Genome 559
Wi 2009

RNA
Function, Search, Discovery

The Message

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

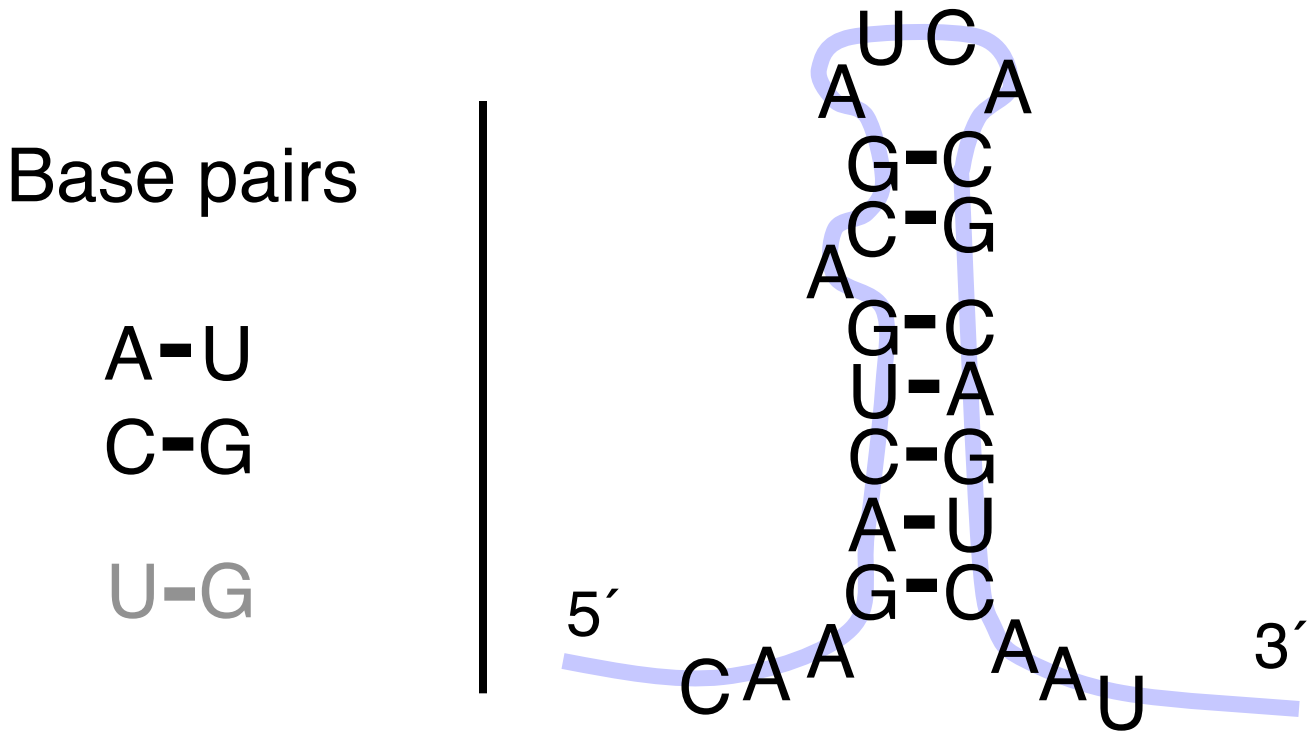
Functionally important, functionally diverse

Structurally complex

New tools required

alignment, discovery, search, scoring, etc.

RNA Secondary Structure:
RNA makes helices too



Usually *single* stranded

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."

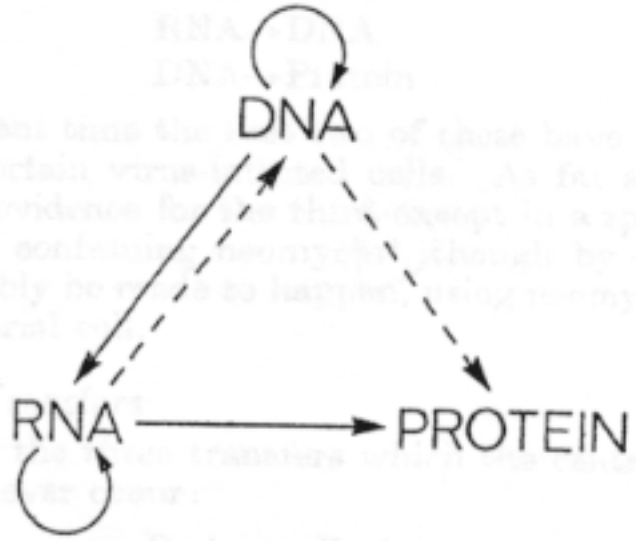
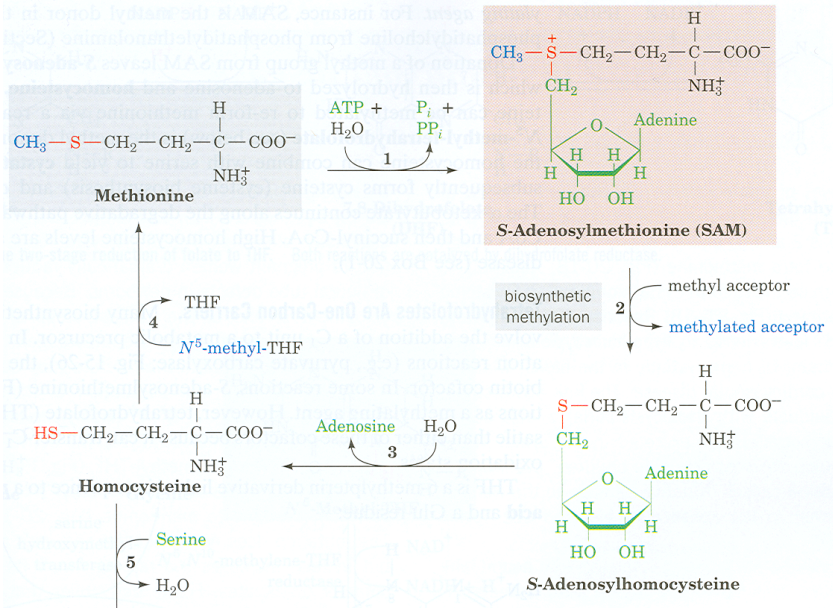
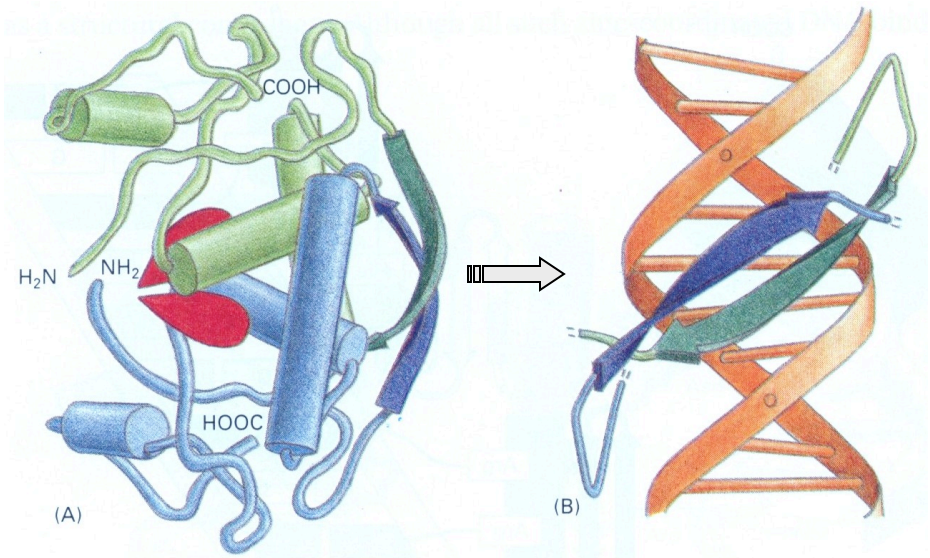


