

# 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>

# Central Dogma of Molecular Biology

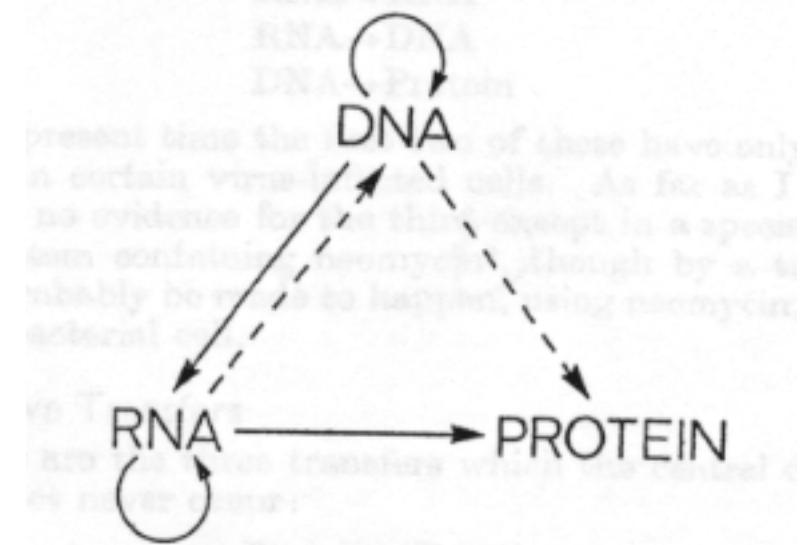
by

FRANCIS CRICK  
MRC Laboratory  
Hills Road,  
Cambridge CB2 2QH

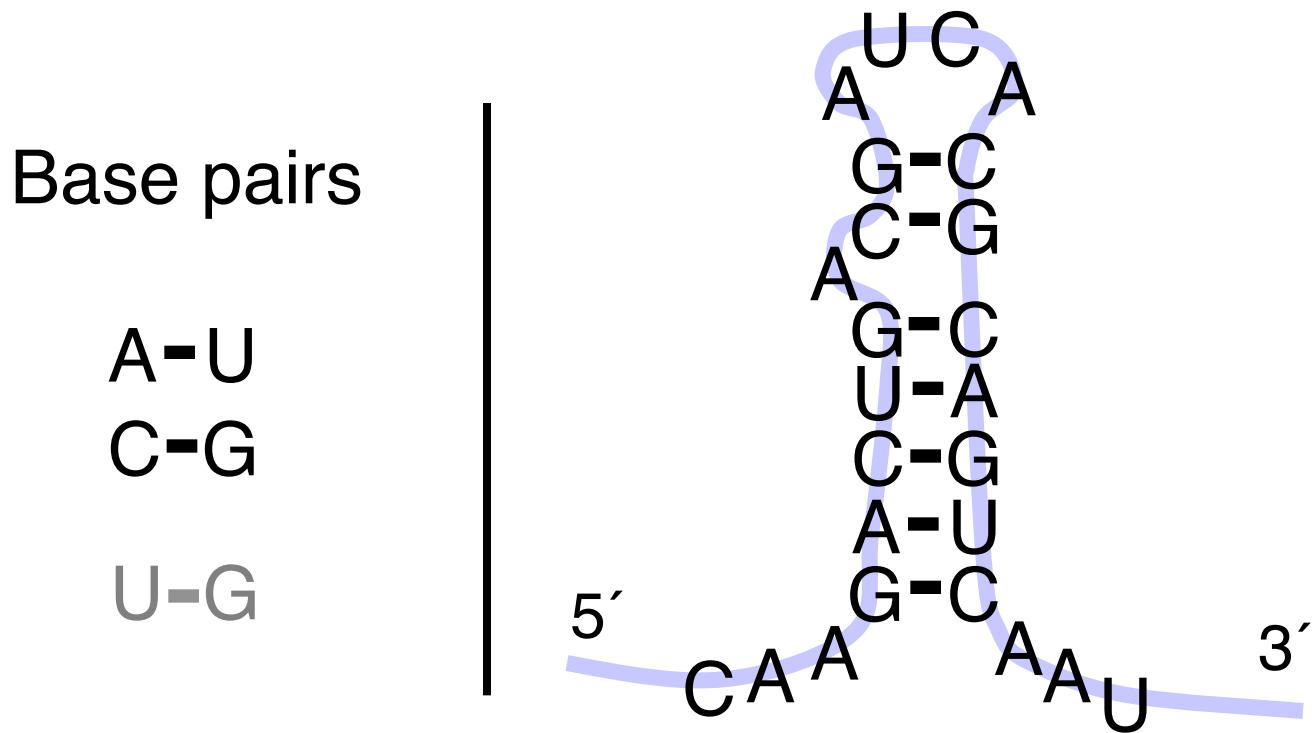
The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.

"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.

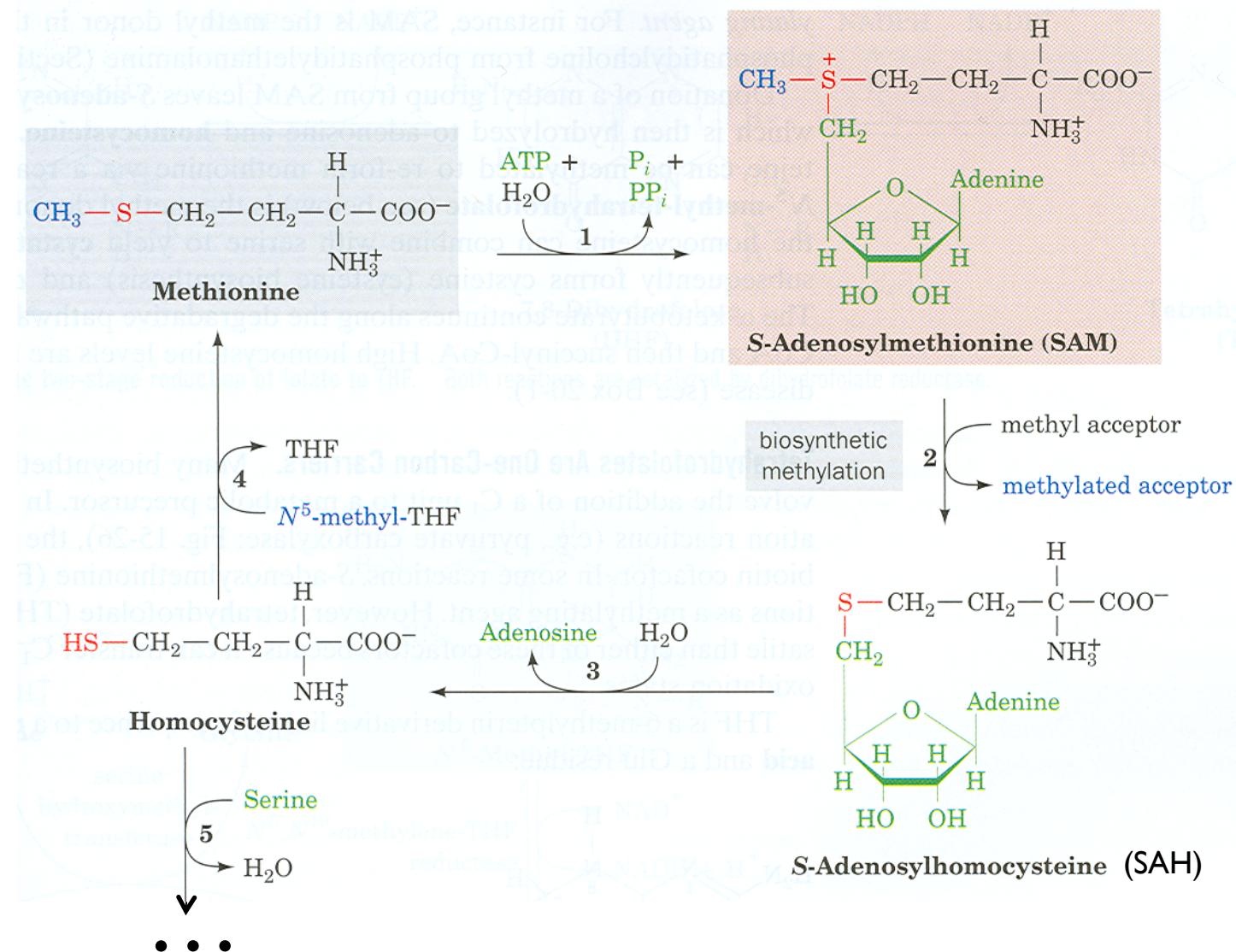


# RNA Secondary Structure: RNA makes helices too

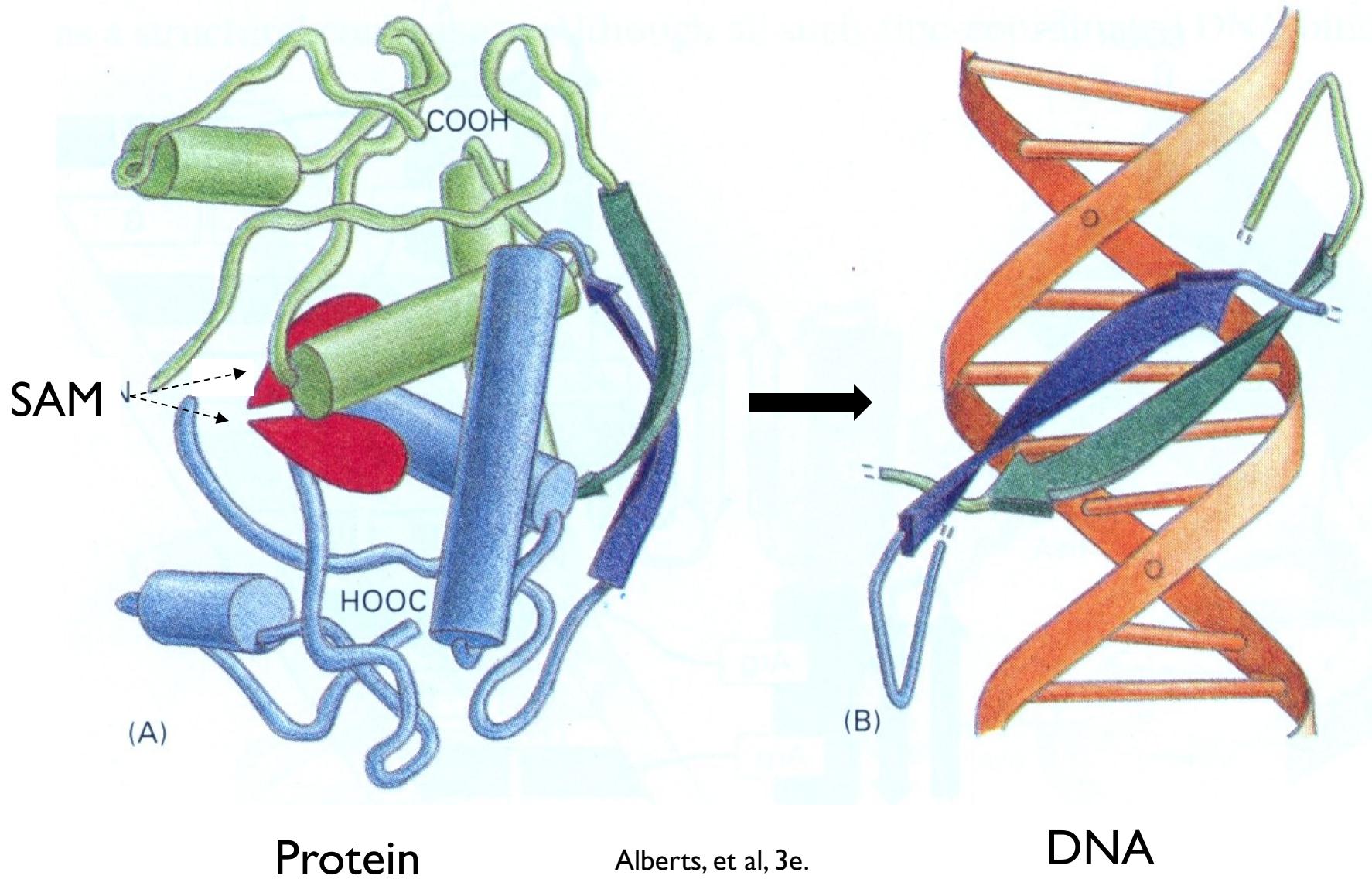


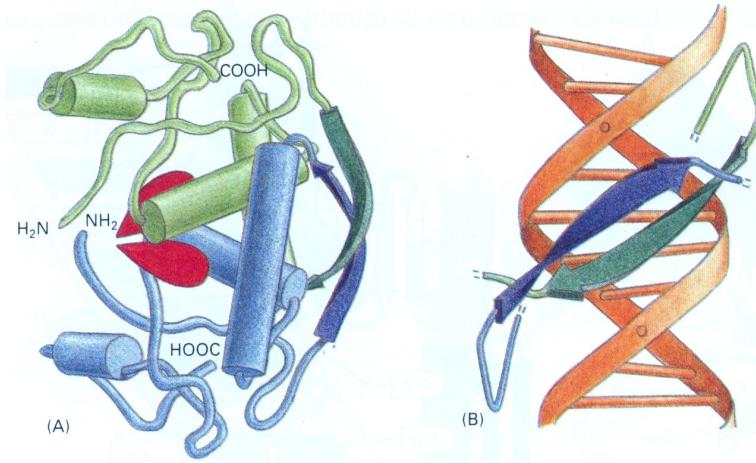
Usually *single* stranded

# Met Pathways



# Gene Regulation: The MET Repressor

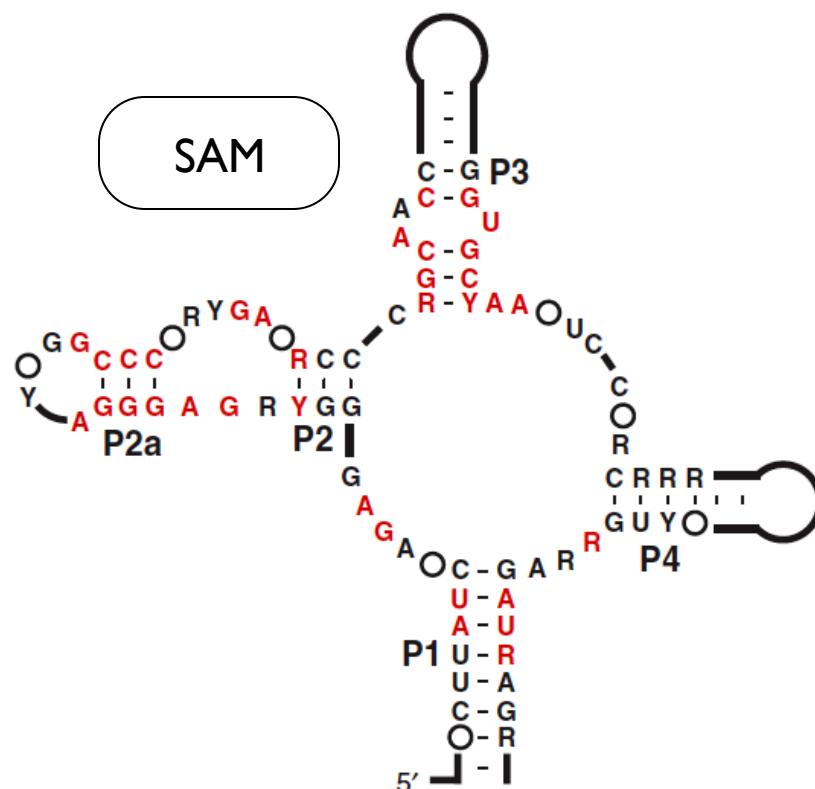




Not the only way!

