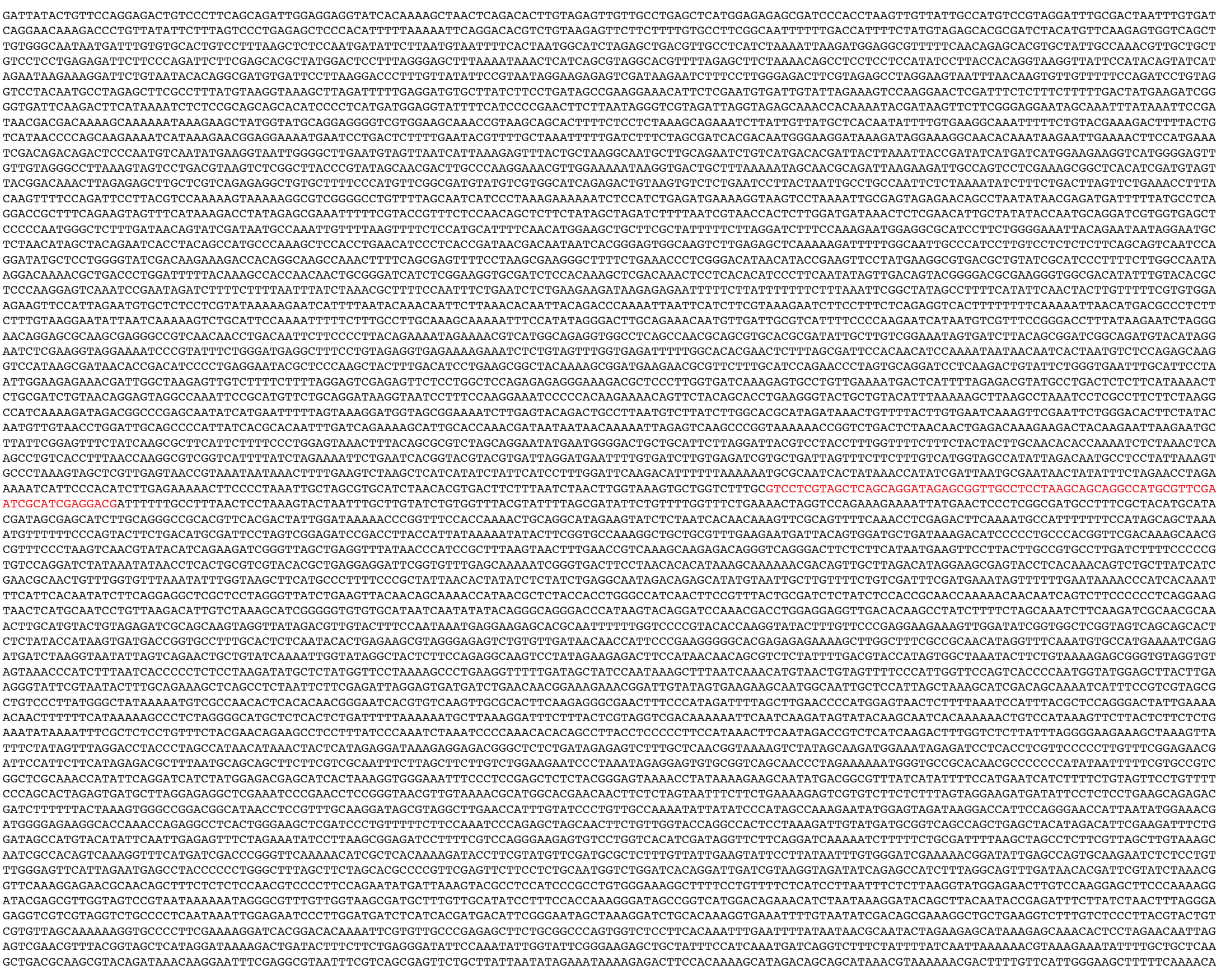
Structural RNA Homology Search and Alignment Using Covariance Models

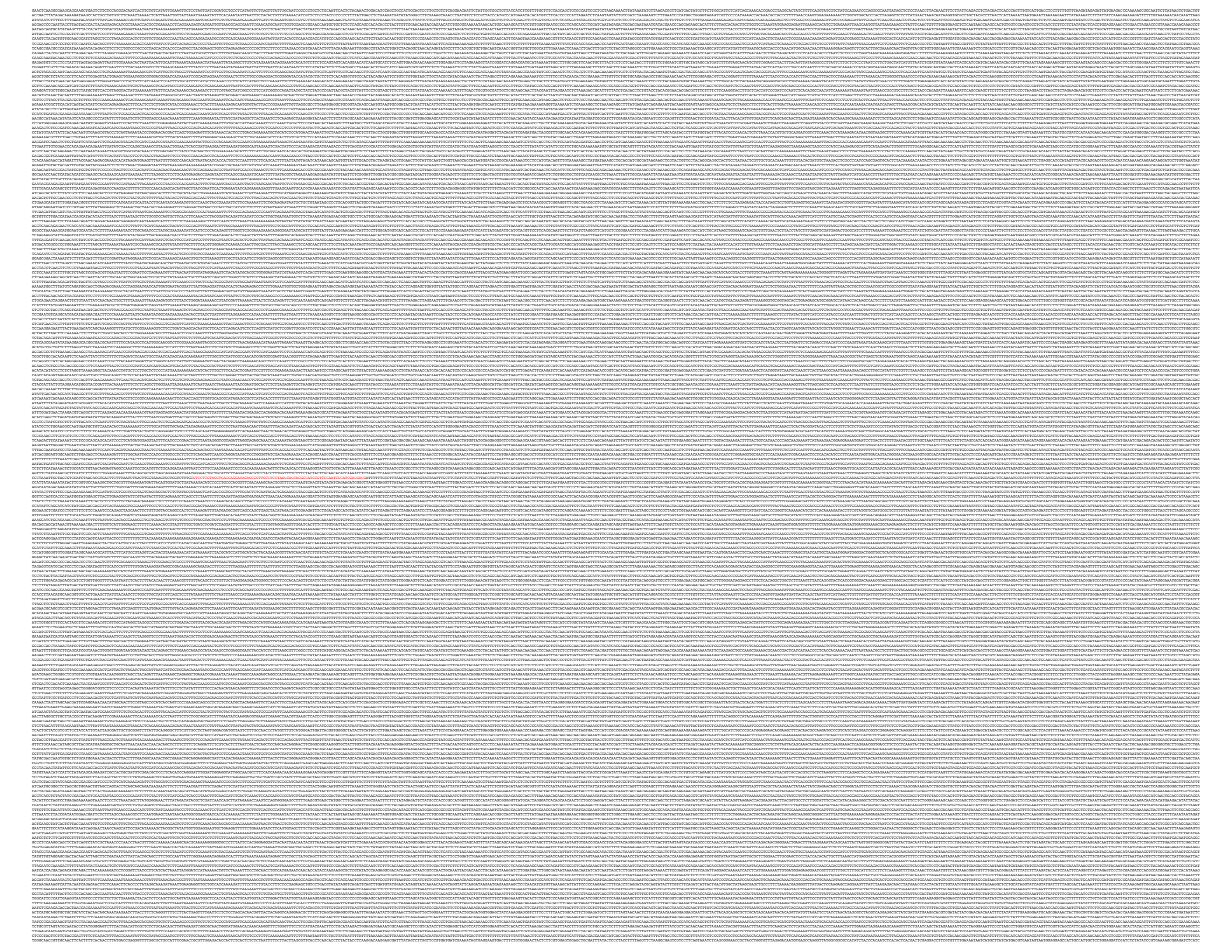
Eric Nawrocki

08.28.09

GTAACCGTAAATAATAAACTTTTGAAGTCTAAGCTCATCATATCTATTCA
TCCTTTGGATTCAAGACATTTTTTAAAAAATGCGCAATCACTATAAACCA
TATCGATTAATGCGAATAACTATATTTCTAGAACCTAGAAAAATCATTCC
CACATCTTGAGAAAAACTTCCCCTAAATTGCTAGCGTGCATCTAACACGT
GACTTCTTTAATCTAACTTGGTAAAGTGCTGGTCTTTGCGTCCTCGTAGC
TCAGCAGGATAGAGCGGTTGCCTCCTAAGCAGCAGGCCATGCGTTCGAAT
CGCATCGAGGACGATTTTTTGCCTTTAACTCCTAAAGTACTAATTTGCTT
GTATCTGTGGTTTACGTATTTTAGCGATATTCTGTTTTGGTTTCTGAAAA
CTAGGTCCAGAAAGAAAATTATGAACTCCCTCGGCGATGCCTTTCGCTAC
ATGCATACGATAGCGAGCATCTTGCAGGGCCGCACGTTCACGACTATTGG
ATAAAAACCCGGTTTCCACCAAAACTGCAGGCATAGAAGTATCTCTAATC
ACAACAAAGTTCGCAGTTTTCAAACCTCGAGACTTCAAAATGCCATTTTT
TTCCATAGCAGCTAAAATGTTTTTTCCCAGTACTTCTGACATGCGATTCC
TAGTCGGAGATCCGACCTTACCATTATAAAAATATACTTCGGTGCCAAAG
GCTGCTGCGTTTGAAGAATGATTACAGTGGATGCTGATAAAGACATCCCC
CTGCCCACGGTTCGACAAAGCAACGCGTTTCCCTAAGTCAACGTATACAT
CAGAAGATCGGGTTAGCTGAGGTTTATAACCCATCCGCTTTAAGTAACTT
TGAACCGTCAAAGCAAGAGACAGGGTCAGGGACTTCTCTTCATAATGAAG
TTCCTTACTTGCCGTGCCTTGATCTTTTCCCCCGTGTCCAGGATCTATAA
ATATAACCTCACTGCGTCGTACACGCTGAGGAGGATTCGGTGTTTGAGCA

GTAACCGTAAATAATAAACTTTTGAAGTCTAAGCTCATCATATCTATTCA
TCCTTTGGATTCAAGACATTTTTTAAAAAATGCGCAATCACTATAAACCA
TATCGATTAATGCGAATAACTATATTTCTAGAACCTAGAAAAATCATTCC
CACATCTTGAGAAAAACTTCCCCTAAATTGCTAGCGTGCATCTAACACGT
GACTTCTTTAATCTAACTTGGTAAAGTGCTGGTCTTTGCGTCCTCGTAGC
TCAGCAGGATAGAGCGGTTGCCTCCTAAGCAGCAGGCCATGCGTTCGAAT
CGCATCGAGGACGATTTTTTGCCTTTAACTCCTAAAGTACTAATTTGCTT
GTATCTGTGGTTTACGTATTTTAGCGATATTCTGTTTTGGTTTCTGAAAA
CTAGGTCCAGAAAGAAAATTATGAACTCCCTCGGCGATGCCTTTCGCTAC
ATGCATACGATAGCGAGCATCTTGCAGGGCCGCACGTTCACGACTATTGG
ATAAAAACCCGGTTTCCACCAAAACTGCAGGCATAGAAGTATCTCTAATC
ACAACAAAGTTCGCAGTTTTCAAACCTCGAGACTTCAAAATGCCATTTTT
TTCCATAGCAGCTAAAATGTTTTTTCCCAGTACTTCTGACATGCGATTCC
TAGTCGGAGATCCGACCTTACCATTATAAAAATATACTTCGGTGCCAAAG
GCTGCTGCGTTTGAAGAATGATTACAGTGGATGCTGATAAAGACATCCCC
CTGCCCACGGTTCGACAAAGCAACGCGTTTCCCTAAGTCAACGTATACAT
CAGAAGATCGGGTTAGCTGAGGTTTATAACCCATCCGCTTTAAGTAACTT
TGAACCGTCAAAGCAAGAGACAGGGTCAGGGACTTCTCTTCATAATGAAG
TTCCTTACTTGCCGTGCCTTGATCTTTTCCCCCGTGTCCAGGATCTATAA
ATATAACCTCACTGCGTCGTACACGCTGAGGAGGATTCGGTGTTTGAGCA





What are we looking for?

Protein-coding genes: DNA $\rightarrow$ mRNA $\rightarrow$ protein

Functional RNA genes: DNA $\rightarrow$ RNA

How can we find genes?

