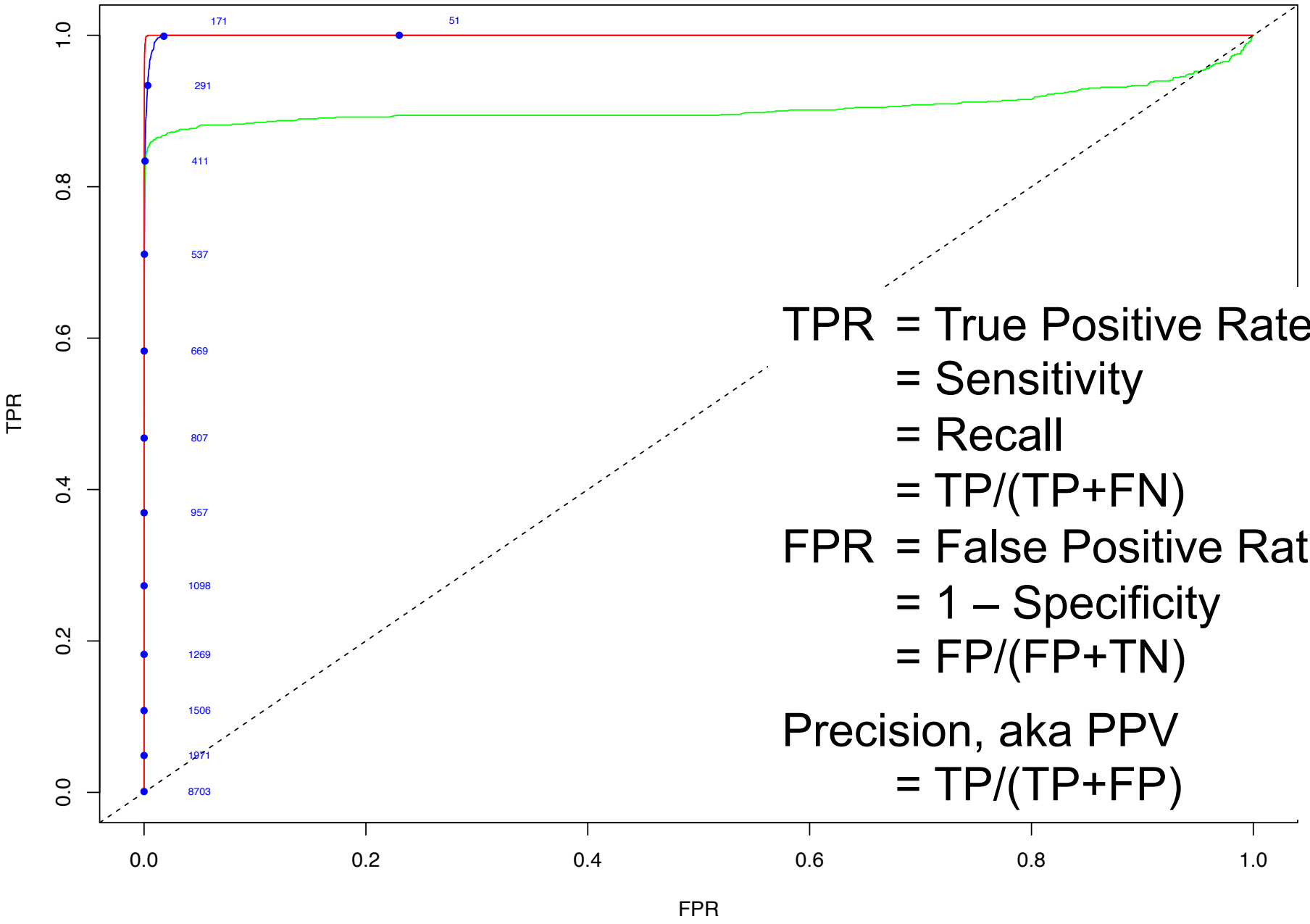
HW5

len	mun	gbk	mmtp	mmp	avg	mm	avg	mm	avg	mm	avg	
3	8381	0	0	0	0.00	0	-0.10	0	0.00	0	0.00	
6	8966	0	0	5092	-0.02	0	-0.10	5458	-0.03	5458	-0.03	5
9	9523	0	0	4946	-0.10	2690	-0.21	5064	-0.10	5064	-0.10	5
12	8622	0	0	4227	-0.15	3143	-0.25	4298	-0.14	4298	-0.14	4
15	7042	0	0	3366	-0.20	2829	-0.28	3458	-0.19	3458	-0.19	3
18	6475	0	0	2971	-0.26	2595	-0.34	3037	-0.25	3037	-0.25	3
21	6223	0	0	2754	-0.33	2500	-0.41	2883	-0.32	2883	-0.32	2
159	142	0	0	9	-7.39	10	-7.91	10	-7.93	10	-7.93	
162	105	1	1	6	-8.75	6	-9.33	6	-9.38	6	-9.38	
165	113	0	0	6	-7.83	7	-8.36	8	-8.34	8	-8.34	
168	100	0	0	6	-7.93	6	-8.41	7	-8.43	7	-8.43	
171	117	0	0	8	-8.21	9	-8.76	9	-8.80	9	-8.80	
174	80	1	1	6	-7.85	6	-8.39	6	-8.33	6	-8.33	
177	103	0	0	8	-7.91	8	-8.44	8	-8.43	8	-8.43	
5253	1	1	1	1	148.45	1	174.37	1	174.80	1	174.80	
8703	1	1	1	1	61.87	1	80.09	1	80.37	1	80.37	

mm	avg	mm	avg	mm	avg	mm	avg	mm	avg	mm	avg
0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
458	-0.03	5458	-0.03	5458	-0.03	5189	-0.02	5458	-0.03	5150	-0.02
064	-0.10	5064	-0.10	5064	-0.10	5226	-0.09	5080	-0.10	5123	-0.10
298	-0.14	4298	-0.14	4298	-0.14	4363	-0.13	4323	-0.14	4325	-0.14
458	-0.19	3458	-0.19	3458	-0.19	3471	-0.18	3485	-0.18	3464	-0.18
037	-0.25	3037	-0.25	3037	-0.25	3061	-0.25	3058	-0.24	3051	-0.25
883	-0.32	2883	-0.32	2883	-0.32	2852	-0.31	2906	-0.31	2887	-0.31
10	-7.93	10	-7.93	11	-7.60	9	-7.90	10	-7.77	10	-7.90
6	-9.38	6	-9.38	7	-9.08	6	-9.37	7	-9.21	6	-9.38
8	-8.34	8	-8.34	8	-7.99	7	-8.35	9	-8.17	8	-8.30
7	-8.43	7	-8.43	9	-8.09	8	-8.44	7	-8.28	6	-8.39
9	-8.80	9	-8.80	10	-8.49	9	-8.78	9	-8.63	9	-8.77
6	-8.33	6	-8.33	6	-8.00	6	-8.37	6	-8.17	6	-8.26
8	-8.43	8	-8.43	10	-8.16	8	-8.45	8	-8.25	9	-8.41
1	174.80	1	174.80	1	174.80	1	172.22	1	177.83	1	180.67
1	80.37	1	80.37	1	80.37	1	71.00	1	86.31	1	84.79

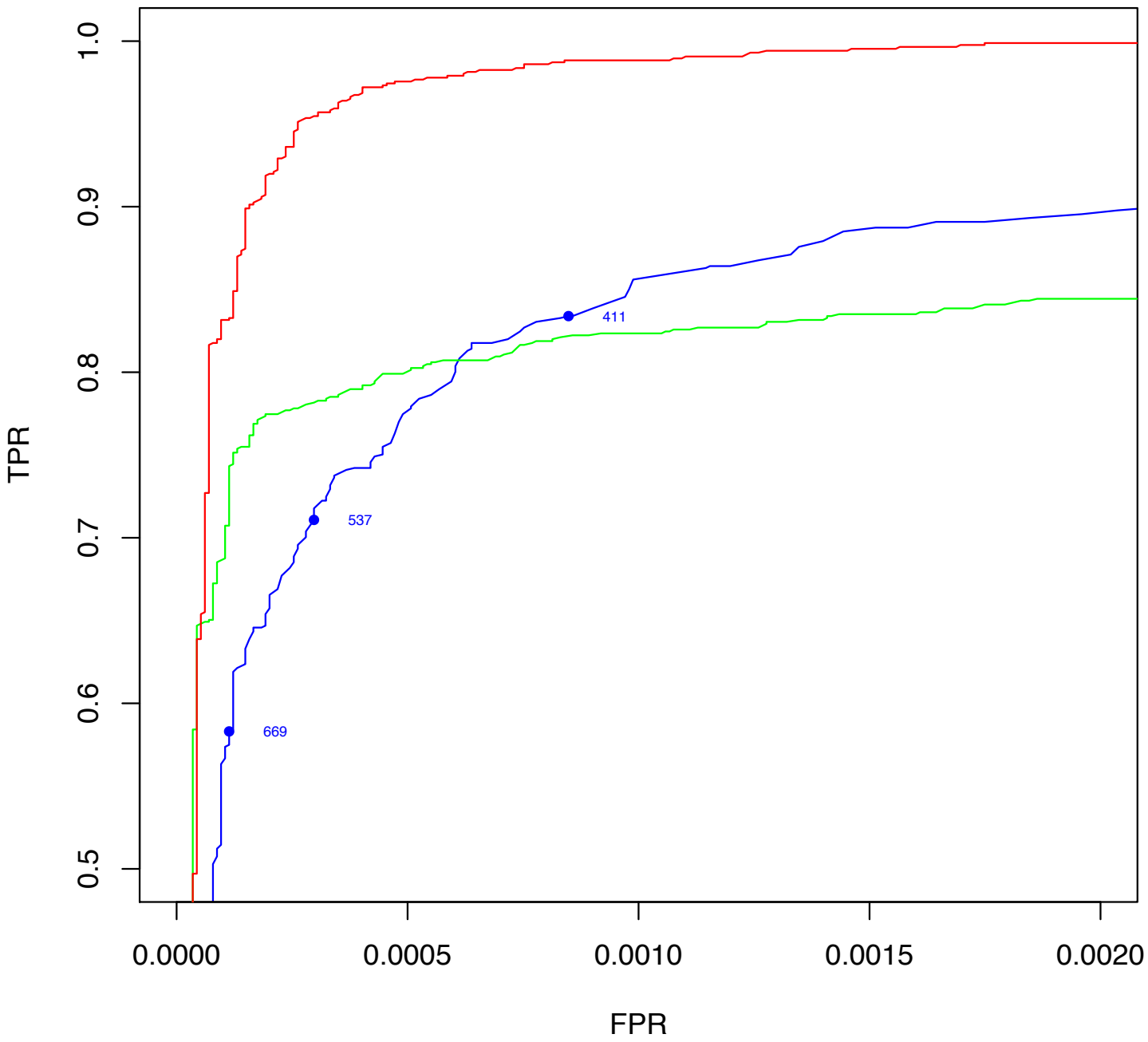
ROC

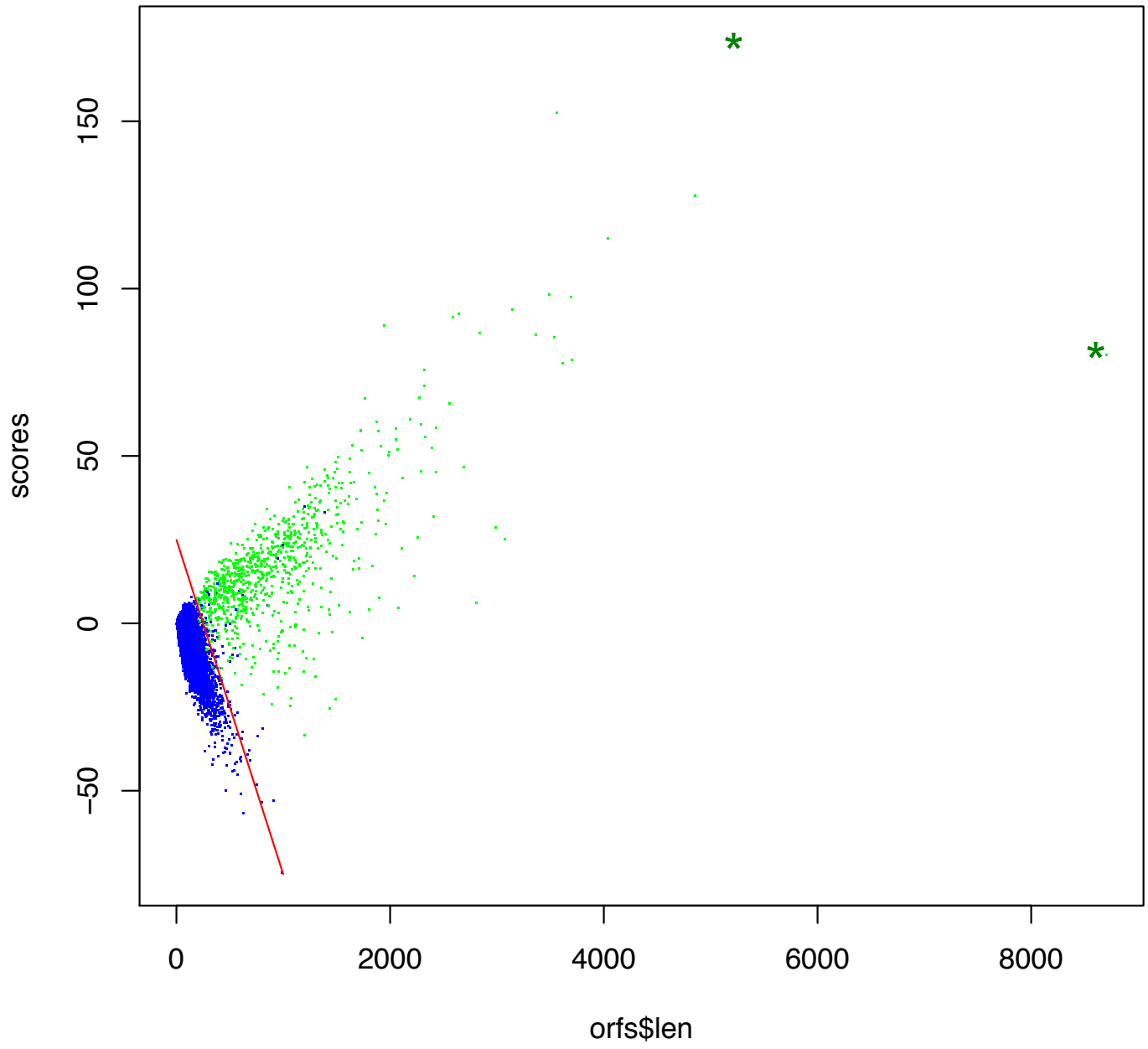


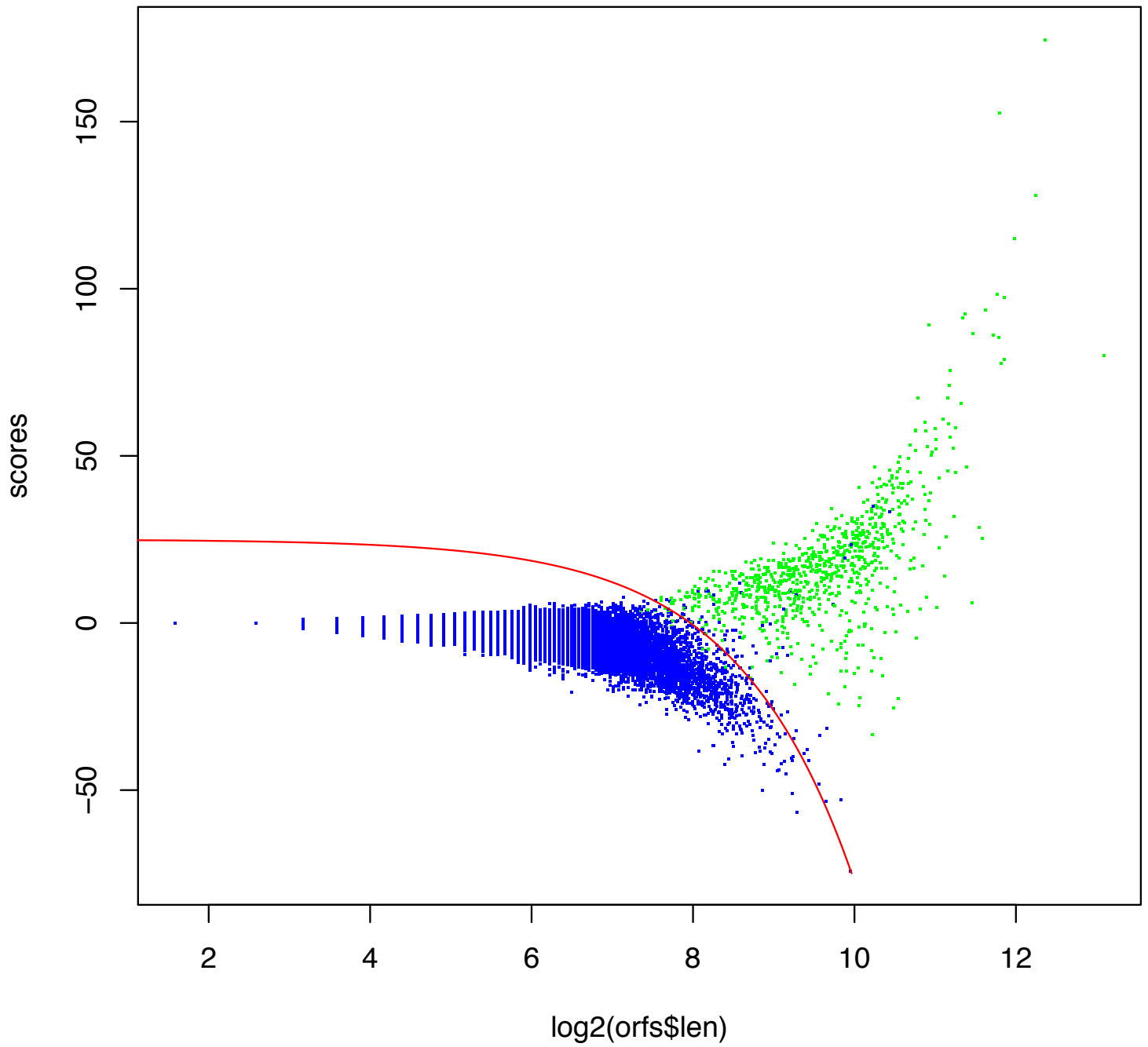
TPR = True Positive Rate
= Sensitivity
= Recall
= TP/(TP+FN)