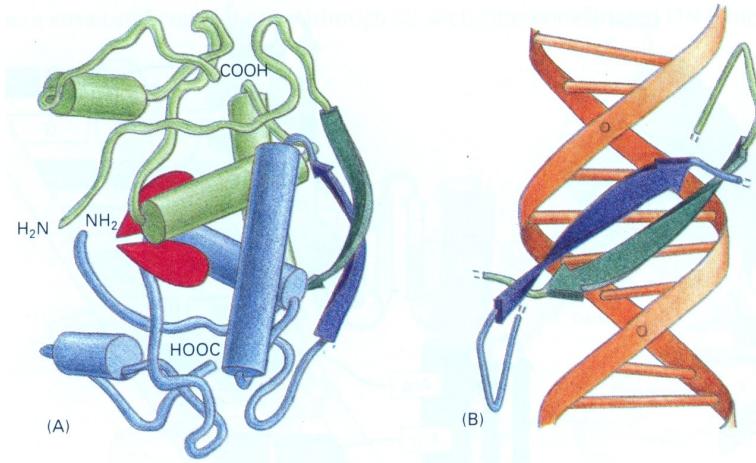
Protein  
way

Riboswitch  
alternative



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

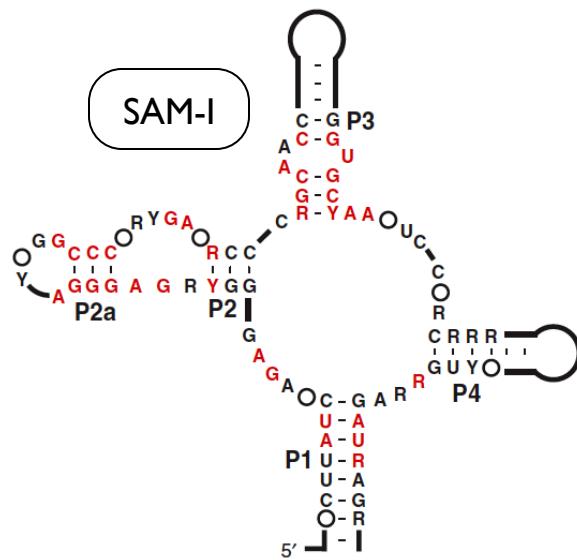
Alberts, et al, 3e.



Not the only way!

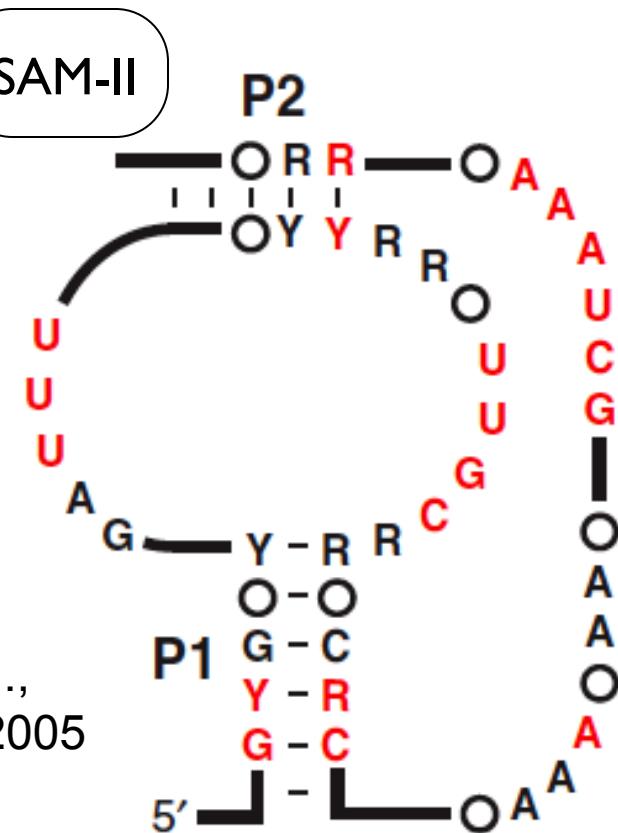
Protein  
way

Riboswitch  
alternatives

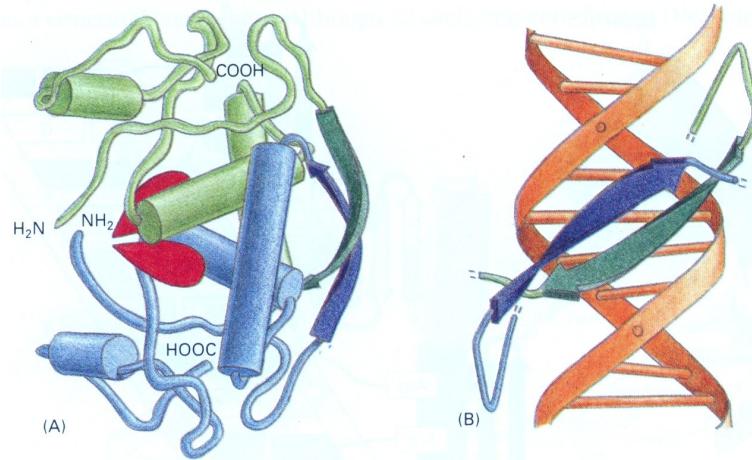


Grundy, Epshteyn, Winkler  
et al., 1998, 2003

Corbino et al.,  
Genome Biol. 2005



Alberts, et al., 3e.

