Modeling and Searching for Non-Coding RNA

W.L. Ruzzo

http://www.cs.washington.edu/homes/ruzzo

http://www.cs.washington.edu/homes/ruzzo/courses/gs541/09sp

GENOME 541

"… *protein* and *DNA* sequence analysis … to determine the "periodic table of biology," i.e., the list of *proteins* …, which can be regarded as the first stage in…"

No mention of RNA…

The Message

Cells make lots of ~~RNA~~ *noncoding* RNA

Functionally important, functionally diverse

Structurally complex

New tools required

alignment, discovery, search, scoring, etc.

The Outline

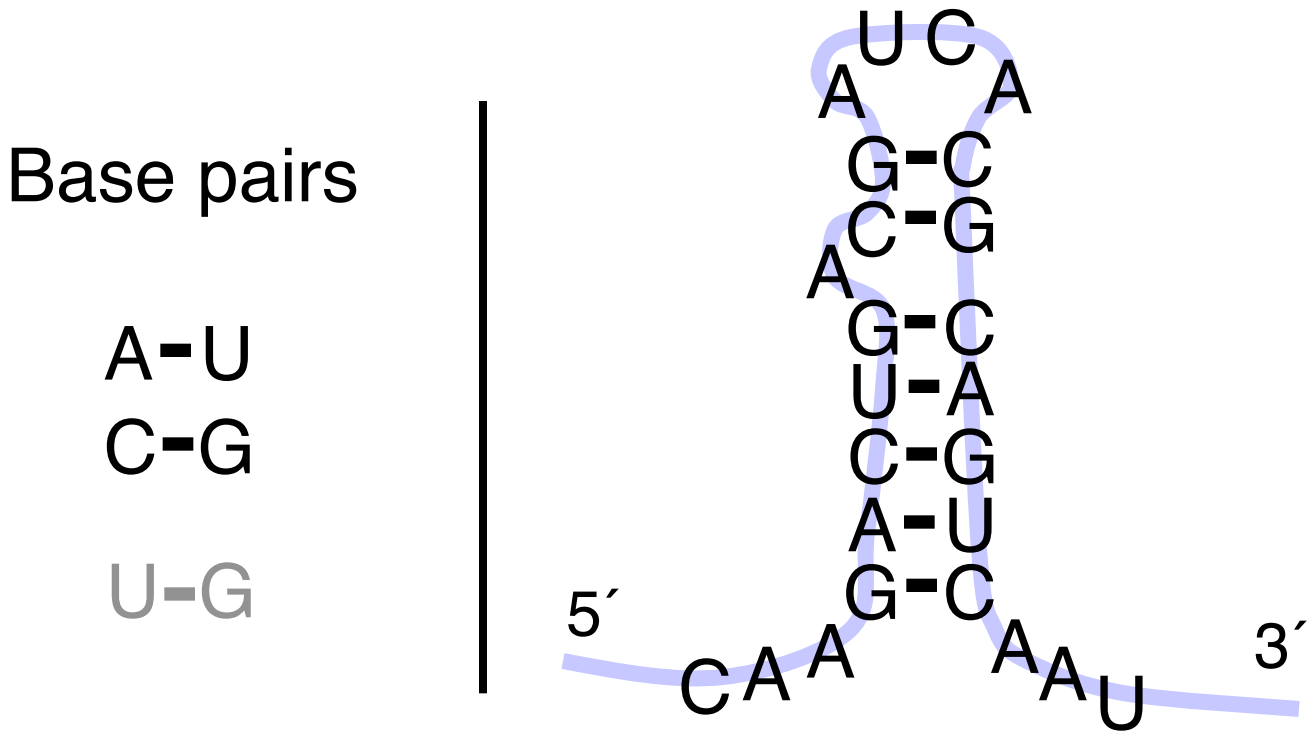
The problem: noncoding RNA

Why: it's important

Some results

Some methods

RNA Secondary Structure:
RNA makes helices too



Usually *single* stranded

RNA: Interest

NATURE VOL. 227 AUGUST 8 1970

Central Dogma of Molecular Biology

by
FRANCIS CRICK
MRC Laboratory
Hills Road,
Cambridge CB2 2QH

The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable over-simplification."

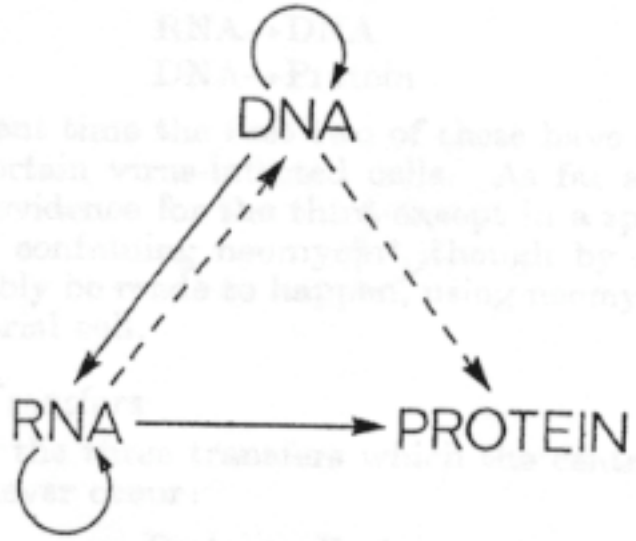


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

“Classical” RNAs

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

a handful of others

Bacteria

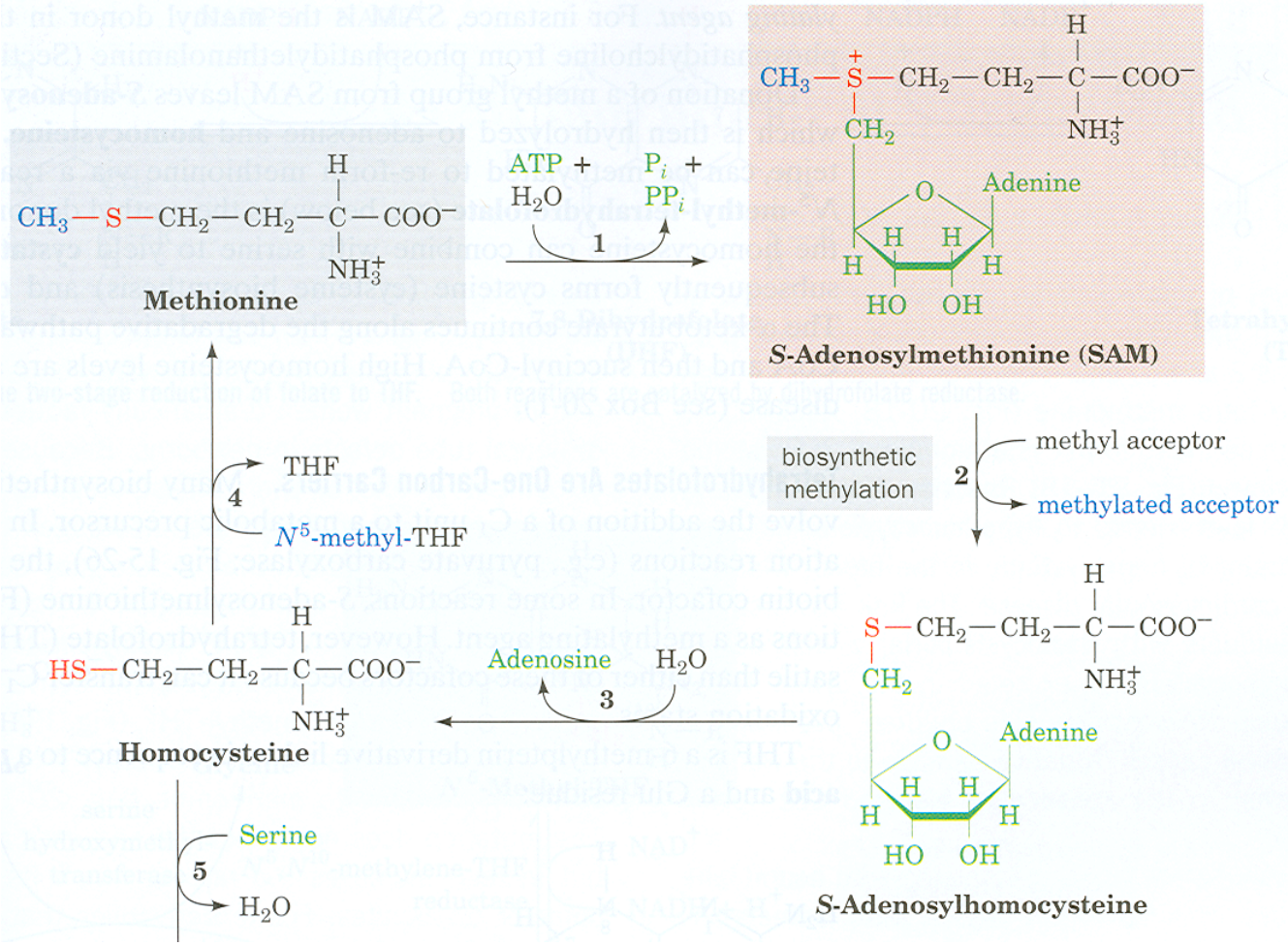
Triumph of proteins

80% of genome is coding DNA

Functionally diverse

- receptors
- motors
- catalysts
- regulators (Monod & Jakob, Nobel prize 1965)
- …

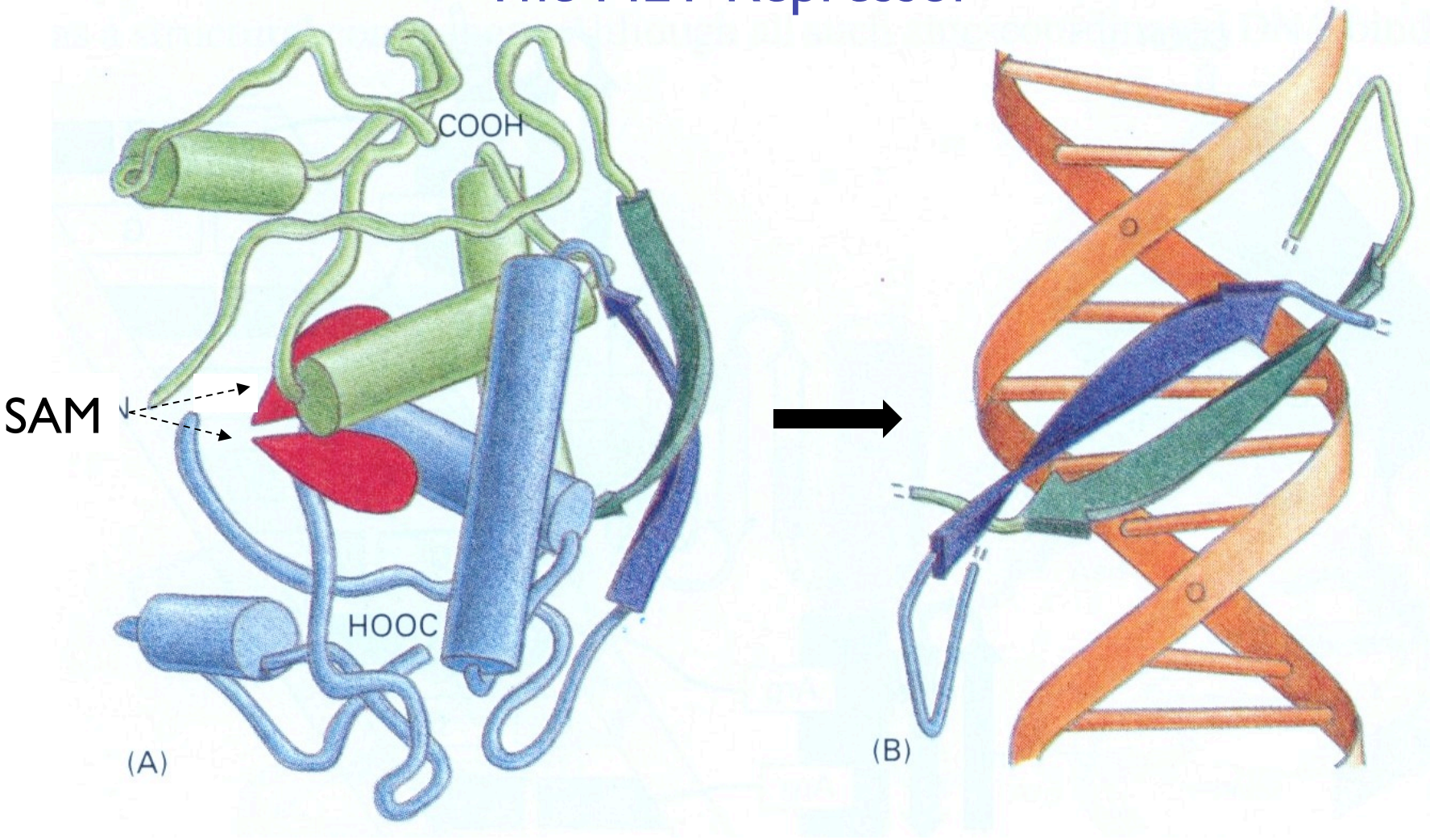
Proteins Catalyze Biochemistry: Met Pathways



• • •

Proteins Regulate Biochemistry:

The MET Repressor

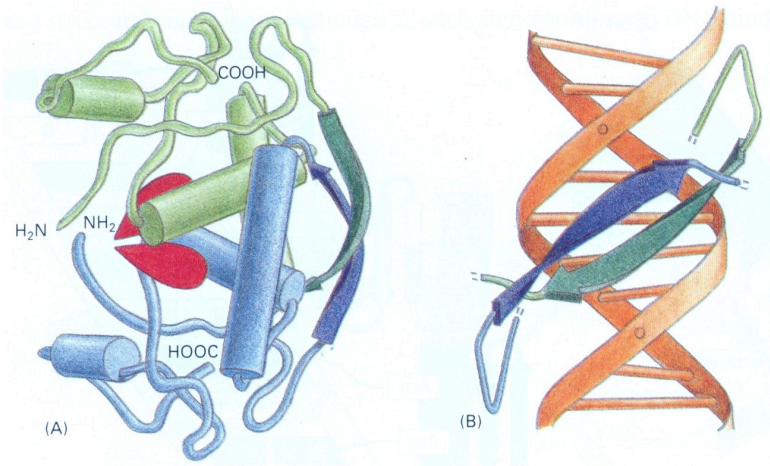


Protein

Alberts, et al, 3e.

DNA

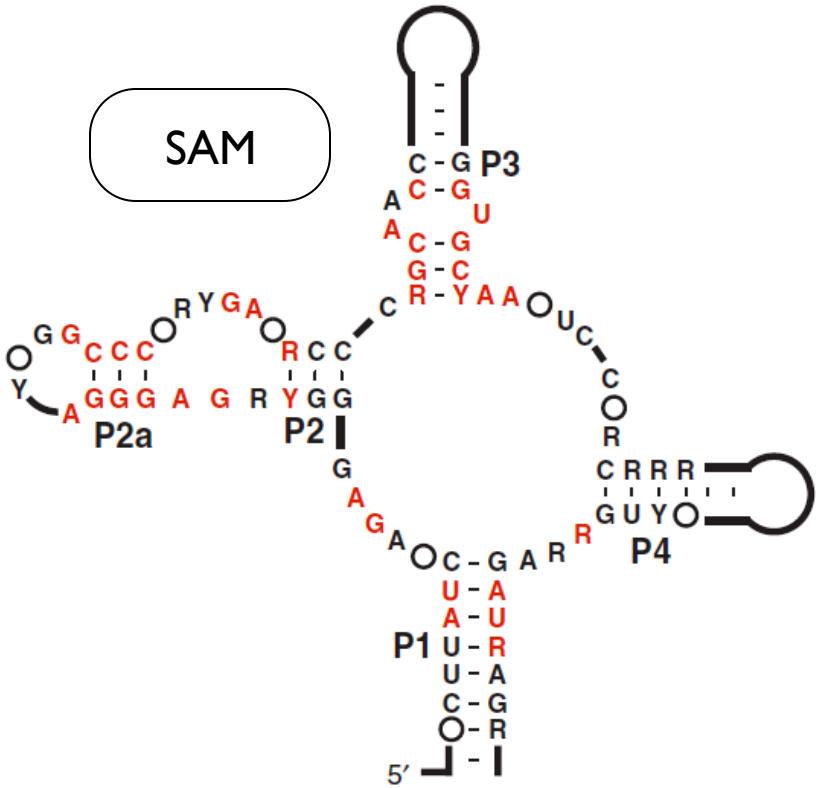
Alberts, et al, 3e.



Not the only way!

Protein way

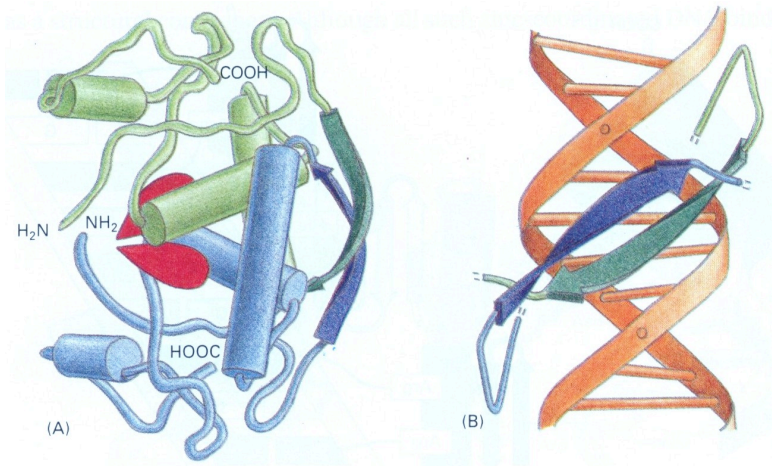
Riboswitch alternative



Grundy & Henkin, Mol. Microbiol 1998
Epshtein, et al., PNAS 2003
Winkler et al., Nat. Struct. Biol. 2003

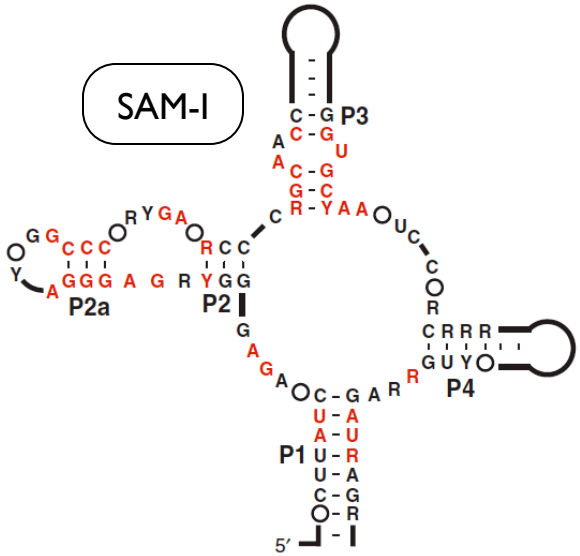
Not the only way!

Alberts, et al, 3e.

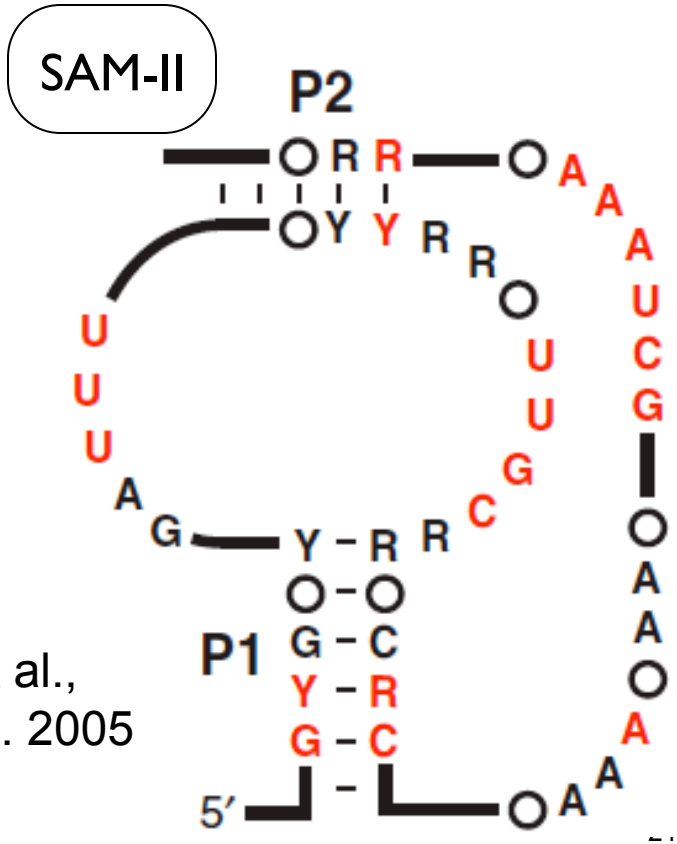


← Protein way

Riboswitch alternatives →



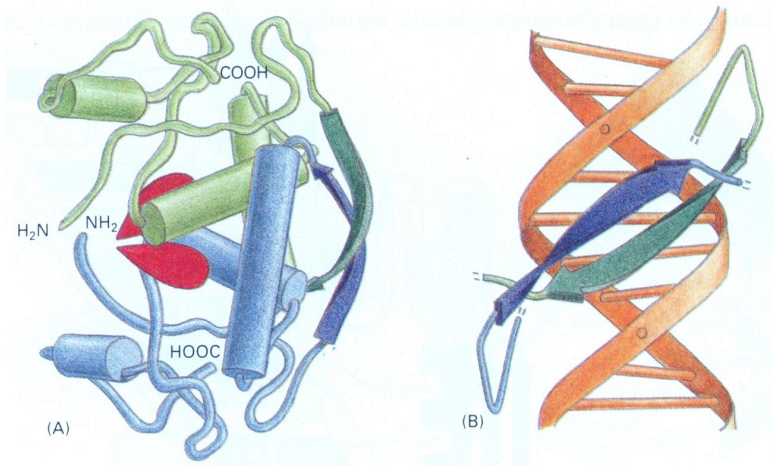
Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

Not the only way!