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

Proteins catalyze & regulate biochemistry

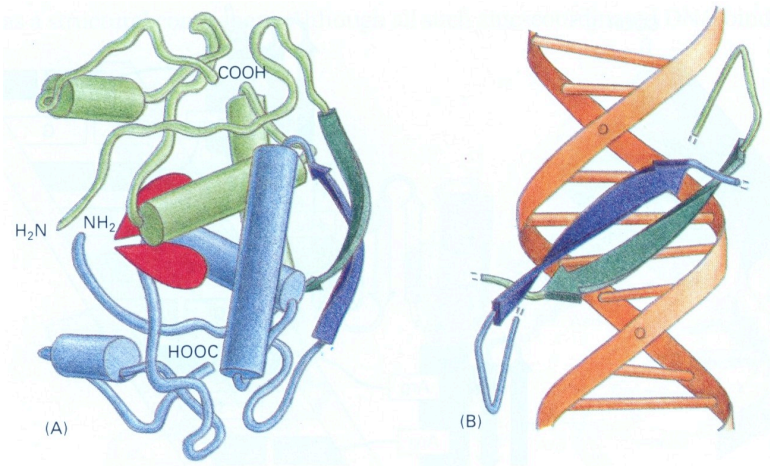


SAM



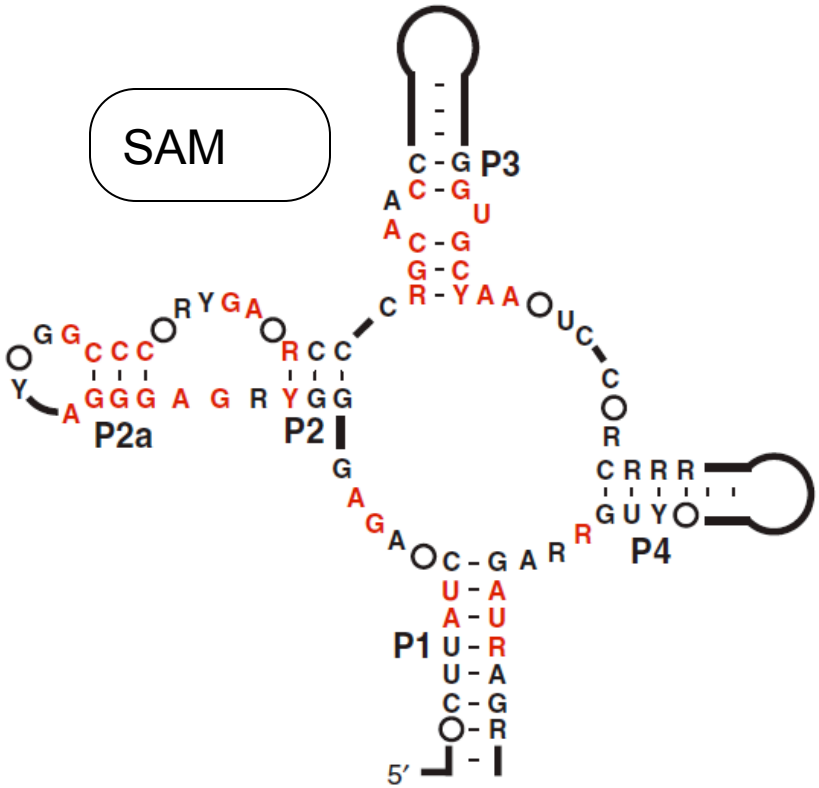
The Met Repressor

Alberts, et al, 3e.



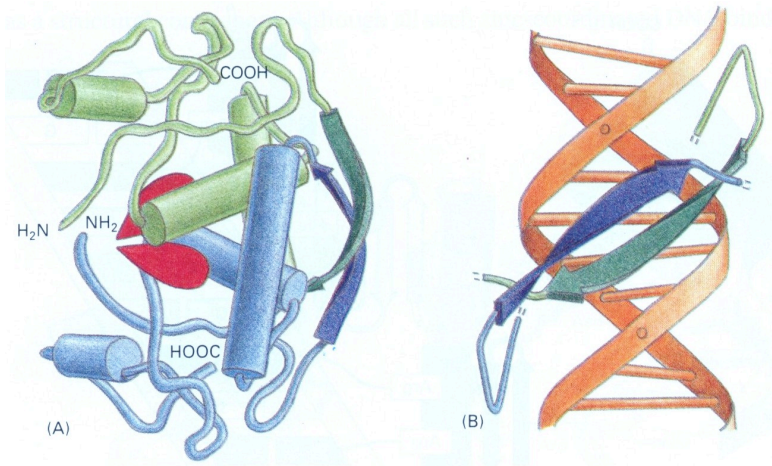
The protein way

Riboswitch alternative



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

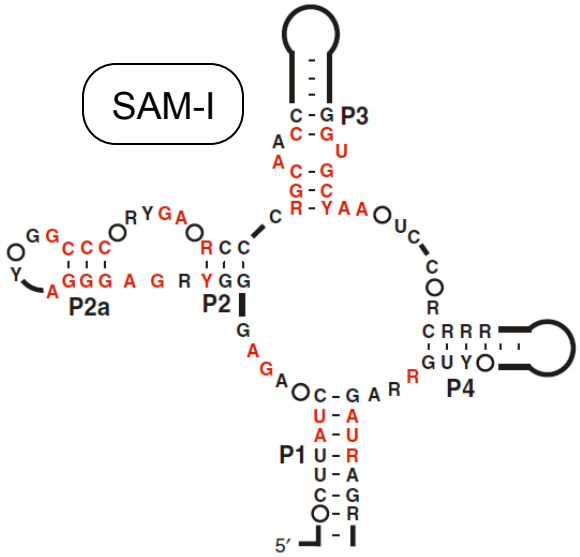
Alberts, et al, 3e.



The protein way

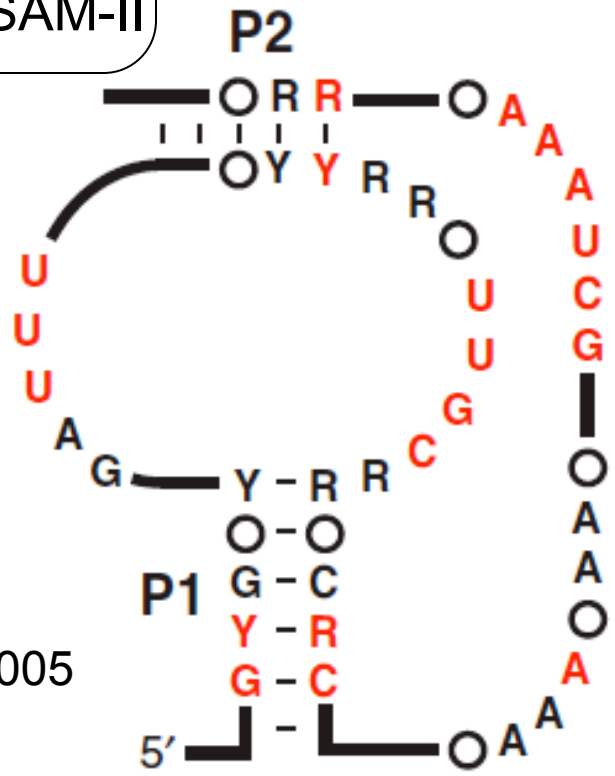
Riboswitch alternatives

SAM-I



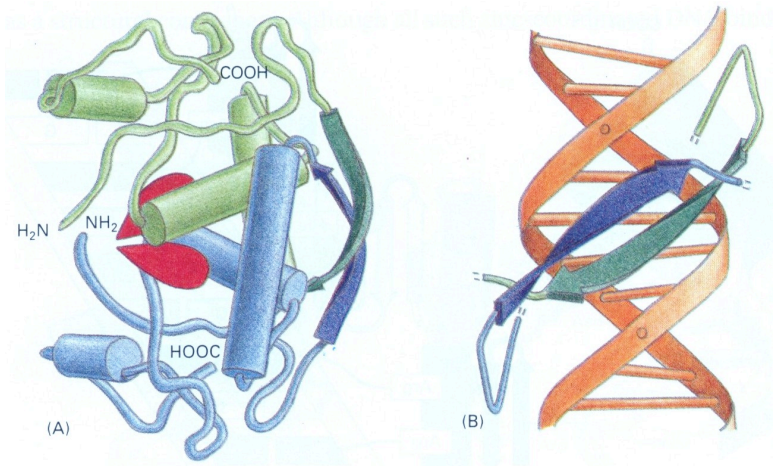
Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



Corbino et al., Genome Biol. 2005

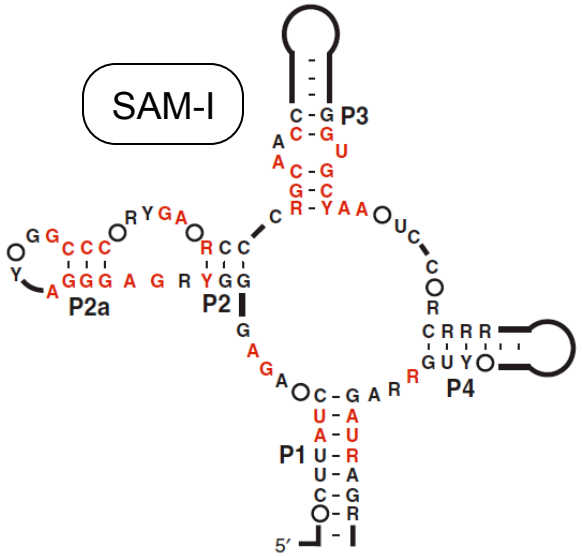
Alberts, et al, 3e.



