RNA Search and Motif Discovery

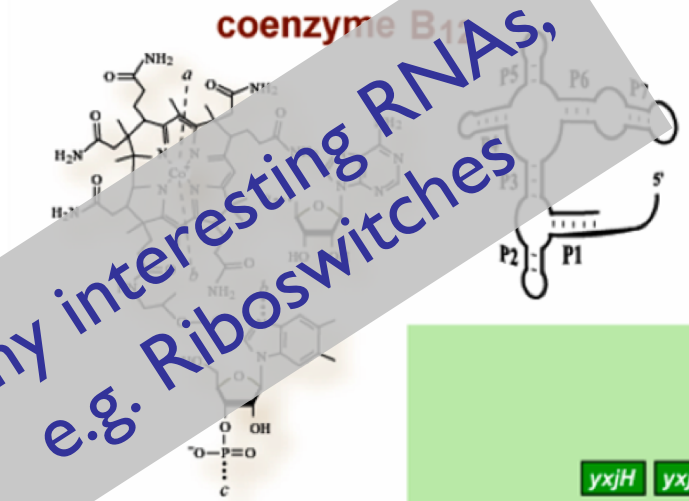
CSEP 527
Computational Biology

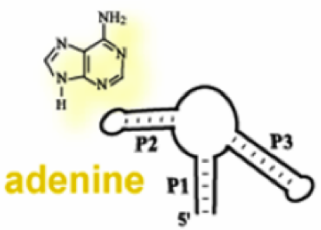
Previous Lecture

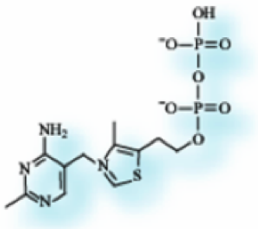
Many biologically interesting roles for RNA

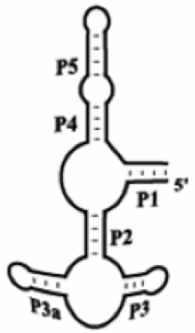
RNA secondary structure prediction

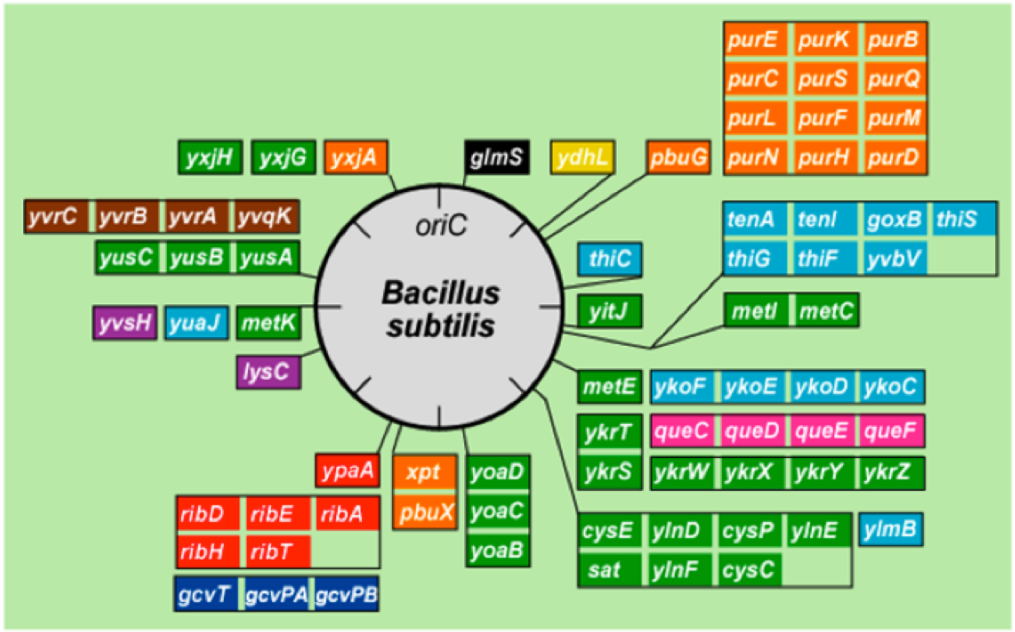
Many interesting RNAs, e.g. Riboswitches

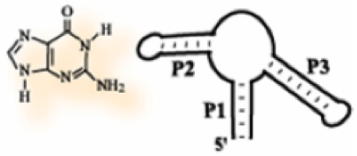


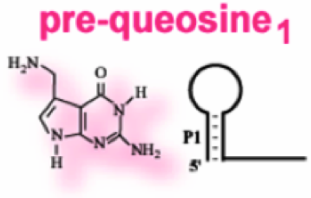


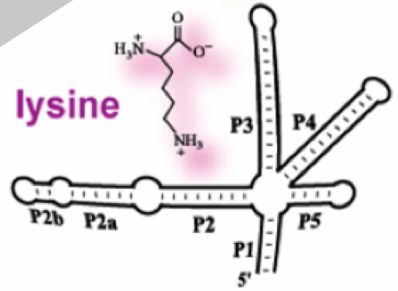


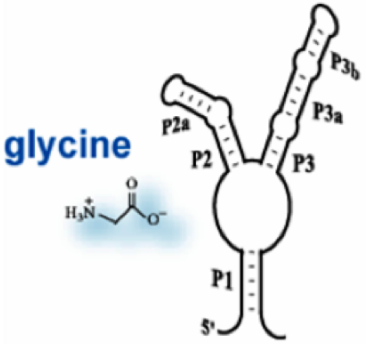


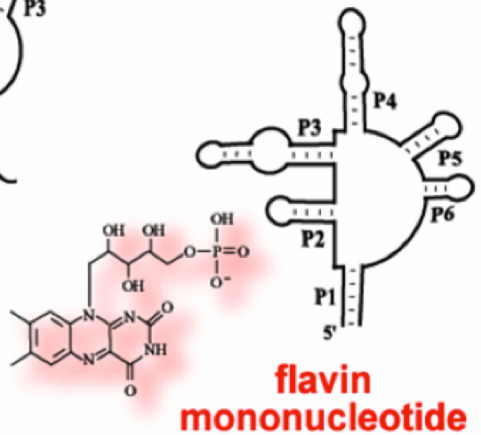




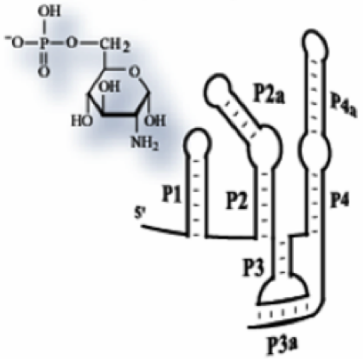


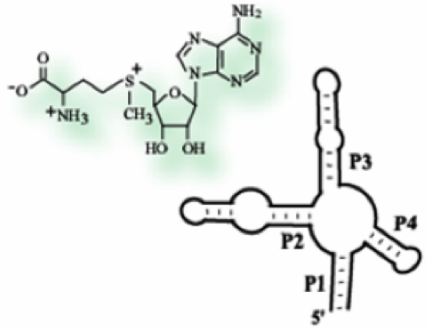






flavin mononucleotide





S-adenosyl-methionine

Structure Prediction

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

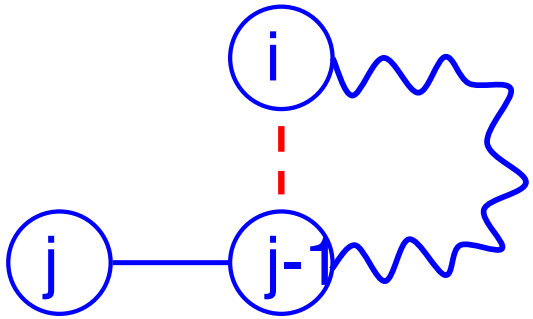
Partition Function
- + finds all folds
- - ignores pseudoknots

"Optimal pairing of r_i ... r_j"

Two possibilities

j Unpaired:

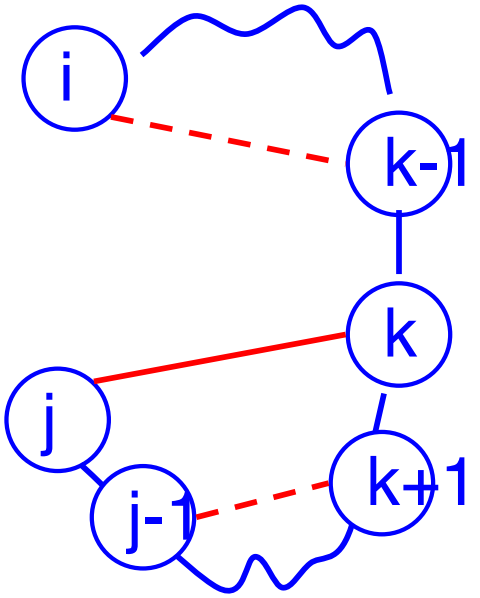
Find best pairing of r_i ... r_{j-1}



j Paired (with some k):

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

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

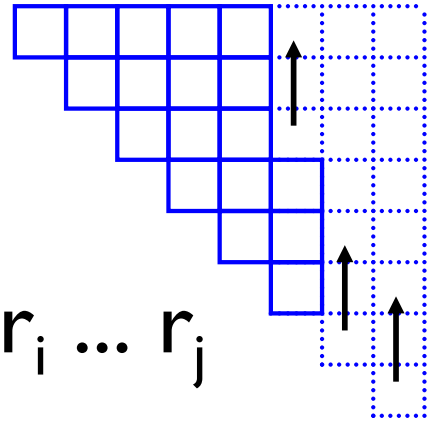


Why is it slow?

Why do pseudoknots matter?

Structure Prediction

Nussinov: Computation Order



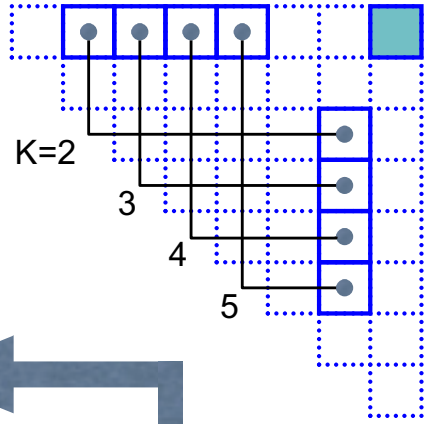
Or -energy

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: $O(n^3)$

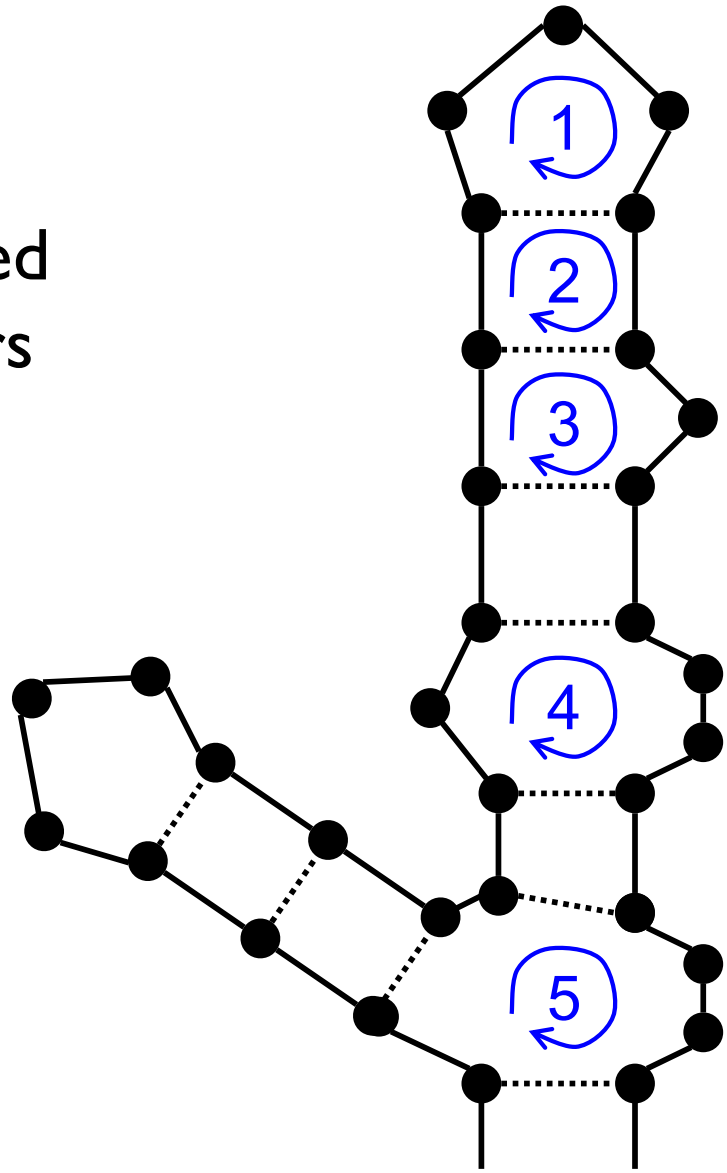
Loop-based energy version is better; recurrences similar, slightly messier

Loop-based Energy Minimization

Detailed experiments show it's more accurate to model based on *loops*, rather than just pairs

Loop types

1. Hairpin loop
2. Stack
3. Bulge
4. Interior loop
5. Multiloop



Single Seq Prediction Accuracy

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

Estimates suggest ~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

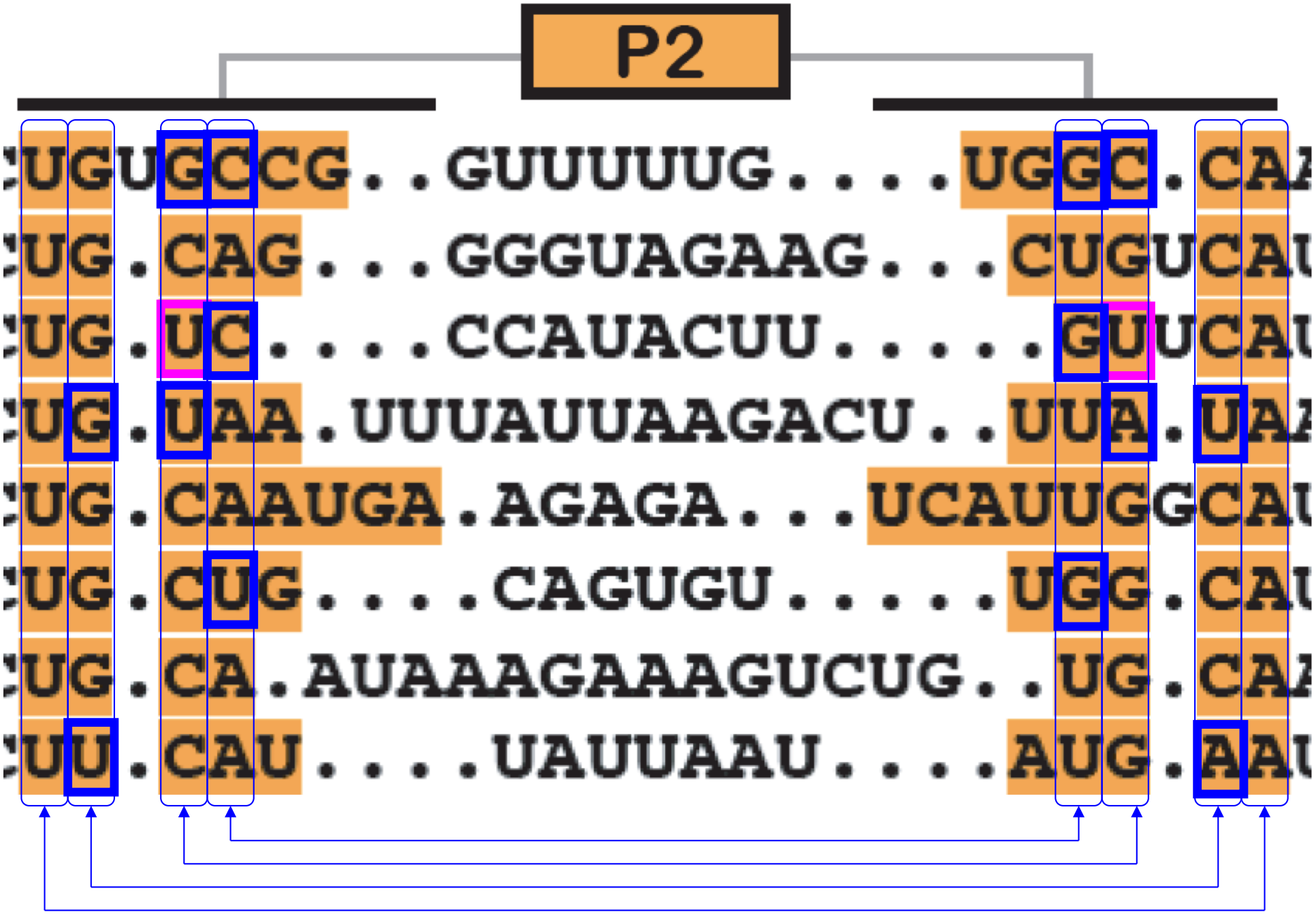
+ handles all pairings (potentially incl. pseudoknots)
- requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars

Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

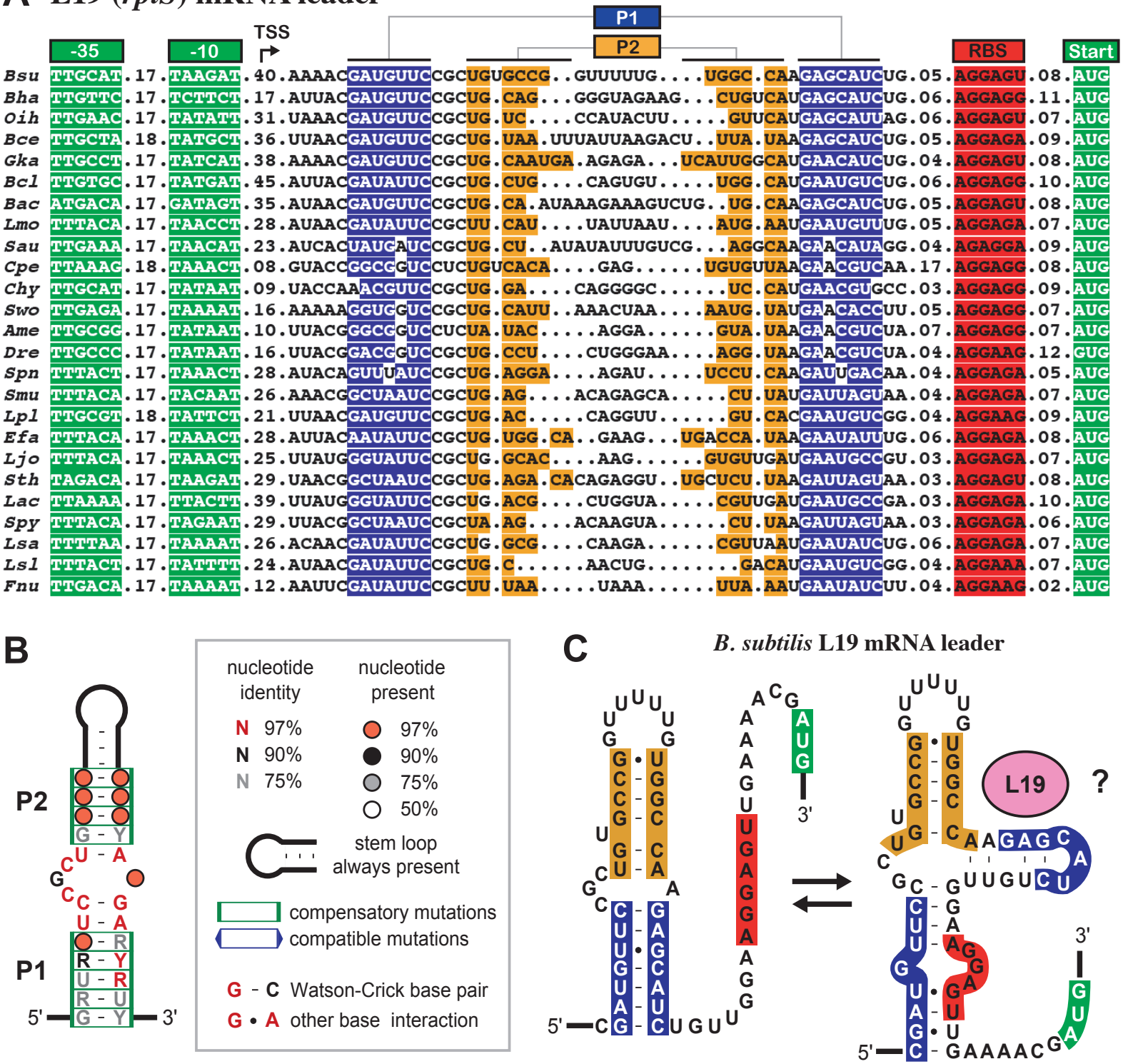
Today



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)