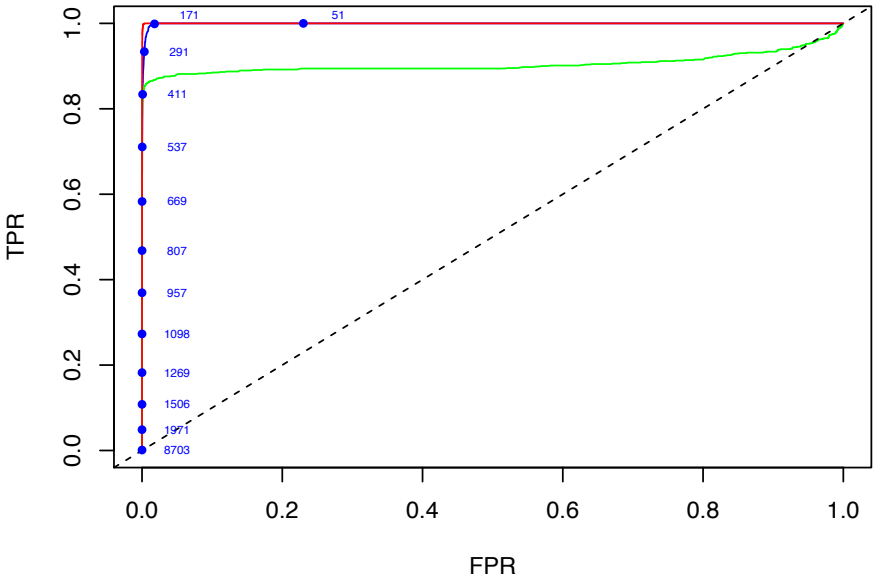
FPR = False Positive Rate
= 1 – Specificity
= FP/(FP+TN)

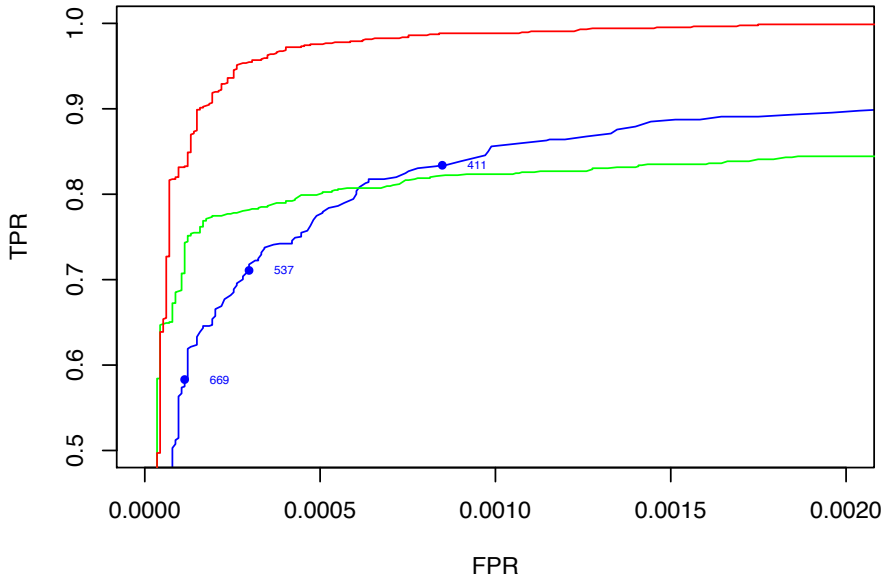
Precision, aka PPV
= TP/(TP+FP)

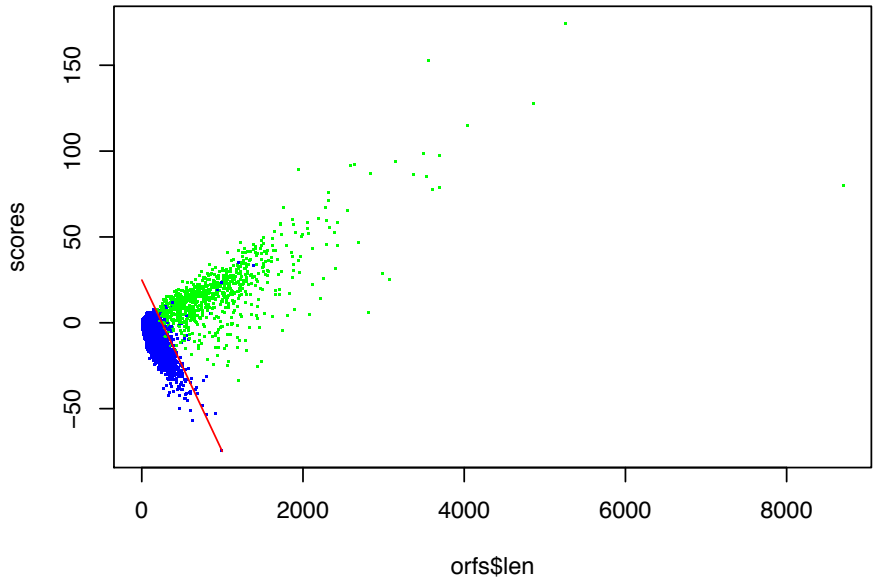

TPR
1.0
0.9
0.8
0.7
0.6
0.5
0.0000
0.0005
0.0010
0.0015
0.0020
FPR
411
537
669

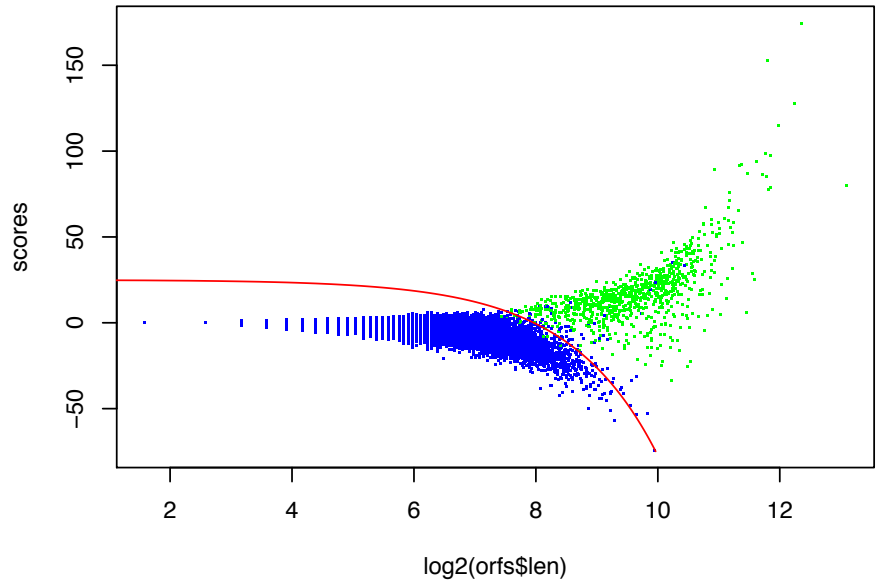

scores
150
100
50
0
−50
0
2000
4000
6000
8000
orfs$len


scores
150
100
50
0
−50
2
4
6
8
10
12
log2(orfs$len)


TPR
FPR
171
51
291
411
537
669
807
957
1098
1269
1506
1971
8703

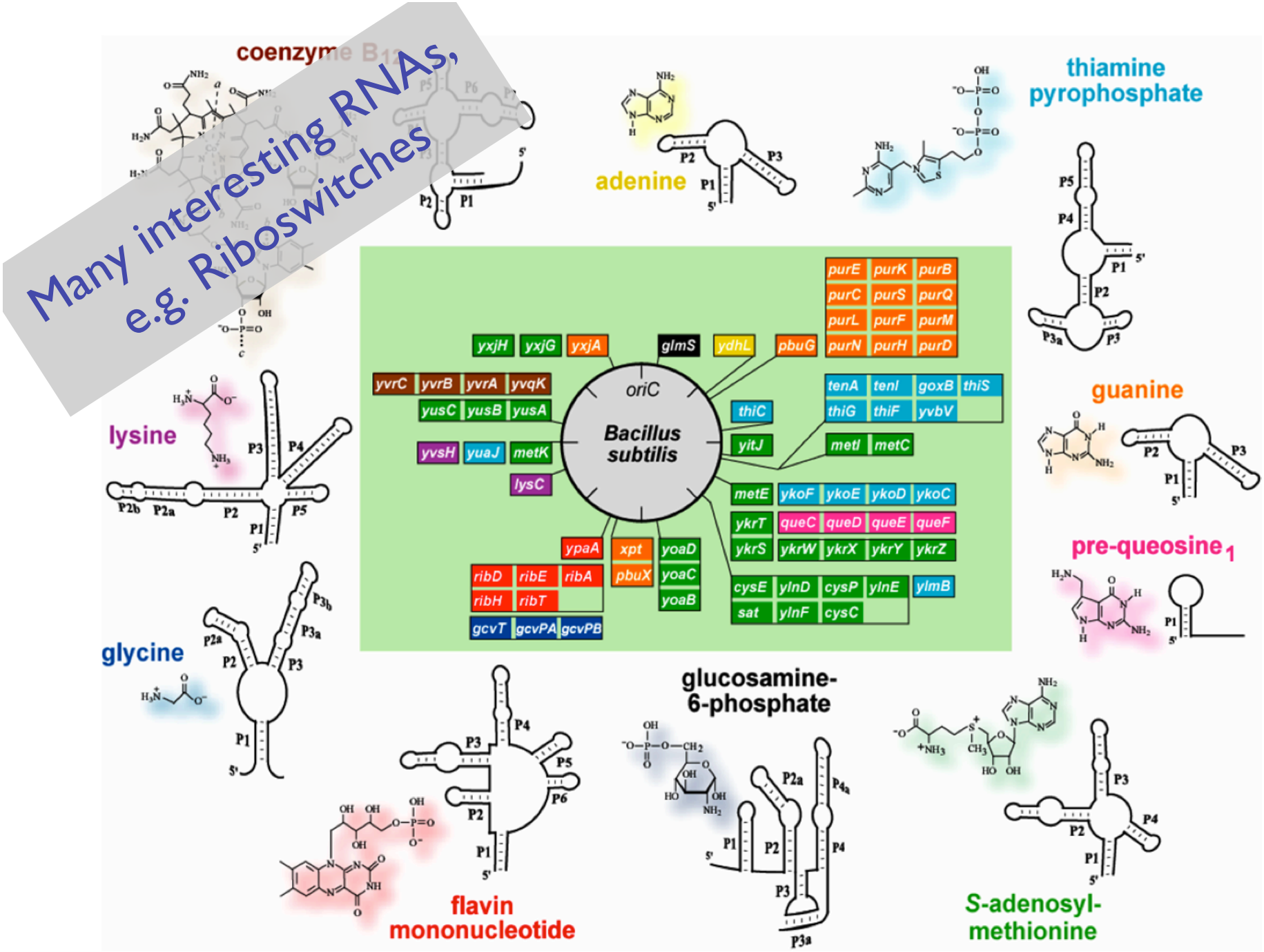

TPR
FPR
411
537
669


scores
orfs$len


scores
log2(orfs$len)

RNA Search and Motif Discovery

CSEP 590 A
Computational Biology


Many interesting RNAs, e.g. Riboswitches
coenzyme B12
adenine
thiamine pyrophosphate
guanine
lysine
glycine
pre-queosine 1
glucosamine-6-phosphate
flavin mononucleotide
S-adenosyl-methionine
oriC
Bacillus subtilis
yxjH yxjG yxjA
glmS
ydhL
pbuG
purE purK purB
purC purS purQ
purL purF purM
purN purH purD
yvrC yvrB yvrA yvqK
yusC yusB yusA
yvsH yuaJ metK
lysC
thiC
yitJ
tenA tenI goxB thiS
thiG thiF yvbV
metI metC
metE ykoF ykoE ykoD ykoC
ykrT queC queD queE queF
ykrS ykrW ykrX ykrY ykrZ
ypaA xpt yoaD
pbuX yoaC
yoaB
ribD ribE ribA
ribH ribT
gcvT gcvPA gcvPB
cysE ylnD cysP ylnE ylmB
sat ylnF cysC

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

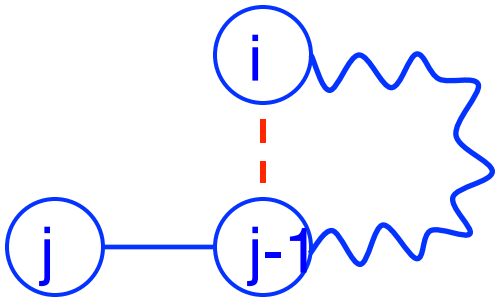
Structure Prediction

"Optimal pairing of r_i ... r_j"

Two possibilities

j Unpaired:

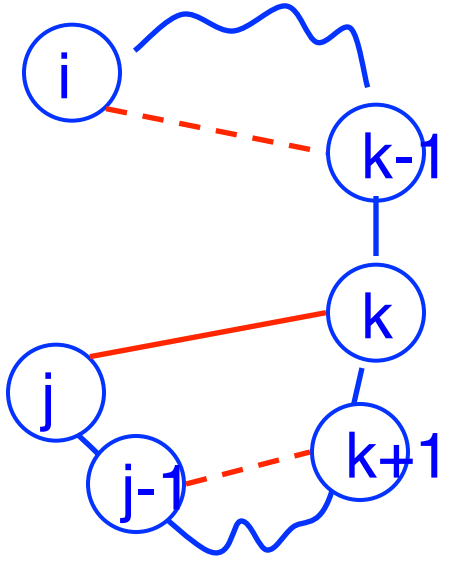
Find best pairing of r_i ... r_{j-1}



j Paired (with some k):

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

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

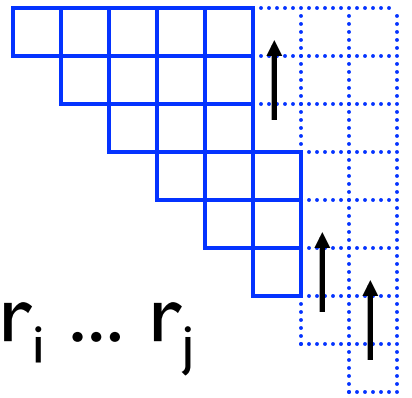


Why is it slow?

Why do pseudoknots matter?

