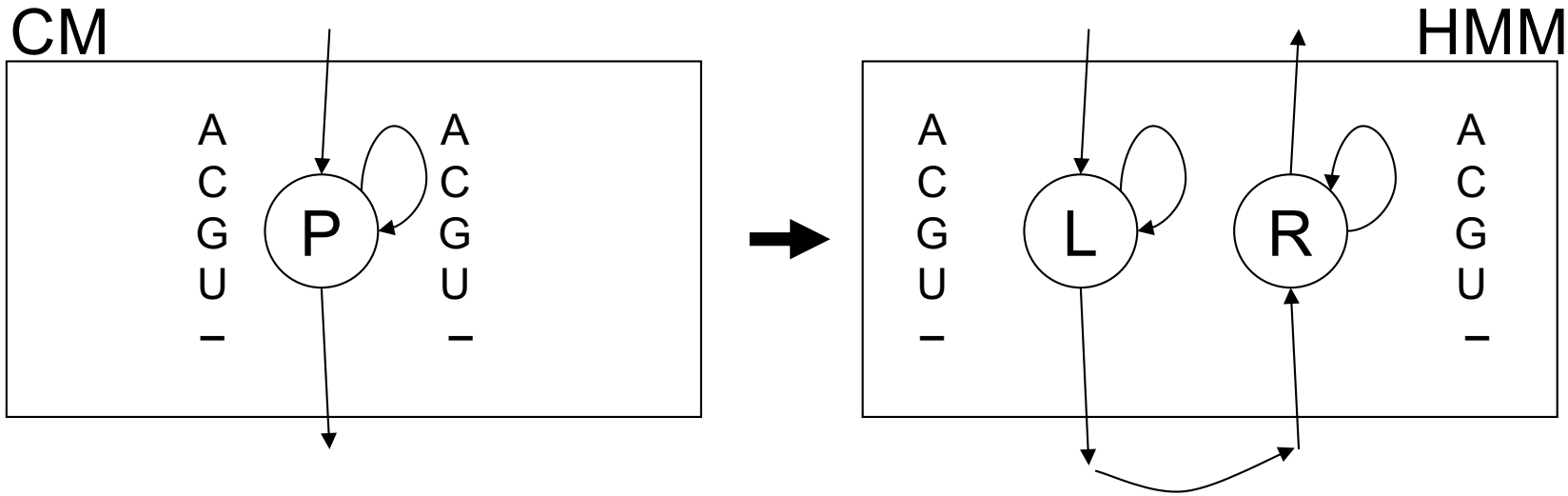
CM to HMM



25 emisions per state

5 emissions per state, 2x states

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

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

NB: HMM not a prob. model

Rigorous Filtering

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

Any scores satisfying the linear inequalities give rigorous filtering

Proof:

CM Viterbi path score

≤ "corresponding" HMM path score

≤ Viterbi HMM path score

(even if it does not correspond to *any* CM path)

Some scores filter better

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

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

Option 1:

$L_U = R_A = R_G = 2$

Option 2:

$L_U = 0, R_A = 1, R_G = 4$

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

Expected Score

Minimizing $E(L_i, R_i)$
(subject to linear constraints)

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:

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

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

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

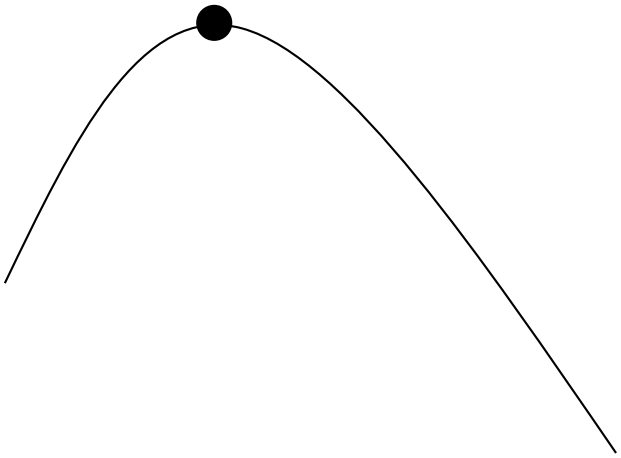
(but better ways, often)

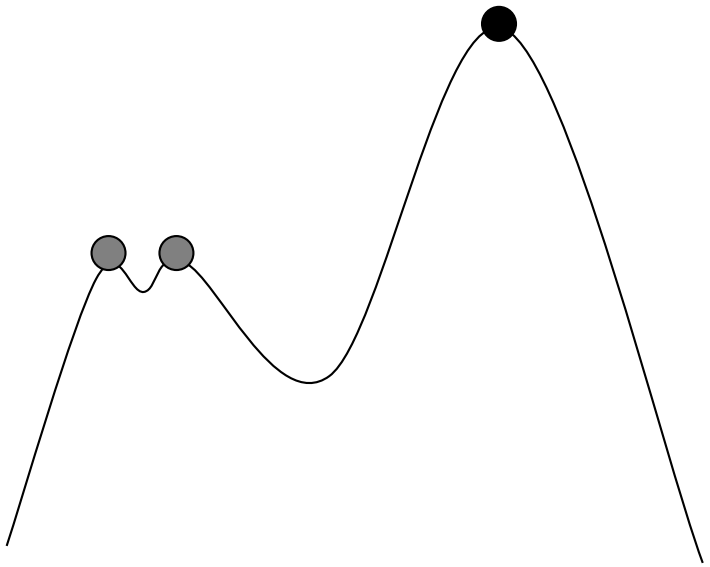
Nonconvex:

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

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

Motif Discovery

RNA Motif Discovery

- CM's are great, but where do they come from?
- Key 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.