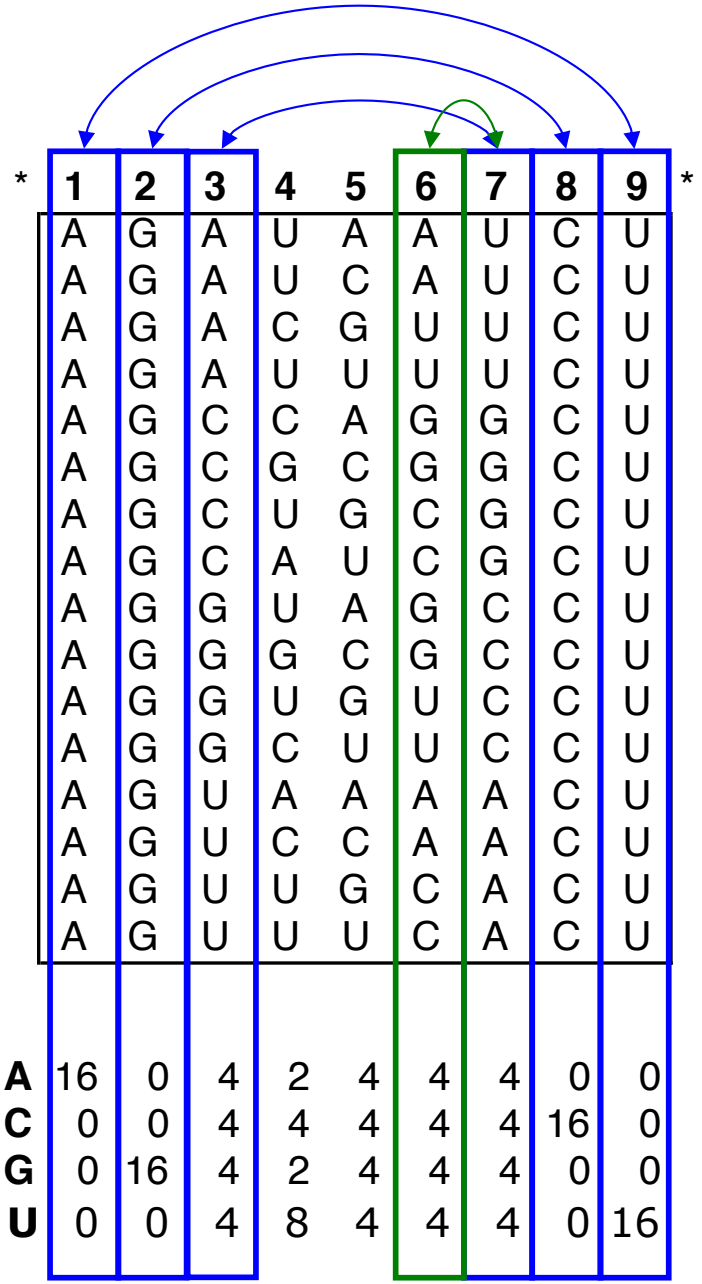
Max when *no* seq conservation but perfect pairing

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)



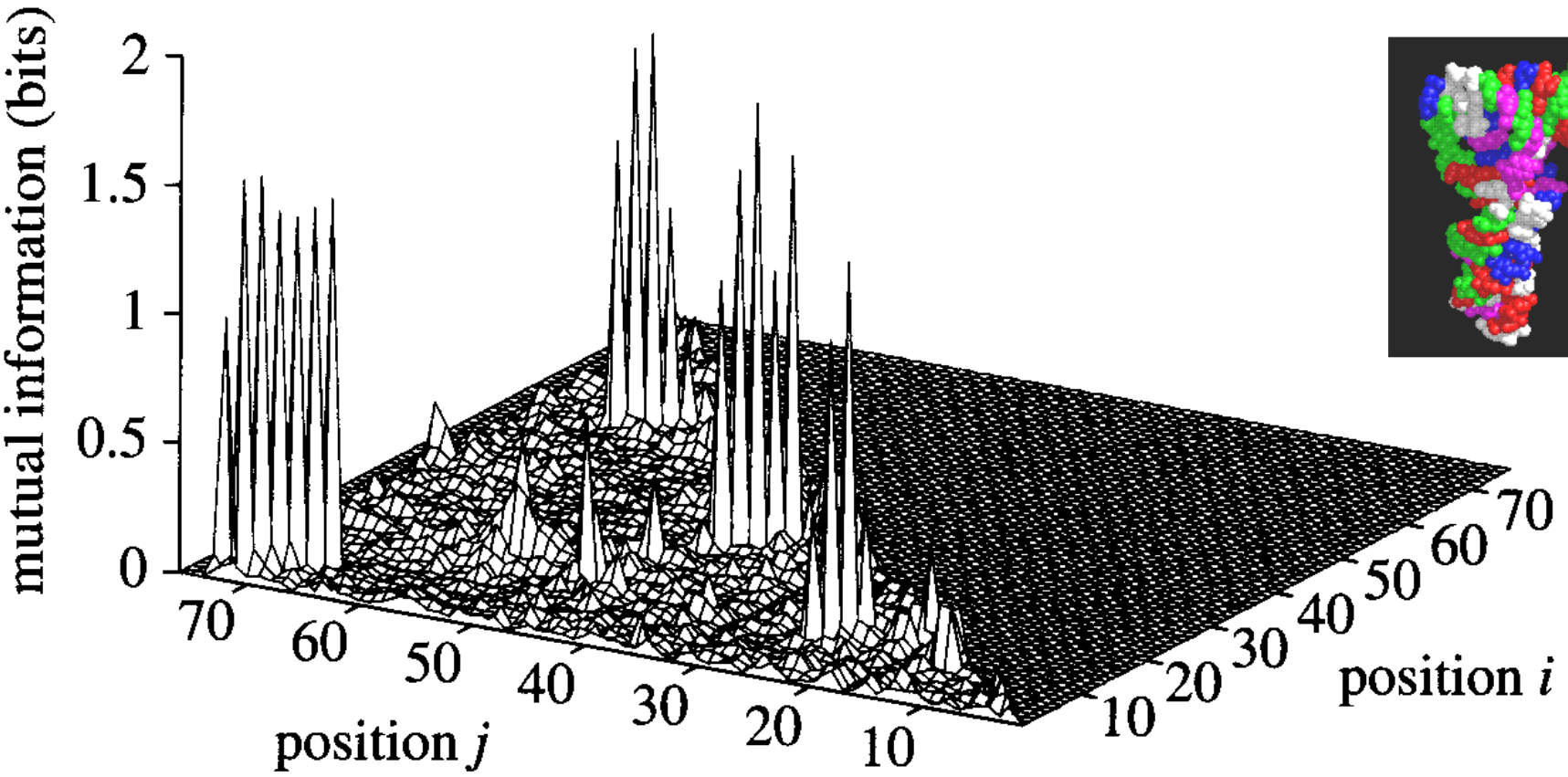
*	1	2	3	4	5	6	7	8	9	*
	A	G	A	U	A	A	U	C	U	
	A	G	A	U	C	A	U	C	U	
	A	G	A	C	G	U	U	C	U	
	A	G	A	U	U	U	U	C	U	
	A	G	C	C	A	G	G	C	U	
	A	G	C	G	C	G	G	C	U	
	A	G	C	U	G	C	G	C	U	
	A	G	C	A	U	C	G	C	U	
	A	G	G	U	A	G	C	C	U	
	A	G	G	G	C	G	C	C	U	
	A	G	G	U	G	U	C	C	U	
	A	G	G	C	U	U	C	C	U	
	A	G	U	A	A	A	A	C	U	
	A	G	U	C	C	A	A	C	U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

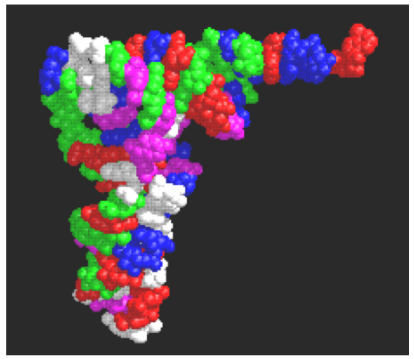
MI:	1	2	3	4	5	6	7	8	9
9	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	0		
7	0	0	**2**	**0.30**	**0**	**1**			
6	0	0	**1**	**0.55**	**1**				
5	0	0	**0**	**0.42**					
4	0	0	**0.30**						
3	0	0							
2	0								
1									

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.





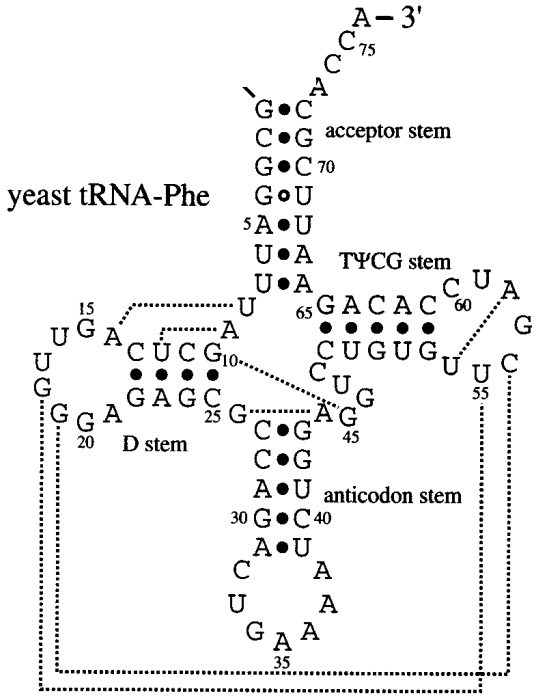


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary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

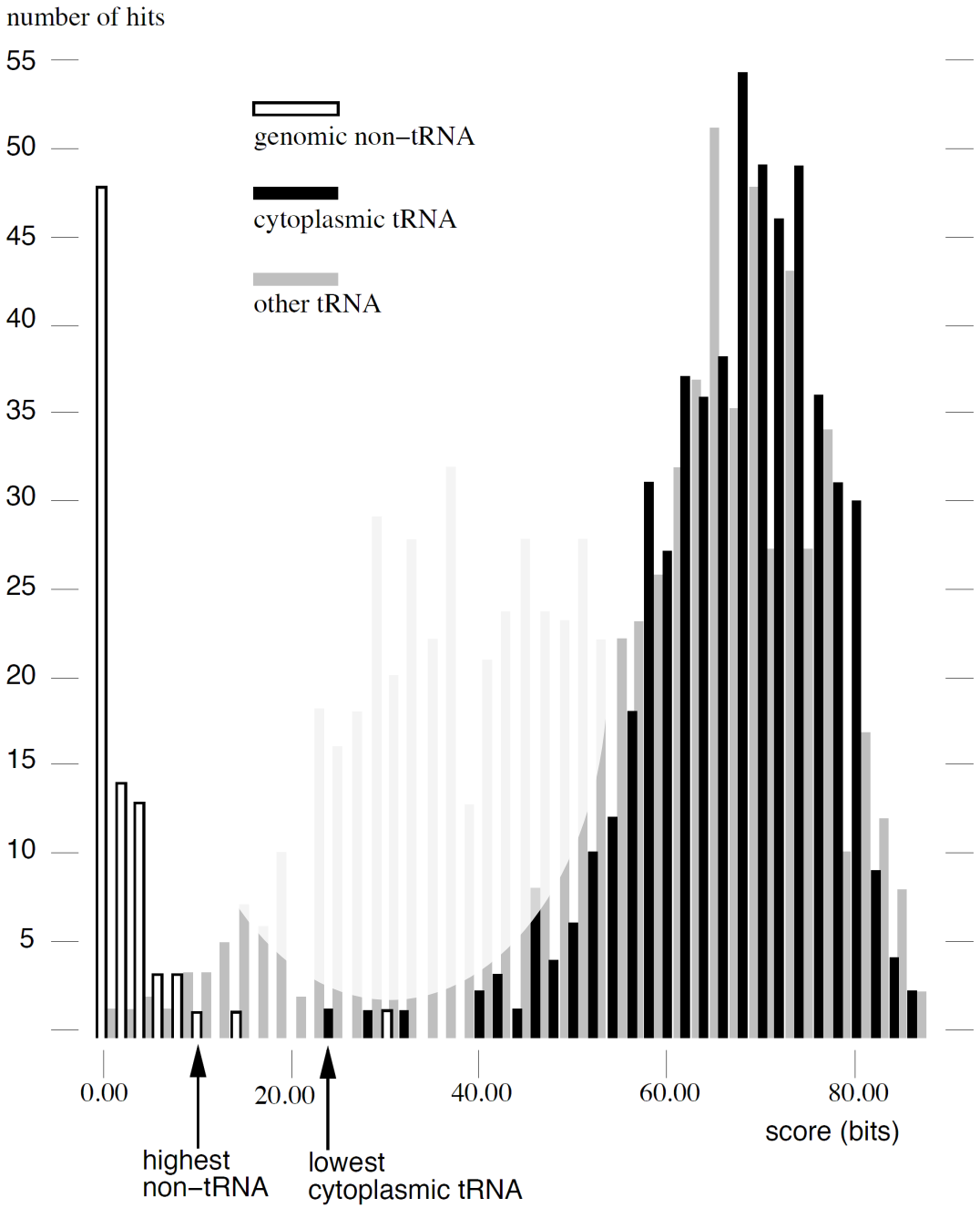
As with HMMs, given a sequence:

- 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

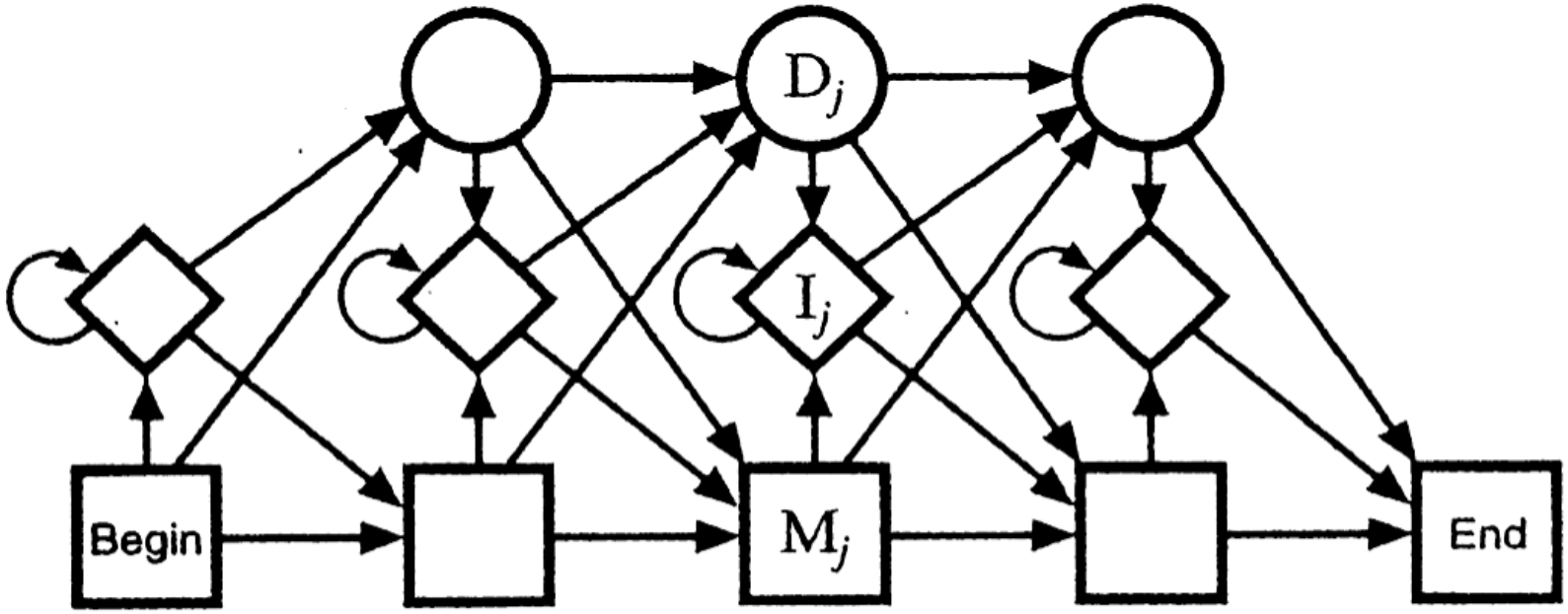


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

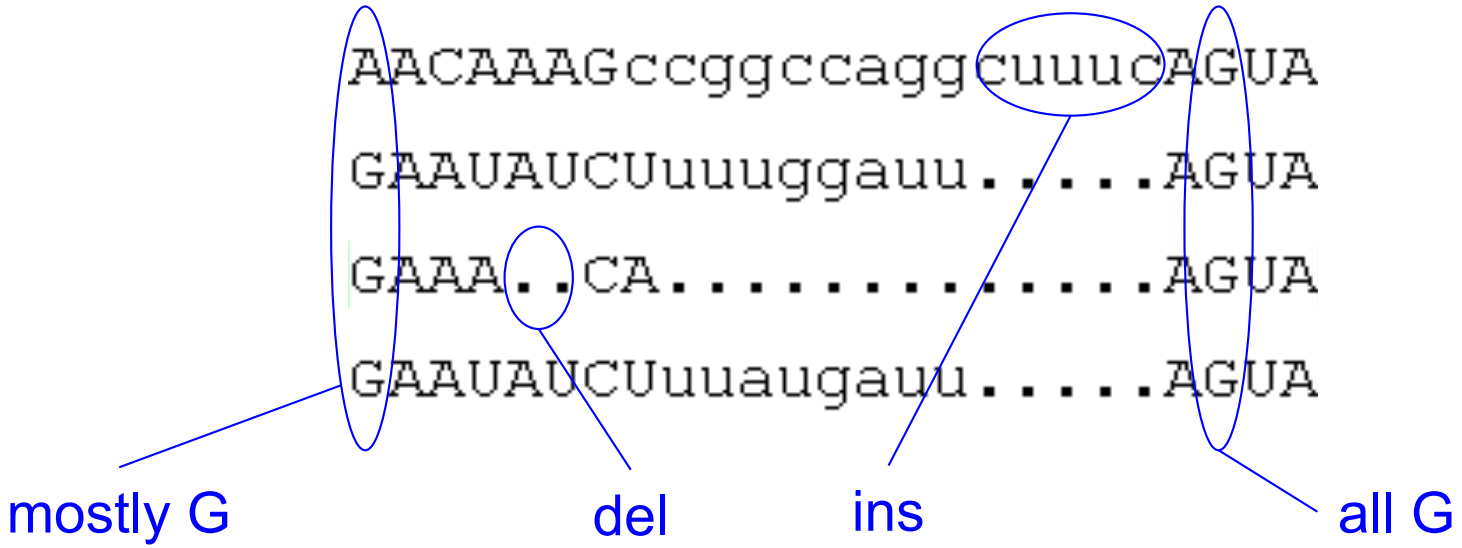
D_j: Delete states (silent - no emission)

How to model an RNA "Motif"?