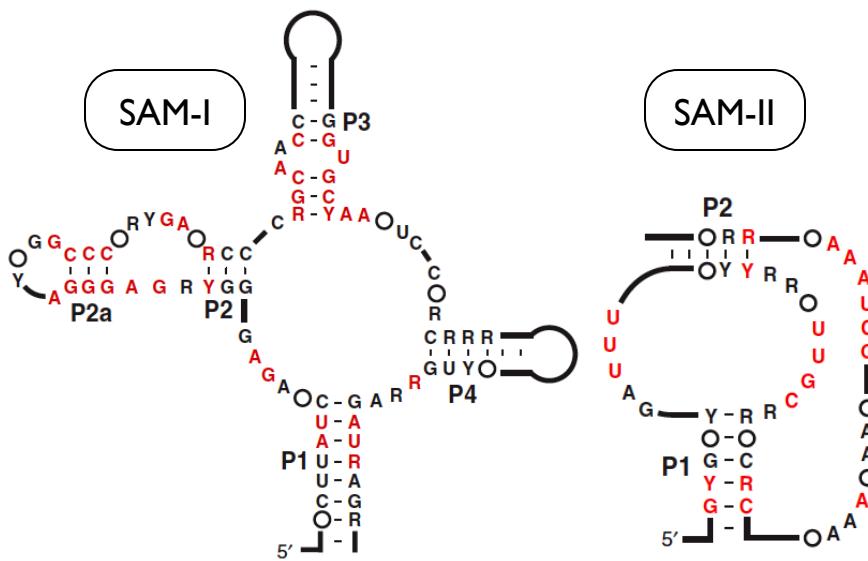


# Not the only way!

# Protein way

# Riboswitch alternatives

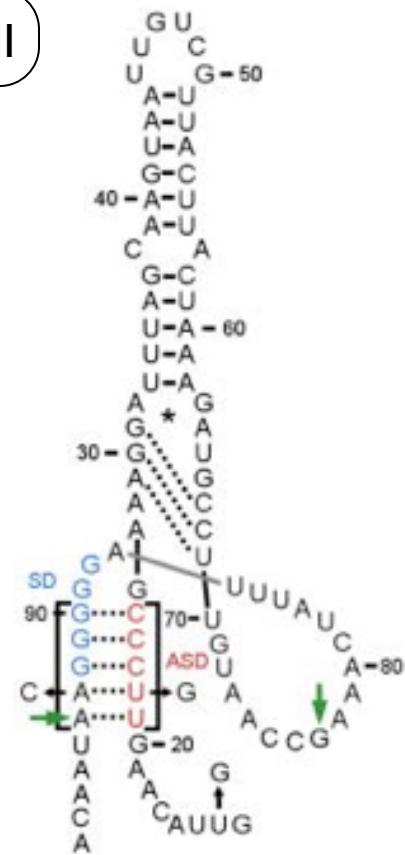
SAM-III

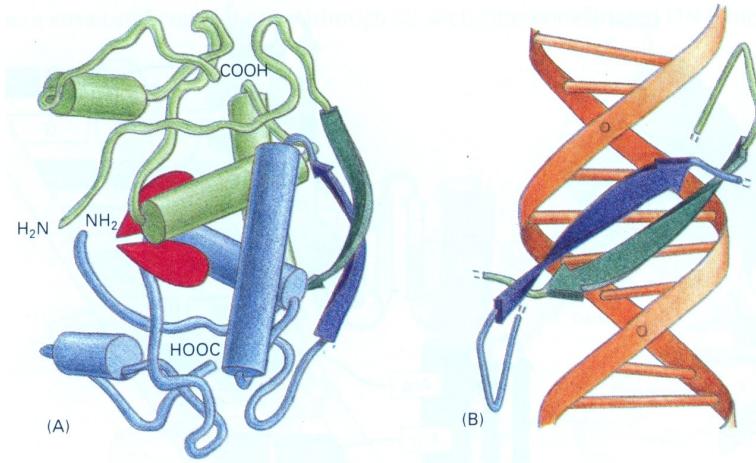


Grundy, Epshteyn, Winkler  
et al., 1998, 2003

Corbino et al.,  
Genome Biol. 2005

Fuchs et al.,  
NSMB 2006

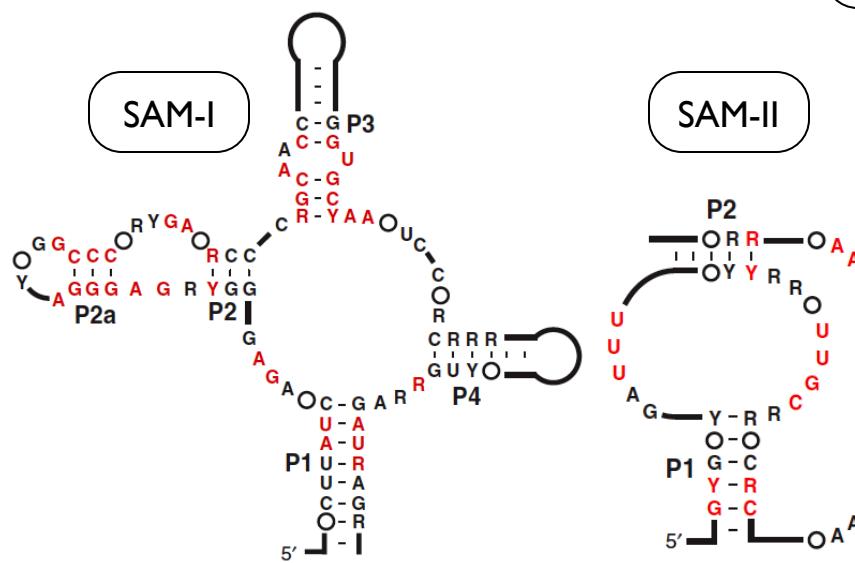




# Not the only way!

Protein  
way

Riboswitch  
alternatives

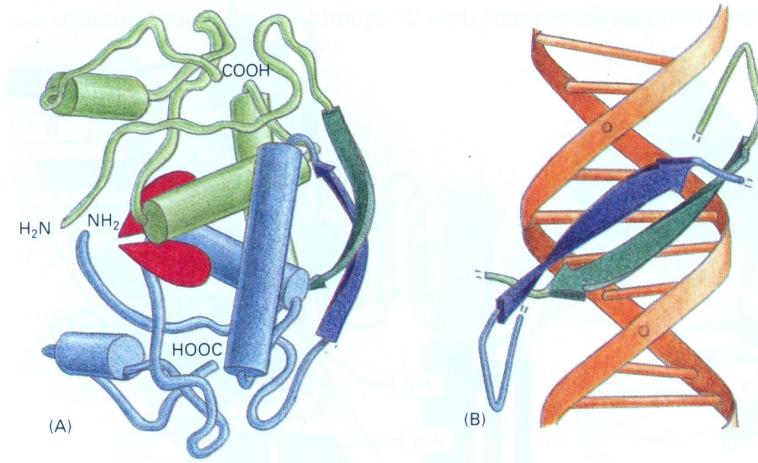


Grundy, Epshteyn, Winkler  
et al., 1998, 2003

Corbino et al.,  
Genome Biol. 2005

Fuchs et al.,  
NSMB 2006

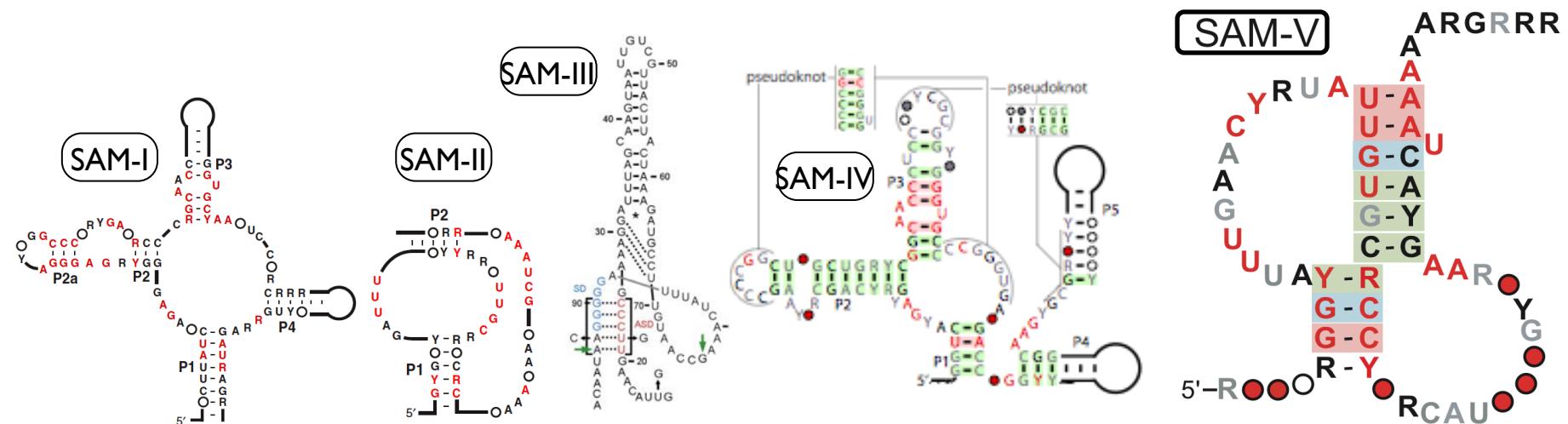
Weinberg et al.,  
RNA 2008



# Not the only way!

Protein  
way

Riboswitch  
alternatives



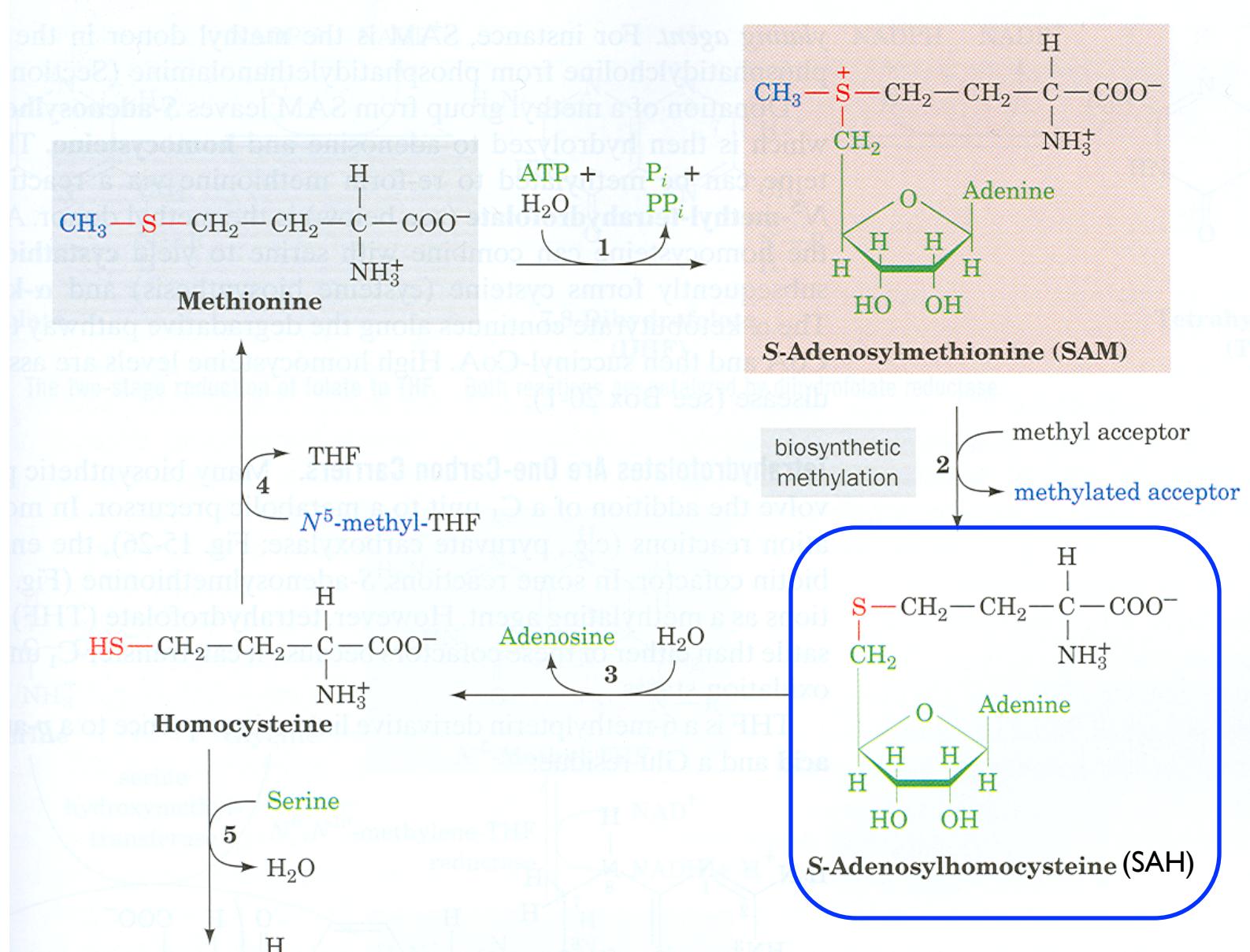
Grundy, Epshtain,  
Winkler  
et al., 1998, 2003

