Some stories about computing and riboswitches

Walter L. (Larry) Ruzzo

Computer Science & Engineering and Genome Sciences, Univ. of Washington
Fred Hutchinson Cancer Research Center
Seattle, USA

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

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."

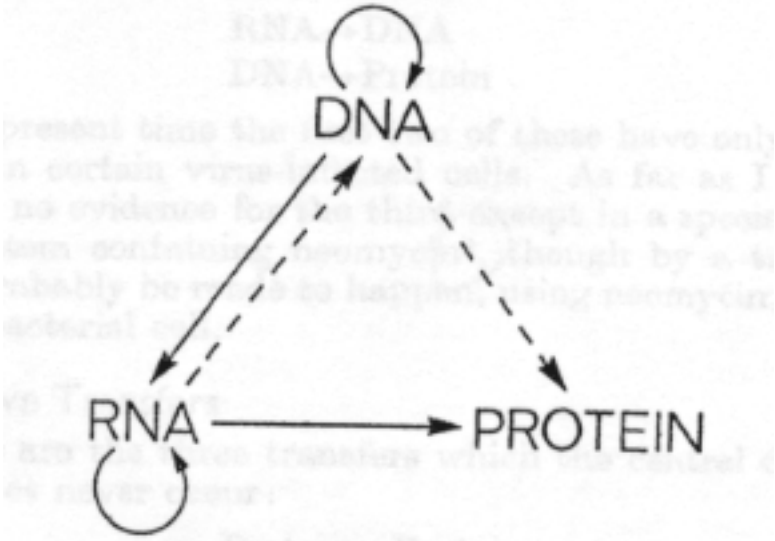
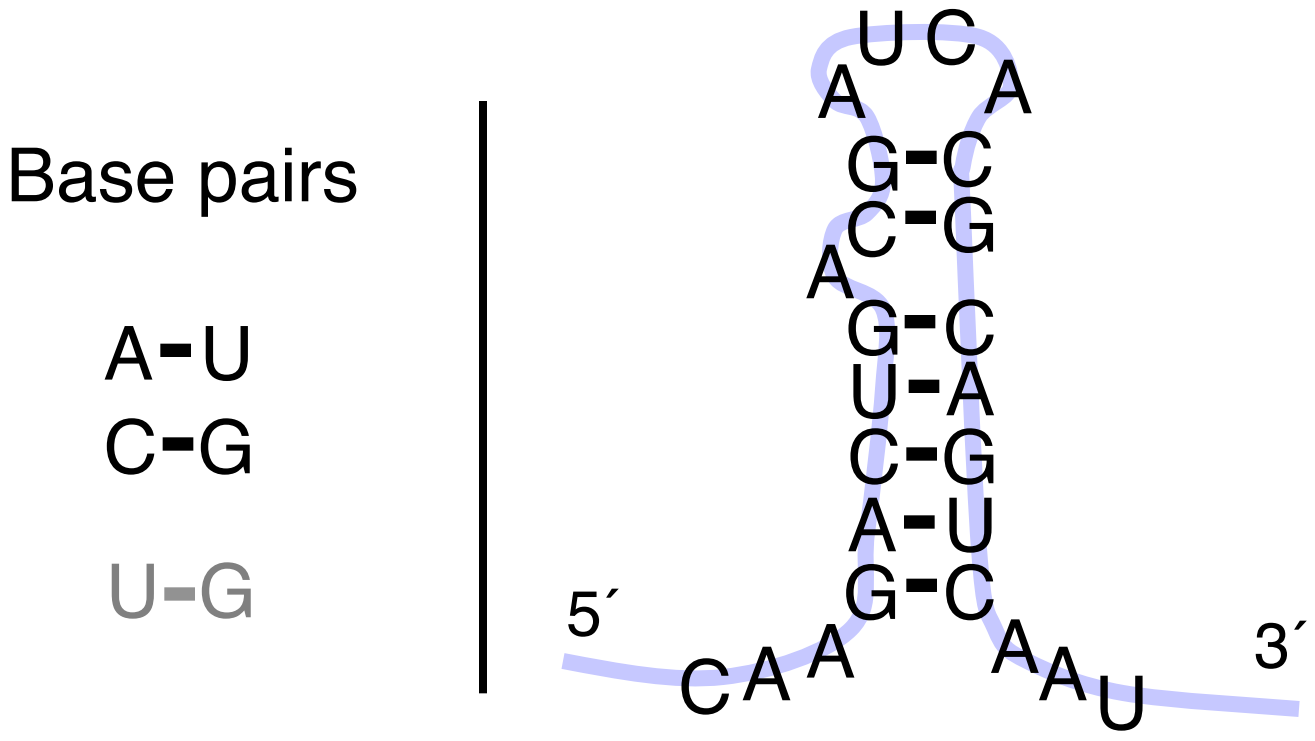


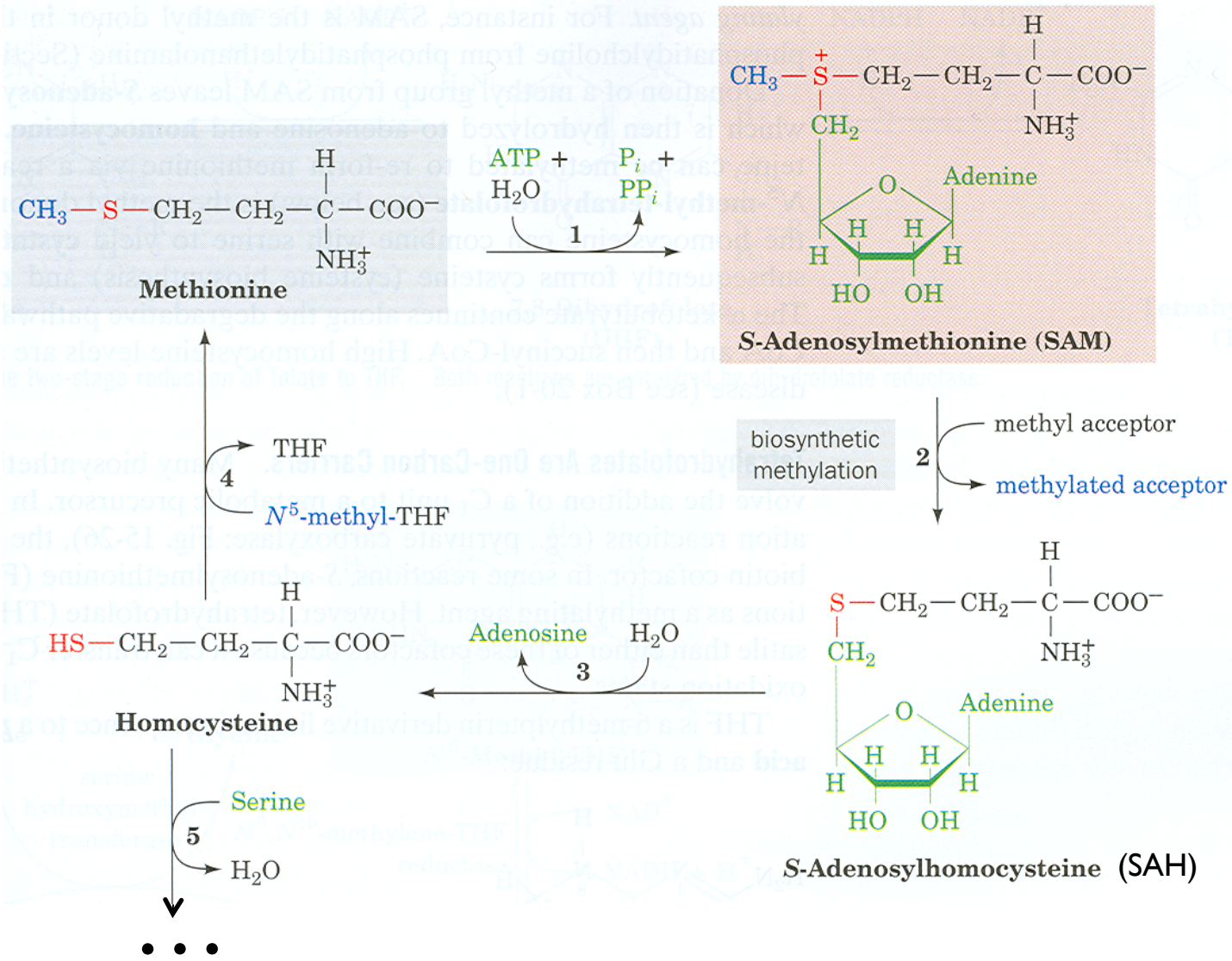
Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

RNA Secondary Structure:
RNA makes helices too

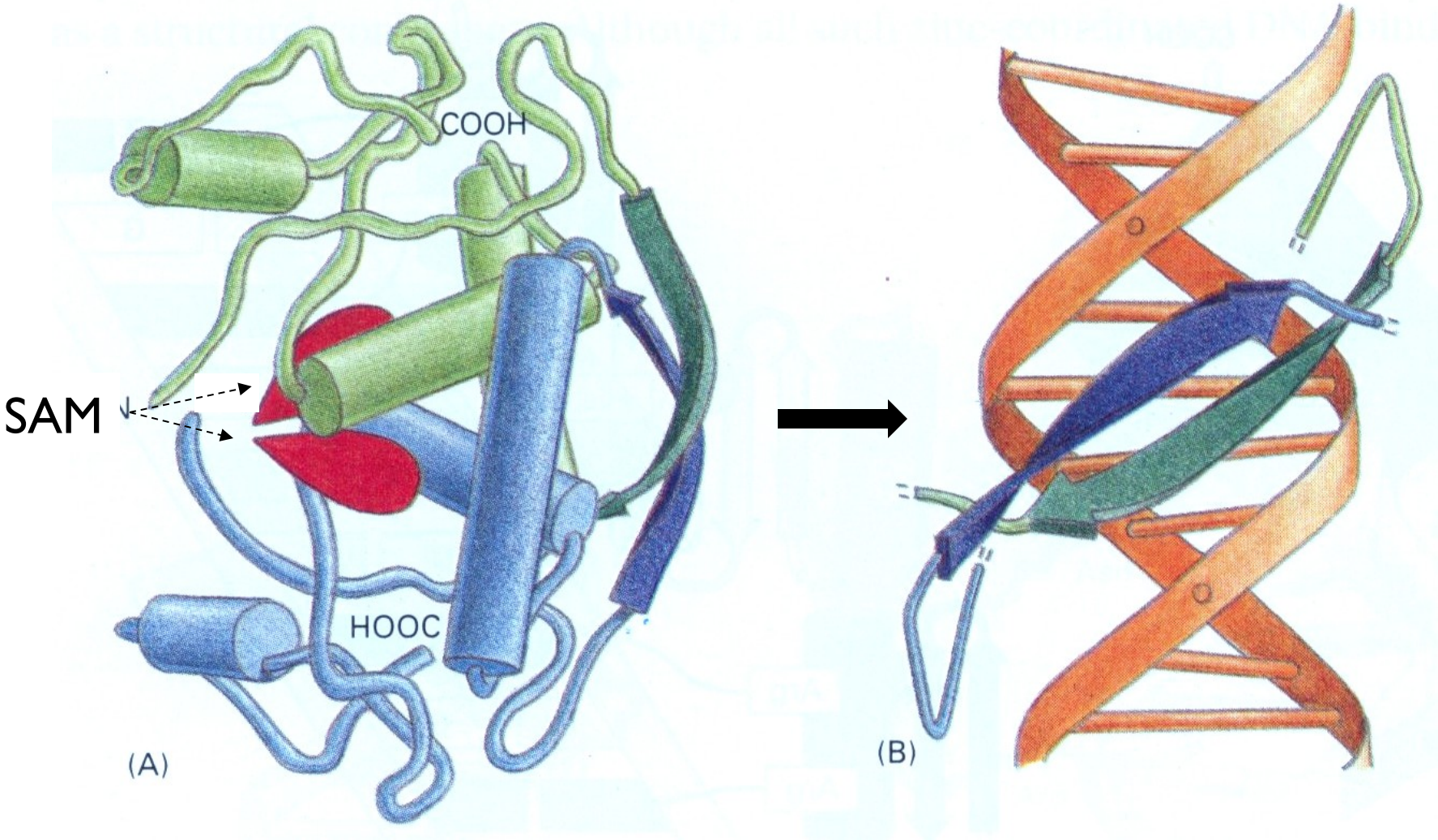


Usually *single* stranded

Met Pathways



Gene Regulation: The MET Repressor



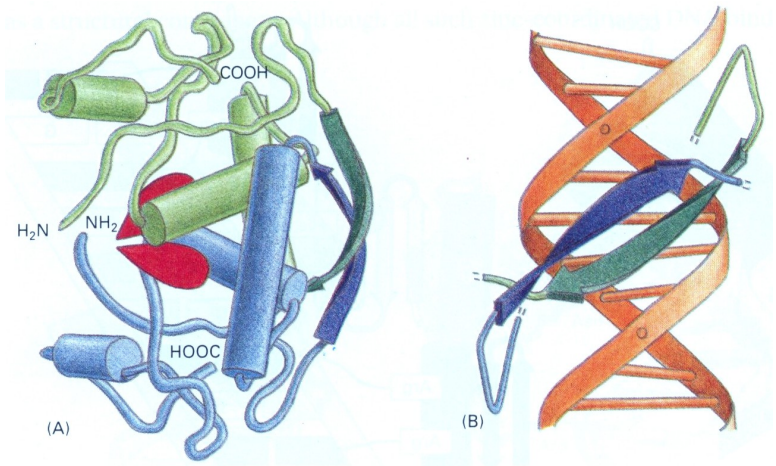
Protein

Alberts, et al, 3e.