Conceptually, start with a profile HMM:

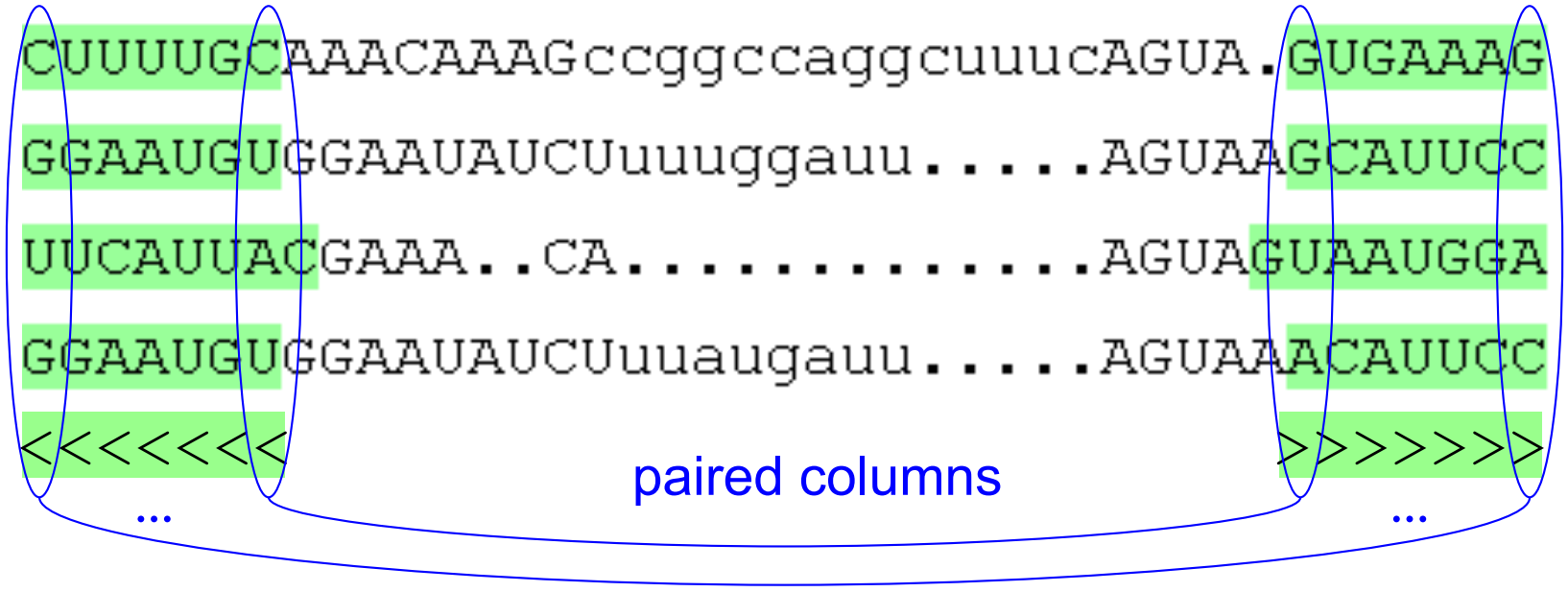
from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position

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

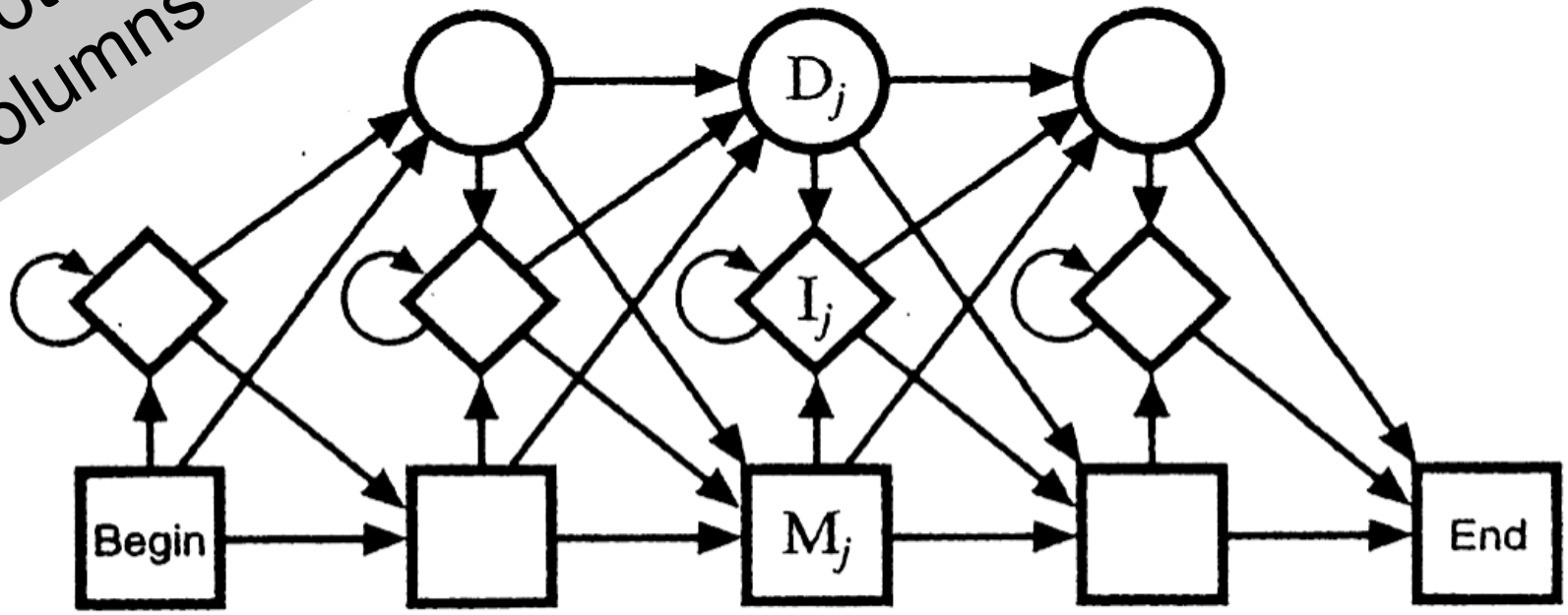


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

D_j: Delete states (silent - no emission)

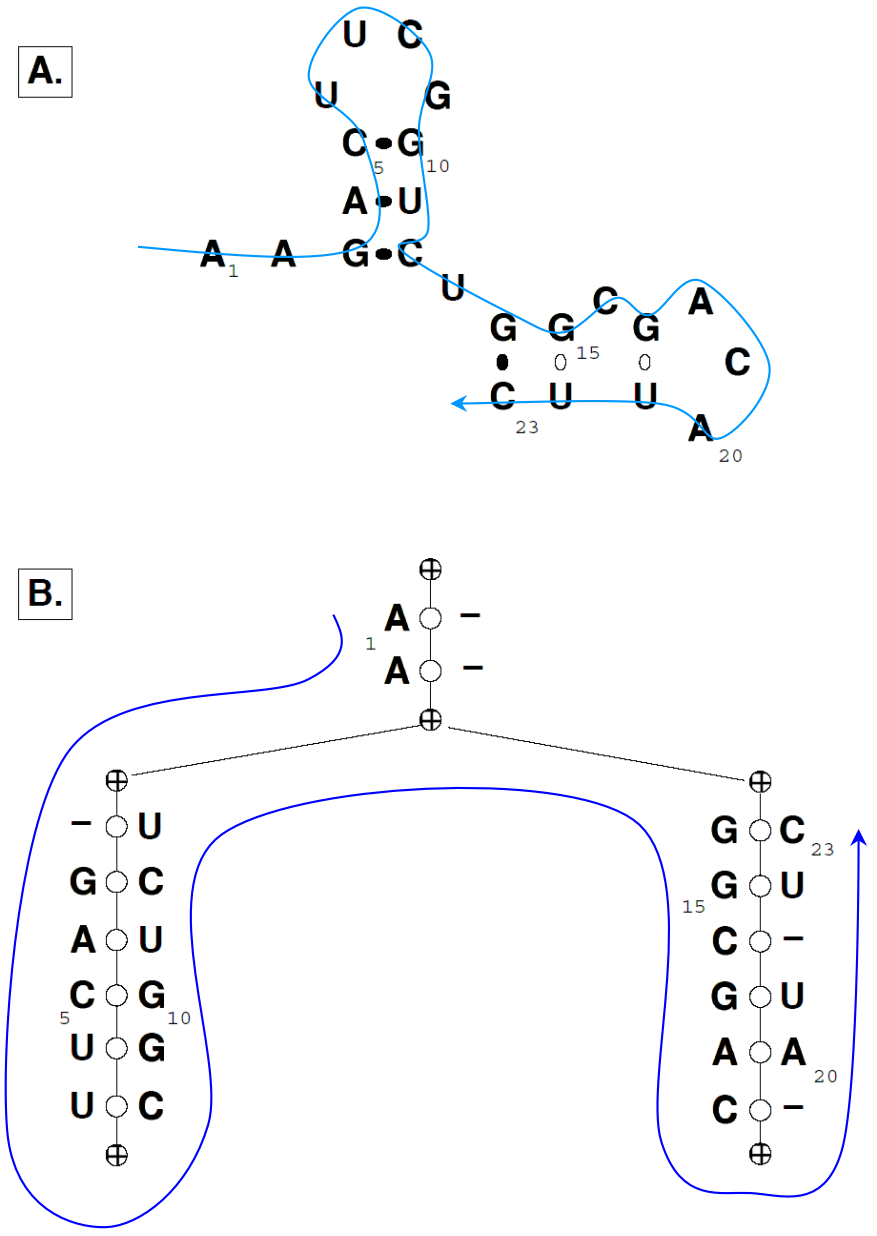
CM Structure

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)



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

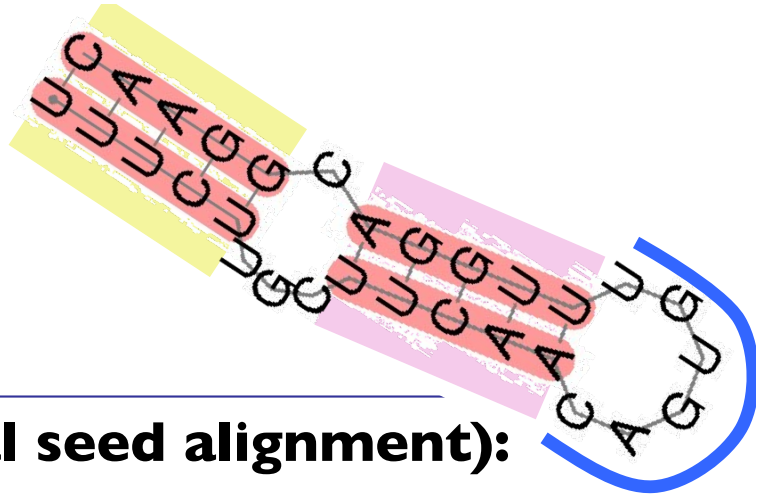
disallowing / allowing pseudoknots

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

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

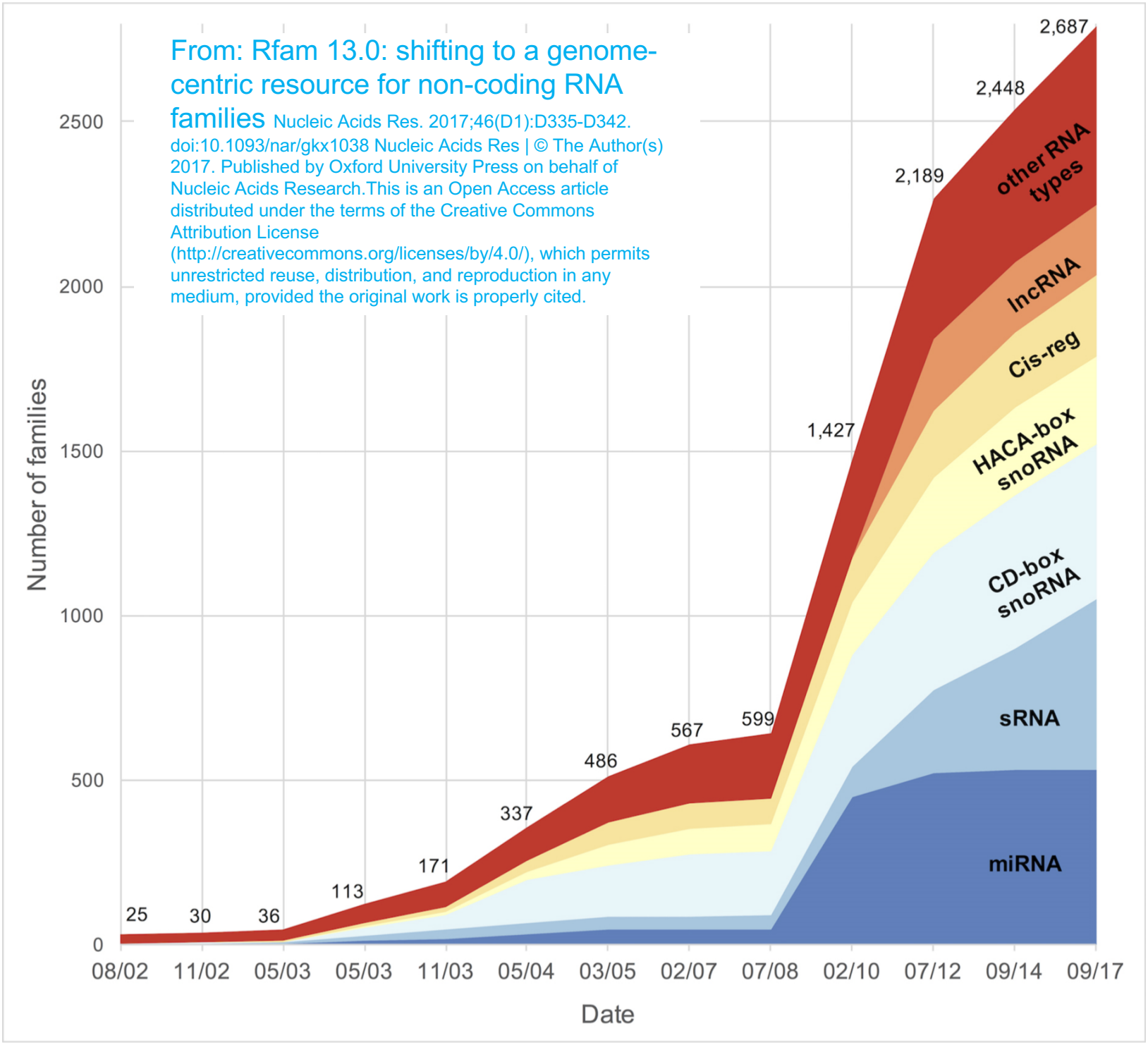
Rfam – an RNA family DB

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	


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 Nucleic Acids Res | © The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Number of families
2500
2000
1500
1000
500
0
Date
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

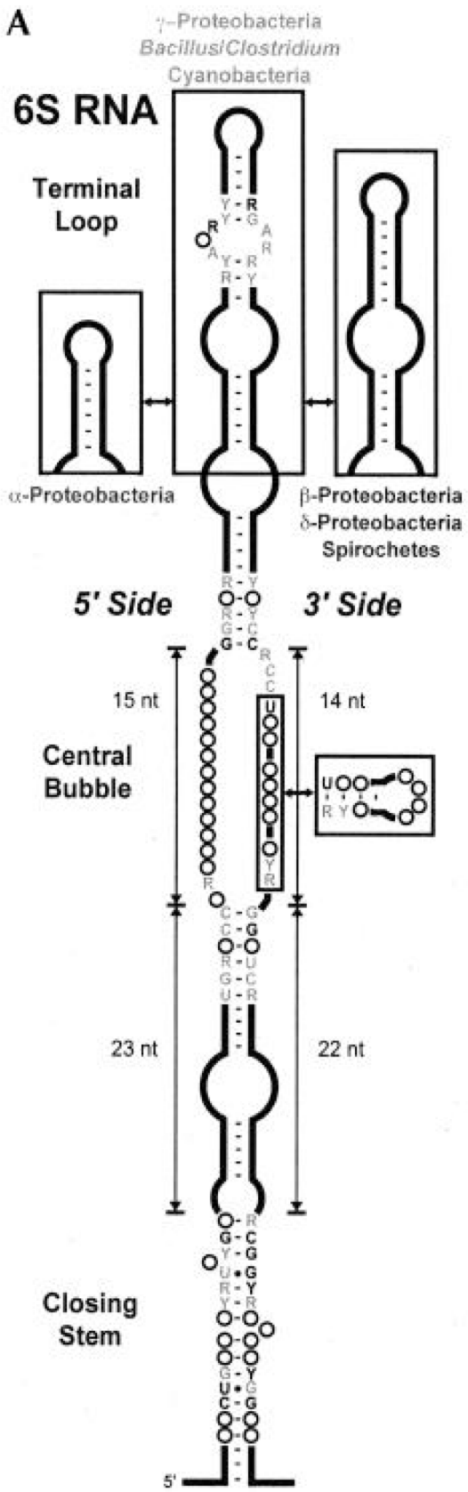
"Homolog" – similar by descent from common ancestor

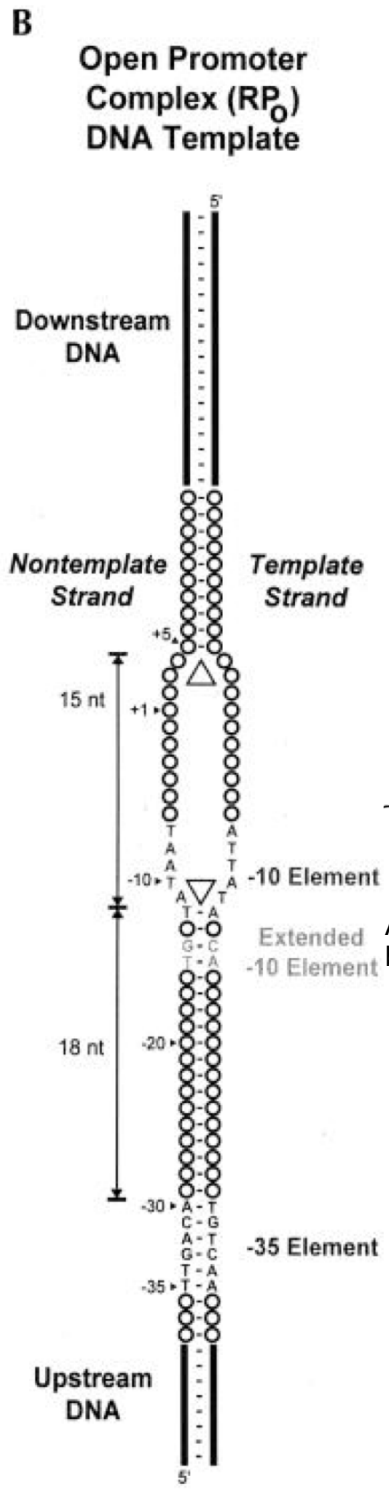
Sequence-based

- Smith-Waterman
- FASTA
- BLAST

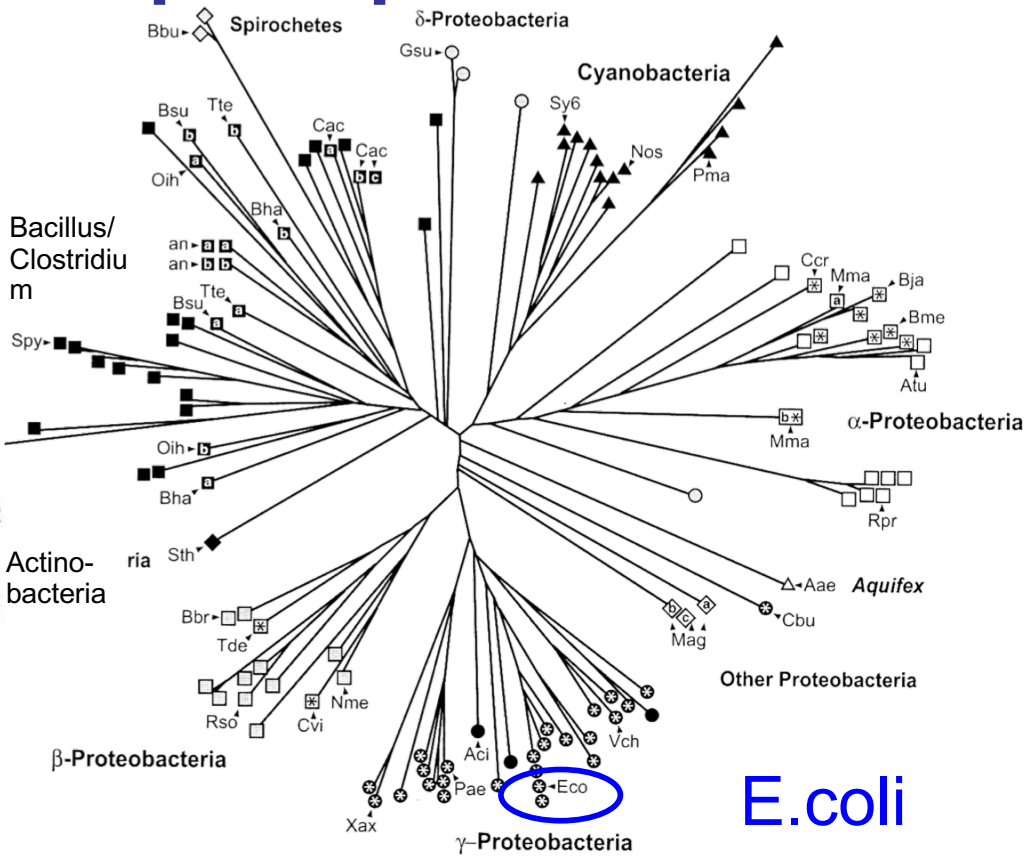
For RNA, sharp decline in sensitivity at ~60-70% identity