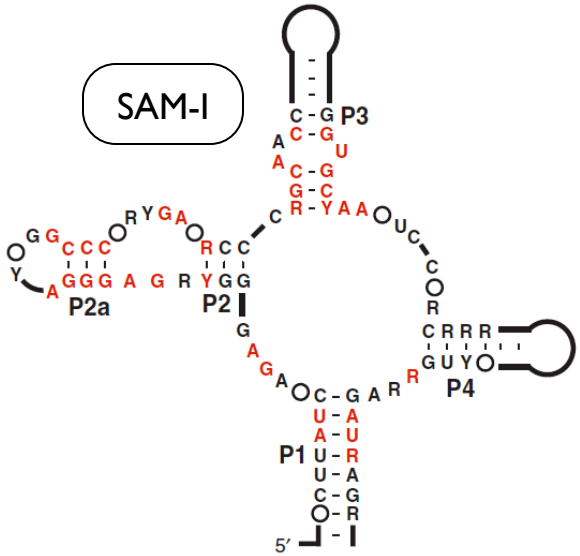
Alberts, et al, 3e.



Protein way

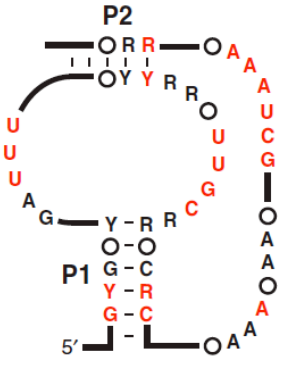
Riboswitch alternatives

SAM-I



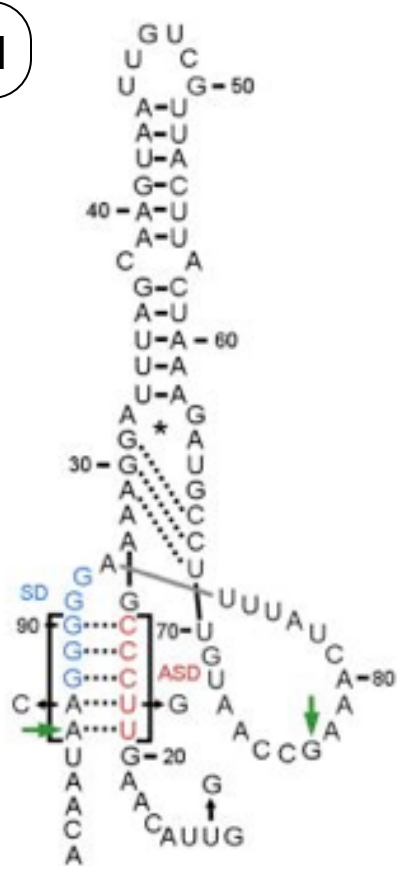
Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



Corbino et al., Genome Biol. 2005

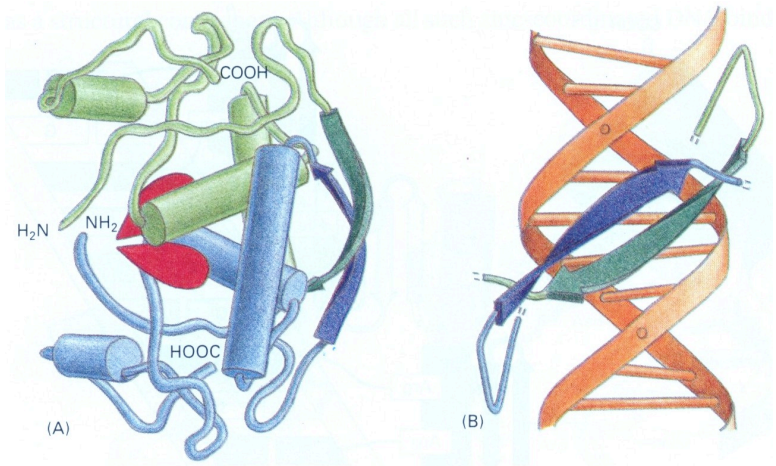
SAM-III



Fuchs et al., NSMB 2006

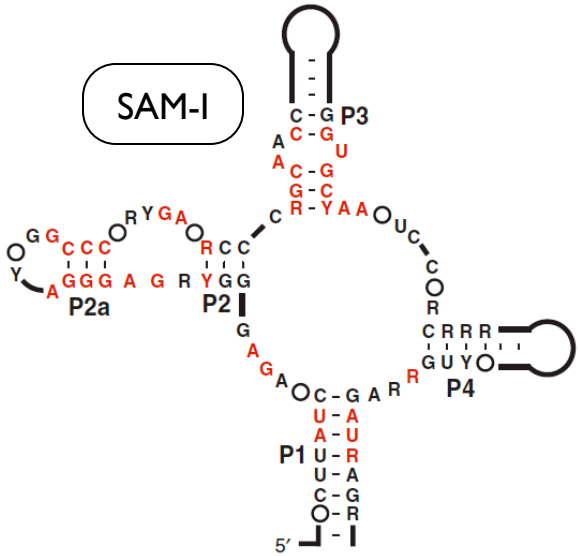
Not the only way!

Alberts, et al, 3e.

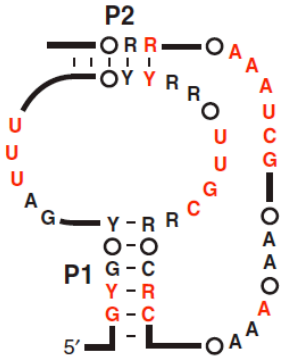


← Protein way

Riboswitch alternatives ↓

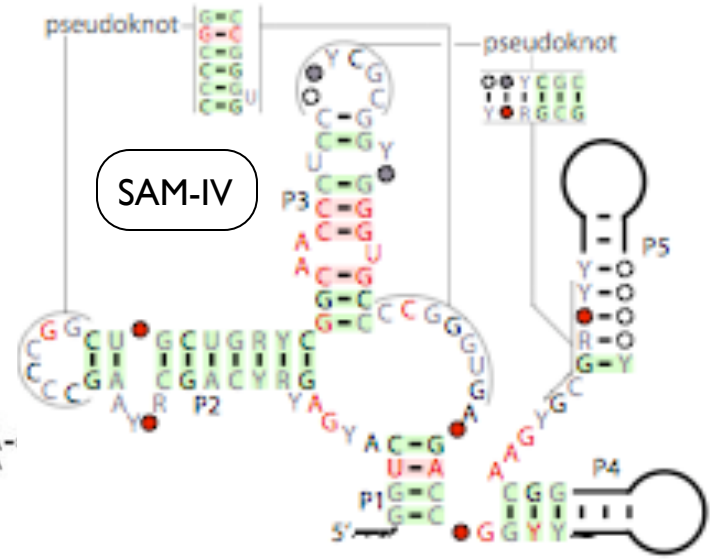


Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

Fuchs et al., NSMB 2006



Weinberg et al., RNA 2008

Tip of the iceberg?

Approximately 20 riboswitches known

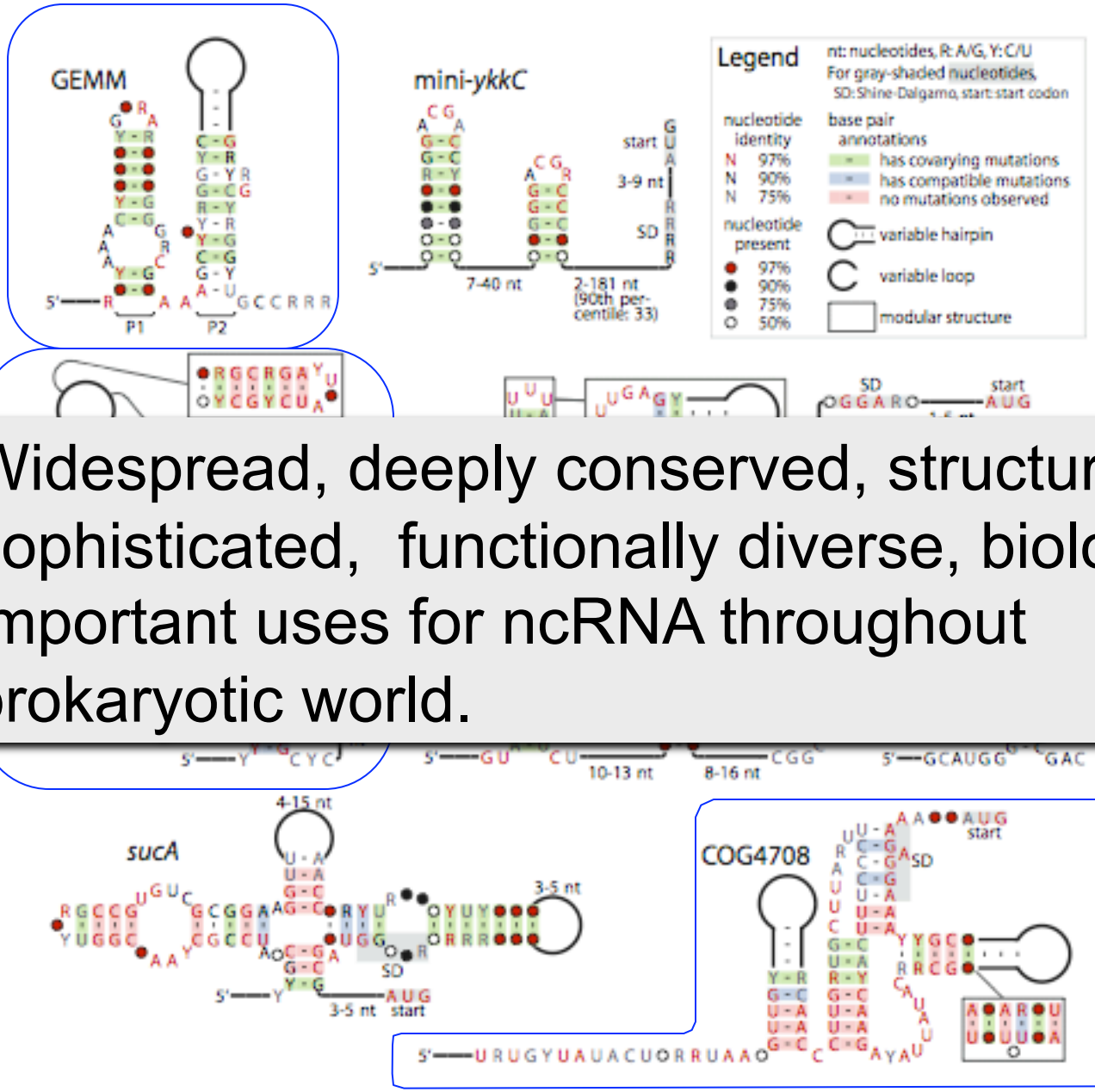
Regulate ~20,000 operons in sequenced organisms

All found since 2003

More ribo-regulators are known than protein transcription factors in many species

Growing number of *trans*-acting elements, too

E.g., a recent RNA-seq study found >500 small RNA and 127 antisense RNA in *V. cholerae*



Widespread, deeply conserved, structurally sophisticated, functionally diverse, biologically important uses for ncRNA throughout prokaryotic world.

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

Vertebrates

Bigger, more complex genomes

<2% coding

But >5% conserved in sequence?

And 50-90% transcribed?

And *structural* conservation, if any, invisible
(without proper alignments, etc.)

What's going on?

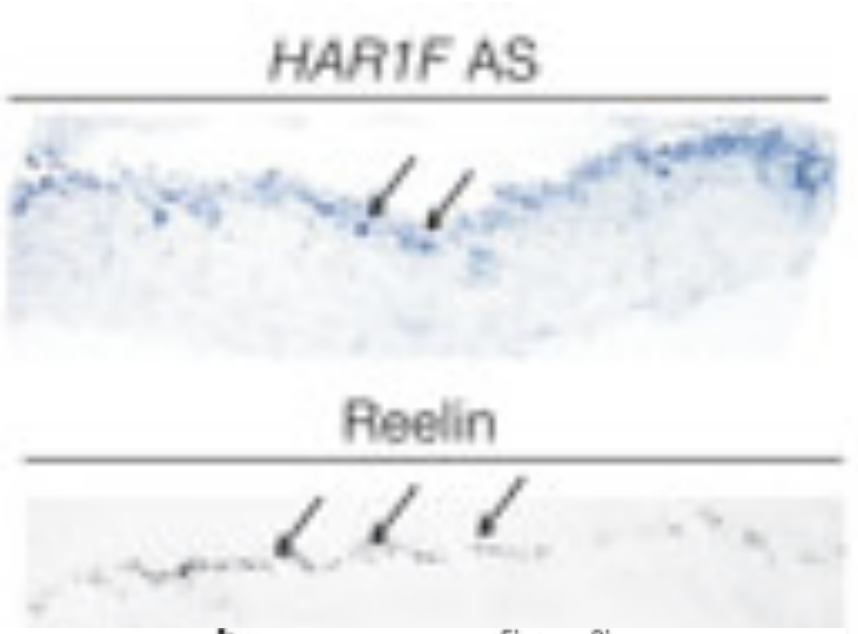
Vertebrate ncRNAs

mRNA, tRNA, rRNA, … of course

PLUS:

snRNA, spliceosome, snoRNA, teleomerase, microRNA, RNAi, SECIS, IRE, piwi-RNA, XIST (X-inactivation), ribozymes, …

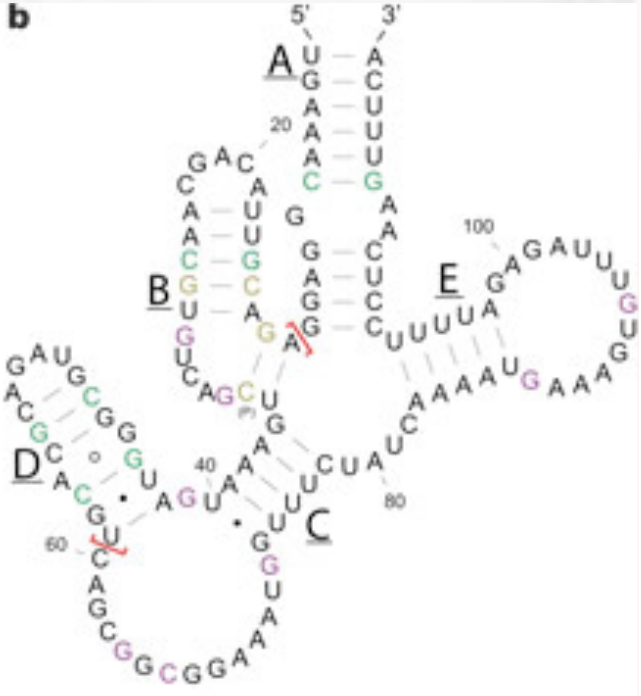
Fastest Human Gene?



a

Position	20	30	40	50
Human	AGACGTTACAGCAACGTGTCAGCTGAAATGATGGGCGTAGACGCACGT			
Chimpanzee	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Gorilla	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Orang-utan	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Macaque	AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT			
Mouse	AGAAATTACAGCAATTTATCAGCTGAAATTATAGGTGTAGACACATGT			
Dog	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Cow	AGAAATTACAGCAATTCATCAGCTGAAATTATAGGTGTAGACACATGT			
Platypus	ATAAATTACAGCAATTTATCAAATGAAATTATAGGTGTAGACACATGT			
Opossum	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Chicken	AGAAATTACAGCAATTTATCAACTGAAATTATAGGTGTAGACACATGT			
Fold	((((((((.....)))))).....))[[[[[.((((.(((....))))).))))			
Pair symbol	lmnopqr rqpon ml rstuvwx xwvutsr			

b



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

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

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.

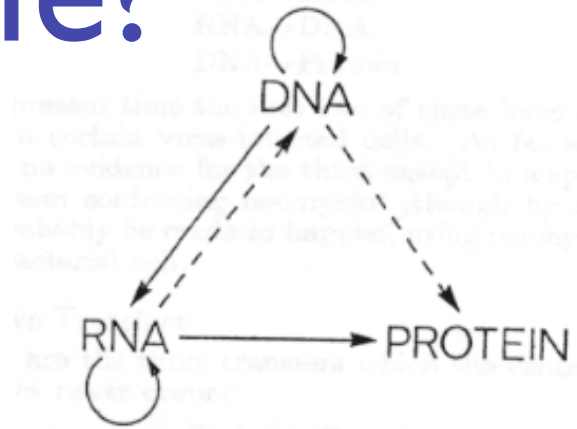
1800 candidates from 36970 (of 100,000) pairs

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)

Origin of Life?



Life needs

information carrier: DNA

molecular machines, like enzymes: Protein

making proteins needs DNA + RNA + proteins

making (duplicating) DNA needs proteins

Horrible circularities! How could it have arisen in an abiotic environment?

Origin of Life?

RNA can carry information, too

RNA double helix; RNA-directed RNA polymerase

RNA can form complex structures

RNA enzymes exist (ribozymes)

RNA can control, do logic (riboswitches)

The "RNA world" hypothesis:
1st life was RNA-based

Some extant RNAs are relicts of that origin, some are "modern" inventions

"Classical" RNAs

tRNA - transfer RNA (~61 kinds, ~ 75 nt)

rRNA - ribosomal RNA (~4 kinds, 120-5k nt)

snRNA - small nuclear RNA (splicing: U1, etc, 60-300nt)

RNaseP - tRNA processing (~300 nt)

RNase MRP - rRNA processing; mito. rep. (~225 nt)

SRP - signal recognition particle; membrane targeting (~100-300 nt)

SECIS - selenocysteine insertion element (~65nt)

6S - ? (~175 nt)

Semi-classical RNAs
(discovery in mid 90's)

tmRNA - resetting stalled ribosomes

Telomerase - (200-400nt)

snoRNA - small nucleolar RNA (many varieties; 80-200nt)

Recent discoveries

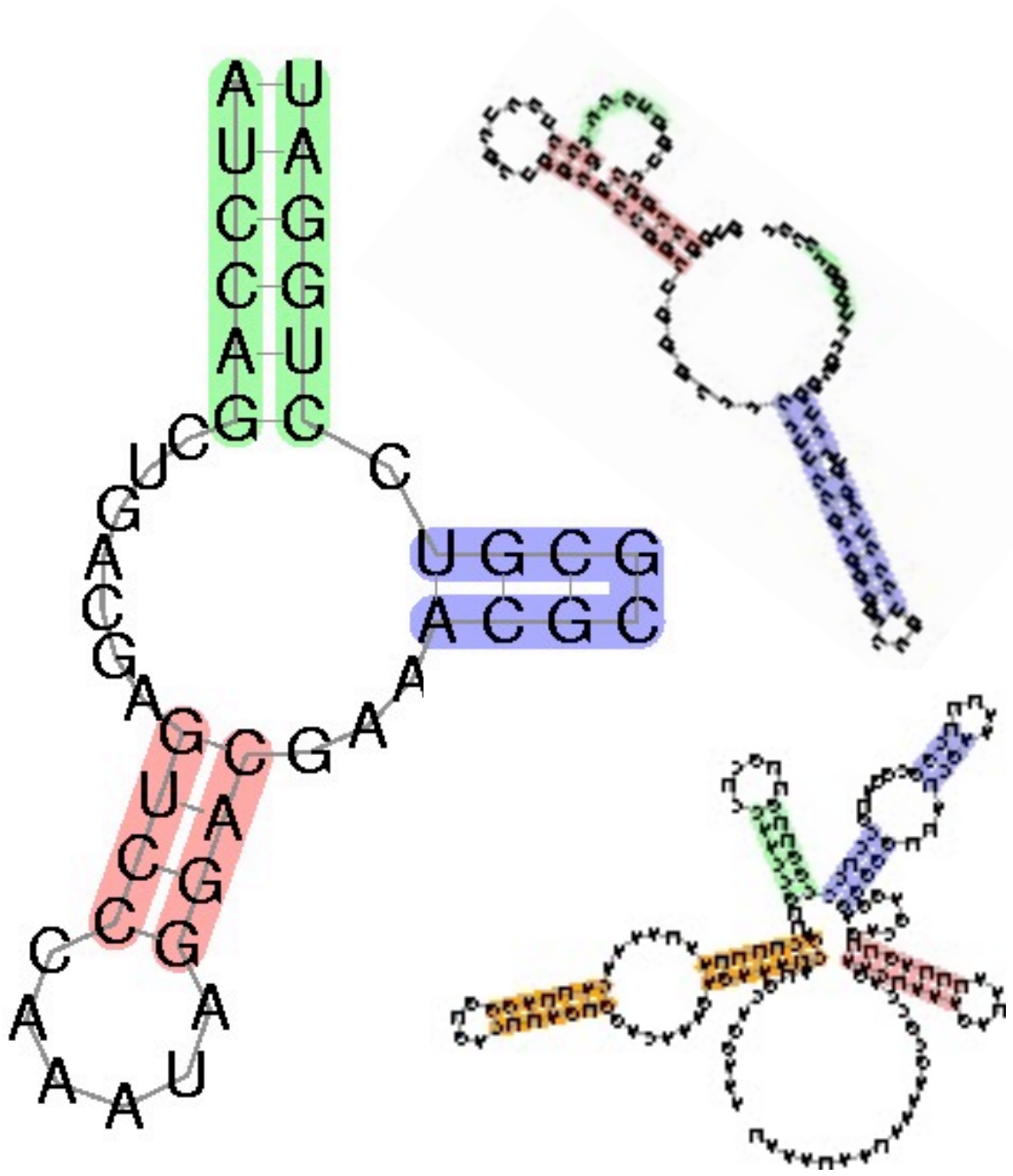
siRNA (Nobel prize 2006: Fire & Mello)
microRNAs (Lasker prize 2008:
Ambros, Baulcombe & Ruvkun)
riboswitches
many ribozymes
regulatory elements
…
Hundreds of families

Rfam release 1, 1/2003: 25 families, 55k instances
Rfam release 9, 7/2008: 603 families, 896k instances
Rfam release 9.1, 1/2009: 1372 families, ??? instances

Why?