Structure Prediction

Nussinov: Computation Order



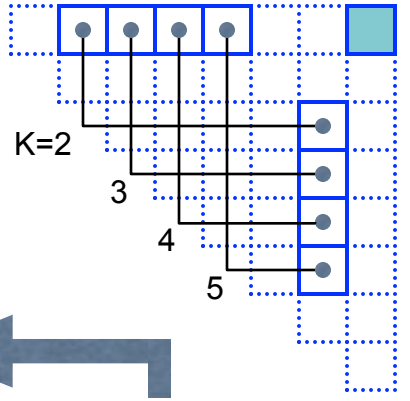
Or energy

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

Today

Structure prediction via comparative analysis

Covariance Models (CMs) represent
    RNA sequence/structure motifs

Fast CM search

Motif Discovery

Applications in prokaryotes & vertebrates

Structure Prediction

Approaches, II

Comparative sequence analysis

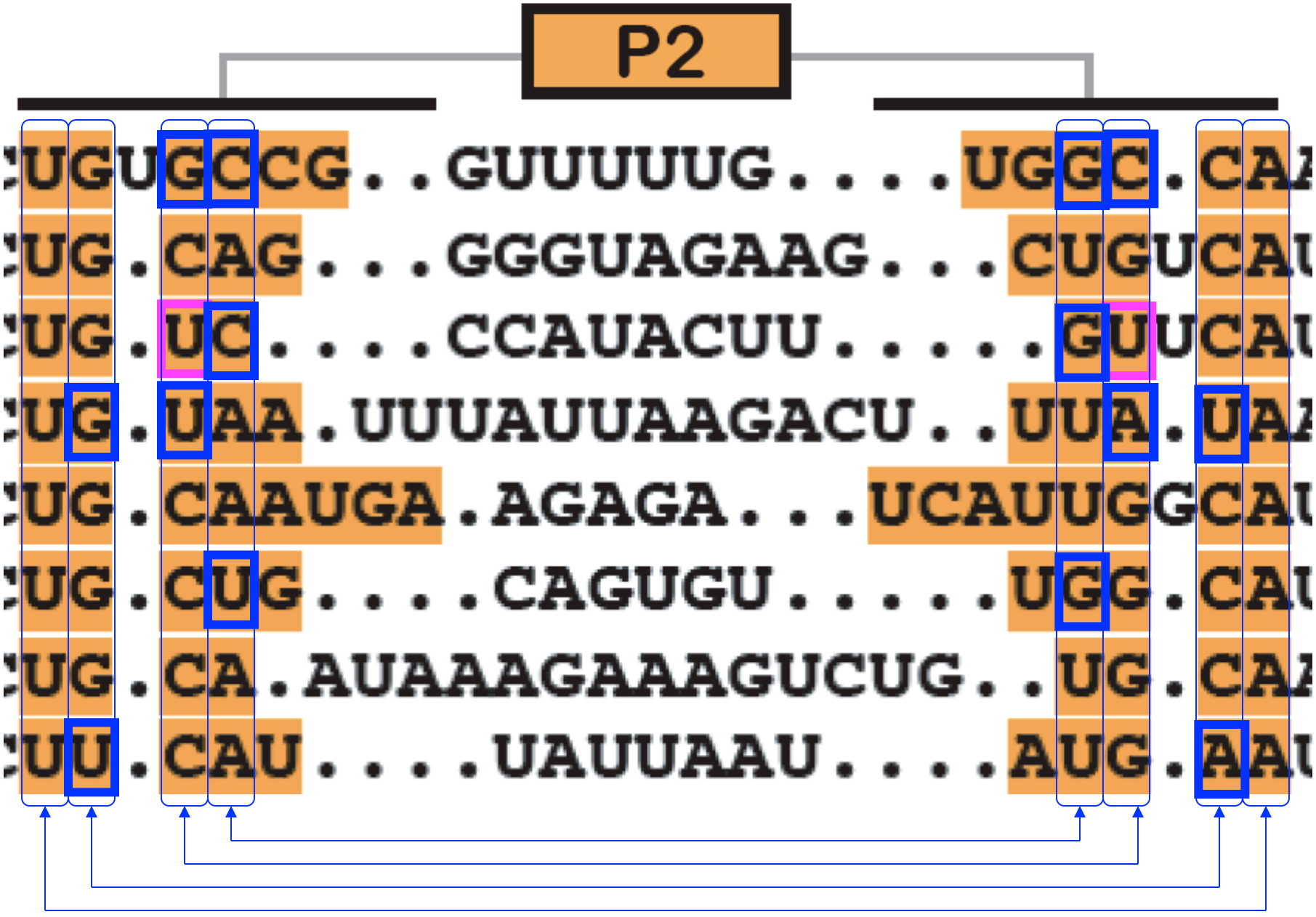
+ handles all pairings (potentially incl. pseudoknots)
- requires several (many?) aligned, appropriately diverged sequences

Stochastic Context-free Grammars

Roughly combines min energy & comparative, but no pseudoknots

Physical experiments (x-ray crystalography, NMR)

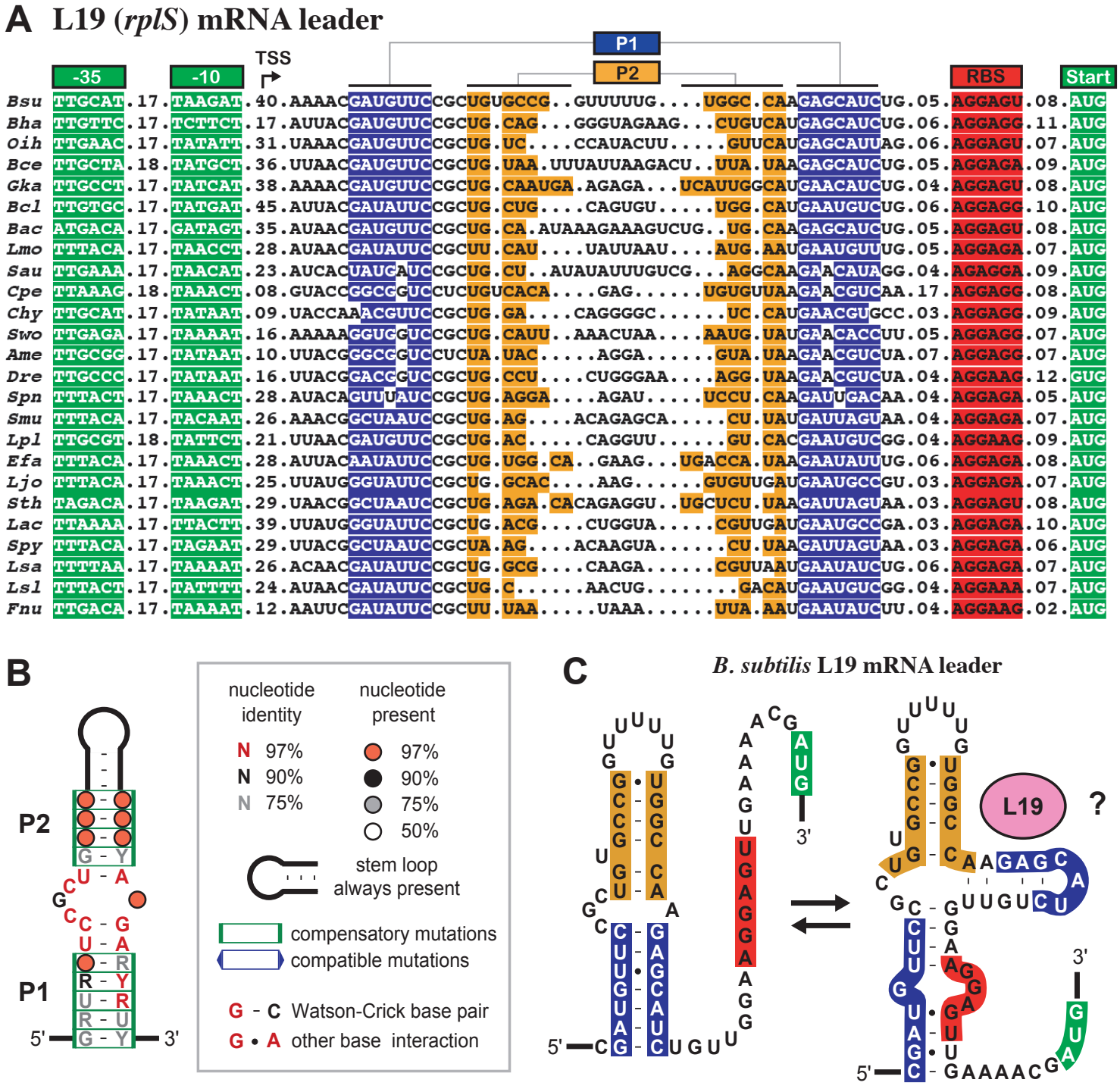
Today



Covariation is strong evidence for base pairing

Example: Ribosomal Autoregulation:

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



Mutual Information

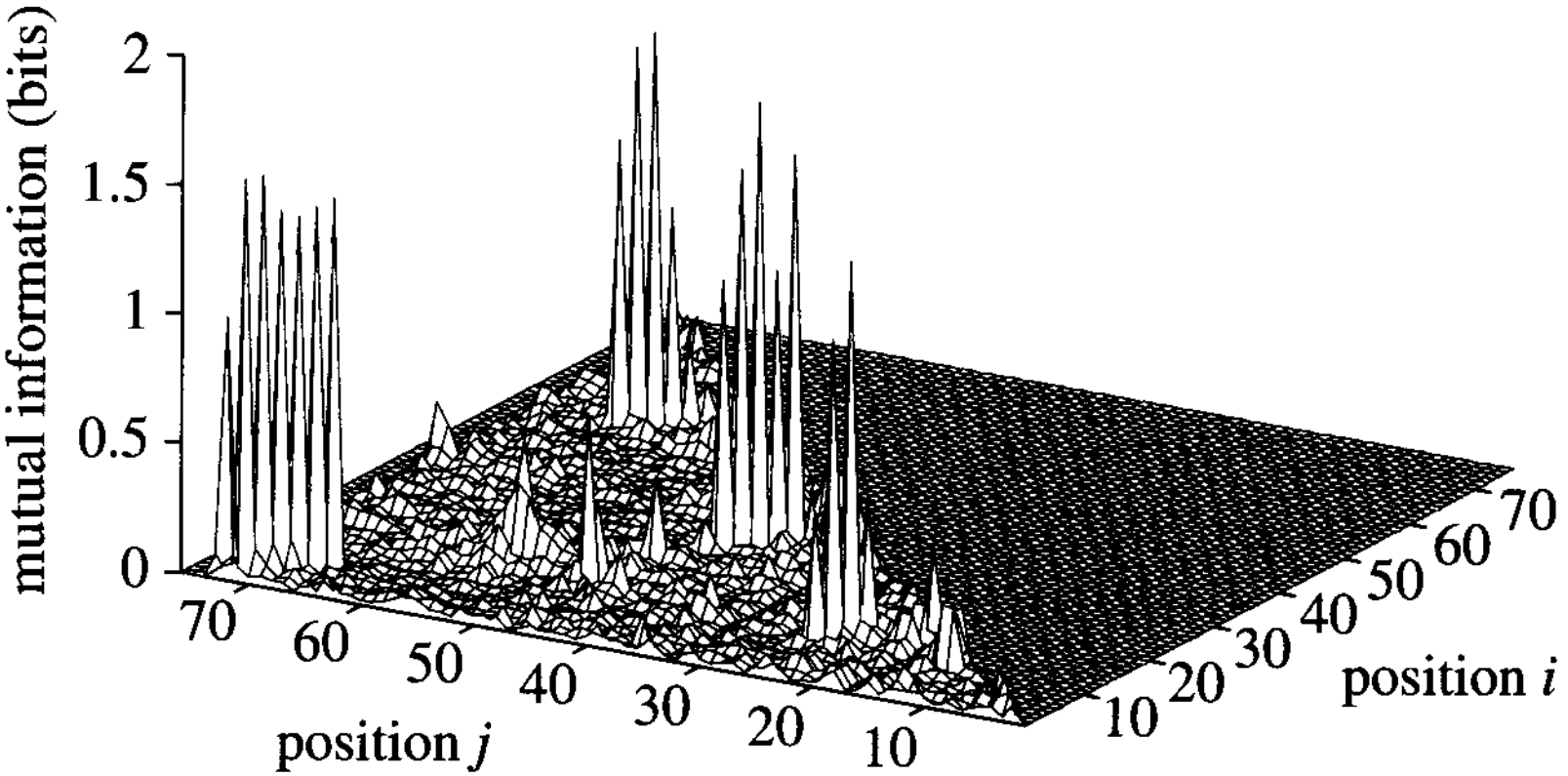
$$M_{ij} = \sum_{xi,xj} f_{xi,xj} \log_2 \frac{f_{xi,xj}}{f_{xi} f_{xj}}; \quad 0 \le M_{ij} \le 2$$

Max when *no* seq conservation but perfect pairing

MI = expected score gain from using a pair state (below)

Finding optimal MI, (i.e. opt pairing of cols) is hard(?)

Finding optimal MI *without pseudoknots* can be done by dynamic programming



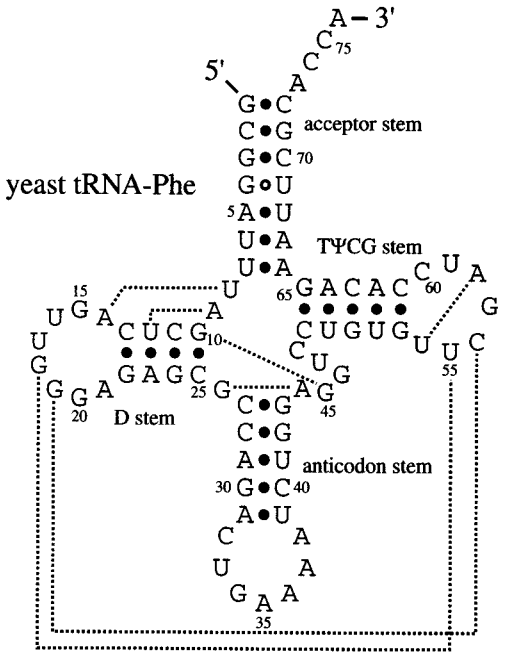


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary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

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
- 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 - current favorite)

Given sufficient data, model construction comparable to, but not quite as good as, human experts

Some quantitative info on importance of pseudoknots and other tertiary features

Probabilistic Model Search

As with HMMs, given a sequence, you calculate likelihood ratio that the model could generate the sequence, vs a background model

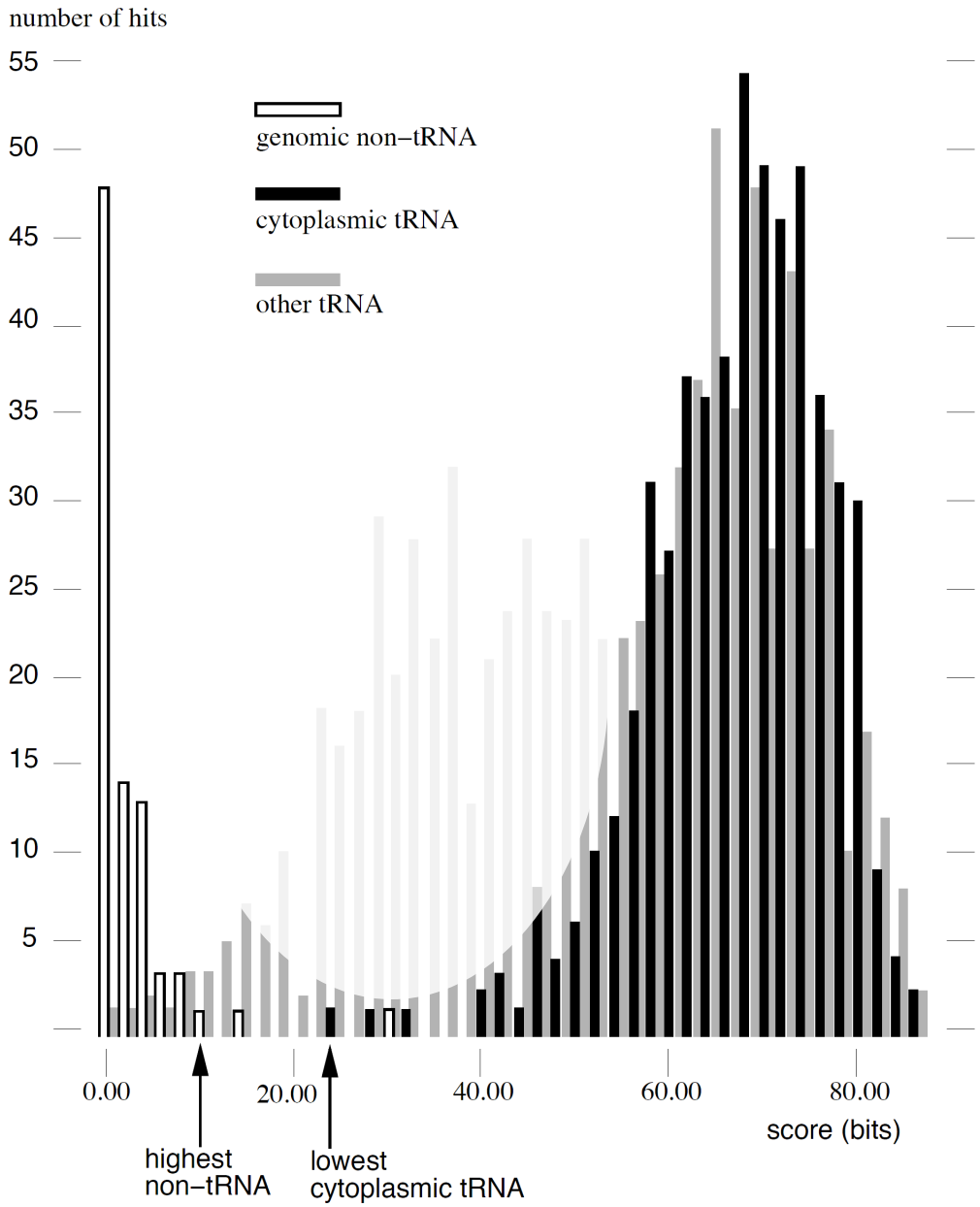
You set a score threshold

Anything above threshold → a "hit"

Scoring:

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

Example: searching for tRNAs



Profile Hmm Structure

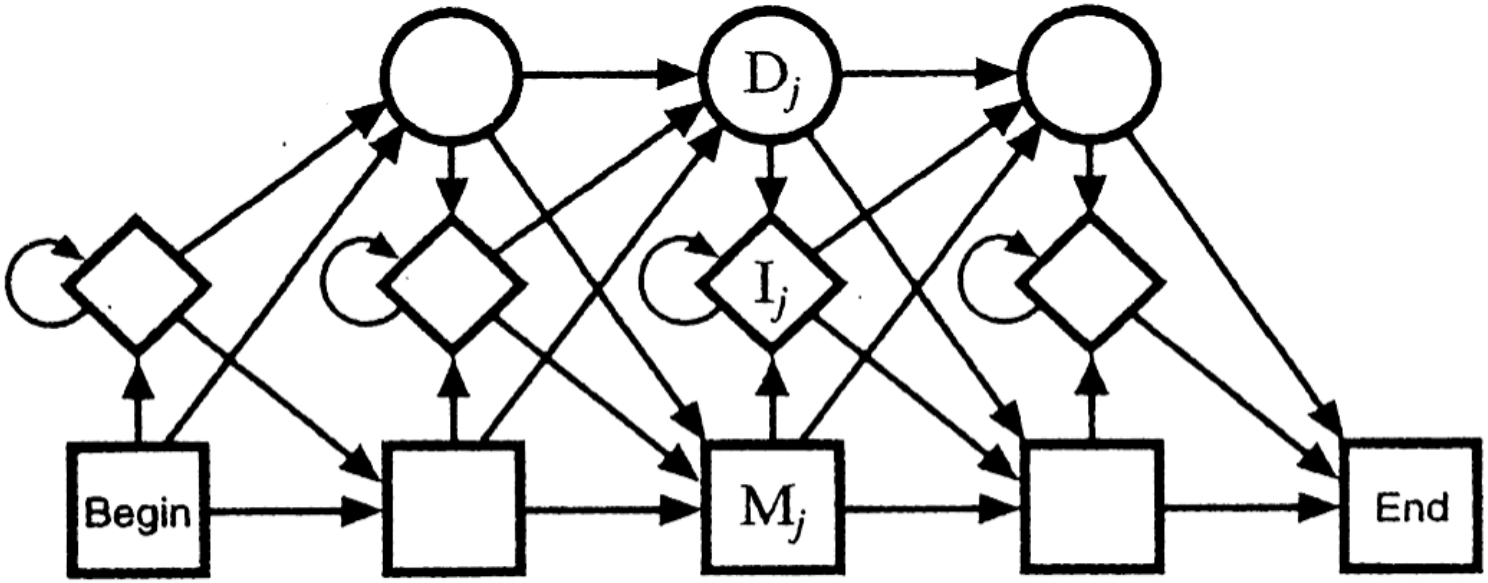


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

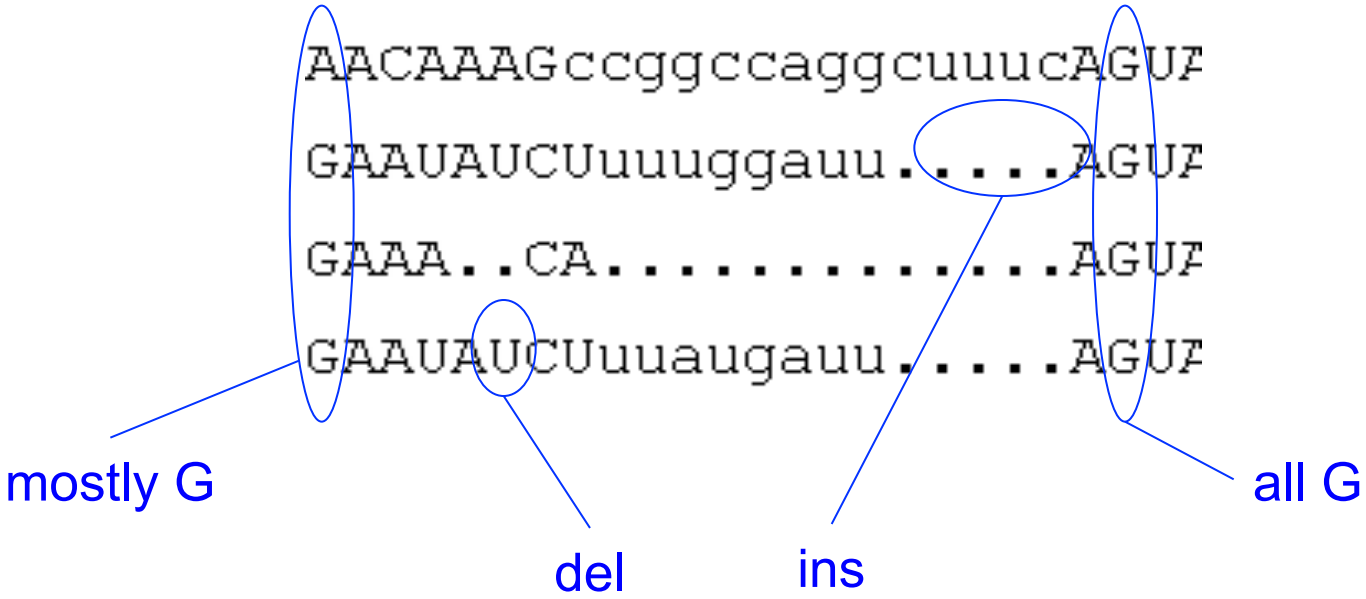
D_j: Delete states (silent - no emission)

How to model an RNA "Motif"?

Conceptually, start with a profile HMM:

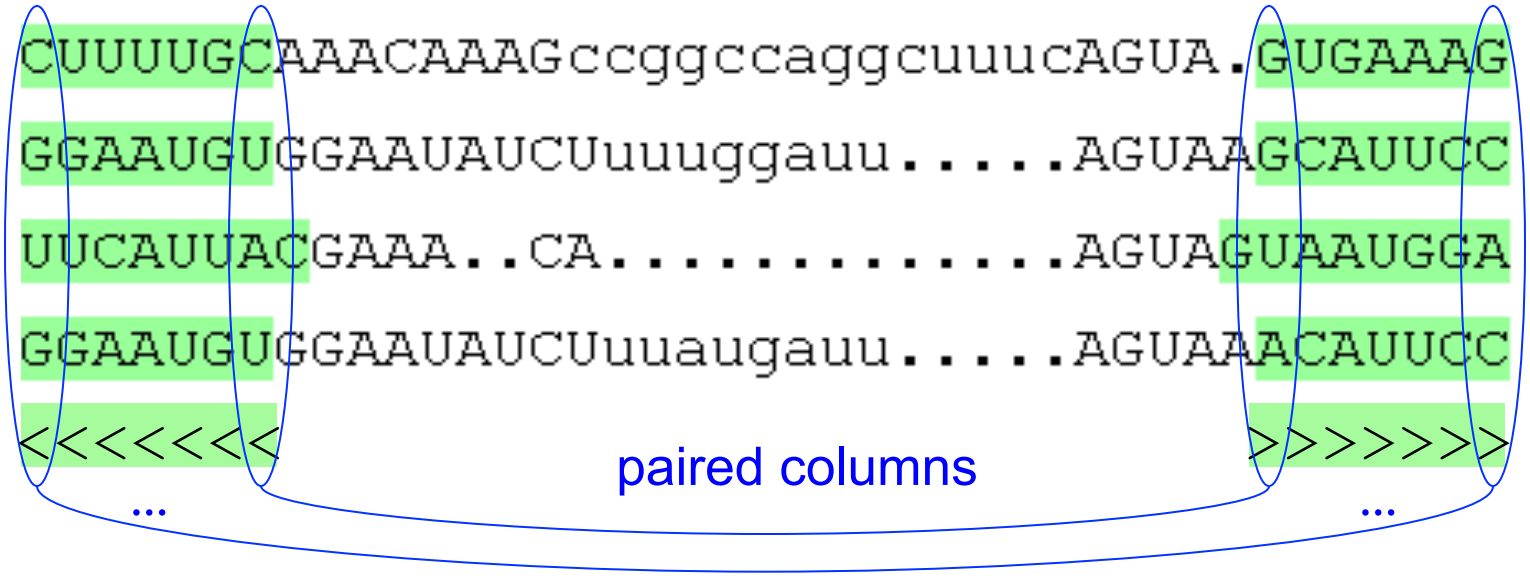
from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position

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

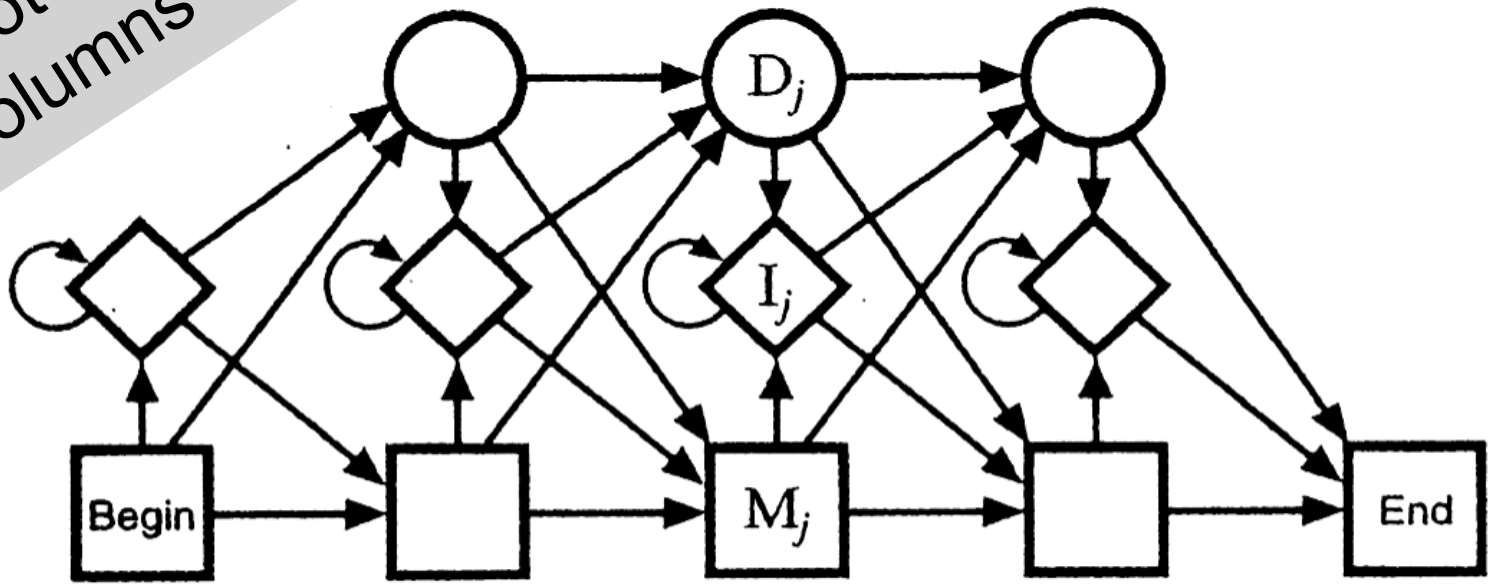


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (20 emission probabilities)

I_j: Insert states (Background emission probabilities)

D_j: Delete states (silent - no emission)

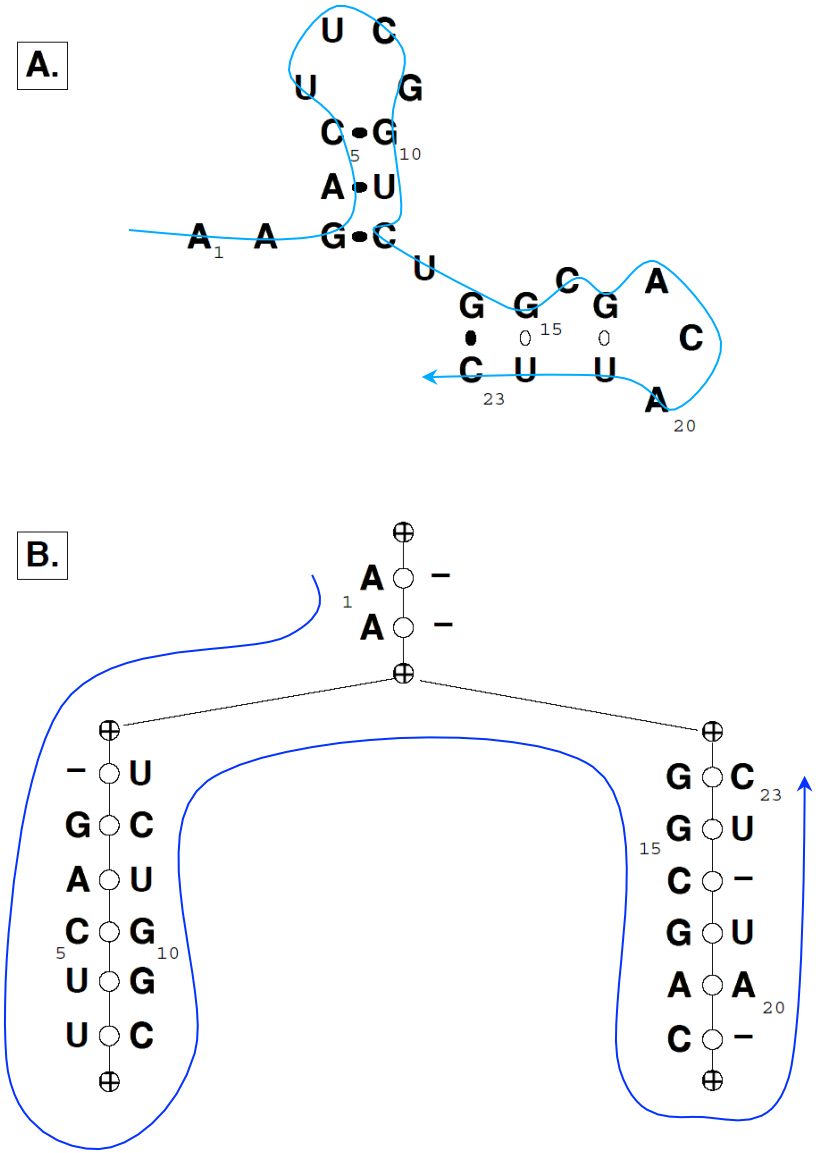
CM Structure

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)

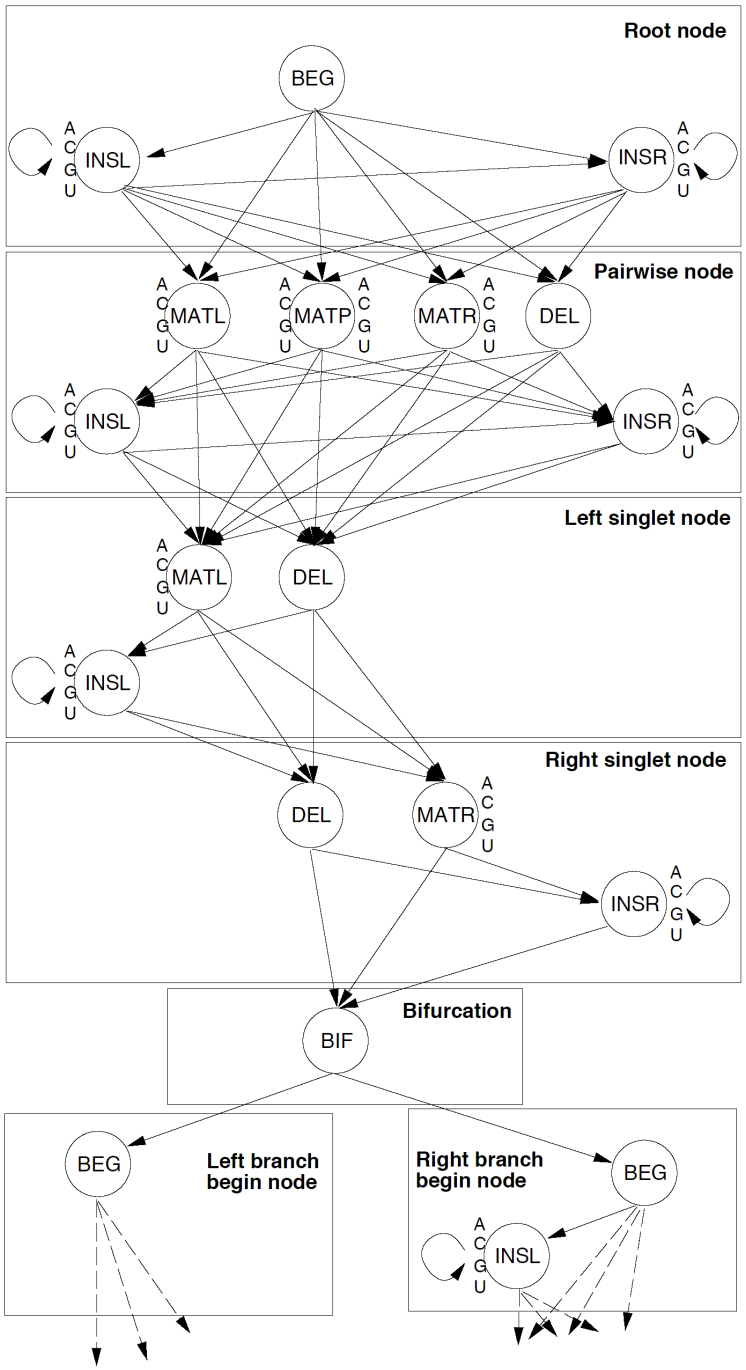


Overall CM Architecture

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

An Important Application: Rfam

Rfam – an RNA family DB

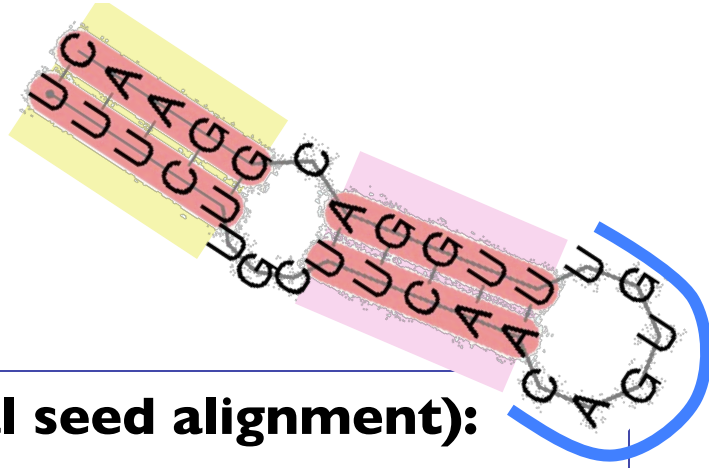
Griffiths-Jones, et al., NAR '03, '05, '08, '11, '12