Approaches

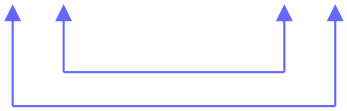
Align-First: Align sequences, then look for common structure

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

Joint: Do both together

"Align First" Approach: Predict Struct from Multiple Alignment

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



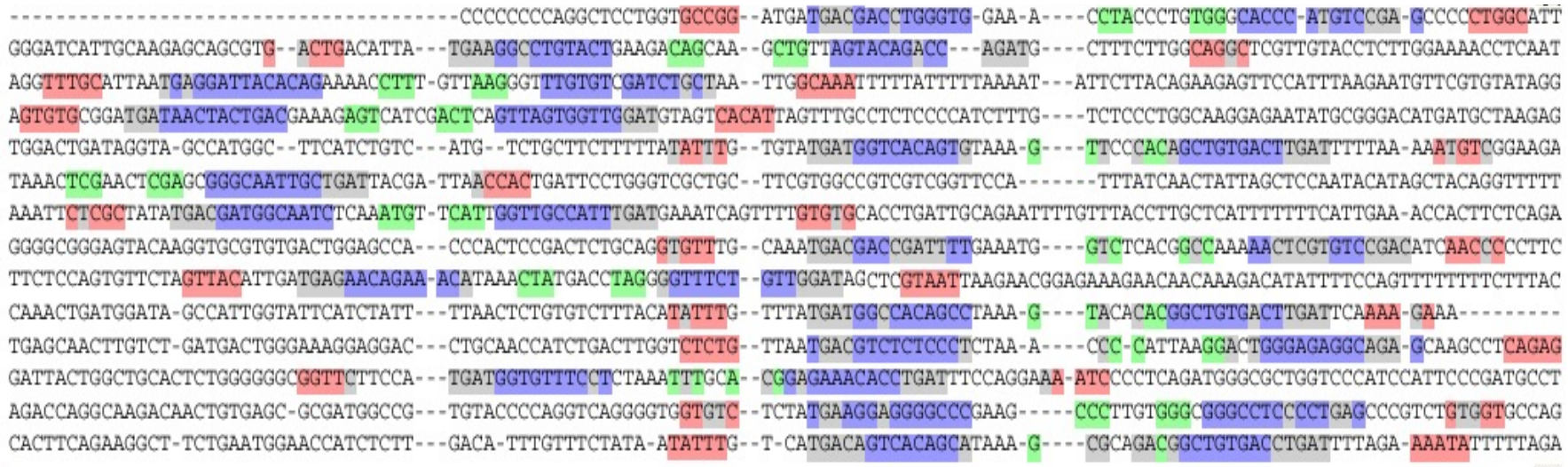
Compensatory mutations reveal structure (core of "comparative sequence analysis") *but* usual alignment algorithms penalize them (twice)

Pitfall for sequence-alignment-first approach

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions



same-colored boxes *should* be aligned

Approaches

Align-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

Our Approach: CMfinder

RNA motifs from unaligned sequences

Simultaneous *local* alignment, folding and CM-based motif description via an EM-style learning procedure