RNA's fold, and function

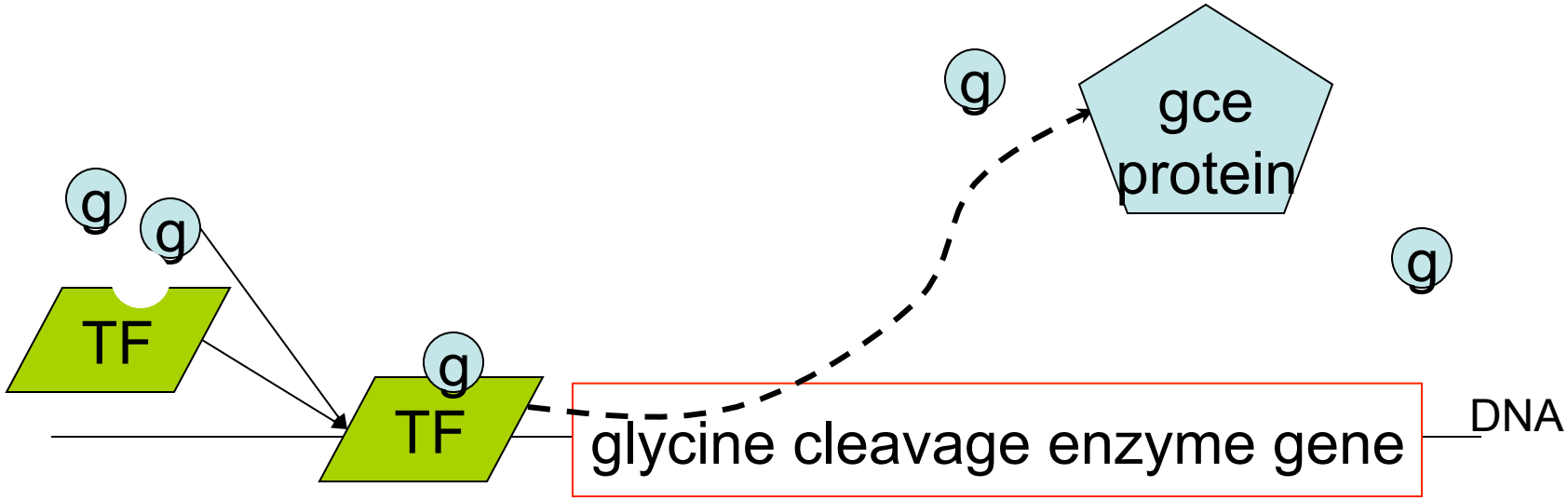
Nature uses what works



Example: Glycine Regulation

How is glycine level regulated?

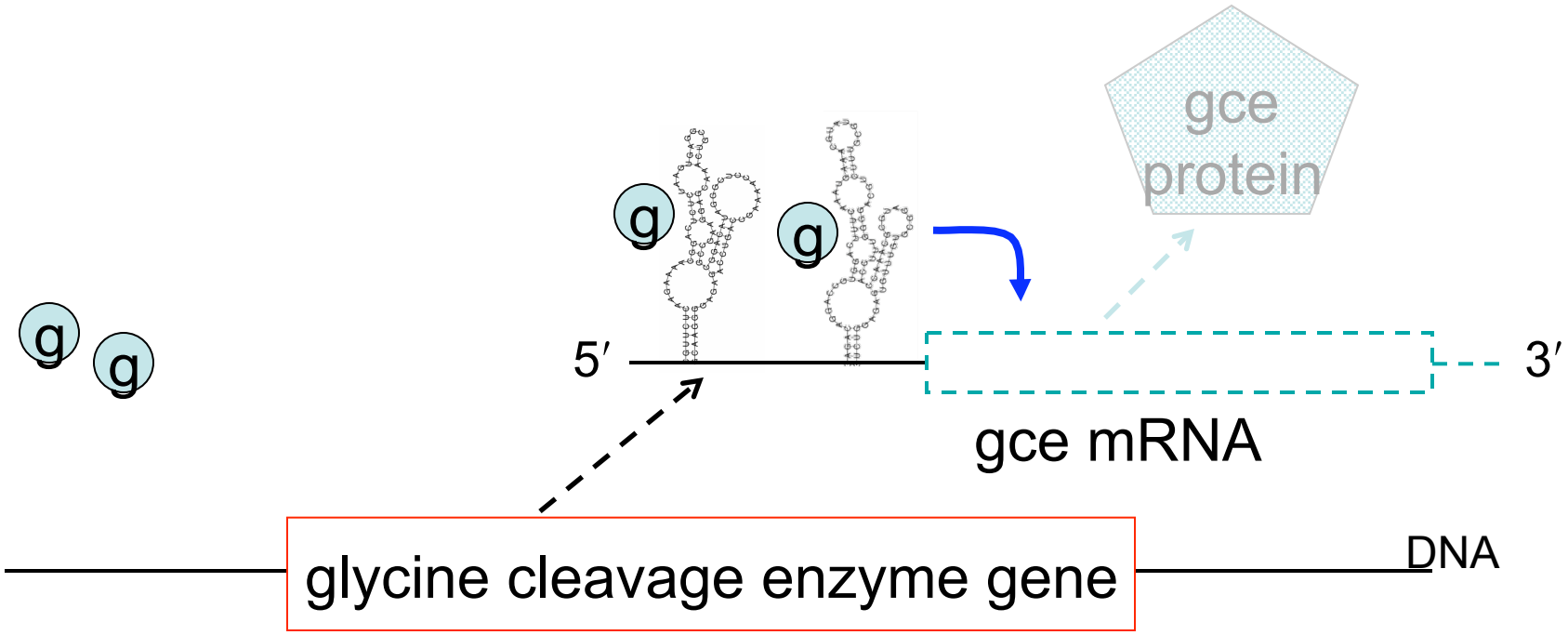
Plausible answer:



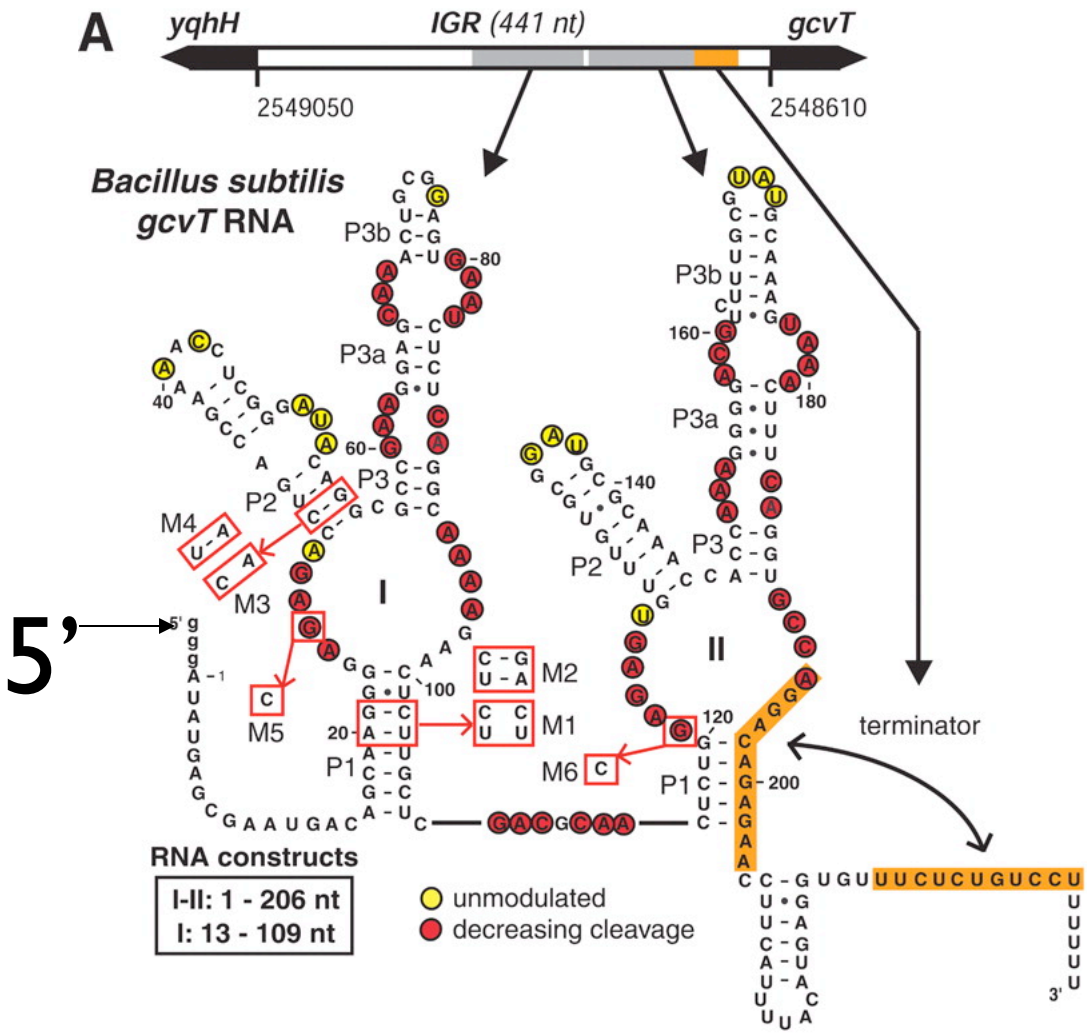
transcription factors (proteins) bind to DNA to turn nearby genes on or off

The Glycine Riboswitch

Actual answer (in many bacteria):



Mandal et al. Science 2004



(Mandal, Lee, Barrick, Weinberg, Emilsson, Ruzzo, Breaker, *Science* 2004)

5’ → gcvT ORF → 3’

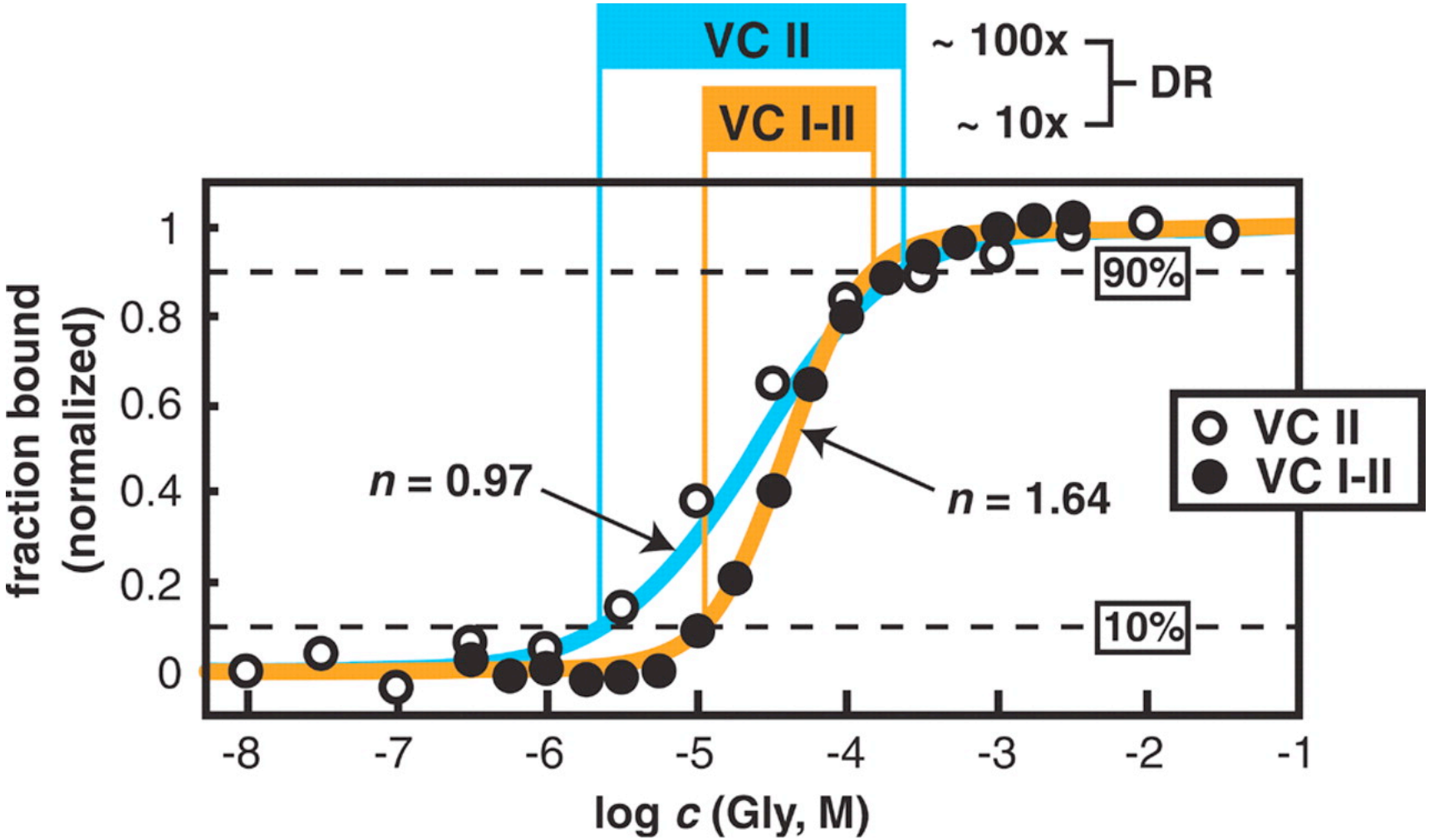


Fig. 3. Cooperative binding of two glycine molecules by the VC I-II RNA. Plot depicts the fraction of VC II (open) and VC I-II (solid) bound to ligand versus the concentration of glycine. The constant, n, is the Hill coefficient for the lines as indicated that best fit the aggregate data from four different regions (fig. S3). Shaded boxes demark the dynamic range (DR) of glycine concentrations needed by the RNAs to progress from 10%- to 90%-bound states.

Riboswitches

~ 20 ligands known; multiple nonhomologous solutions for some

dozens to hundreds of instances of each

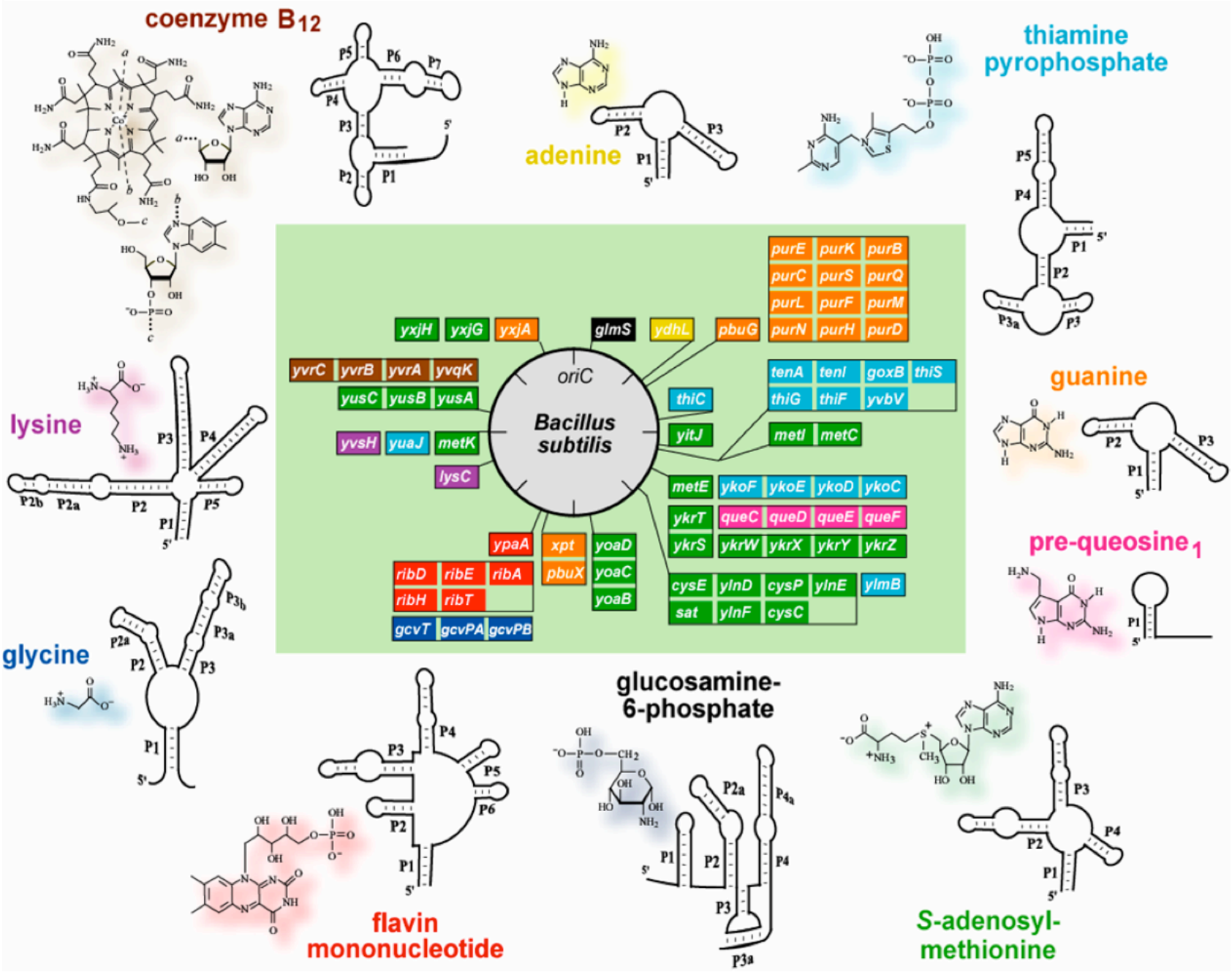
TPP known in archaea & eukaryotes

one known in bacteriophage

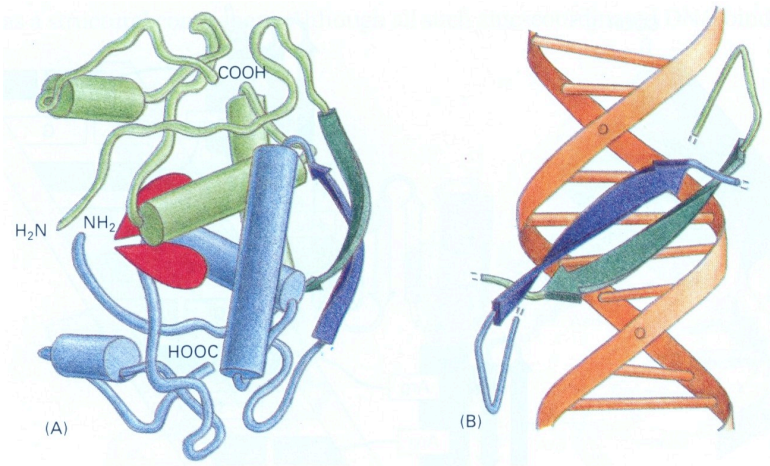
on/off; transcription/translation; splicing; combinatorial control

In some bacteria, more riboregulators identified than protein TFs

all found since ~2003


coenzyme B12
thiamine pyrophosphate
adenine
guanine
lysine
glycine
pre-queosine1
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

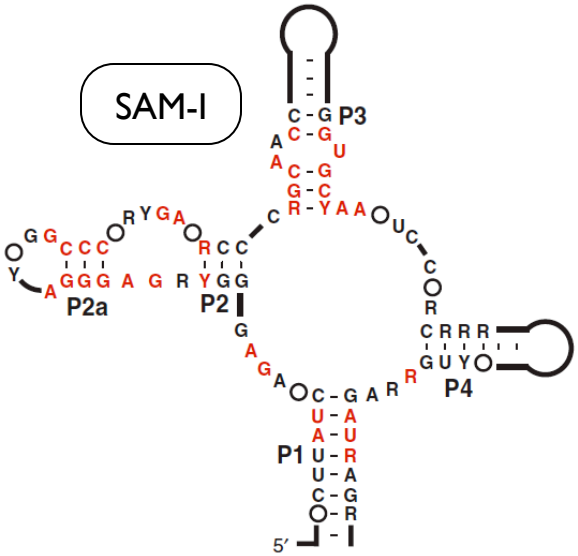
Alberts, et al, 3e.