DNA

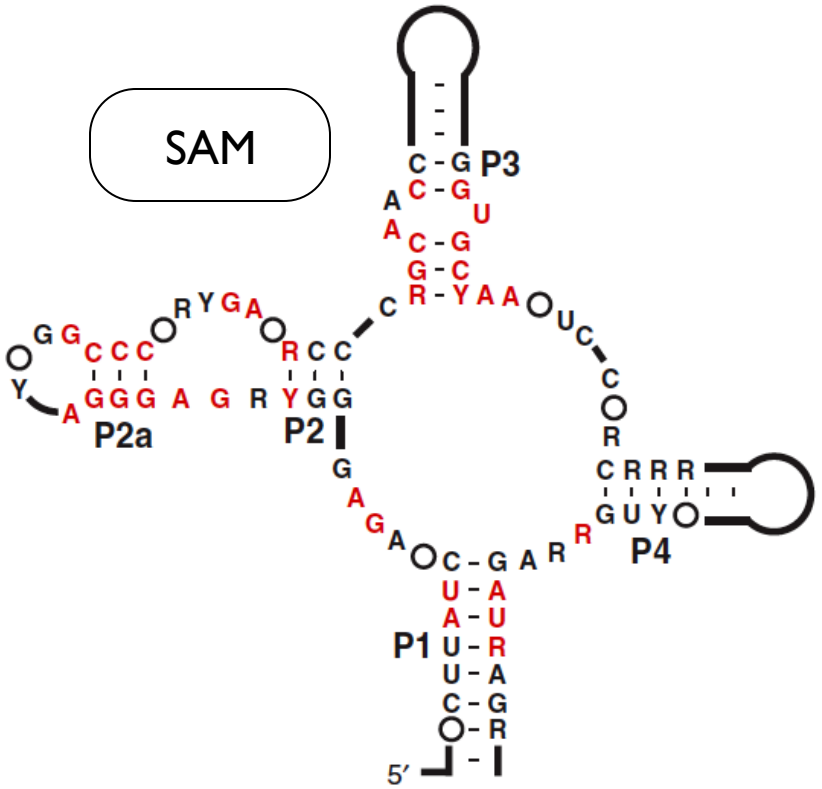
Not the only way!

Alberts, et al, 3e.



Protein way

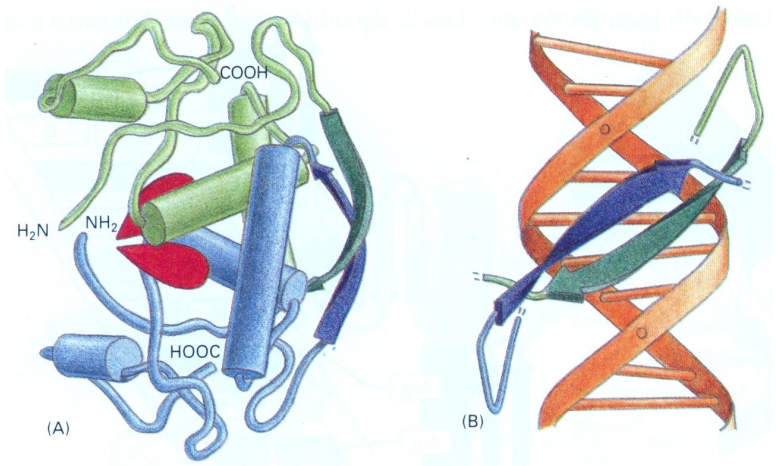
Riboswitch alternative



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

Not the only way!

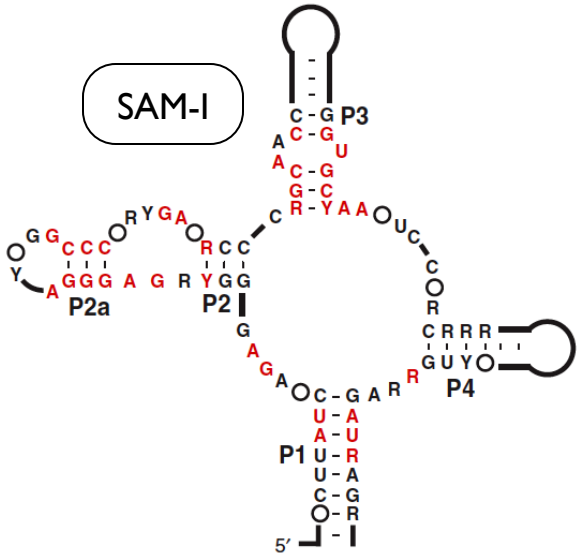
Alberts, et al, 3e.



Protein way

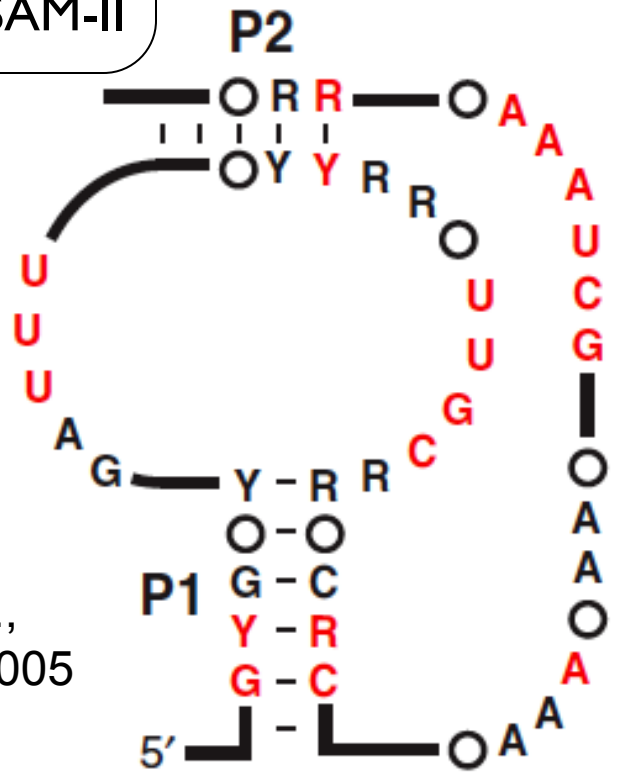
Riboswitch alternatives

SAM-I



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

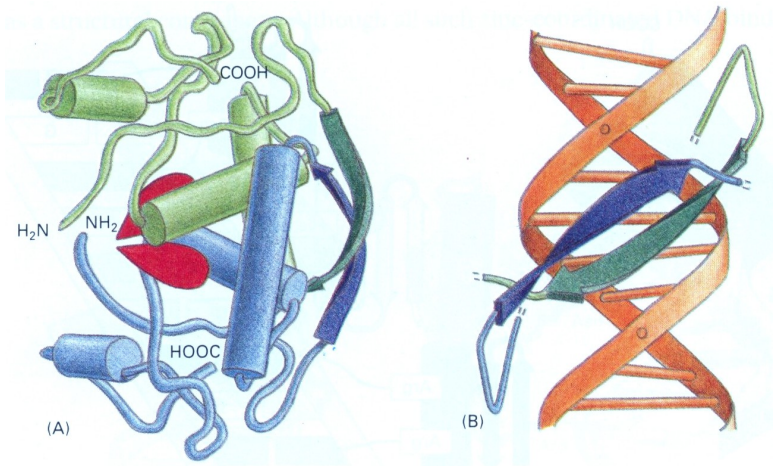
SAM-II



Corbino et al., Genome Biol. 2005

Not the only way!

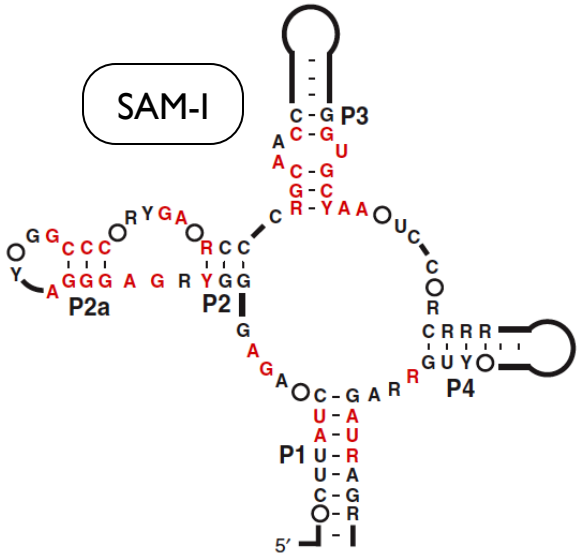
Alberts, et al, 3e.



Protein way

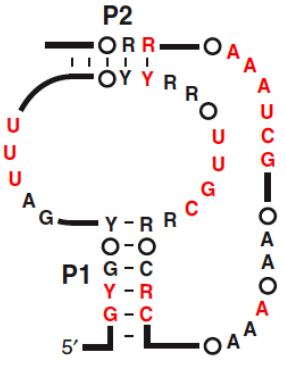
Riboswitch alternatives

SAM-I



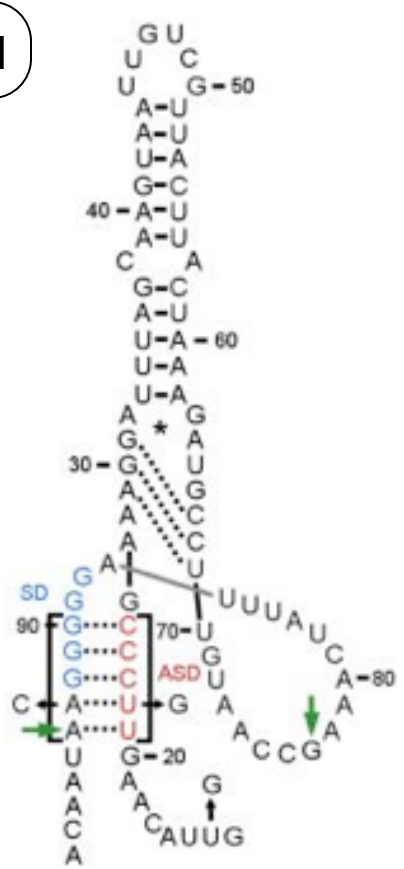
Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



Corbino et al., Genome Biol. 2005

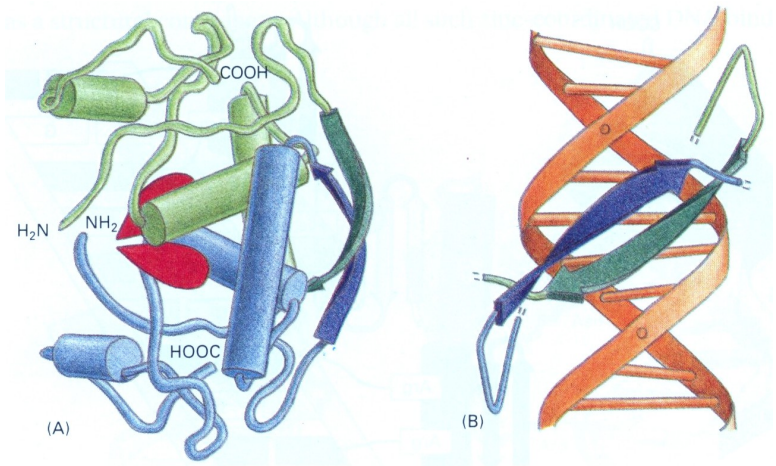
SAM-III



Fuchs et al., NSMB 2006

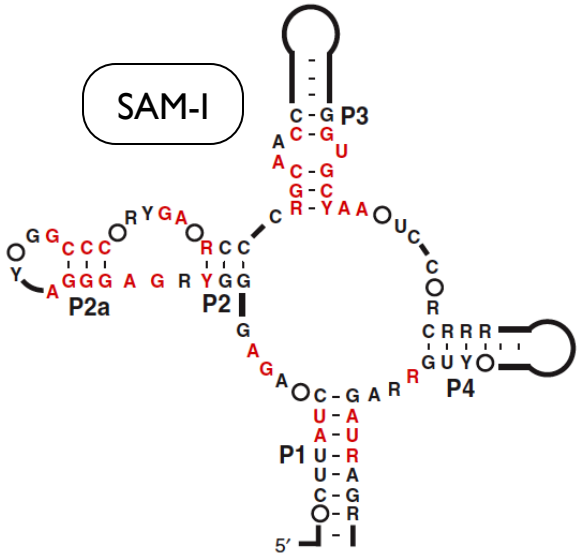
Not the only way!

Alberts, et al, 3e.