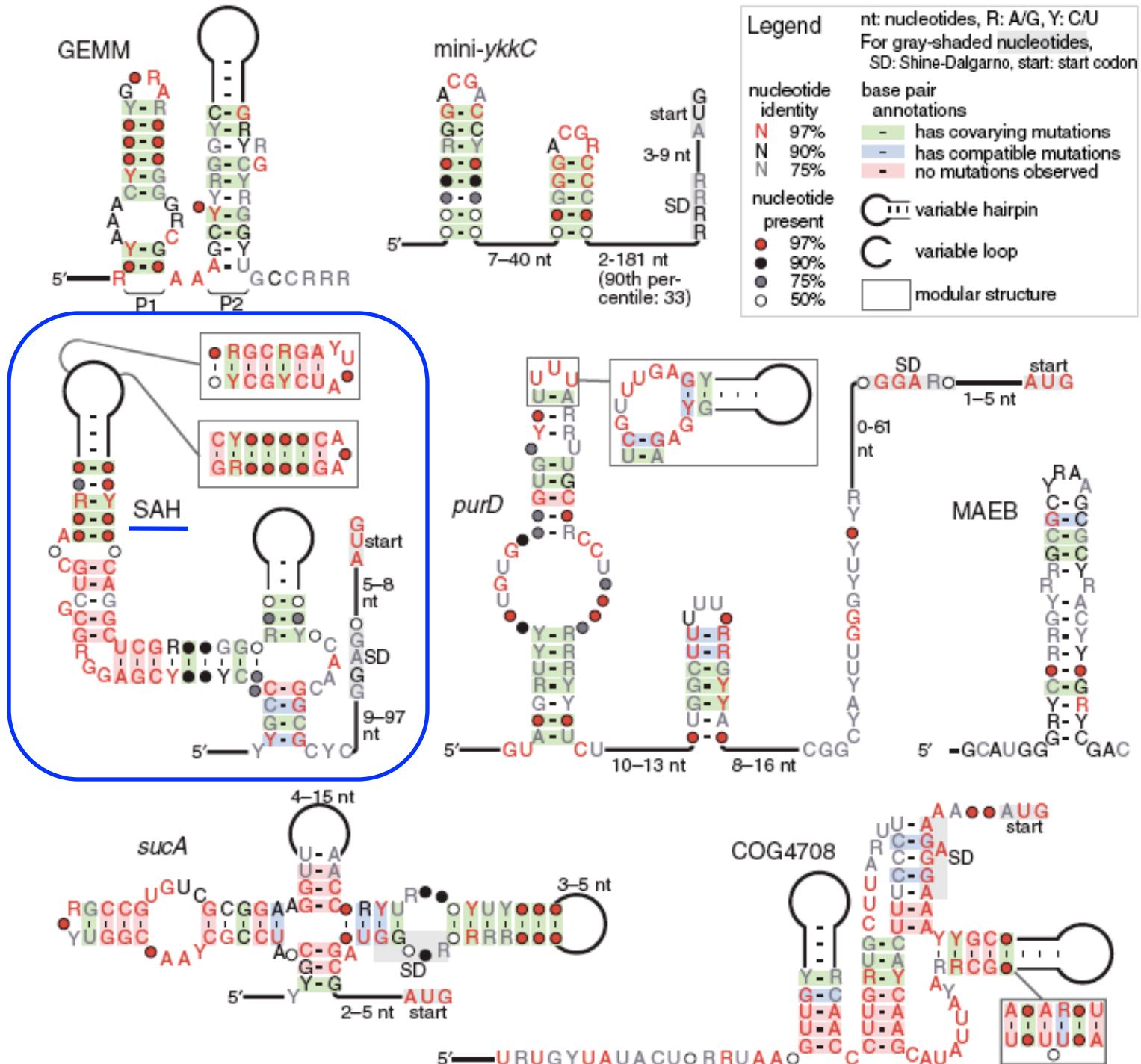
Corbino et  
al.,  
Genome  
Biol. 2005

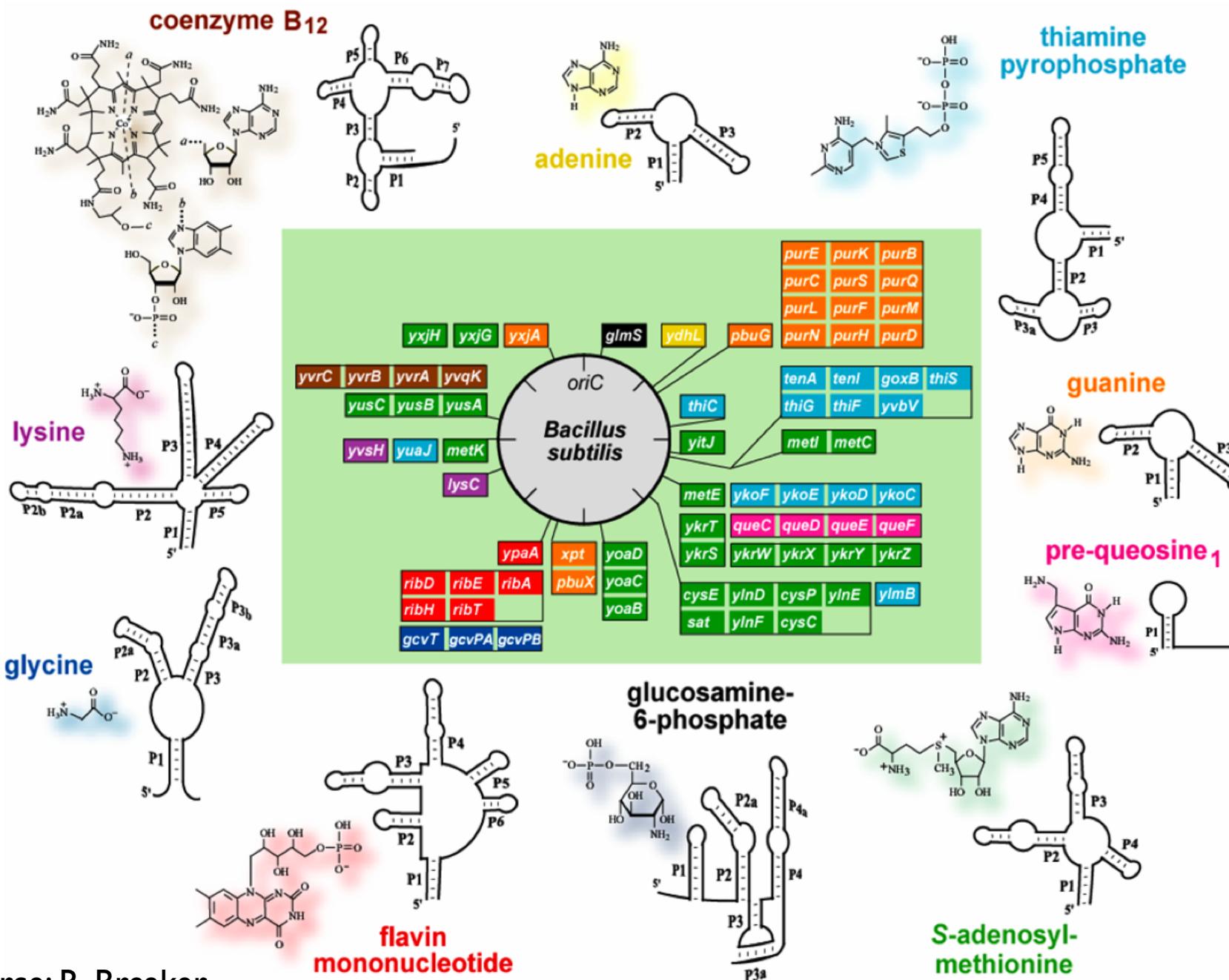
Fuchs  
et al.,  
NSMB  
2006

Weinberg  
et al.,  
RNA 2008

Meyer, et al., BMC  
Genomics 2009

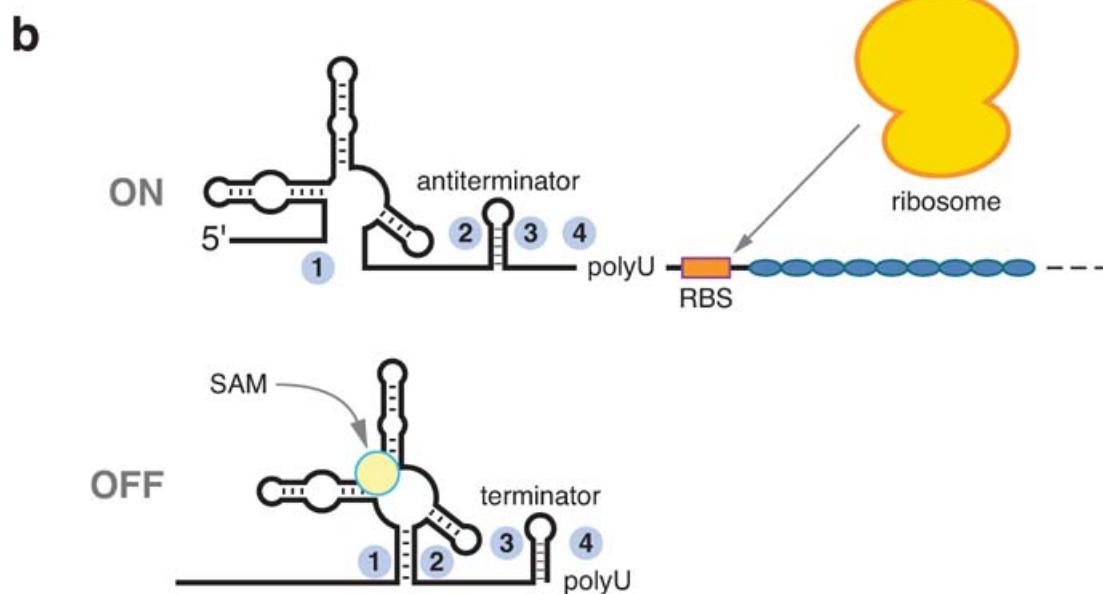
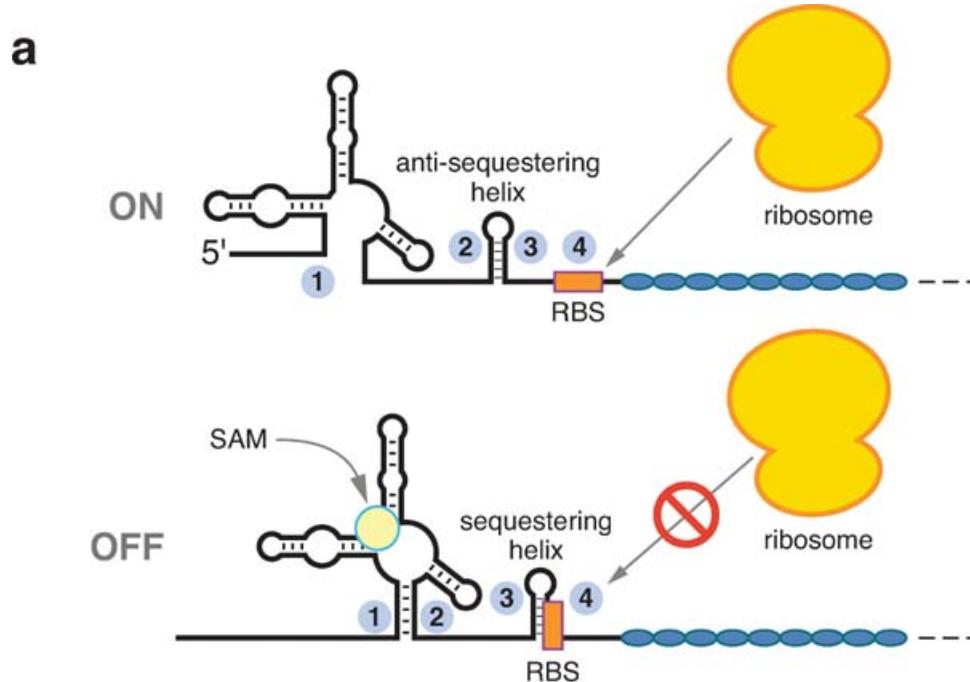






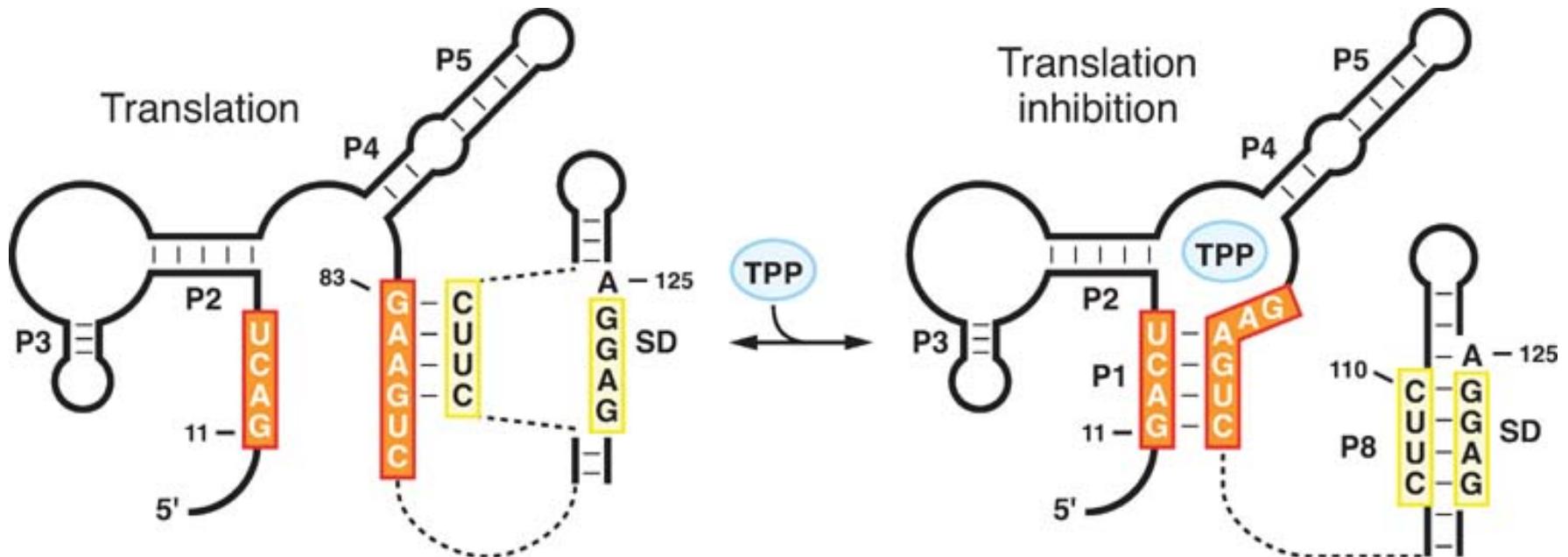
Source: R. Breaker

# Translational Control



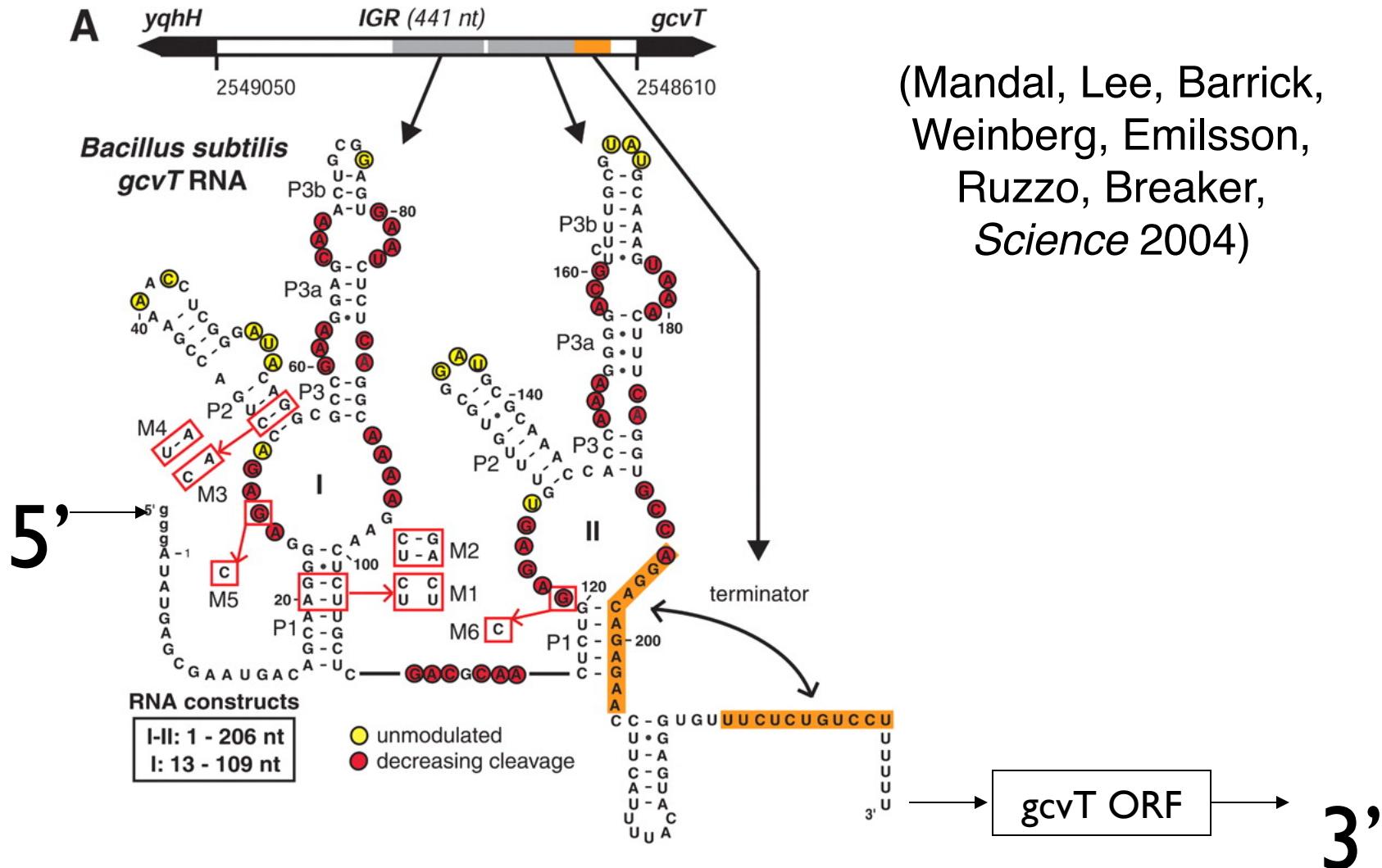
# Transcriptional Control

# Detail of Translational Control

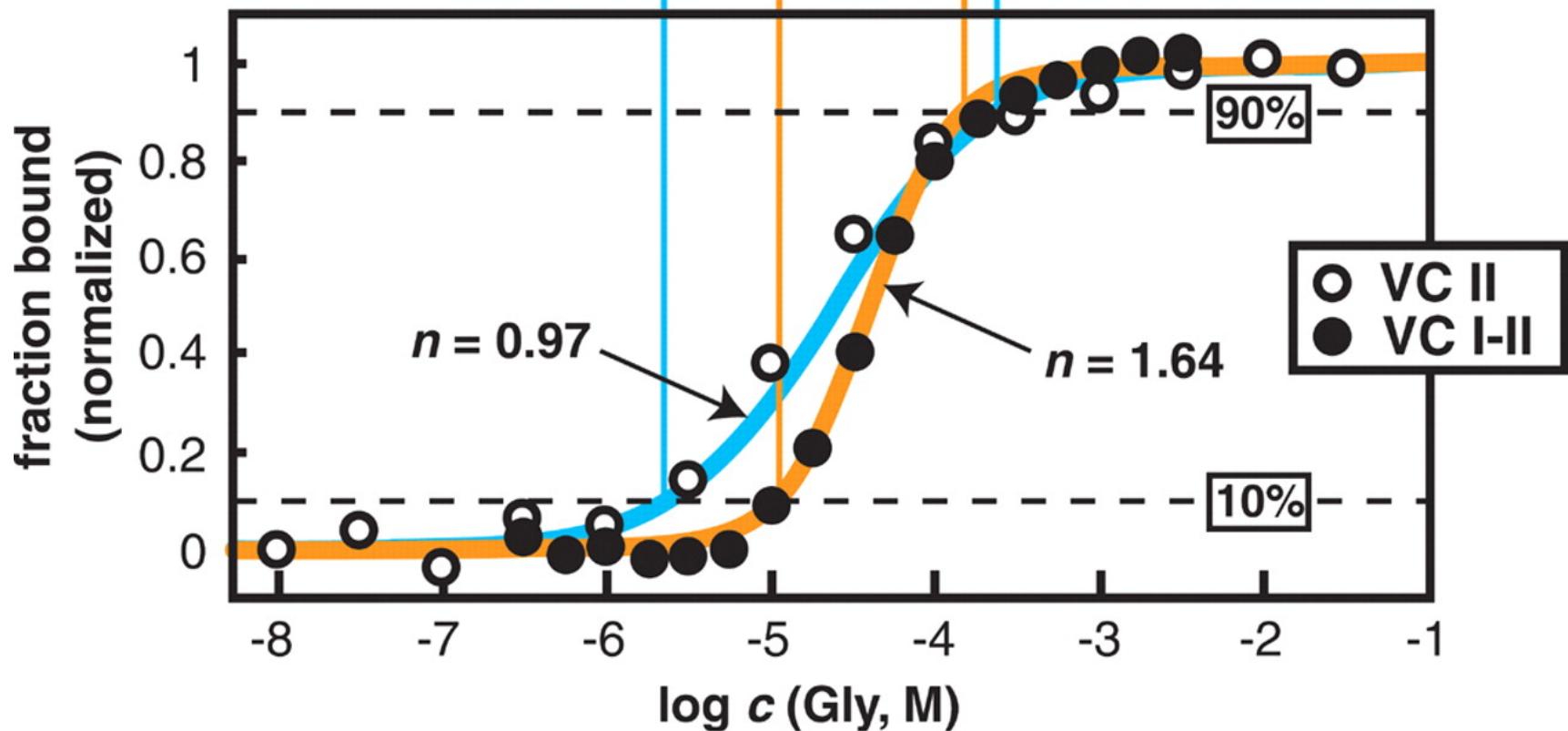


Winkler WC, Breaker RR. 2005.  
Annu. Rev. Microbiol. 59:487–517

# The Glycine Riboswitch



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 mark 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