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 9.1, 1/09:	1372 families, 1148k instances	
Rel 10.0, 1/10:	1446 families, 3193k instances	~160GB
Rel 11.0, 8/12:	2208 families, 6125k instances	~320GB

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

An Important Need: Faster Search

Homology search

"Homolog" – similar by descent from common ancestor

Sequence-based

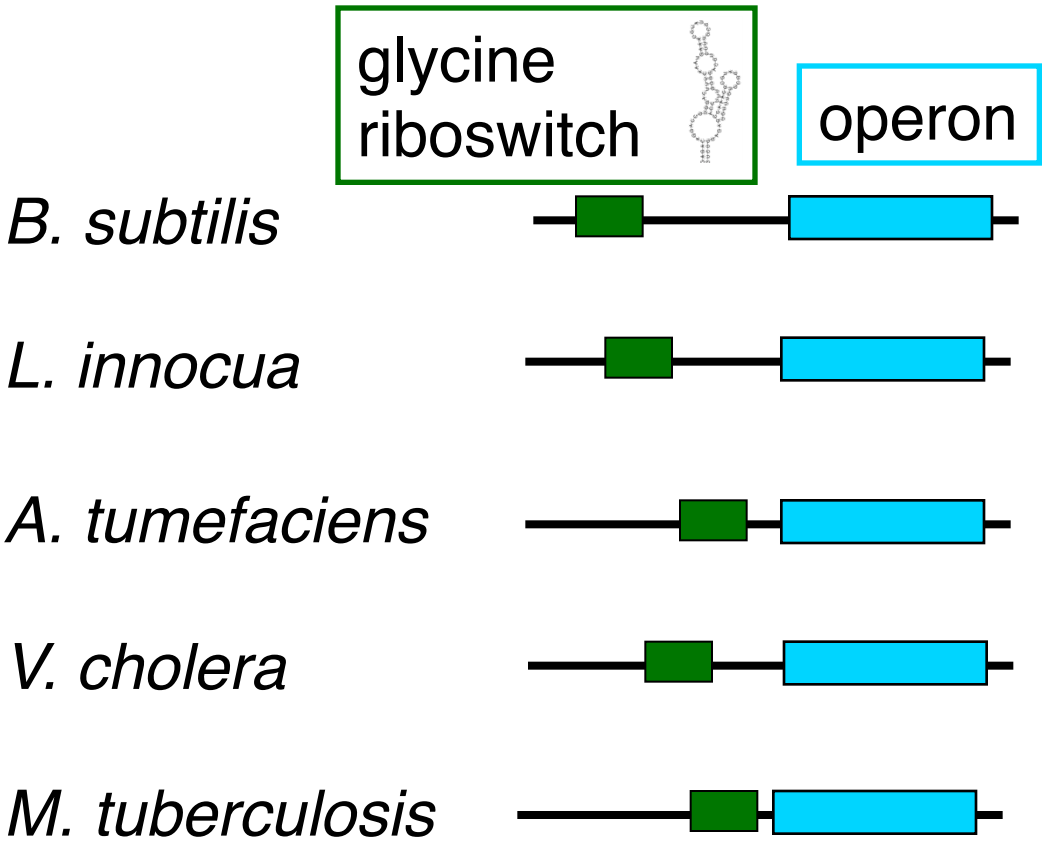
- Smith-Waterman
- FASTA
- BLAST

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

So, use structure, too

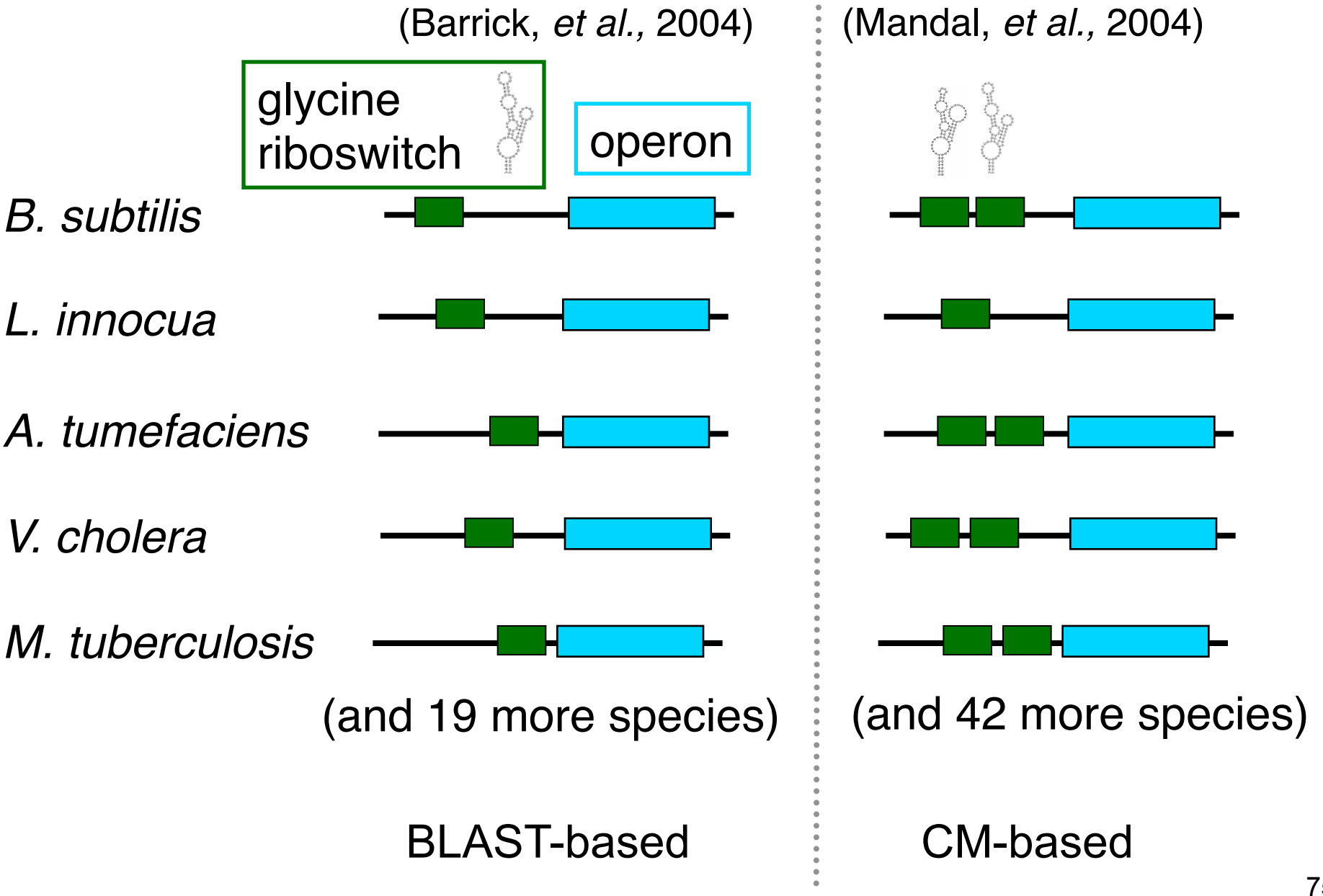
Impact of RNA homology search

(Barrick, *et al.,* 2004)



(and 19 more species)

Impact of RNA homology search



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

Zasha Weinberg

& W.L. Ruzzo

Recomb ‘04, ISMB ‘04, Bioinfo ‘06

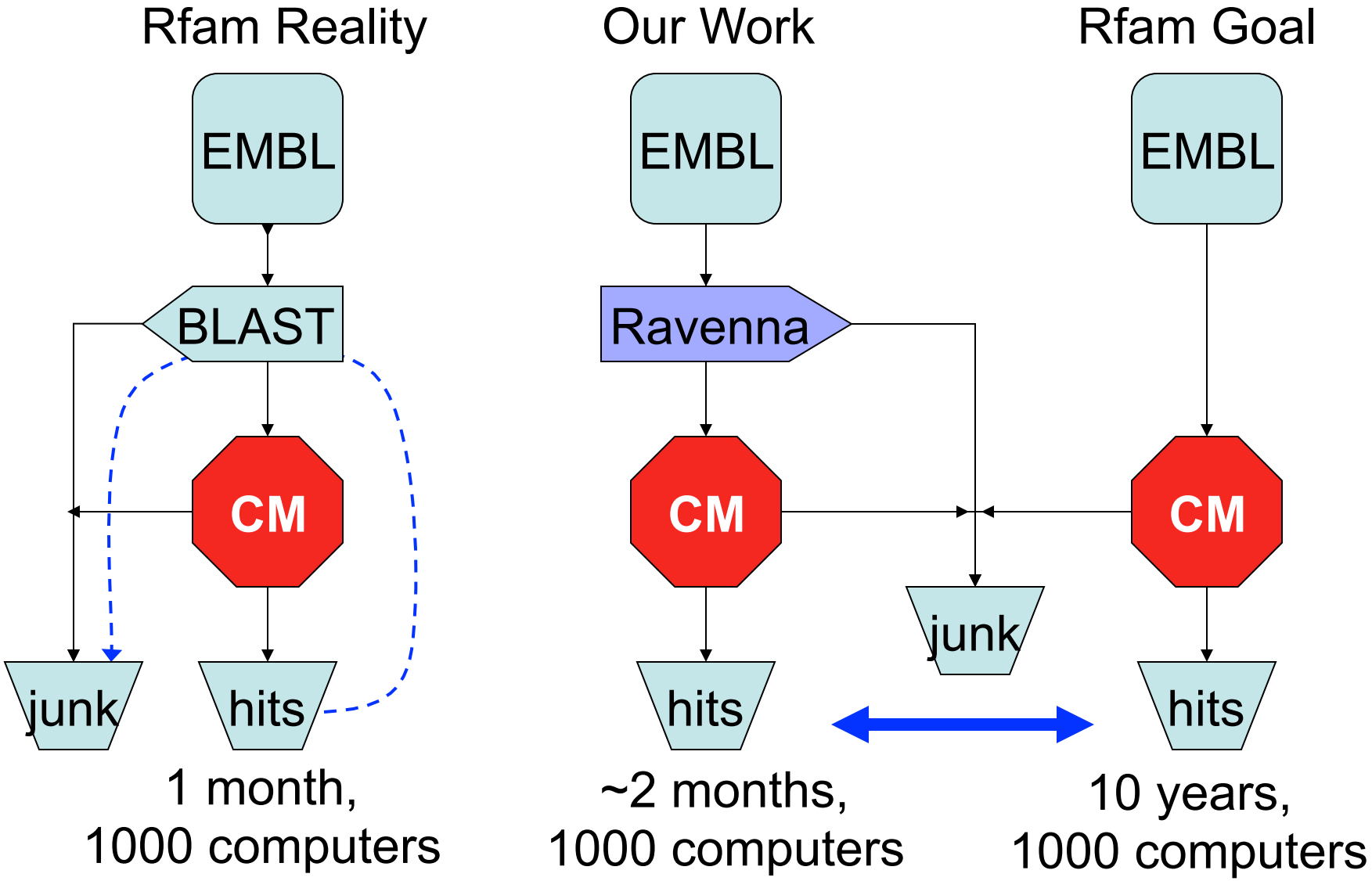
RaveNnA: Genome Scale RNA Search

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

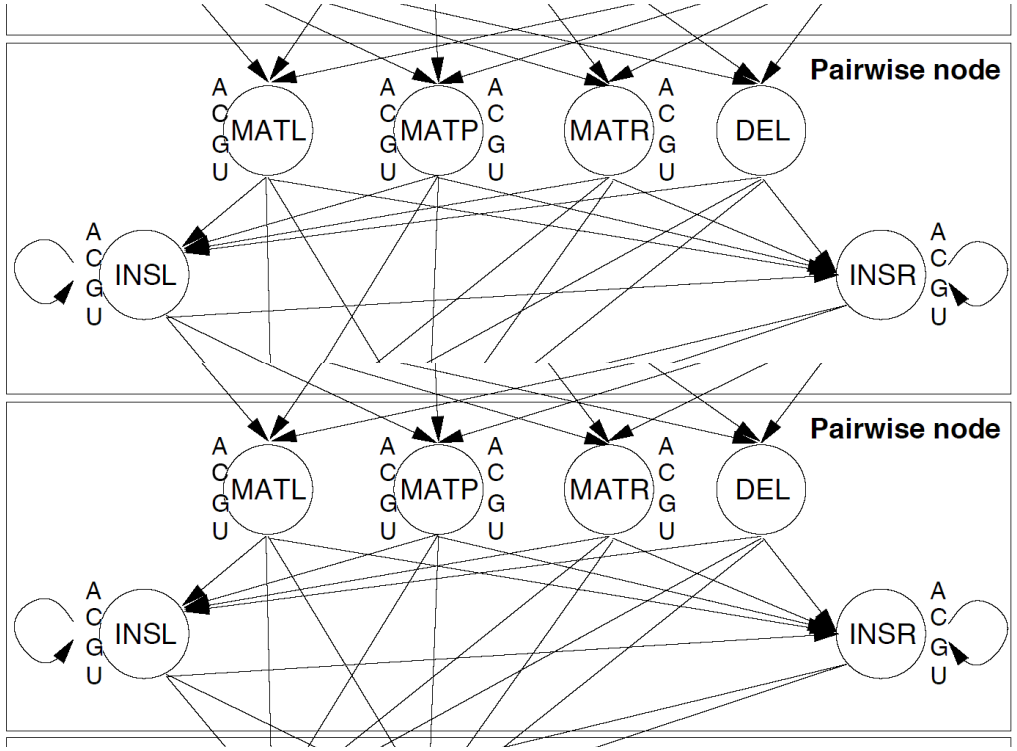
- Drop structure from CM to create a (faster) HMM
- Use that to pre-filter sequence;
- Discard parts where, provably, CM score < threshold;
- Actually run CM on the rest (the promising parts)
- Assignment of HMM transition/emission scores is key
  - (a large convex optimization problem)

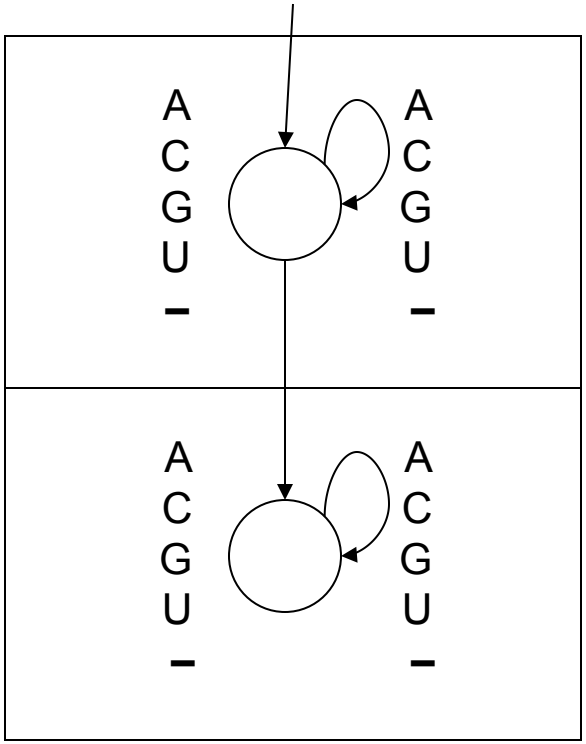
Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