So, use structure, too





6S mimics an open promoter



E.coli

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

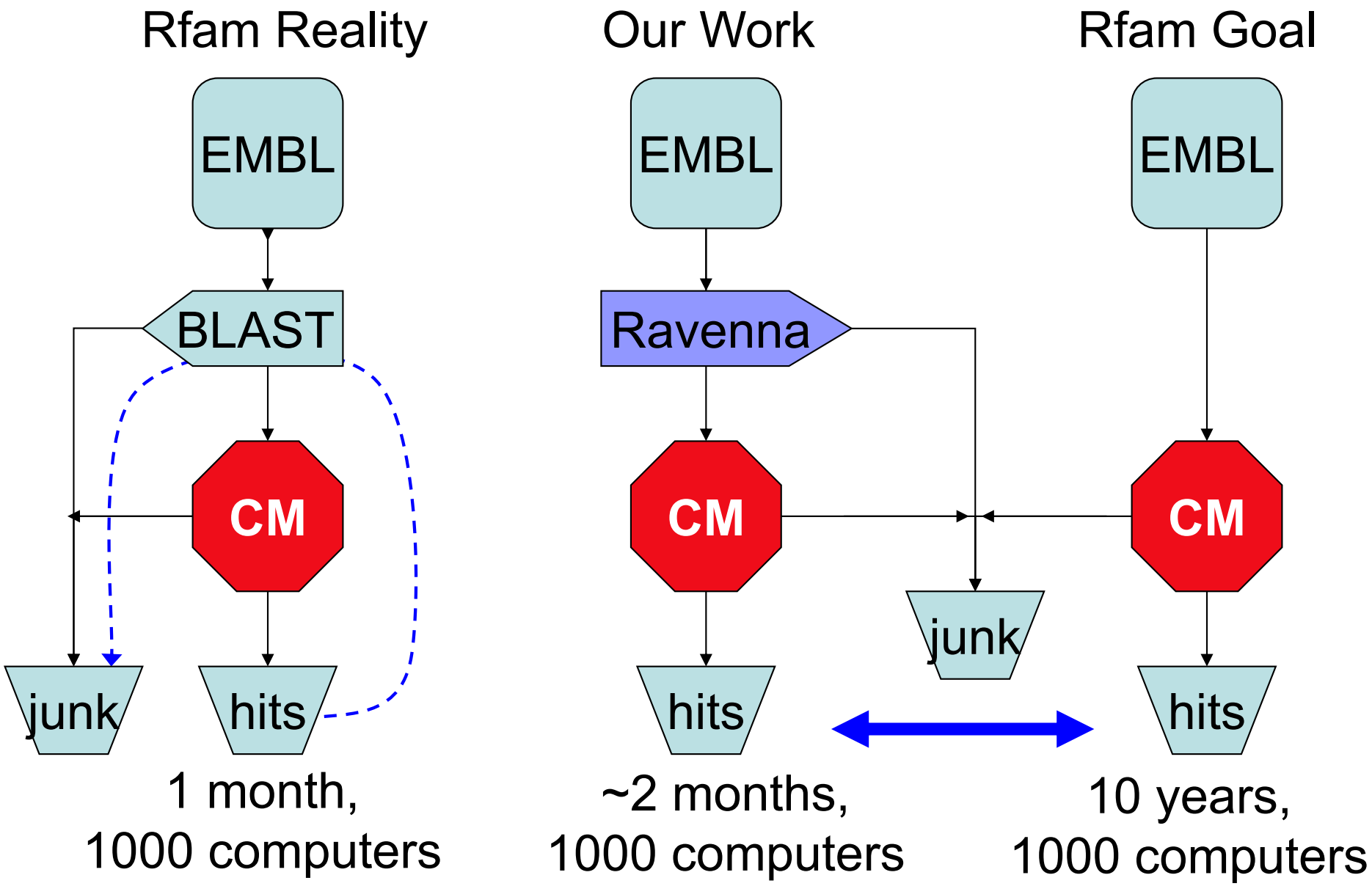
Faster Genome Annotation of Non-coding RNAs Without Loss of Accuracy

Zasha Weinberg

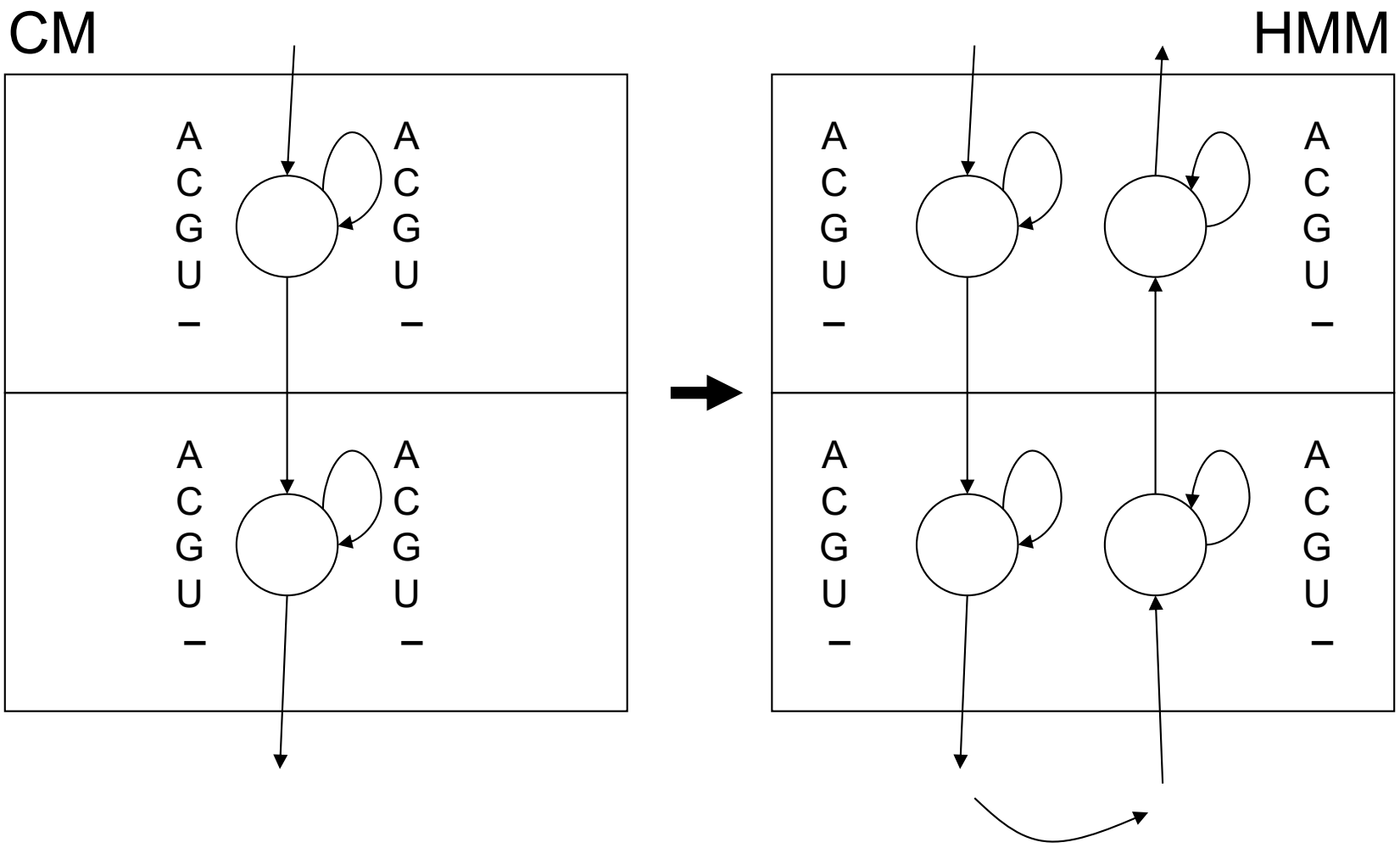
& W.L. Ruzzo

Recomb ‘04, ISMB ‘04, Bioinfo ‘06

CM's are good, but slow



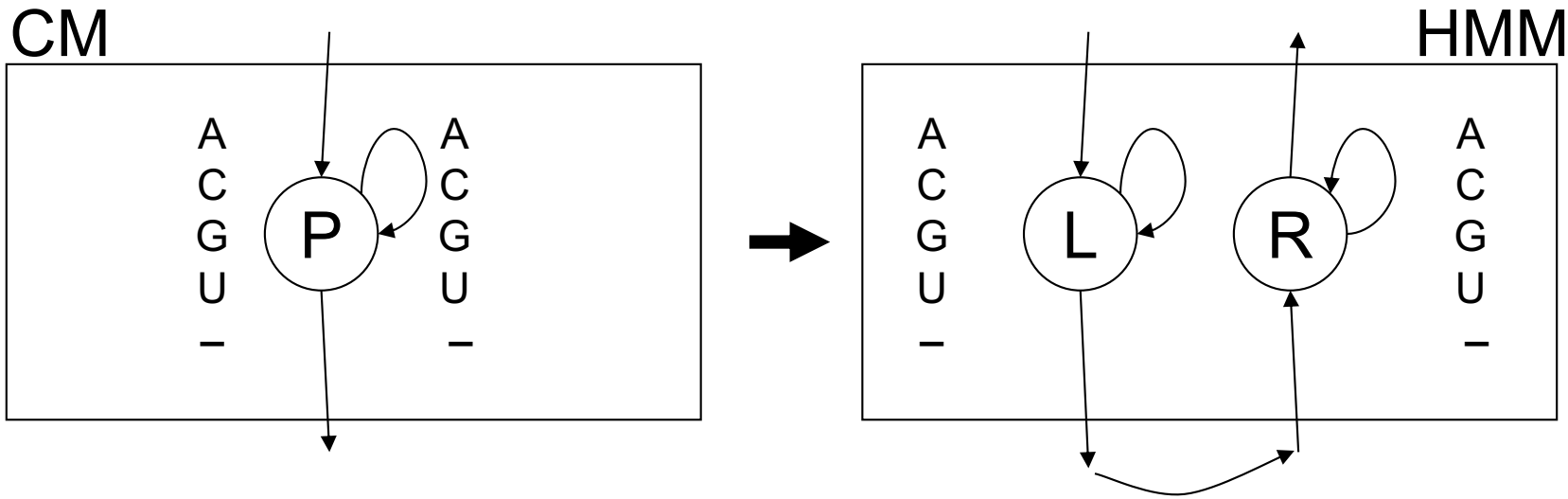
CM to HMM



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

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

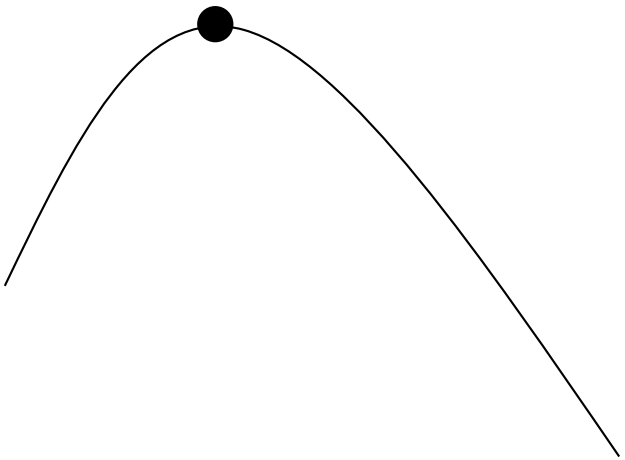
(but better ways, often)

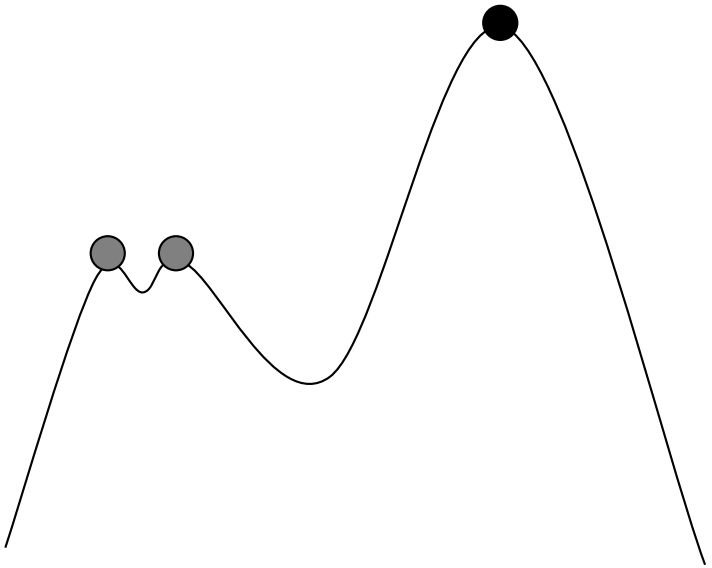
Nonconvex:

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.

RNA Motif Discovery

Would be great if: given 100 complete genomes from diverse species, we could automatically find all the RNAs.

State of the art: that's hopeless

Hope: can we exploit biological knowledge to narrow the search space?

RNA Motif Discovery

More promising problem: given a 10-20 unaligned sequences of a few kb, 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

Orthologs = counterparts in different species

Approaches

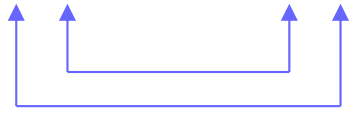
Align-First: Align sequences, then look for common structure

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 …



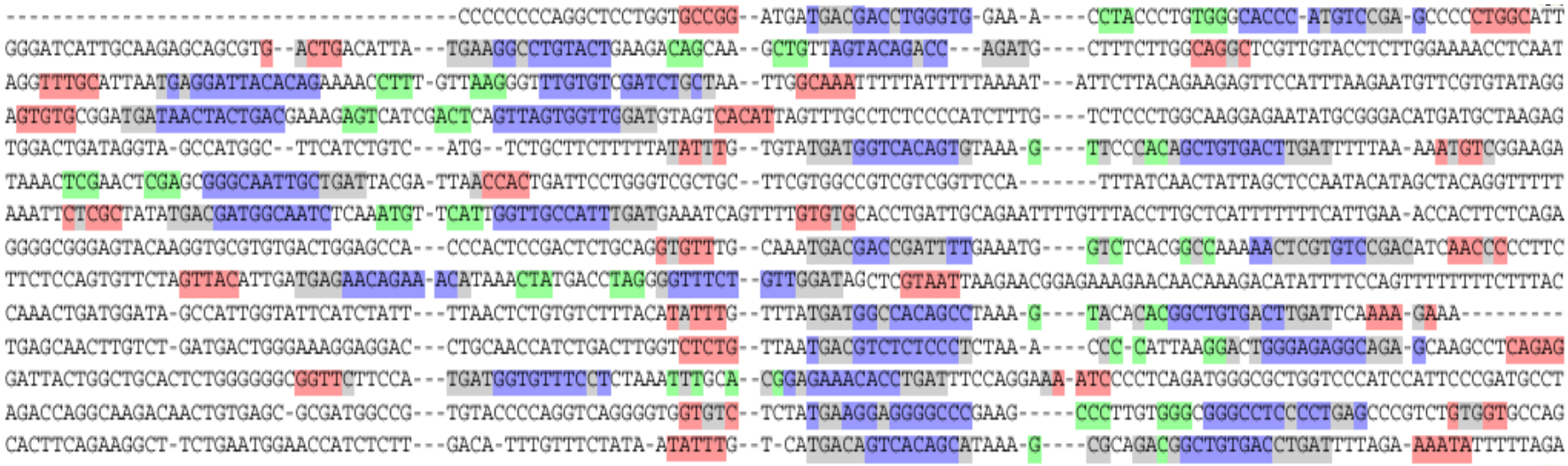
Compensatory mutations reveal structure (core of "comparative sequence analysis") *but* usual alignment algorithms penalize them (twice)