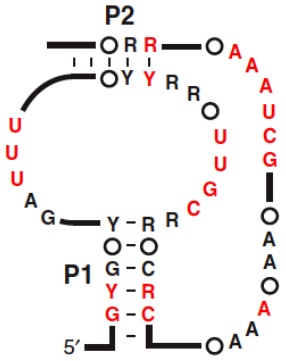


← Protein way

Riboswitch alternatives ↓

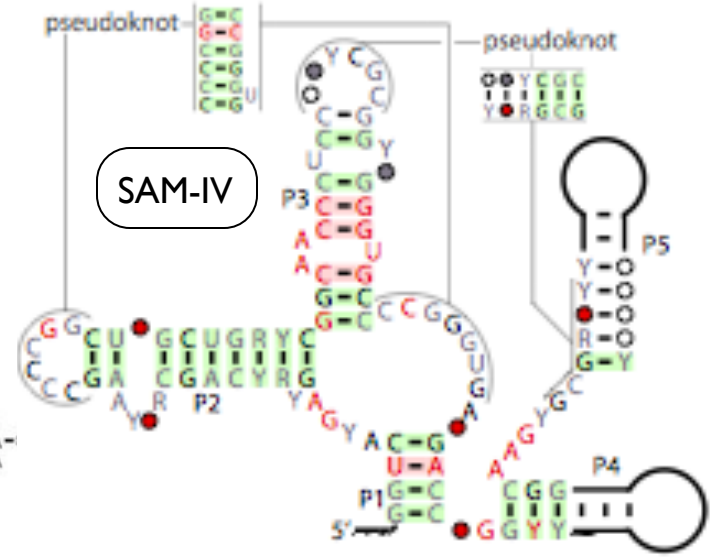


Grundy, Epshtein, Winkler et al., 1998, 2003



Corbino et al., Genome Biol. 2005

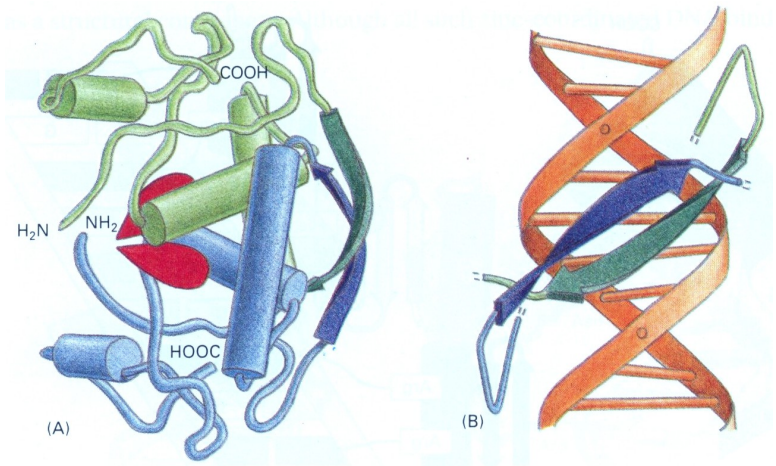
Fuchs et al., NSMB 2006



Weinberg et al., RNA 2008

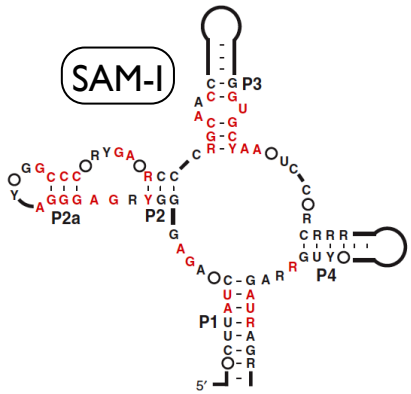
Not the only way!

Alberts, et al, 3e.