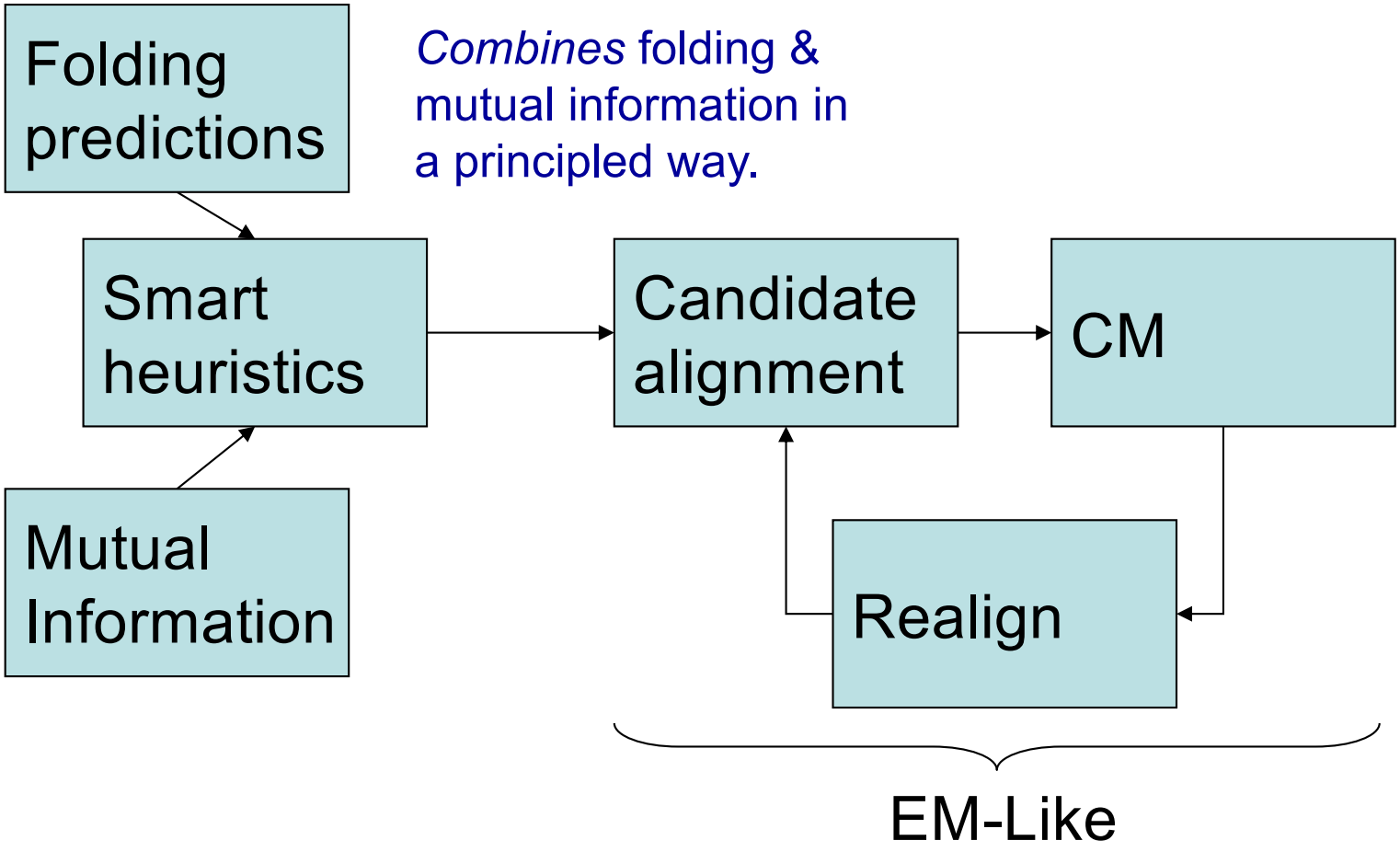
- Sequence conservation exploited, but not required
- Robust to inclusion of unrelated and/or flanking sequence
- Reasonably fast and scalable
- Produces a probabilistic model of the motif that can be directly used for homolog search

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

CMFinder

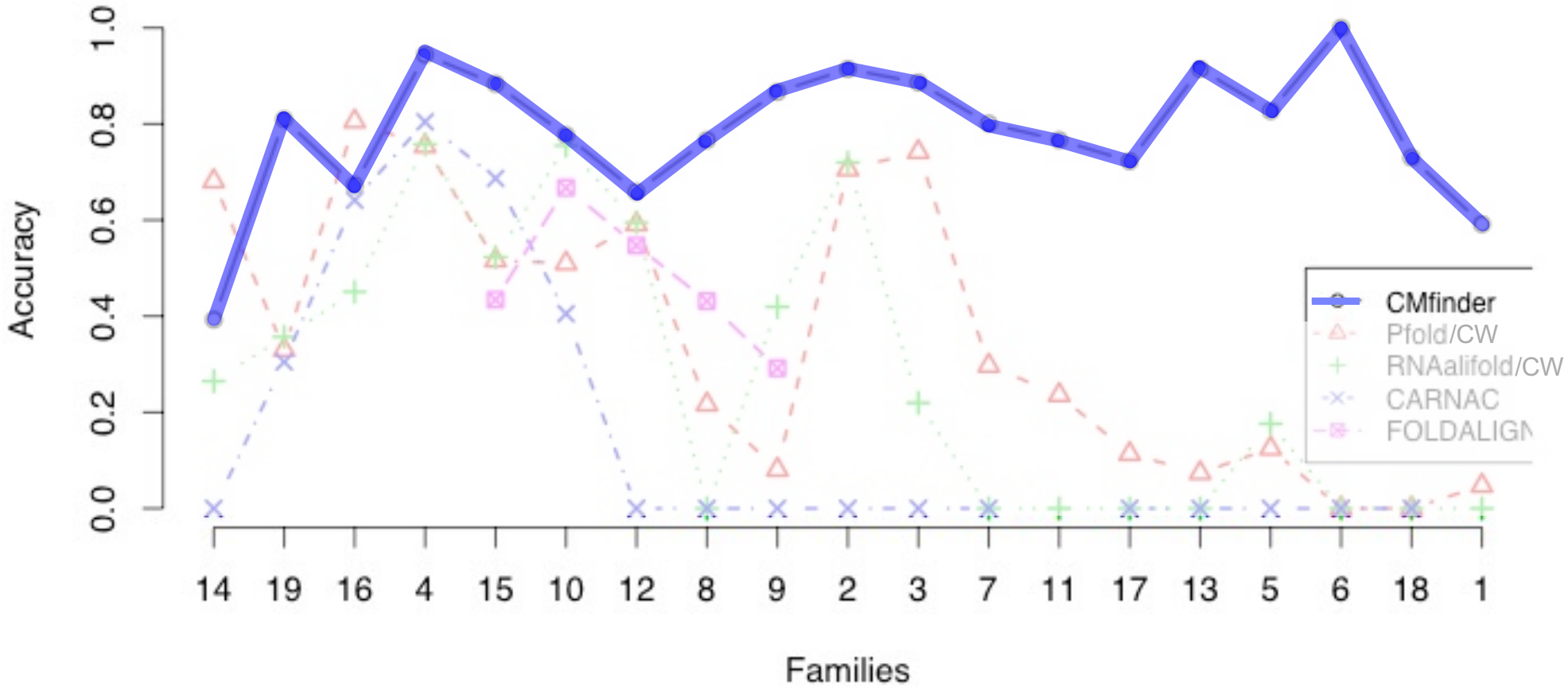
Simultaneous alignment, folding & motif description