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

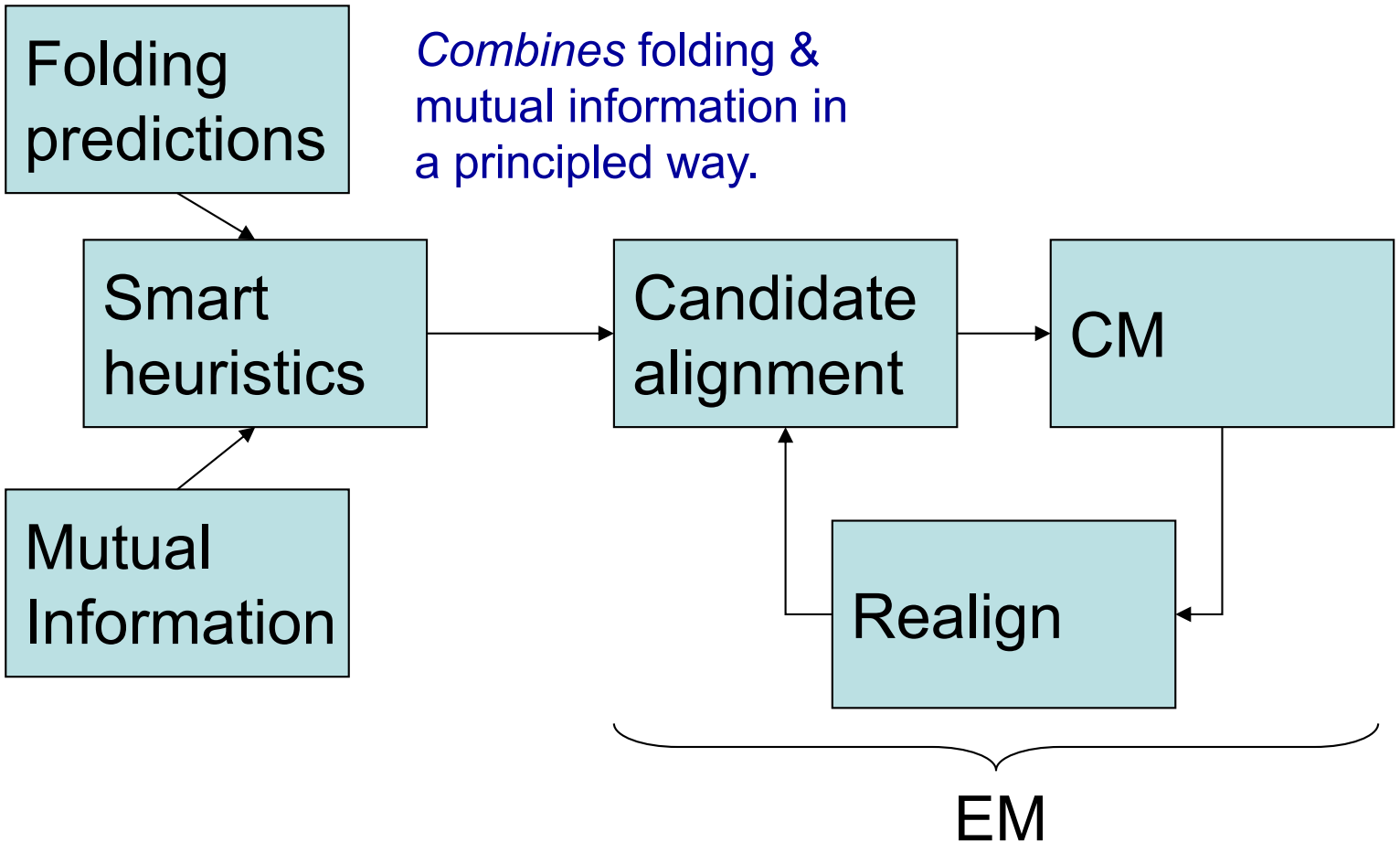
- Sequence conservation exploited, but not required
- Robust to inclusion of unrelated and/or flanking sequence
- Reasonably fast and scalable
- Produces a probabilistic model of the motif that can be directly used for homolog search

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

CMFinder

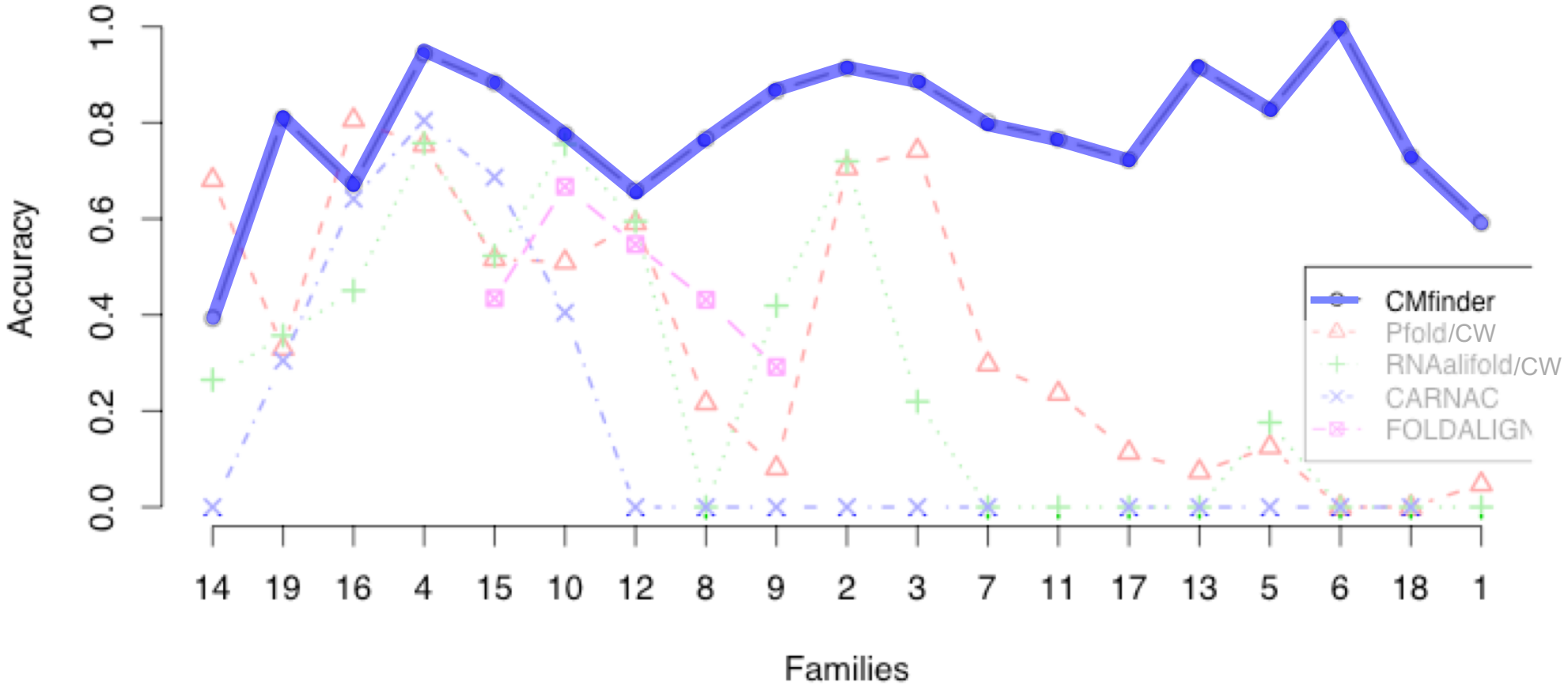
Simultaneous alignment, folding & motif description

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]

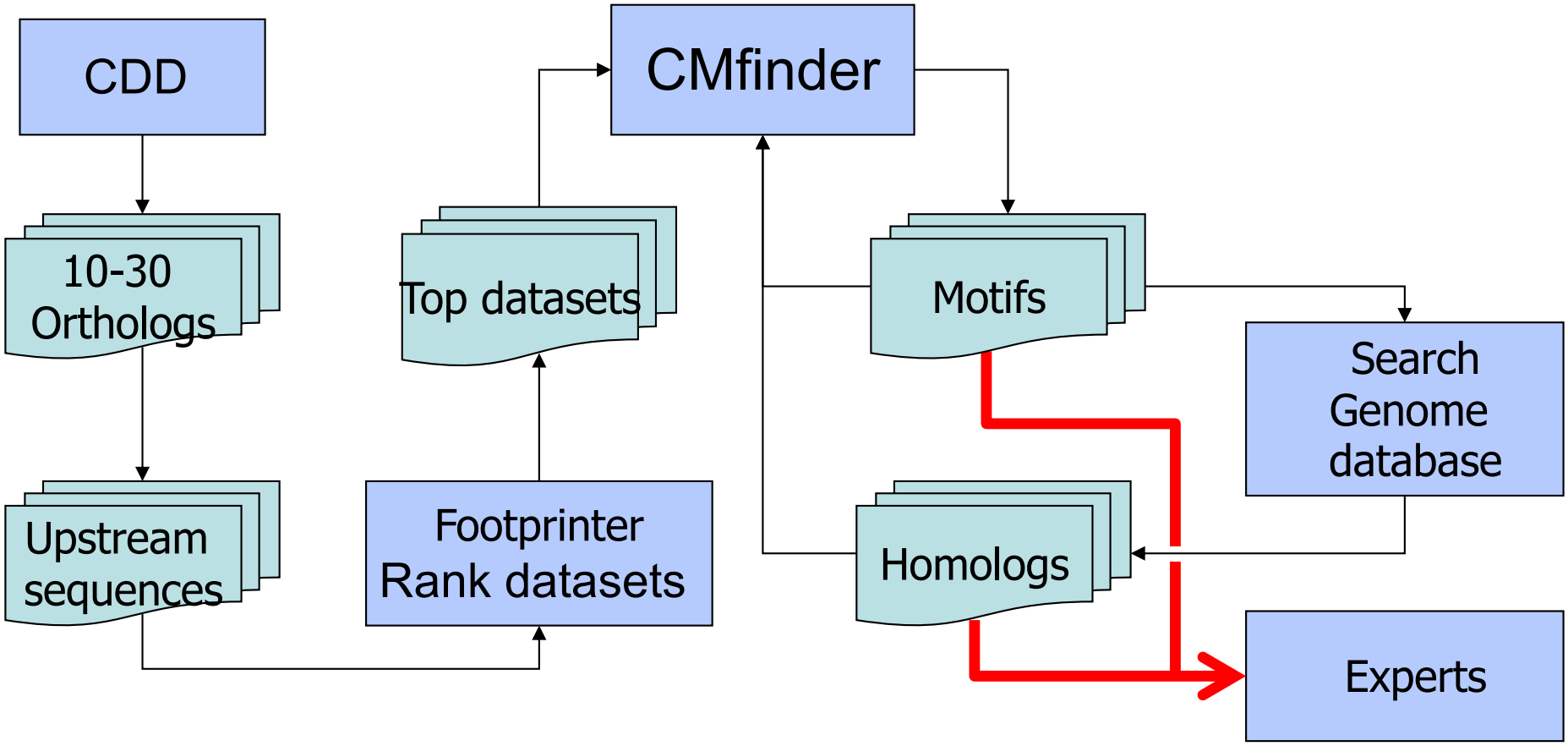
Published online 9 July 2007

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]

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 Comput Biol. 3(7): e126, July 6, 2007.

Semi-automated Example

Started with 16 genes orthologous to folC in *B. subtilis*

Found 9 sharing good structural motif

Searched all bacterial genomes for this motif

Found 234 hits

Realigned these to refine structural motif

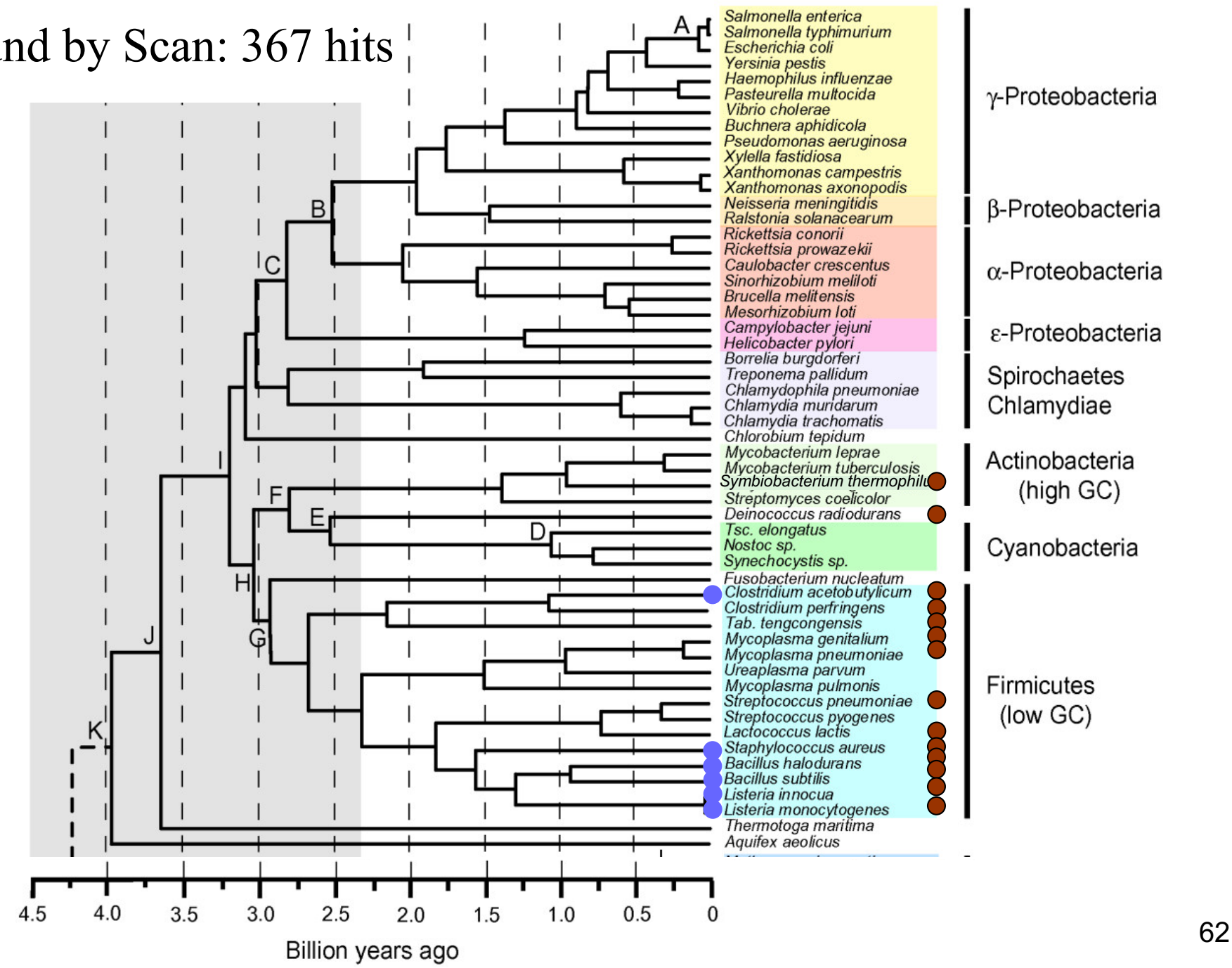
Found 367 hits

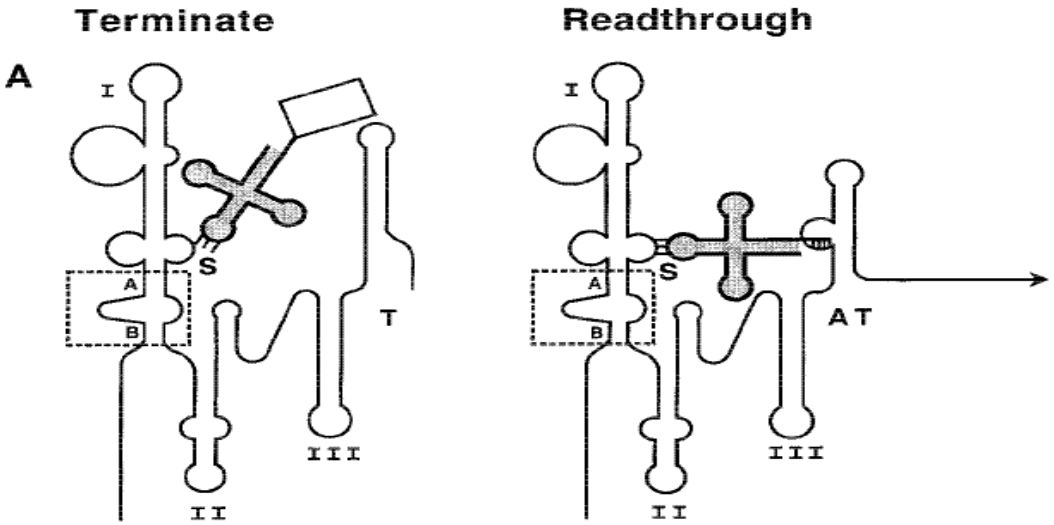
257 match RFAM's T-box

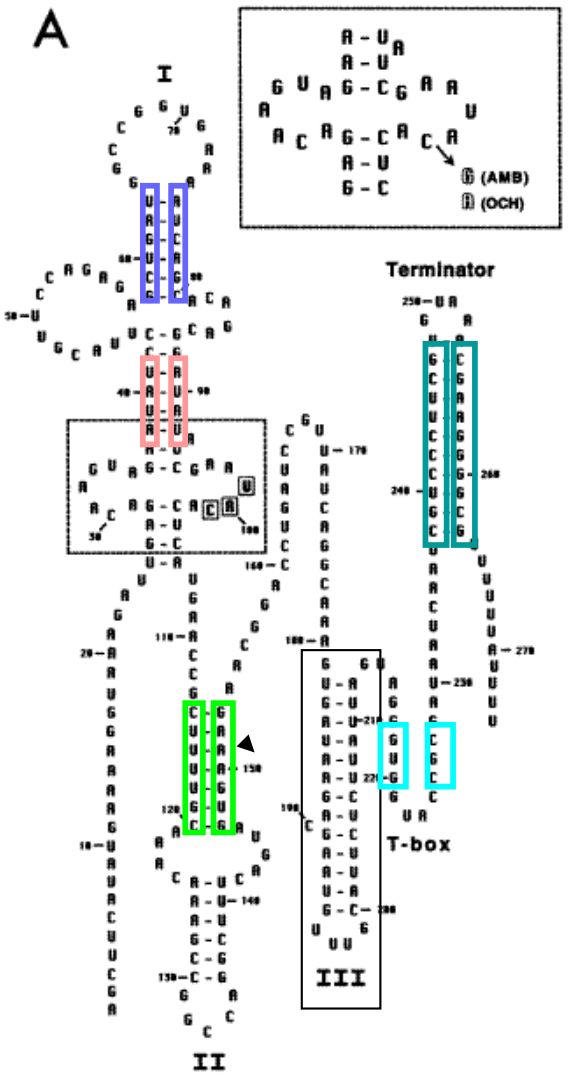
(Based on hand-curated alignment of 67 knowns)

62/110 "false positives" are probable true positives (upstream of annotated tRNA-synthetase genes)

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








```
NC_000964.1 AUAUC.CUUACGU..UCCAGAGAGCUGAUGGCCGGUGAAA.AUCAGCACAGACGGAUAUA
NC_004722.1 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUU
NC_004193.1 AAAAGUAGAACCG.AUCUAGCGAAUUGAGGAU.GGUGUGAGCUCAGUGC.GGAAAGCUUU
NC_003997.3 CAAAU.GUCGUUUcUUAUAGAGAGUCGAUGGUUGGUGGAA.AUCGAUAG..AAACAGUUU

NC_000964.1 CGAA..UACACUCAUGAACCGCUUUUGCAAACAAAGccggccaggcuuucAGUA.GUGAA
NC_004722.1 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuuggauu.....AGUAAGCAUU
NC_004193.1 AGAAAAUC.ACUCUUGAGUU.UUCAUUACGAAA..CA............AGUAGUAAUG
NC_003997.3 UGAA..UCCAUCCUGGAAU..GGAAUGUGGAAUAUCUuuaugauu.....AGUAAACAUU

NC_000964.1 aCGGAC.CUGAUCCGUUAUCAGGCAAAGUGGUACCGCGAUAAUCAAUCGUCCCUUCGUGUAAaCGAAGGGGCGUUU
NC_004722.1 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAUaAGGGACGGGAGUU
NC_004193.1 .CGGUUcAUC.UCCGUUAUCGAUCUUAGUGGUACCGCGA.....GUCUUCUCGUCCCUUUU..GGGAUUAGAAGGC
NC_003997.3 .CGGUG.AAGAGCCGUUAUU...UCuAGUGGCAACGCGG..GUUAACUCCCGUCCCUUUAUAUaAGGGACGGGAGUU
```