The protein way

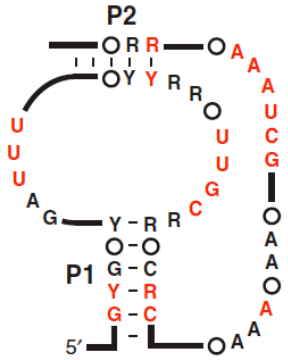
Riboswitch alternatives

SAM-I



Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II

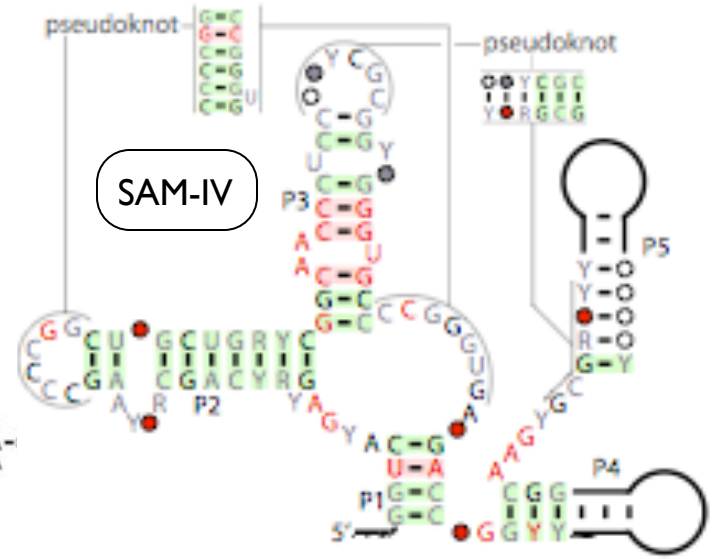


Corbino et al., Genome Biol. 2005

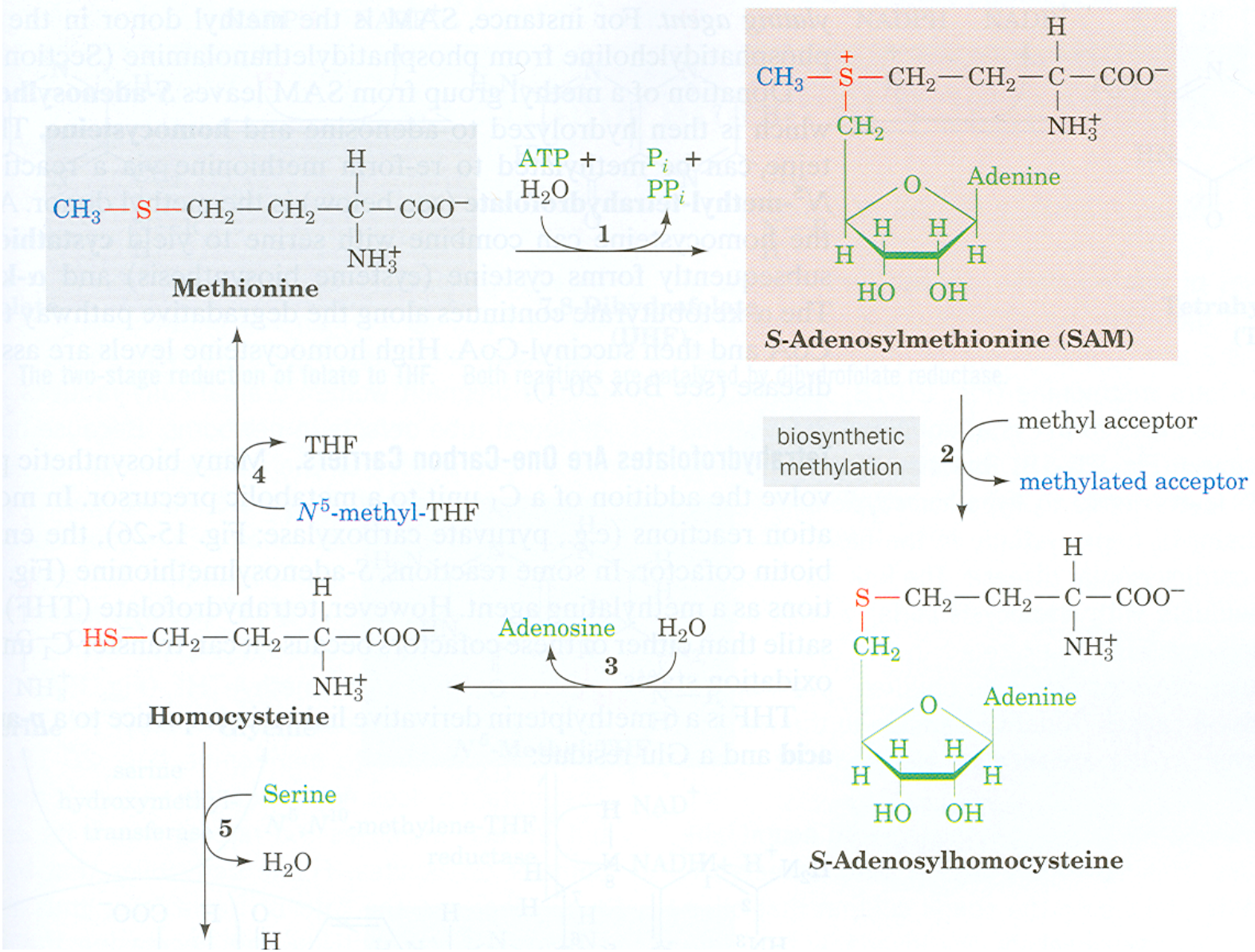
SAM-III

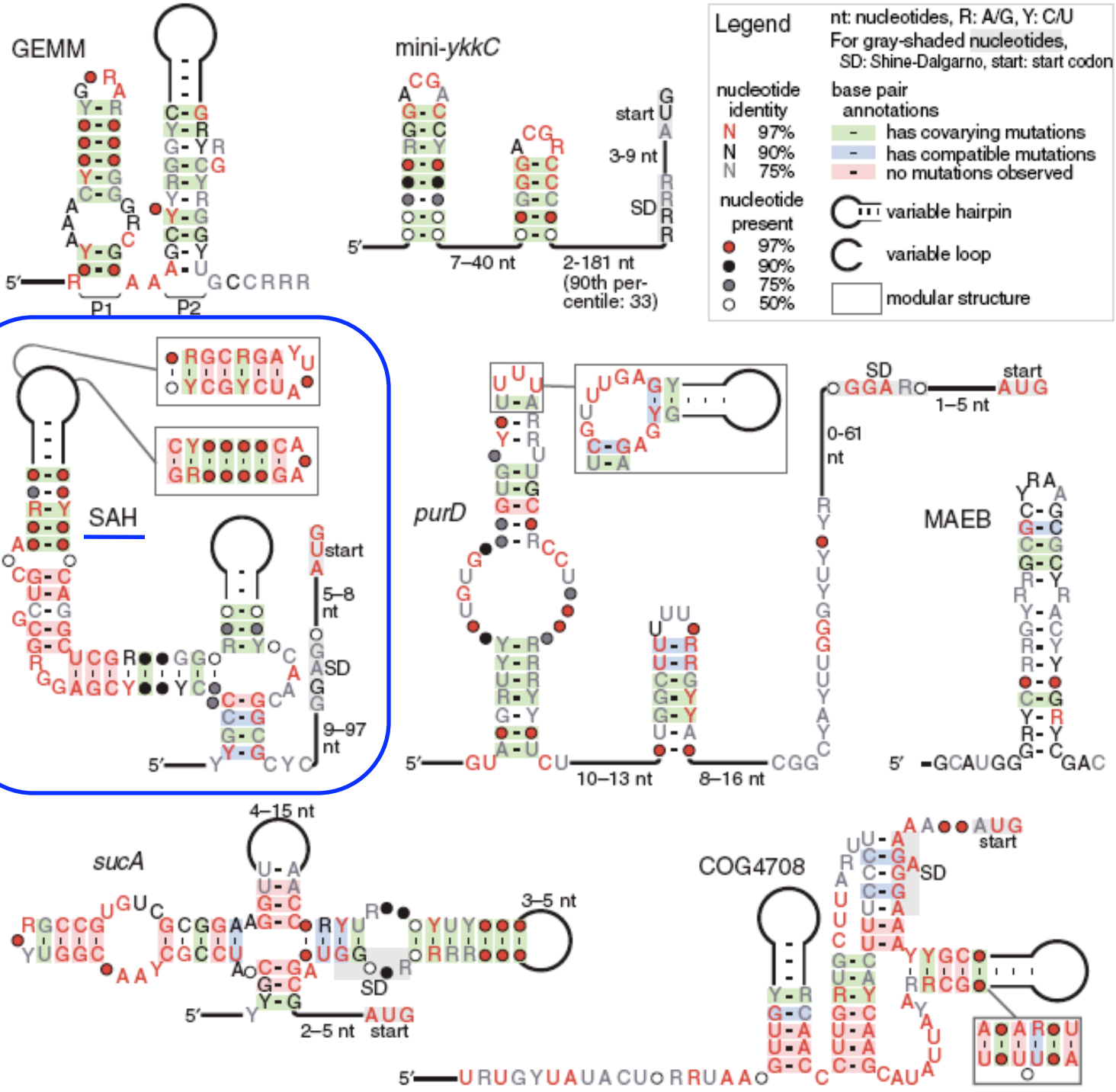
Fuchs et al., NSMB 2006

SAM-IV



Weinberg et al., RNA 2008


Methionine
ATP + H2O
Pi + PPi
1
S-Adenosylmethionine (SAM)
Adenine
biosynthetic methylation
2
methyl acceptor
methylated acceptor
THF
4
N5-methyl-THF
Homocysteine
Adenosine
H2O
3
S-Adenosylhomocysteine
Serine
5
H2O


GEMM
P1
P2
mini-ykkC
start
3-9 nt
SD
7–40 nt
2-181 nt
(90th per-
centile: 33)
Legend
nt: nucleotides, R: A/G, Y: C/U
For gray-shaded nucleotides,
SD: Shine-Dalgarno, start: start codon
nucleotide identity
N 97%
N 90%
N 75%
base pair annotations
has covarying mutations
has compatible mutations
no mutations observed
nucleotide present
97%
90%
75%
50%
variable hairpin
variable loop
modular structure
SAH
start
5–8 nt
SD
9–97 nt
purD
10–13 nt
8–16 nt
SD
start
1–5 nt
0-61 nt
MAEB
sucA
4–15 nt
3–5 nt
SD
2–5 nt
start
COG4708
start
SD

Wanted

Good structure prediction tools

Good motif descriptions/models

Good, fast search tools

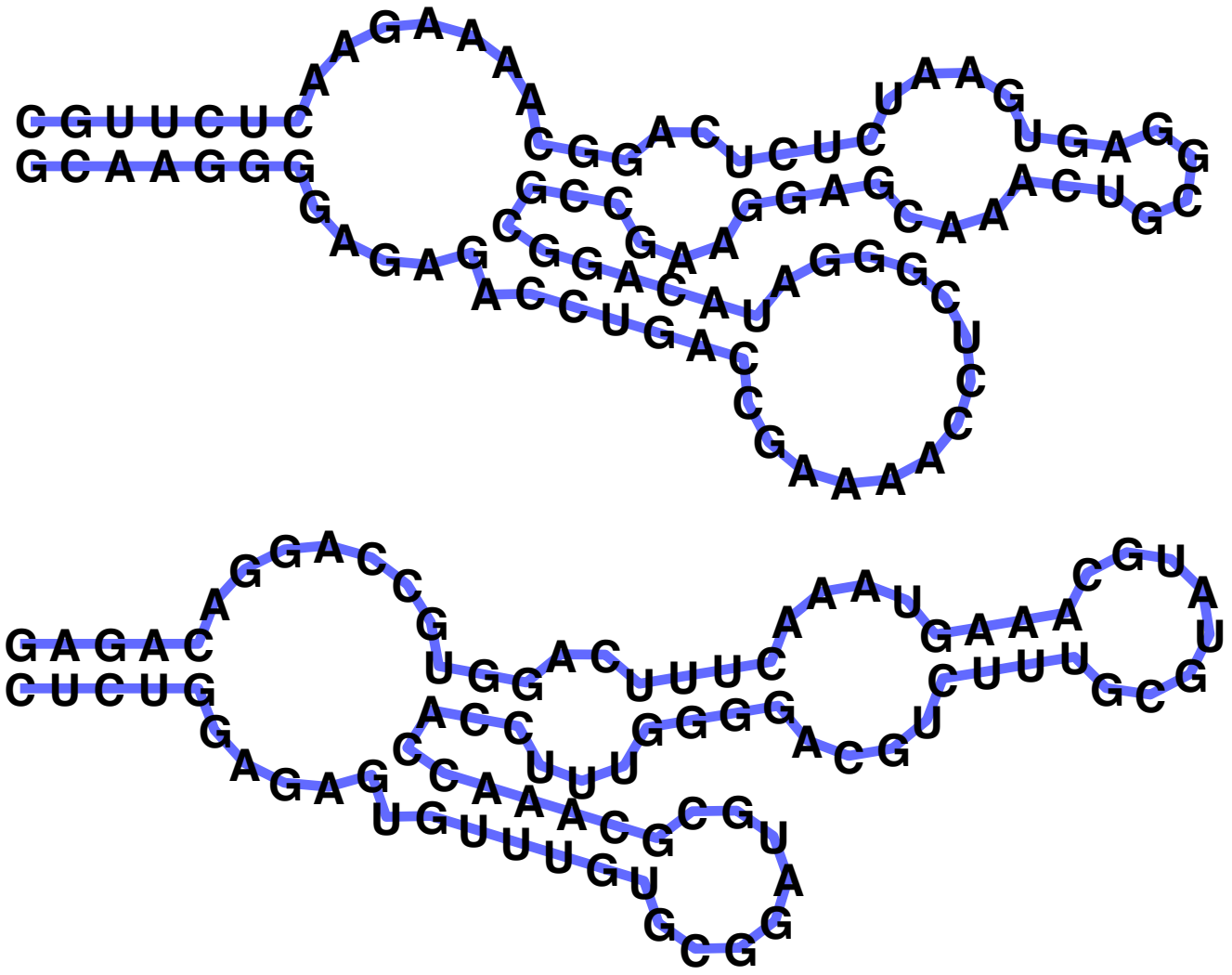
("RNA BLAST", etc.)

Good, fast motif discovery tools

("RNA MEME", etc.)

Importance of structure makes last 3 hard

Why is RNA hard to deal with?



A: Only 29% identity! *Structure* often trumps *sequence*

Structure Prediction

RNA Structure

Primary Structure: Sequence

Secondary Structure: Pairing

Tertiary Structure: 3D shape

Single Seq Secondary Structure Prediction

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

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

Definitely useful, but obviously imperfect

Motif Description

"RNA sequence analysis using covariance models"

Eddy & Durbin

Nucleic Acids Research, 1994
vol 22 #11, 2079-2088

(see also, Ch 10 of Durbin *et al.*)

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

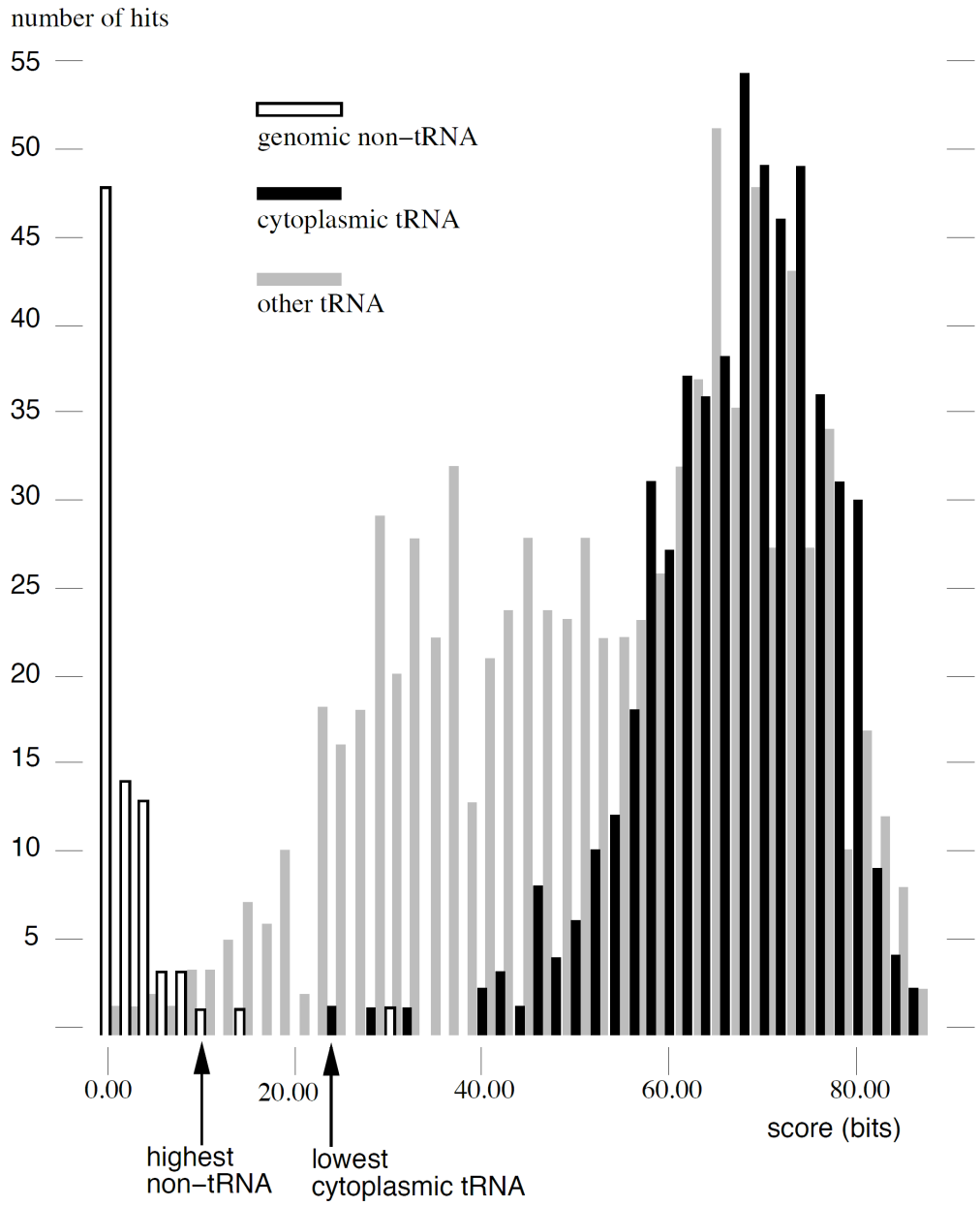
You set a score threshold

Anything above threshold → a "hit"

Scoring:

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

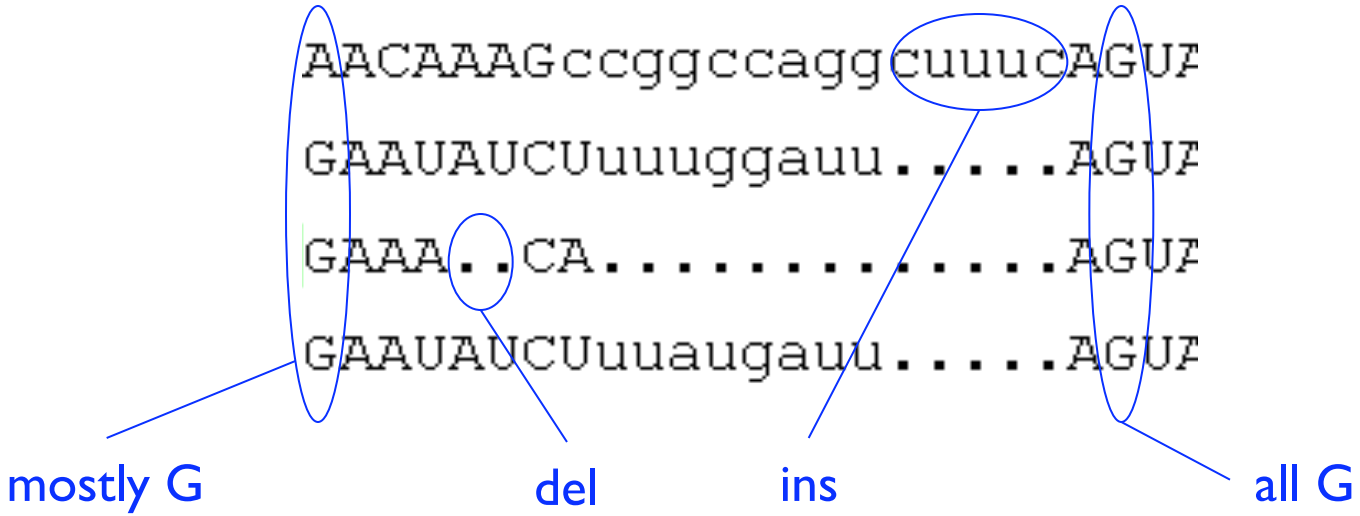
Example: searching for tRNAs



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



Profile Hmm Structure

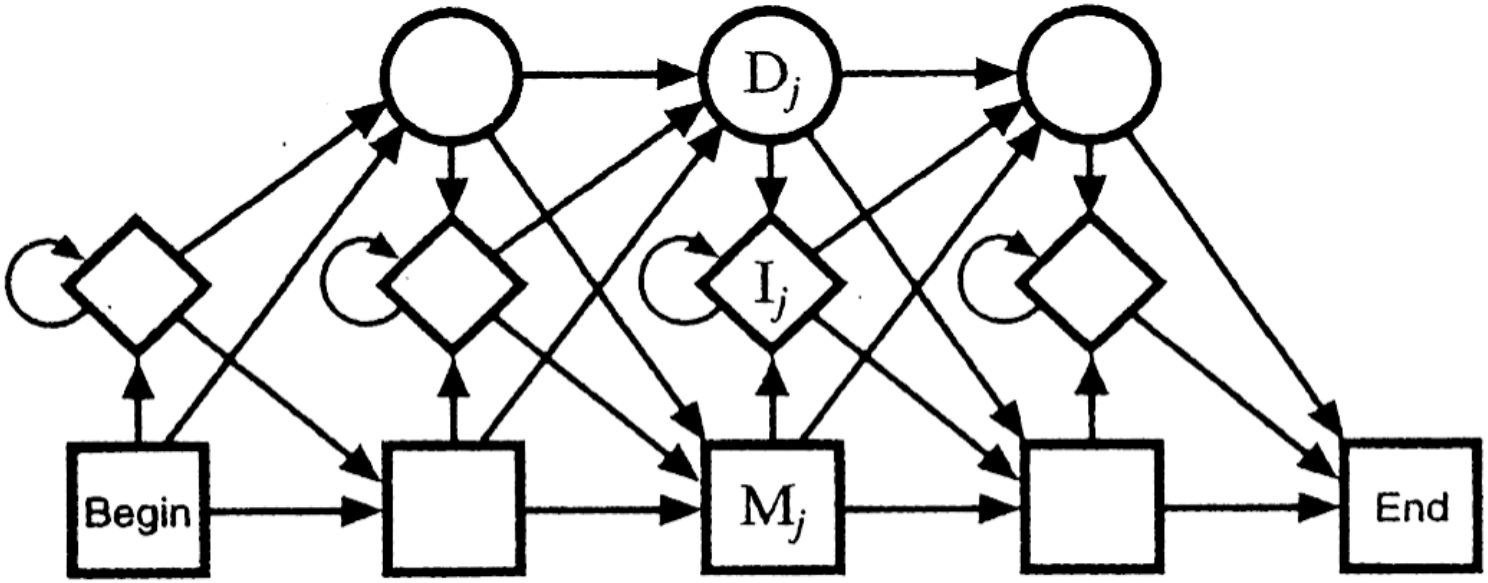


Figure 5.2 *The transition structure of a profile HMM.*

M_j: Match states (4 emission probabilities)

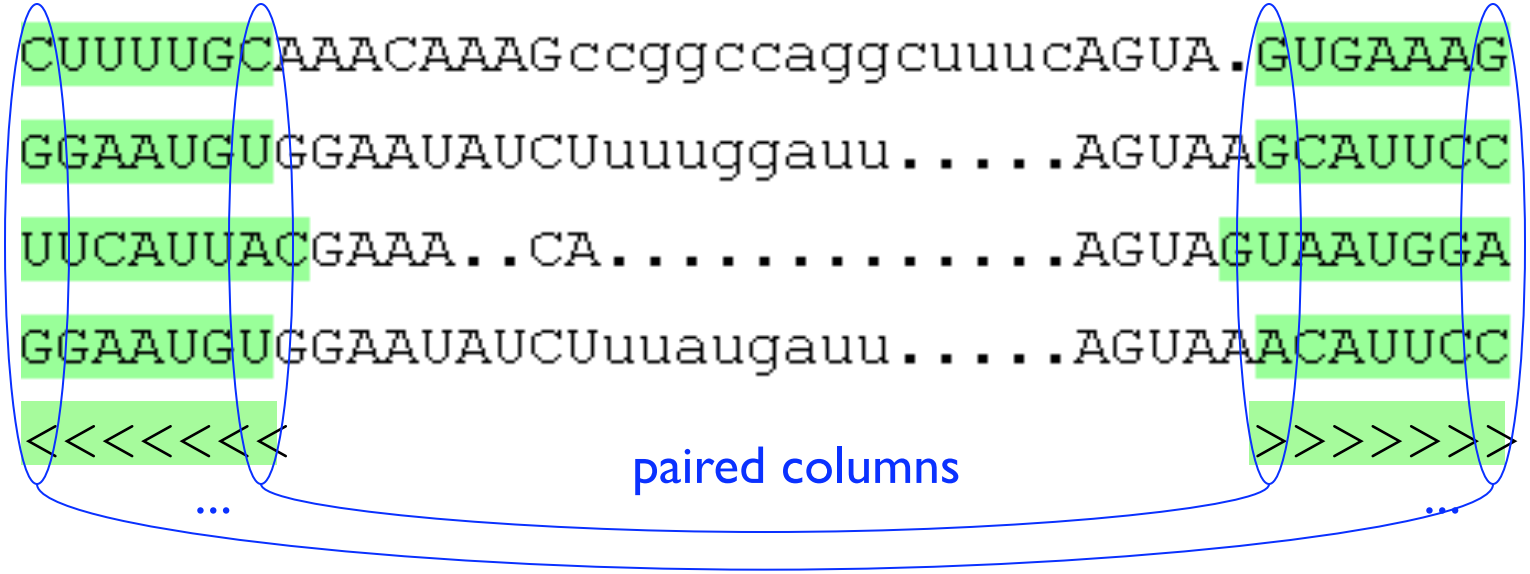
I_j: Insert states (Background emission probabilities)