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Discovery in Bacteria

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

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

Zizhen Yao[1*], Jeffrey Barrick[2¤], Zasha Weinberg[3], Shane Neph[1,4], Ronald Breaker[2,3,5], Martin Tompa[1,4], Walter L. Ruzzo[1,4]

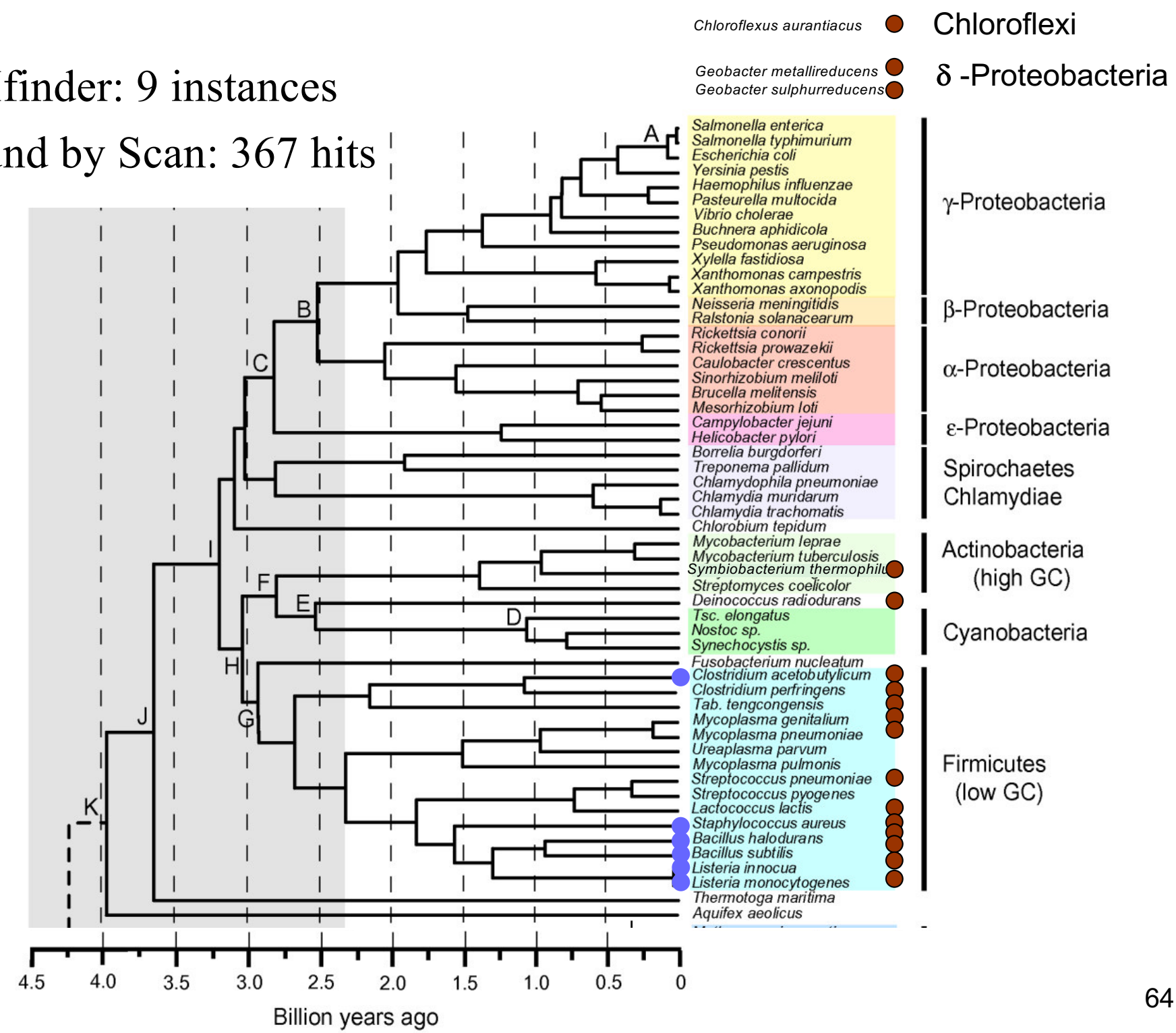
Published online 9 July 2007

Nucleic Acids Research, 2007, Vol. 35, No. 14 4809–4819
doi:10.1093/nar/gkm487

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

Zasha Weinberg[1,*], Jeffrey E. Barrick[2,3], Zizhen Yao[4], Adam Roth[2], Jane N. Kim[1], Jeremy Gore[1], Joy Xin Wang[1,2], Elaine R. Lee[1], Kirsten F. Block[1], Narasimhan Sudarsan[1], Shane Neph[5], Martin Tompa[4,5], Walter L. Ruzzo[4,5] and Ronald R. Breaker[1,2,3]