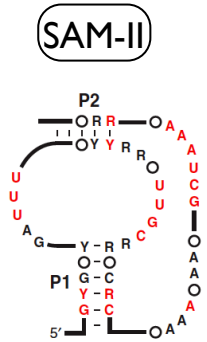


← Protein way

Riboswitch alternatives ↓

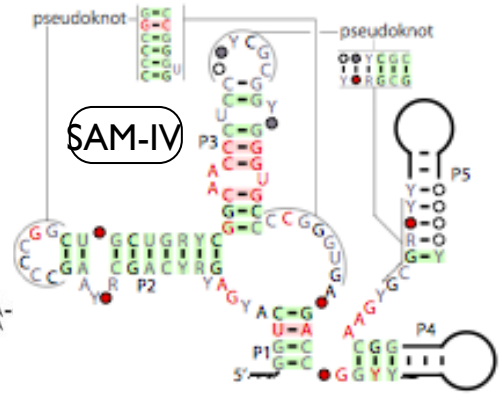


Grundy, Epshtein, Winkler et al., 1998, 2003

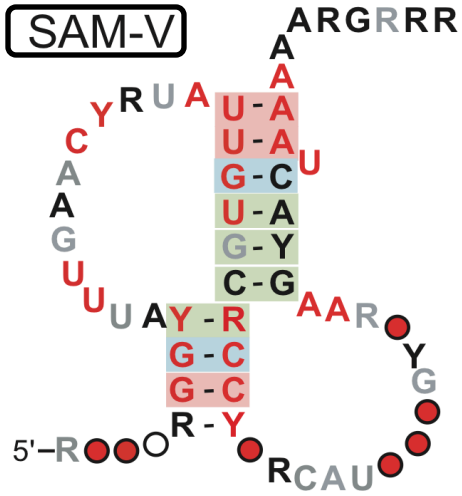


Corbino et al., Genome Biol. 2005

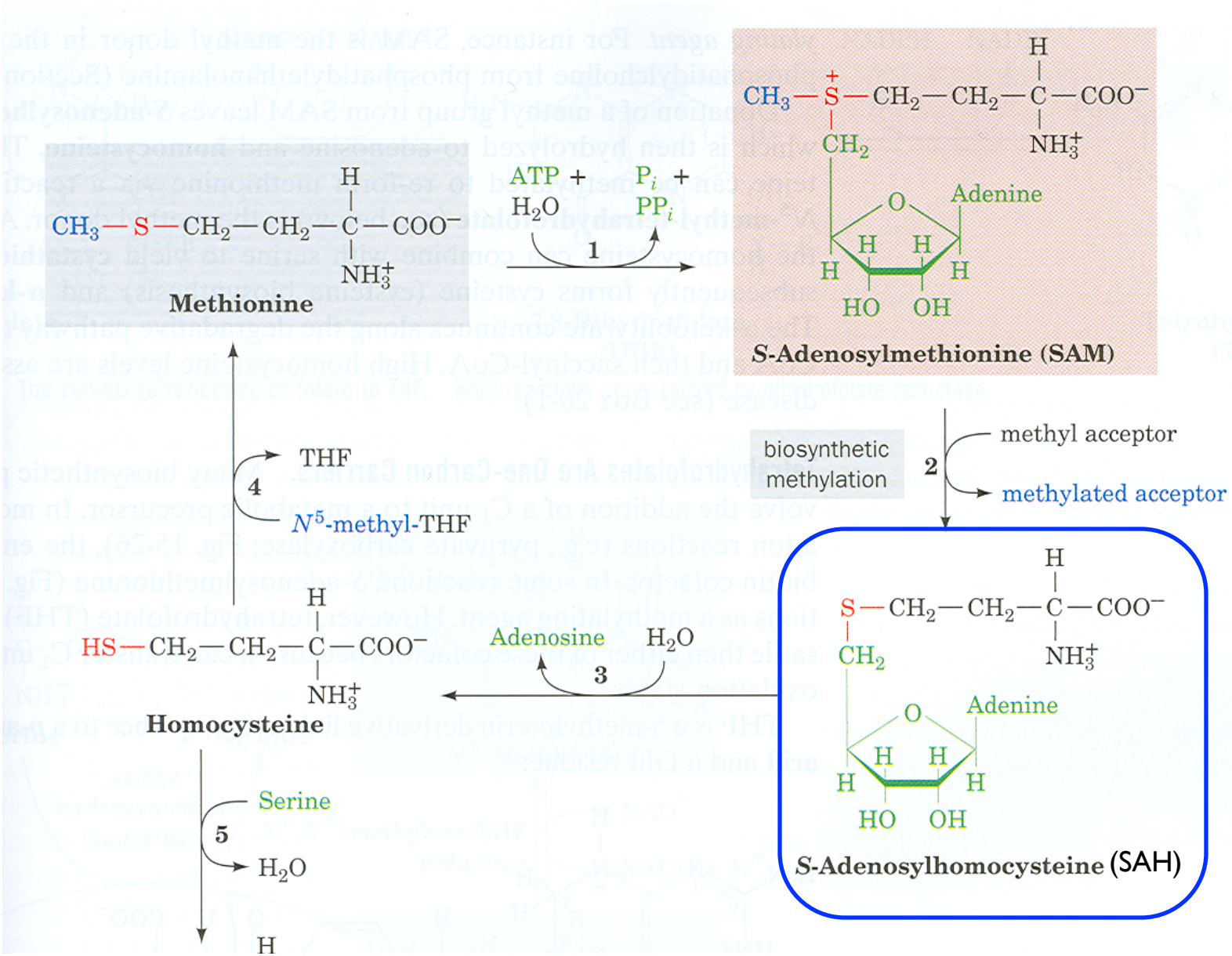
Fuchs et al., NSMB 2006

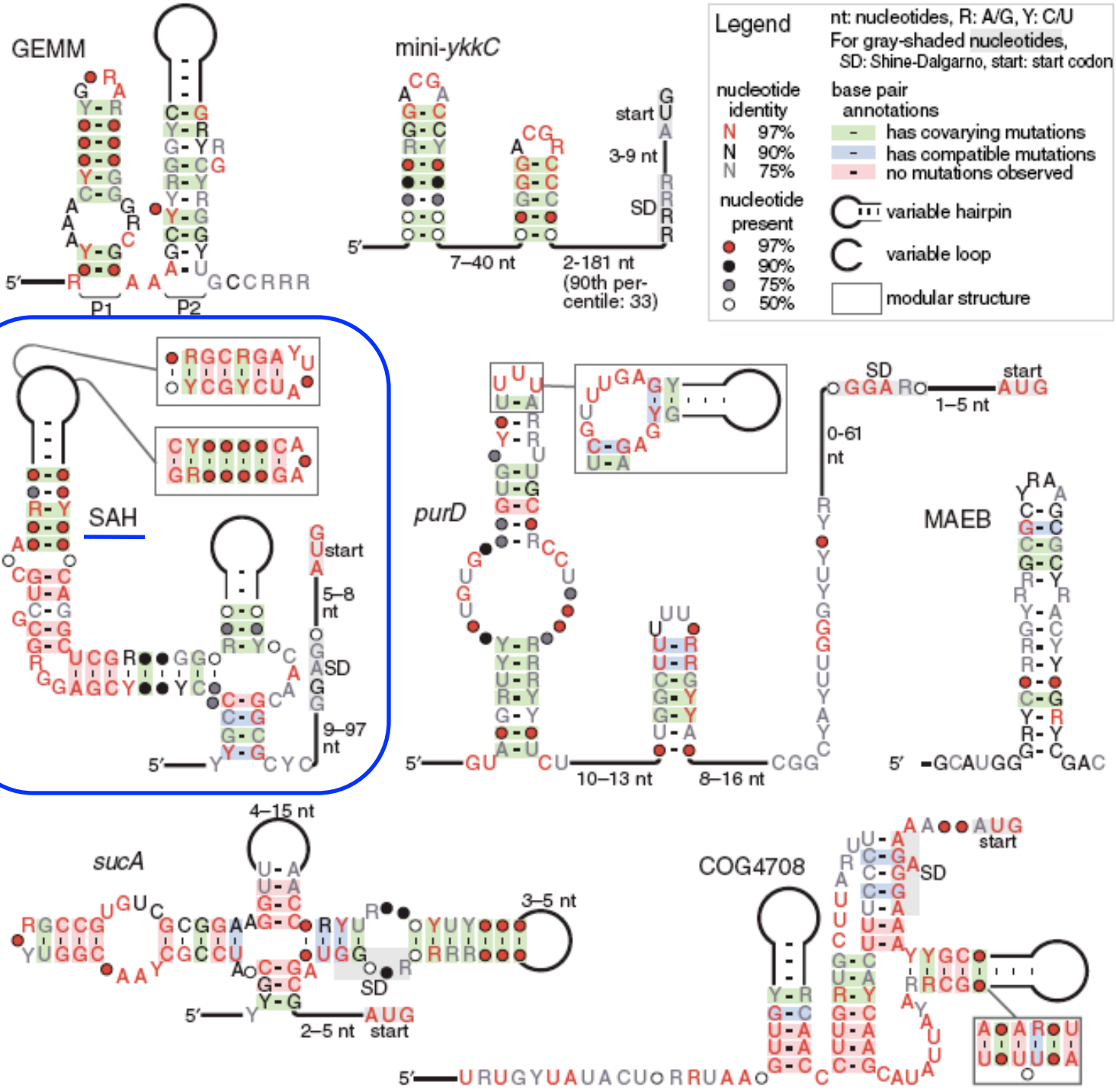


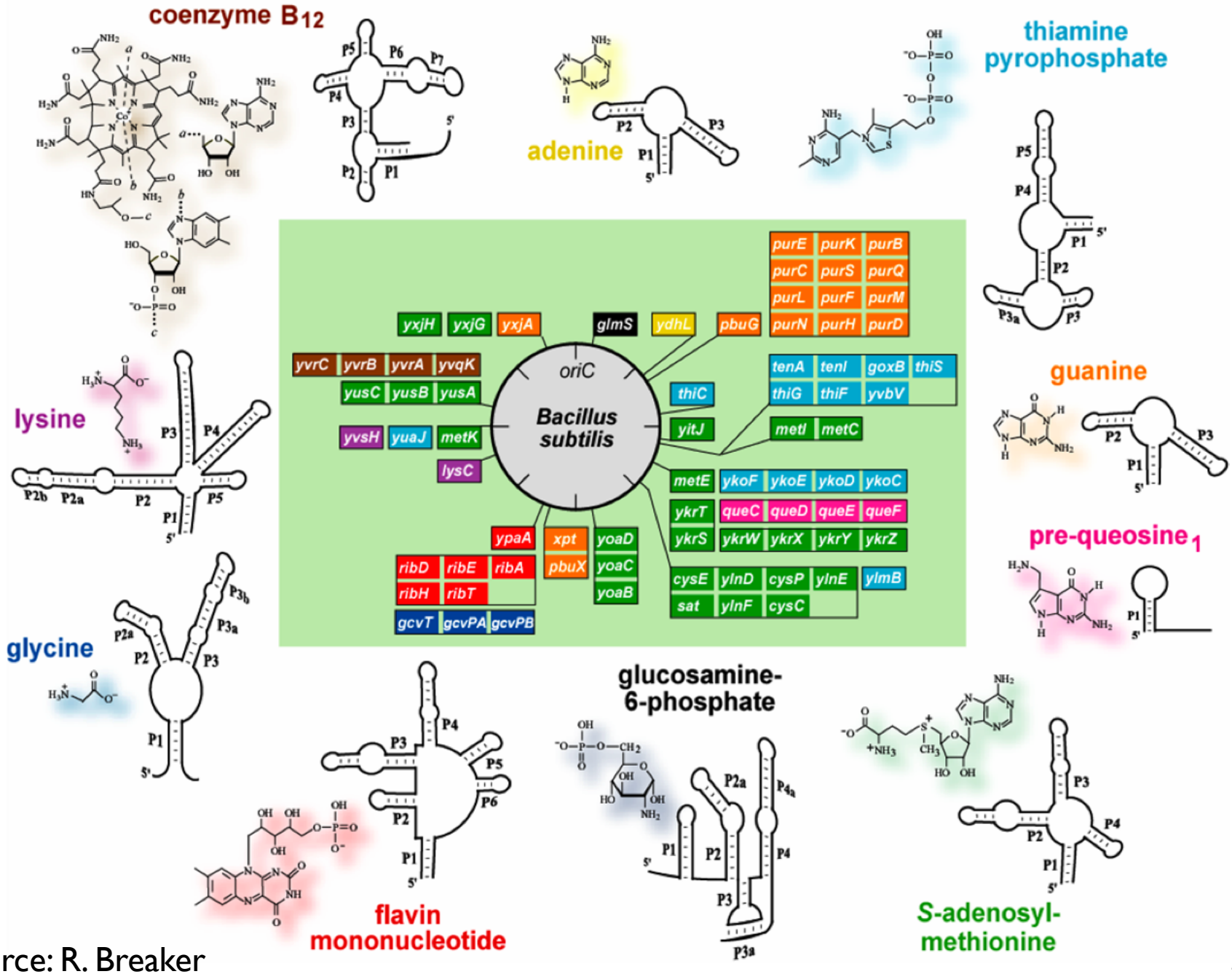
Weinberg et al., RNA 2008



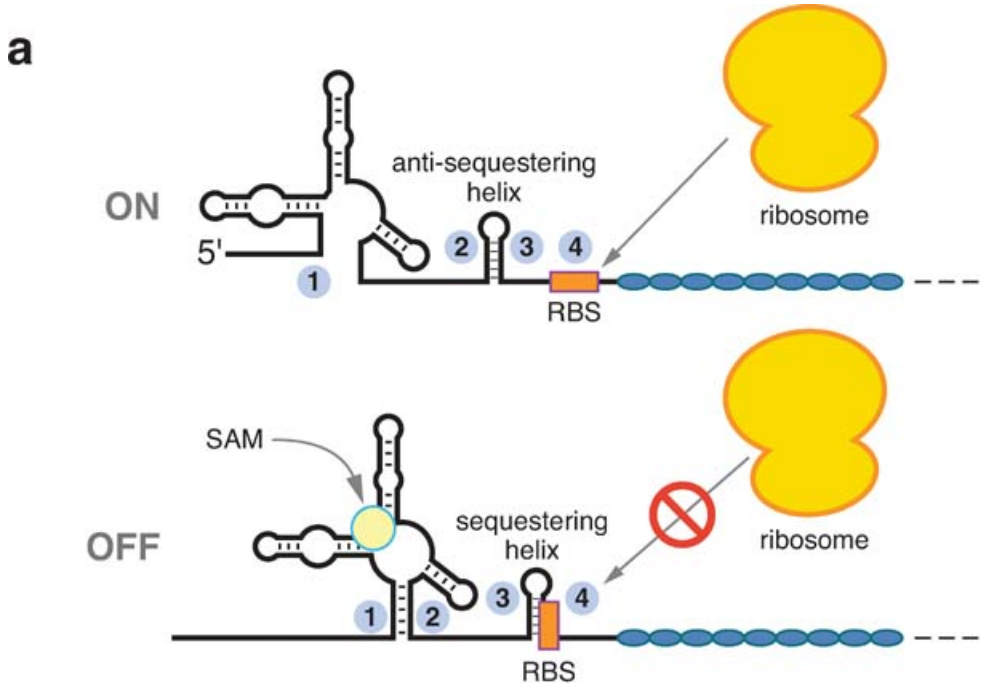
Meyer, etal., BMC Genomics 2009


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

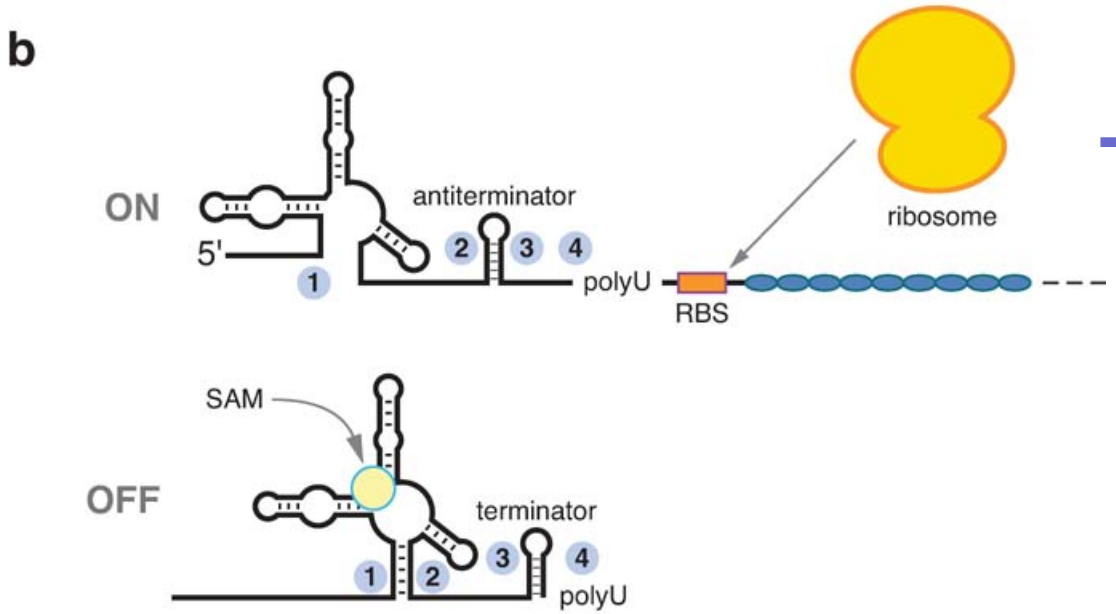

GEMM
P1
P2
mini-ykkC
start
3-9 nt
SD
7–40 nt
2-181 nt
(90th per-
centile: 33)
Legend
nt: nucleotides, R: A/G, Y: C/U
For gray-shaded nucleotides,
SD: Shine-Dalgarno, start: start codon
nucleotide identity
N 97%
N 90%
N 75%
base pair annotations
has covarying mutations
has compatible mutations
no mutations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



Source: R. Breaker



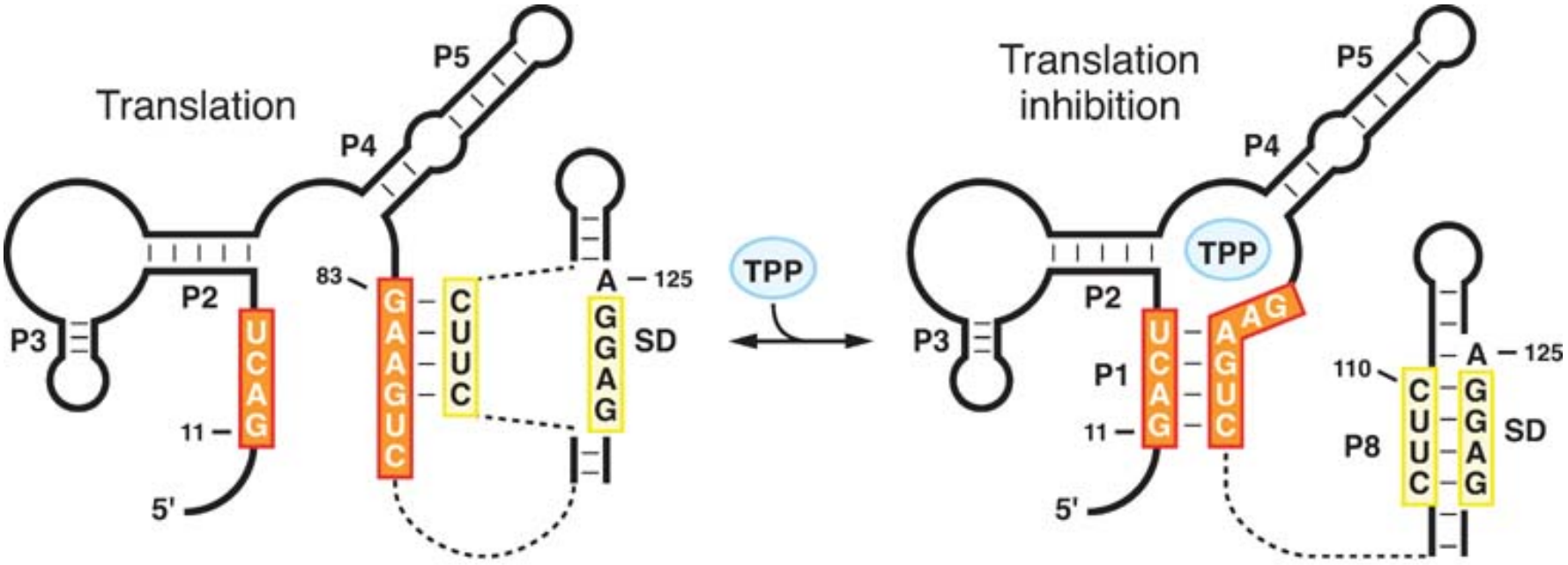
Translational Control



Transcriptional Control

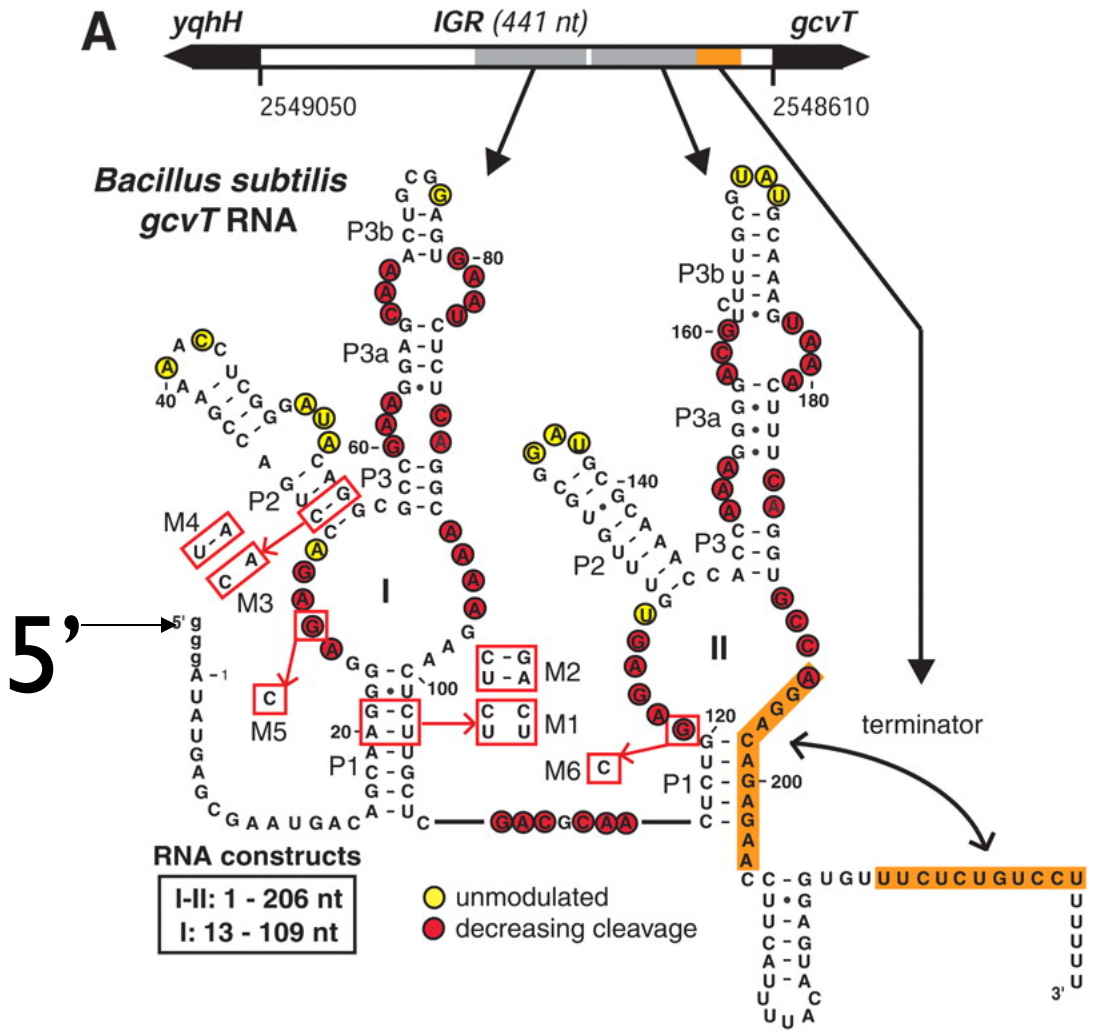
Winkler WC, Breaker RR. 2005.
Annu. Rev. Microbiol. 59:487–517