Overall Pipeline & Processing Times

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

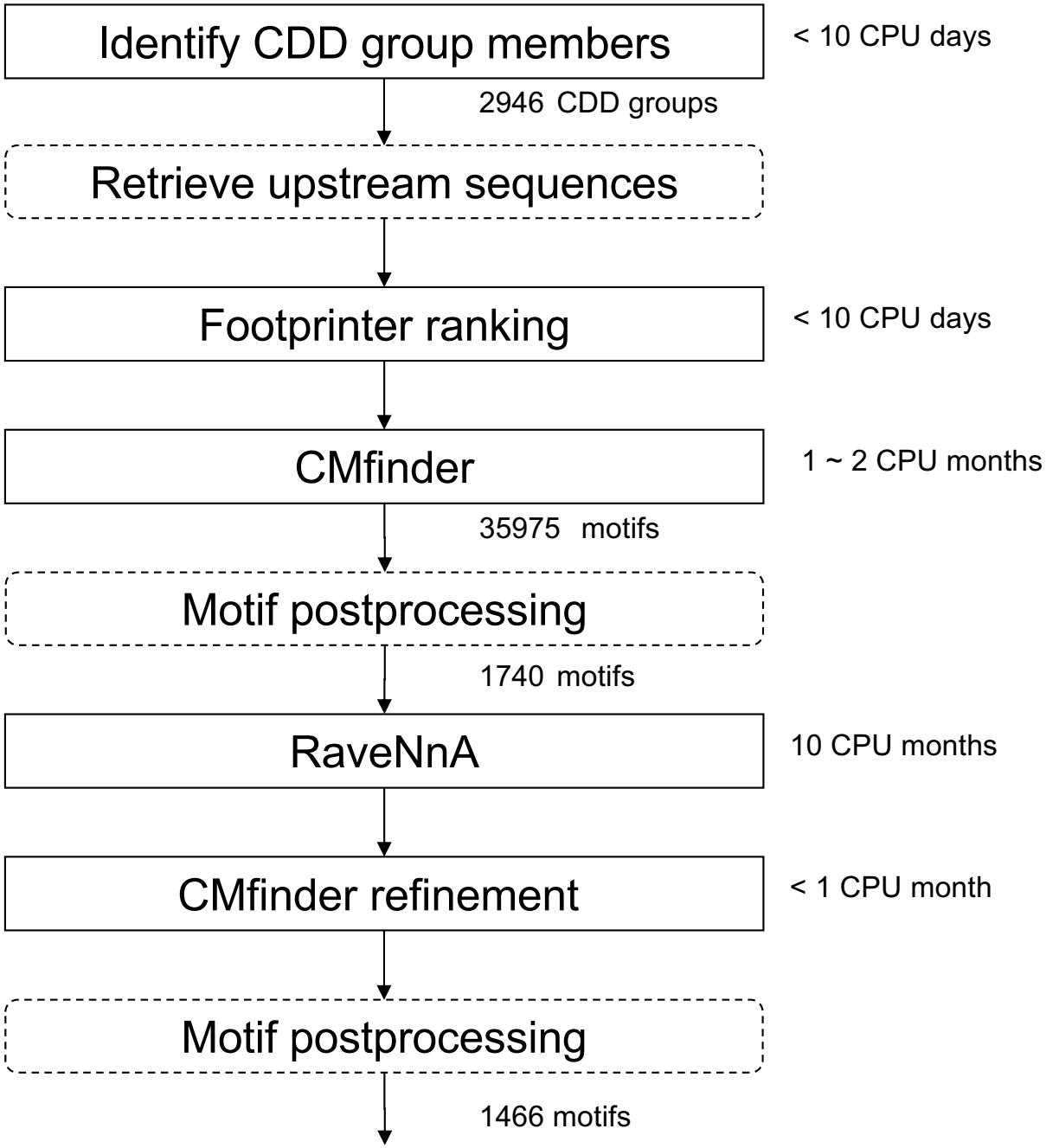


Table 1: Motifs that correspond to Rfam families

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

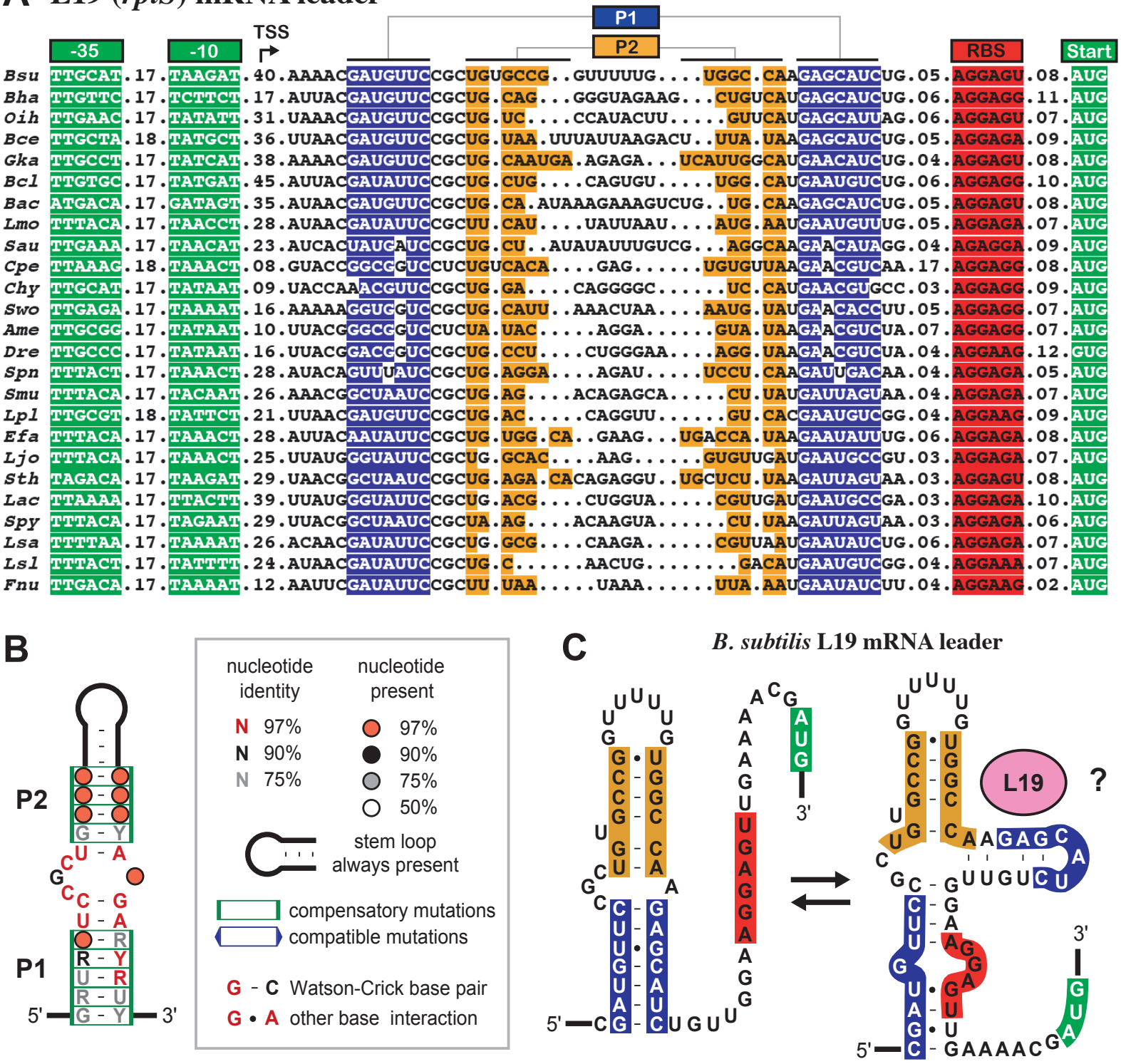
Tbl 2: Prediction accuracy compared to prokaryotic subset of Rfam full alignments.
Membership: # of seqs in overlap between our predictions and Rfam's, the sensitivity (Sn) and specificity (Sp) of our membership predictions. Overlap: the avg len of overlap between our predictions and Rfam's (nt), the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

Table 3: High ranking motifs not found in Rfam

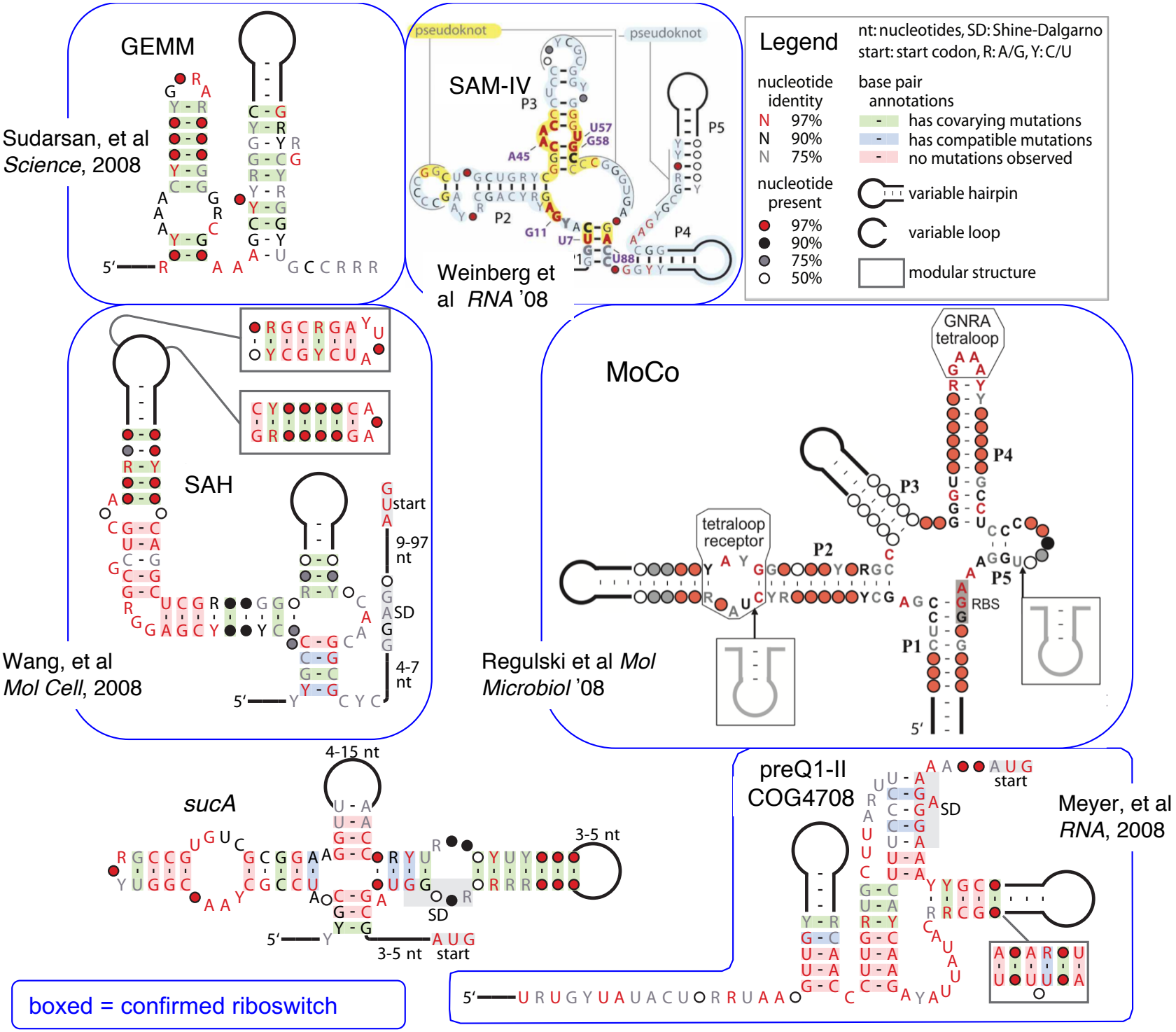
Rank	#	CDD	Gene: Description	Annotation
6	69	28178	DHOase IIa: Dihydroorotase	PyrR attenuator [22]
15	33	10097	RplL: Ribosomal protein L7/L1	L10 r-protein leader; see Supp
19	36	10234	RpsF: Ribosomal protein S6	S6 r-protein leader
22	32	10897	COG1179: Dinucleotide-utilizing enzymes	6S RNA [25]
27	27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

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.

CSH PRESS GENOME RESEARCH

The identification and functional annotation of RNA structures conserved in vertebrates

Stefan E. Seemann,[1,2] Aashiq H. Mirza,[1,3,10] Claus Hansen,[1,4,10]
Claus H. Bang-Berthelsen,[1,5,10,11] Christian Garde,[1,6,10]
Mikkel Christensen-Dalsgaard,[1,4] Elfar Torarinsson,[1] Zizhen Yao,[7]
Christopher T. Workman,[1,6] Flemming Pociot,[1,3] Henrik Nielsen,[1,4]
Niels Tommerup,[1,4] Walter L. Ruzzo,[1,8,9] and Jan Gorodkin[1,2]

Genome Res. 2017 27: 1371-1383 originally published online May 9, 2017
Access the most recent version at doi:10.1101/gr.208652.116

Outline

There is *A LOT* of noncoding expression

Significance remains controversial

What could help clarify? – *conserved 2^d structure* (not seq)

Several groups have tried

+ genome-wide, rather than cell type/state-specific RNAseq

– high FDR

Our improved screen:

better scoring, better null, realignment

Results –

selection, conserved expression, conserved structures, SNP association

⇒enhancer/promoter

Conclusion

Motivation

<2% of the human genome codes for protein

<25% is in protein coding genes (cds+introns)

But recent estimates say *50-90% transcribed*

Functional? Or "transcriptional noise"?

Lots of ncRNA

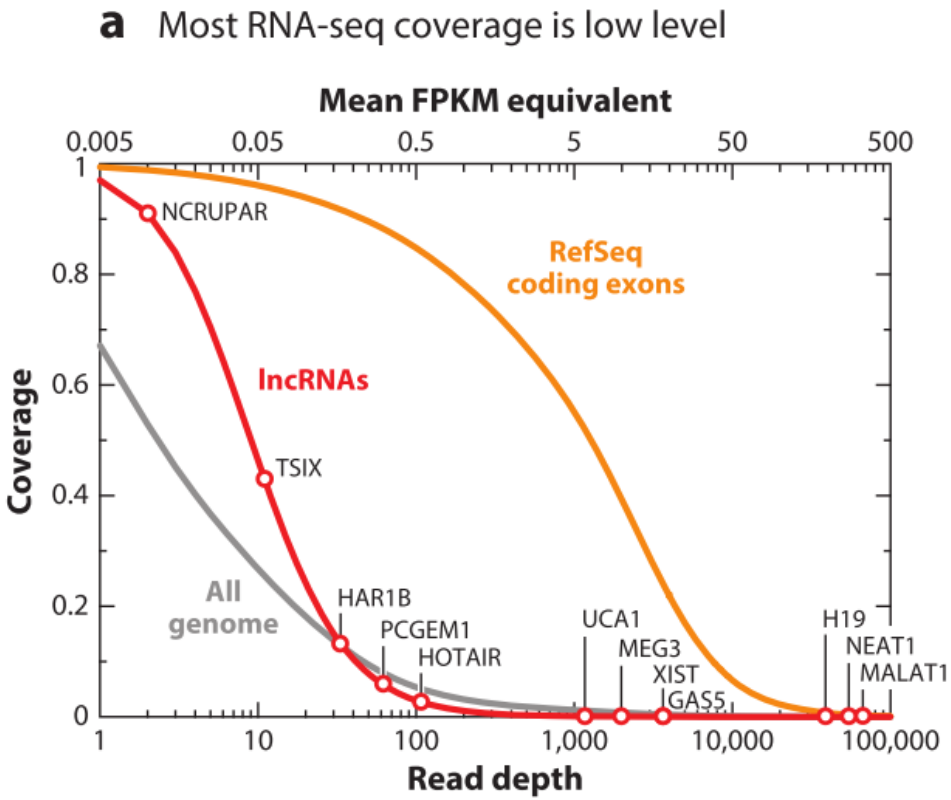
GENCODE version 23 (March 2015):

- 19,797 protein-coding genes
- 15,931 long non-coding RNAs; 9,882 small non-coding RNAs

Lots of ncRNA;but low expr

GENCODE version 23 (March 2015):

- 19,797 protein-coding genes
- 15,931 long non-coding RNAs; 9,882 small non-coding RNAs

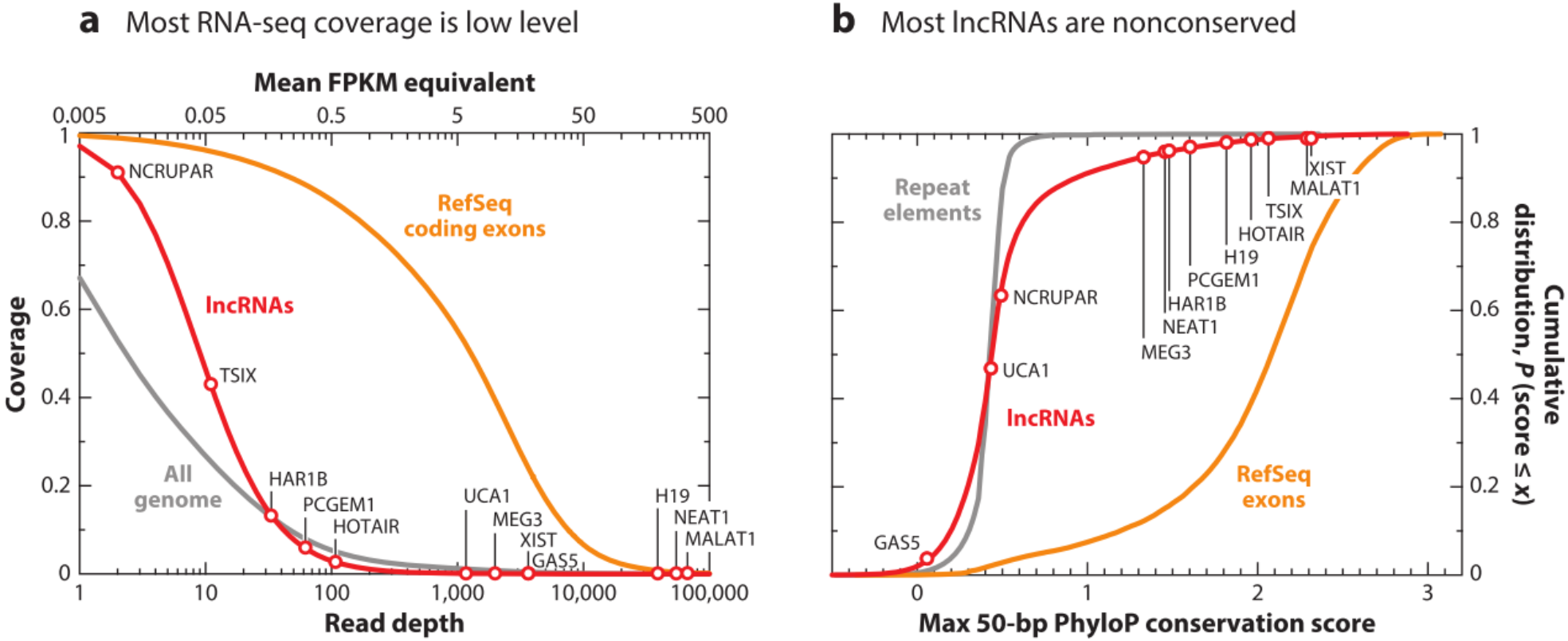


[Eddy (2013) *Annu Rev Biophys*]

Lots of ncRNA;but low expr, consv

GENCODE version 23 (March 2015):

- 19,797 protein-coding genes
- 15,931 long non-coding RNAs; 9,882 small non-coding RNAs



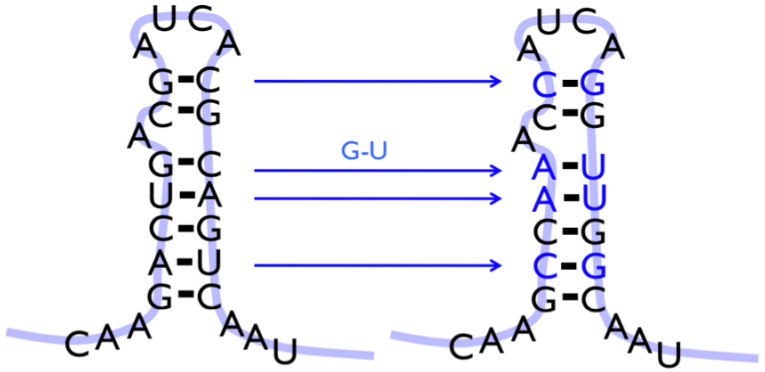
[Eddy (2013) *Annu Rev Biophys*]

Conservation

Above is *Sequence-level* conservation

But secondary structure plays an important role in biogenesis and/or activity of most ncRNAs (that we understand)

What about conservation of structure?