- CMfinder: 9 instances
- Found by Scan: 367 hits



Overall Pipeline & Processing Times

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

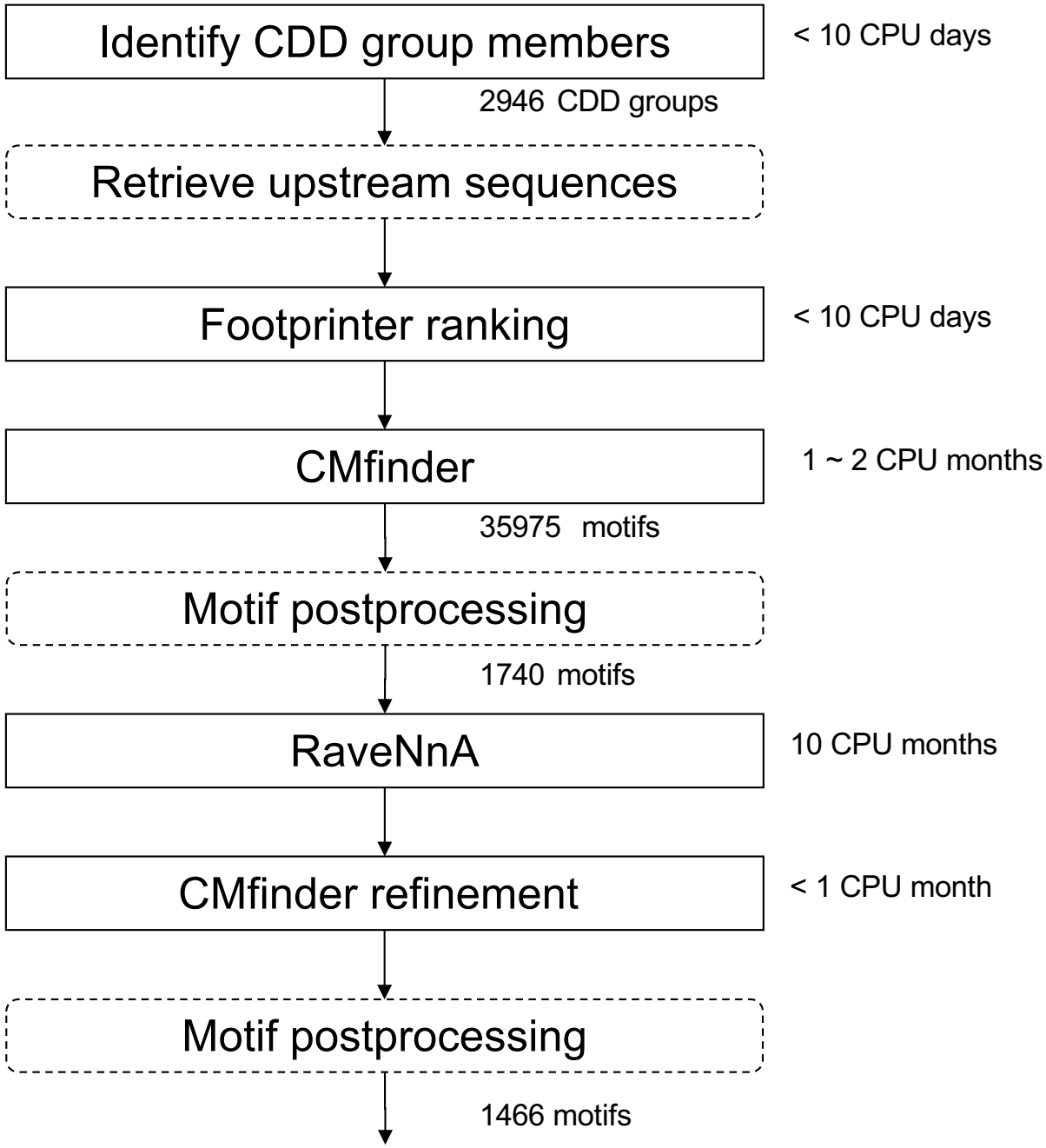


Table 1: Motifs that correspond to Rfam families

Rank			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