Detail of Translational Control



Winkler WC, Breaker RR. 2005.
Annu. Rev. Microbiol. 59:487–517

The Glycine Riboswitch



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

5'

gcvT ORF

3'

Cooperativity →
10x sharper switch

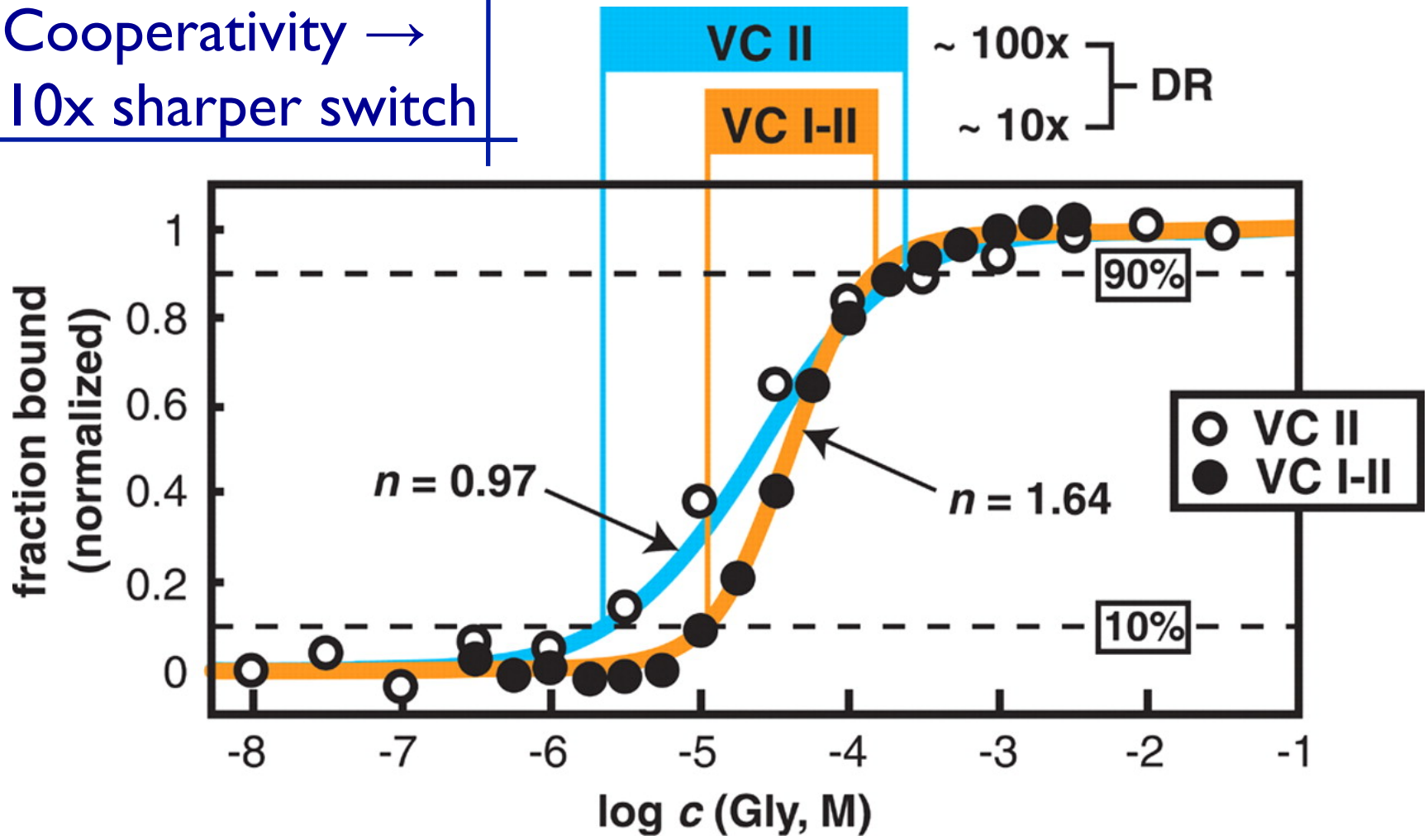


Fig. 3. Cooperative binding of two glycine molecules by the VC I-II RNA. Plot depicts the fraction of VC II (open) and VC I-II (solid) bound to ligand versus the concentration of glycine. The constant, n, is the Hill coefficient for the lines as indicated that best fit the aggregate data from four different 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

UTR structure that directly senses/binds small molecules & regulates mRNA

widespread in prokaryotes

some in eukaryotes & archaea

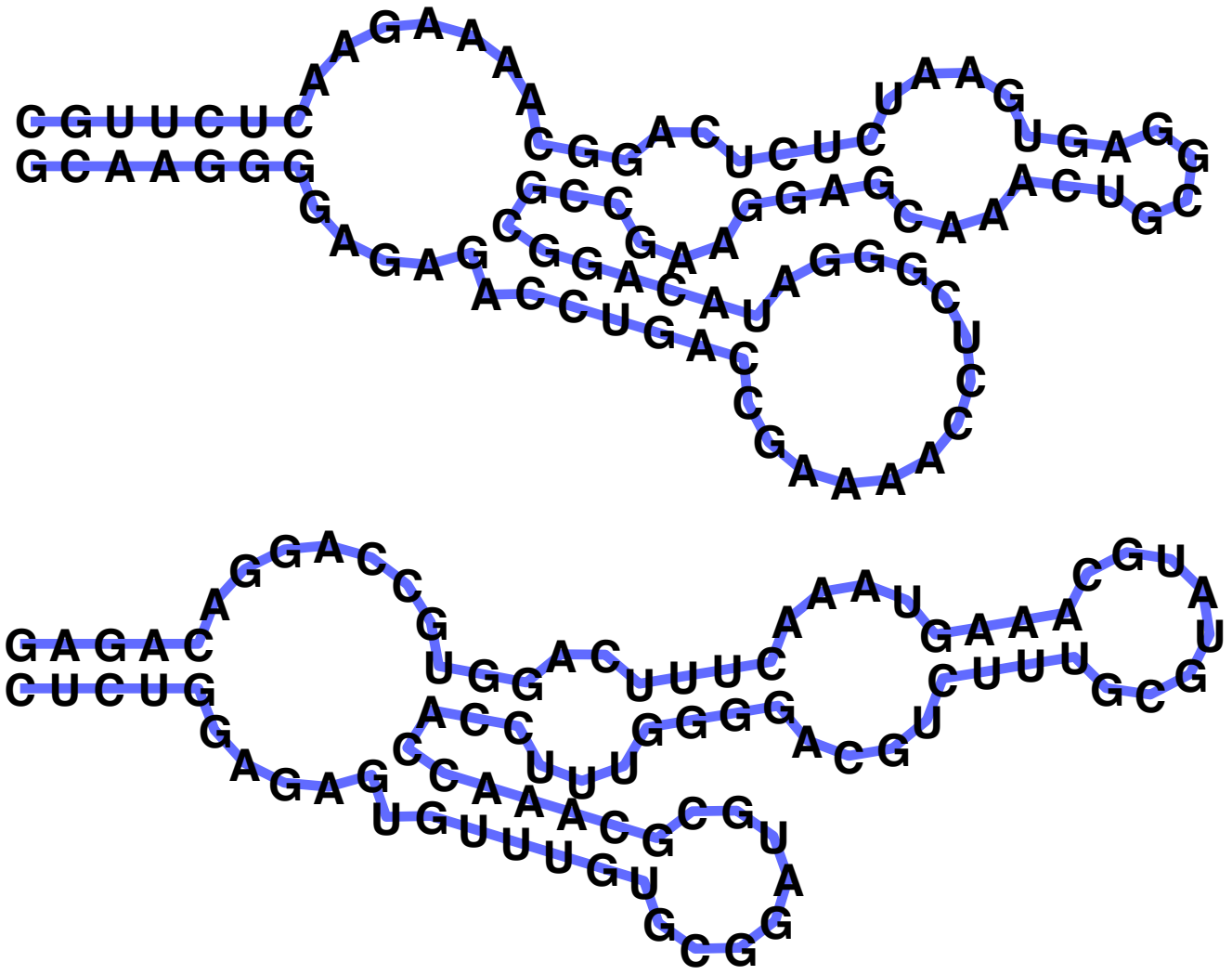
~ 20 ligands known; multiple nonhomologous solutions for some (e.g. SAM)

dozens to hundreds of instances of each

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

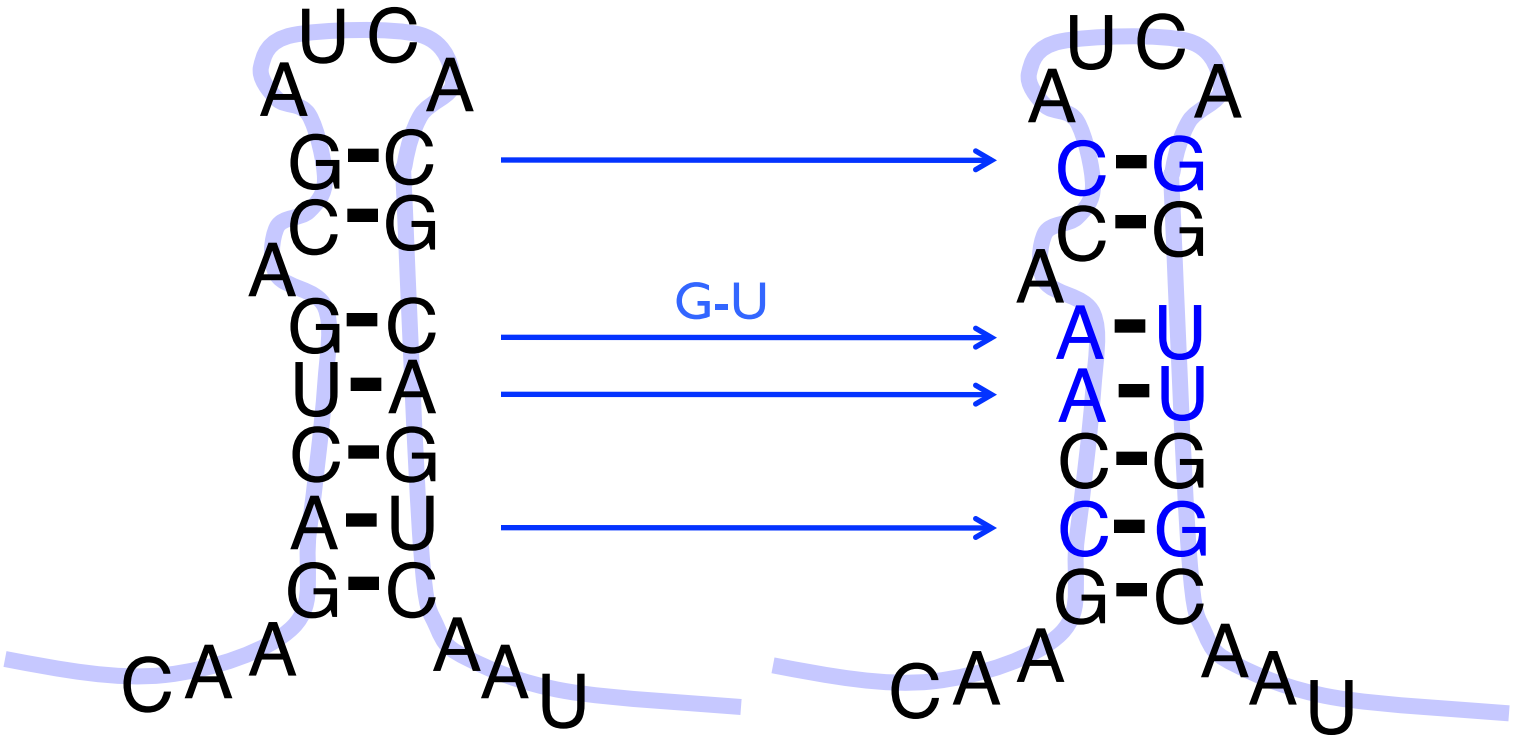
all found since ~2003; most via bioinformatics

Why is RNA hard to deal with?