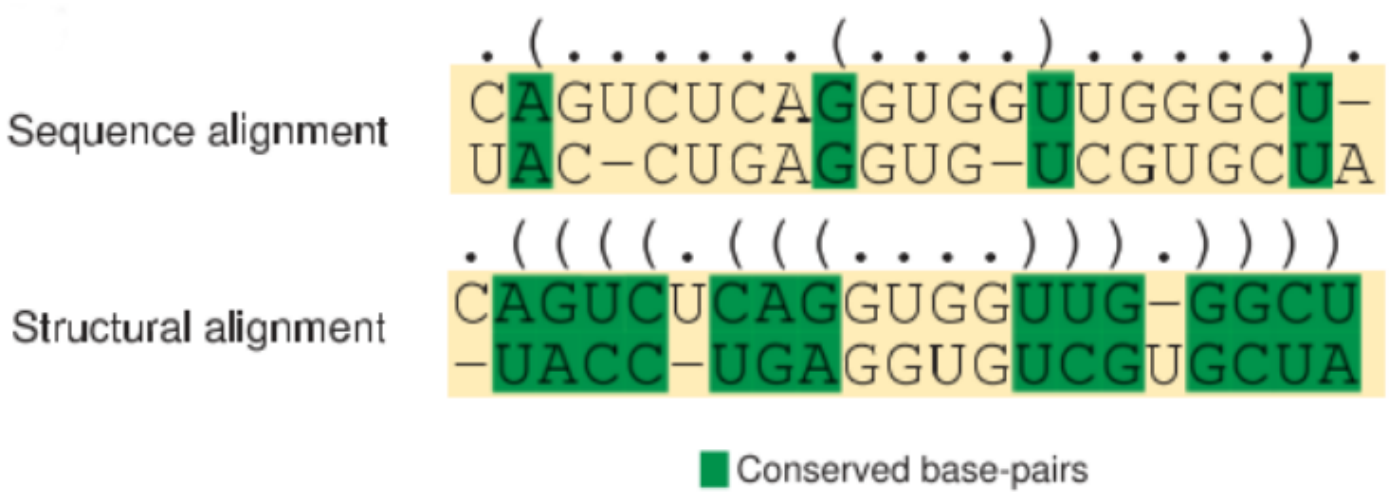
Genome-wide screens find many conserved RNA structures

Previous screens for RNA structure prediction in vertebrate genomes:

- AliFoldZ [Washietl (2007) *Genome Res*]
- RNAz [Gruber (2010) *Pacific Symposium on Biocomputing*]
- Evofold [Parker (2011) *Genome Res*]
- RNAz + SISSIz [Smith (2013) *NAR*]

+ whole genome
– high FDR

Limitation of comparative analysis based on multiple sequence alignments:



	Base matches	Basepair matches
Sequence alignment	13	2
Structural alignment	6	7

New genome-wide screen: Methods

17-way vertebrate alignments from MultiZ

Ignore nucleotide-level alignment but hope alignment blocks will contain orthologous regions

Align with Cmfinder

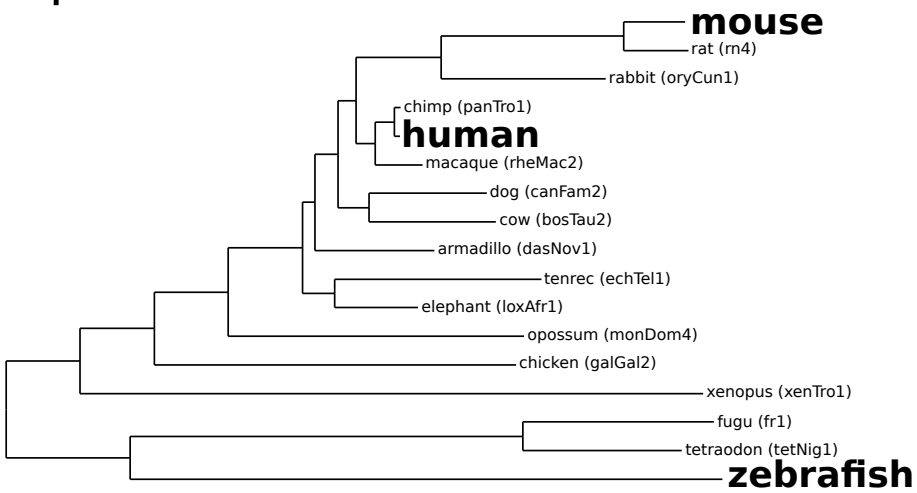
Score motifs (phylogenetically informed scores based on separate substitution matrices for single- and double-stranded positions)

Estimate FDR base on di-nucleotide controlled shuffling of alignments, with regression-based correction of important effects like GC content

… thousands of CPU years pass …

Genome-wide screen of human for conserved RNA structures

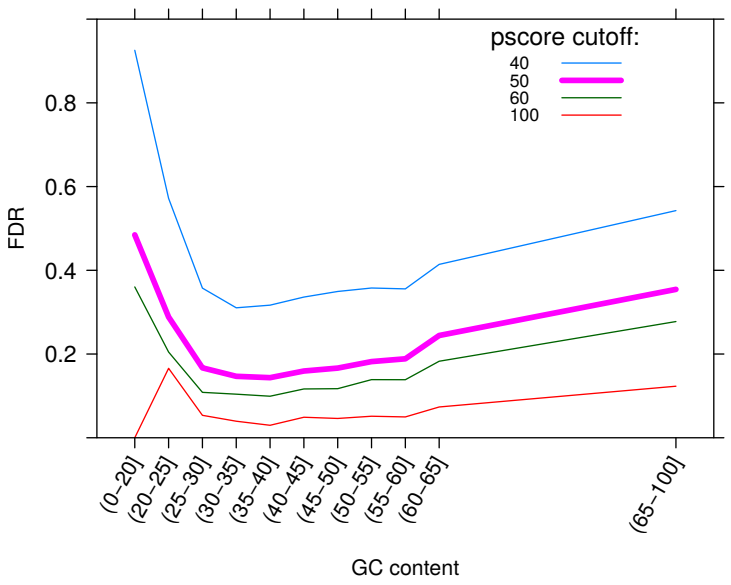
Input:

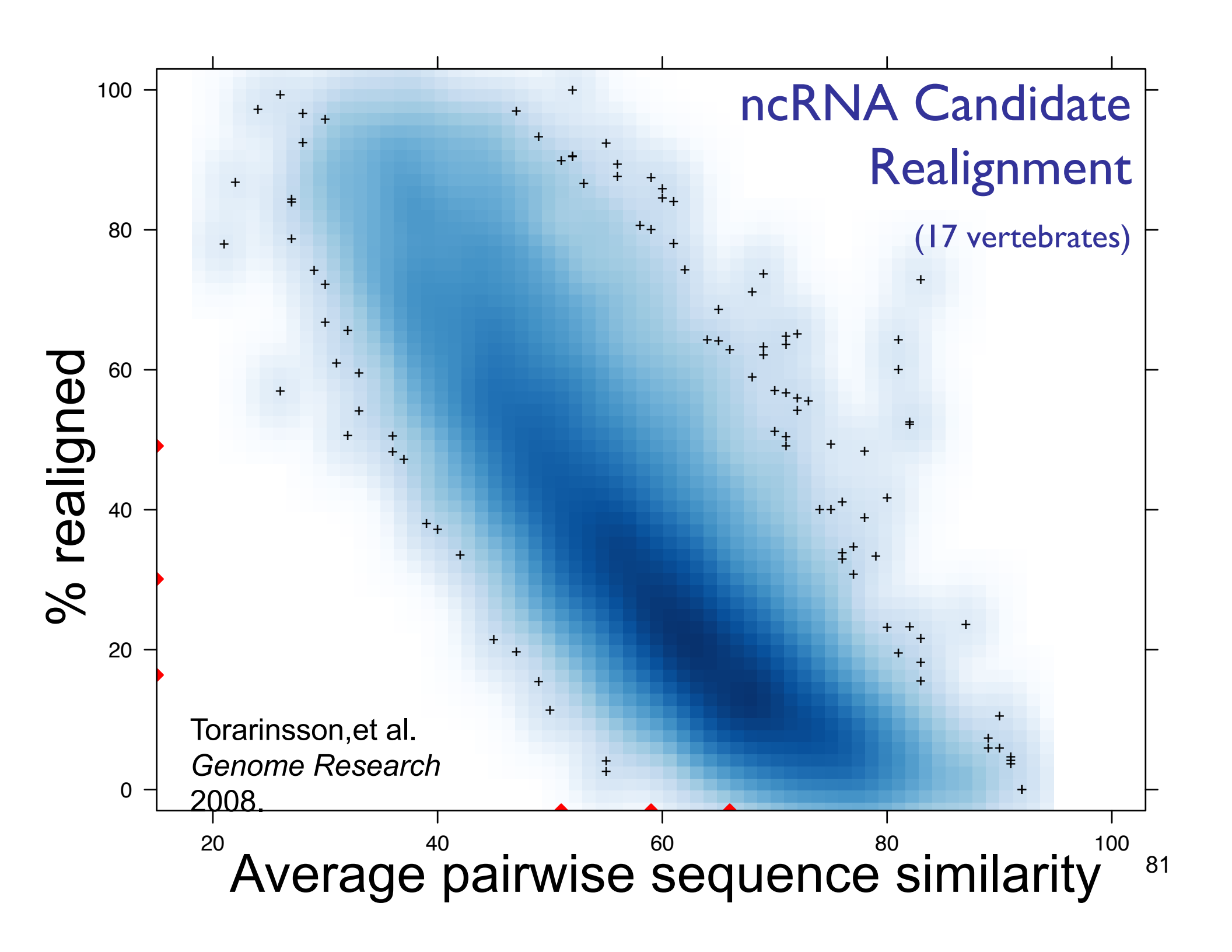


- human centered 17-way MULTIZ alignments
- 50% of human genome
- 50% have low conservation according to PhastCons

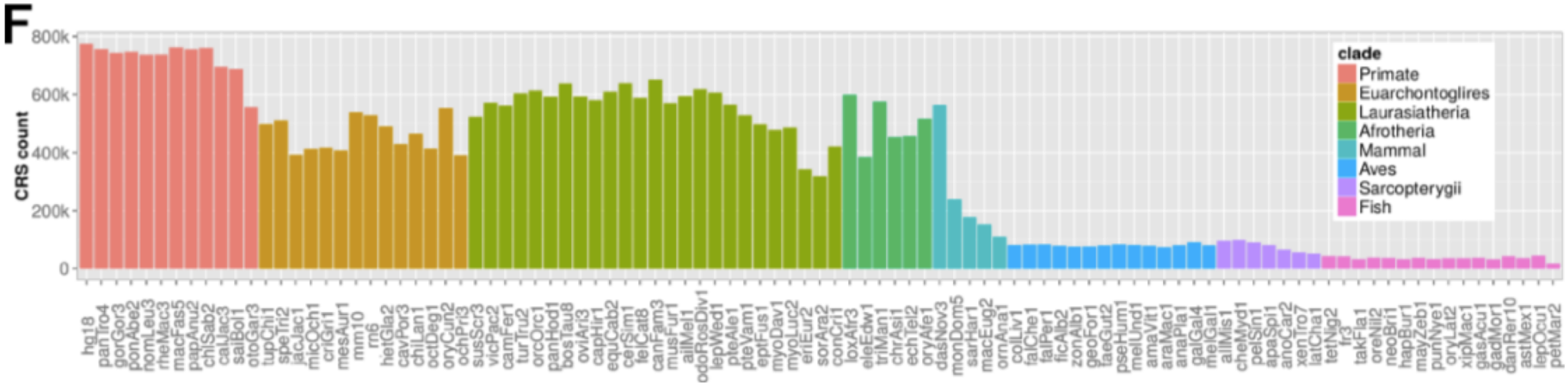
Prediction results:

- 780k conserved RNA structures (CRSs) from 520k regions
- estimated FDR ~15% (GC content range 20%-65%)
- sequence identity: 60.2%
- length: 69bp (longest: 497bp)



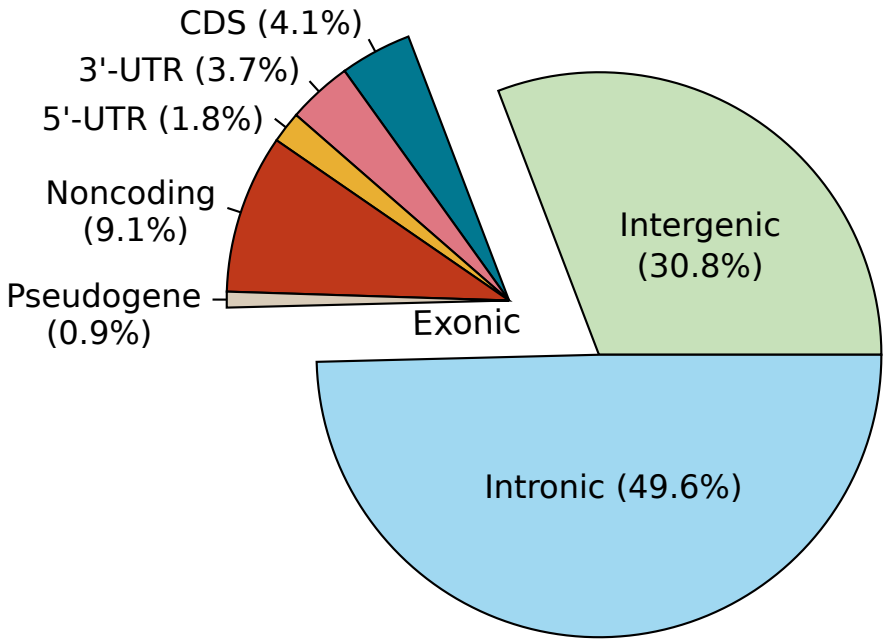

ncRNA Candidate Realignment
(17 vertebrates)
% realigned
Average pairwise sequence similarity
0
20
40
60
80
100
20
40
60
80
100
Torarinsson,et al.
Genome Research
2008.

Broad Conservation

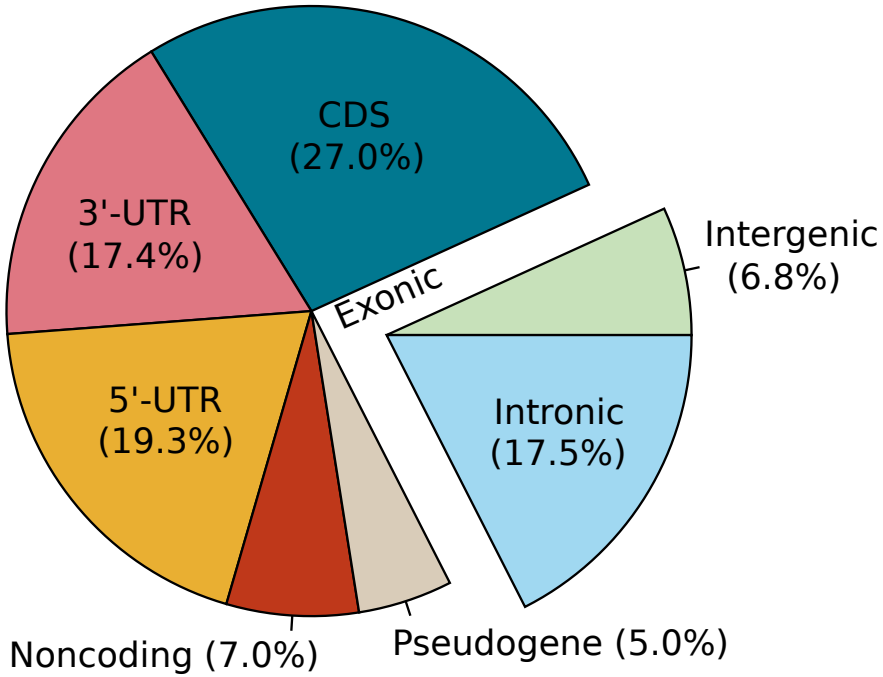


Genomic annotation of CRSs

Absolute:

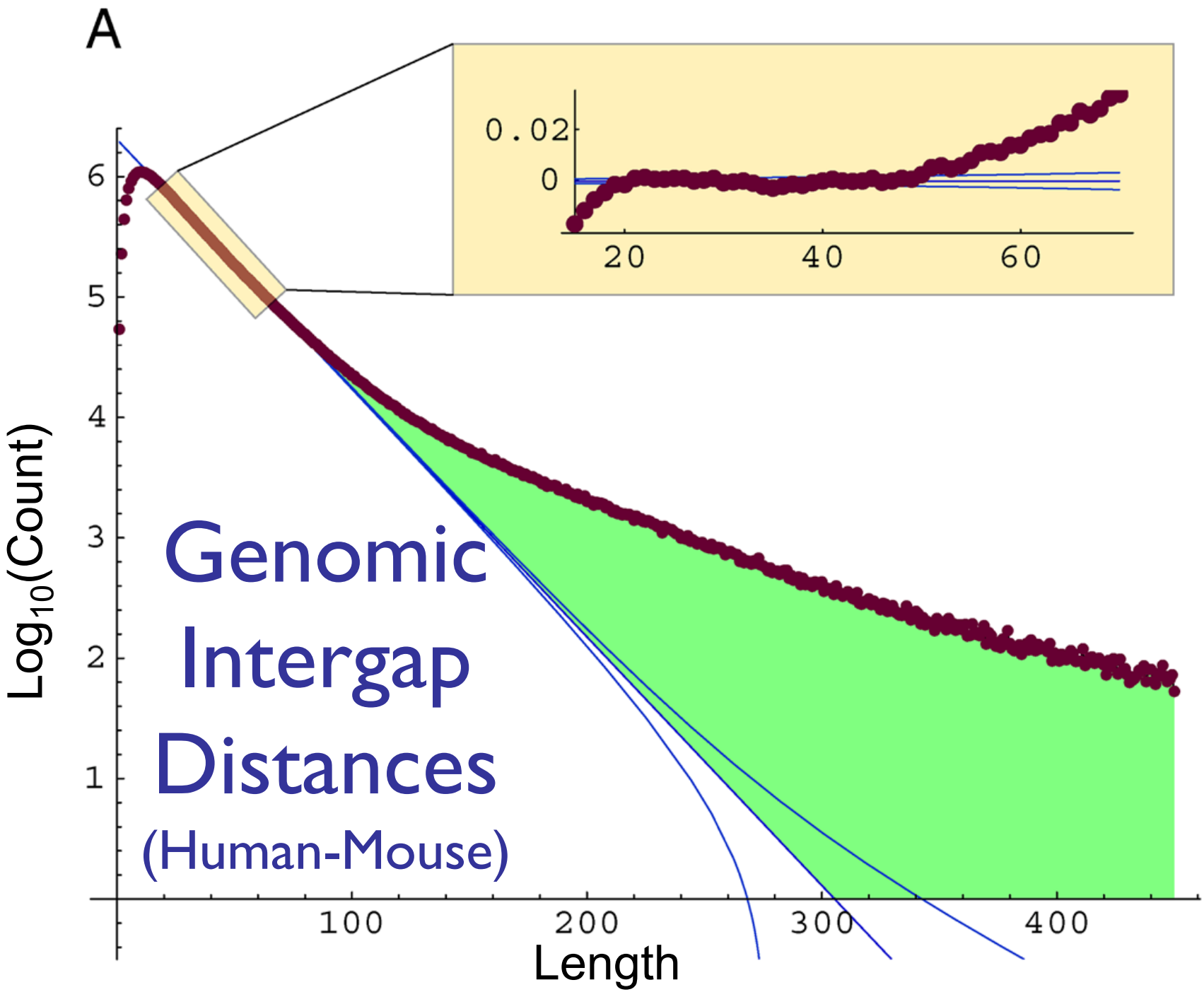


Relative:



CRSs are enriched to overlap

- known ncRNAs (*e.g.* pre-miRNAs, tRNA, snoRNAs and lncRNAs)
- protein binding sites (CLIP RBP)



Genome-Wide Identification of Human Functional DNA Using a Neutral Indel Model
Gerton Lunter, Chris P. Ponting, Jotun Hein, PLoS Comput Biol 2006, 2(1): e5.