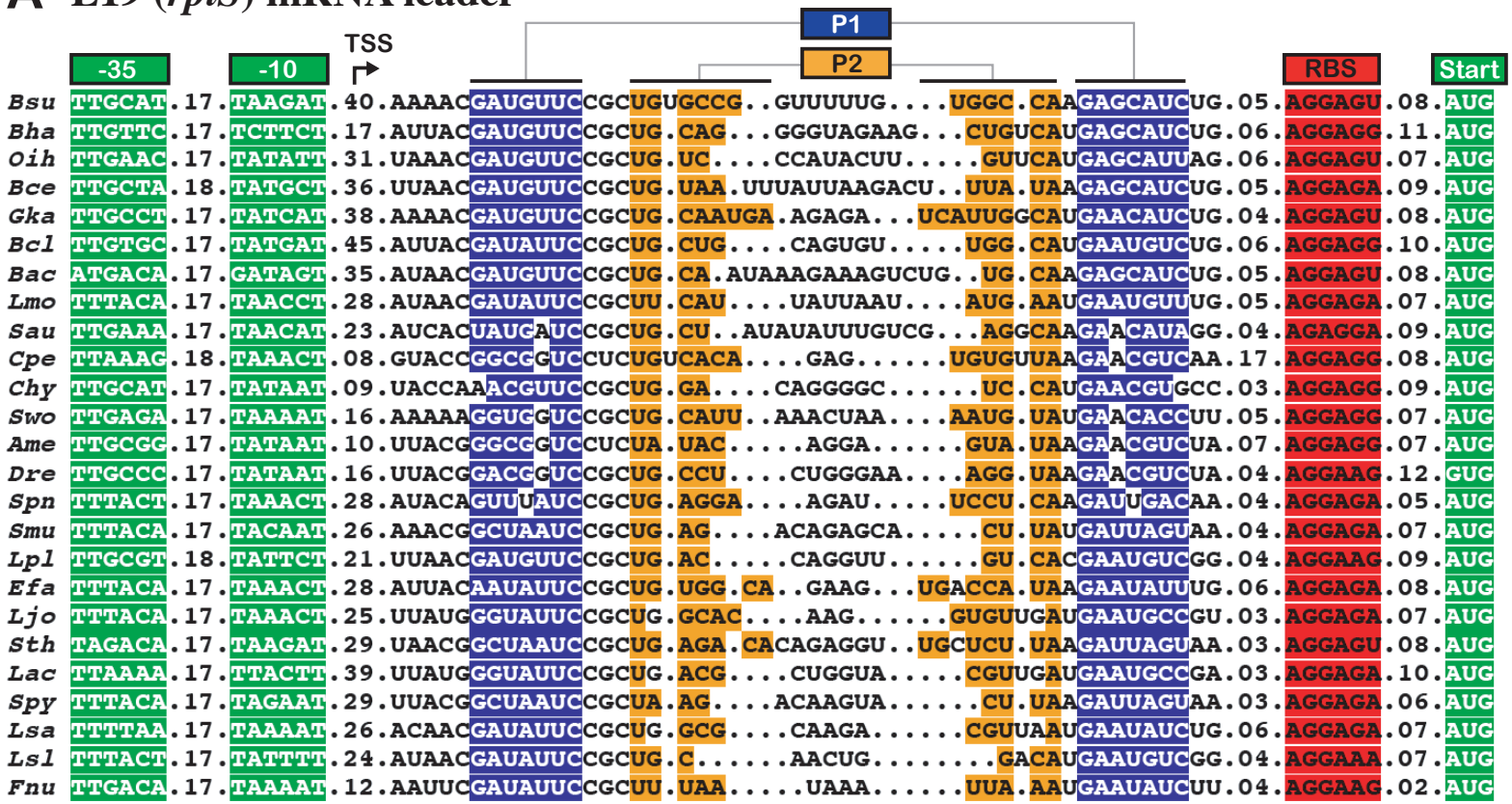
Table 1: Motifs that correspond to Rfam families. "Rank": the three columns show ranks for refined motif clusters after genome scans ("RAV"), CMfinder motifs before genome scans ("CMF"), and FootPrinter results ("FP"). We used the same ranking scheme for RAV and CMF. "Score"