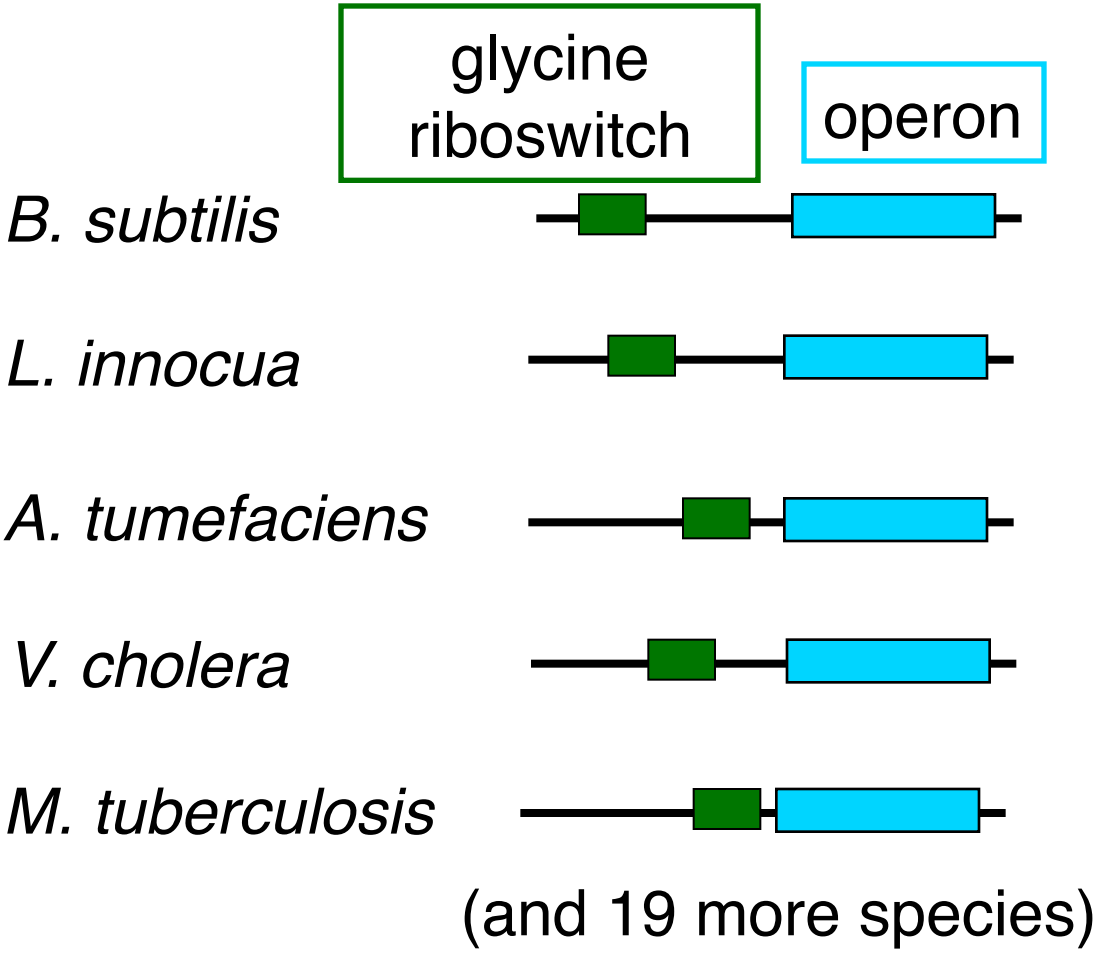
A: *Structure* often more important than *sequence*

RNA Secondary Structure: can be fixed while sequence evolves



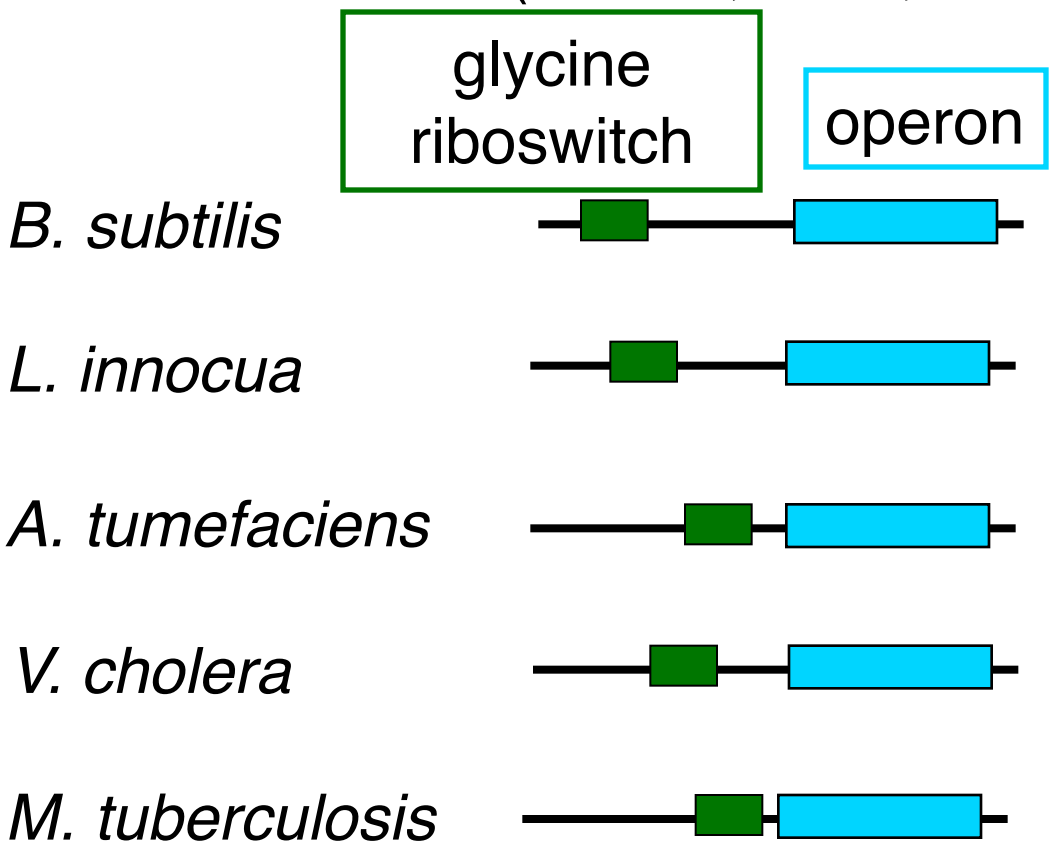
Impact of RNA homology search

(Barrick, *et al.*, 2004)



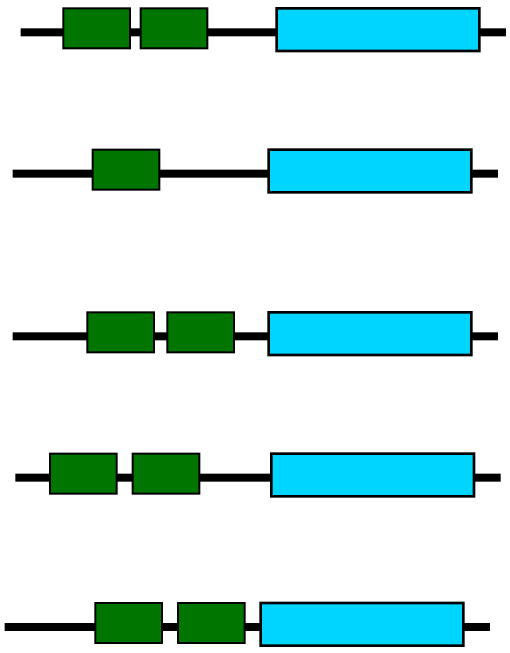
Impact of RNA homology search

(Barrick, *et al.,* 2004)

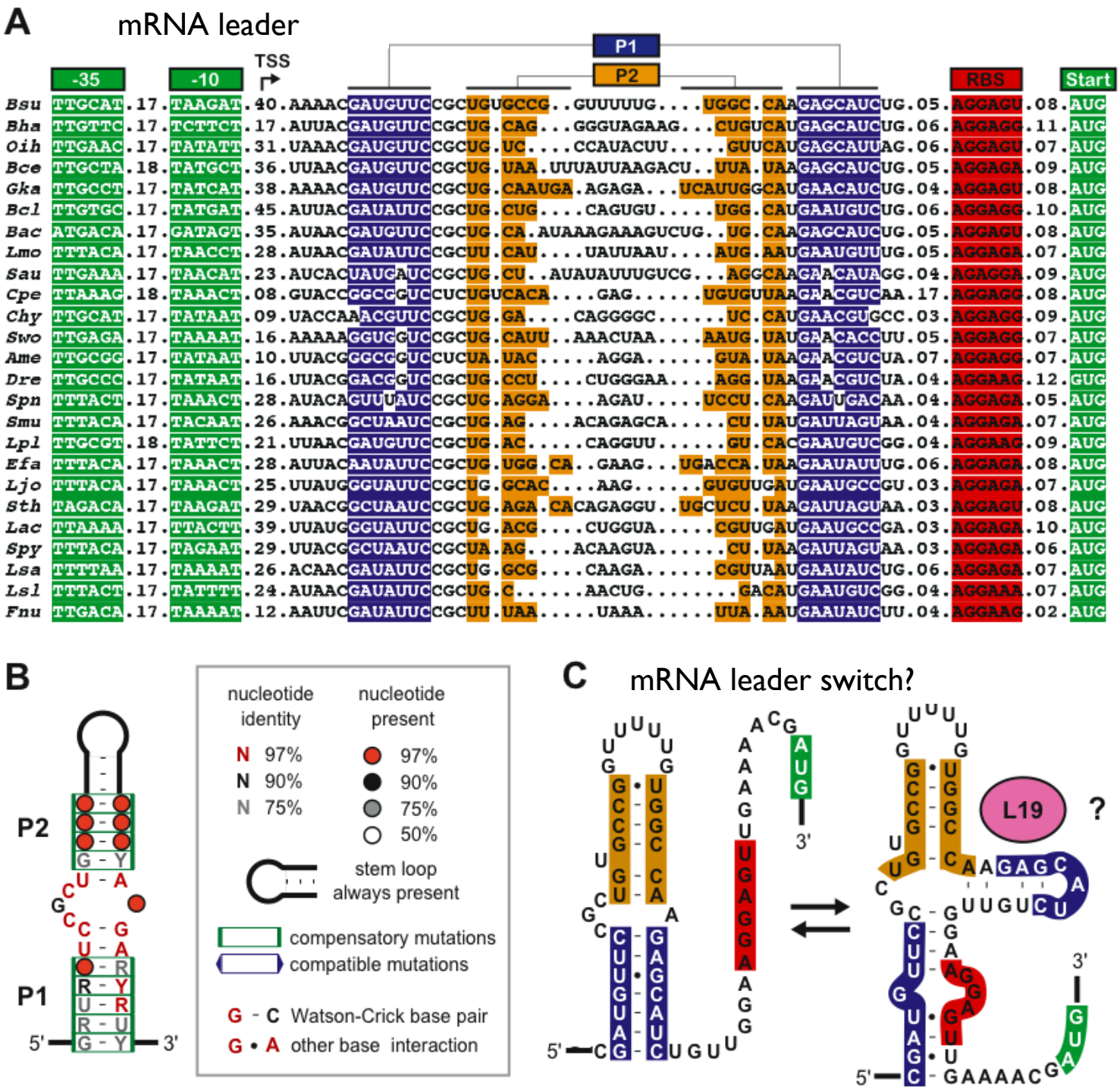


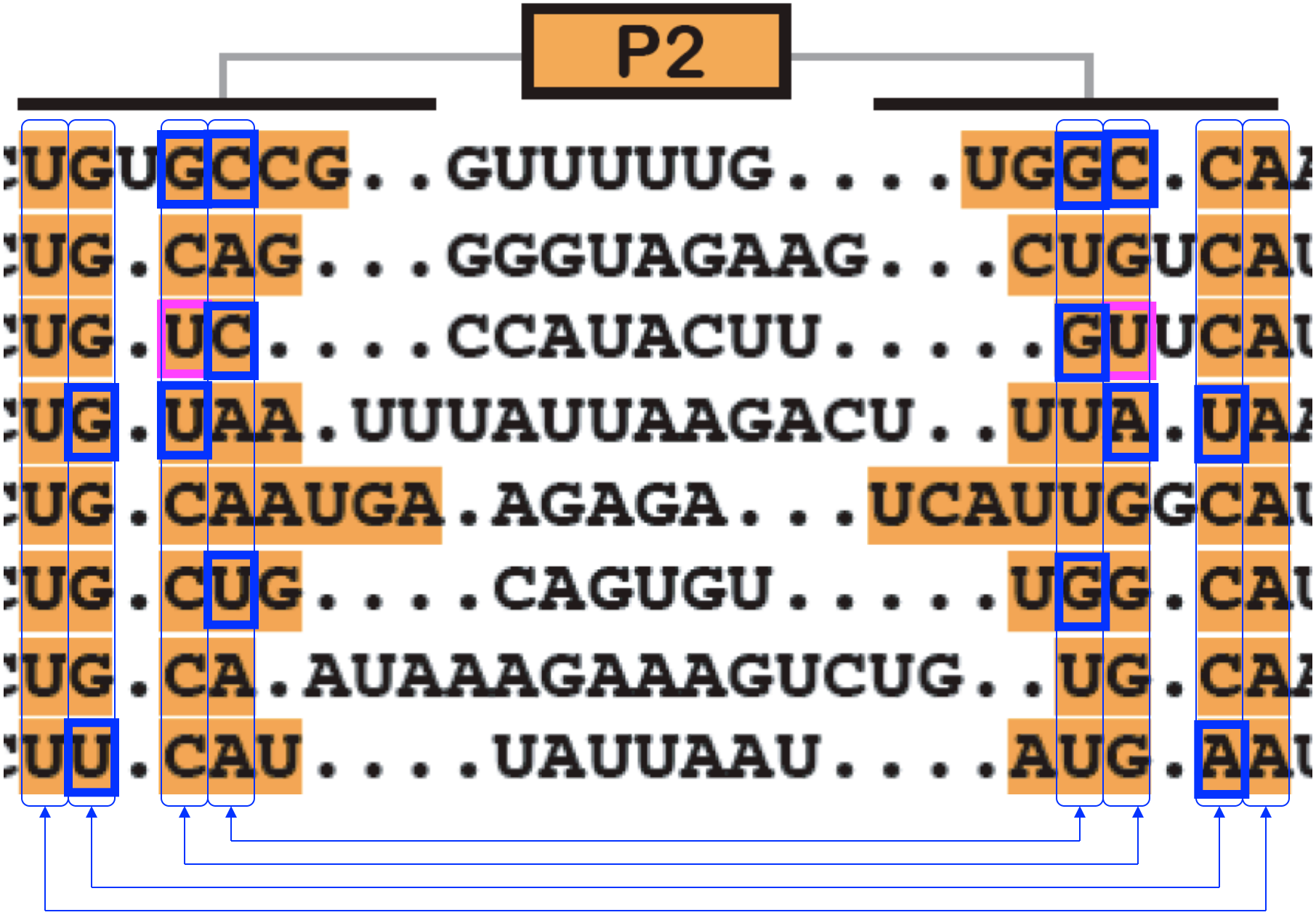
(and 19 more species)