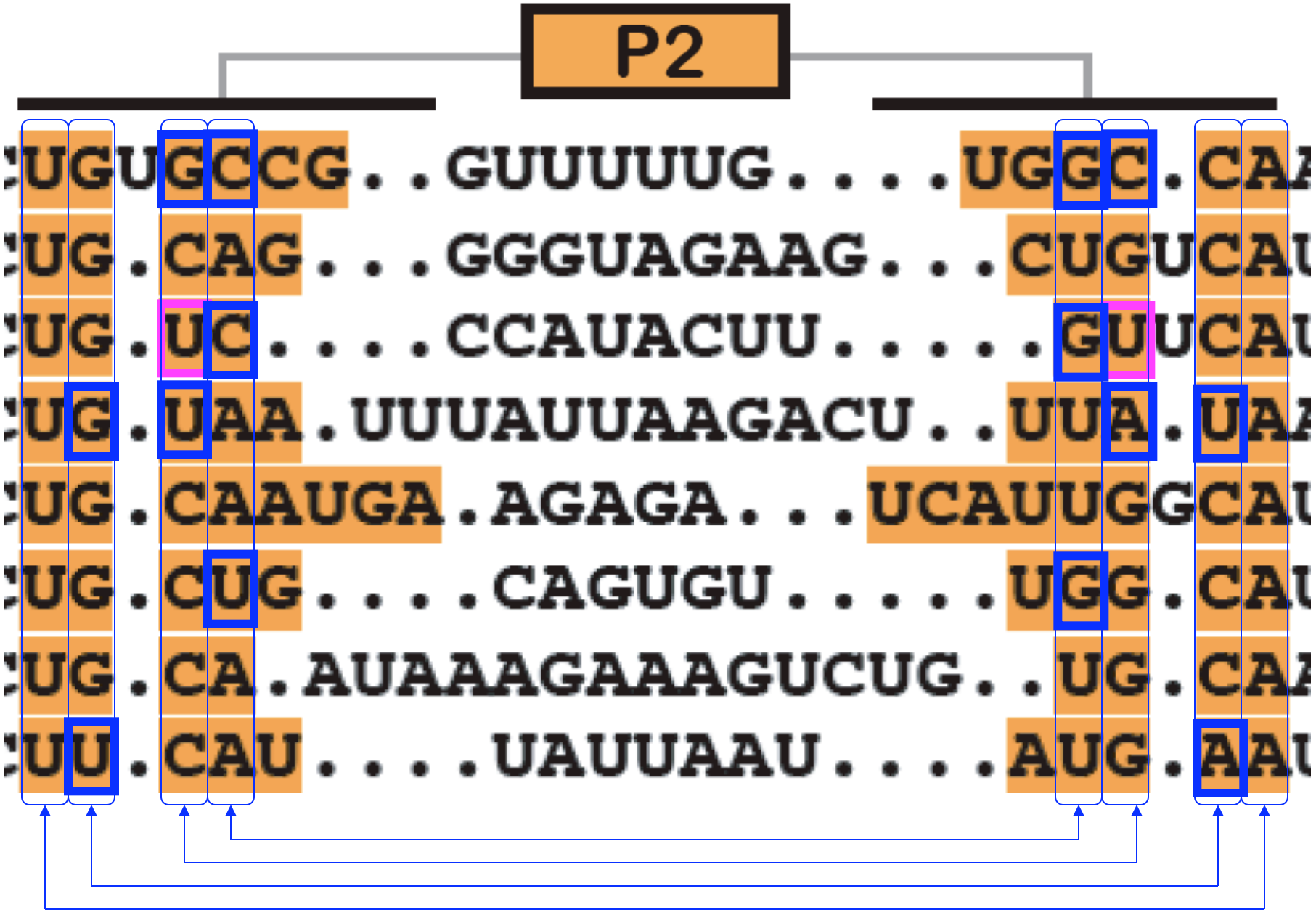
D_j: Delete states (silent - no emission)

How to model an RNA "Motif"?

Covariance Models (aka "profile SCFG")

Probabilistic models, like profile HMMs, but adding "column pairs" and pair emission probabilities for base-paired regions





Covariation is strong evidence for base pairing

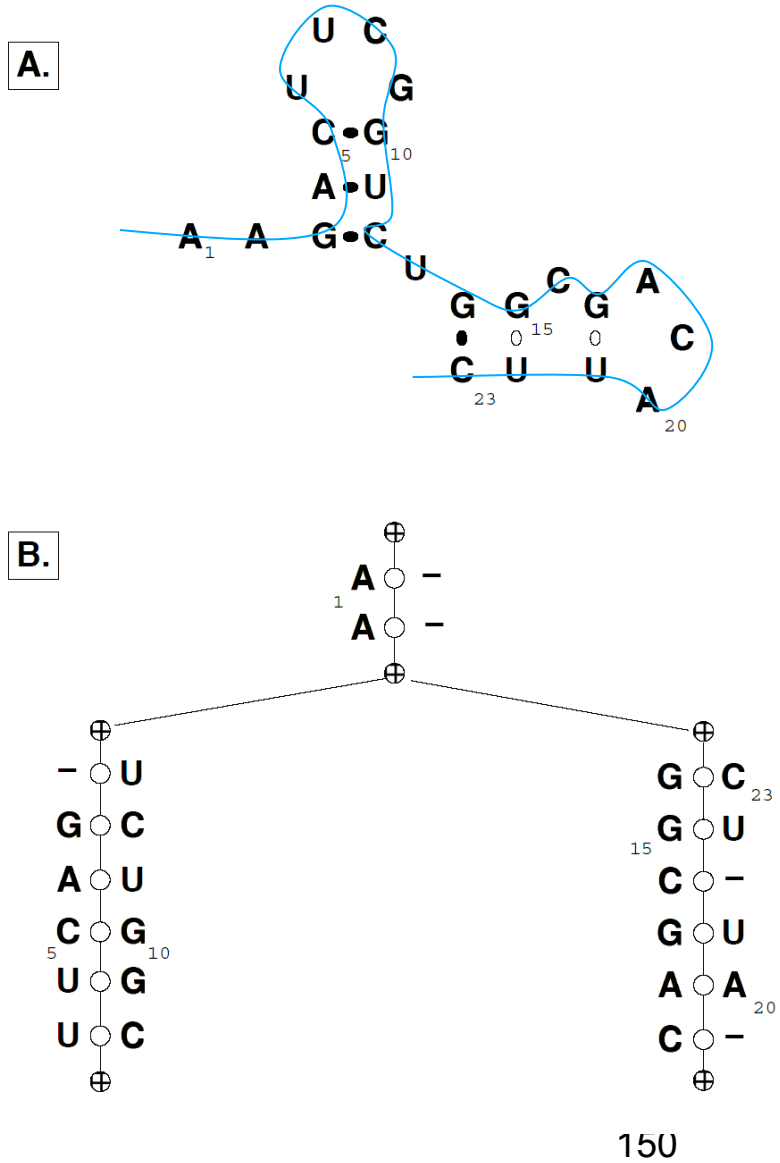
CM Structure

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)



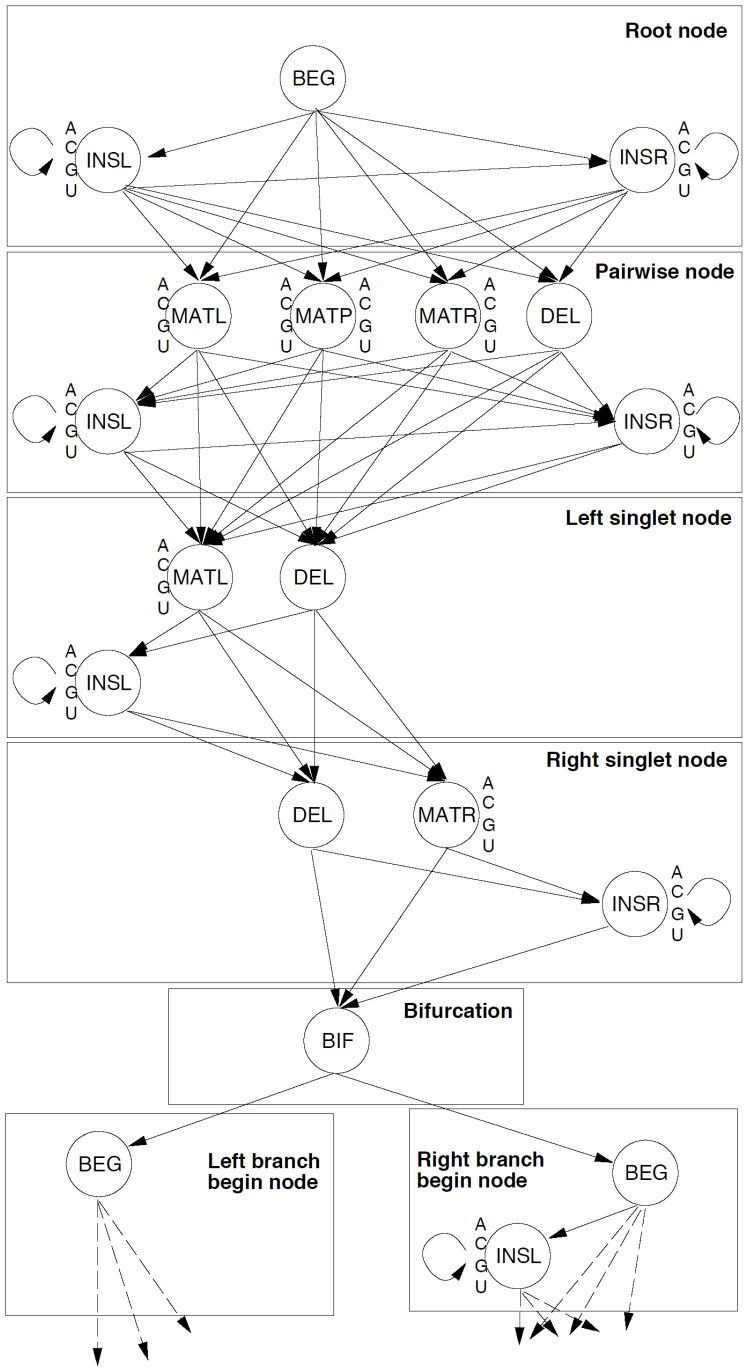
Overall CM Architecture

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

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

Viterbi, cont.

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

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

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

compare: O(qn) for profile HMM

Mutual Information

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

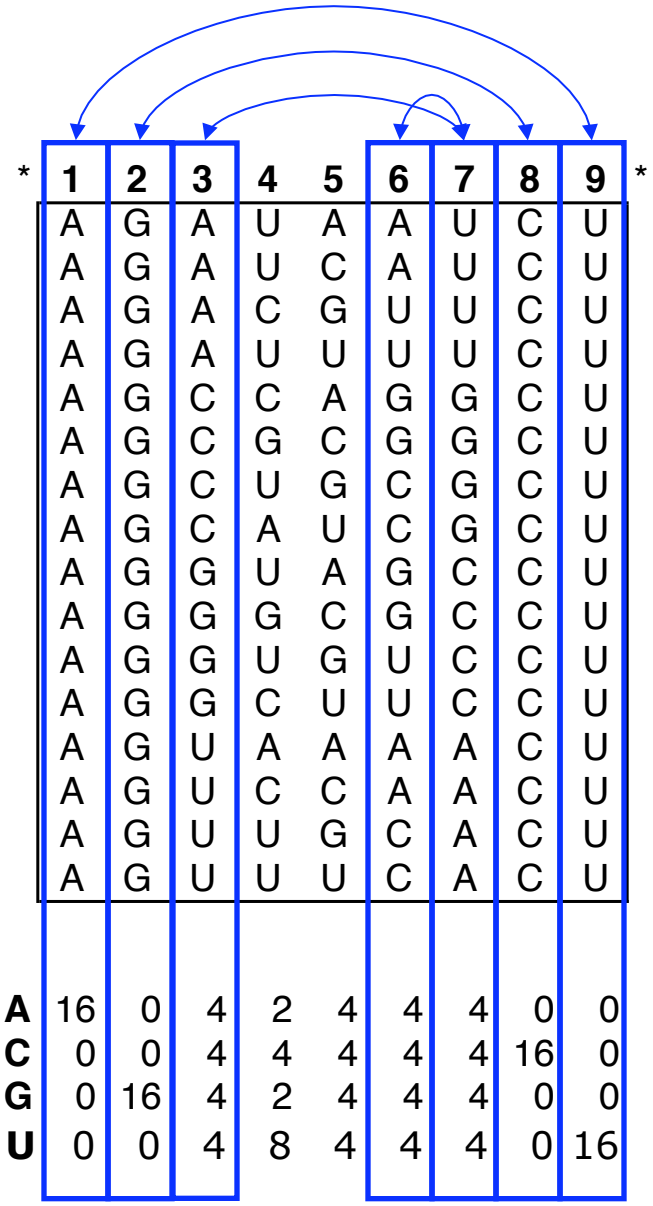
Max when *no* seq conservation but perfect pairing

MI = expected score gain from using a pair state

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

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

M.I. Example (Artificial)



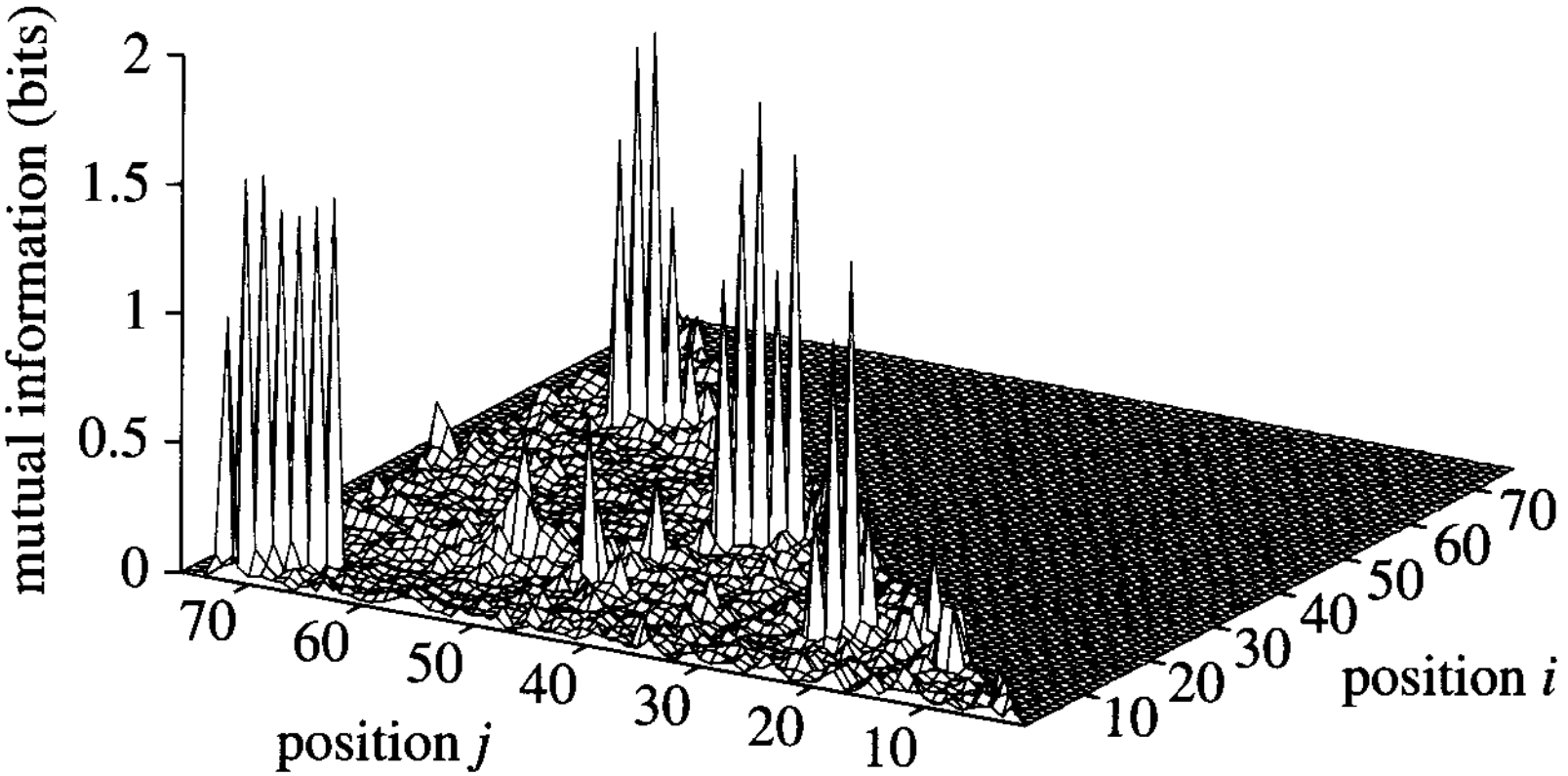
*	1	2	3	4	5	6	7	8	9	*
	A	G	A	U	A	A	U	C	U	
	A	G	A	U	C	A	U	C	U	
	A	G	A	C	G	U	U	C	U	
	A	G	A	U	U	U	U	C	U	
	A	G	C	C	A	G	G	C	U	
	A	G	C	G	C	G	G	C	U	
	A	G	C	U	G	C	G	C	U	
	A	G	C	A	U	C	G	C	U	
	A	G	G	U	A	G	C	C	U	
	A	G	G	G	C	G	C	C	U	
	A	G	G	U	G	U	C	C	U	
	A	G	G	C	U	U	C	C	U	
	A	G	U	A	A	A	A	C	U	
	A	G	U	C	C	A	A	C	U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



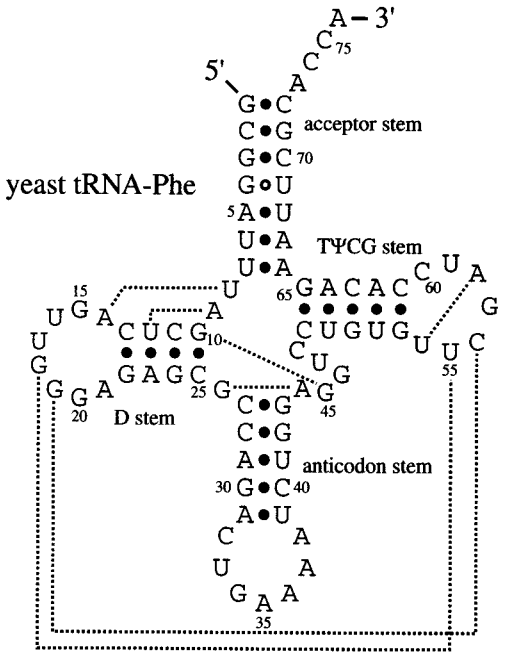


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary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

Modeling and Searching for Non-Coding RNA

W.L. Ruzzo

http://www.cs.washington.edu/homes/ruzzo

http://www.cs.washington.edu/homes/ruzzo/courses/gs541/09sp

Outline

- Whirlwind tour of ncRNA search & discovery
  - RNA motif description (Covariance Model Review)
  - Motif search
    - Rigorous & heuristic filtering
  - Motif discovery
  - Applications
    - Prokaryotes
    - Vertebrates
  - Open problems

An Important Application: Rfam

Rfam – an RNA family DB

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

Biggest scientific computing user in Europe - 1000 cpu cluster for a month per release

Rapidly growing:

Rel 1.0, 1/03: 25 families, 55k instances

Rel 7.0, 3/05: 503 families, >300k instances

Rel 9.0, 7/08: 603 families, 896k instances

Rel 9.1, 1/09: 1372 families, ??? instances

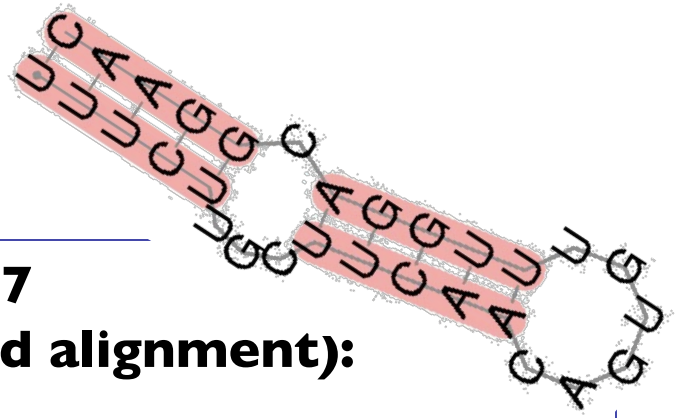
Rfam

Input (hand-curated):

- MSA "seed alignment"
- SS_cons
- Score Threshold
- Window Len W

Output:

- CM
- scan results & "full alignment"