Gene family: group of evolutionarily related (*homologous*) genes in different genomes

Homology search: given one or more homologs of a family, find more

Homologous genes often share similar functions, structures and sequences

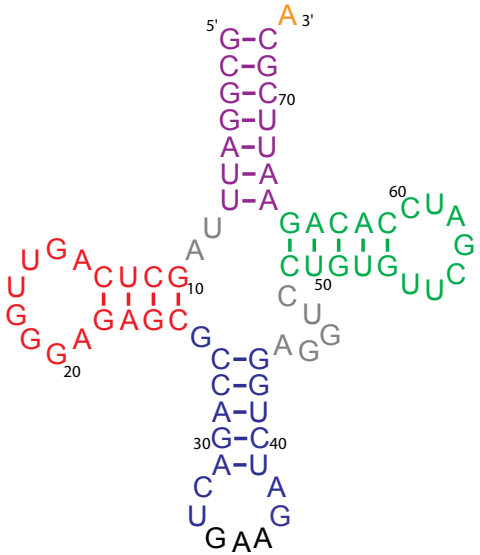
Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



3-dimensional structure

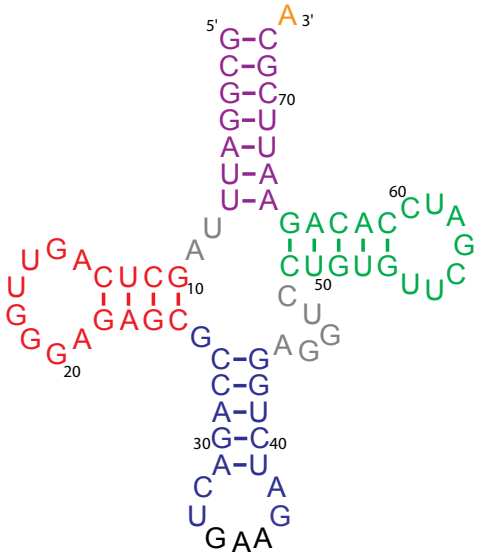
Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



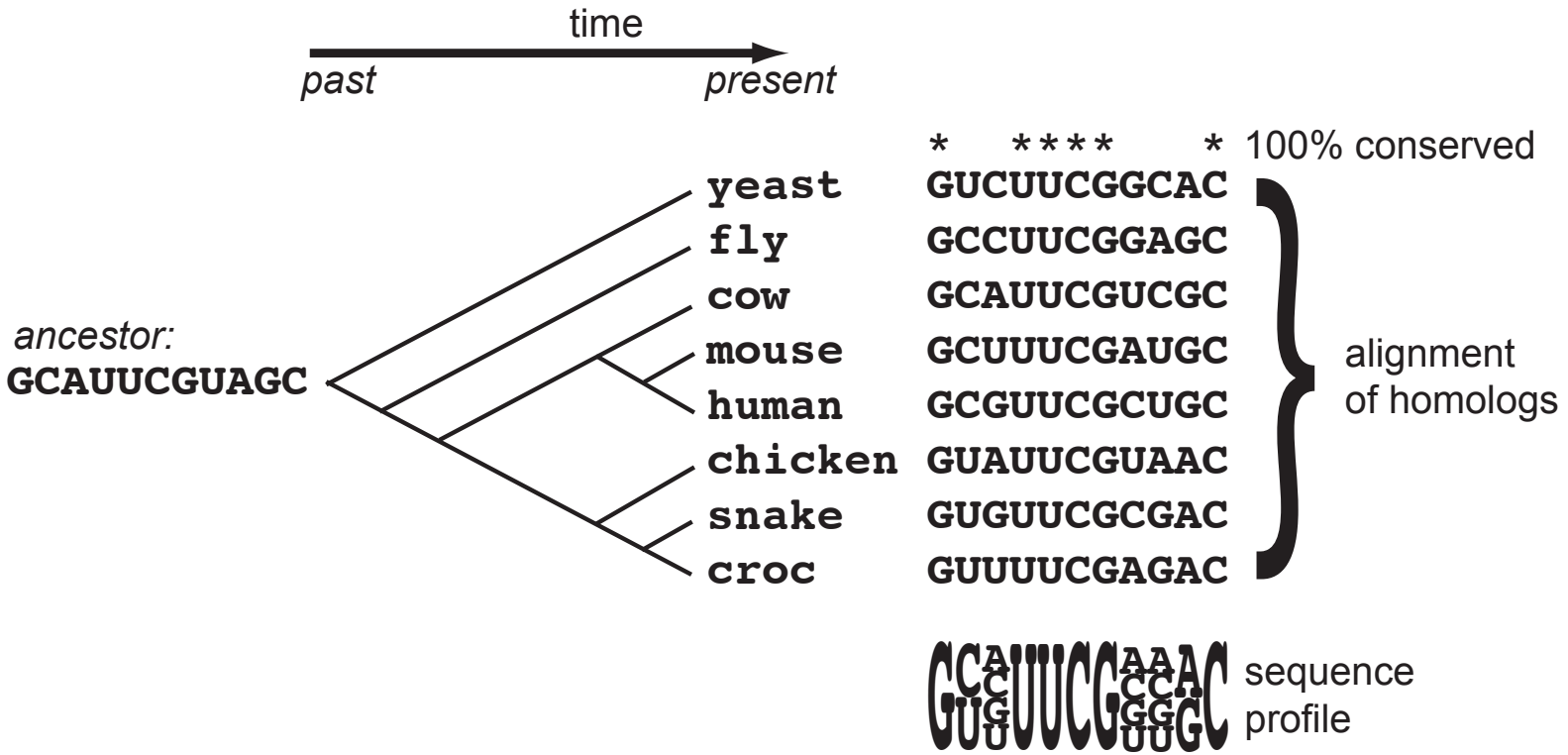
3-dimensional structure

BLAST: given a single sequence, search genomes for similar sequences.

Homologous proteins and RNAs conserve different sequence and structural features to different degrees.

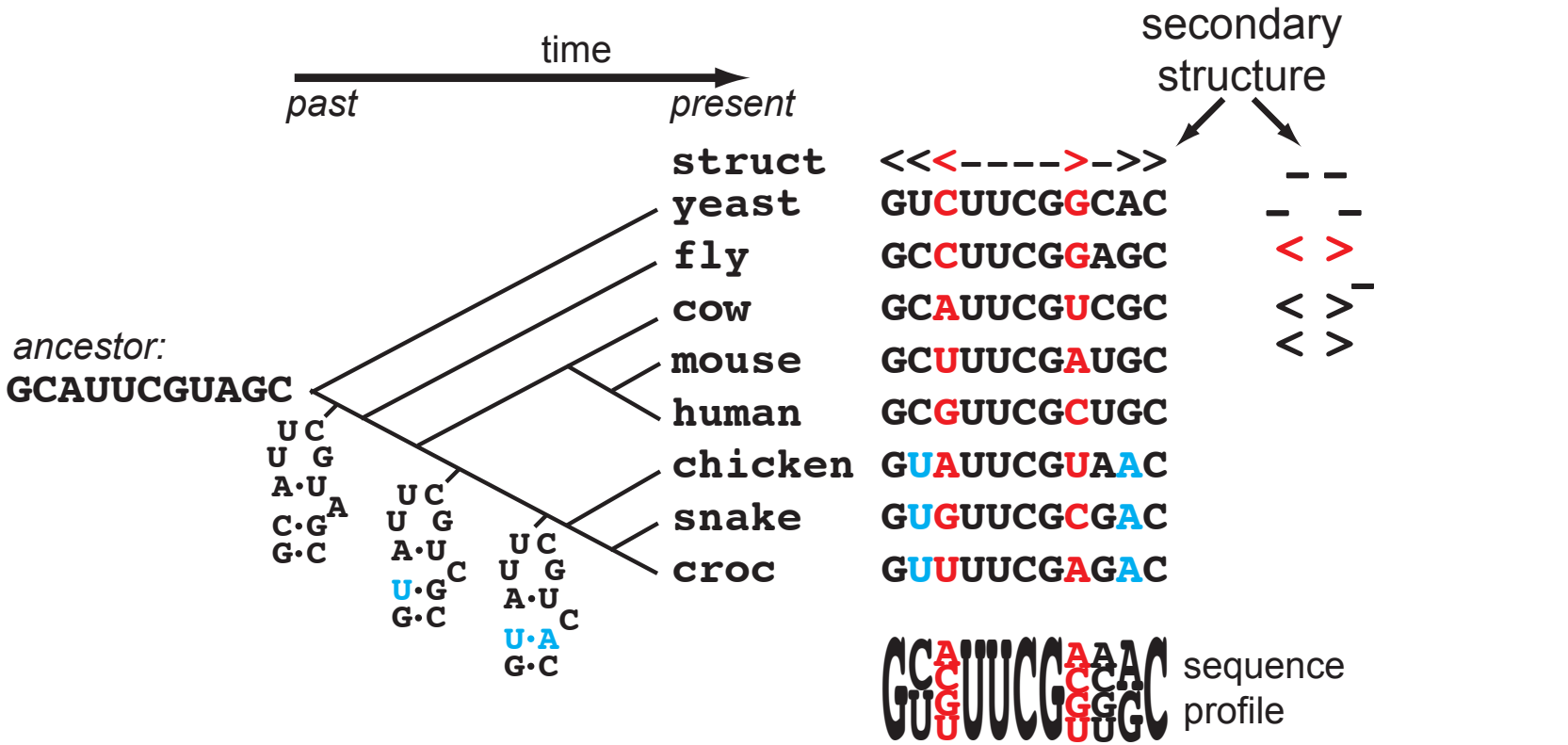
Sequence conservation provides information for homology searches

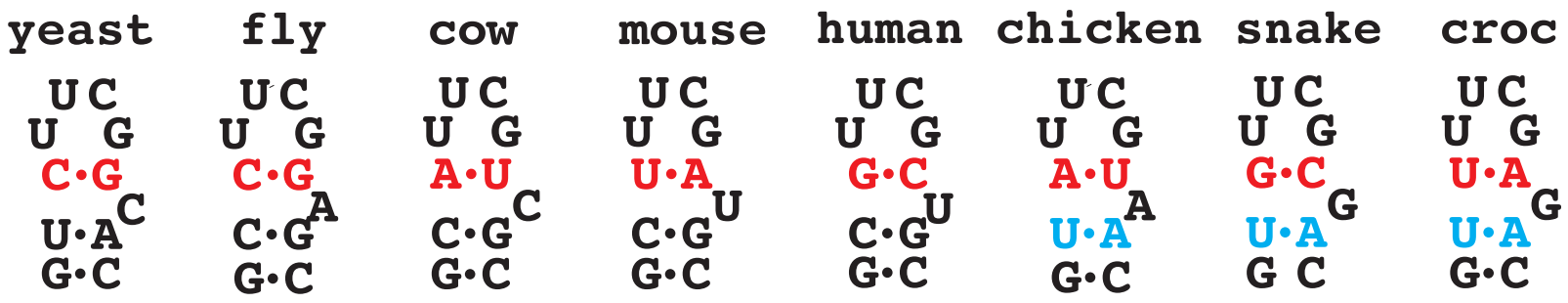
Conservation levels vary across alignment columns.



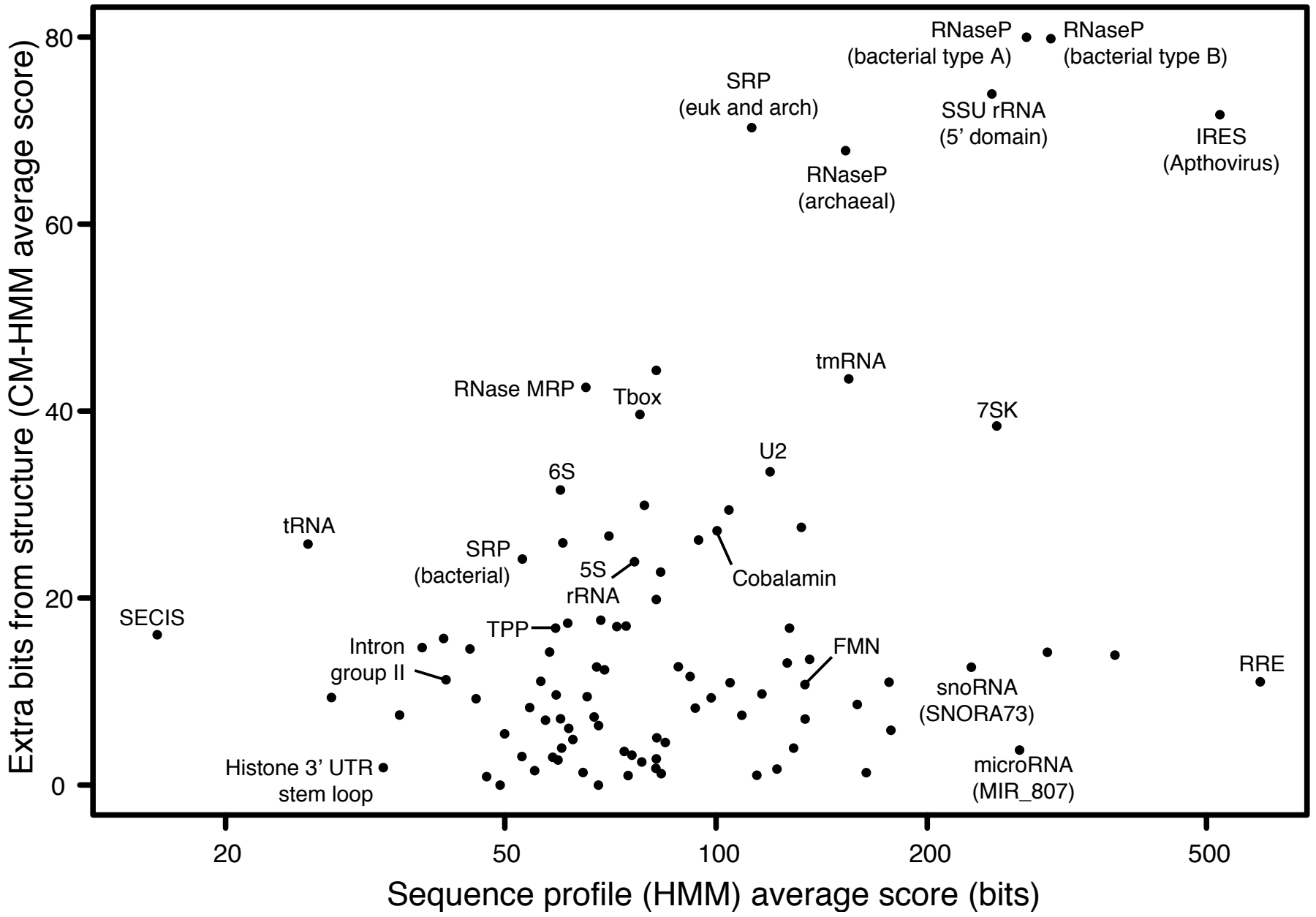
Structure conservation provides additional information

Base-paired positions covary
to maintain Watson-Crick complementarity.