CM's are good, but slow

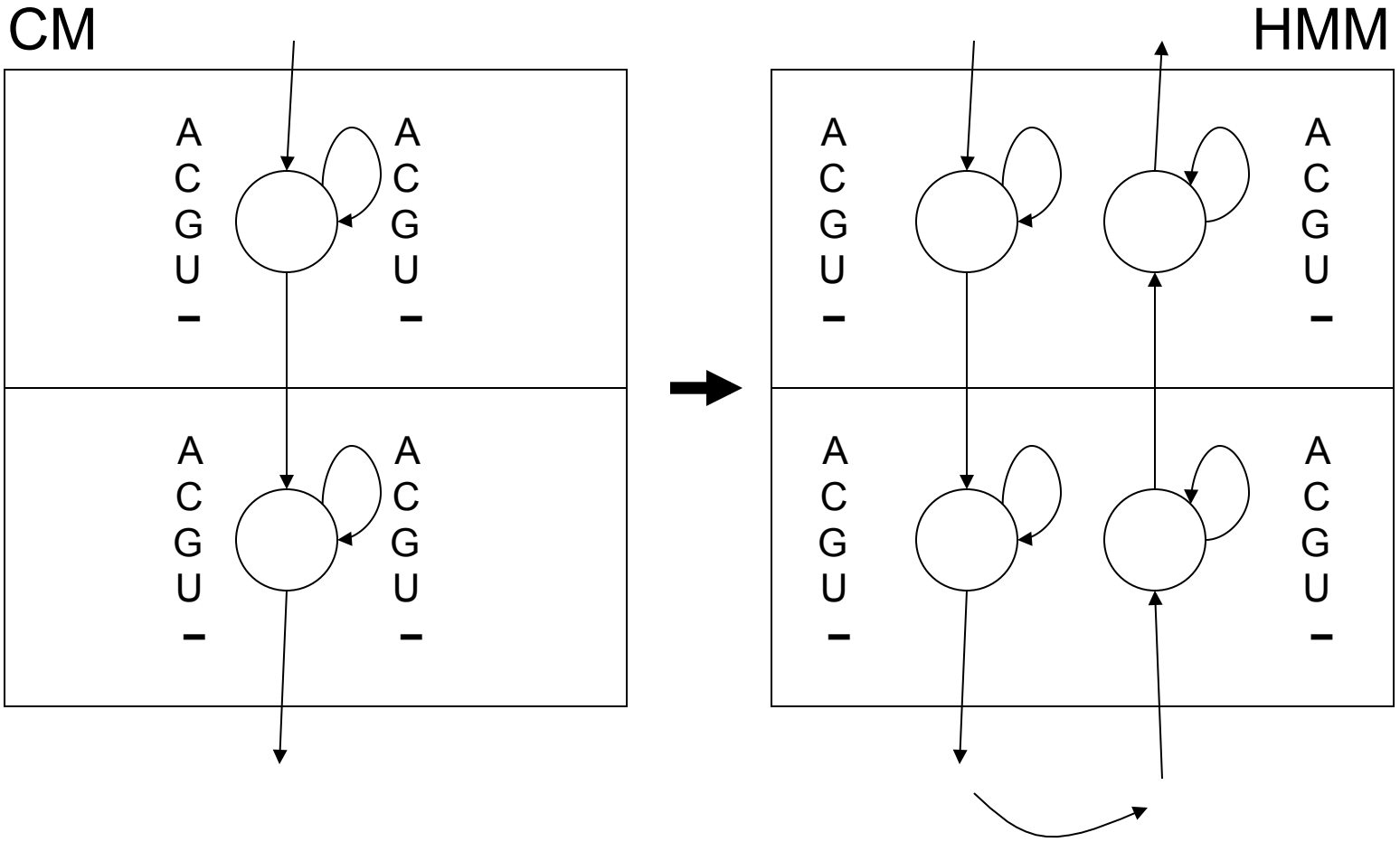


Oversimplified CM
(for pedagogical purposes only)





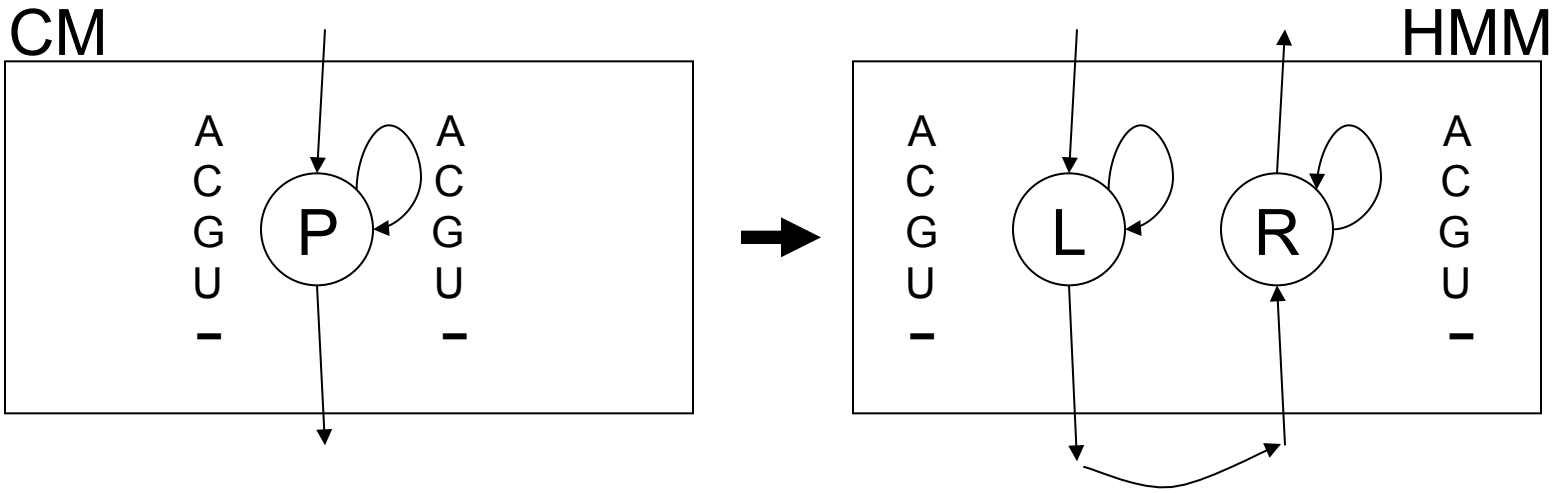
CM to HMM



25 emisions per state

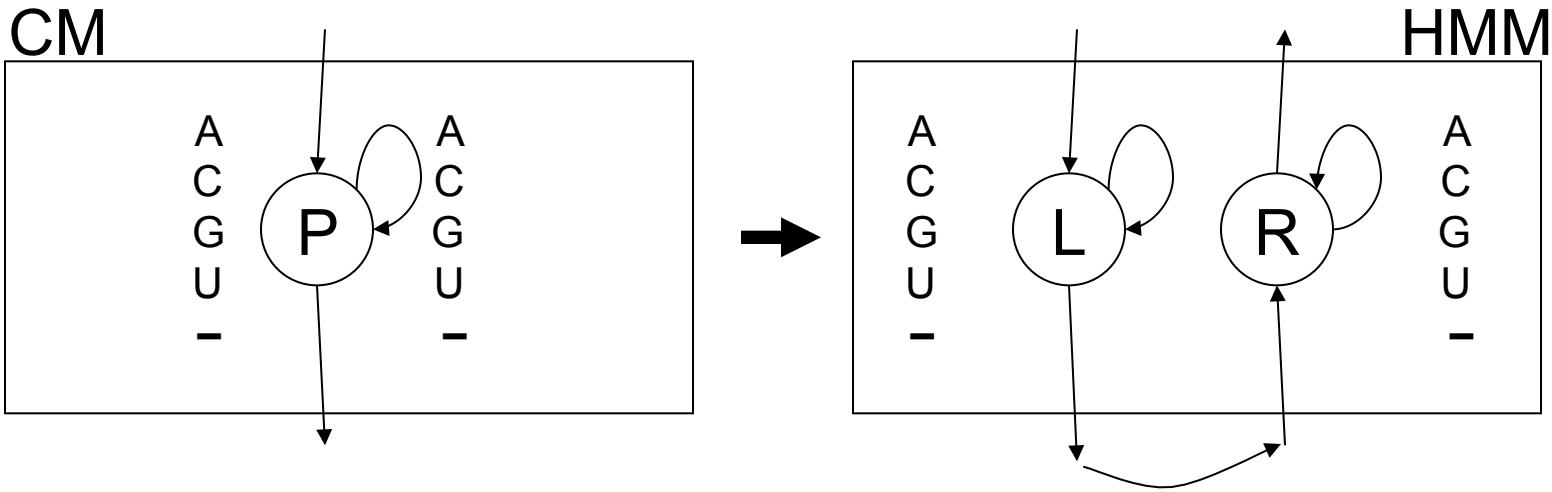
5 emissions per state, 2x states

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

Key Issue: 25 scores → 10



Need: log Viterbi scores CM ≤ HMM

$P_{AA} \le L_A + R_A$ $P_{CA} \le L_C + R_A$ …

$P_{AC} \le L_A + R_C$ $P_{CC} \le L_C + R_C$ …

$P_{AG} \le L_A + R_G$ $P_{CG} \le L_C + R_G$ …

$P_{AU} \le L_A + R_U$ $P_{CU} \le L_C + R_U$ …

$P_{A-} \le L_A + R_-$ $P_{C-} \le L_C + R_-$ …

NB: HMM not a prob. model

Rigorous Filtering

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

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)

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

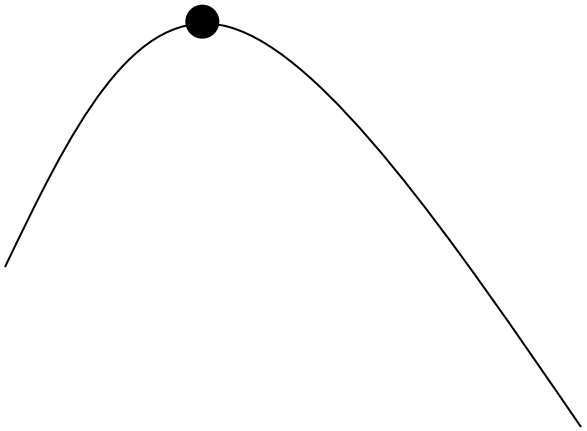
Problem sizes:

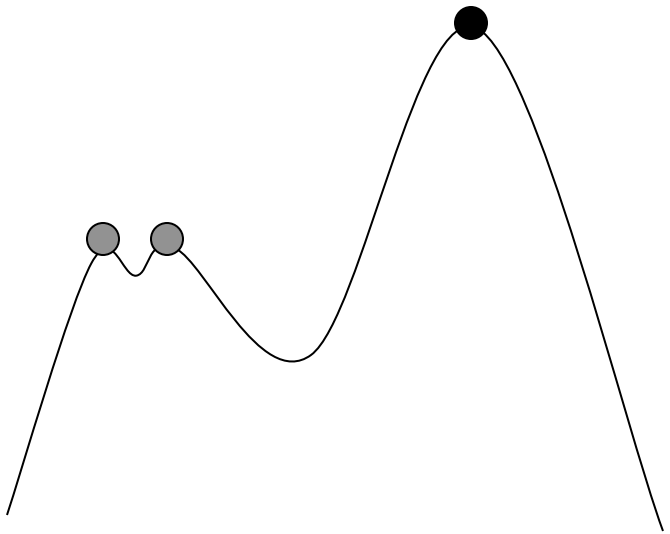
- 1000-10000 variables
- 10000-100000 inequality constraints

"Convex" Optimization

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

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

Motif Discovery

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

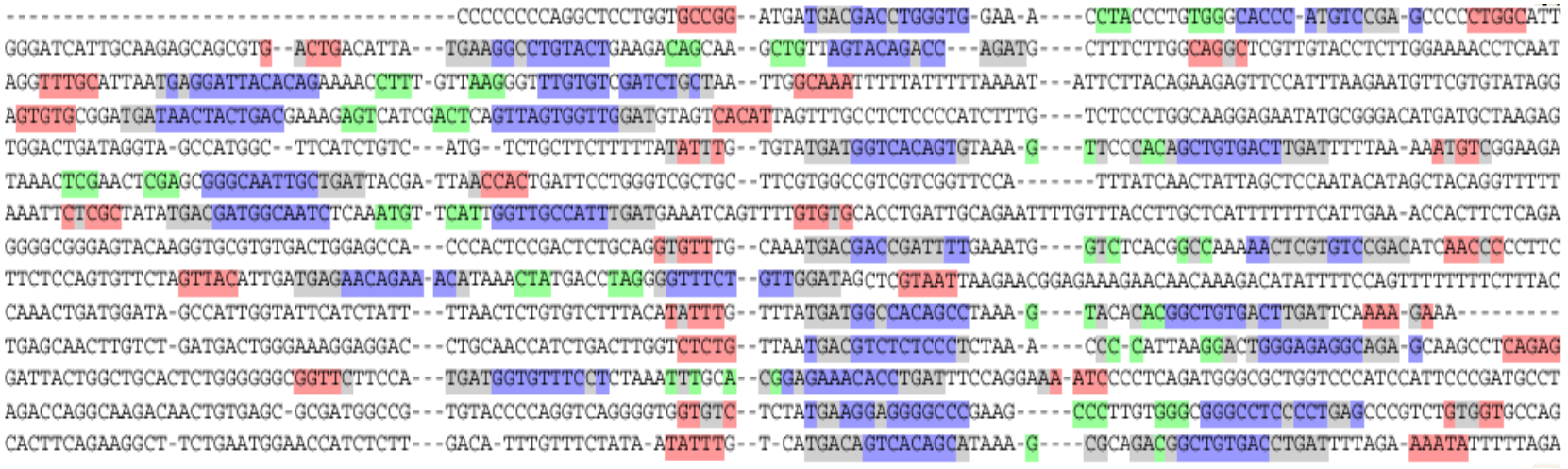
Joint: Do both together

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

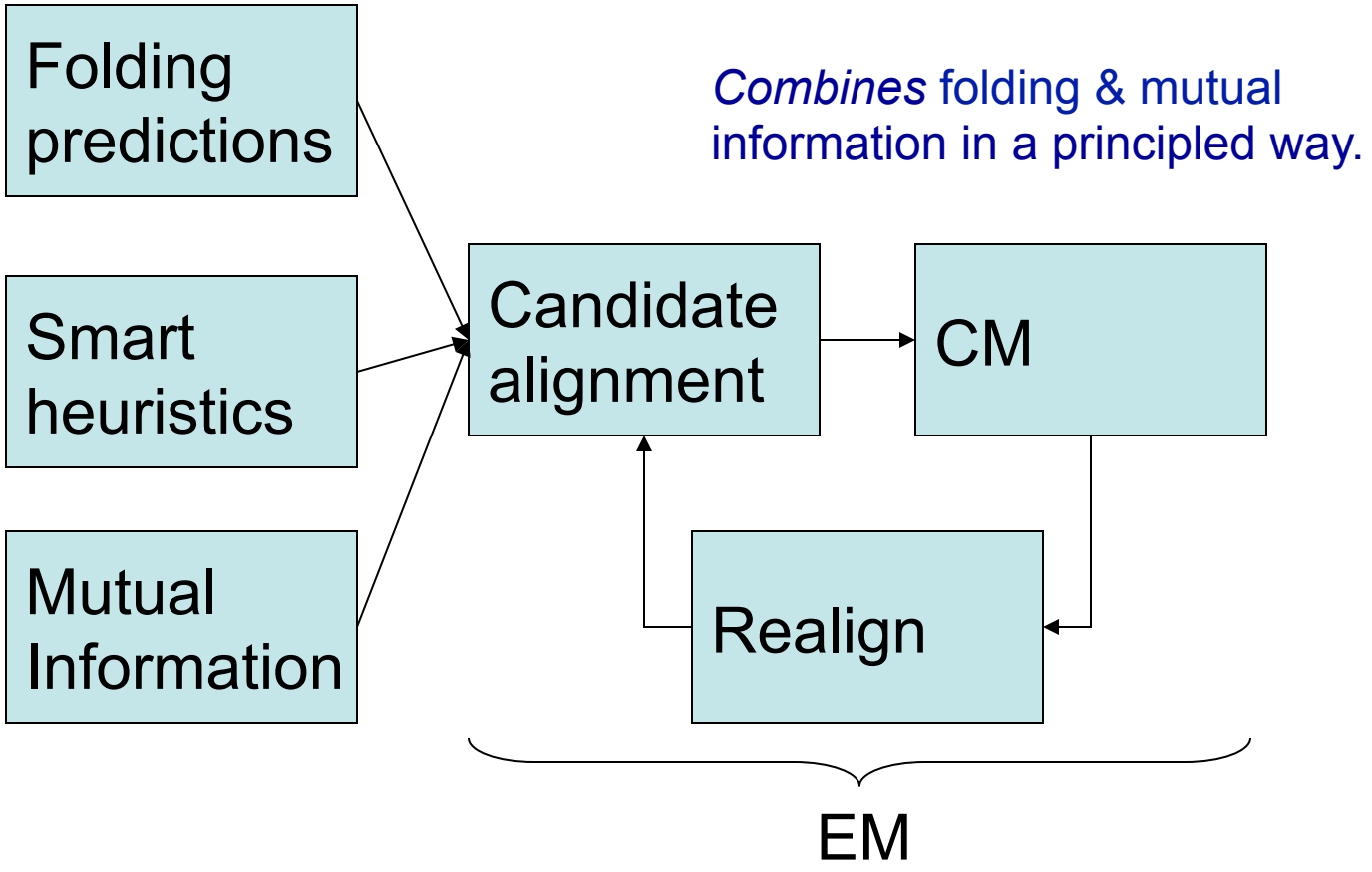
Simultaneous *local* alignment, folding and CM-based motif description using an EM-style learning procedure