IRE RF00037
(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

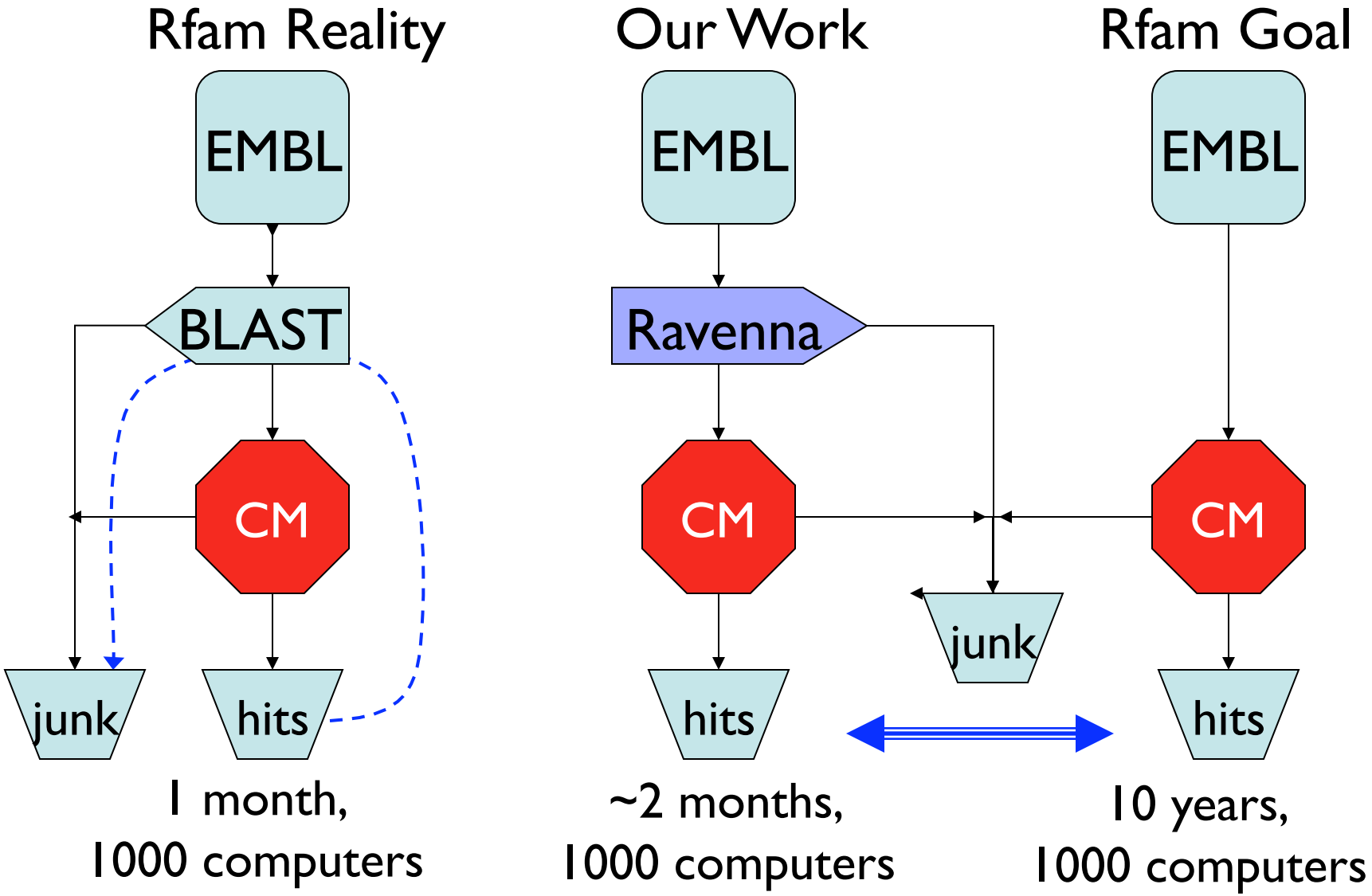
RaveNnA: Genome Scale RNA Search

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

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

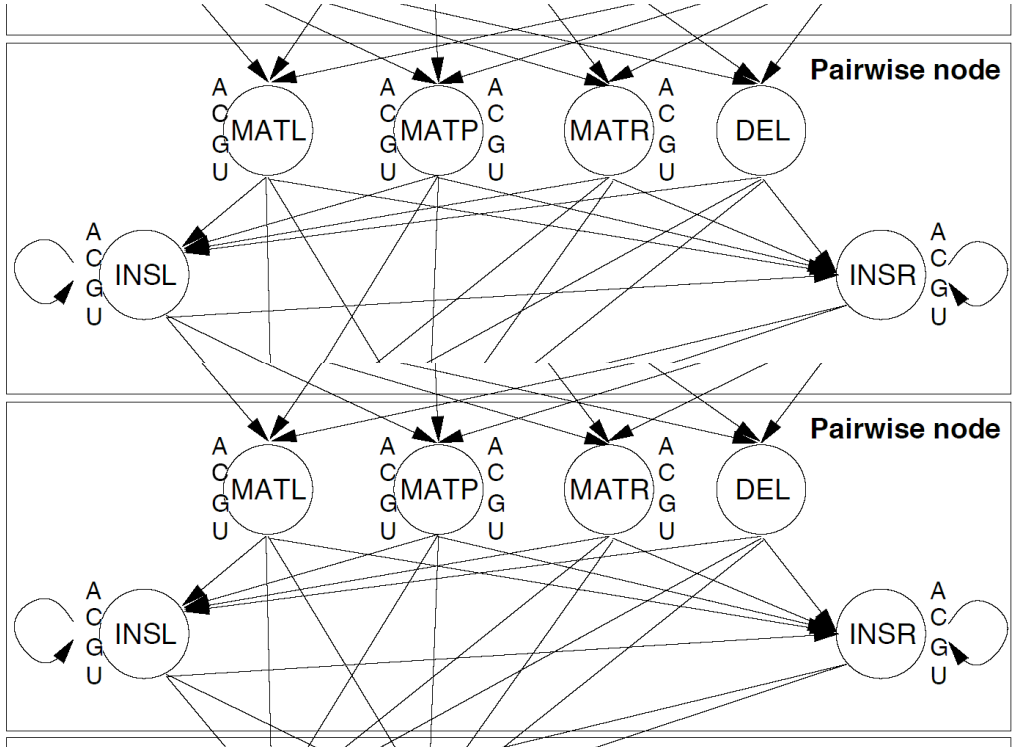
Weinberg & Ruzzo, *Bioinformatics*, 2004, 2006

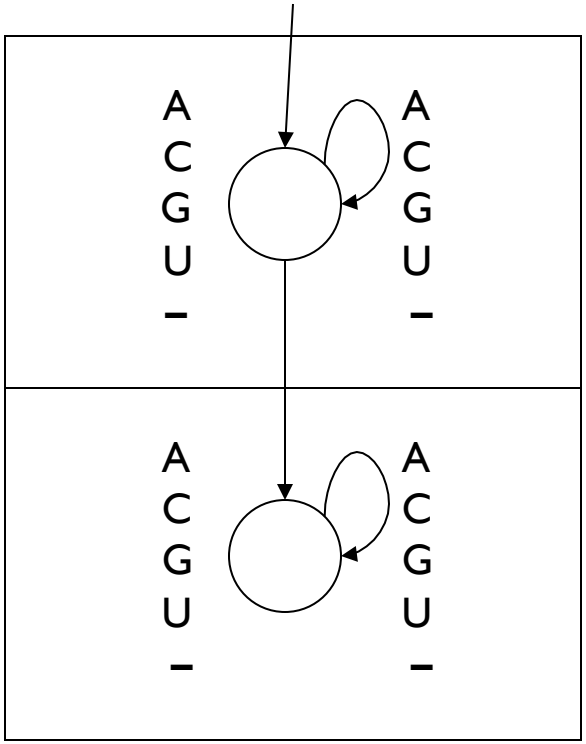
CM's are good, but slow



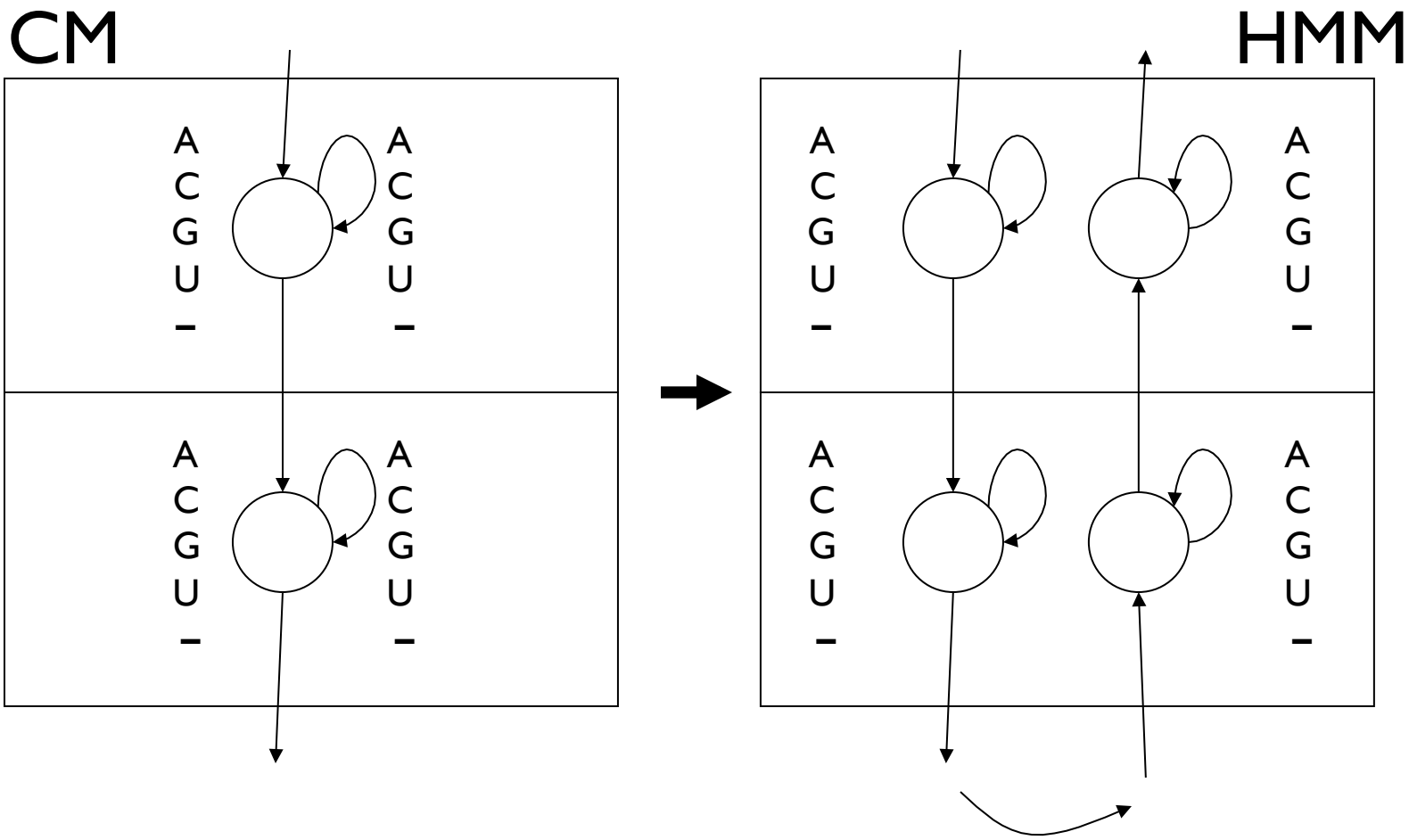
Oversimplified CM

(for pedagogical purposes only)





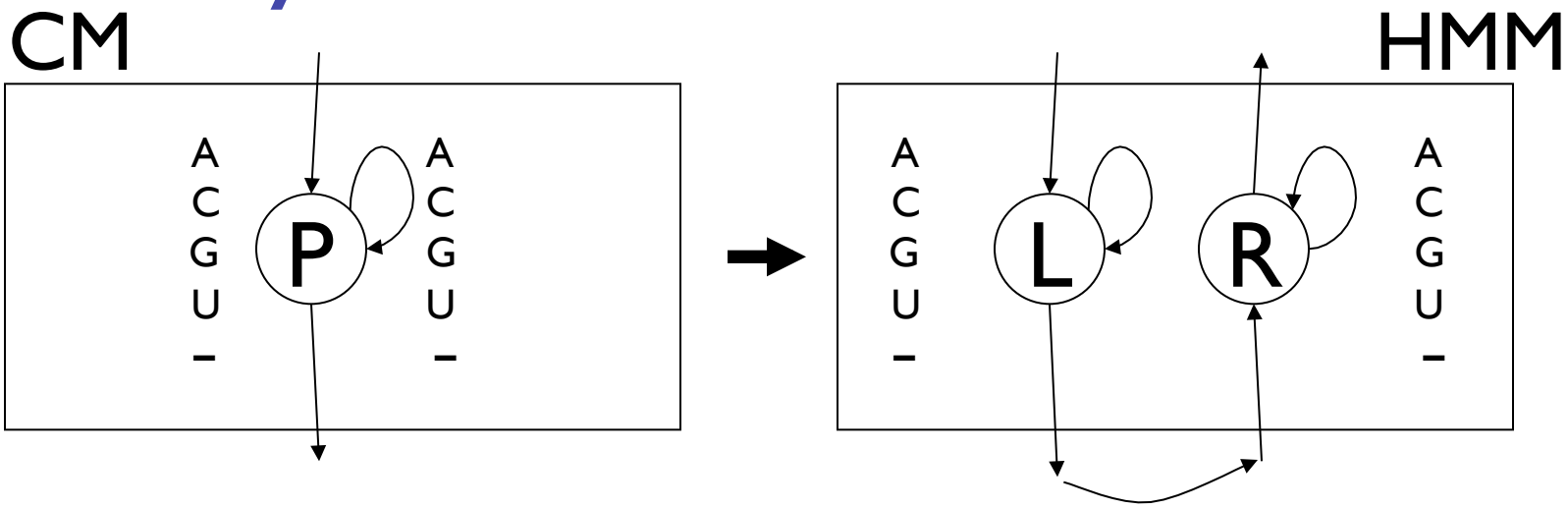
CM to HMM



25 emisions per state

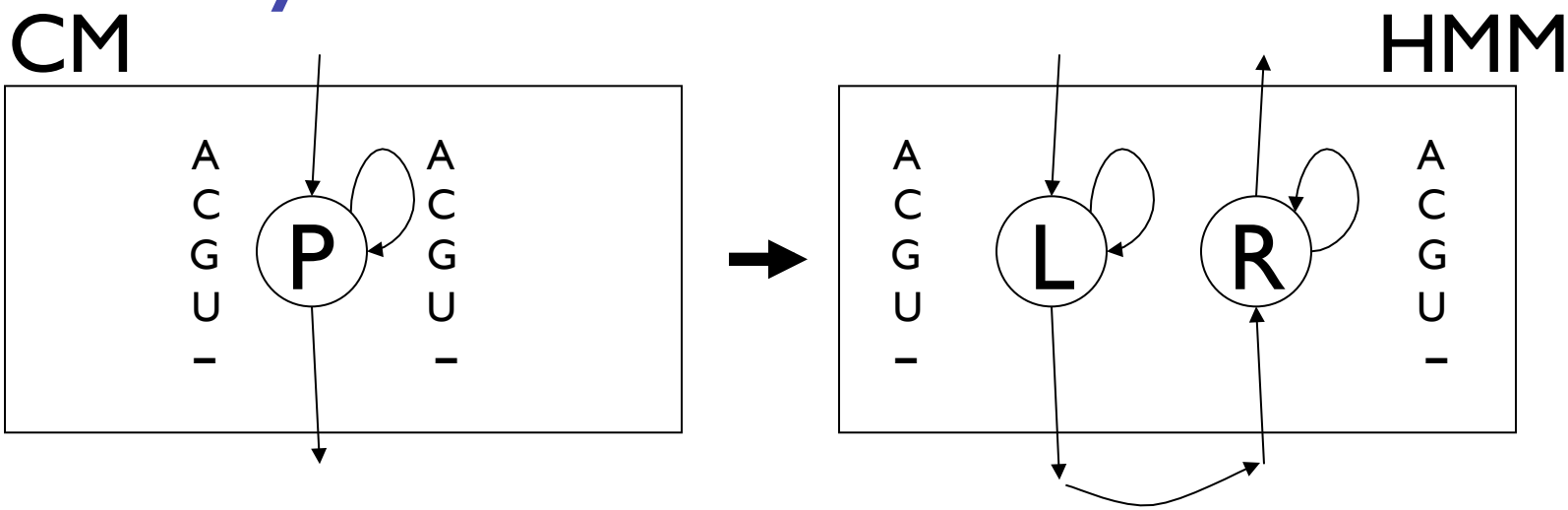
5 emissions per state, 2x states

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

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

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: $L_U + (R_A + R_G)/2 = 2.5$

What should the scores be?

Convex optimization problem

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

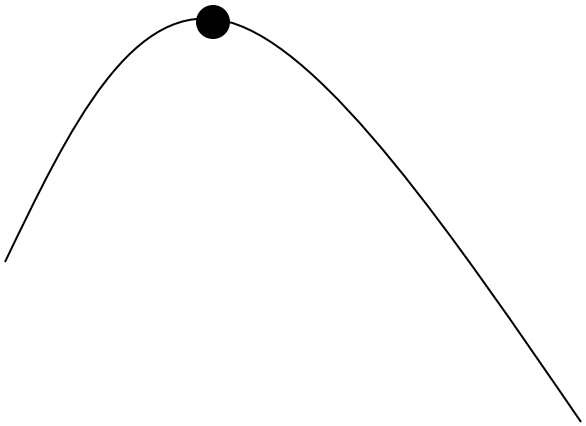
Problem sizes:

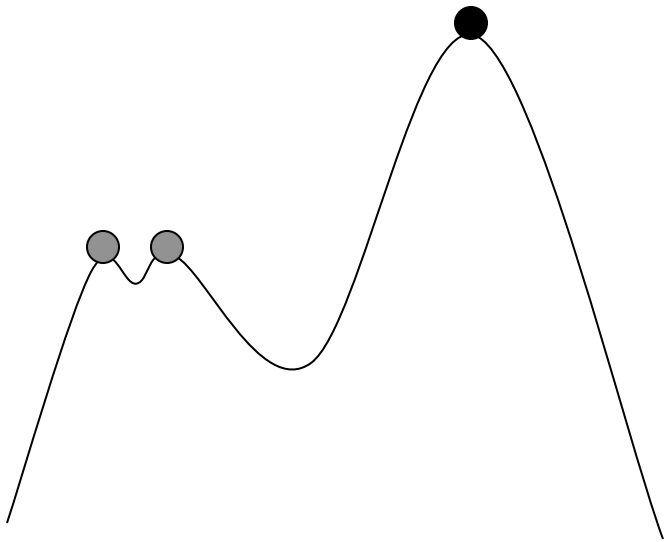
- 1000-10000 variables
- 10000-100000 inequality constraints

“Convex” Optimization

Convex:
local max = global max;
simple “hill climbing” works

Nonconvex:
can be many local maxima,
<< global max;
“hill-climbing” fails





Estimated Filtering Efficiency

(139 Rfam 4.0 families)

Filtering fraction	# families (compact)	# families (expanded)
< 10^{-4}	105	110
10^{-4} - 10^{-2}	8	17
.01 - .10	11	3
.10 - .25	2	2
.25 - .99	6	4
.99 - 1.0	7	3

≈ break even

~100x speedup

Averages 283 times faster than CM

Results: New ncRNA's?

Name	# found BLAST + CM	# found rigorous filter + CM	# new
Pyrococcus snoRNA	57	180	123
Iron response element	201	322	121
Histone 3' element	1004	1106	102
Purine riboswitch	69	123	54
Retron msr	11	59	48
Hammerhead I	167	193	26
Hammerhead III	251	264	13
U4 snRNA	283	290	7
S-box	128	131	3
U6 snRNA	1462	1464	2
U5 snRNA	199	200	1
U7 snRNA	312	313	1

Motif Discovery

RNA Motif Discovery

Typical problem: given a ~10-20 unaligned sequences of ~1kb, most of which contain instances of one RNA motif of, say, 150bp -- find it.