Levels of sequence and structure conservation in RNA families



Eddy lab software for profile probabilistic models

(since 1994)

	sequence profiles	sequence and structure profiles
models	profile HMMs	covariance models (CMs)
software	HMMER	Infernal (prev. COVE)
main use	proteins	RNAs
database	Pfam (9318 families)	Rfam (1371 families)
performance for RNAs	faster but less accurate	slower but more accurate

http://hmmer.janelia.org

Eddy, SR. PLoS Comp. Biol., 4:e1000069, 2008.

Eddy. SR. Bioinformatics, 14:755-763, 1998.

http://infernal.janelia.org

Eddy SR, Durbin R. Nucleic Acids Research, 22:2079-2088, 1994.

Eddy SR, BMC Bioinformatics, 3:18, 2002.

Profile HMMs: sequence family models built from alignments

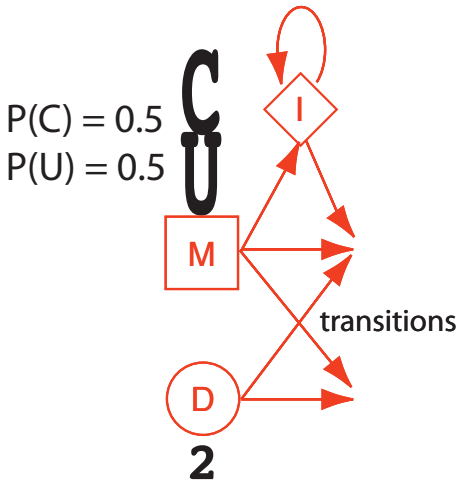
yeast	GUCUUCGGCAC
fly	GCCUUCGGAGC
cow	GCAUUCGUCGC
mouse	GCUUUCGAUGC
human	GCGUUCGCUGC
chicken	GUAUUCGUAAC
snake	GUGUUCGCGAC
croc	GUUUUCGAGAC

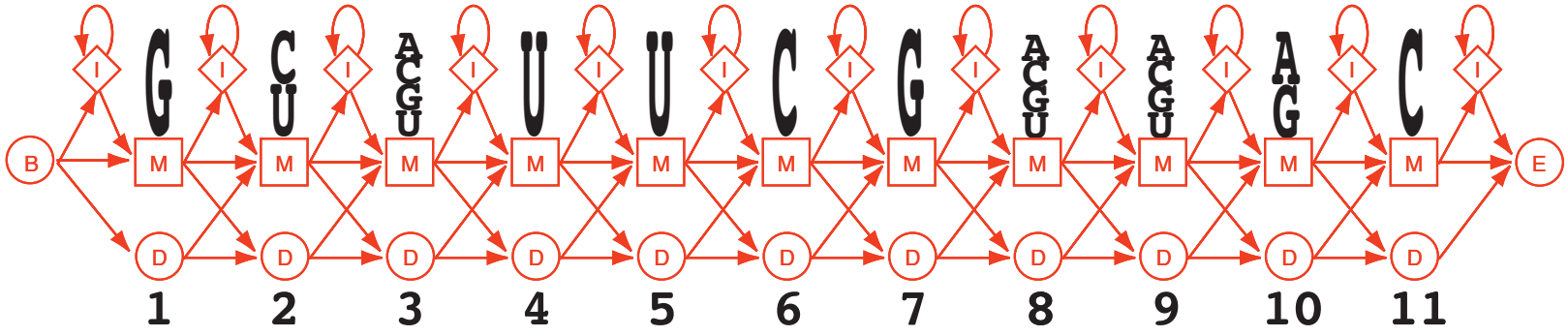
One HMM node per alignment column

3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

Node for column 2:





Profile HMMs: sequence family models built from alignments


```
yeast     GUCUUCGGCAC
fly       GCCUUCGGAGC
cow       GCAUUCGUCGC
mouse     GCUUUCGAUGC
human     GCGUUCGCUGC
chicken   GUAUUCGUAAC
snake     GUGUUCGCGAC
croc      GUUUUCGAGAC
```

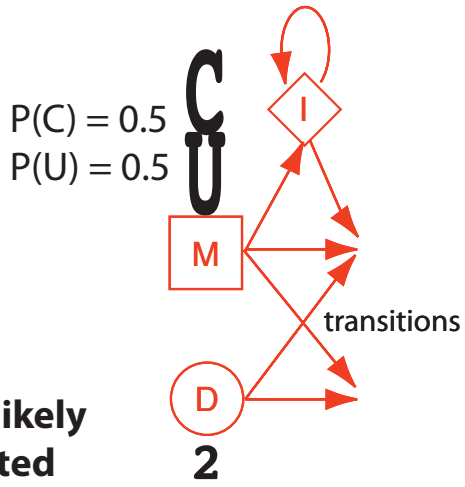
One HMM node per alignment column

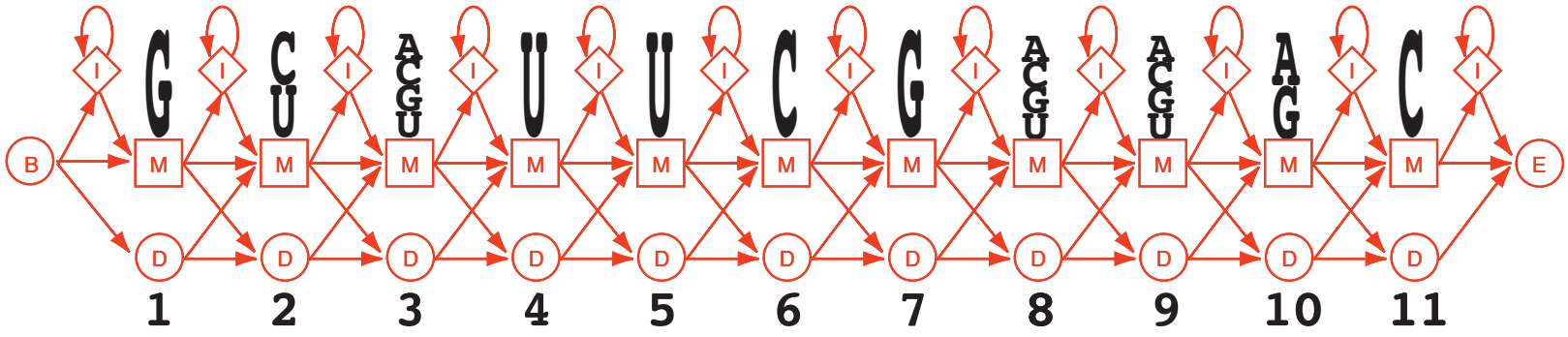
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

Given a sequence, the most likely path that could have generated that sequence can be computed.

Node for column 2:





Profile HMMs: sequence family models built from alignments


```
yeast     GUCUUCGGCAC
fly       GCCUUCGGAGC
cow       GCAUUCGUCGC
mouse     GCUUUCGAUGC
human     GCGUUCGCUGC
chicken   GUAUUCGUAAC
snake     GUGUUCGCGAC
croc      GUUUUCGAGAC
worm      GCGUUCGCGGC
```

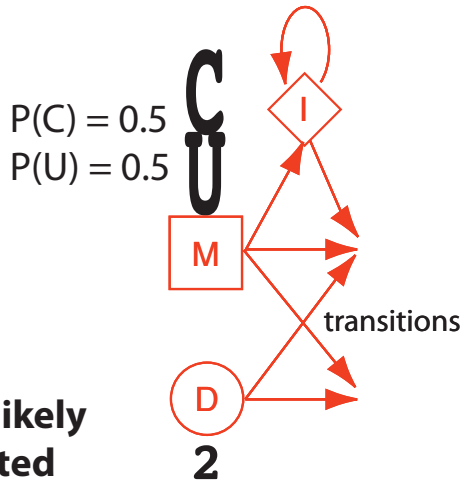
One HMM node per alignment column

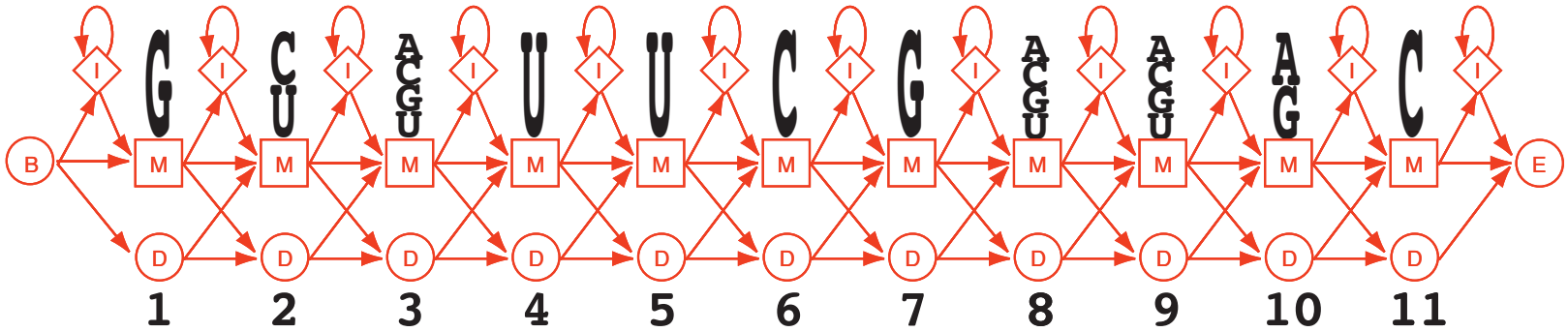
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

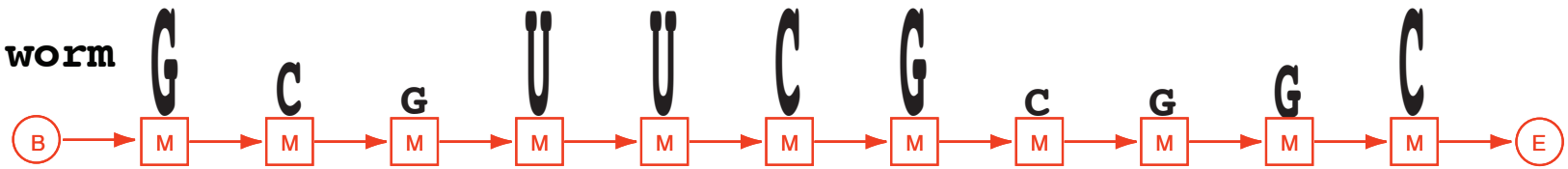
HMMs generate homologous sequences.

Given a sequence, the most likely path that could have generated that sequence can be computed.