The protein way

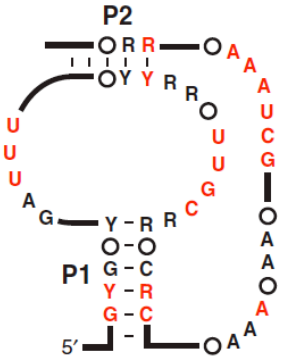
Riboswitch alternatives

SAM-I



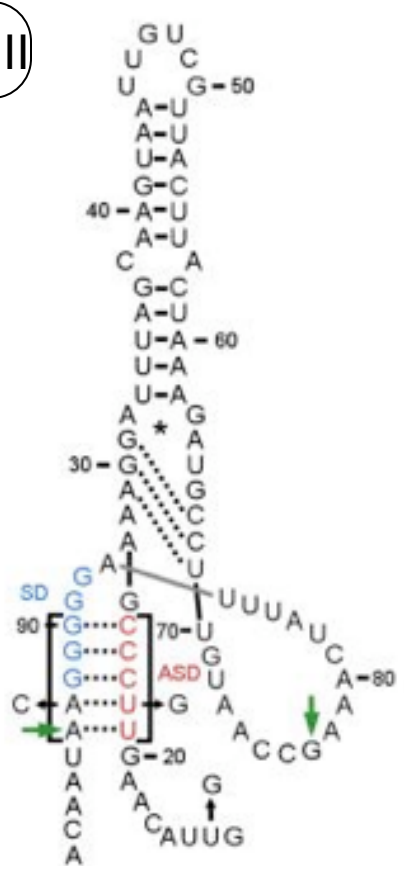
Grundy, Epshtein, Winkler et al., 1998, 2003

SAM-II



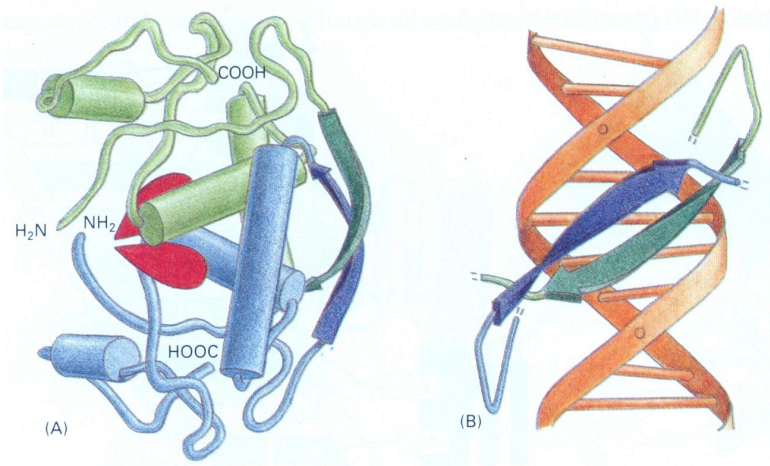
Corbino et al., Genome Biol. 2005

SAM-III



Fuchs et al., NSMB 2006

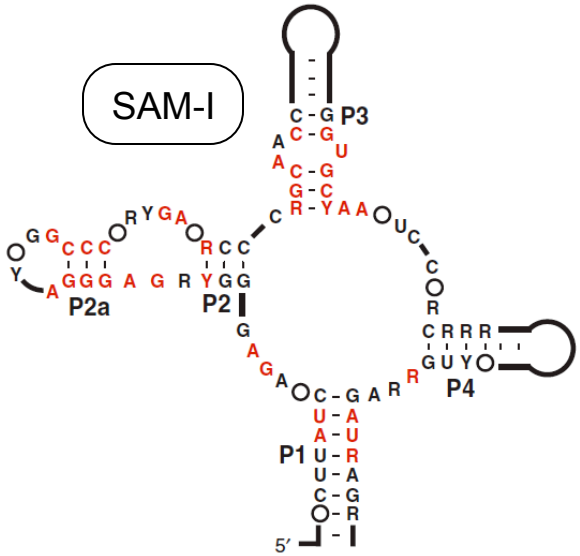
Alberts, et al, 3e.