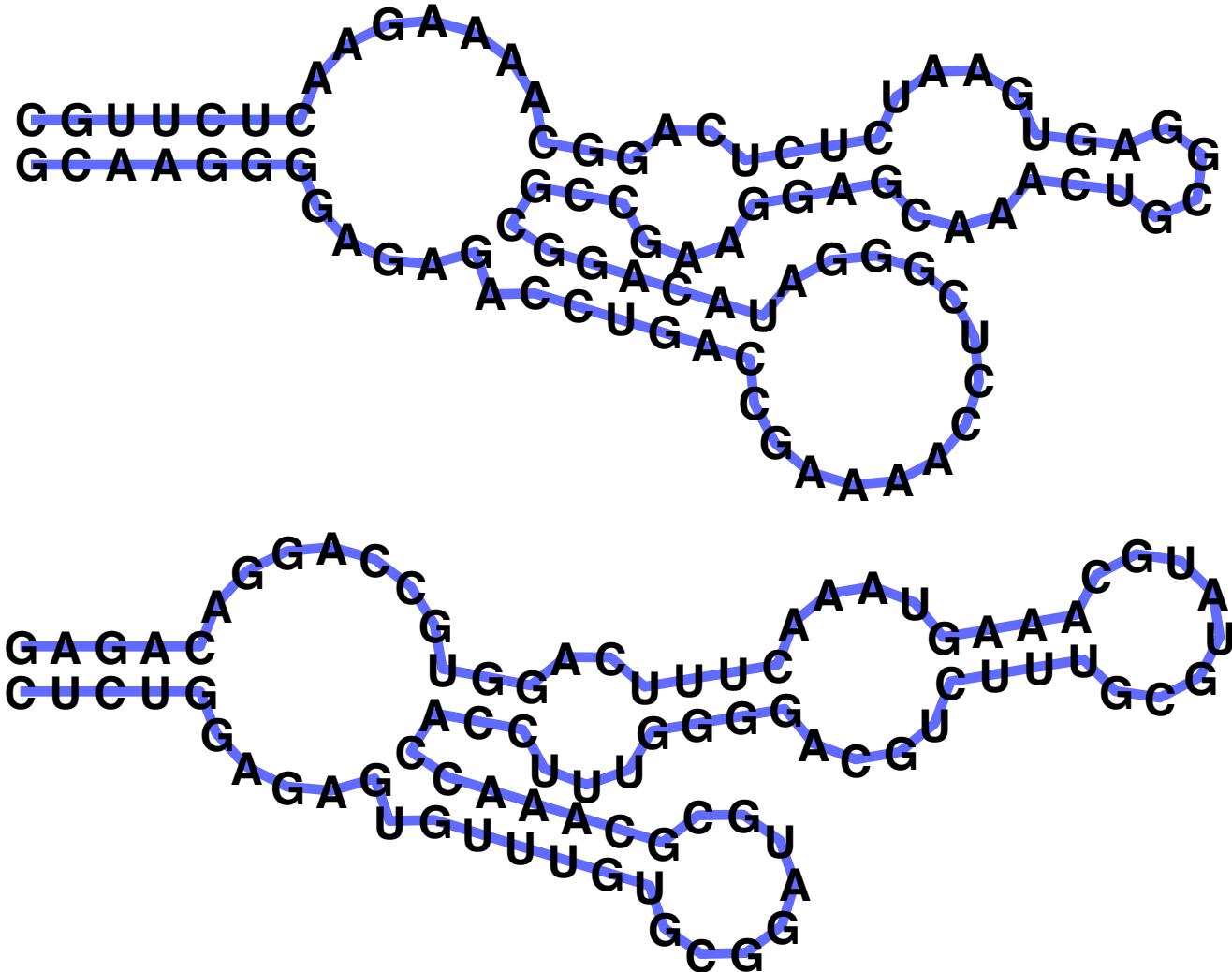
~ 20 ligands known; multiple nonhomologous solutions for some (e.g. SAM)

dozens to hundreds of instances of each

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

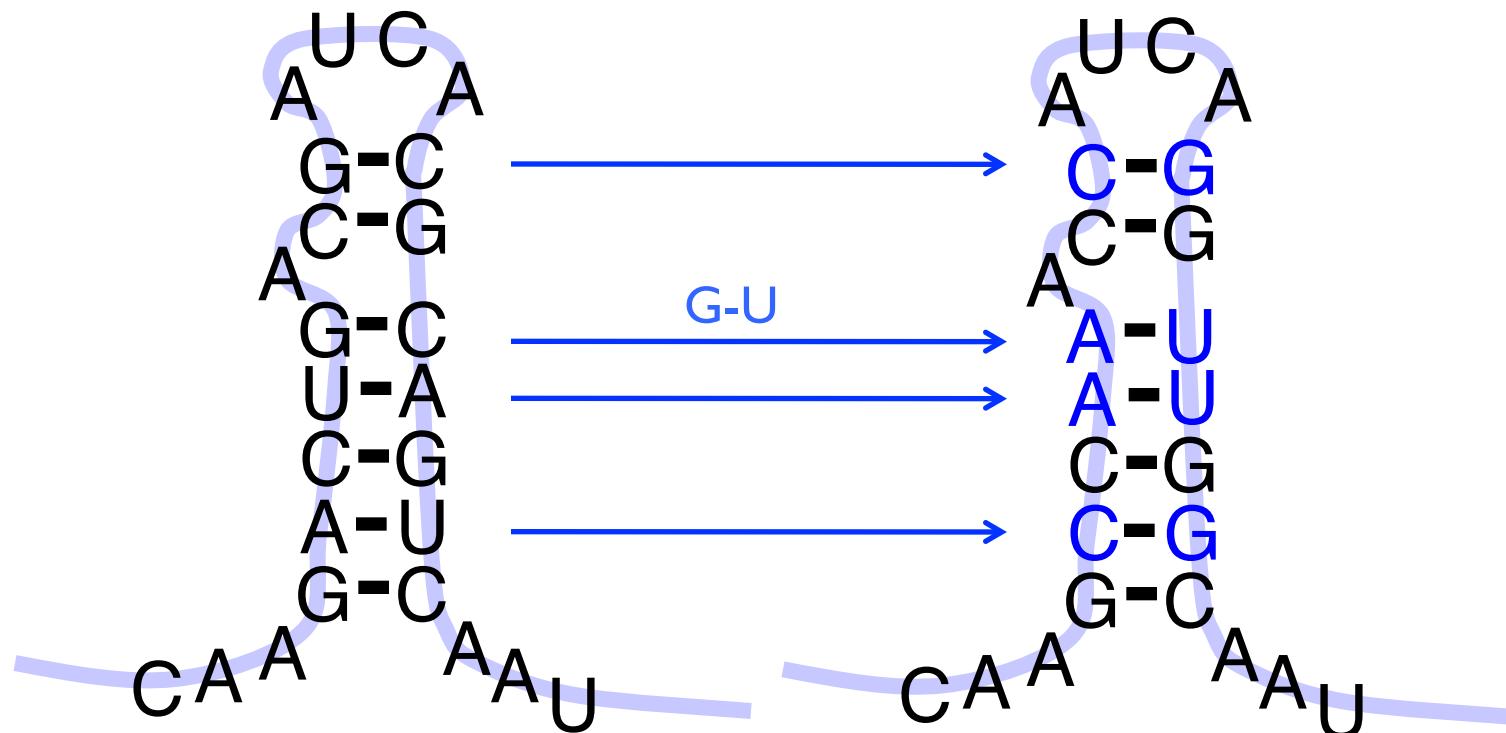
all found since ~2003; most via bioinformatics

# Why is RNA hard to deal with?

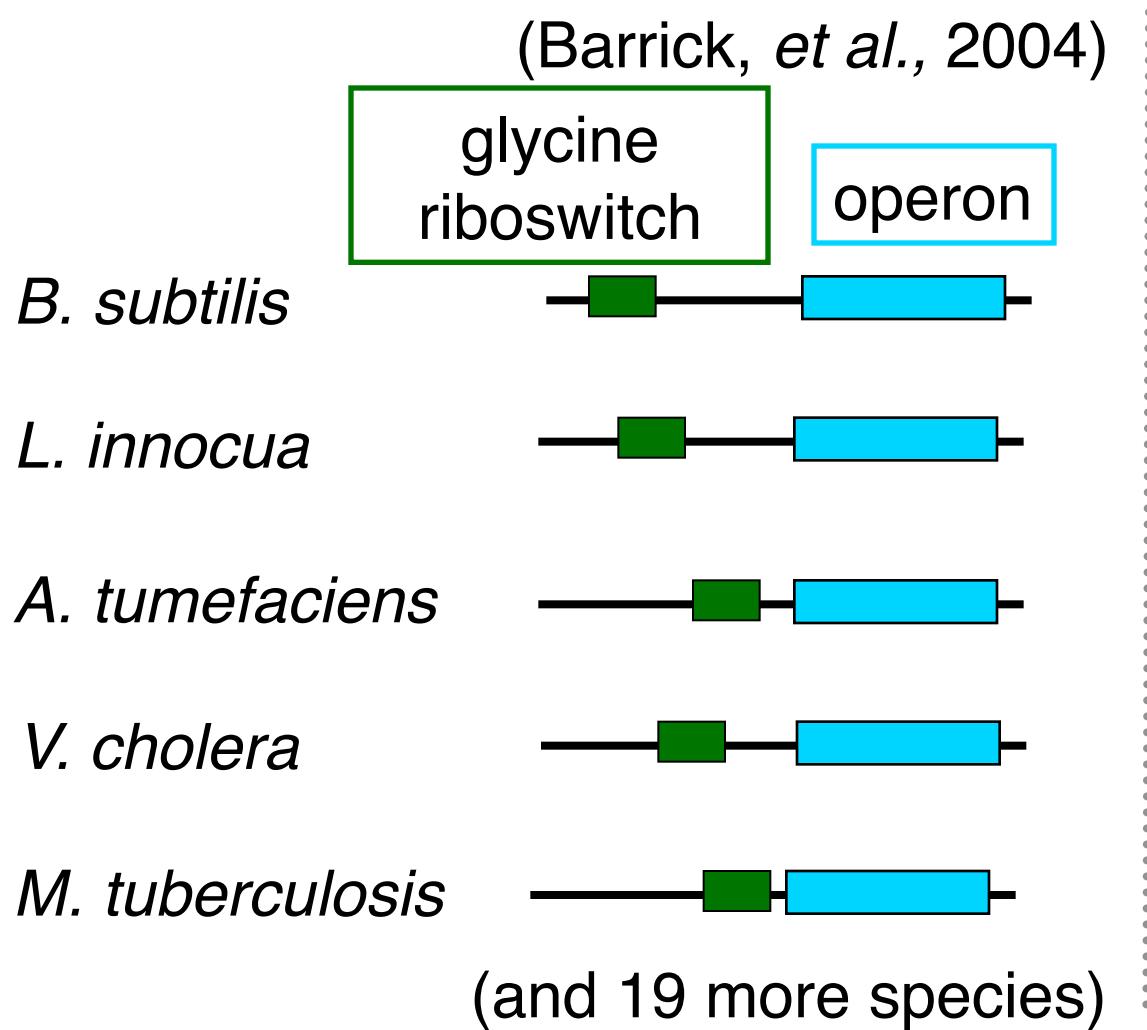


A: *Structure often more important than sequence*<sub>19</sub>

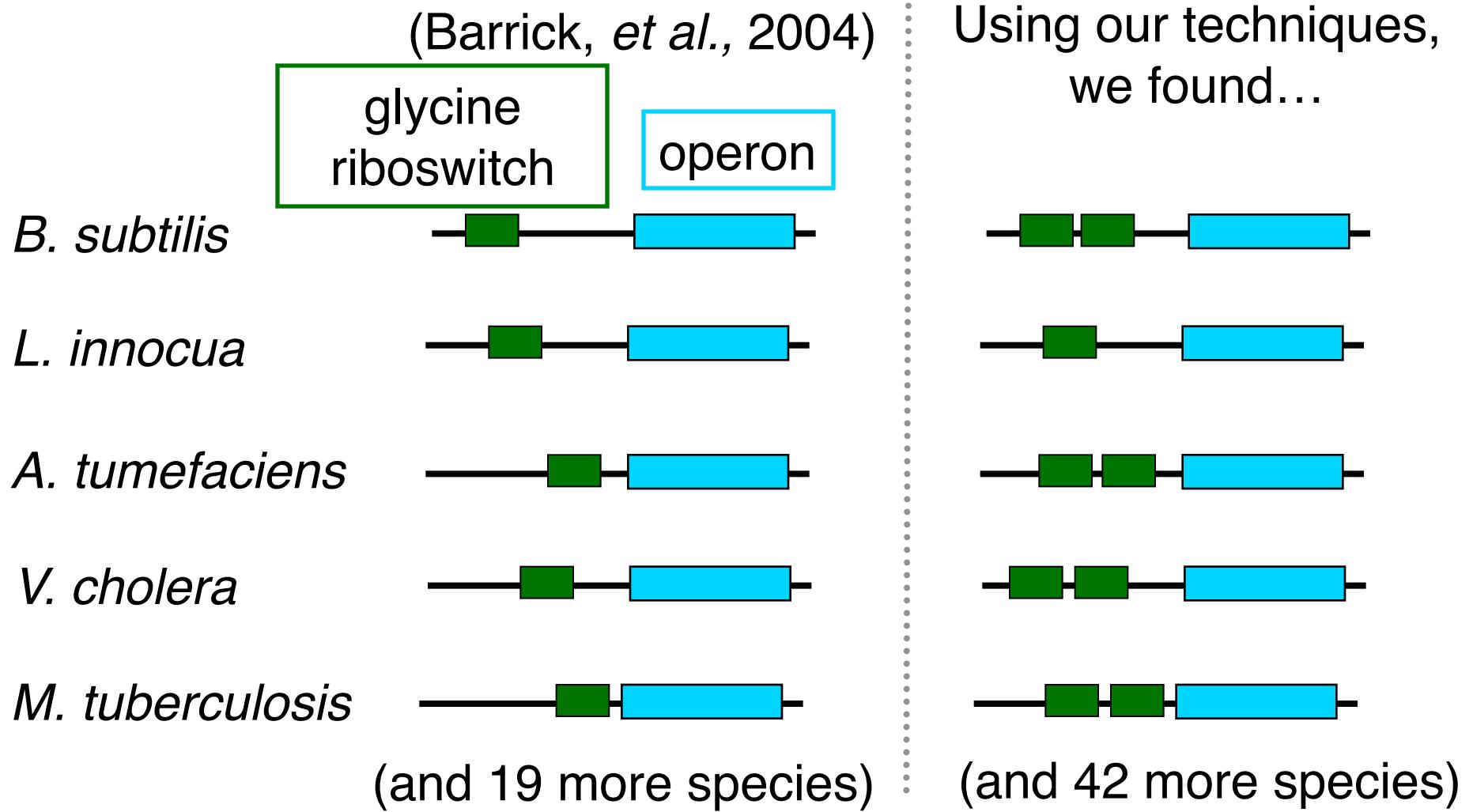
# RNA Secondary Structure: can be fixed while sequence evolves