Node for column 2:







Profile HMMs: sequence family models built from alignments


```
yeast     GU.CUUCGGCAC
fly       GC.CUUCGGAGC
cow       GC.AUUCGUCGC
mouse     GC.UUUCGAUGC
human     GC.GUUCGCUGC
chicken   GU.AUUCGUAAC
snake     GU.GUUCGCGAC
croc      GU.UUUCGAGAC
worm      GC.GUUCGCGGC
corn      GUGAUUCGU.GC
```

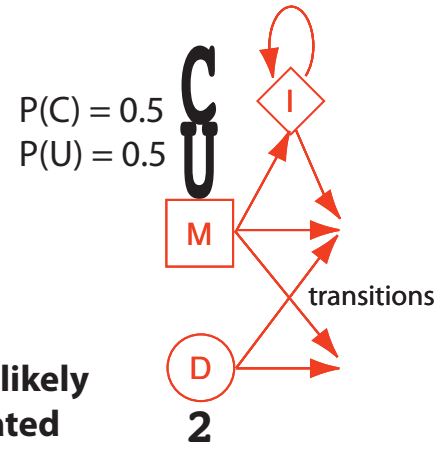
One HMM node per alignment column

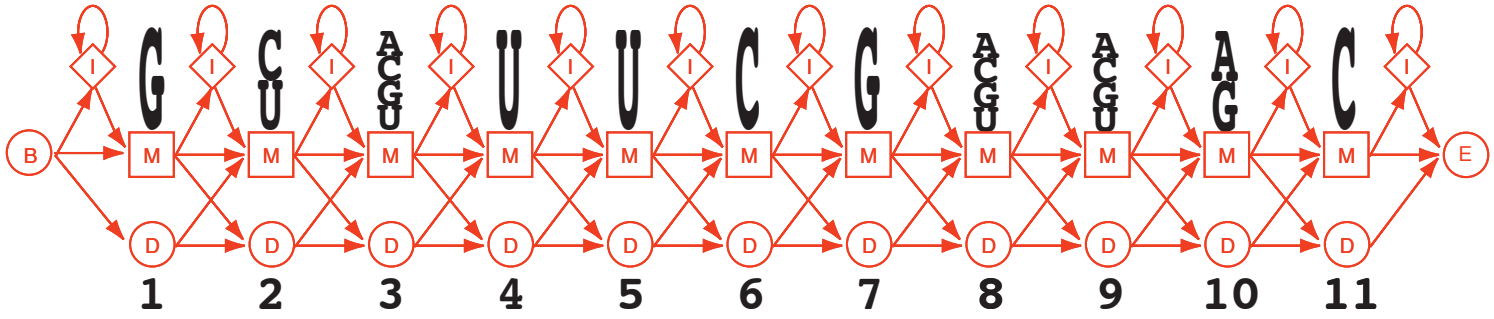
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

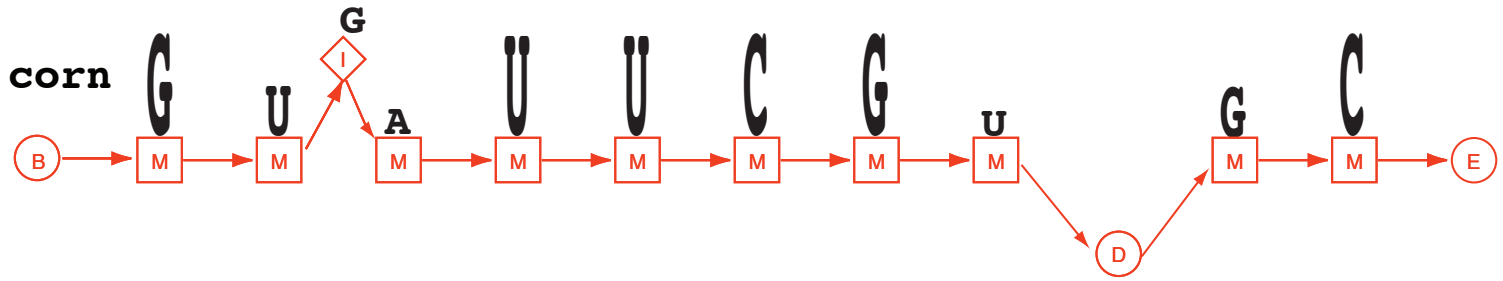
HMMs generate homologous sequences.

Given a sequence, the most likely path that could have generated that sequence can be computed.

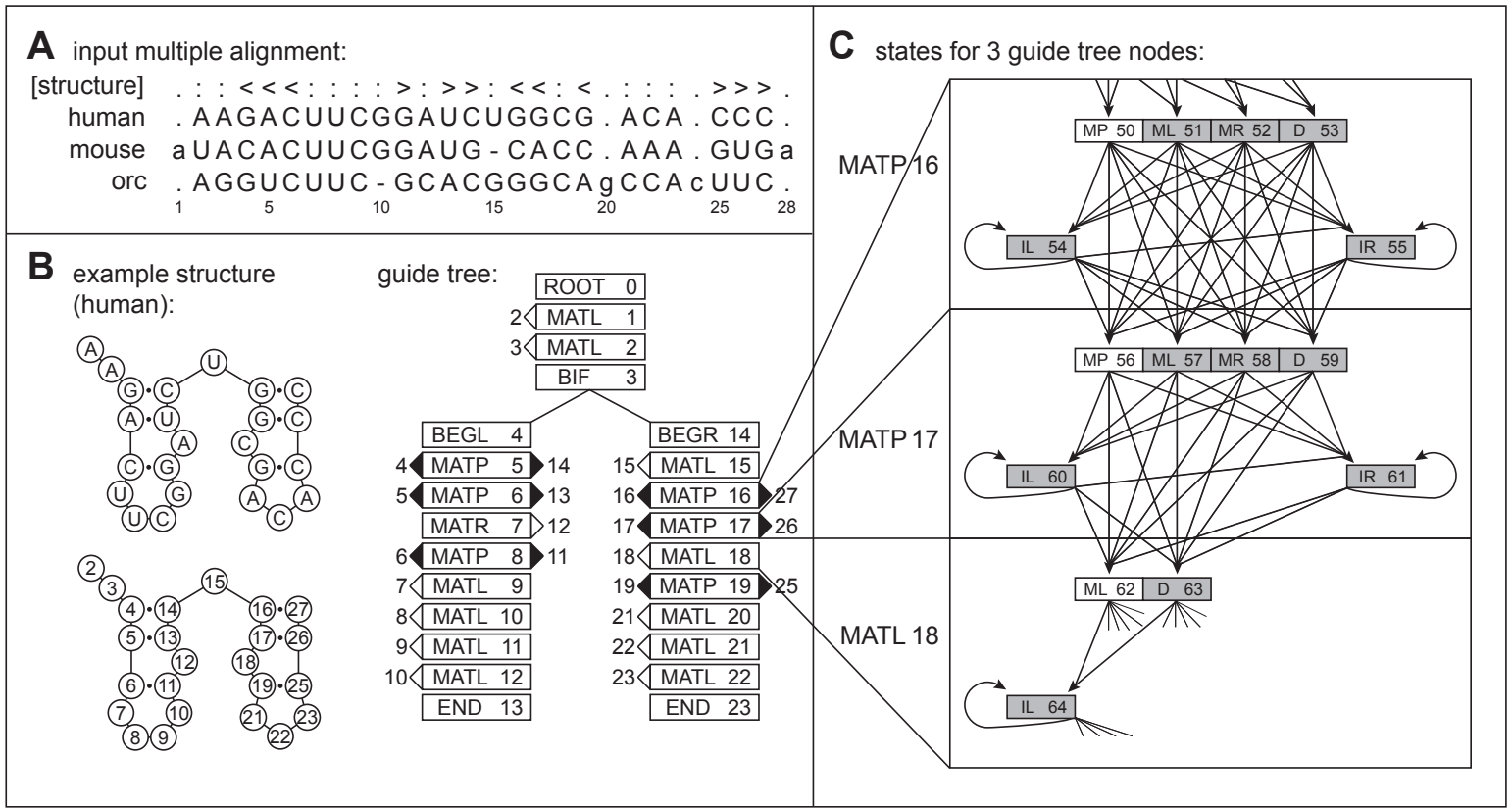
Node for column 2:

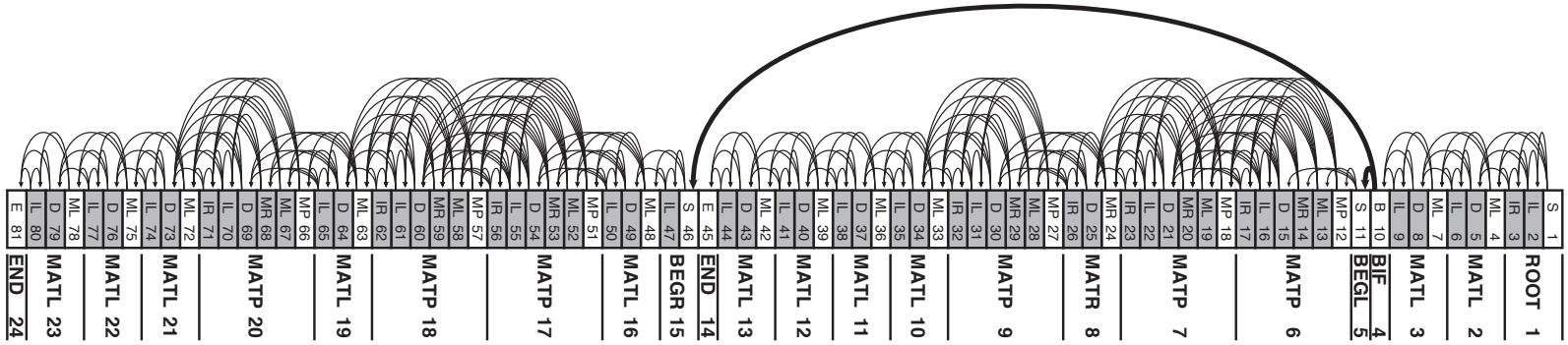






Covariance models (CMs) are built from structure-annotated alignments





My work on Infernal

- Infernal version 0.55 existed when I entered the lab.
- I have developed and implemented methods for:
  - improving the accuracy of CM searches
  - accelerating CM searches
  - accelerating CM alignment

    Nawrocki EP. Eddy SR. PLoS Comput. Biol., 3:e56, 2007.

    Nawrocki EP. Kolbe DL. Eddy SR. Bioinformatics, 25:1335-1337, 2009.

My work on Infernal

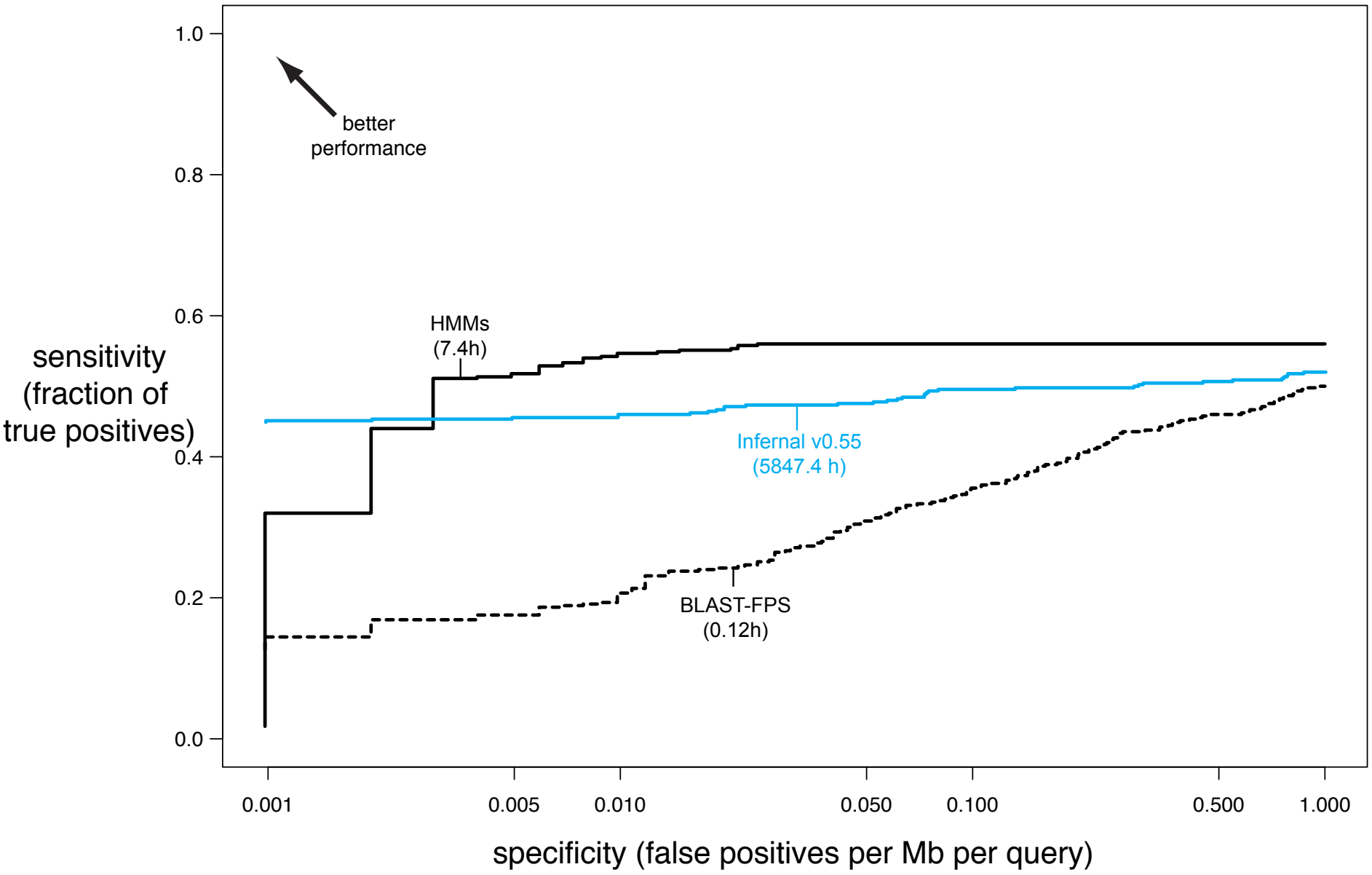
- Infernal version 0.55 existed when I entered the lab.
- I have developed and implemented methods for:
  - improving the accuracy of CM searches
  - accelerating CM searches
  - accelerating CM alignment

    Nawrocki EP, Eddy SR. PLoS Comput. Biol., 3:e56, 2007.