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

Approaches

Align sequences, then look for common structure

Predict structures, then try to align them

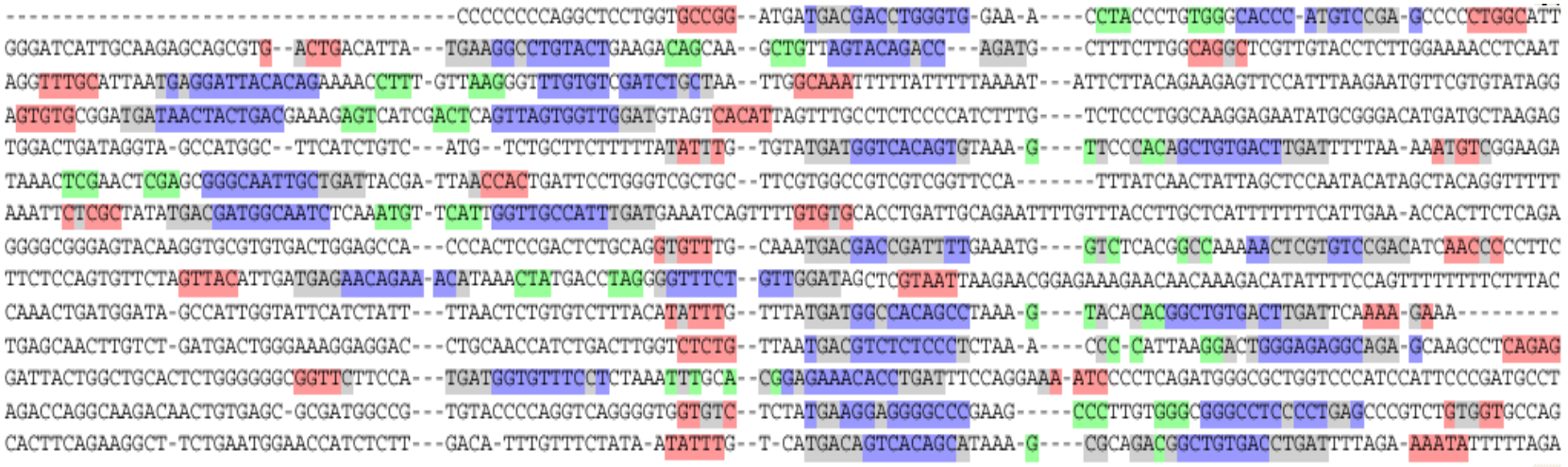
Do both together

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 sequences, then look for common structure

Predict structures, then try to align them

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

Do both together

Our Approach: CMfinder

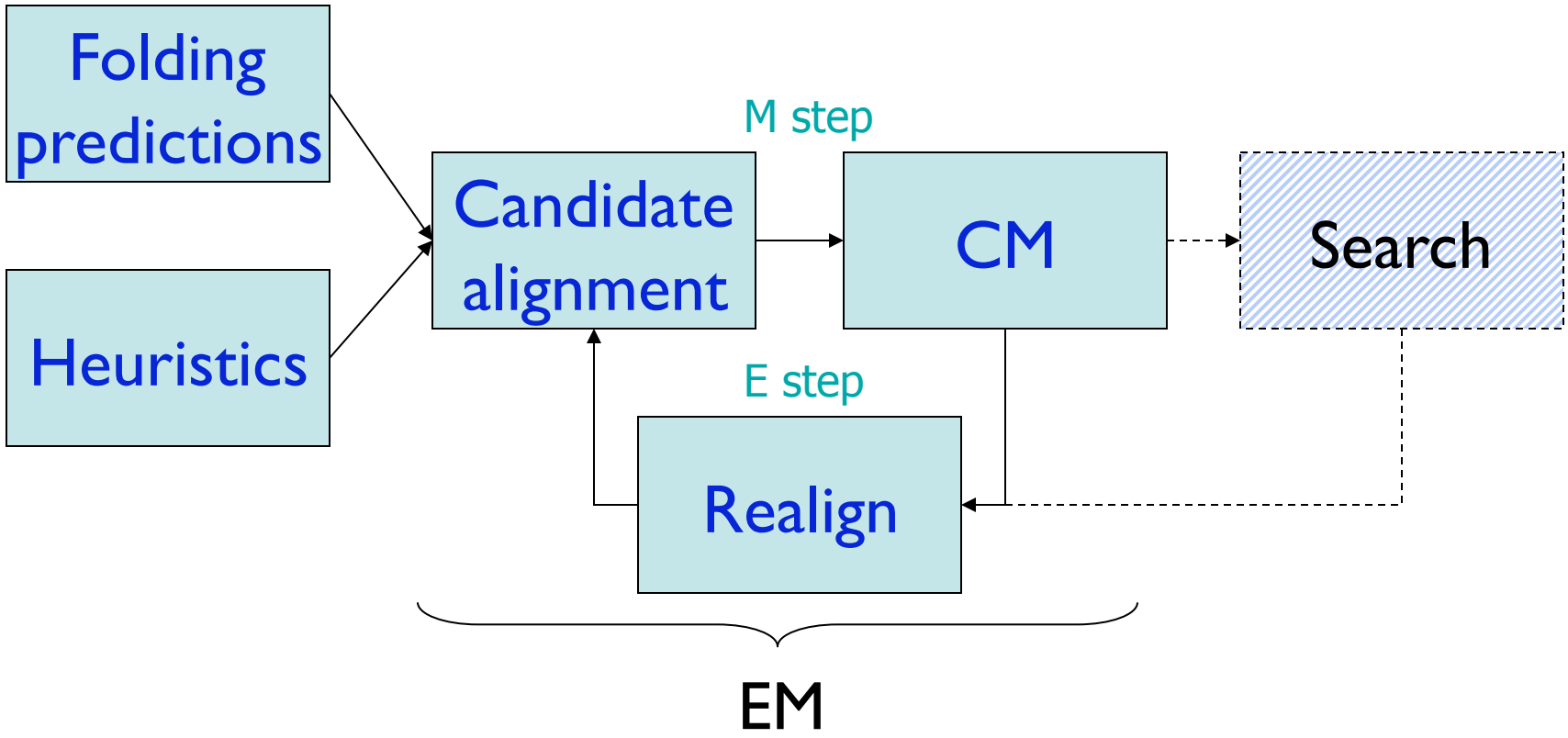
RNA motifs from unaligned sequences

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 Outline



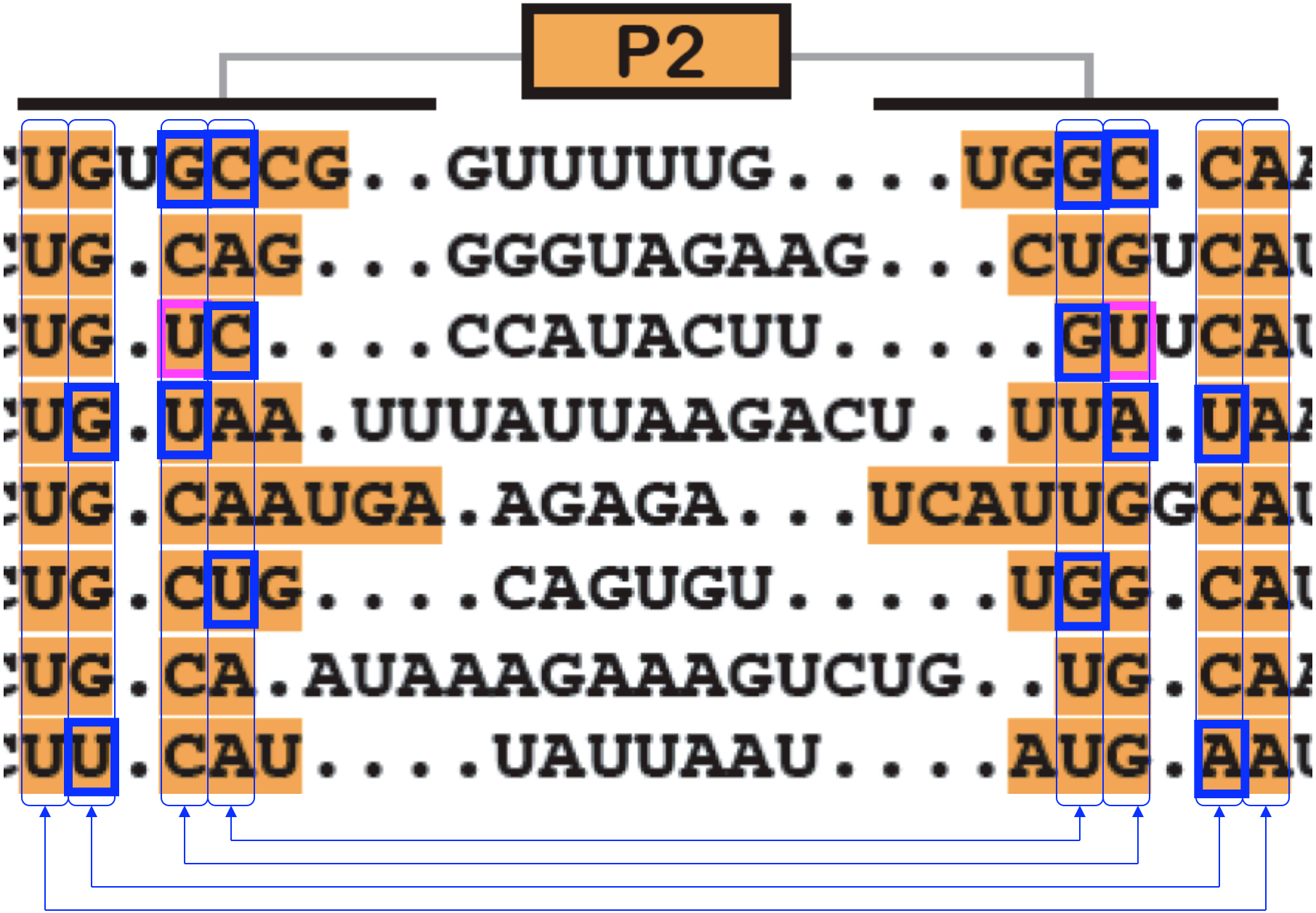
M-step uses M.I. + folding energy for structure prediction

Structure Inference

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

We combine:

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


P2
UGUGCCG..GUUUUUG....UGGC.CA
UG.CAG...GGGUAGAAG...CUGUCA
UG.UC....CCAUACUU.....GUUCA
UG.UAA.UUUAUUAAGACU..UUA.UA
UG.CAAUGA.AGAGA...UCAUUGGCA
UG.CUG....CAGUGU.....UGG.CA
UG.CA.AUAAAGAAAGUCUG..UG.CA
UU.CAU....UAUUAAU....AUG.AA

Mutual Information

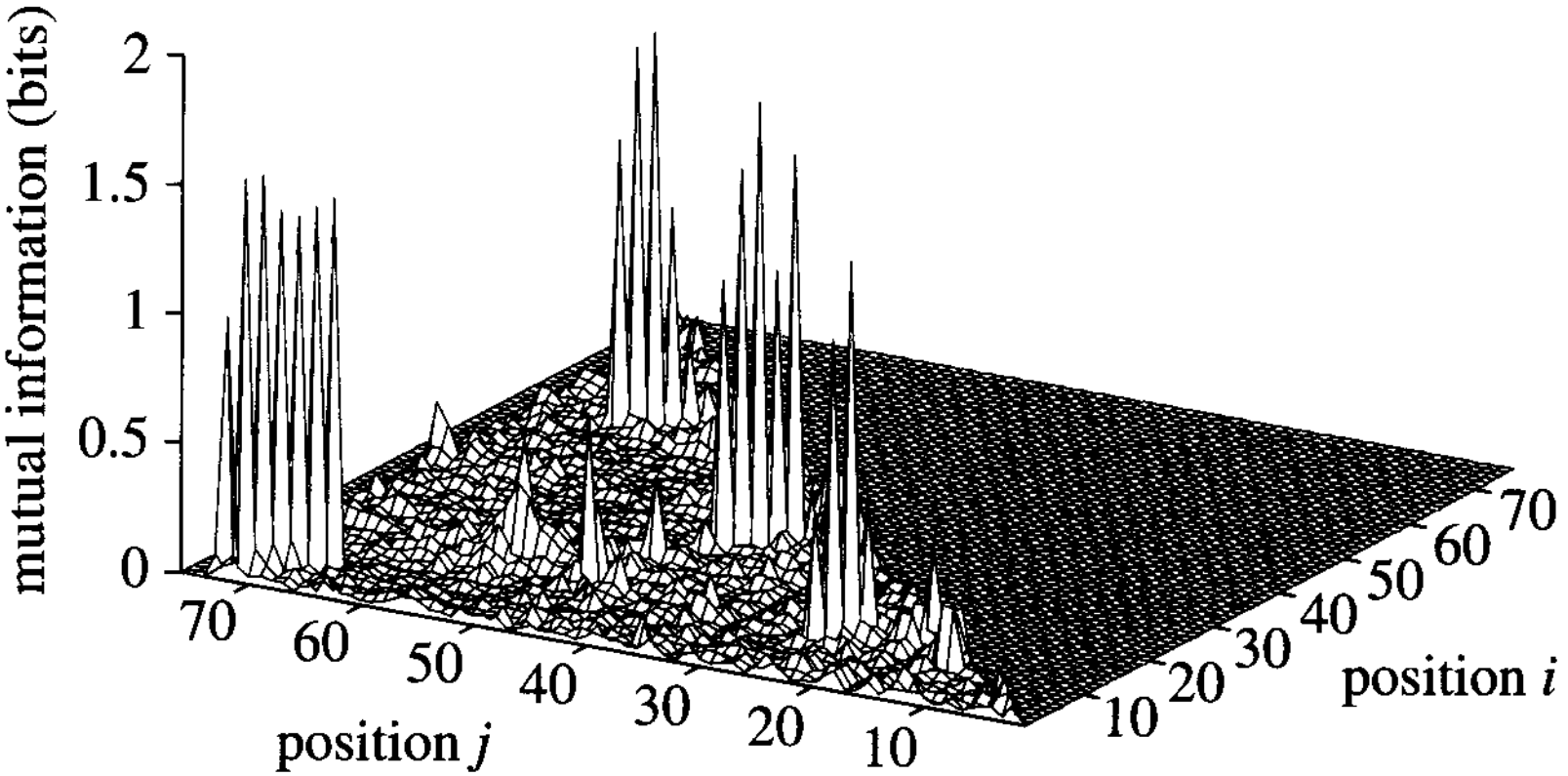
$$M_{ij} = \sum_{xi,xj} f_{xi,xj} \log_2 \frac{f_{xi,xj}}{f_{xi} f_{xj}}; \quad 0 \le M_{ij} \le 2$$

Max when *no* seq conservation but perfect pairing

MI = expected score gain from using a pair state

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

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



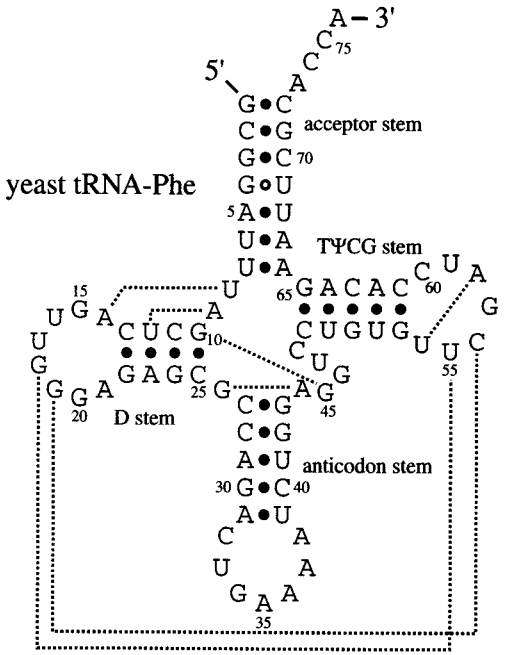
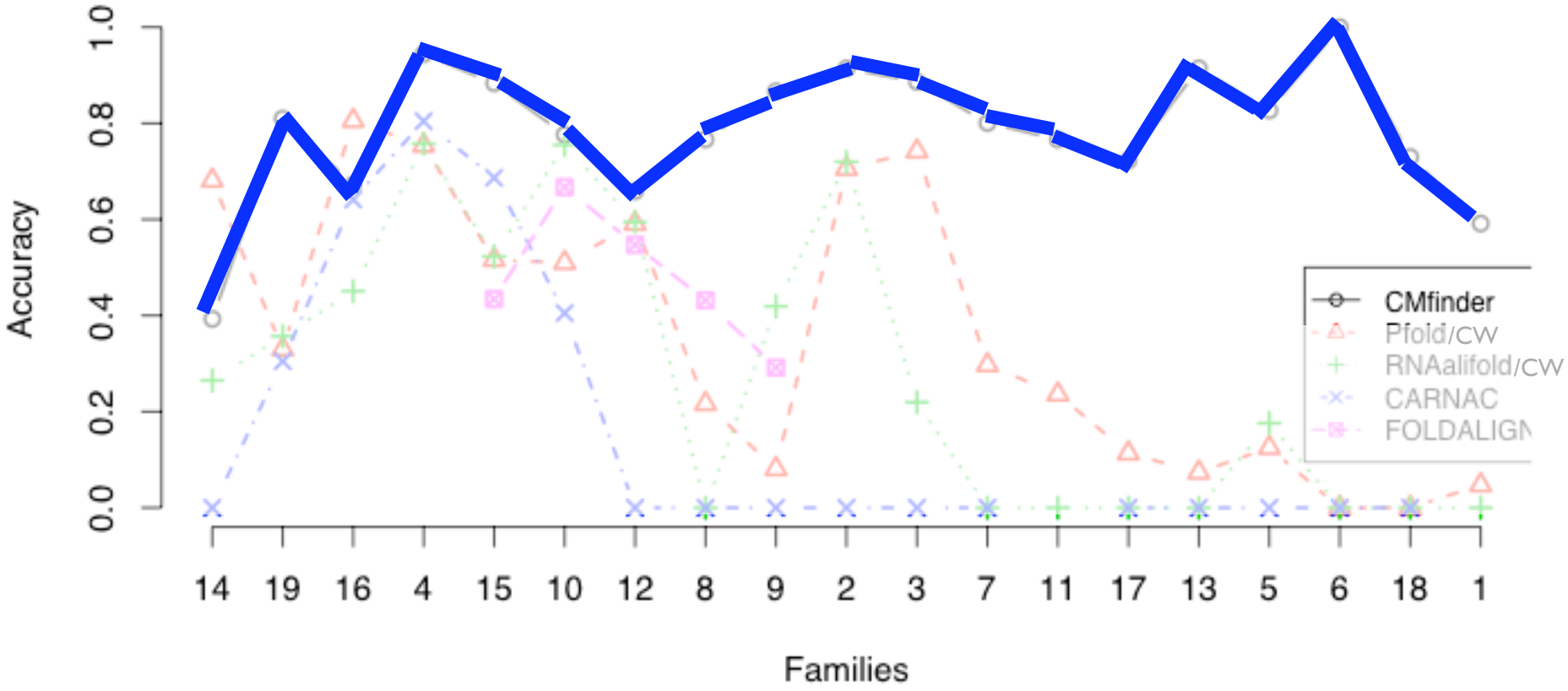


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary contacts observed in the yeast tRNA-Phe crystal structure. Some of these tertiary contacts produce correlated pairs which can be seen weakly in the mutual information plot.*

CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Applications:
ncRNA discovery in
prokaryotes and vertebrates

Key issue in both cases is exploiting prior knowledge to focus on promising data

Application I

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

Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo.
PLoS Computational Biology. 3(7): e126, July 6, 2007.

Right Data: Why/How

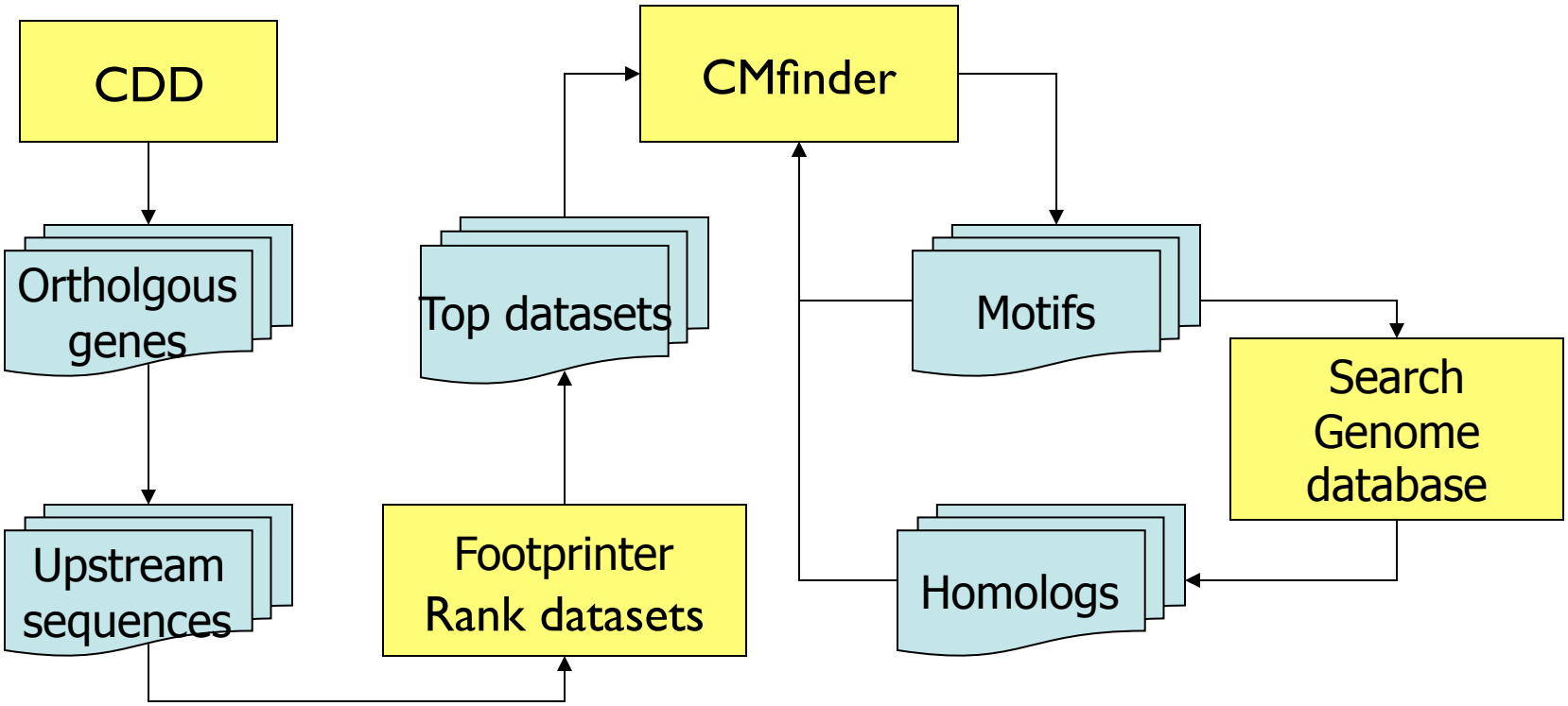
We can recognize, say, 5-10 good examples amidst 20 extraneous ones (but not 5 in 200 or 2000) of length 1k or 10k (but not 100k)

Regulators often near regulatees (protein coding genes), which are usually recognizable cross-species

So, find similar genes ("homologs"), look at adjacent DNA

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

A pipeline for RNA motif genome scans



Yao, Barrick, Weinberg, Neph, Breaker, Tompa and Ruzzo. A Computational Pipeline for High Throughput Discovery of cis-Regulatory Noncoding RNA in Prokaryotes. PLoS Computational Biology. 3(7): e126, July 6, 2007.