Overlap w/ Indel Purified Segments

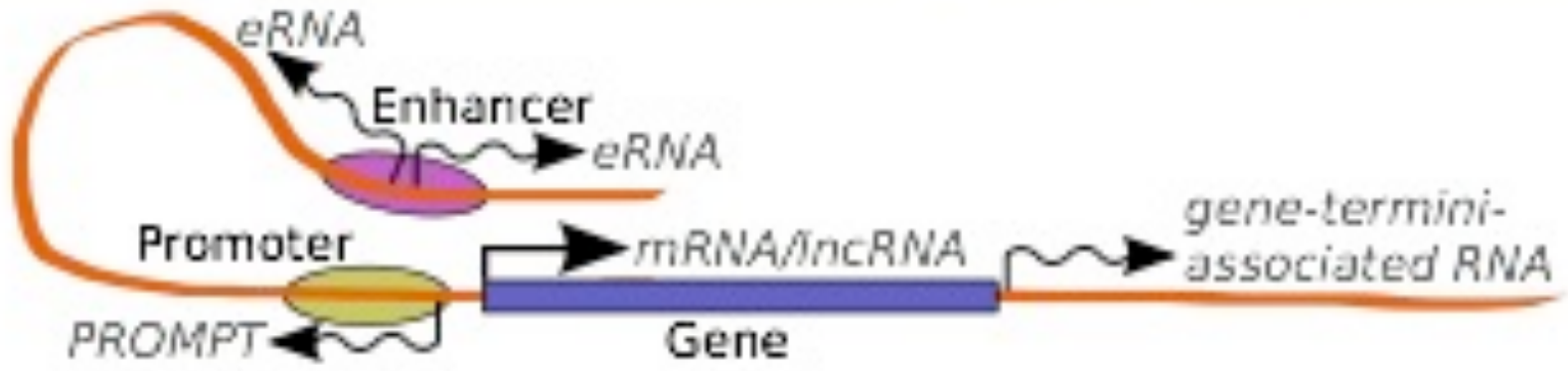
IPS presumed to signal purifying selection

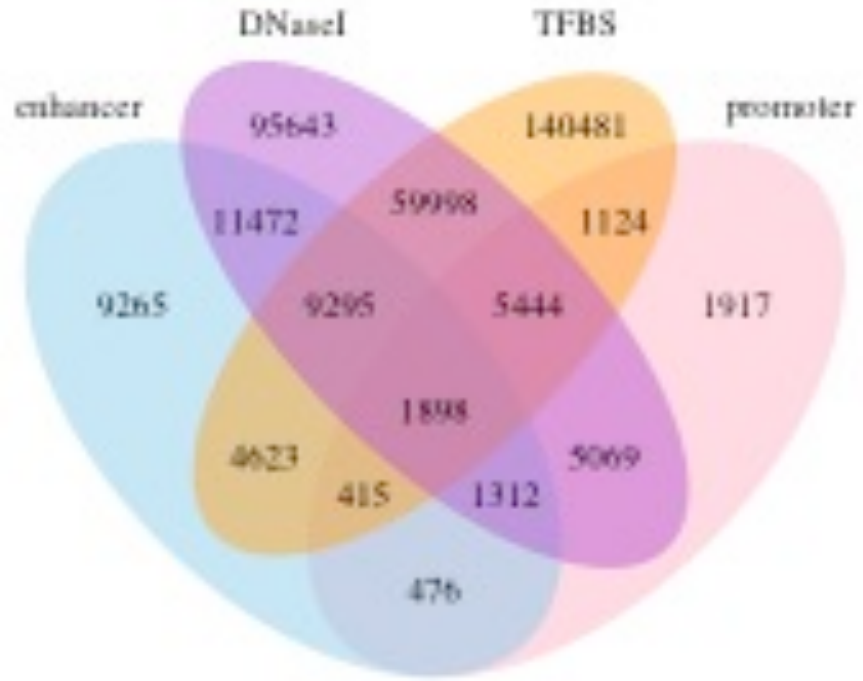
Majority (64%) of candidates have >45% G+C

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%

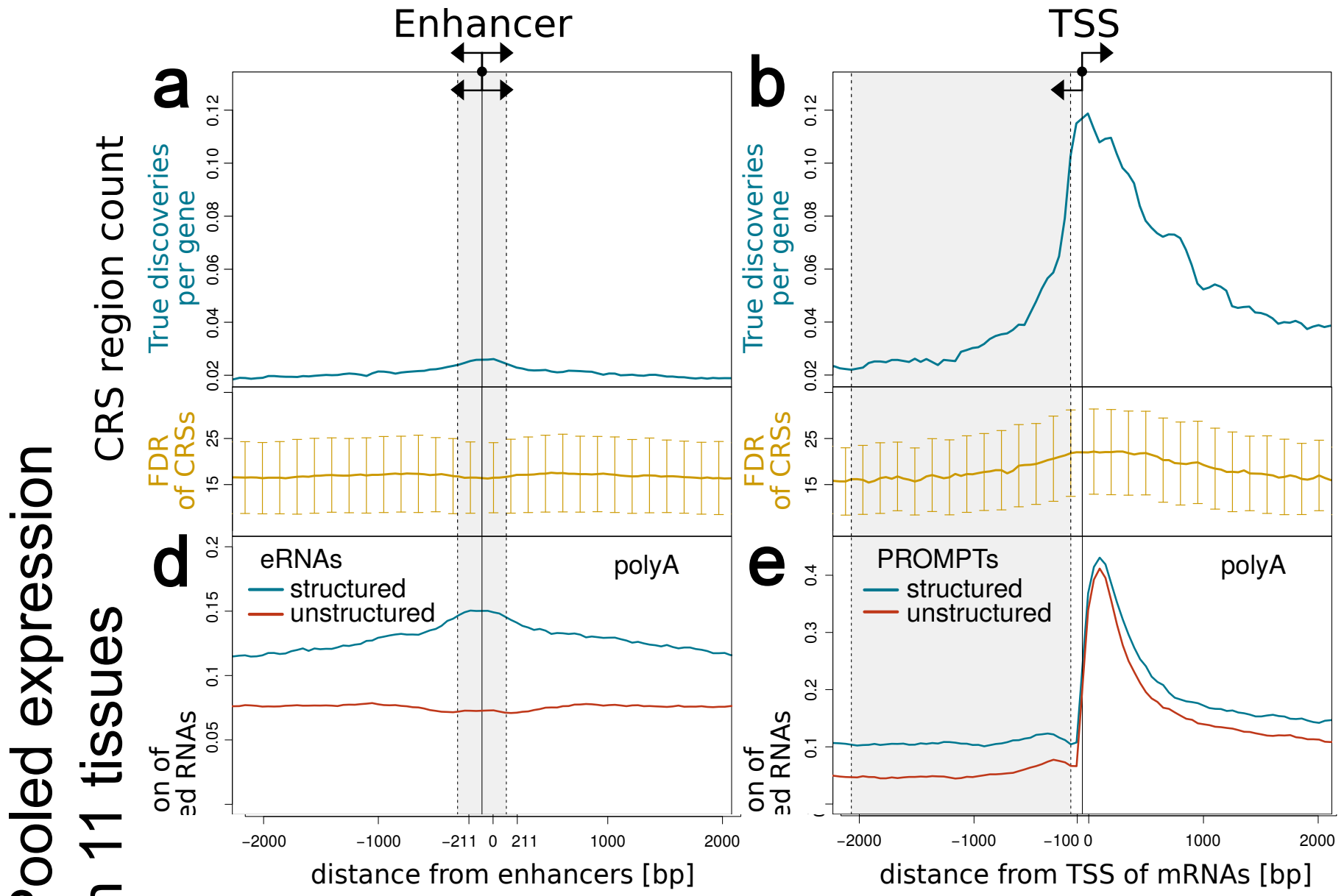
CRSs are located in cis-regulatory regions



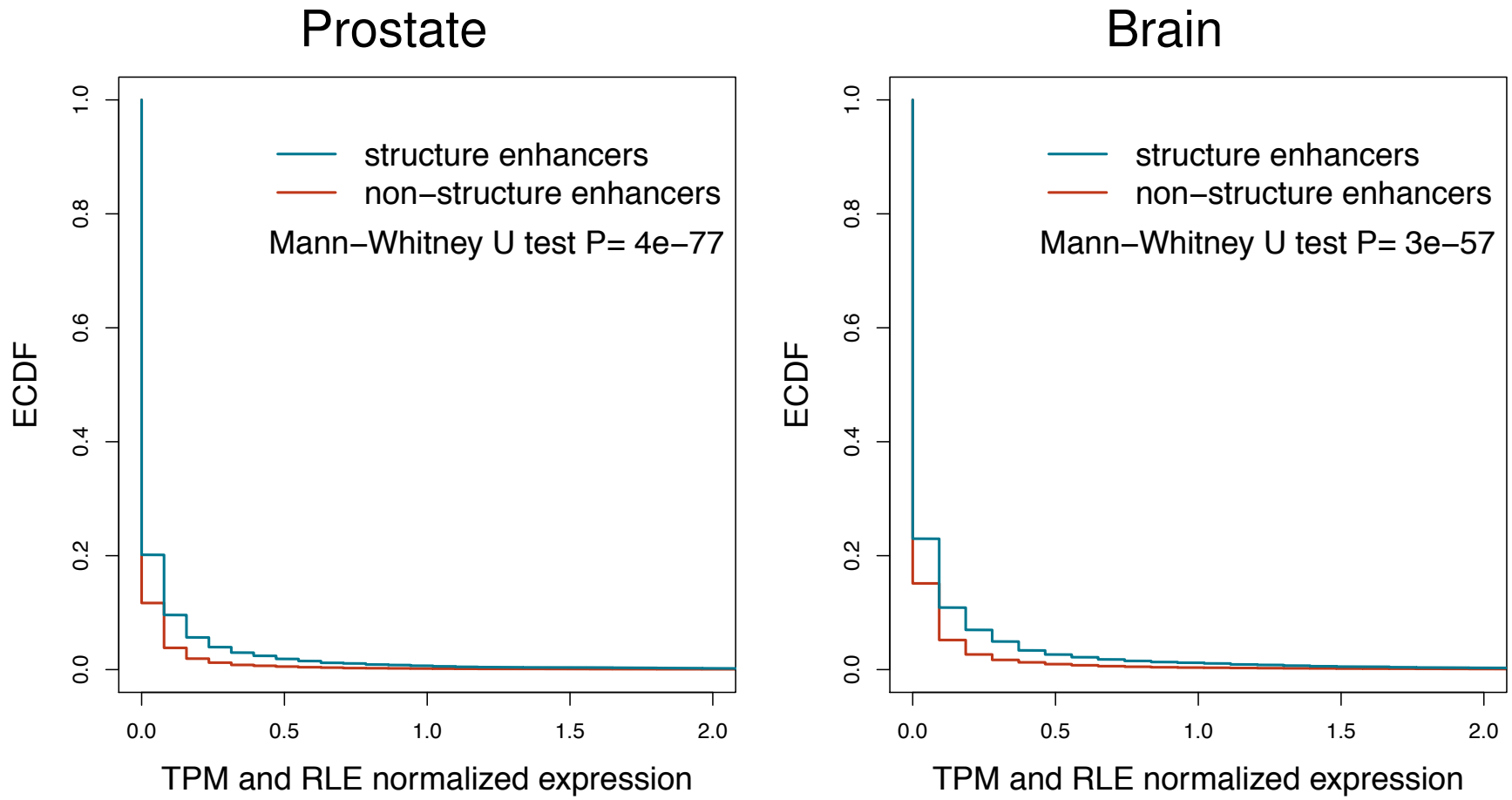


- RNAs transcribed upstream of promoters
  → *from TSS -100bp to -2kb*
- RNAs transcribed downstream of 3' termini of genes
  → *from most distal 3' end +100bp to +2kb*
- enhancers transcribed from specific chromatin marks
  → *ENCODE chromatin segmentation; 100bp to 1kb*

Enriched in cis-regulatory regions



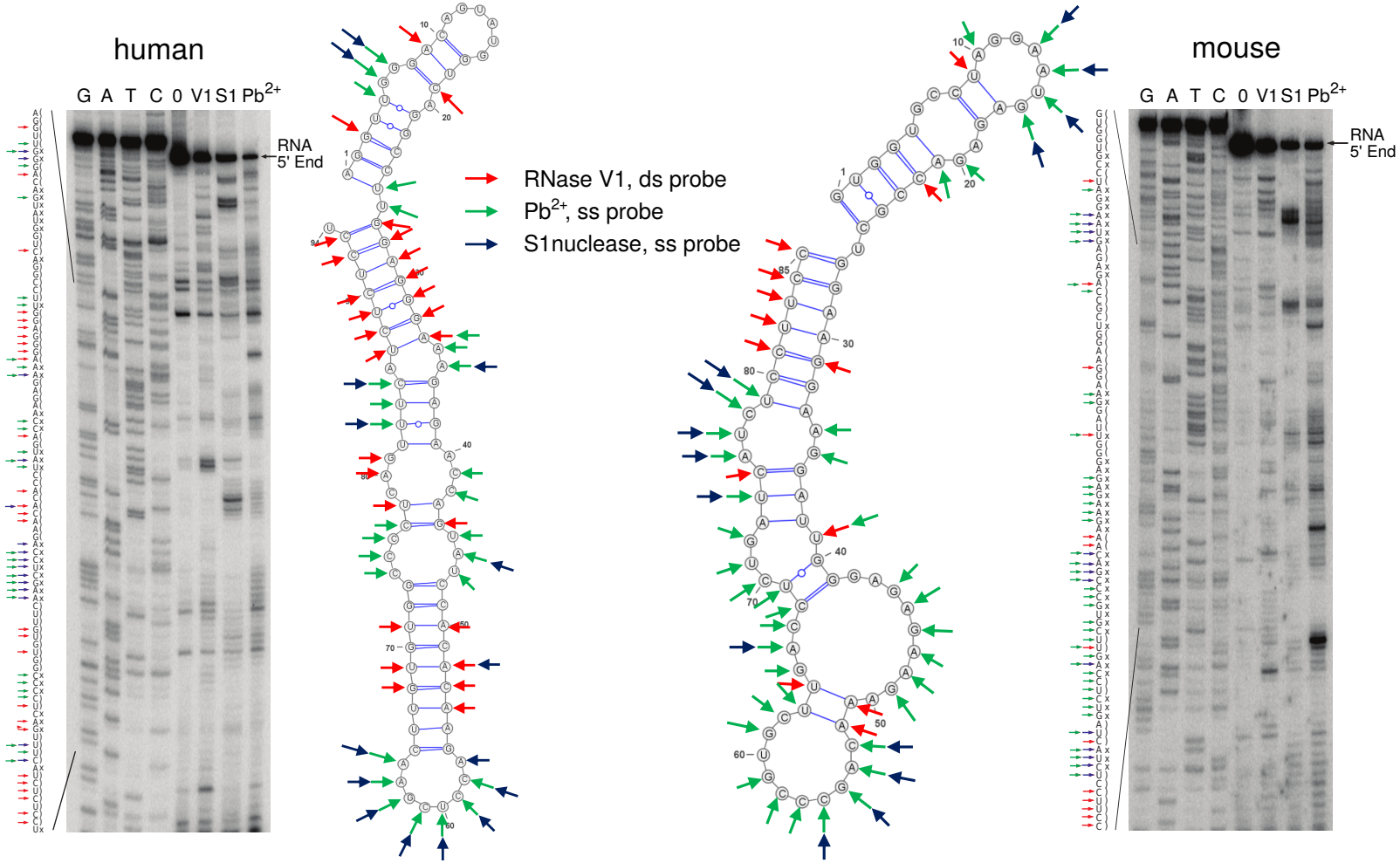
Transcribed enhancer RNAs are enriched for CRSs



6k structured enhancers are compared to 37k non-structured enhancers; expression is measured by CAGE in FANTOM5 [Andersson (2014) *Nature*]

CRSs putatively serve active regulatory function

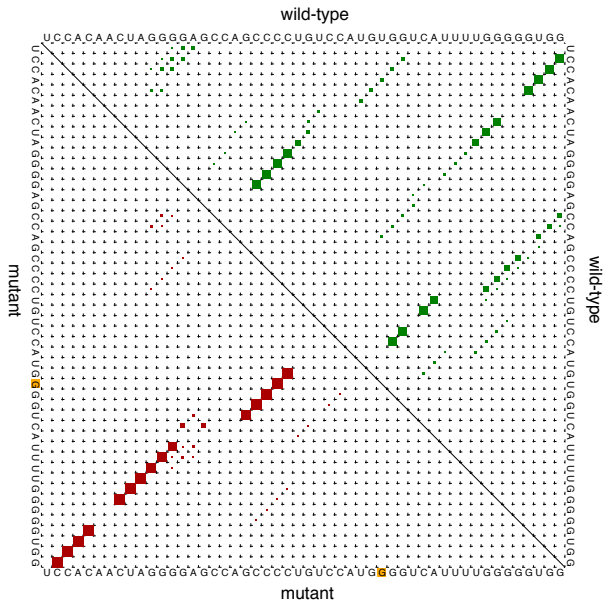
Intergenic M1695693 is potential enhancer RNA (SI=46%; FDR=9.93)

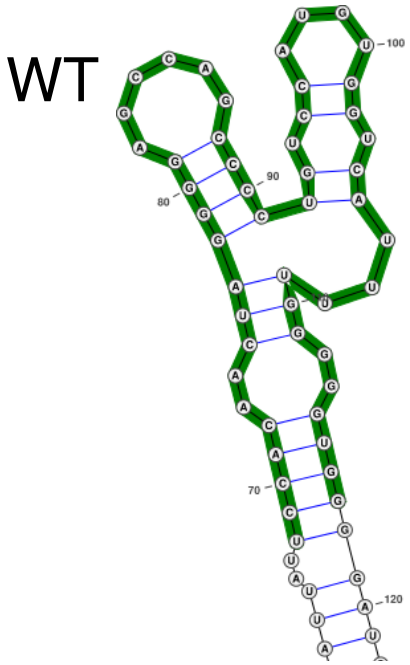


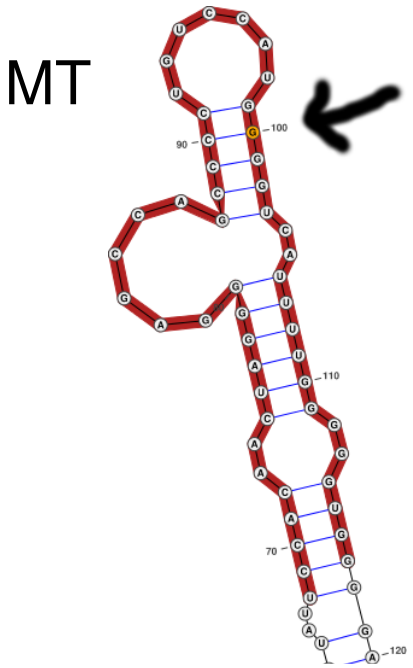
Overlaps RNA binding site of FMR1 (fragile X mental retardation 1; CLIP-seq)

Disease-associated SNPs potentially alter RNA structure

- Vast majority of disease variants (SNPs) identified by GWAS are noncoding
- disease-associated SNPs [Farh (2014)] are enriched for CRSs (OR=89)
- 21% of these SNPs significantly change local RNA structure (RNAsnp [Sabarinathan (2013) *Hum Mutat*; `http://rth.dk/resources/rnasnp/`] ; $p<0.1$)
- An example: CRS/rs2359796 overlaps enhancer region







Seemann et al. 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