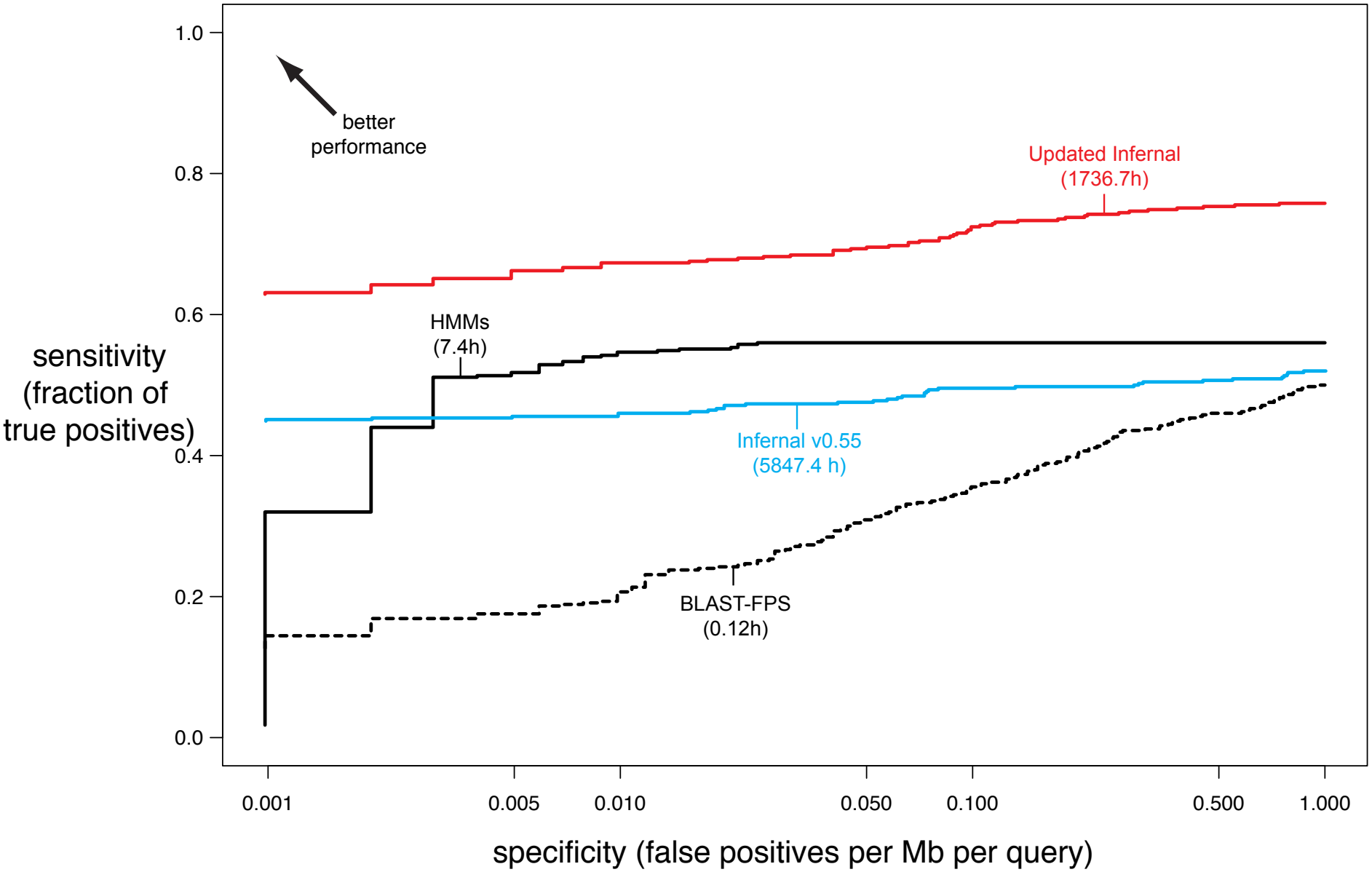
    Nawrocki EP, Kolbe DL, Eddy SR. Bioinformatics, 25:1335-1337, 2009.

- First, I constructed a benchmark to evaluate homology search performance.
- Performance is judged by:
  - sensitivity: correctly inferring homology
  - specificity: correctly ignoring non-homology

Infernal v0.55 does no better than sequence-only profiles (HMMs)



Updated Infernal*† shows significant improvement



*Nawrocki EP. Eddy SR. PLoS Comput. Biol., 3:e56, 2007.

†Based on work on profile HMMs: Johnson S. PhD Thesis, 2006, Karplus K et. al, ISMB, 1995, Sjölander K et. al, CABIOS, 1996, Buhler and Swope (HMMER2, unpublished)

CM searches are especially slow for large RNAs

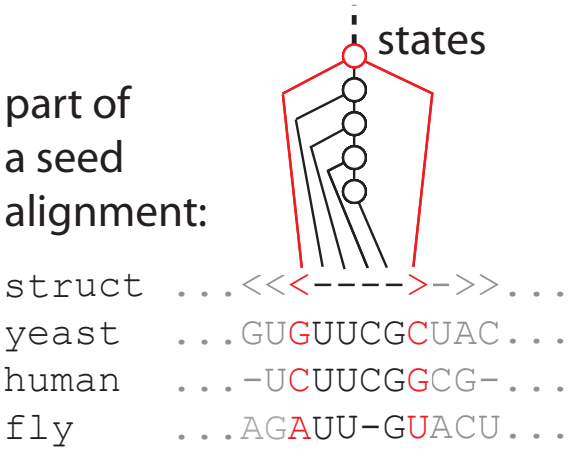
		search (min/Mb)		
family	length	HMM	CM	CM/HMM
tRNA	71	0.34	27.0	79.4
Lysine riboswitch	183	0.80	133.2	166.7
SRP RNA	304	1.32	276.4	214.4
RNaseP RNA	365	1.56	733.4	470.3

Why CM homology search is so slow

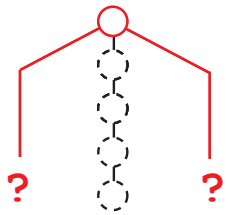
- CM homology search algorithms align/score all subsequences of length $1..W$ as they scan along the target sequence looking for high scoring hits

Example: Finding a hairpin loop

part of a seed alignment:



where's the loop?



NNN...NNNNNN...NNN
1 W

W*(W-1)/2 possibilities

if W=100 ~ 5000 possibilities
if W=200 ~ 20000 possibilities

We could save time by restricting the possible loop lengths considered.

One idea: take advantage of the generative capacity of CMs to generate sequences and examine loop length distribution.

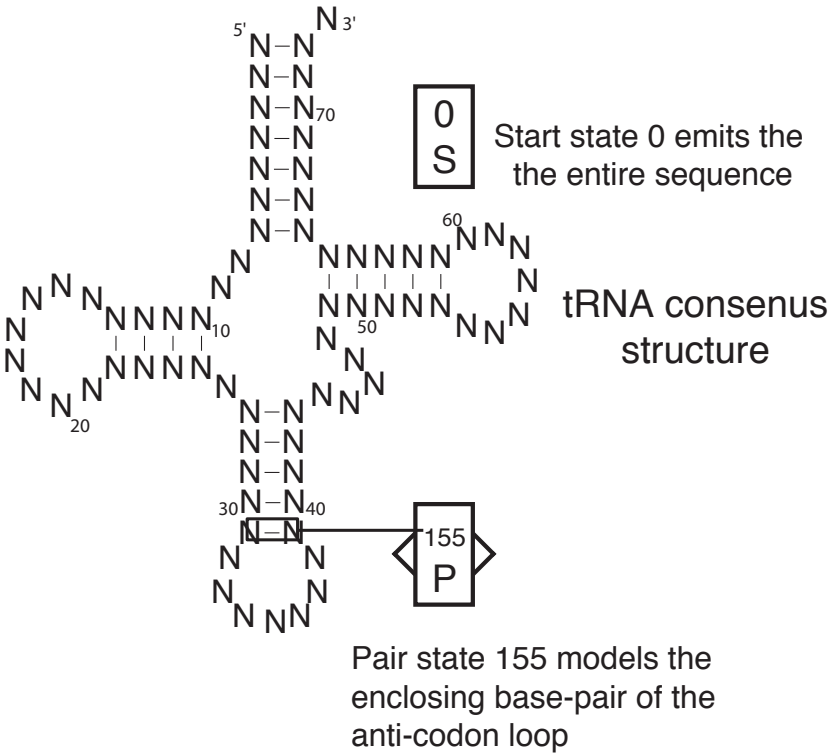
Query-dependent banding (QDB) strategy

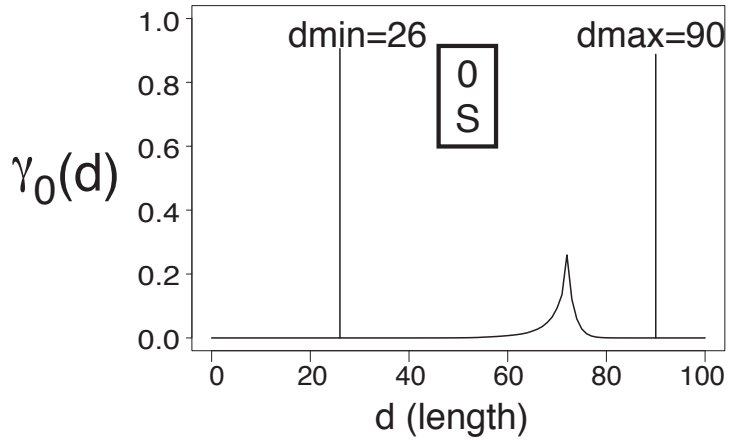
- Calculate $\gamma_v(d)$ probability each state v will emit/align to subsequences of length d, for $d = 0..Z$

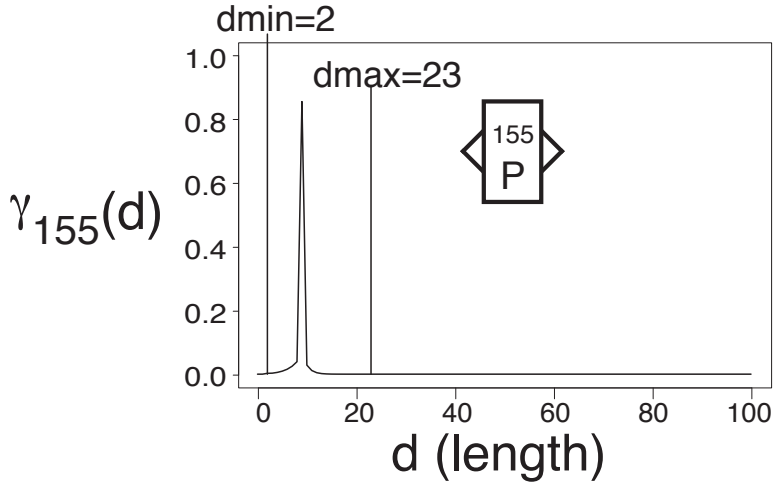
for states $v = M-1$ down to 0:

v = end state (E):	$\gamma_v(0) = 1$ $\gamma_v(d) = 0$	for $d = 1$ to Z
v = bifurcation (B):	$\gamma_v(d) = \sum_{n=0}^{d} \gamma_y(n) * \gamma_z(d-n)$	for $d = 0$ to Z
else ($v = S, P, L, R$):	$\gamma_v(d) = 0$ $\gamma_v(d) = \sum_{y \in C_v} \gamma_y(d - (\Delta_v^L + \Delta_v^R)) * t_v(y)$	for $d = 0$ to $(\Delta_v^L + \Delta_v^R - 1)$ for $d = (\Delta_v^L + \Delta_v^R)$ to Z

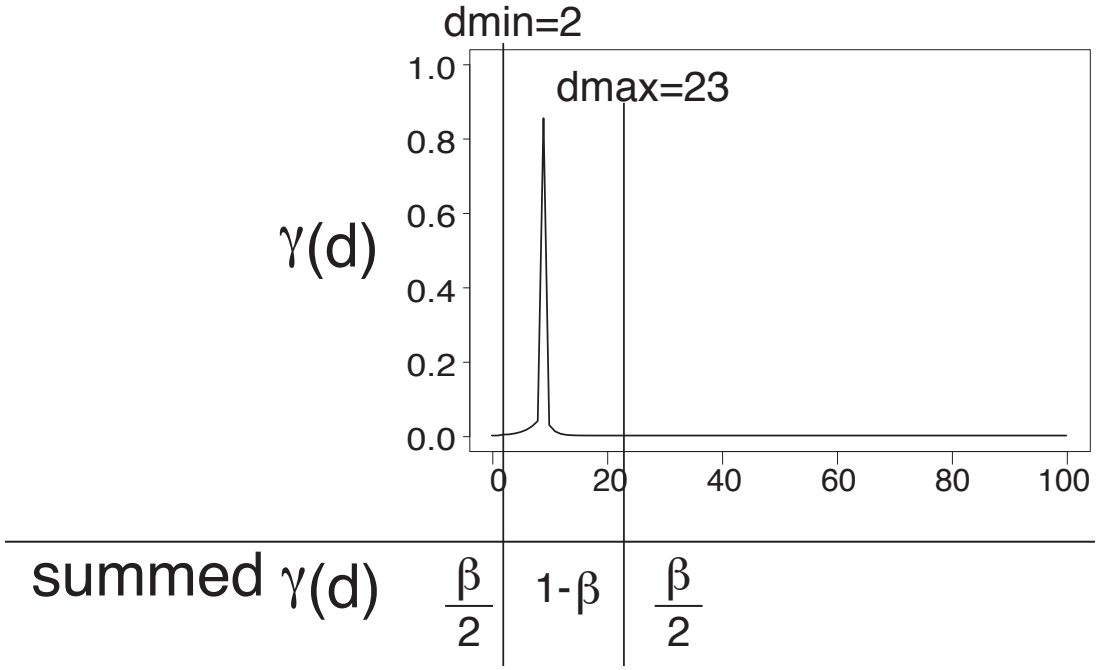
QDBs for a tRNA CM:







The β parameter controls amount of probability loss



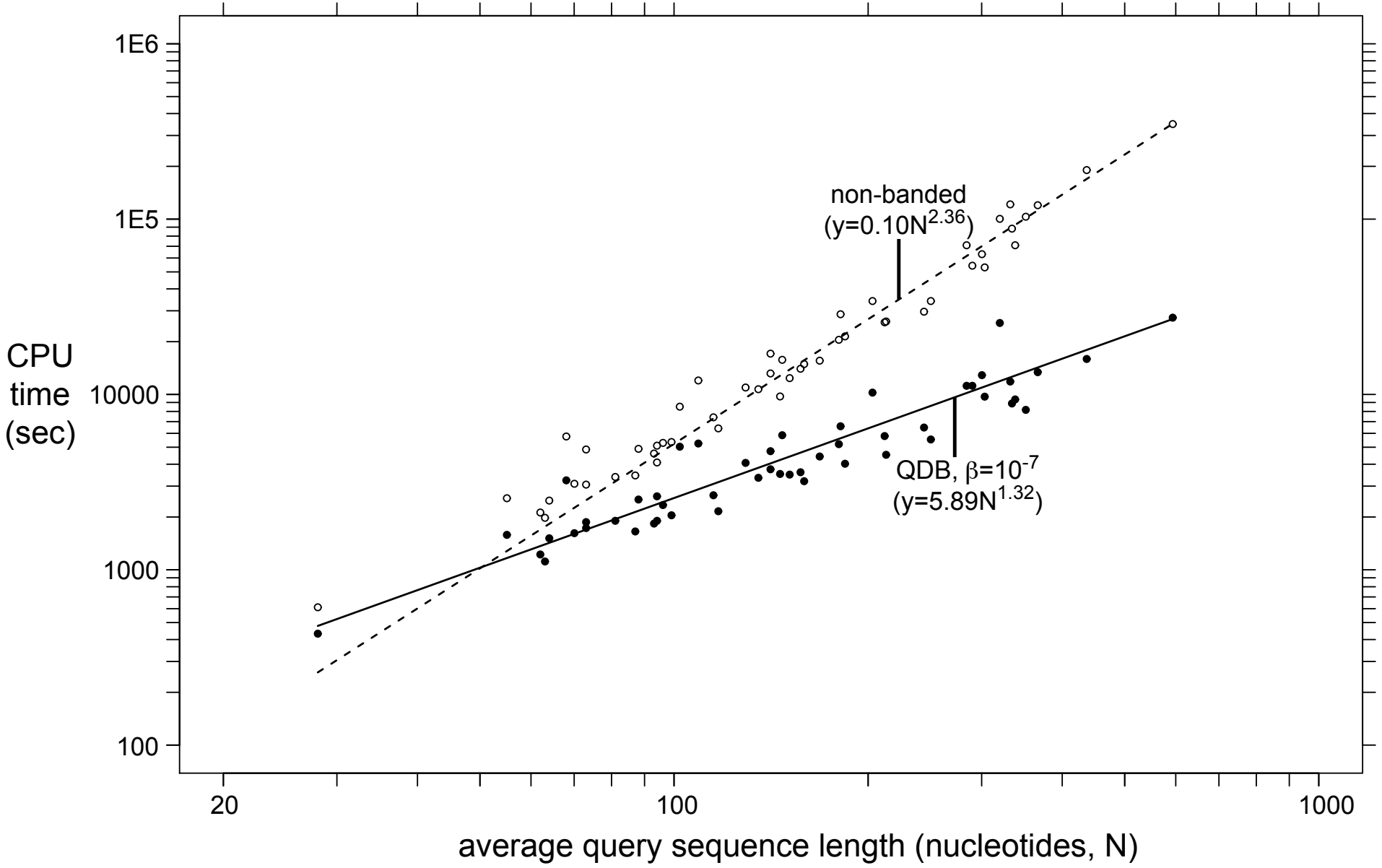
$$\sum_{d=0}^{\text{dmin}-1} \gamma(d) < \frac{\beta}{2}$$

$$\sum_{d=\text{dmin}}^{\text{dmax}} \gamma(d) = 1 - \beta$$

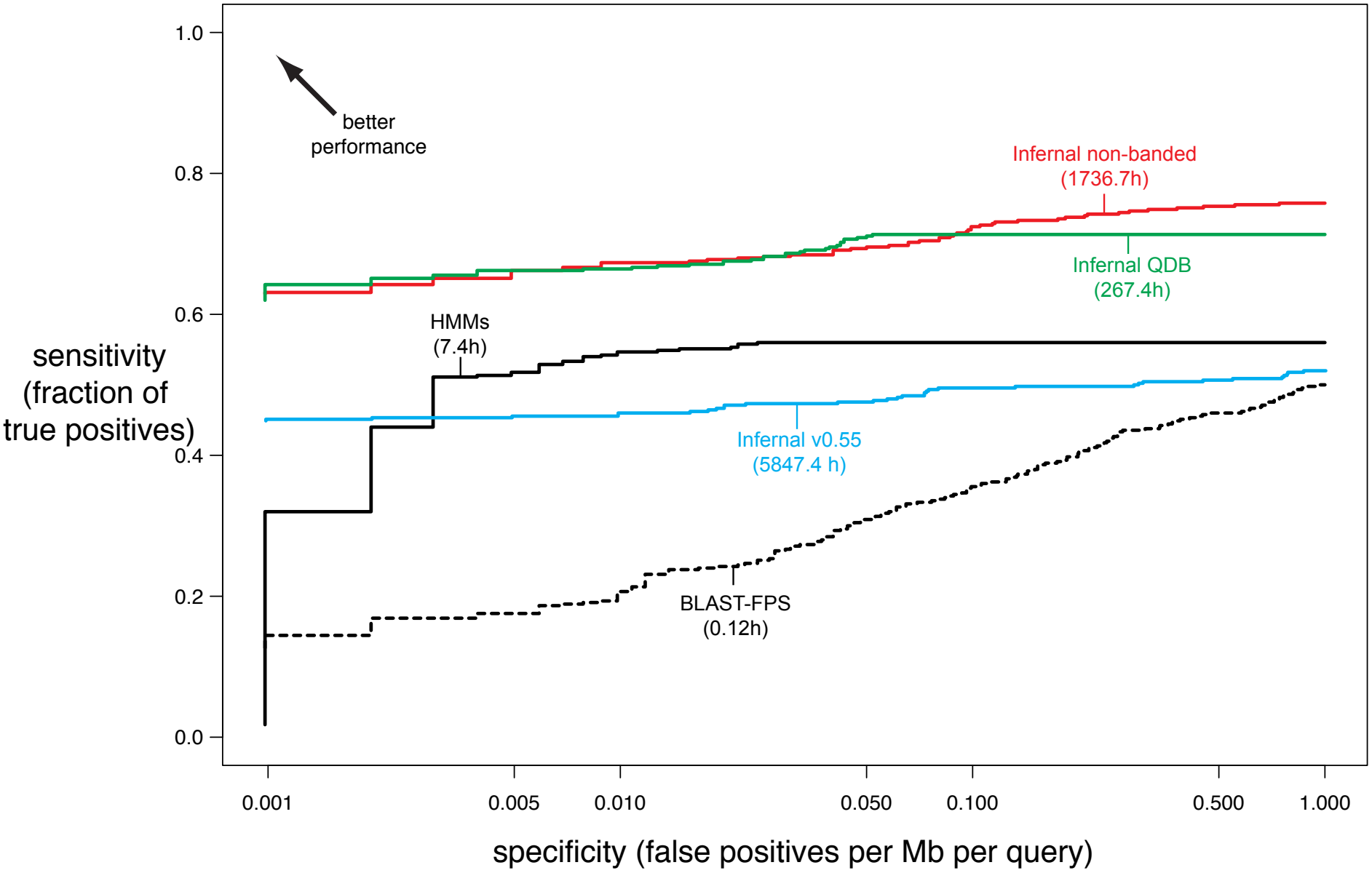
$$\sum_{d=\text{dmax}+1}^{Z} \gamma(d) < \frac{\beta}{2}$$

- β is typically very small
  for example: $0.0000001(10^{-7})$
- Higher β gives more acceleration
  but at larger cost to accuracy