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

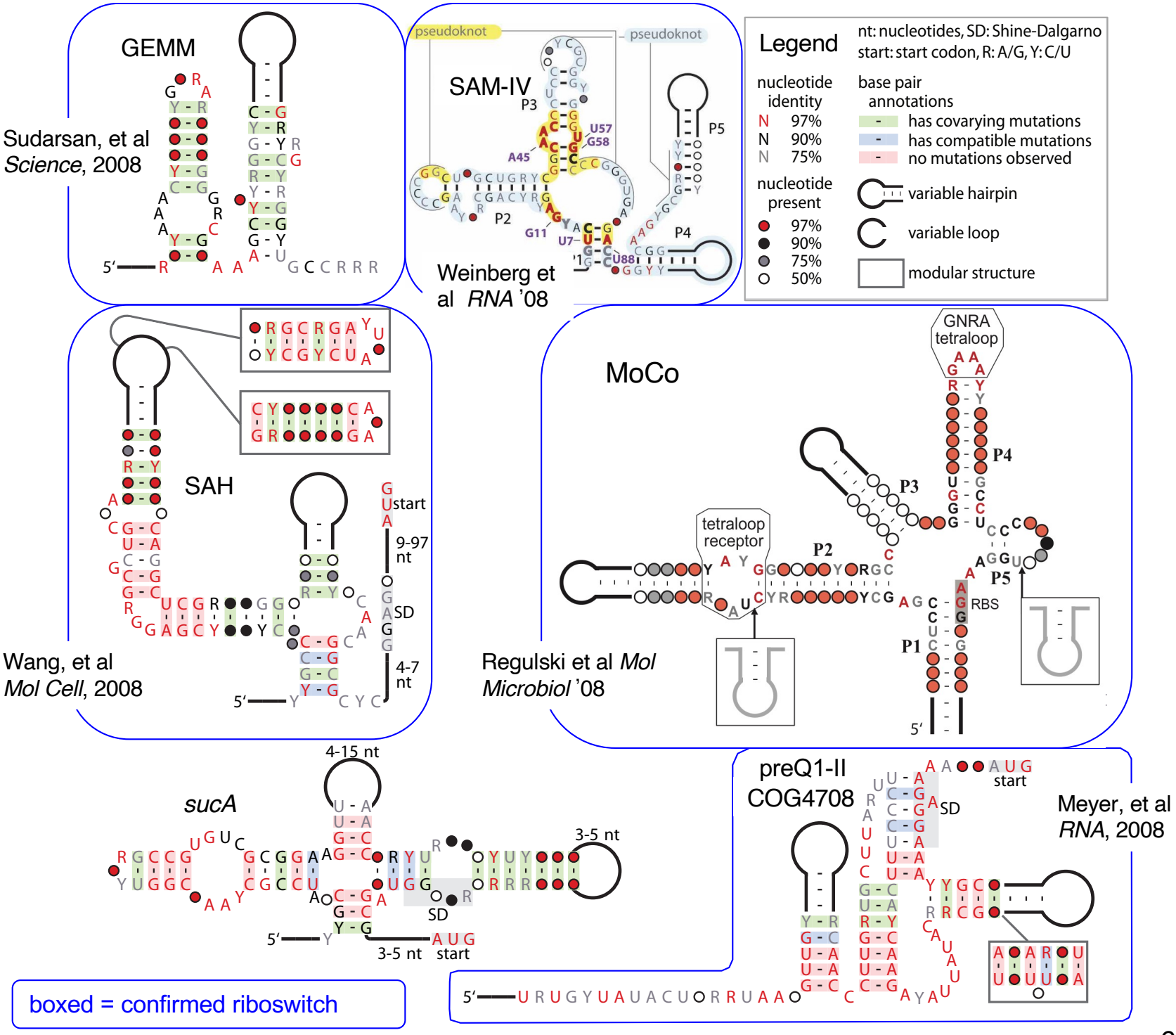
Tbl 2: Prediction accuracy compared to prokaryotic subset of Rfam full alignments.
Membership: # of seqs in overlap between our predictions and Rfam's, the sensitivity (Sn) and specificity (Sp) of our membership predictions. Overlap: the avg len of overlap between our predictions and Rfam's (nt), the 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%.

Example: Ribosomal Autoregulation:

Excess L19 represses L19 (RF00556; 555-559 similar)



Examples: 6 (of 22) Representative motifs



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

Vertebrate ncRNAs

Some Results

Human Predictions

Evofold

S Pedersen, G Bejerano, A Siepel, K Rosenbloom, K Lindblad-Toh, ES Lander, J Kent, W Miller, D Haussler, "Identification and classification of conserved RNA secondary structures in the human genome." PLoS Comput. Biol., 2, #4 (2006) e33.

48,479 candidates (~70% FDR?)

FOLDALIGN

E Torarinsson, M Sawera, JH Havgaard, M Fredholm, J Gorodkin, "Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure." *Genome Res.*, 16, #7 (2006) 885-9.

1800 candidates from 36978 (of 100,000) pairs

RNAz

S Washietl, IL Hofacker, M Lukasser, A Hüttenhofer, PF Stadler, "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." Nat. Biotechnol., 23, #11 (2005) 1383-90.

30,000 structured RNA elements

1,000 conserved across all vertebrates.