## Impact of RNA homology search

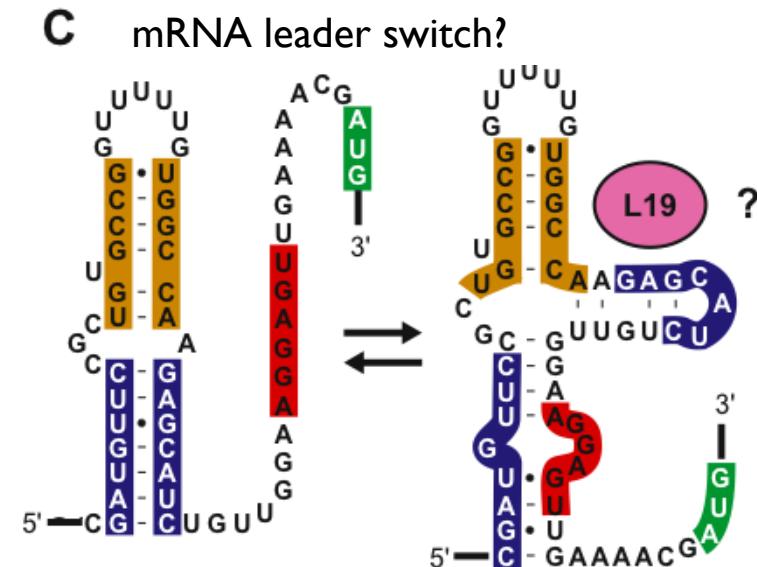
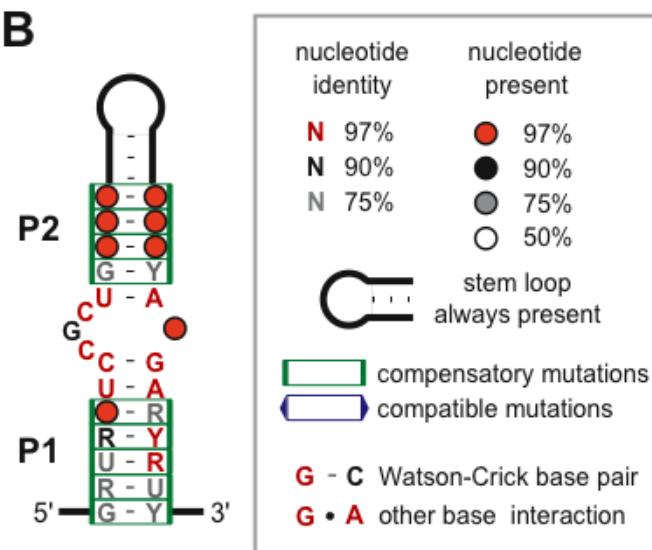


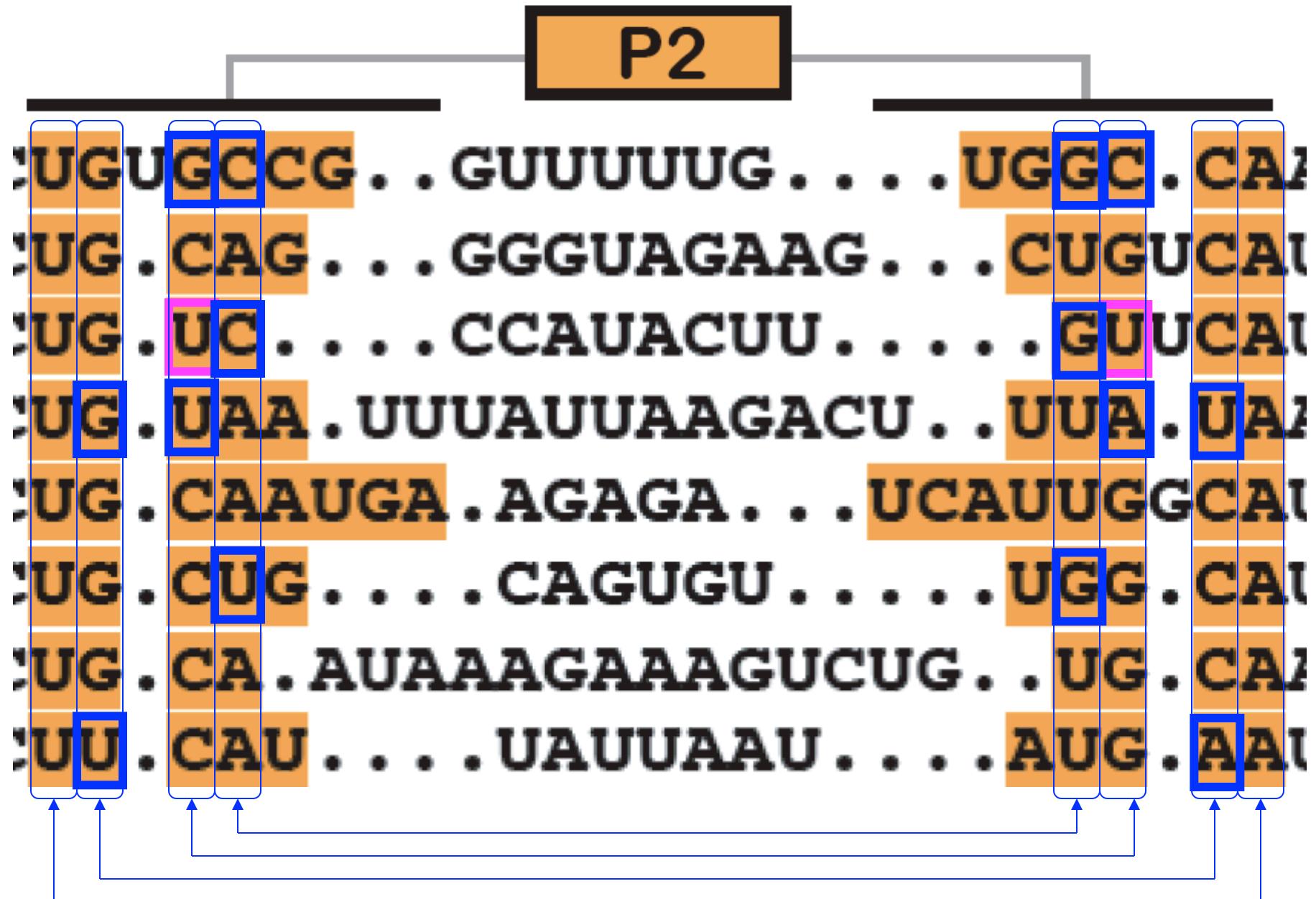
## Impact of RNA homology search



## A mRNA leader

|     | -35     | -10 | TSS<br>↑ | P1         | P2   | RBS         | Start                          |
|-----|---------|-----|----------|------------|--|-------------|--------------------------------|
| Bsu | TTGCAT  | 17. | TAAGAT   | 40. AAAAC  | GAUGUUC CGCUG UGCCG . GUUUUUG . . . UGGC           | CAAGACCAUC  | UG.05. AGGAGU.08. AUG          |
| Bha | TTGTTTC | 17. | TCTTCT   | 17. AUUAC  | GAUGUUC CGCUG CAG . . . GGGUAGAAG . . . CUGUCAUG   | GACCAUC     | UG.06. AGGAGG.11. AUG          |
| Oih | TTGAAC  | 17. | TATATT   | 31. UAAAC  | GAUGUUC CGCUG UC . . . CCAUACUU . . . GUCAUG       | GACCAUUAG   | .06. AGGAGU.07. AUG            |
| Bce | TTGCTA  | 18. | TATGCT   | 36. UUAAC  | GAUGUUC CGCUG UAA . . . UUUAUUAAGACU . . . UUA     | UAA GACCAUC | UG.05. AGGAGA.09. AUG          |
| Gka | TTGCCT  | 17. | TATCAT   | 38. AAAAC  | GAUGUUC CGCUG CAAUGA AGAGA . . . UCAUUGC           | GAACACAU    | UG.04. AGGAGU.08. AUG          |
| Bcl | TTGTGC  | 17. | TATGAT   | 45. AUUAC  | GAUUAUC CGCUG CUG . . . CAGUGU . . . UGG           | CAUGAAUGUC  | UG.06. AGGAGG.10. AUG          |
| Bac | ATGACA  | 17. | GATAGT   | 35. AUUAC  | GAUGUUC CGCUG CA . . . AUAAAAGAAAGUCUG . . . UG    | CAAGACCAUC  | UG.05. AGGAGU.08. AUG          |
| Lmo | TTTACA  | 17. | TAACCT   | 28. AUUAC  | GAUUAUC CGCUU CAU . . . UAUUAAU . . . AUG          | AAUGAAUGUU  | UG.05. AGGAGA.07. AUG          |
| Sau | TTGAAA  | 17. | TAACAT   | 23. AUCAC  | UAUGAUC CGCUG CU . . . AUAAUUAUUGUCG . . . AGGCAAG | RACAUAGG    | .04. AGAGCA.09. AUG            |
| Cpe | TTAAAG  | 18. | TAACAT   | 08. GUACC  | GGCGGU CUCUGUCACAC . . . GAG . . . UGUGUUAAGA      | ACGUCAA     | .17. AGGAGG.08. AUG            |
| Chy | TTGCAT  | 17. | TATAAT   | 09. UACCAA | ACGUUC CGCUG GA . . . CAGGGGC . . . UC             | CAUGAACGU   | GCC.03. AGGAGG.09. AUG         |
| Swo | TTGAGA  | 17. | AAAAAT   | 16. AAAAA  | GGUGGU CCGUG CAUU . . . AACUAA . . . AAUG          | UACACC      | UU.05. AGGAGG.07. AUG          |
| Ame | TTGGGG  | 17. | TATAAT   | 10. UUACG  | GGCGGU CUCUA UAC . . . AGGA . . . GUA              | UAA GACGU   | UA.07. AGGAGG.07. AUG          |
| Dre | TTGCC   | 17. | TATAAT   | 16. UUACG  | GACGGU CGCUG CCU . . . CUGGGAA . . . AGG           | UAA GACGU   | CGUUA.04. AGGAAC.12. GUG       |
| Spn | TTTACT  | 17. | TAACAT   | 28. AUAC   | AGGUUAUC CGCUG AGGA . . . AGAU . . . UCCU          | CAAGAU      | UGACAA.04. AGGAGA.05. AUG      |
| Smu | TTTACA  | 17. | TACAAT   | 26. AAACG  | GUUAUC CGCUG AG . . . ACAGAGCA . . . CU            | UAU         | GAUUAGUAA.04. AGGAGA.07. AUG   |
| Lpl | TTGCGT  | 18. | TATTCT   | 21. UUAC   | GAUGUUC CGCUG AC . . . CAGGUU . . . GU             | CAC         | GAUUGUCGG.04. AGGAGC.09. AUG   |
| Efa | TTTACA  | 17. | TAACAT   | 28. AUUAC  | AAUAUUC CGCUG UGG CA . . . GAAG . . . UGACCA       | UAA         | GAUAAUUUG.06. AGGAGA.08. AUG   |
| Ljo | TTTACA  | 17. | TAACAT   | 25. UUAUC  | GGGUAAUC CGCUG GCAC . . . AAG . . . GUGU           | UGA         | UAGAAUGCCGU.03. AGGAGA.07. AUG |
| Sth | TAGACA  | 17. | TAAGAT   | 29. UAACC  | GGCUAAUC CGCUG AGA CA . . . CAGAGGU . . . UGCUCU   | UAA         | GAUUAGUAA.03. AGGAGU.08. AUG   |
| Lac | TTAAAA  | 17. | TTACTT   | 39. UUAUC  | GGGUAAUC CGCUG ACC . . . CUGGUA . . . CGU          | UGA         | UAGAAUGCCGA.03. AGGAGA.10. AUG |
| Spy | TTTACA  | 17. | TAGAAT   | 29. UUACG  | GUUAUC CGCUG AG . . . ACAAGUA . . . CU             | UAA         | GAUUAGUAA.03. AGGAGA.06. AUG   |
| Lsa | TTTTAA  | 17. | AAAAAT   | 26. ACAAC  | GAUUAUC CGCUG GCG . . . CAAGA . . . CGU            | AAU         | UAGAAUACUG.06. AGGAGA.07. AUG  |
| Lsl | TTTACT  | 17. | TATTTT   | 24. AUUAC  | GAUUAUC CGCUG C . . . AACUG . . . GACAU            | GAU         | UAGAAUGUCGG.04. AGGAAA.07. AUG |
| Fnu | TTGACA  | 17. | TAACAT   | 12. AAUUC  | GAUUAUC CGCUU UAA . . . UAAA . . . UUA             | AAU         | GAUUAUCUU.04. AGGAAG.02. AUG   |