Empirical time complexity of CM homology search



QDB sacrifices very little sensitivity and gives 6-fold speedup

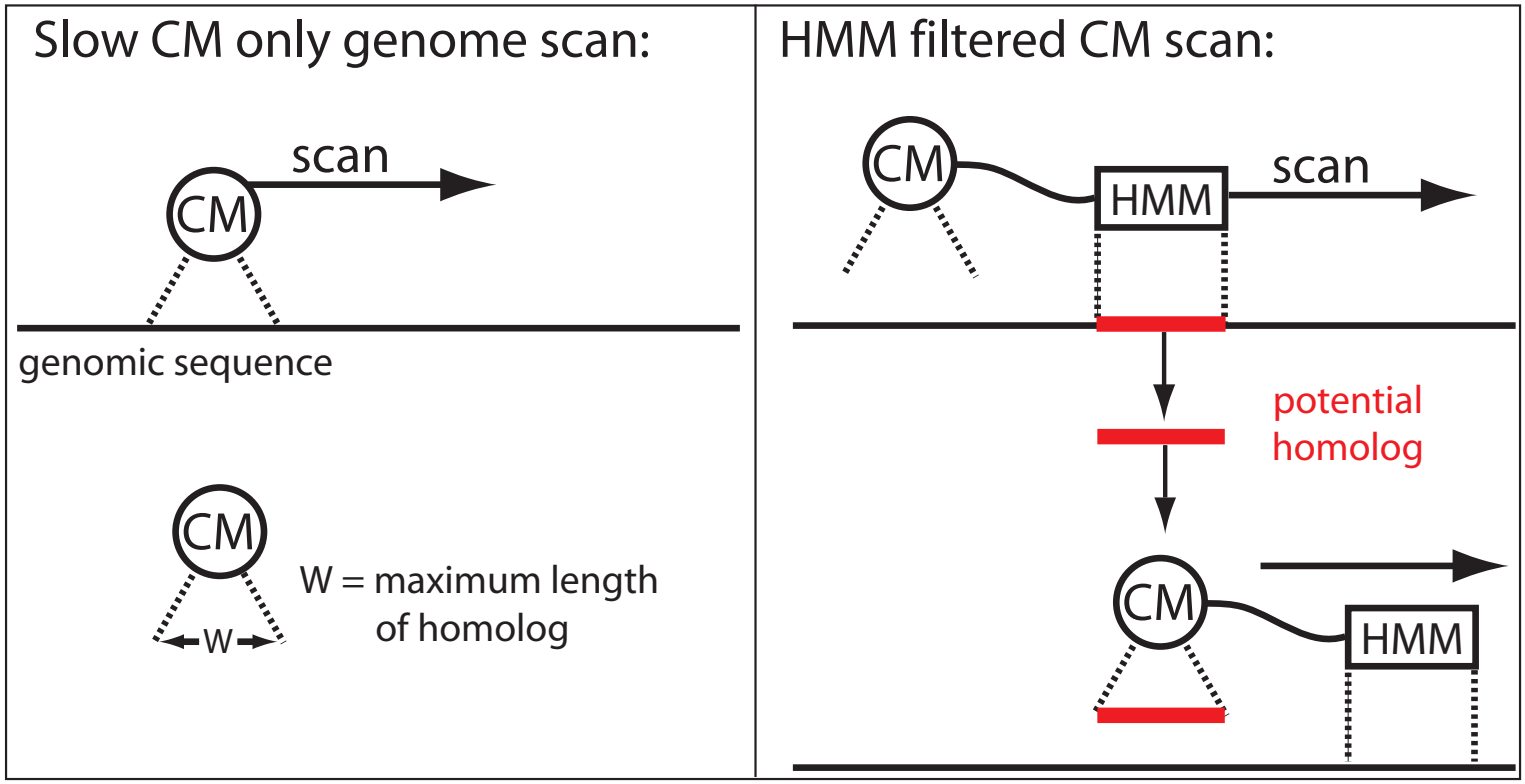


CM homology searches are still slow

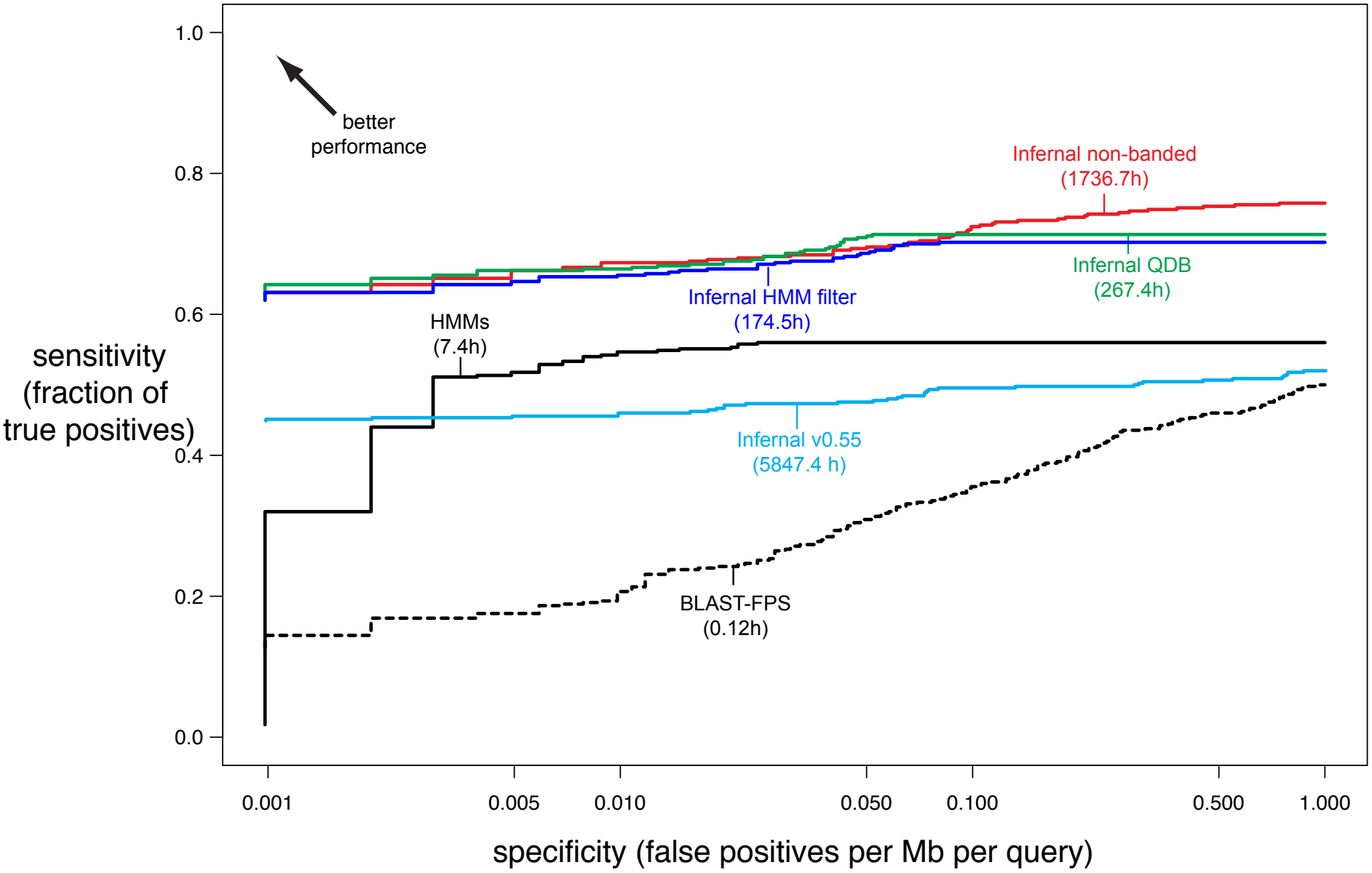
		search (min/Mb)			
family	length	HMM	QDB CM	QDB CM/HMM	non-banded CM/HMM
tRNA	71	0.34	9.6	28.2	79.4
Lysine riboswitch	183	0.80	33.8	42.3	166.7
SRP RNA	304	1.32	50.5	38.3	214.4
RNaseP RNA	365	1.56	81.6	52.3	470.3

Filtering as a complementary acceleration strategy

- Main idea: search database with faster method first, hits above some threshold survive the filter and are searched with the slow CM.
- Weinberg and Ruzzo developed HMM filters for faster searches
- Others have also worked on this (Sun & Buhler, 2008; Zhang & Bafna, 2006)



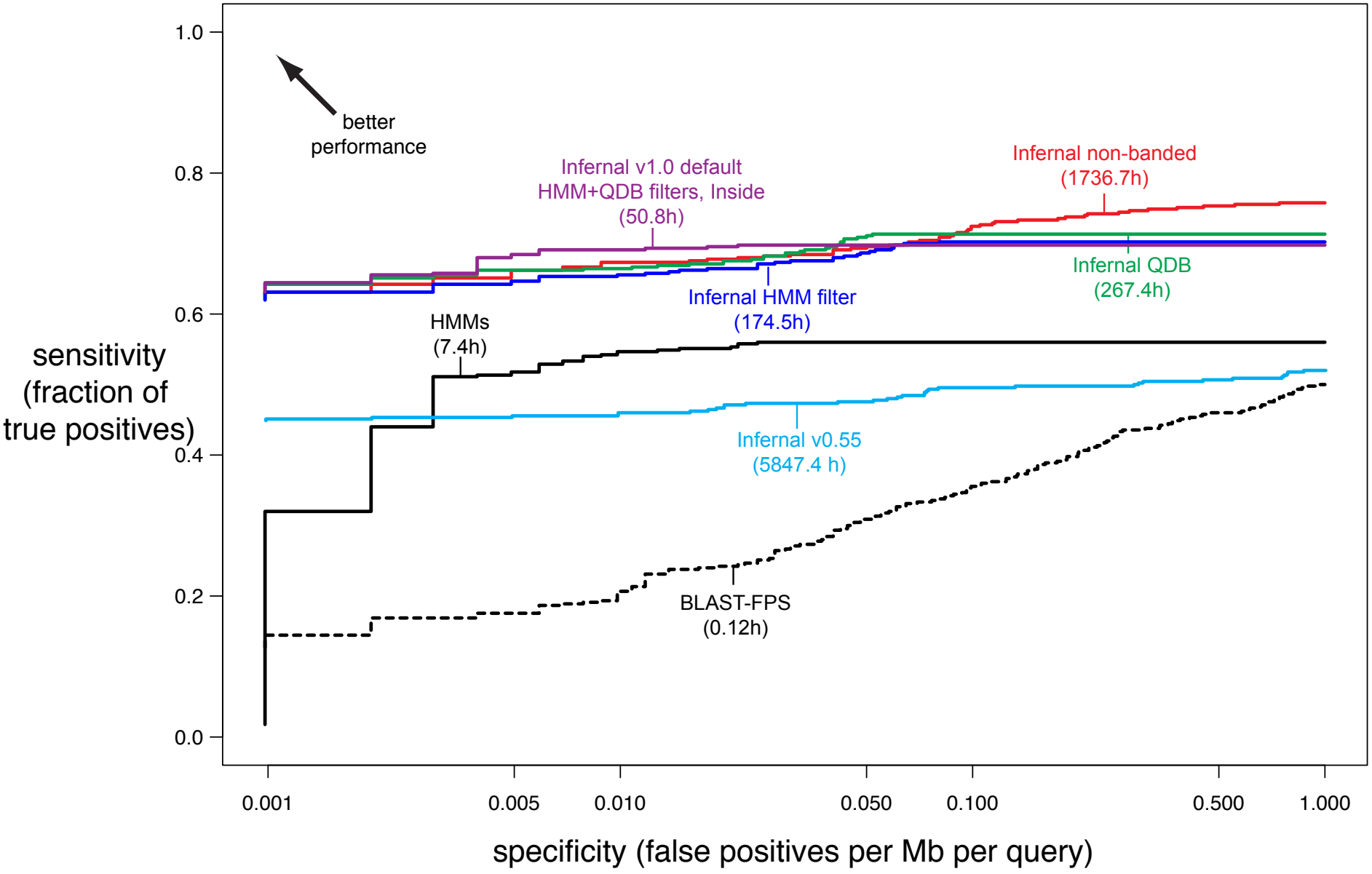
HMM filters achieve 10-fold speedup at very small cost to accuracy



Combining QDB and HMM filters yields greater acceleration

The more powerful, slower Inside algorithm is used post-filtering.

Infernal is now 100-fold faster and significantly more sensitive than v0.55.

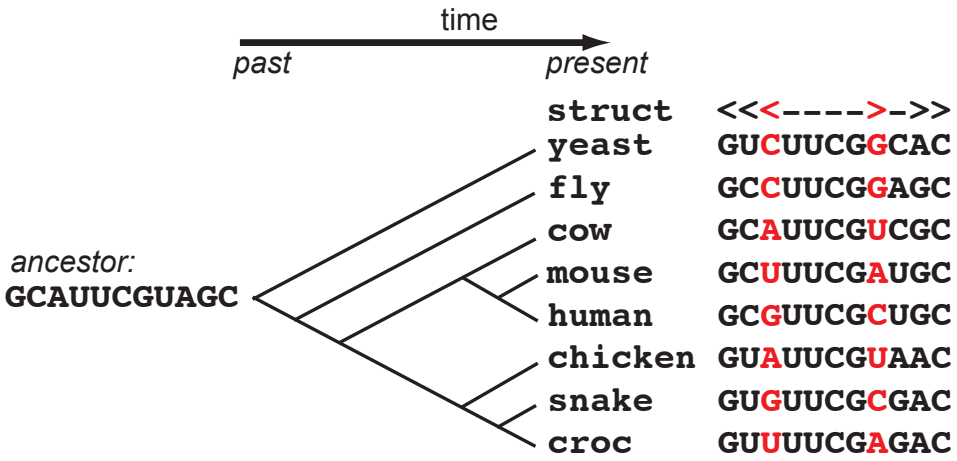


CMs are now nearly as fast as HMMs (usually)

		search (min/Mb)			
family	length	HMM	HMM+QDB filtered CM	HMM+QDB filtered CM/HMM	non-banded CM/HMM
tRNA	71	0.34	8.8	25.9	79.4
Lysine riboswitch	183	0.80	2.2	2.8	166.7
SRP RNA	304	1.32	6.0	4.5	214.4
RNaseP RNA	365	1.56	1.8	1.2	470.3

Structural RNA alignment using CMs

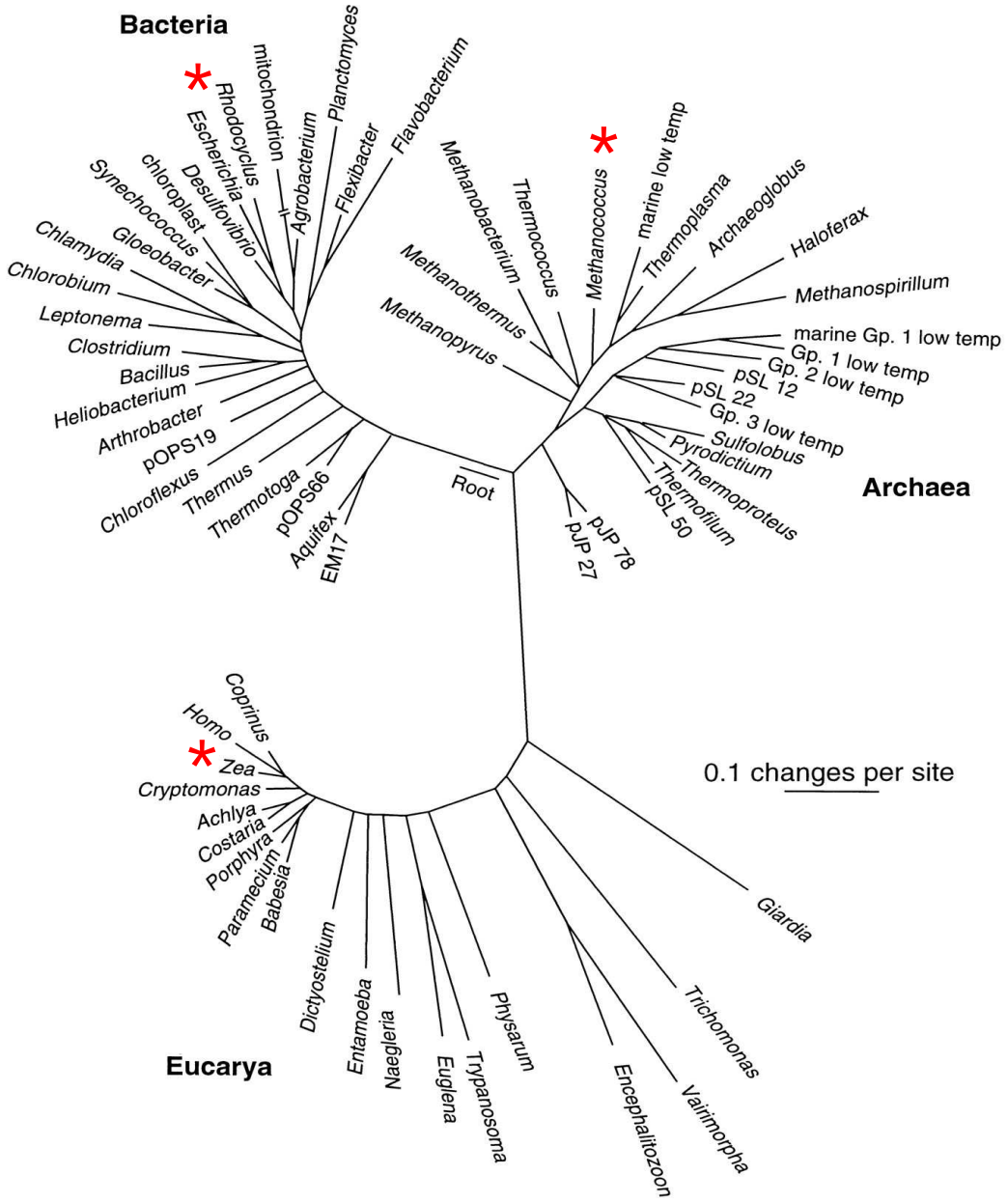
- CMs can also be used to create structural alignments of homologous RNAs.
- Given known homologs, place homologous residues in the same columns.