Using our techniques, we found…



(and 42 more species)


A
mRNA leader
P1
P2
-35
-10
TSS
RBS
Start
B
P2
P1
5'
3'
nucleotide identity
nucleotide present
N 97%
N 90%
N 75%
97%
90%
75%
50%
stem loop always present
compensatory mutations
compatible mutations
G - C Watson-Crick base pair
G • A other base interaction
C
mRNA leader switch?
L19
?
5'
3'



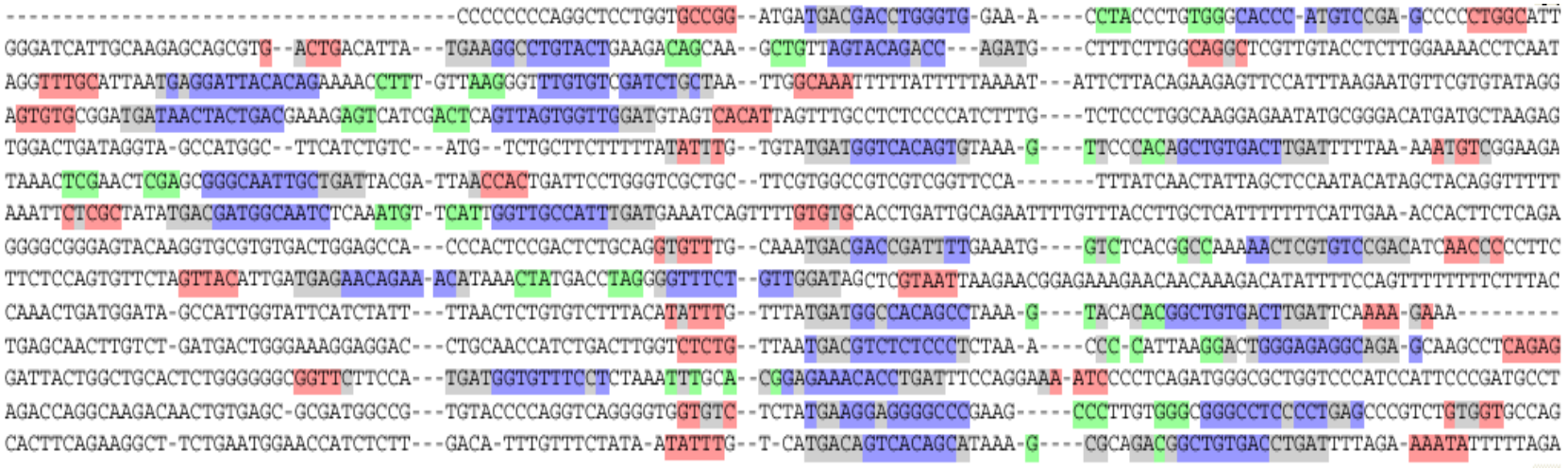
Covariation is strong evidence for base pairing

Alignment Matters

Structural conservation ≠ Sequence conservation

Alignment without structure information is unreliable

CLUSTALW alignment of SECIS elements with flanking regions

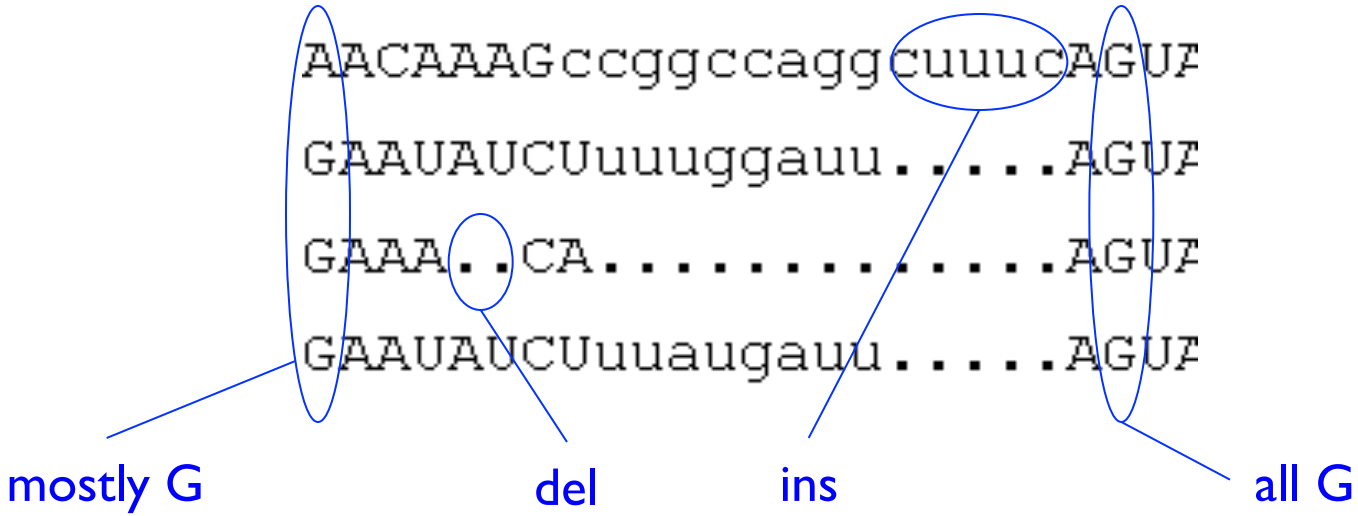


same-colored boxes *should* be aligned

How to model an RNA "Motif"?

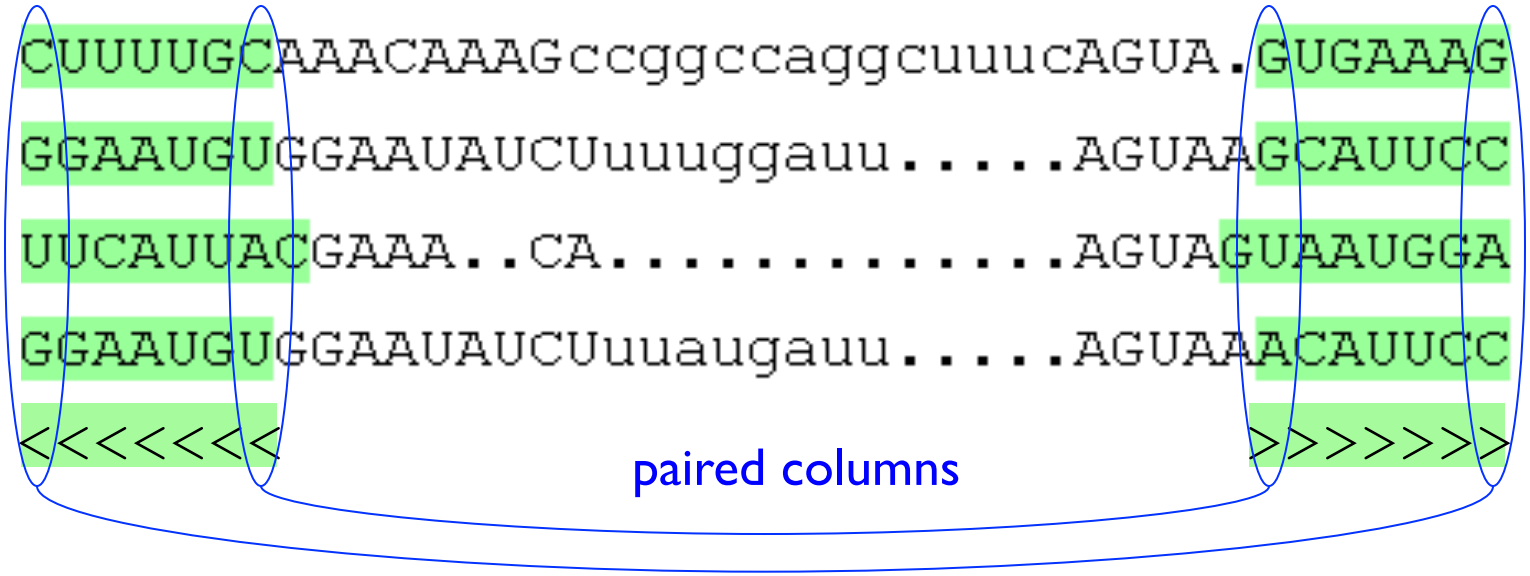
Conceptually, start with a profile HMM:

- 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"?

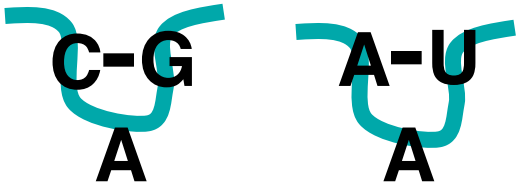
Add "column pairs" and pair emission probabilities for base-paired regions



Covariance Models
(specialized stochastic CFGs)

Sequences

CAG or AAU



CM

$S_1 \rightarrow cS_2g \mid aS_2u$

$S_2 \rightarrow a$

Example parse of CAG

$S_1 \rightarrow cS_2g \rightarrow cag$

Stochastic context-free grammar

CM

P=0.25 P=0.75

$S_1 \rightarrow cS_2g \mid aS_2u$

$S_2 \rightarrow a$

P=1

Example parse of CAG

$S_1 \rightarrow cS_2g \rightarrow cag$

$0.25 \times 1 = 0.25$

Classification

Is Pr(parse of CAG) ≥ threshold

(e.g., vs Pr(CAG in null model)

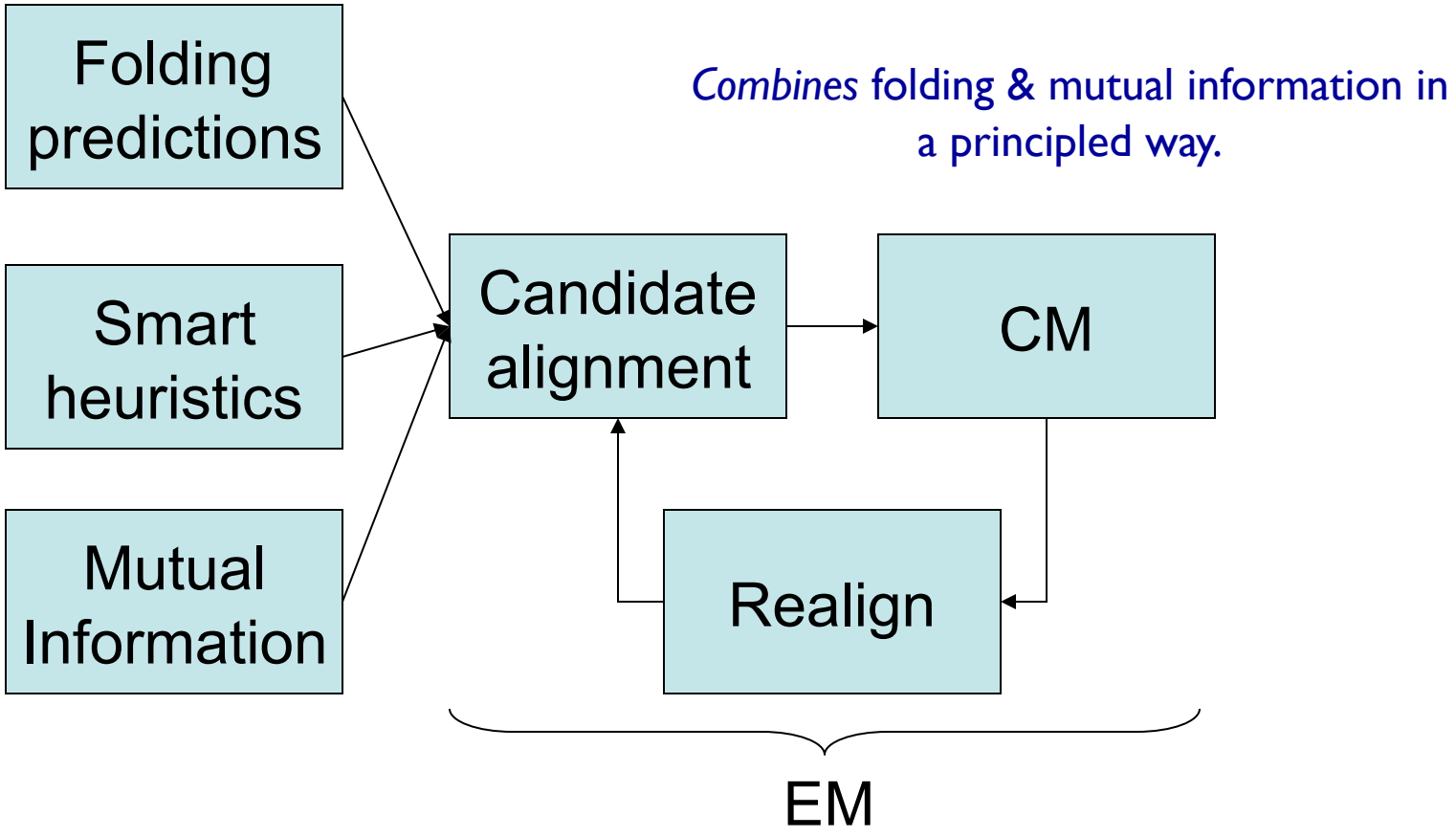
Application:
cis-regulatory ncRNA discovery in prokaryotes