The protein way

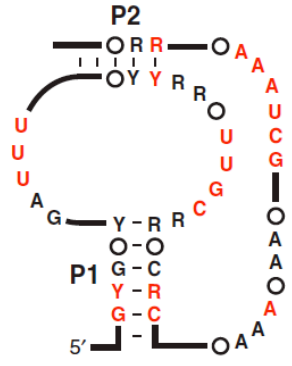
Riboswitch alternatives

SAM-I



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

SAM-II

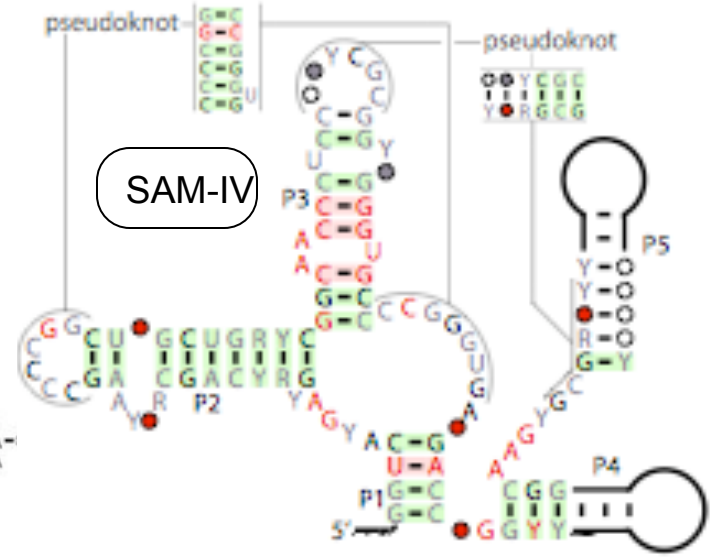


Corbino et al., Genome Biol. 2005

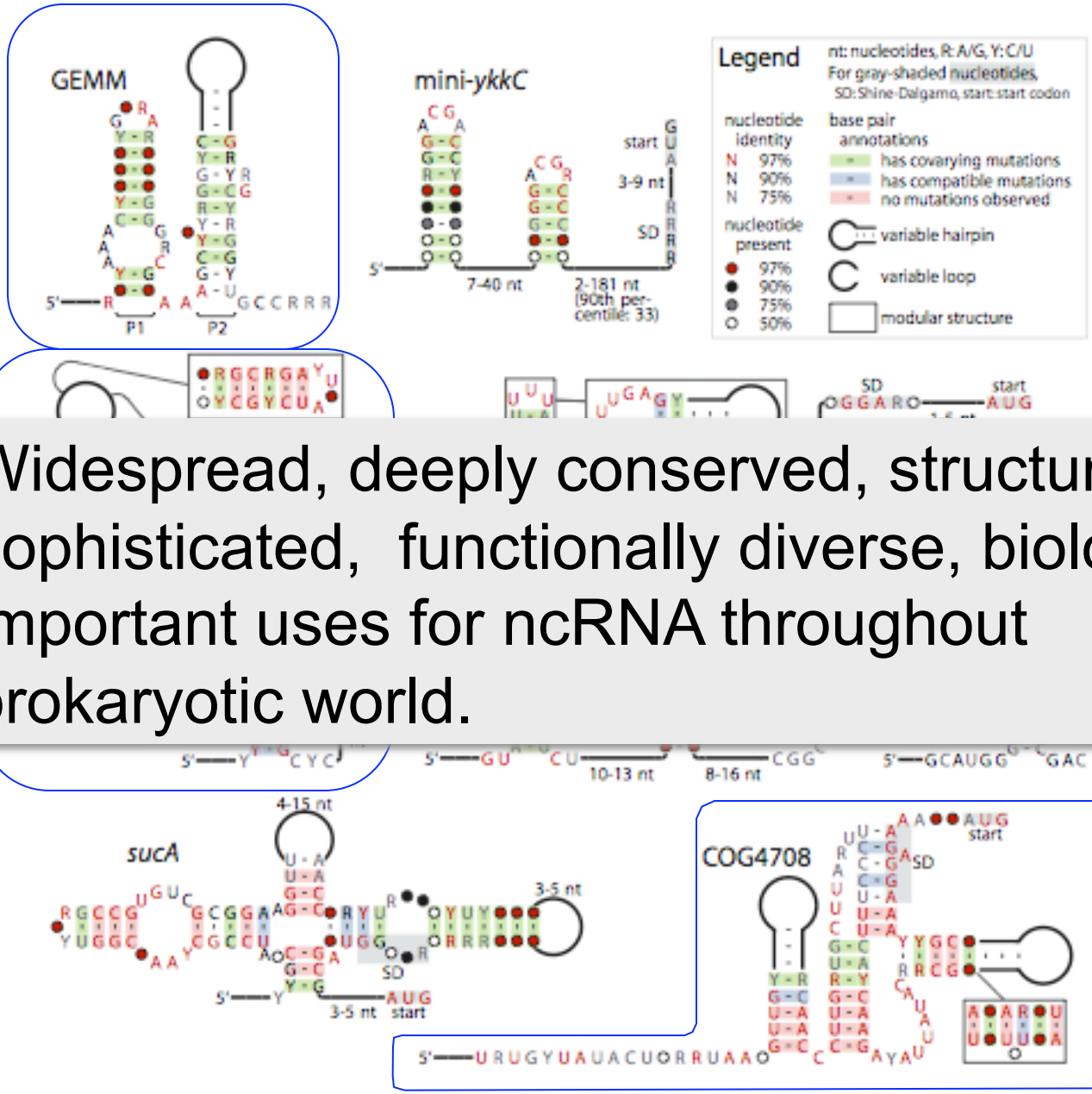
SAM-III

Fuchs et al., NSMB 2006

SAM-IV



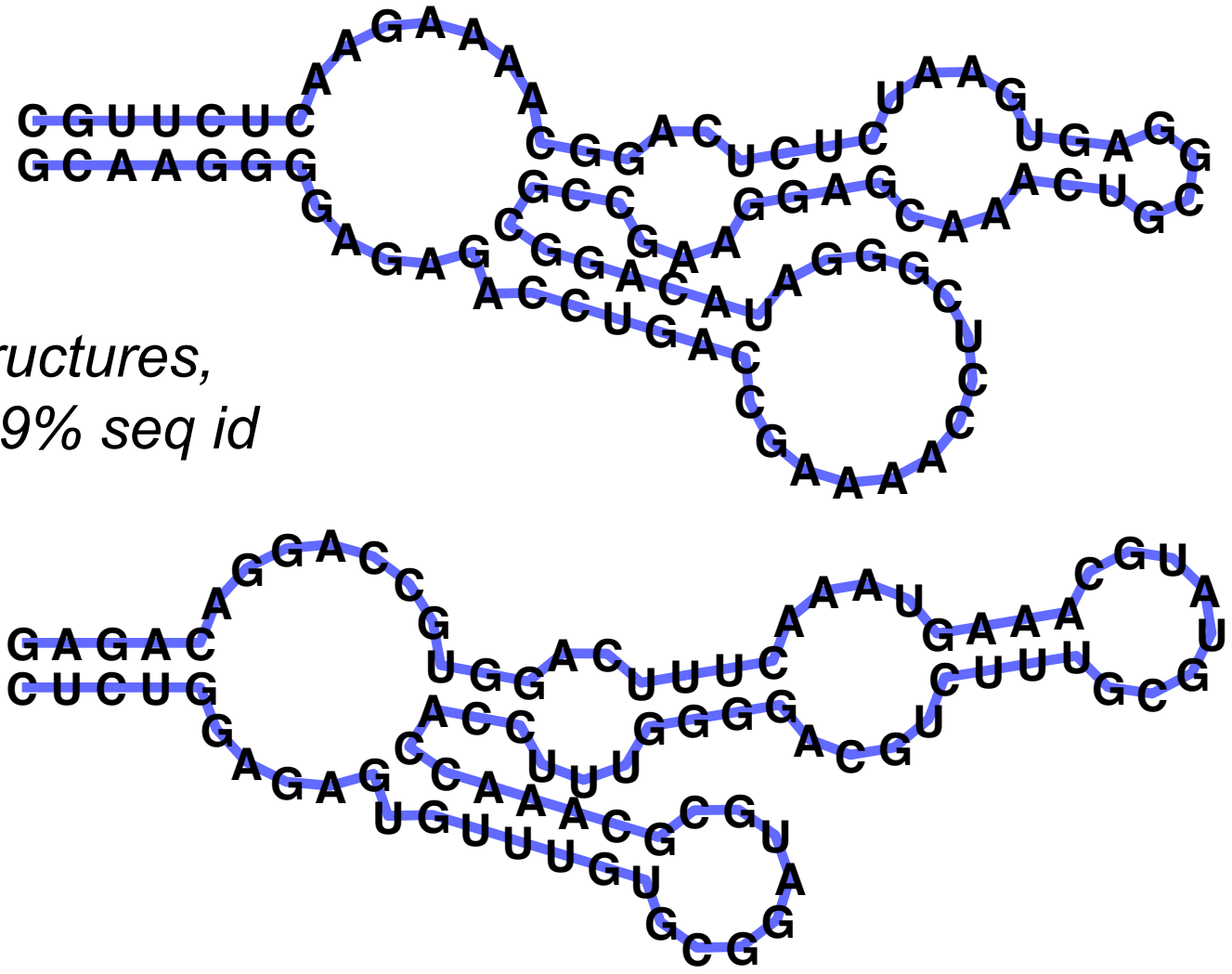
Weinberg et al., RNA 2008



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.

Why is RNA hard to deal with?