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

CMFinder

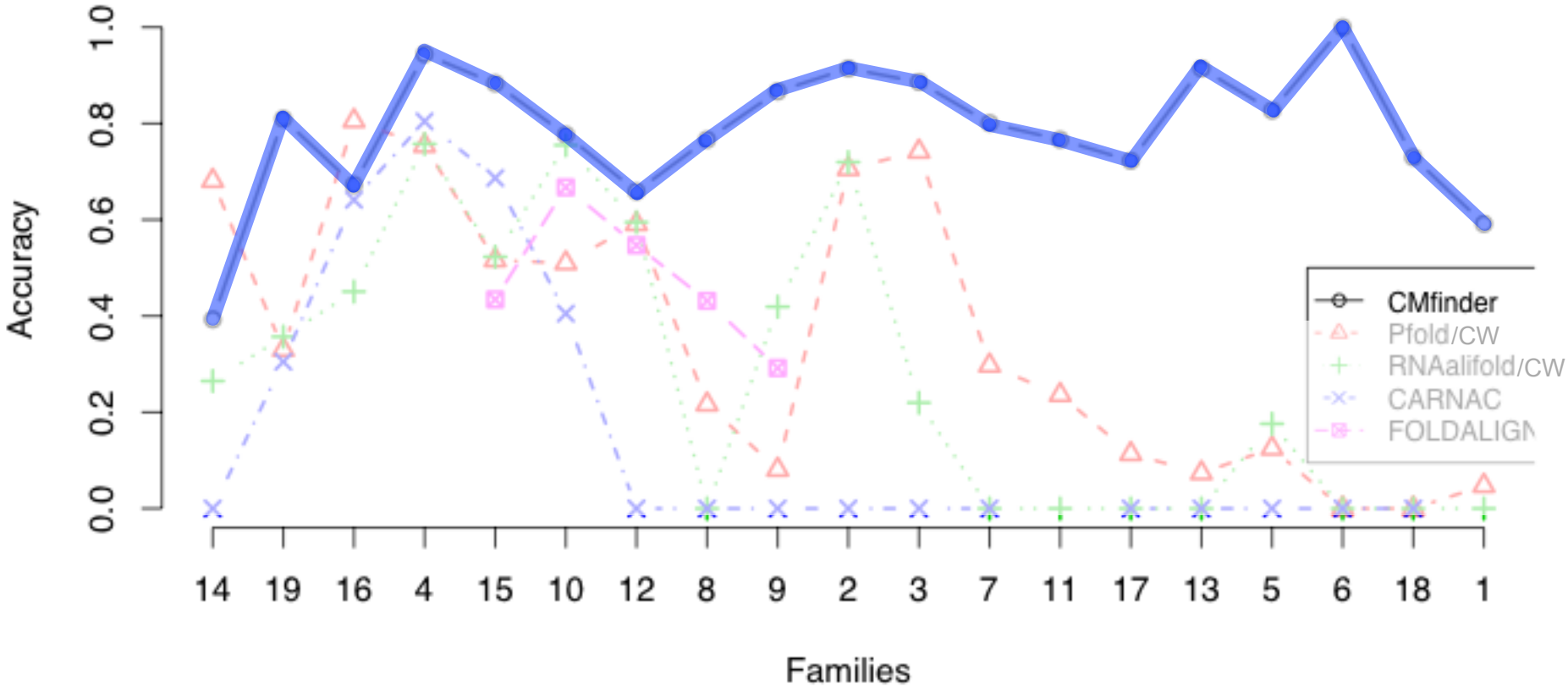
Simultaneous alignment, folding & motif description

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]

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, look near similar genes ("homologs")

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

(Not strategy used in vertebrates - 1000x larger genomes)

Processing Times

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

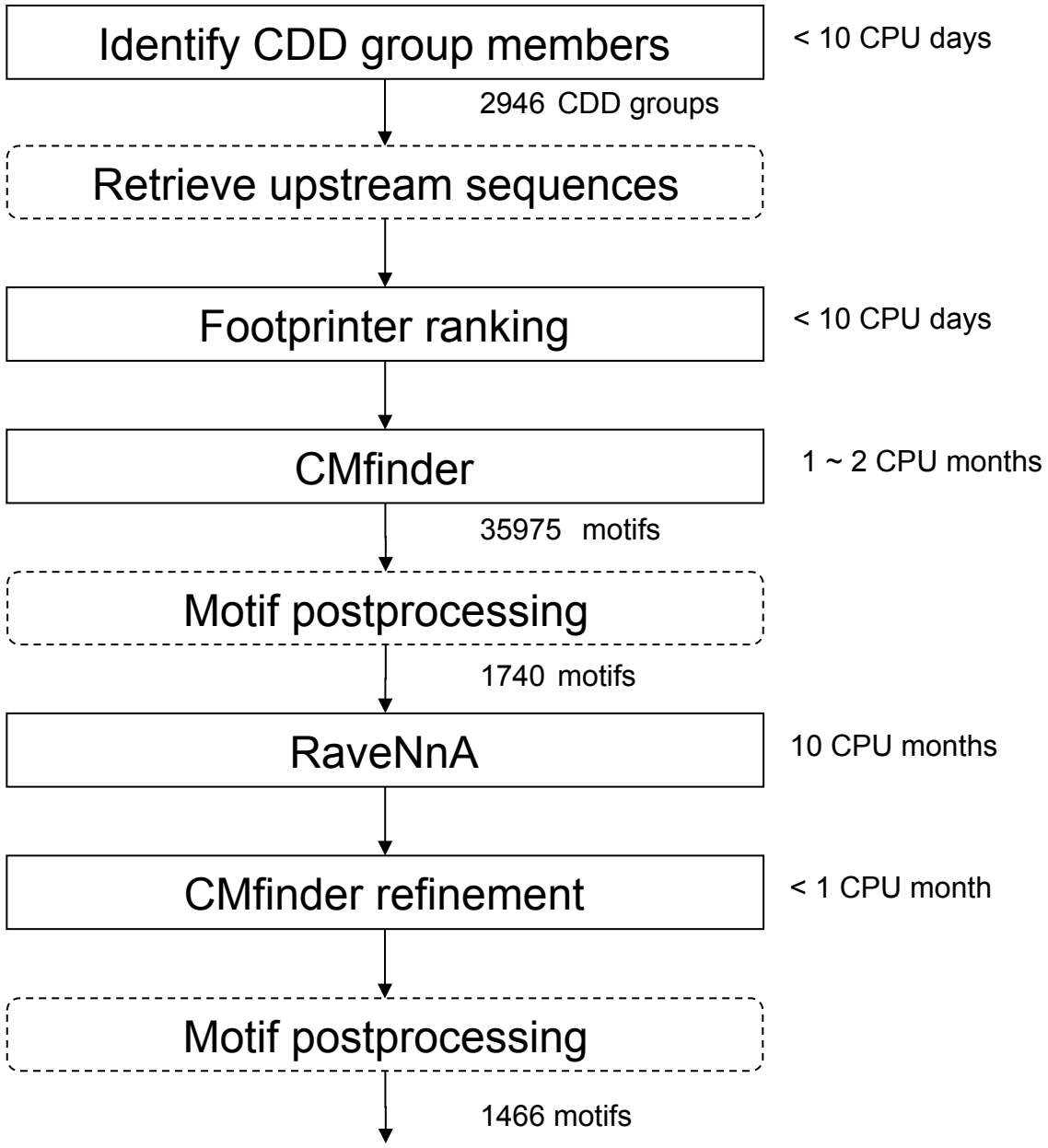


Table 1: Motifs that correspond to Rfam families

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

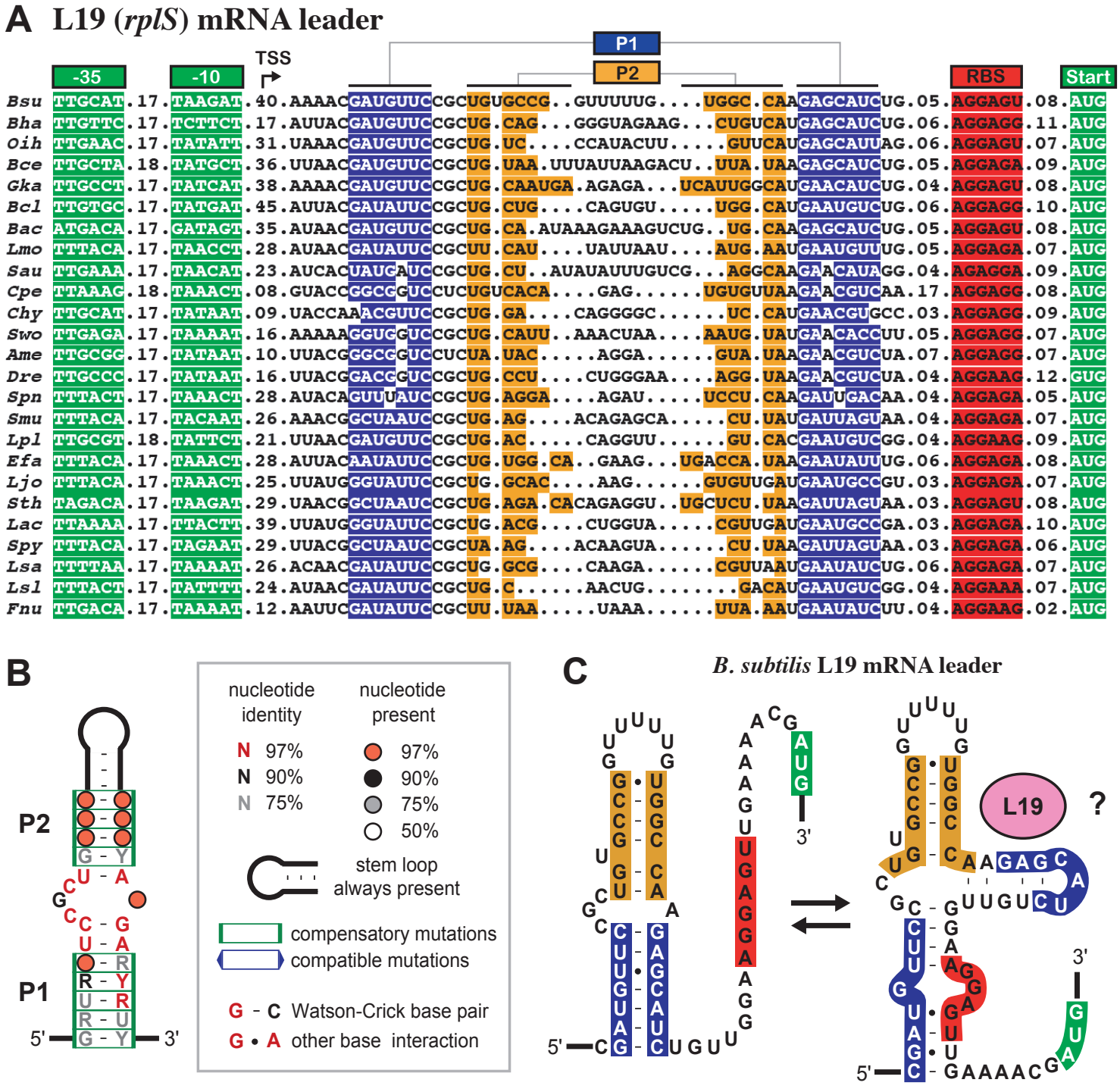
Table 1: Motifs that correspond to Rfam families. "Rank": the three columns show ranks for refined motif clusters after genome scans ("RAV"), CMfinder motifs before genome scans ("CMF"), and FootPrinter results ("FP"). We used the same ranking scheme for RAV and CMF. "Score"

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

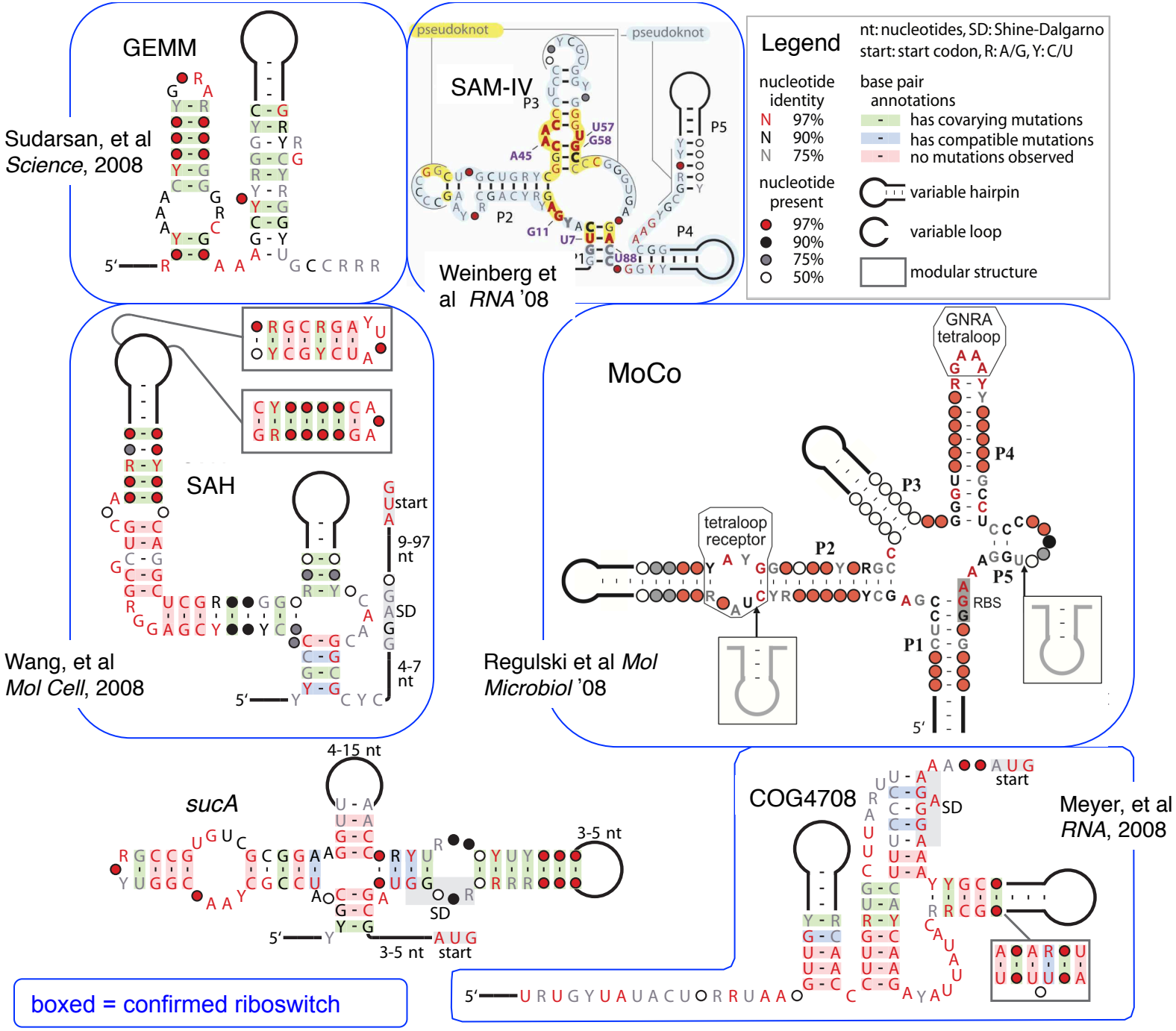
Tbl 2: Prediction accuracy compared to prokaryotic subset of Rfam full alignments.
Membership: # of seqs in overlap between our predictions and Rfam's, the sensitivity (Sn) and specificity (Sp) of our membership predictions. Overlap: the avg len of overlap between our predictions and Rfam's (nt), the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

RNAz

S Washietl, IL Hofacker, M Lukasser, A Hutenhofer, 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
~1000 conserved across *all* vertebrates.
~1/3 in introns of known genes, ~1/6 in UTRs
~1/2 located far from any known gene

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 36970 (of 100,000) pairs

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

6500 candidates in ENCODE alone (better FDR, but still high)

Some details below

Thousands of Predictions

CMfinder Search in Vertebrates

Extract ENCODE* Multiz alignments

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

Apply CMfinder to both strands.