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

The image displays a CLUSTALW multiple sequence alignment. It consists of two rows of sequence text. The top row is a reference sequence and the bottom row is a query sequence. Colored boxes are used to highlight structural conservation. In the reference sequence, the boxes are mostly aligned with each other. In the query sequence, the boxes are also aligned with the reference sequence, but the sequence itself is very different. This illustrates that sequence conservation does not necessarily mean structural conservation.

```
-----CCCCCCCCAGGCCCTGGTGCAGG--ATGATGACGACCTGGGTG-GAA-A---CCTACCCCTGTGGGCACCC-ATGTCGA-CCCCCCTGGCATT
GGGATCATTCAGCAAGAGCAGCGTG--ACTGACATTA---TGAAGGGCTGTACTGAAGACAGCAA--GCTGTTAGTACAGACC--AGATG---CTTCTTGGCAGGCTCGTTGACCTCTGGAAAACCTCAAT
AGGTTTGCATTAATGAGGATTACACAGAAAACCTT-GTTAAGGGTTGTGATCTGCTAA-TTGGCAAAATTTTATTTTAAAT---ATTCTTACAGAAGAGTCCATTAAAGAATGTTGTGATAGG
AGTGTGCGGATGATACTACTGACGAAAGACTCATCGACTCAGTTAGTGGGTGATGTAGTCACATTAGTTGCCCTCCCCCATCTTG---TCTCCCTGGCAAGGAGAATATGCCGACATGATGCTAAGAG
TGGACTGATAGGTA-GCCATGGC--TTCATCTGTC---ATG---TCTGCTTCTTTATATTG--TGTATGATGGTCACAGTGAAAG---TTCCCACAGCTGTGACTTGATTTAA-AAATGTCGGAAGA
TAAACTCGAACCTGGAGCGGGCAATTGCTGATTACGA-TTAACCACGTATCCCTGGGTCGCTGC--TTCTGGCCGTGCTCGGTTCCA-----TTTATCAACTATTAGCTCCAATACATAGCTACAGGTTTTT
AAATTCTCGCTATATGACGATGCCAATCTAAATGT-TCATTGGTTGCCATTGATGAAATCAGTTTGTTGACCTGCAAGAATTGTTGCTCATTTCATTGAA-ACCACTTCTCAGA
GGGGCGGGAGTACAAGGTGCGTGTGACTGGAGCCA---CCCACTCCGACTCTGCAGGTGTTG--CAAATGACGACCGATTGAAATG---GTCTCACGGCCAAAACCTCGTGTCCGACATCAACCCCTTC
TTCTCCAGTGTCTAGTTACATTGATGAGAACAGAA-ACATAAACTATGACCTAGGGGTTCT- GTTGGATAGCTCTAAATTAAAGAACGGAGAAAGAACAAACAAAGACATATTTCAGTTTTTTCTTAC
CAAACGTGATGGATA-GCCATTGGTATTCTATCTATT---TTAACTCTGTGCTTTACATTG--TTTATGATGGCCACAGCCTAAA-G---TACACACGGCTGTGACTTGATTCAAA-GAA-----
TGAGCAACTTGTCT-GATGACTGGGAAAGGAGGAC---CTGCAACCCTGACTTGGTCTCTG--TTAATGACGTCTCCCTCTAA-A---CCC-CATTAAGGACTGGGAGAGGCAGA-GCAAGCCTCAGAG
GATTACTGGCTGACTCTGGGGGCGGTTCTTCCA---TGATGGTGTTCCTCTAAATTGCA-CGGAGAAACACCTGATTTCAGGAAA-ATCCCCCTCAGATGGCGCTGGTCCCATCTCCGATGCCT
AGACCAGGCAAGACAACGTGAGC-GCGATGGCG---TGTACCCCAGGTCAAGGGGTGGTGTGCT-TCTATGAAGGAGGGGCCGAAG---CCCTTGTGGCGGGCCTCCCTGAGCCCCGTCTGTGGTGCAG
CACTTCAGAAGGCT-TCTGAATGGAACCCTCTT---GACA-TTGTGTTCTATA-ATATTG--T-CATGACAGTCACAGCATAAA-G---CGCAGACGGCTGTGACCTGATTTAGA-AAATTTTTAGA
```

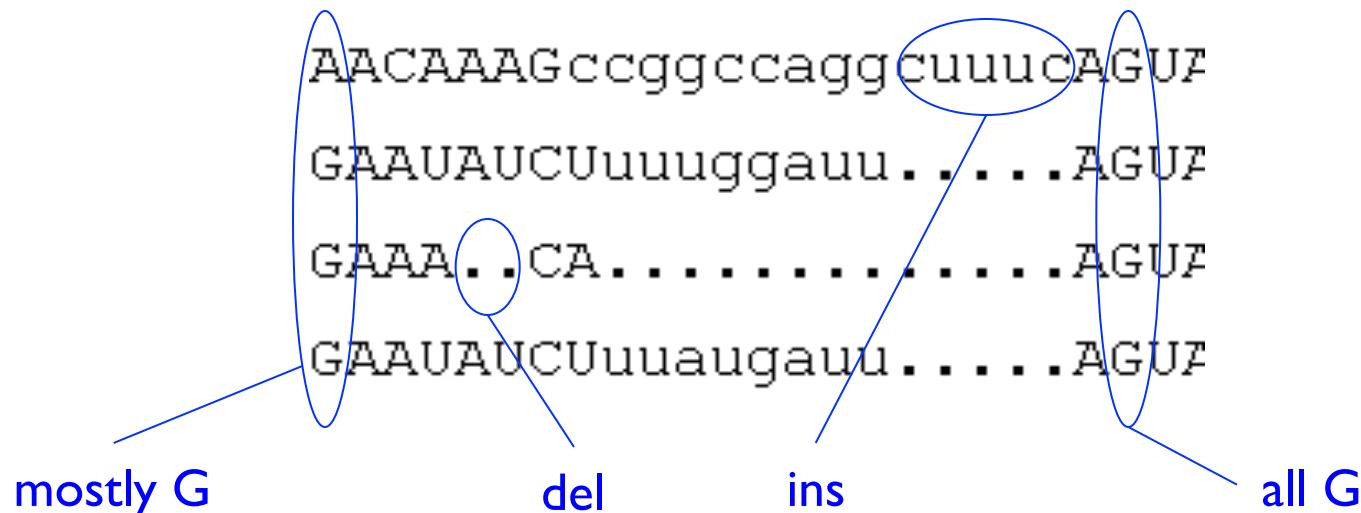
same-colored boxes *should* be aligned

# How to model an RNA “Motif”?

Conceptually, start with a profile HMM:

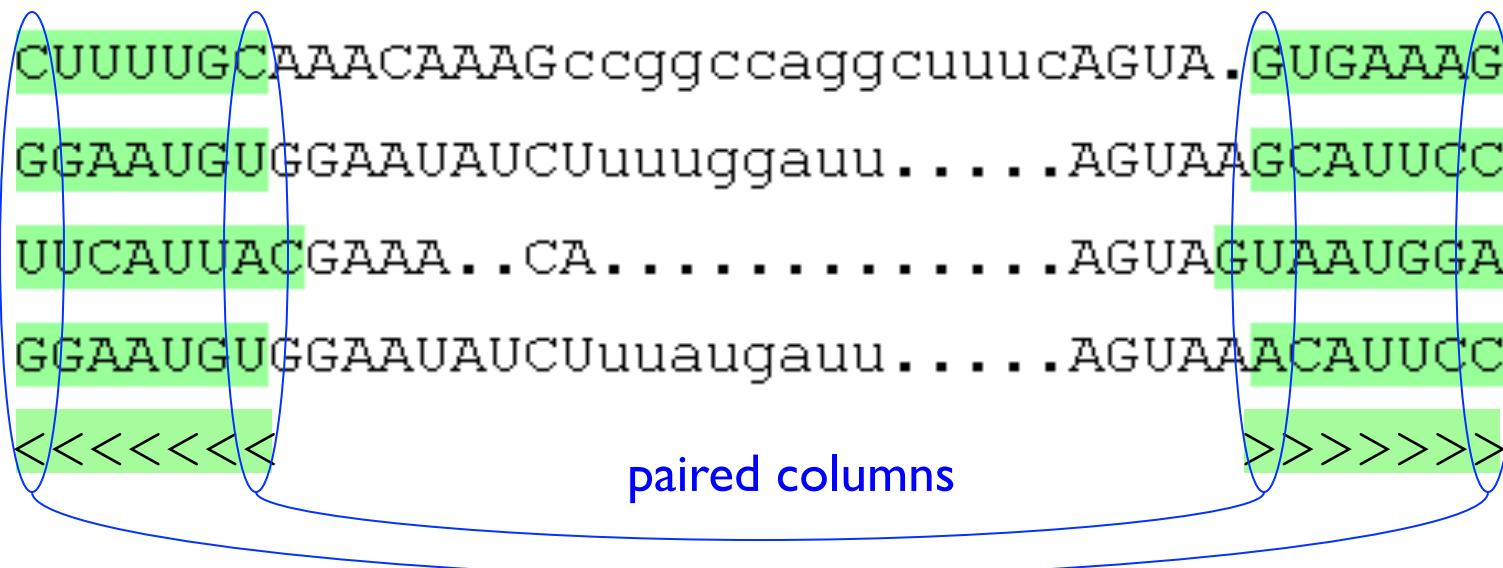
from a multiple alignment, estimate nucleotide/ insert/delete preferences for each position

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”?

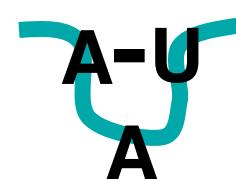
Add “column pairs” and pair emission probabilities for base-paired regions



# Covariance Models (specialized stochastic CFGs)

Sequences

CAG or AAU



CM

$$\begin{aligned} S_1 &\rightarrow cS_2g \mid aS_2u \\ S_2 &\rightarrow a \end{aligned}$$

Example  
parse of CAG

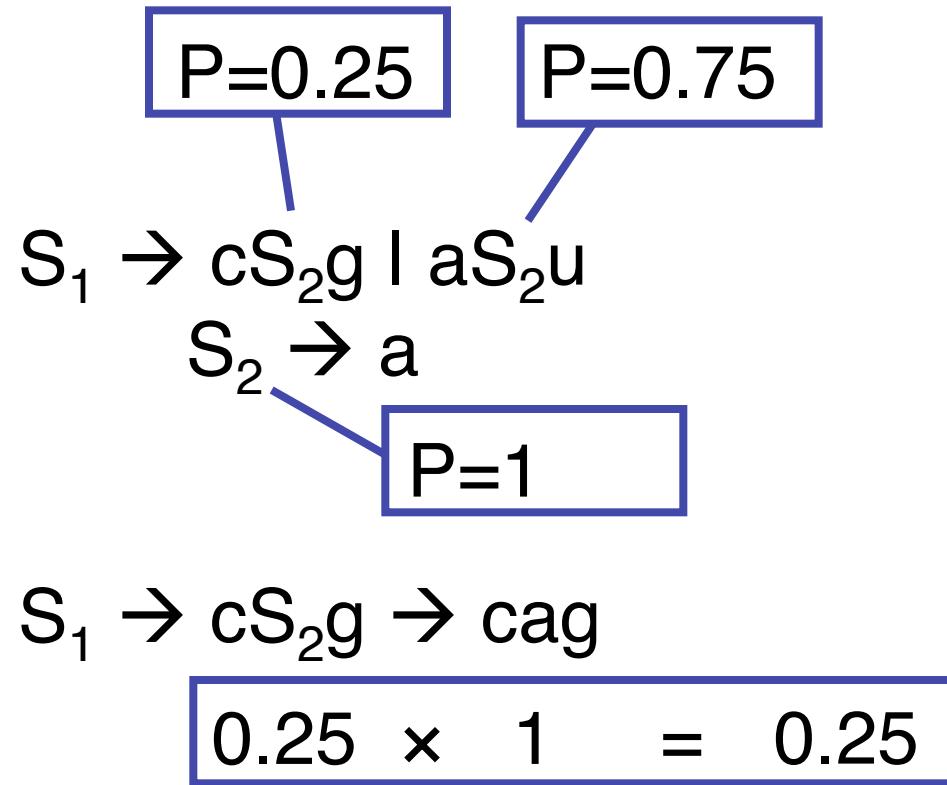
$$S_1 \rightarrow cS_2g \rightarrow cag$$

# *Stochastic context-free grammar*

CM

Example  
parse of CAG

Classification



Is  $\text{Pr}(\text{parse of CAG}) \geq \text{threshold}$   
(e.g., vs  $\text{Pr}(\text{CAG in null model})$ )

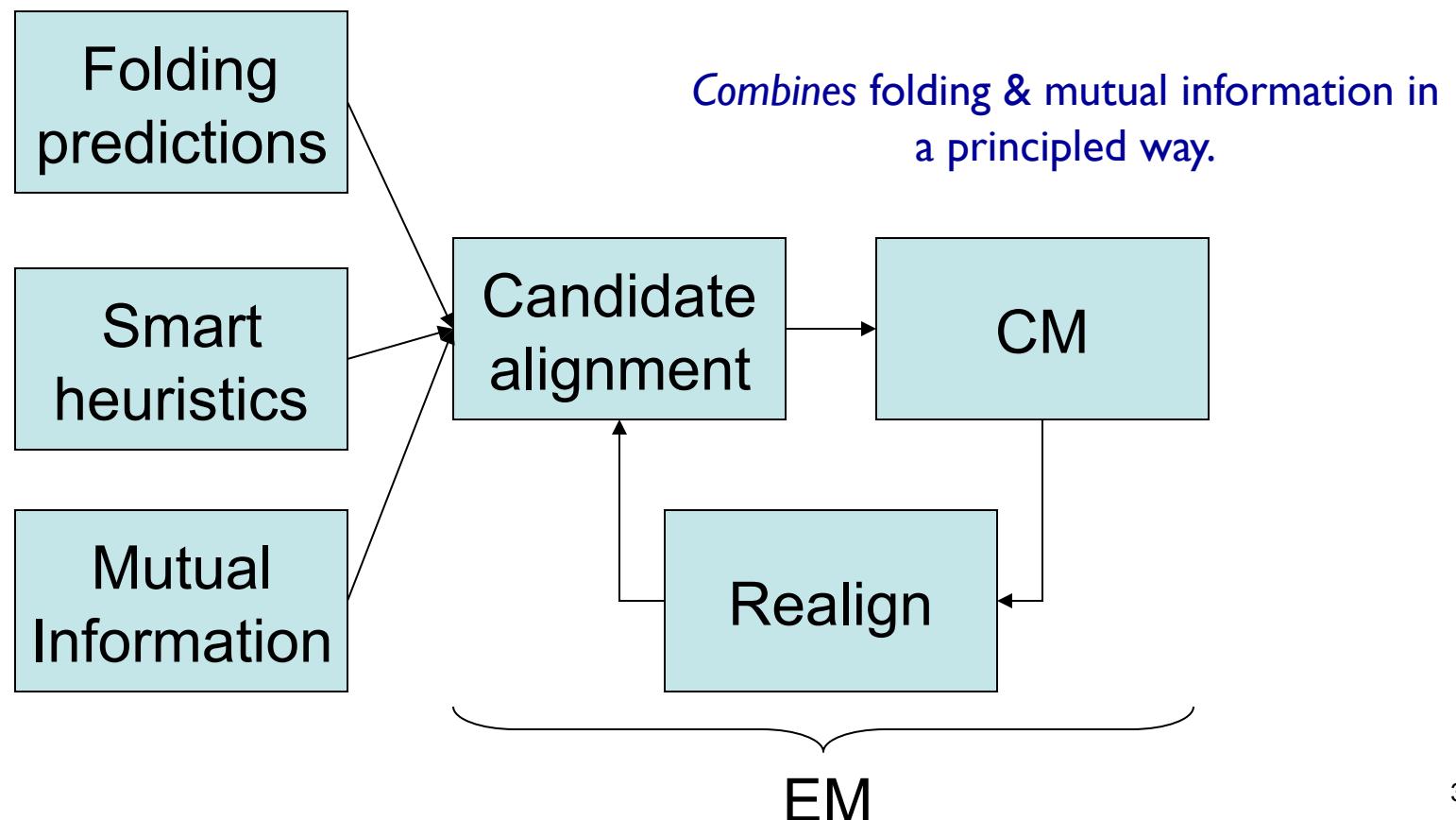
# Application: *cis*-regulatory ncRNA discovery in prokaryotes