Similar structures,
but only 29% seq id

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

Motif Description & Inference

RNA Motif Models

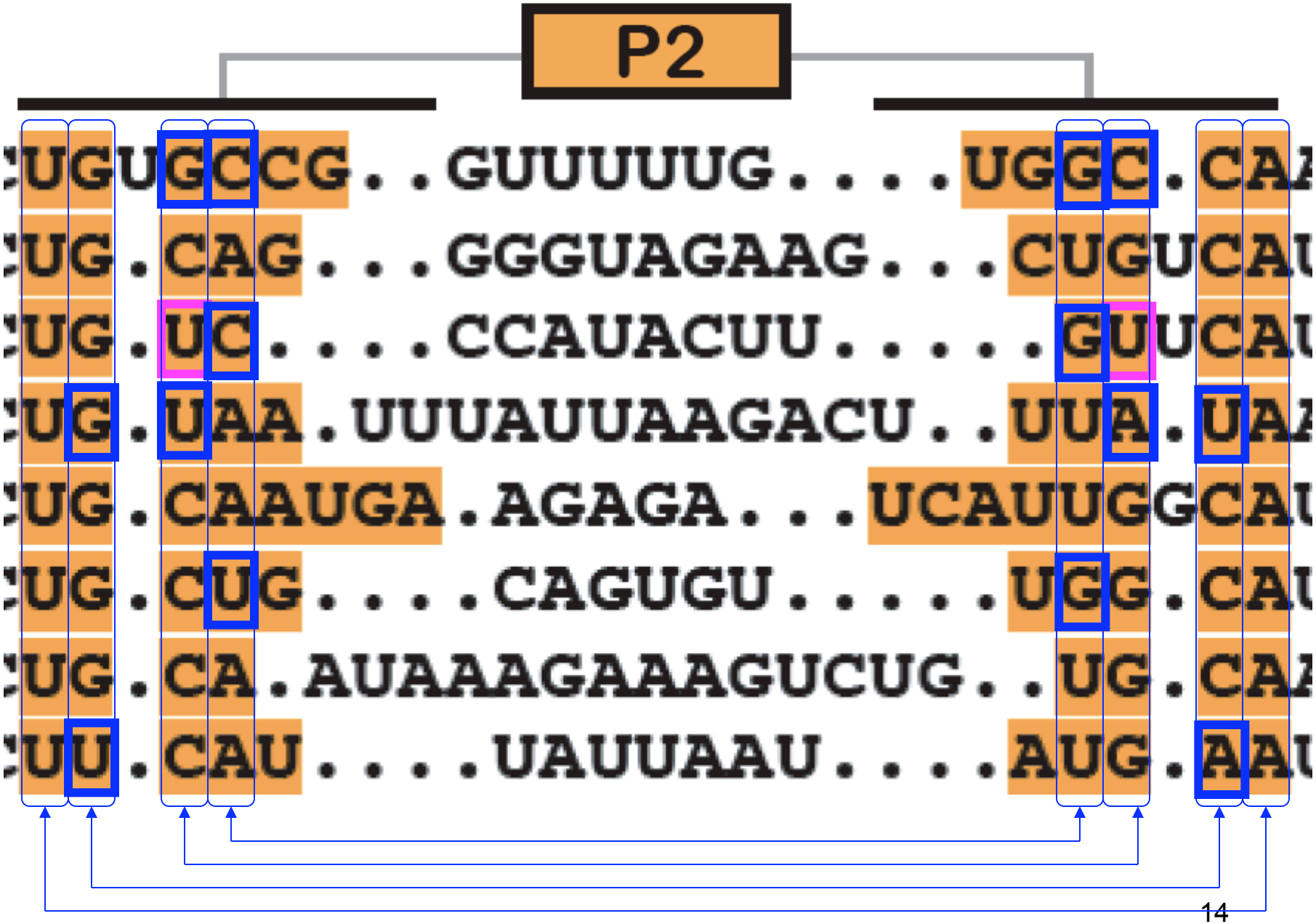
"Covariance Models" (Eddy & Durbin 1994)

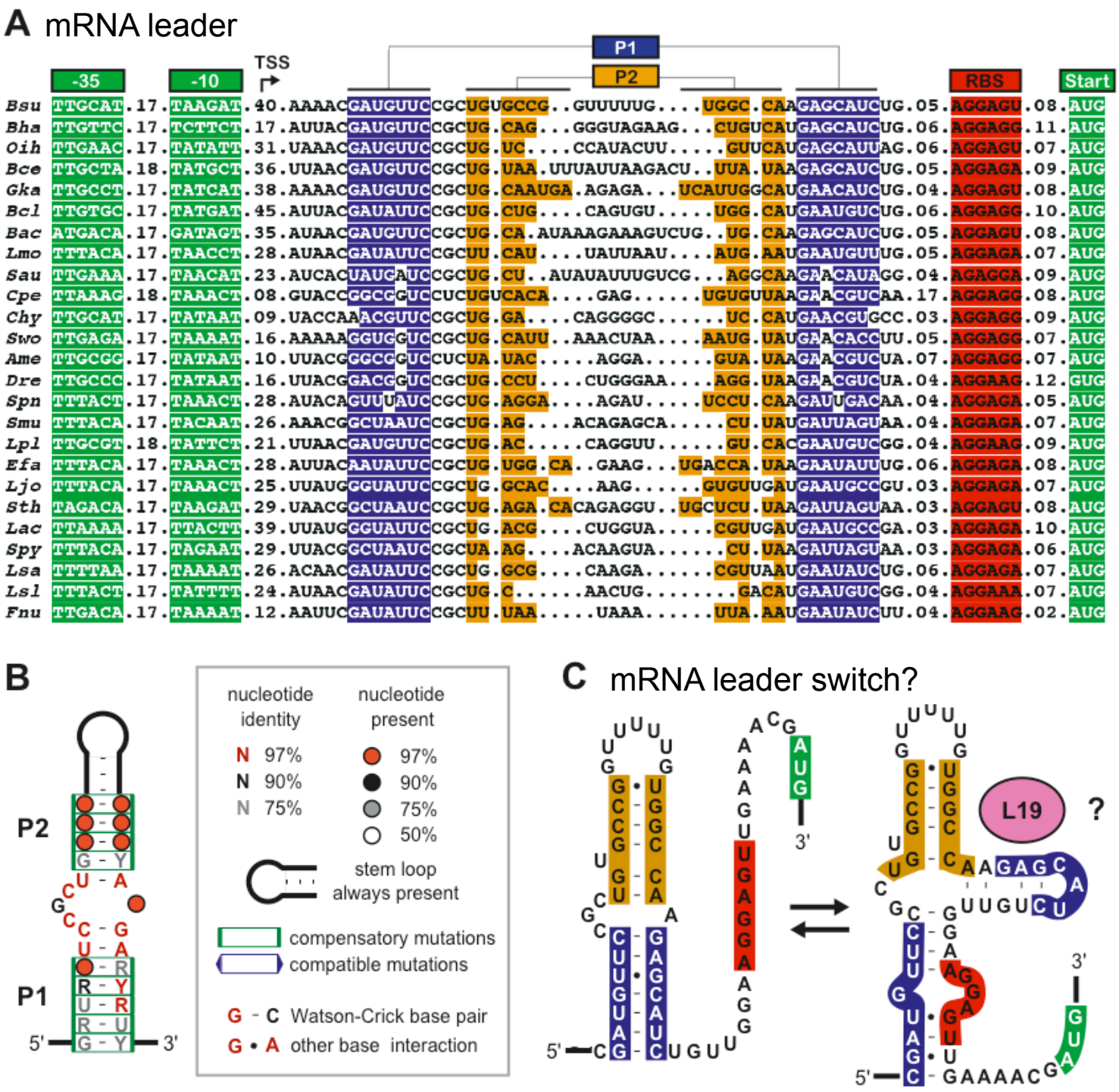
- aka profile stochastic context-free grammars
- aka hidden Markov models on steroids