Key issue is
exploiting prior knowledge
to focus on promising data

CMFinder

Simultaneous alignment, folding & motif description

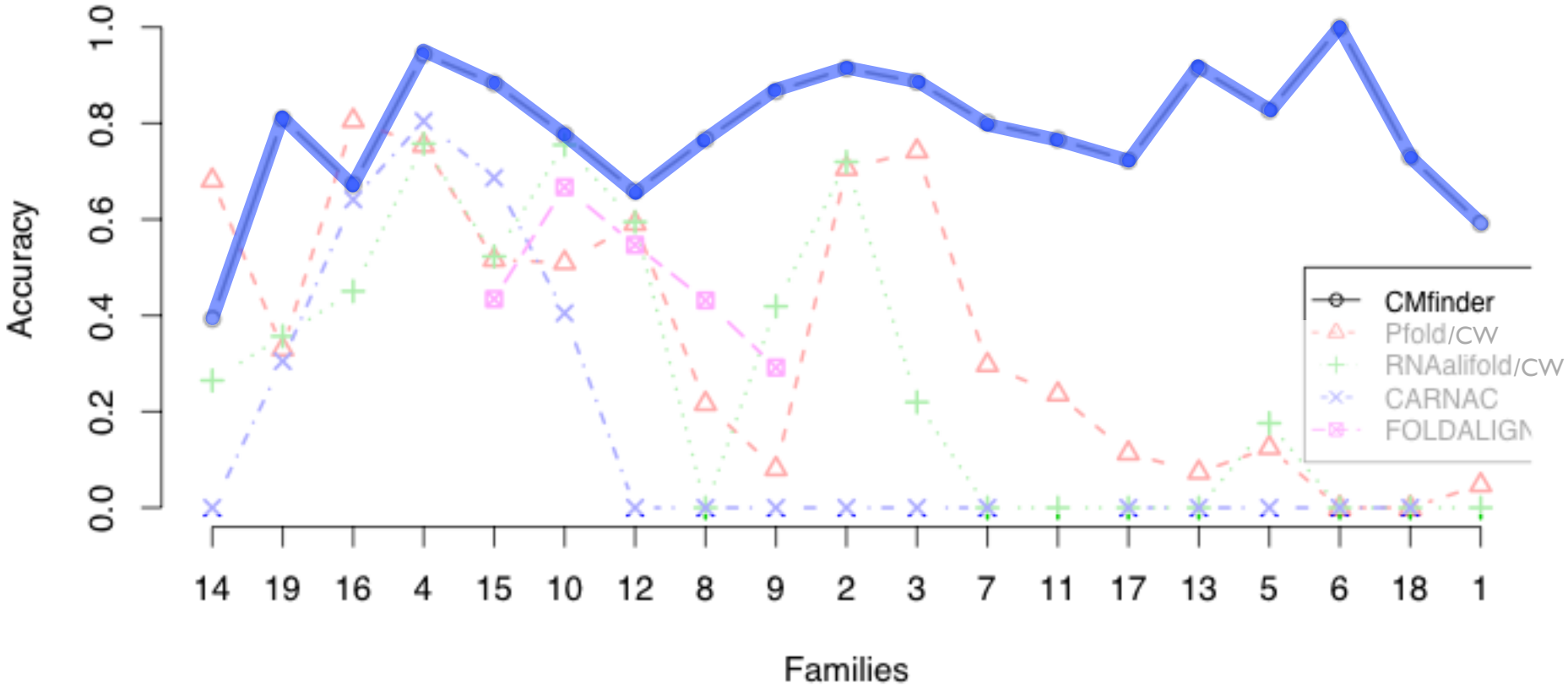
Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006



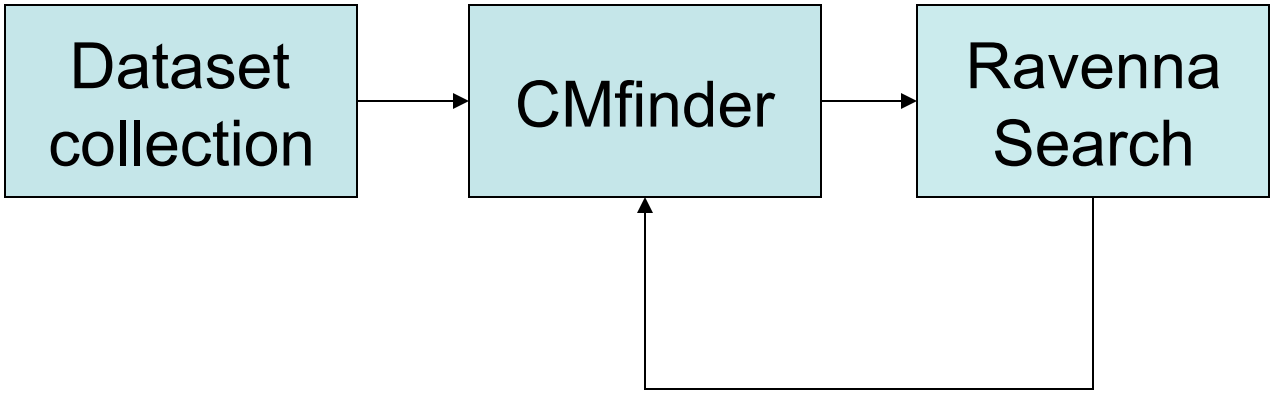
Combines folding & mutual information in a principled way.

CMfinder Accuracy

(on Rfam families *with* flanking sequence)



Use the Right Data; Do Genome Scale Search



Right Data:

- 5-10 examples amidst 20 extraneous ones OK; (but not 5 in 200 or 2000)
- length 1k (not 100k)

How:

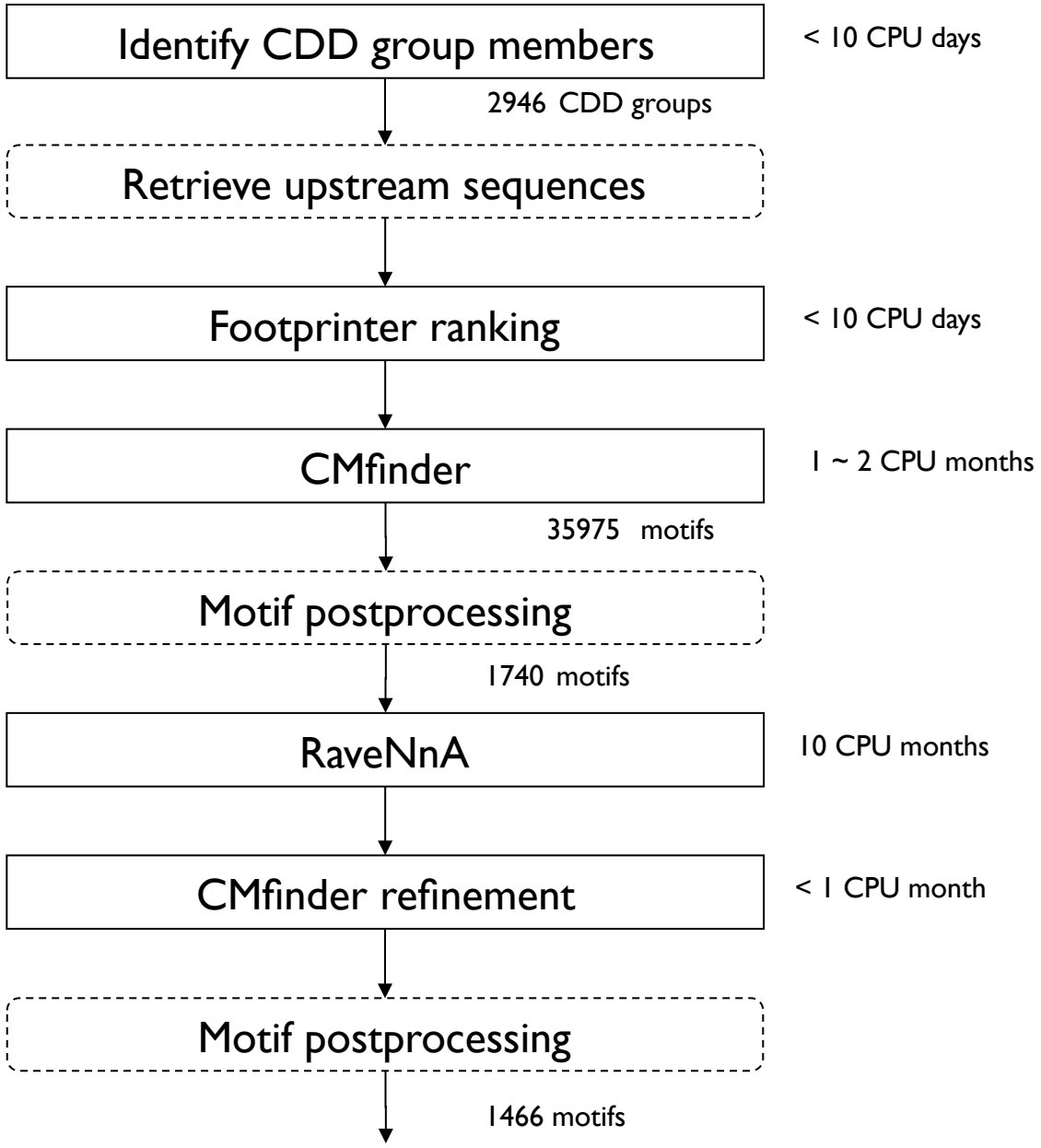
- Regulators near regulatees
- Get UTRs of homologs

Genome Scale Search:

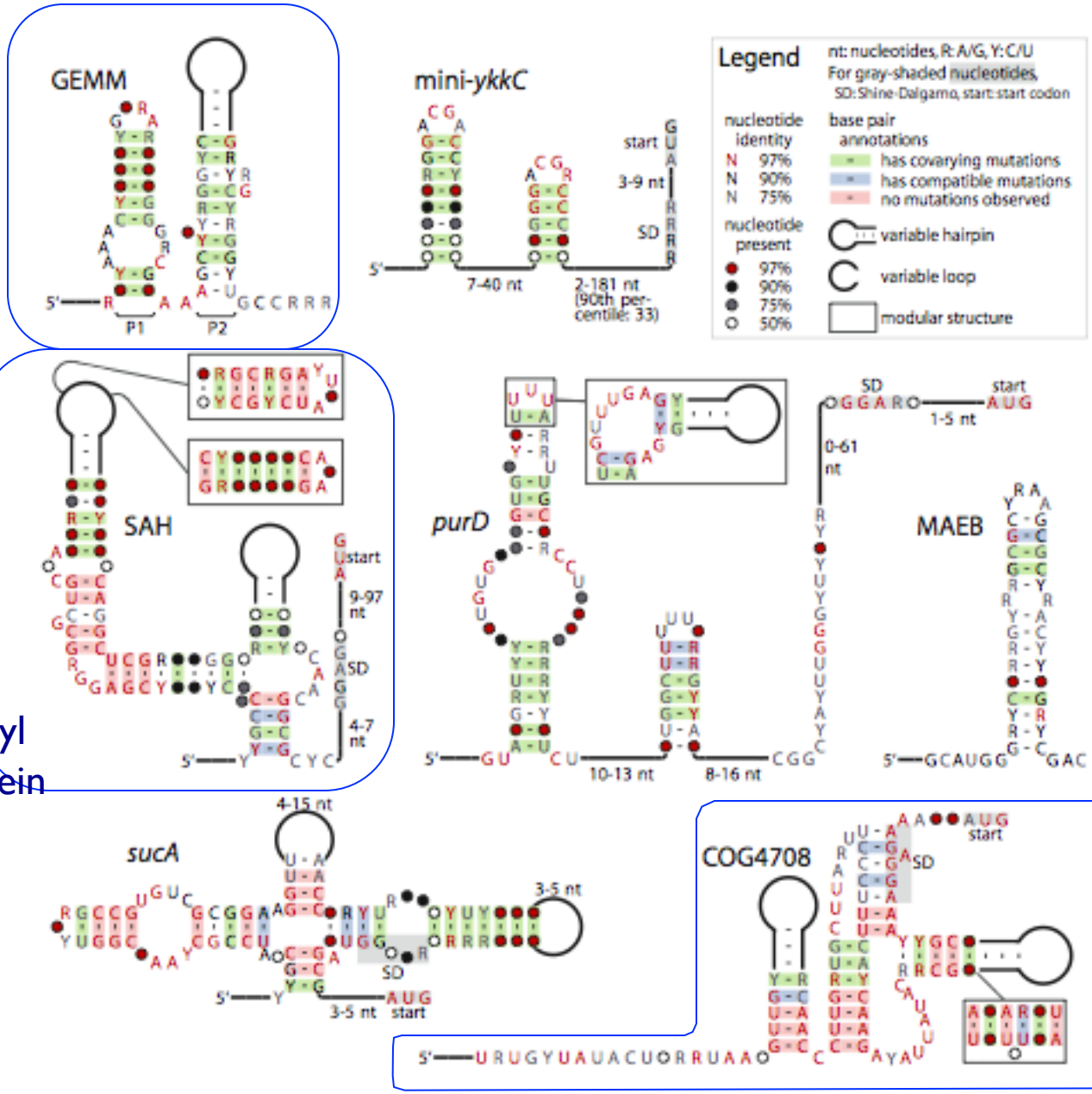
- Many riboswitches are present in ~5 copies per genome
- More examples = better model + clues to function

Processing Times

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



cyclic di-GMP



S-adenosyl homocystein

boxed = confirmed riboswitch (+2 more:

- S-adenosyl methionine
- molybdenum cofactor)

queuosine precursor

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

Riboswitch Summary

RNA elements that control ("switch") gene expression, *without* involvement of (transcription factor) proteins

Varied mechanism: Transcriptional, translational, on, off, combinatorial... Aptamer/expression platform.

Large diversity: Dozens of ligands, multiple aptamers for some, many operons, hundreds of species

Computationally challenging search/discovery

Many open problems!