Key issue is  
*exploiting prior knowledge*  
to focus on promising data

# CMFinder

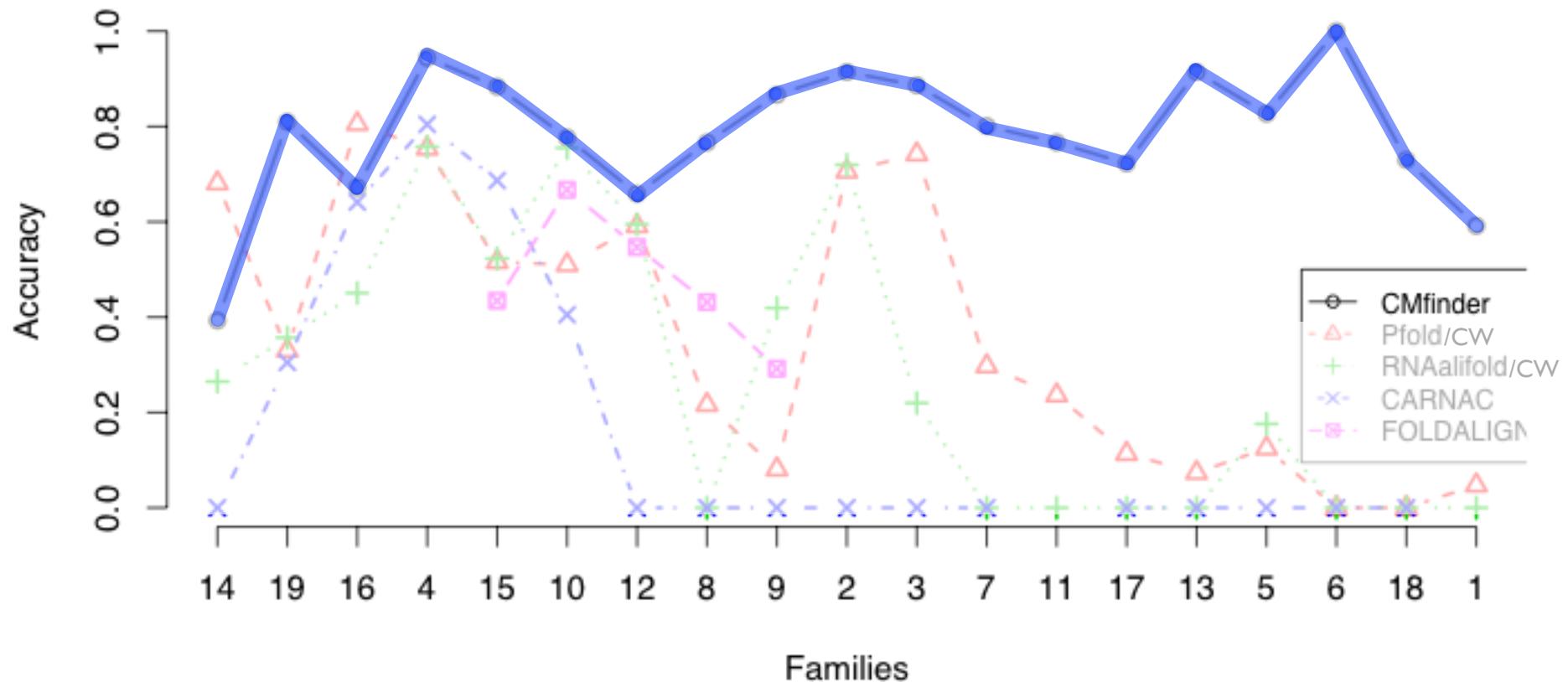
Simultaneous alignment, folding & motif description

Yao, Weinberg & Ruzzo, *Bioinformatics*, 2006

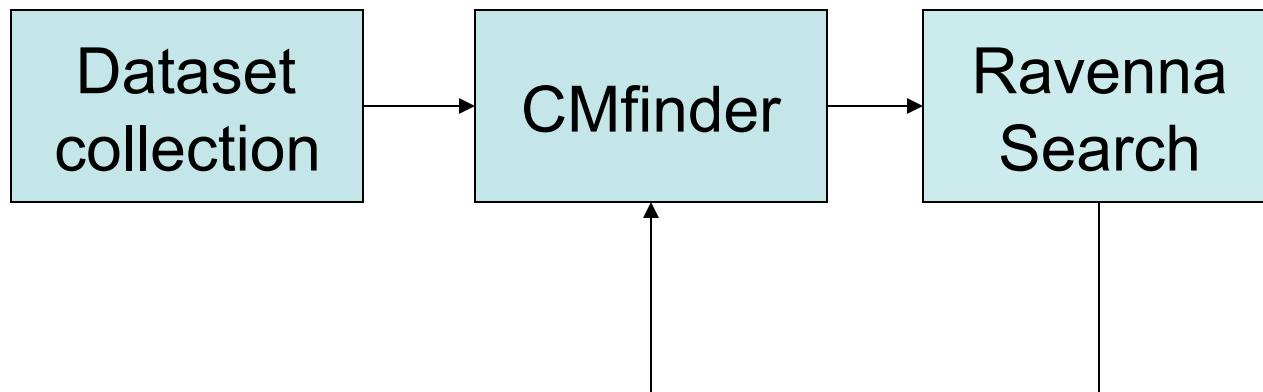


# CMfinder Accuracy

(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:

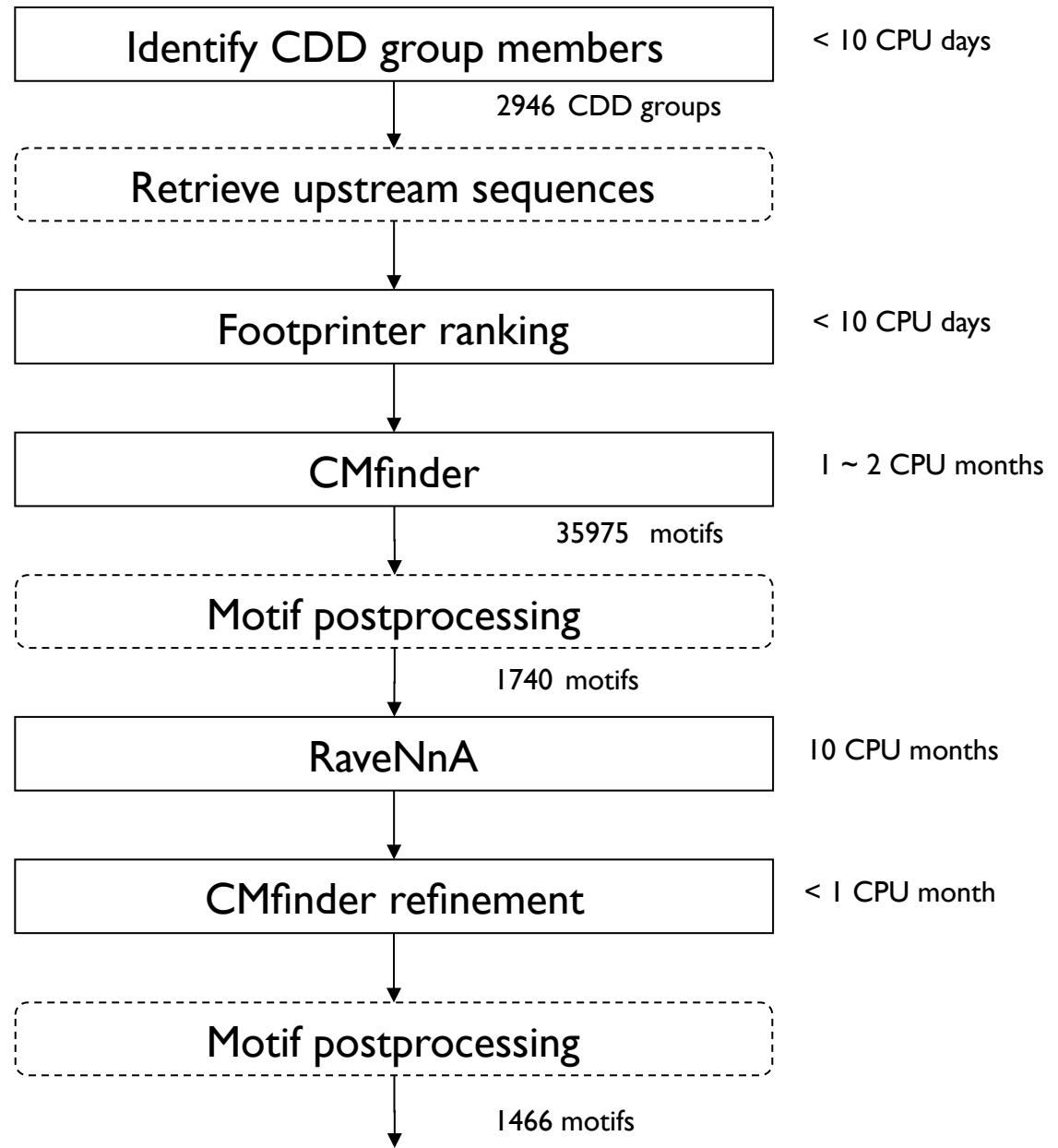
- Regulators near regulatees
- Get UTRs of homologs

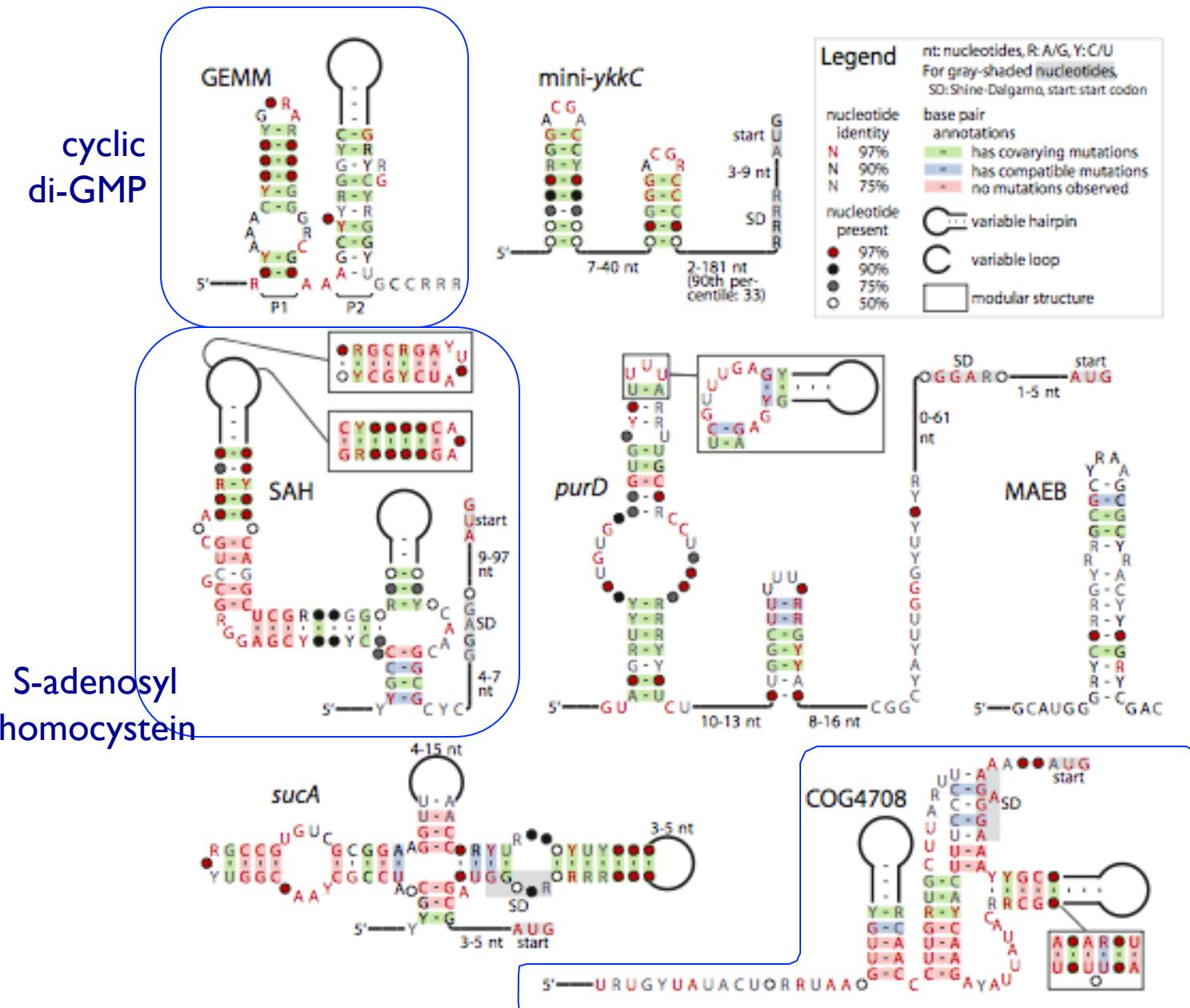
## Genome Scale Search:

- 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





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!