Model position-specific nucleotide preferences *and* base-pair preferences

Pro: accurate

Con: model building hard, search sloooow


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


A mRNA leader
P1
P2
TSS
-35
-10
RBS
Start
Bsu TTGCAT.17.TAAGAT.40.AAAACGAUGUUCCGCUGUGCCG..GUUUUUG....UGGC.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Bha TTGTTC.17.TCTTCT.17.AUUACGAUGUUCCGCUG.CAG...GGGUAGAAG...CUGUCAUGAGCAUCUG.06.AGGAGG.11.AUG
Oih TTGAAC.17.TATATT.31.UAAACGAUGUUCCGCUG.UC....CCAUACUU.....GUUCAUGAGCAUUAG.06.AGGAGU.07.AUG
Bce TTGCTA.18.TATGCT.36.UUAACGAUGUUCCGCUG.UAA.UUUAUUAAGACU..UUA.UAAGAGCAUCUG.05.AGGAGA.09.AUG
Gka TTGCCT.17.TATCAT.38.AAAACGAUGUUCCGCUG.CAAUGA.AGAGA...UCAUUGGCAUGAACAUCUG.04.AGGAGU.08.AUG
Bcl TTGTGC.17.TATGAT.45.AUUACGAUAUUCCGCUG.CUG....CAGUGU.....UGG.CAUGAAUGUCUG.06.AGGAGG.10.AUG
Bac ATGACA.17.GATAGT.35.AUAACGAUGUUCCGCUG.CA.AUAAAGAAAGUCUG..UG.CAAGAGCAUCUG.05.AGGAGU.08.AUG
Lmo TTTACA.17.TAACCT.28.AUAACGAUAUUCCGCUU.CAU....UAUUAAU....AUG.AAUGAAUGUUUG.05.AGGAGA.07.AUG
Sau TTGAAA.17.TAACAT.23.AUCACUAUGAUCCGCUG.CU..AUAUAUUUGUCG...AGGCAAGAACAUAGG.04.AGAGGA.09.AUG
Cpe TTAAAG.18.TAAACT.08.GUACCGGCGGUCCUCUGUCACA....GAG......UGUGUUAAGAACGUCAA.17.AGGAGG.08.AUG
Chy TTGCAT.17.TATAAT.09.UACCAAACGUUCCGCUG.GA....CAGGGGC.....UC.CAUGAACGUGCC.03.AGGAGG.09.AUG
Swo TTGAGA.17.TAAAAT.16.AAAAAGGUGGUCCGCUG.CAUU..AAACUAA....AAUG.UAUGAACACCUU.05.AGGAGG.07.AUG
Ame TTGCGG.17.TATAAT.10.UUACGGGCGGUCCUCUA.UAC.....AGGA.....GUA.UAAGAACGUCUA.07.AGGAGG.07.AUG
Dre TTGCCC.17.TATAAT.16.UUACGGACGGUCCGCUG.CCU....CUGGGAA....AGG.UAAGAACGUCUA.04.AGGAAG.12.GUG
Spn TTTACT.17.TAAACT.28.AUACAGUUUAUCCGCUG.AGGA....AGAU.....UCCU.CAAGAUUGACAA.04.AGGAGA.05.AUG
Smu TTTACA.17.TACAAT.26.AAACGGCUAAUCCGCUG.AG....ACAGAGCA.....CU.UAUGAUUAGUAA.04.AGGAGA.07.AUG
Lpl TTGCGT.18.TATTCT.21.UUAACGAUGUUCCGCUG.AC.....CAGGUU.....GU.CACGAAUGUCGG.04.AGGAAG.09.AUG
Efa TTTACA.17.TAAACT.28.AUUACAAUAUUCCGCUG.UGG.CA..GAAG...UGACCA.UAAGAAUAUUUG.06.AGGAGA.08.AUG
Ljo TTTACA.17.TAAACT.25.UUAUGGGUAUUCCGCUG.GCAC....AAG.....GUGUUGAUGAAUGCCGU.03.AGGAGA.07.AUG
Sth TAGACA.17.TAAGAT.29.UAACGGCUAAUCCGCUG.AGA.CACAGAGGU..UGCUCU.UAAGAUUAGUAA.03.AGGAGU.08.AUG
Lac TTAAAA.17.TTACTT.39.UUAUGGGUAUUCCGCUG.ACG....CUGGUA.....CGUUGAUGAAUGCCGA.03.AGGAGA.10.AUG
Spy TTTACA.17.TAGAAT.29.UUACGGCUAAUCCGCUA.AG....ACAAGUA.....CU.UAAGAUUAGUAA.03.AGGAGA.06.AUG
Lsa TTTTAA.17.TAAAAT.26.ACAACGAUAUUCCGCUG.GCG....CAAGA.....CGUUAAUGAAUAUCUG.06.AGGAGA.07.AUG
Lsl TTTACT.17.TATTTT.24.AUAACGAUAUUCCGCUG.C.....AACUG.....GACAUGAAUGUCGG.04.AGGAAA.07.AUG
Fnu TTGACA.17.TAAAAT.12.AAUUCGAUAUUCCGCUU.UAA.....UAAA.....UUA.AAUGAAUAUCUU.04.AGGAAG.02.AUG
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'

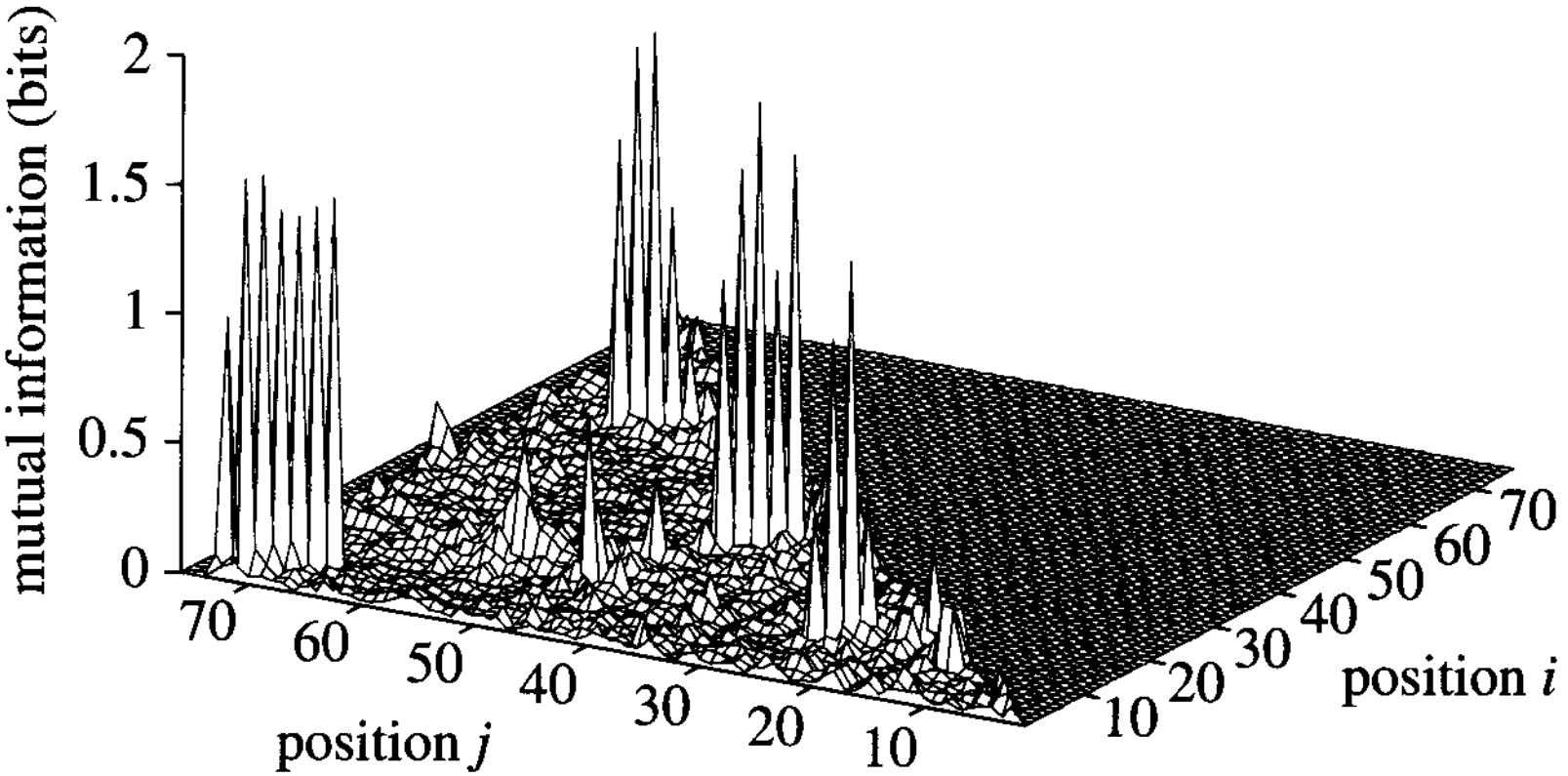
Mutual Information

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

Max when *no* seq conservation but perfect pairing;

Expected score gain from modeling i & j as paired.

Given columns, finding optimal pairing *without pseudoknots* can be done by dynamic programming



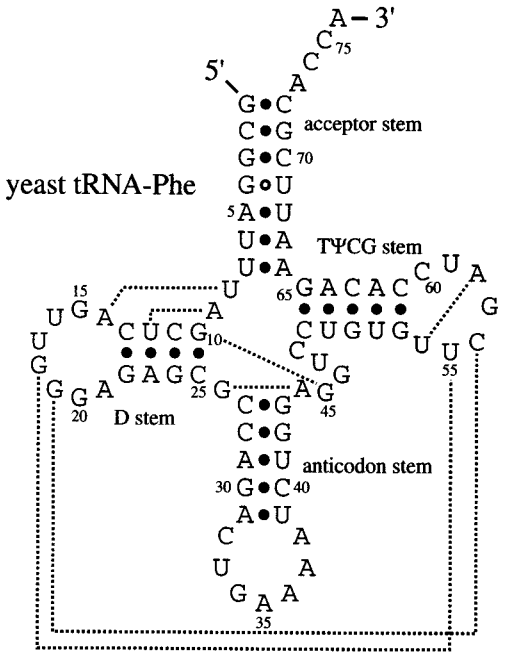


Figure 10.6 *A mutual information plot of a tRNA alignment (top) shows four strong diagonals of covarying positions, corresponding to the four stems of the tRNA cloverleaf structure (bottom; the secondary structure of yeast phenylalanine tRNA is shown). Dashed lines indicate some of the additional tertiary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