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

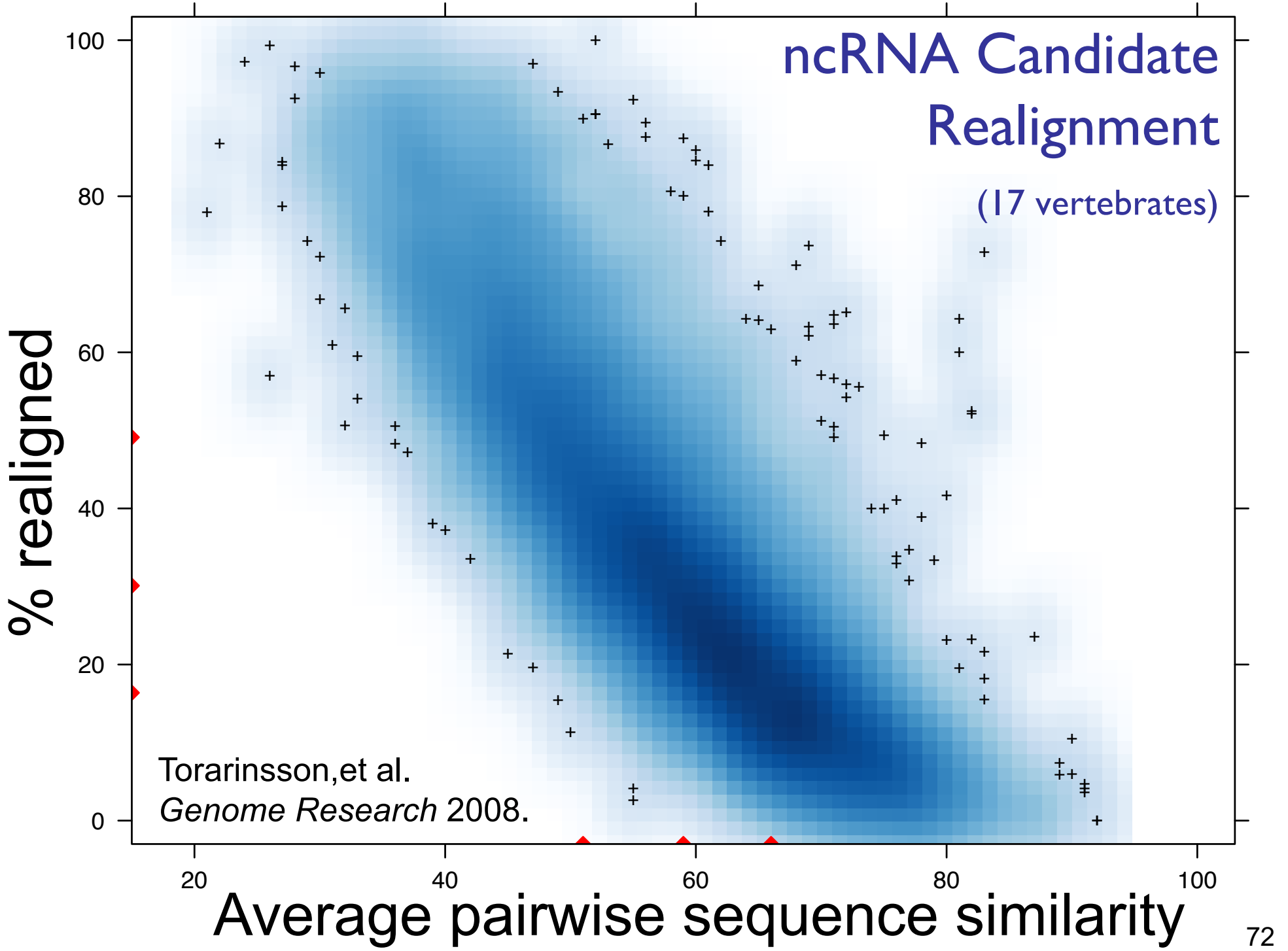
~1/2 located far from any known gene

CMfinder

Torarinsson, Yao, Wiklund, Bramsen, Hansen, Kjems, Tommerup, Ruzzo and Gorodkin. Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions. *Genome Research*, Feb 2008, 18(2):242-251 PMID: 18096747

Seemann, Mirza, Hansen, Bang-Berthelsen, Garde, Christensen-Dalsgaard, Torarinsson, Yao, Workman, Pociot, Nielsen, Tommerup, Ruzzo, Gorodkin. The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res, Aug 2017, 27(8):1371-1383 PMID: 28487280.

Thousands of Predictions


ncRNA Candidate Realignment
(17 vertebrates)
% realigned
Average pairwise sequence similarity
0
20
40
60
80
100
20
40
60
80
100
Torarinsson,et al.
Genome Research 2008.

Summary

After careful control of FDR,

Widespread structured RNA prediction

Evidence for conservation

Evidence for expression

Evidence for elevated expression of structured vs non-structured in CDS contexts

Hypothesis: cis-regulatory roles at these loci

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