Genome Scale Search: Why

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

In most close relatives

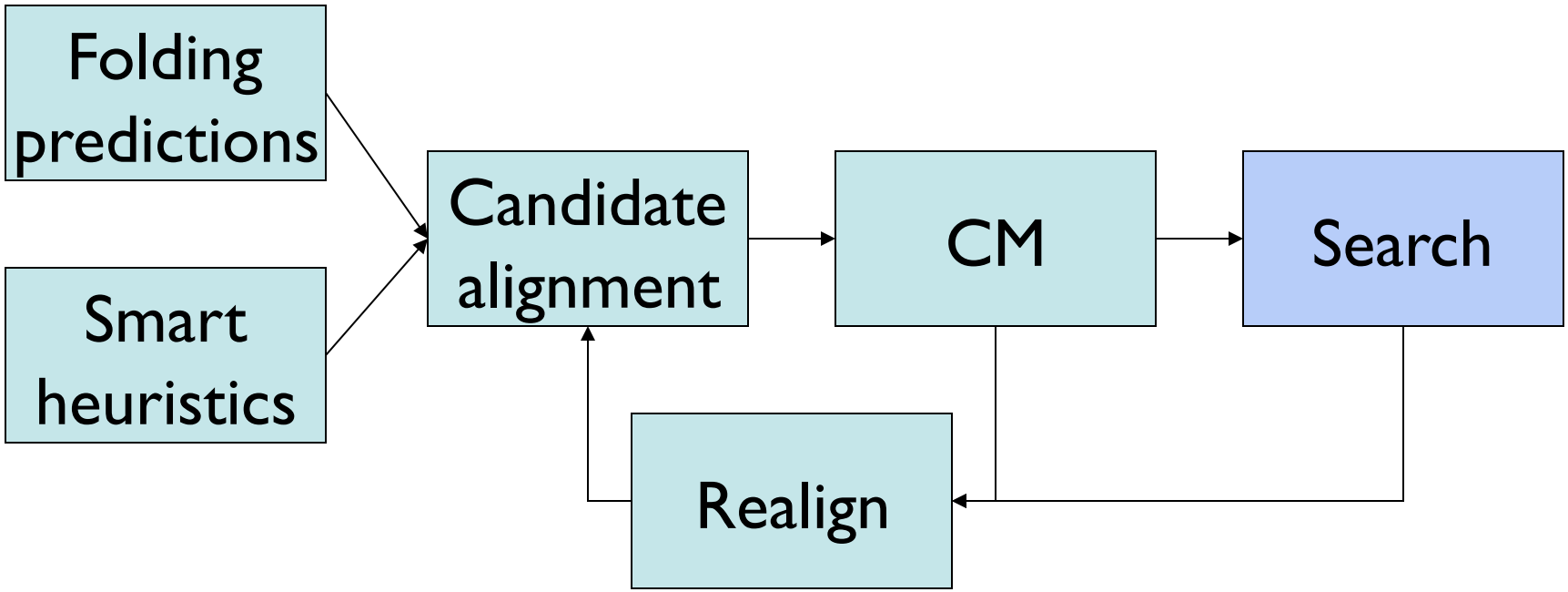
More examples give better model, hence even more examples, fewer errors

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

But inclusion of non-examples can degrade motif…

Genome Scale Search

CMfinder is directly usable for/with search



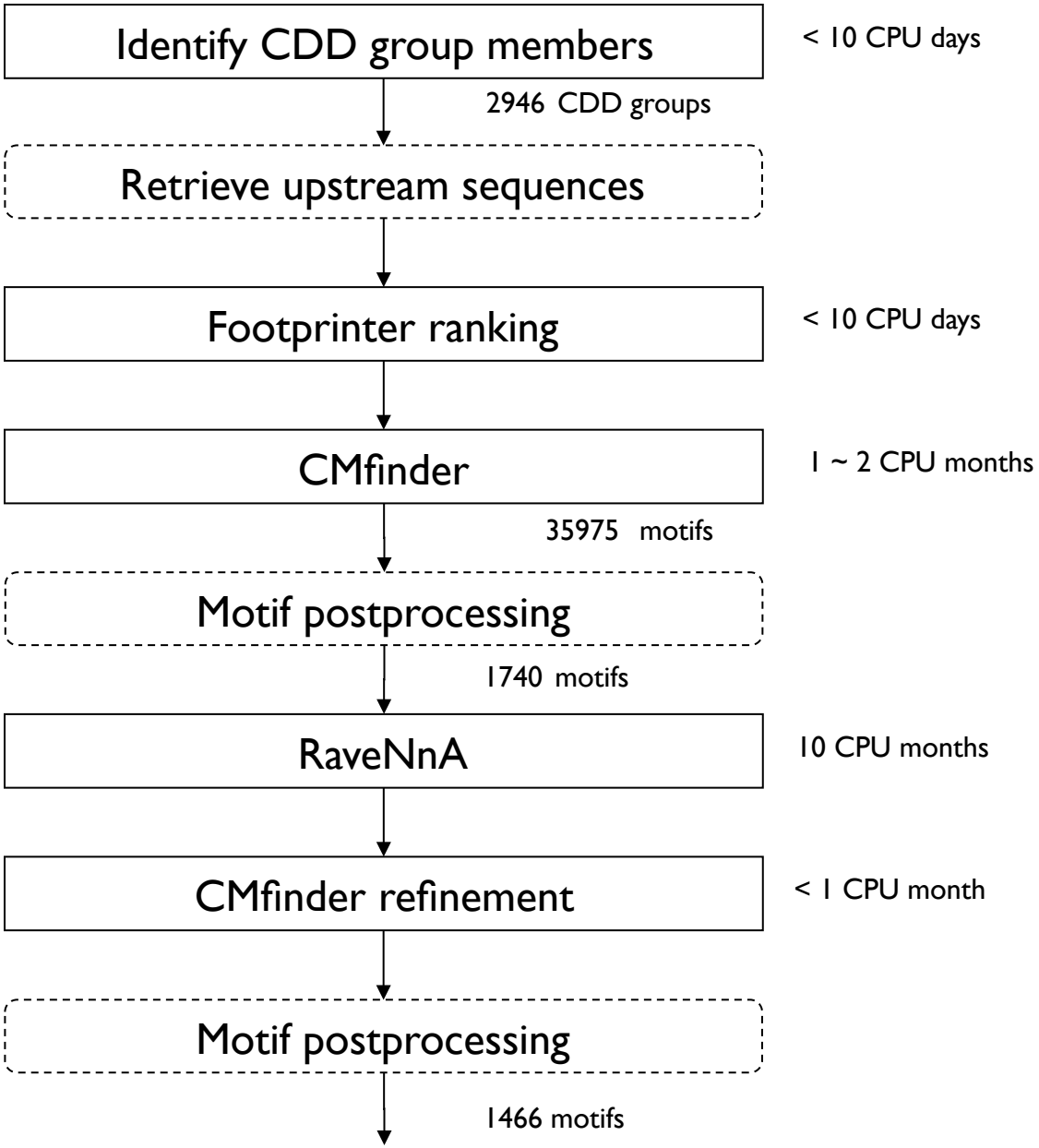

Identify CDD group members
< 10 CPU days
2946 CDD groups
Retrieve upstream sequences
Footprinter ranking
< 10 CPU days
CMfinder
1 ~ 2 CPU months
35975 motifs
Motif postprocessing
1740 motifs
RaveNnA
10 CPU months
CMfinder refinement
< 1 CPU month
Motif postprocessing
1466 motifs

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"

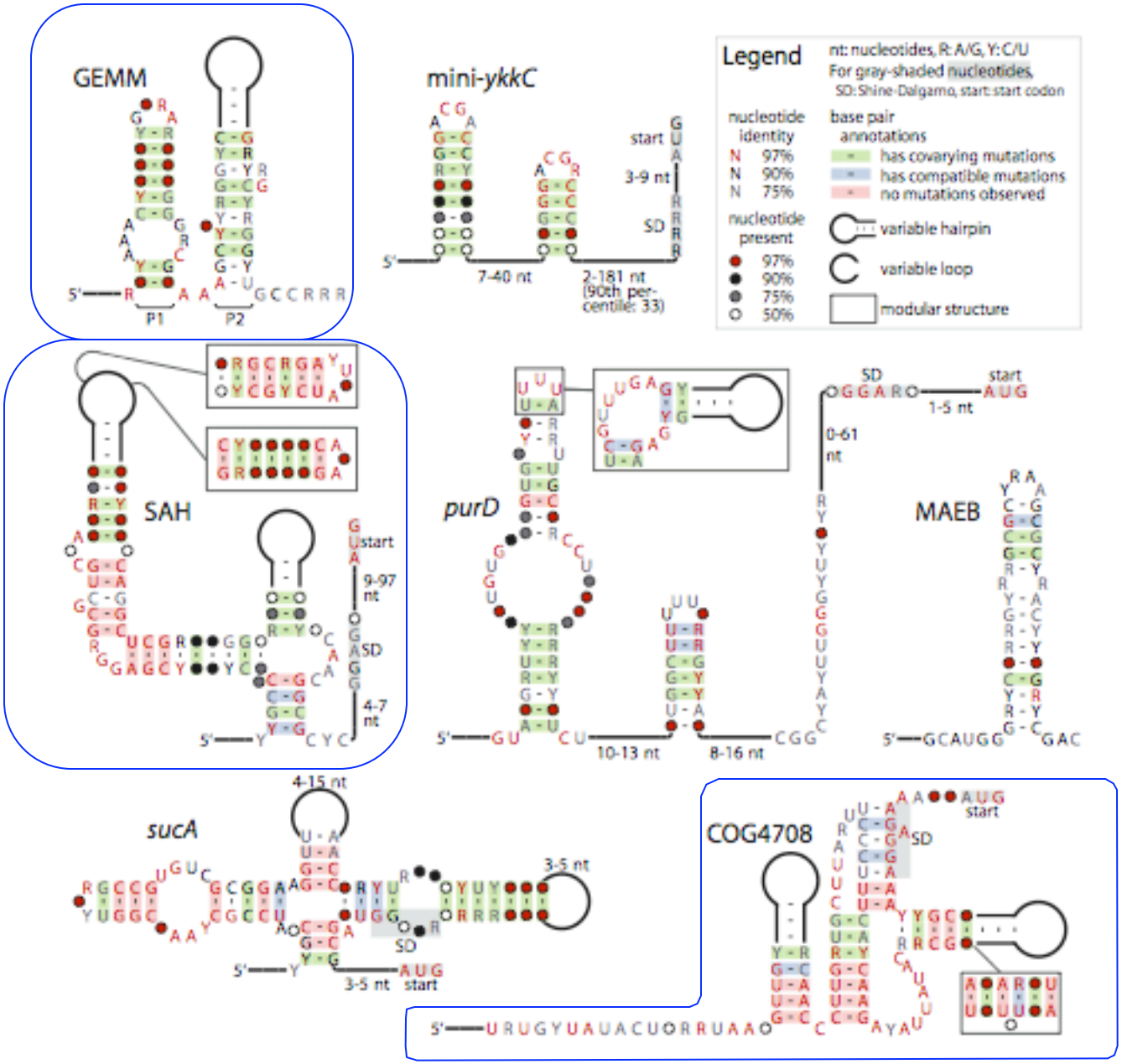
Rfam		Membership			Overlap			Structure		
		#	Sn	Sp	nt	Sn	Sp	bp	Sn	Sp
RF00174	Cobalamin	183	0.74[1]	0.97	152	0.75	0.85	20	0.60	0.77
RF00504	Glycine	92	0.56[1]	0.96	94	0.94	0.68	17	0.84	0.82
RF00234	glmS	34	0.92	1.00	100	0.54	1.00	27	0.96	0.97
RF00168	Lysine	80	0.82	0.98	111	0.61	0.68	26	0.76	0.87
RF00167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

Application II

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

Weinberg, Barrick, Yao, Roth, Kim, Gore, Wang, Lee, Block, Sudarsan, Neph, Tompa, Ruzzo and Breaker.
Nucl. Acids Res., July 2007 35: 4809-4819.



boxed = confirmed riboswitch (+2 more)

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

New Riboswitches

(all lab-verified)

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

Application III

ncRNAs in Vertebrates

ncRNA discovery in Vertebrates

Natural approach : Align, Fold, Score

Previous studies focus on highly conserved regions (Washietl, Pedersen et al. 2007)

- Evofold (Pedersen *et al.* 2006)
- RNAz (Washietl *et al.* 2005)

} Thousands of candidates

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

} Thousands more

Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions

Torarinsson, Yao, Wiklund, Bramsen, Hansen,

Kjems, Tommerup, Ruzzo and Gorodkin

Genome Research, Feb 2008, 18(2):242-251

PMID: 18096747

Search in Vertebrates

Extract ENCODE Multiz alignments

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

Apply CMfinder to both strands.

10,106 predictions, 6,587 clusters.

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

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

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

Overlap w/ Indel Purified Segments

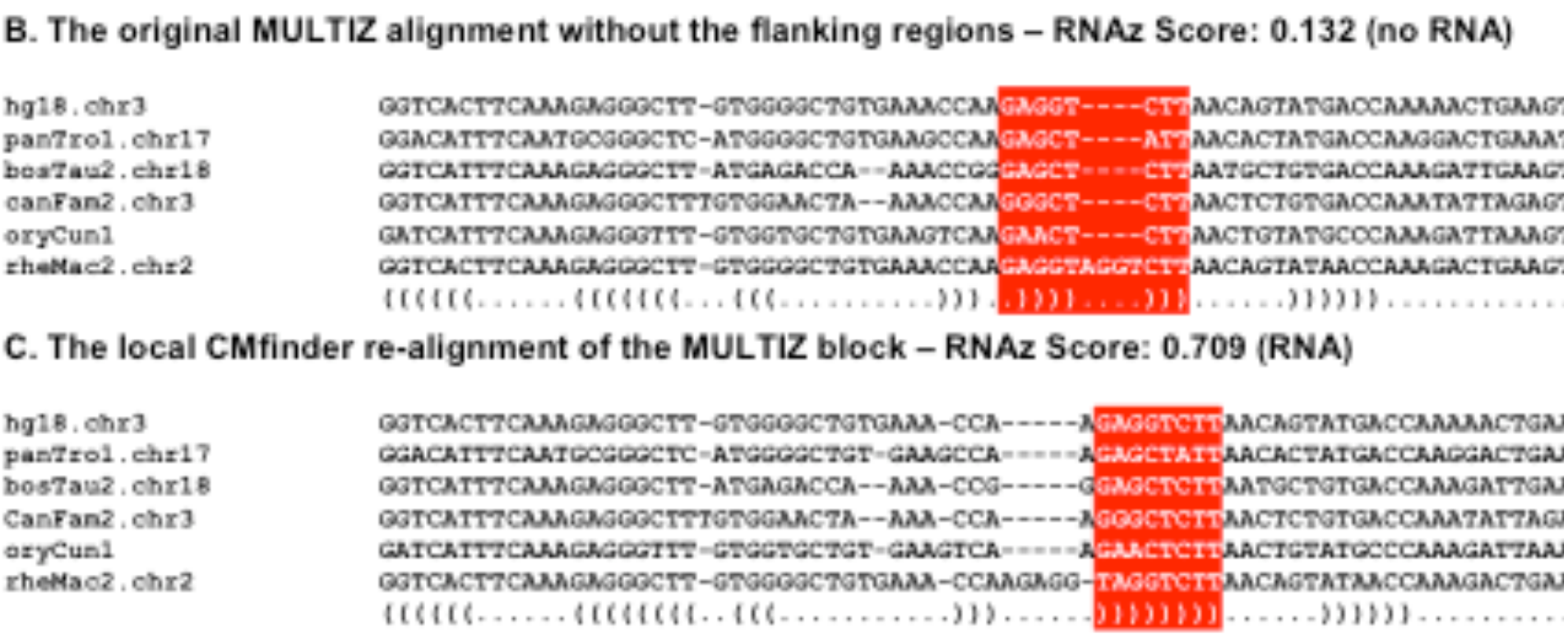
IPS presumed to signal purifying selection

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

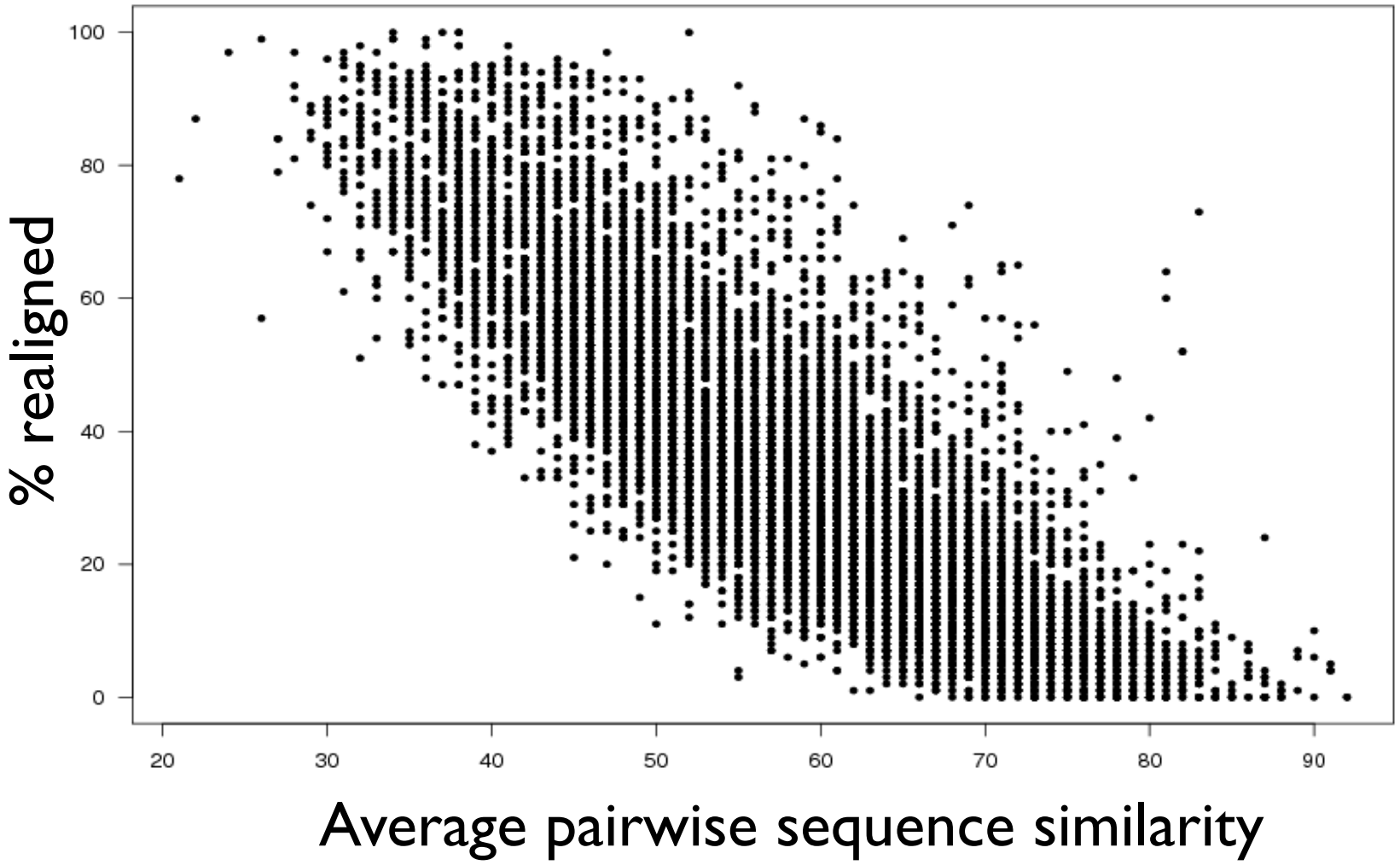
Strong P-value for their overlap w/ IPS

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%

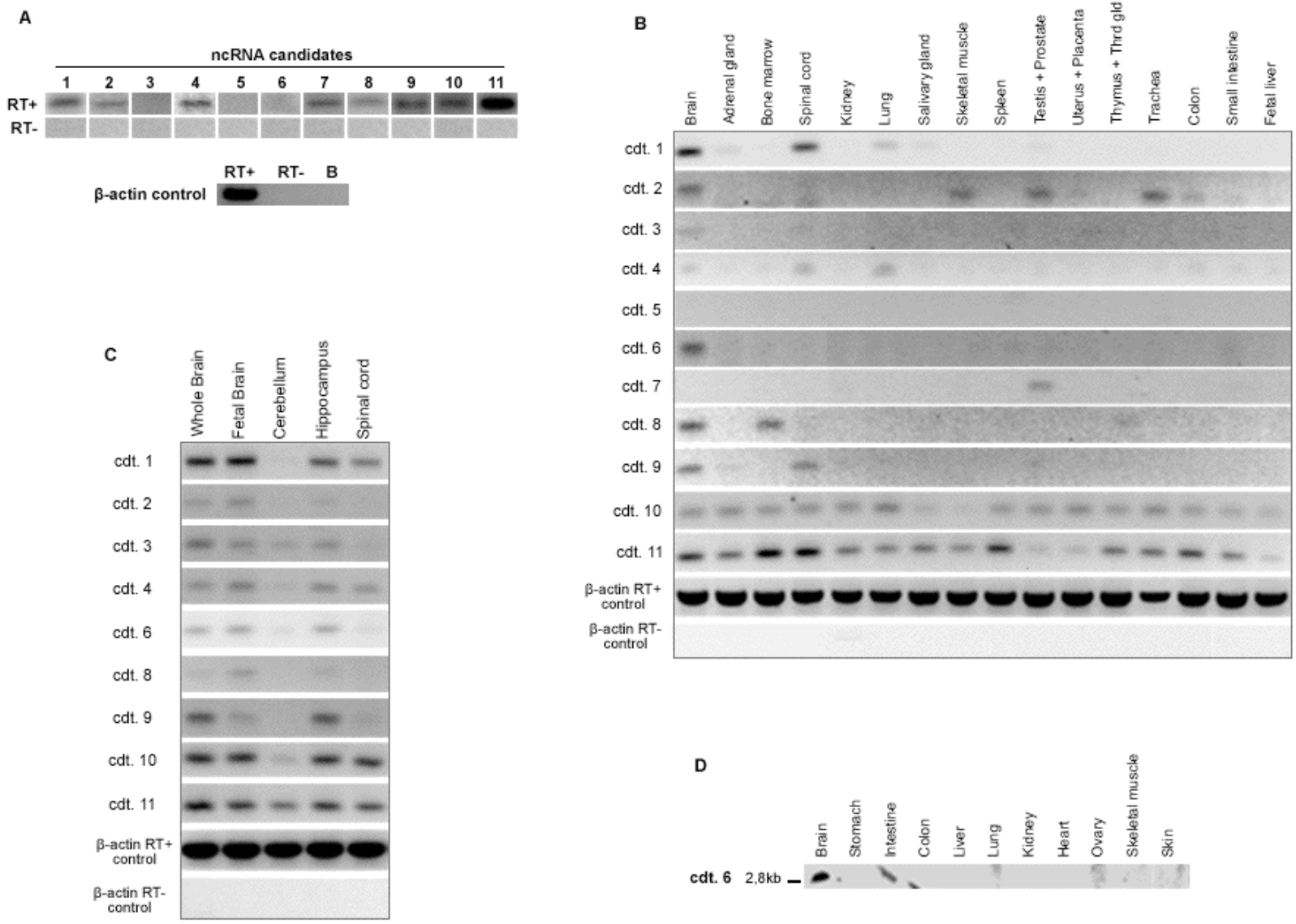
Alignment Matters



Realignment



10 of 11 top (differentially) expressed



Open Problems - Better CM's

Optional- and variable-length stems

Riboswitches & other regulatory RNAs often switch between conformations; better search & alignment exploiting both alternatives?

"Augmented" CM handling pseudoknots probably too slow for scan, but plausibly could be used for alignment

Better use of prior knowledge? (GNRA tetraloops, single-stranded A's, structure motifs, …)

Open Problems - Better algorithms & scoring

- incorporating phylogeny in model construction & scoring
  - e.g. “mutual information” ignores it
- improve scoring by “shuffling”
- other ideas for scan filtering
- comparing & clustering RNA structures
- search/alignment/inference with splicing

Open Problems - Applications & Biology

clustering intergenic sequences, esp prokaryotic

systematic look at eukaryotic UTRs

- how to cluster? how to score?

"swiss-cheese phylogenies"

evidence for selection (no dN/dS)

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

Plenty of room for more!