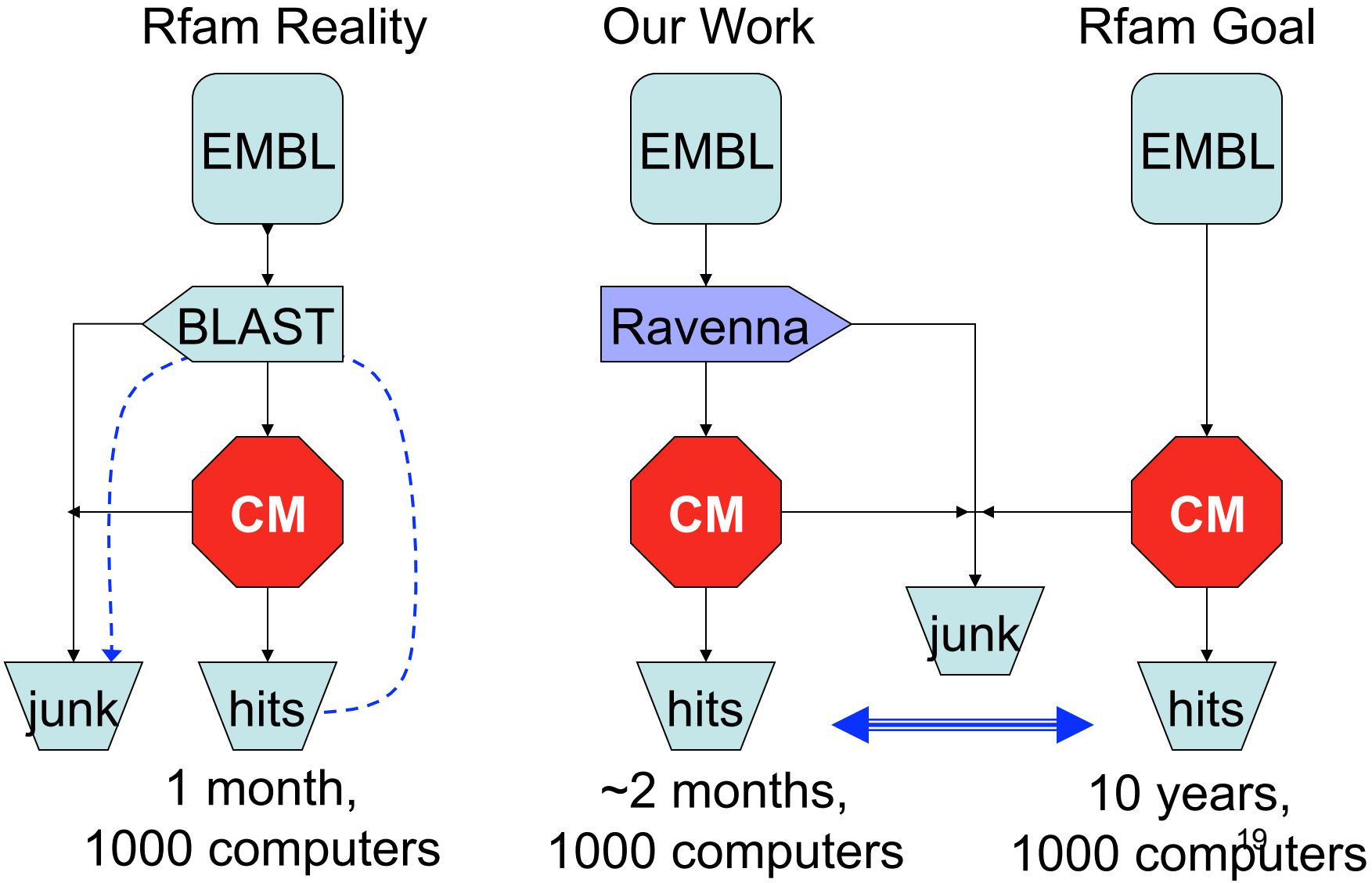
Fast Motif Search

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

Weinberg & Ruzzo

Recomb '04, ISMB '04, Bioinformatics '06

CM's are good, but slow



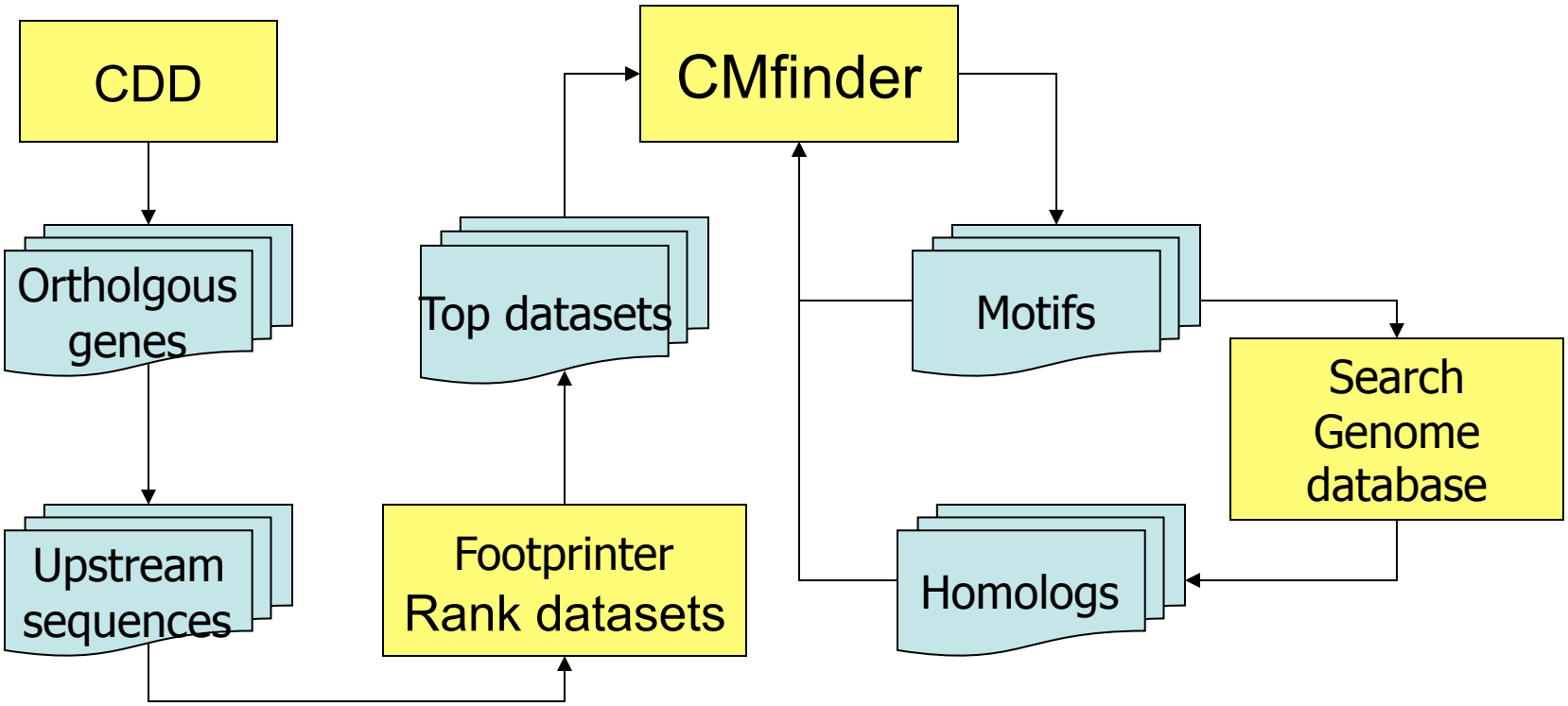
Results: New ncRNA's?