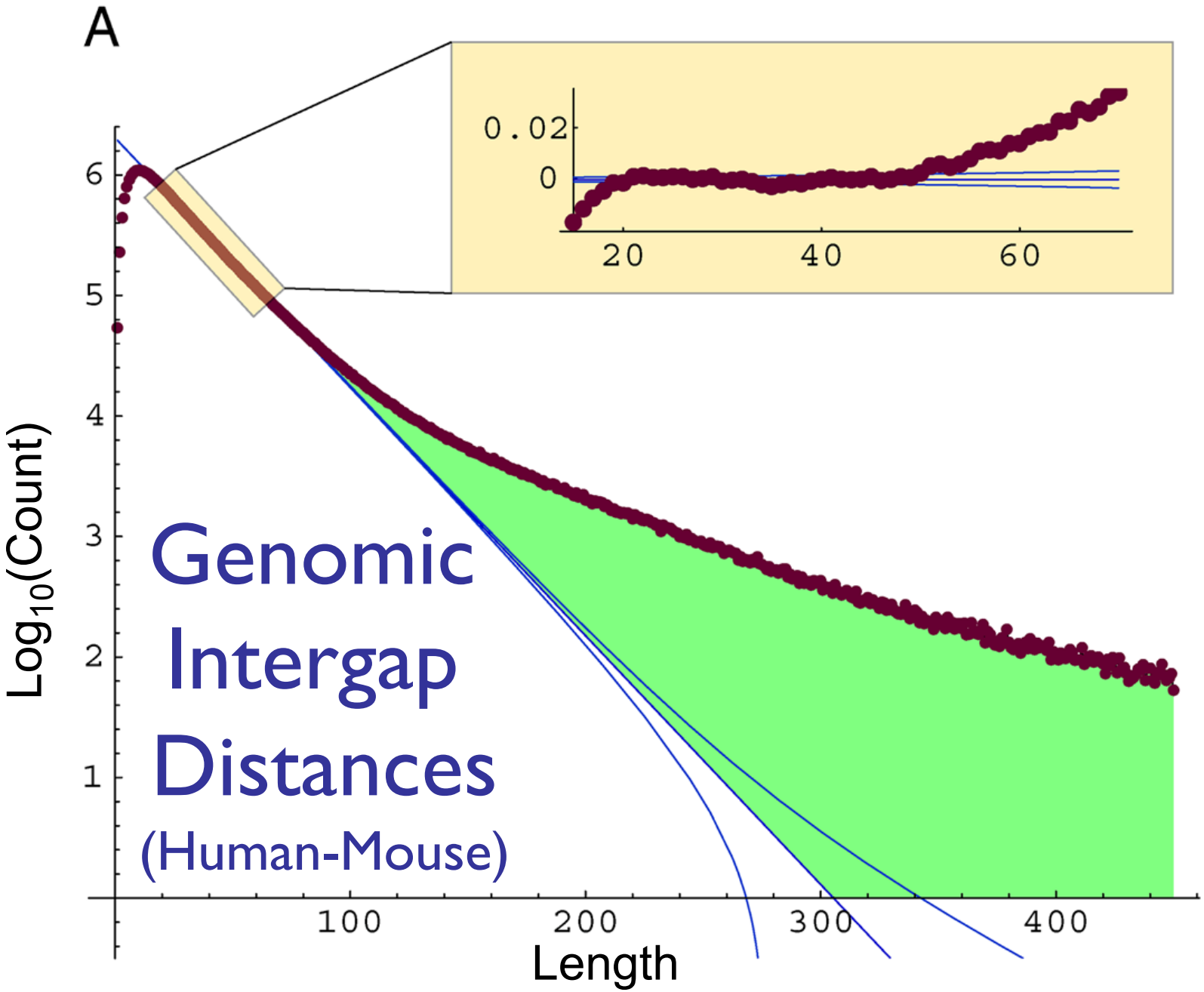
10,106 predictions, 6,587 clusters.

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

* ENCODE: deeply annotated 1% of human genome



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

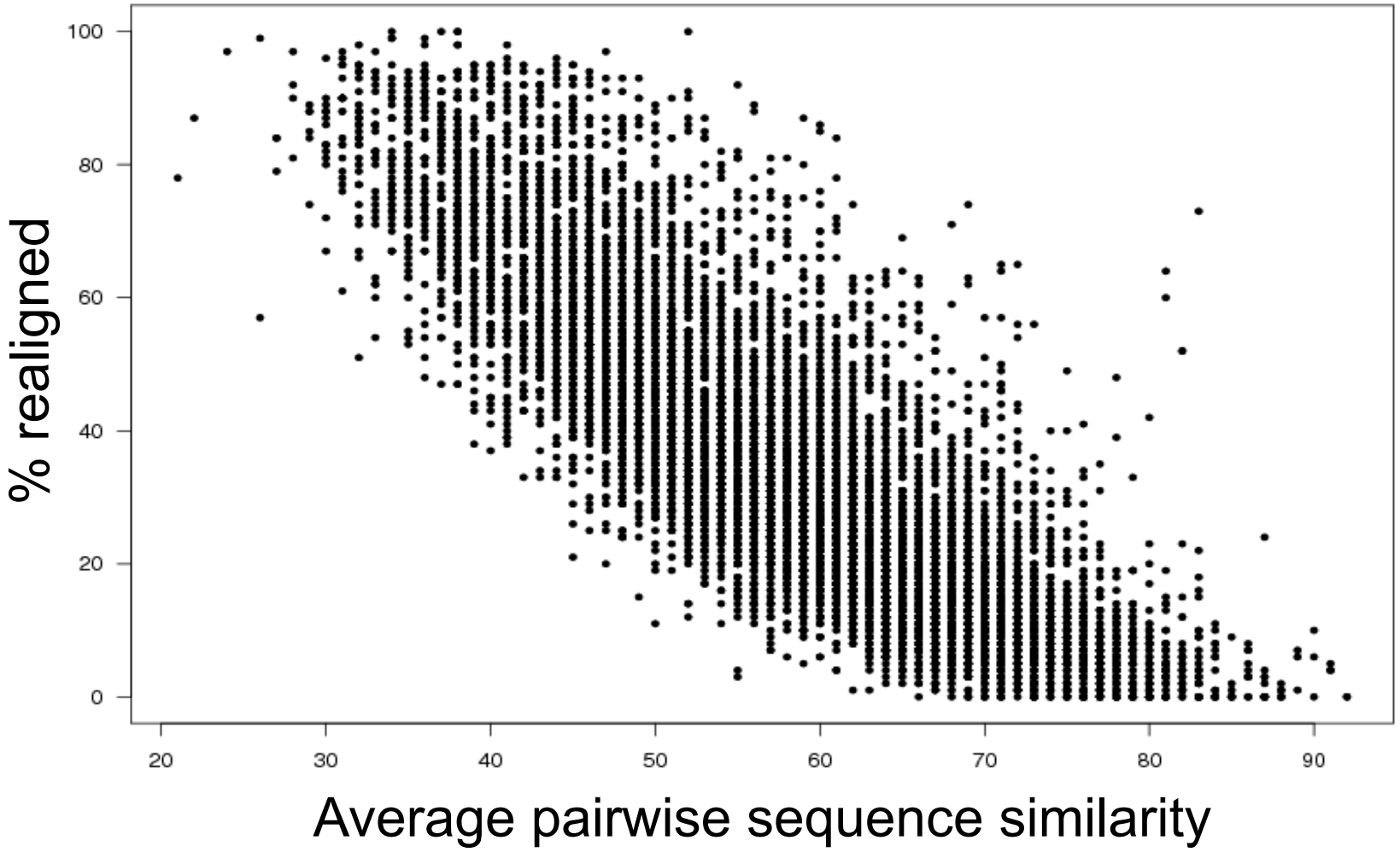
IPS presumed to signal purifying selection

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%

Realignment



Alignment Matters

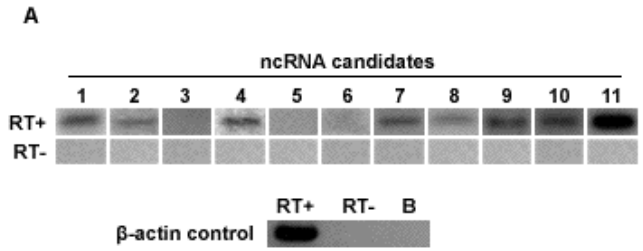
The original MULTIZ alignment without flanking regions. **RNAz Score: 0.132 (no RNA)**

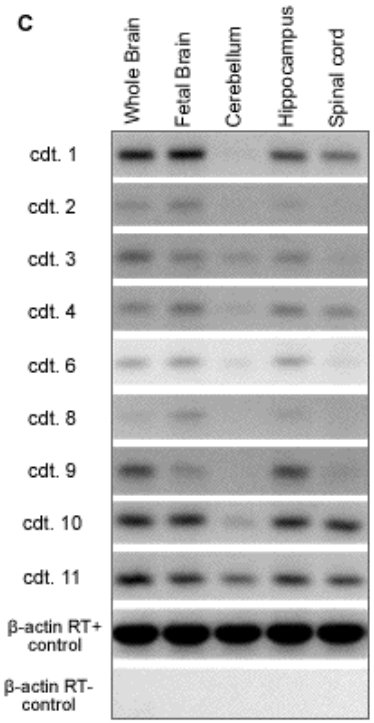
```
Human   GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAACCAAGAGGT----CTTAACAGTATGACCAAAAACTGAAGTT
Chimp   GGACATTTCAATGCGGGCTC-ATGGGGCTGTGAAGCCAAGAGCT----ATTAACACTATGACCAAGGACTGAAATT
Cow     GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAACCGGGAGCT----CTTAATGCTGTGACCAAAGATTGAAGTT
Dog     GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAACCAAGGGCT----CTTAACTCTGTGACCAAATATTAGAGTT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

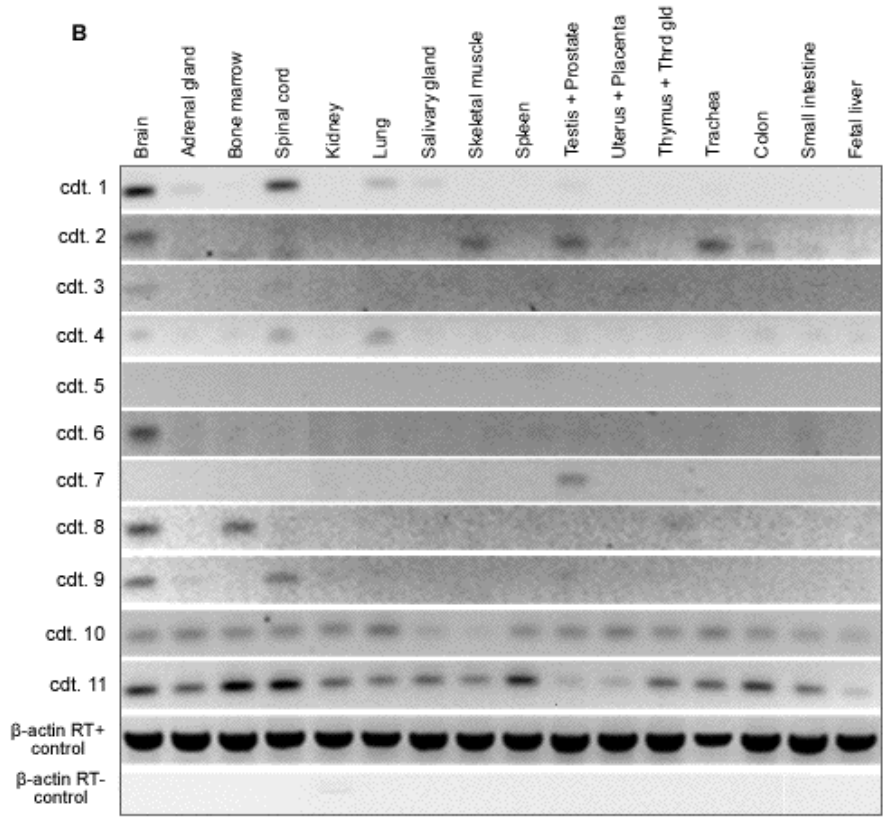
The local CMfinder re-alignment of the MULTIZ block. **RNAz Score: 0.709 (RNA)**

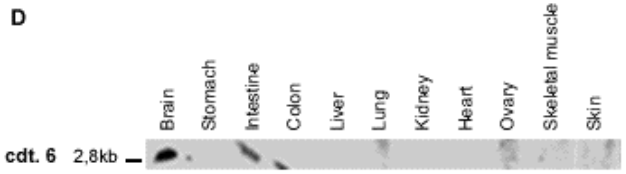
```
Human   GGTCACTTCAAAGAGGGCTT-GTGGGGCTGTGAAA-CCA-----AGAGGTCTTAACAGTATGACCAAAAACTGAAG
Chimp   GGACATTTCAATGCGGGCTC-ATGGGGCTGT-GAAGCCA-----AGAGCTATTAACACTATGACCAAGGACTGAAA
Cow     GGTCATTTCAAAGAGGGCTT-ATGAGACCA--AAA-CCG-----GGAGCTCTTAATGCTGTGACCAAAGATTGAAG
Dog     GGTCATTTCAAAGAGGGCTTTGTGGAACTA--AAA-CCA-----AGGGCTCTTAACTCTGTGACCAAATATTAGAG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

10 of 11 top (differentially) expressed









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

Course Wrap Up

What is DNA? RNA?

How many Amino Acids are there?

Did human beings, as we know them, develop from earlier species of animals?

What are stem cells?

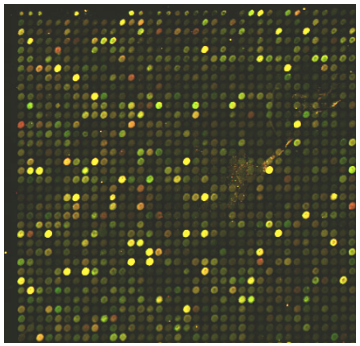
What did Viterbi invent?

What is dynamic programming?

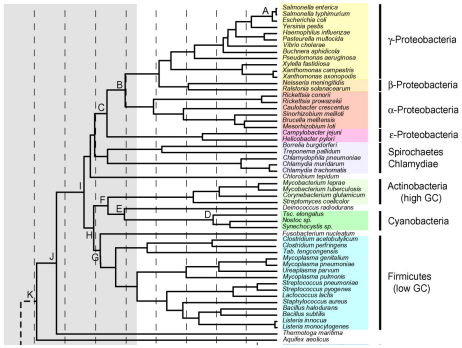
What is a likelihood ratio test?

What is the EM algorithm?

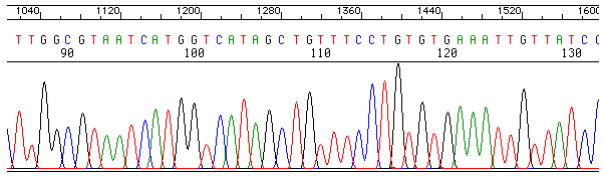
How would you find the maximum of $f(x) = ax^3 + bx^2 + cx + d$ in the interval $-10<x<25$?

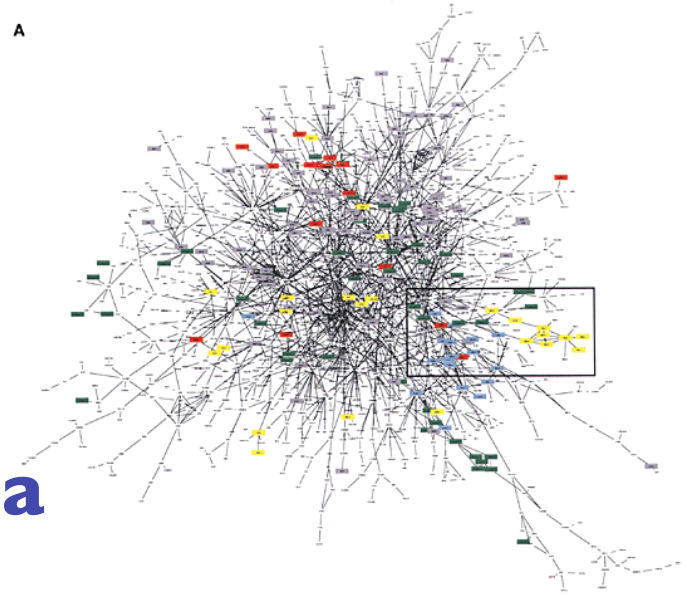


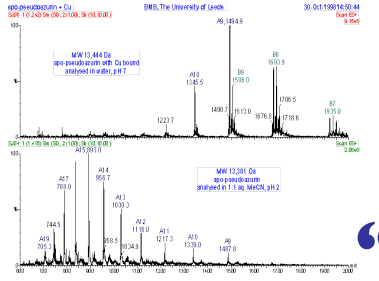
"High-Throughput BioTech"



- Sensors
  - DNA sequencing
  - Microarrays/Gene expression
  - Mass Spectrometry/Proteomics
  - Protein/protein & DNA/protein interaction
- Controls
  - Cloning
  - Gene knock out/knock in
  - RNAi







***Floods* of data**

"Grand Challenge" problems

Exciting Times

Lots to do

Highly multidisciplinary

You'll be hearing a lot more about it

I hope I've given you a taste of it

Thanks!