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

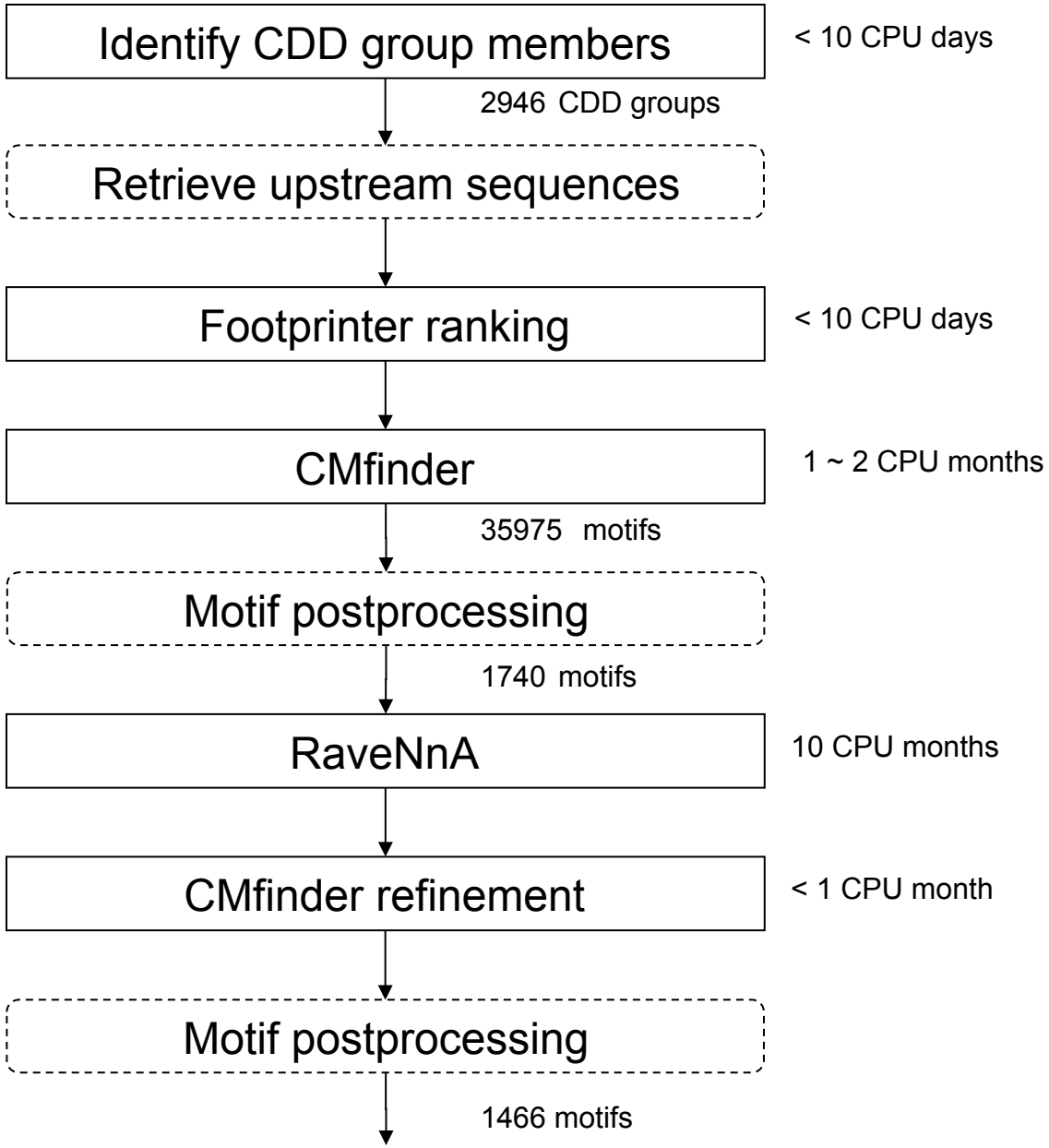
Motif Discovery In Prokaryotes

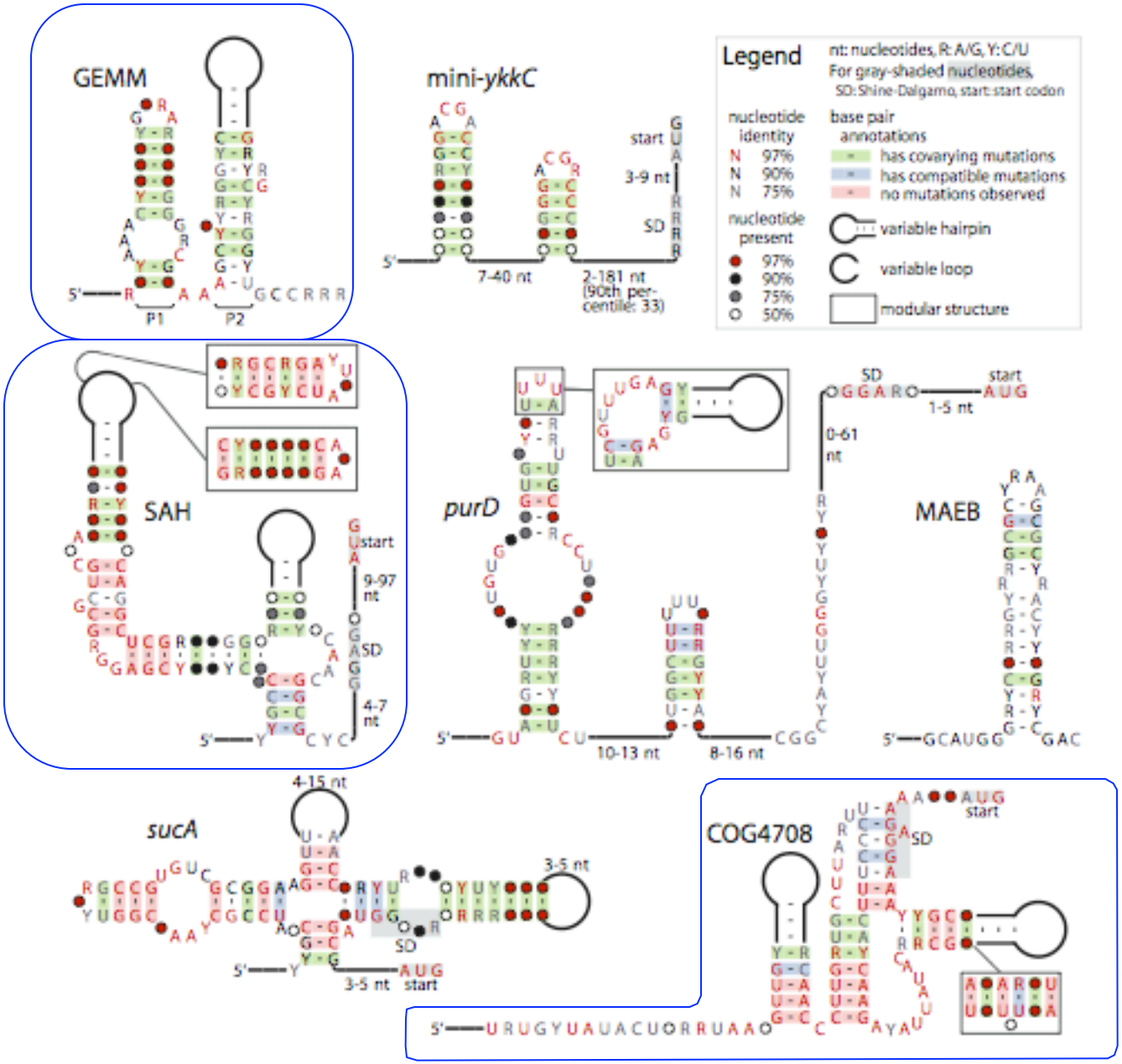
(Vertebrates too, but no time today…
see, e.g., Torarinsson, et al.
Genome Research, Jan 2008)

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.


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



boxed = confirmed riboswitch (+2 more)

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

Summary

ncRNA - apparently widespread, much interest

Covariance Models - powerful but expensive

RaveNnA filtering - search ~100x faster with no/little loss

CMfinder - CM-based motif discovery in unaligned sequences

Pipelines integrating comp and bio for ncRNA discovery

Many vertebrate ncRNAs? *structural,* not seq conservation; functional significance unclear

BIG CPU demands…

Still need for further methods development & application

Course Wrap Up

Modern biology is suddenly very data-rich

Mathematical & computational tools needed

We showed: sequence modeling, alignment & search, phylogeny, linkage/association mapping, some data bases

Python is a good tool for doing much of this

There's lots more!

Check out, e.g., GENOME 540/1, CSE 527…

We hope you enjoyed it.

Thanks!