- Alignments of SSU rRNA have commonly been used for phylogenetic inference.
- However, CM alignment is too slow for SSU alignment.
  Aligning a single SSU sequence takes more than 20 minutes.

Small subunit ribosomal RNA and the tree of life

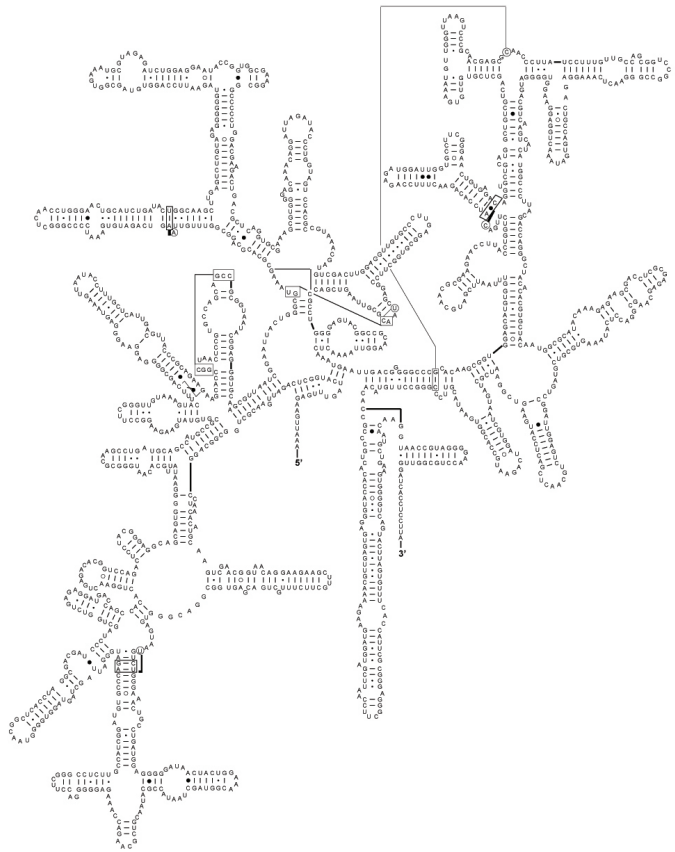
- 1977 - Carl Woese decided to classify all living things phylogenetically
- needed *"a molecule of appropriately broad distribution"* for comparative analysis
- SSU rRNA was chosen
  - universally distributed
  - highly conserved
  - large enough to provide sufficient data



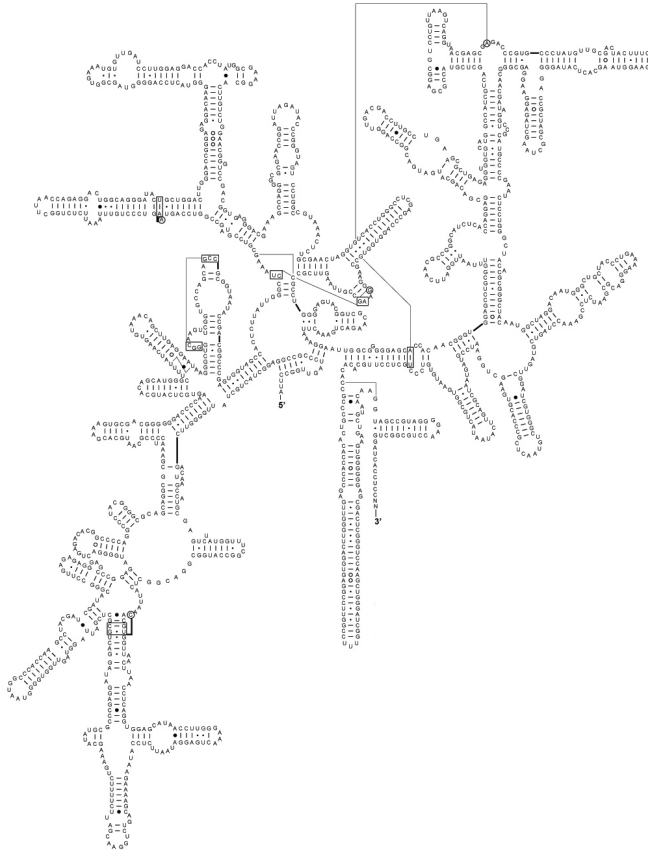
Pace, Science:276; 1997

Universal structural conservation of SSU rRNA

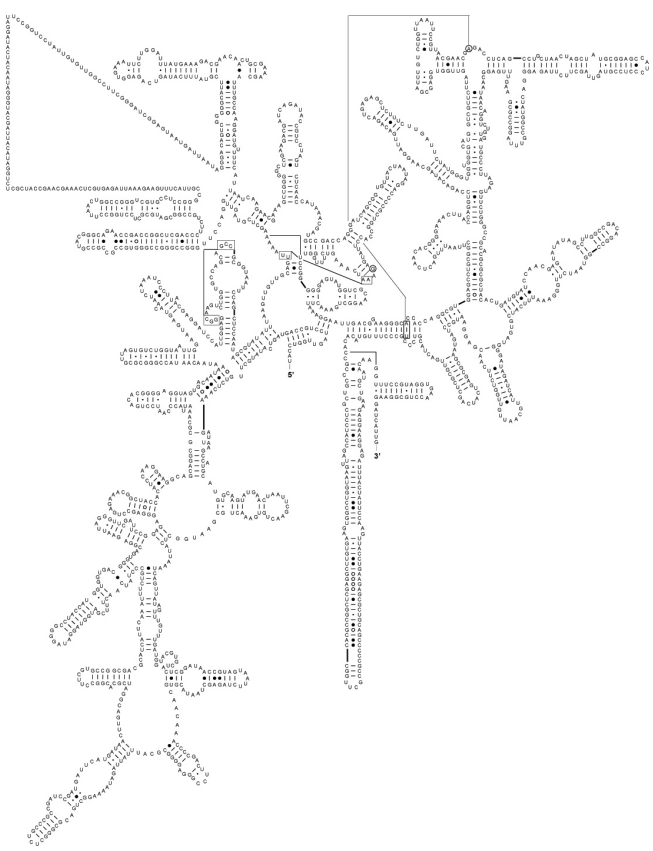
Escherichia coli



Methanococcus vannielii



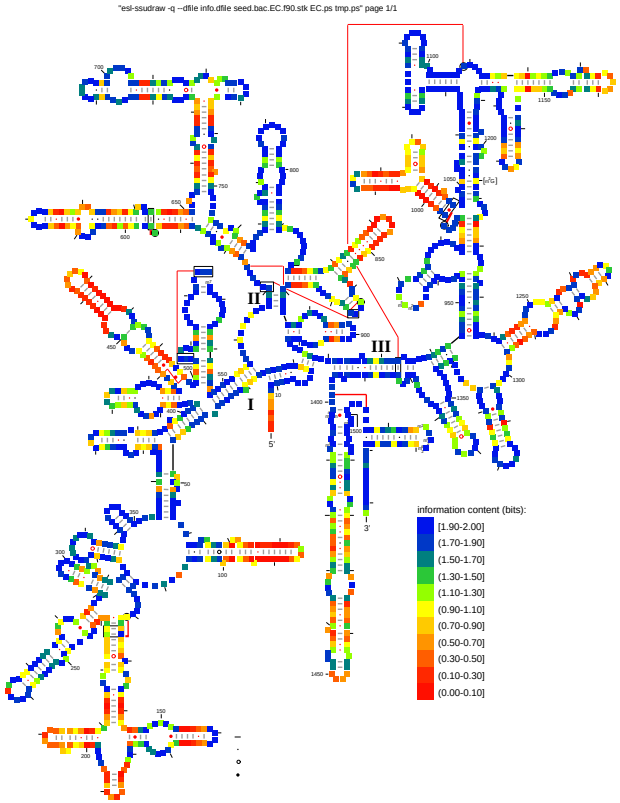
Zea mays



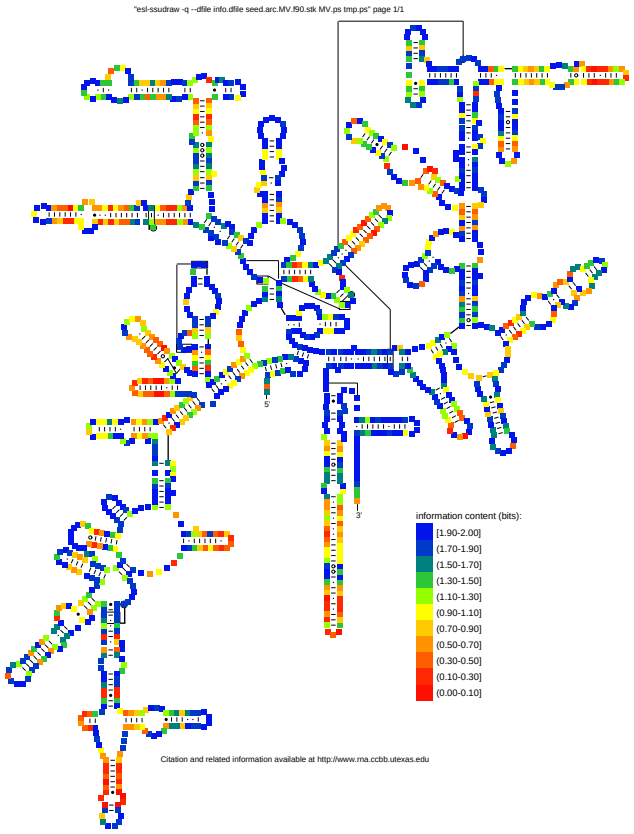
Secondary structure diagrams from:
URL:http://www.rna.ccbb.utexas.edu/

Sequence conservation in SSU rRNA

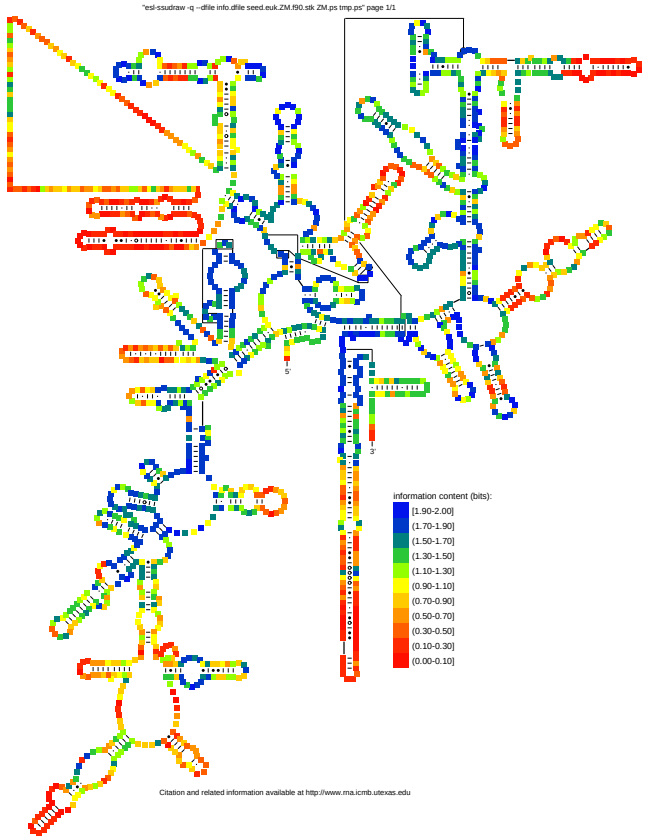
bacteria



archaea



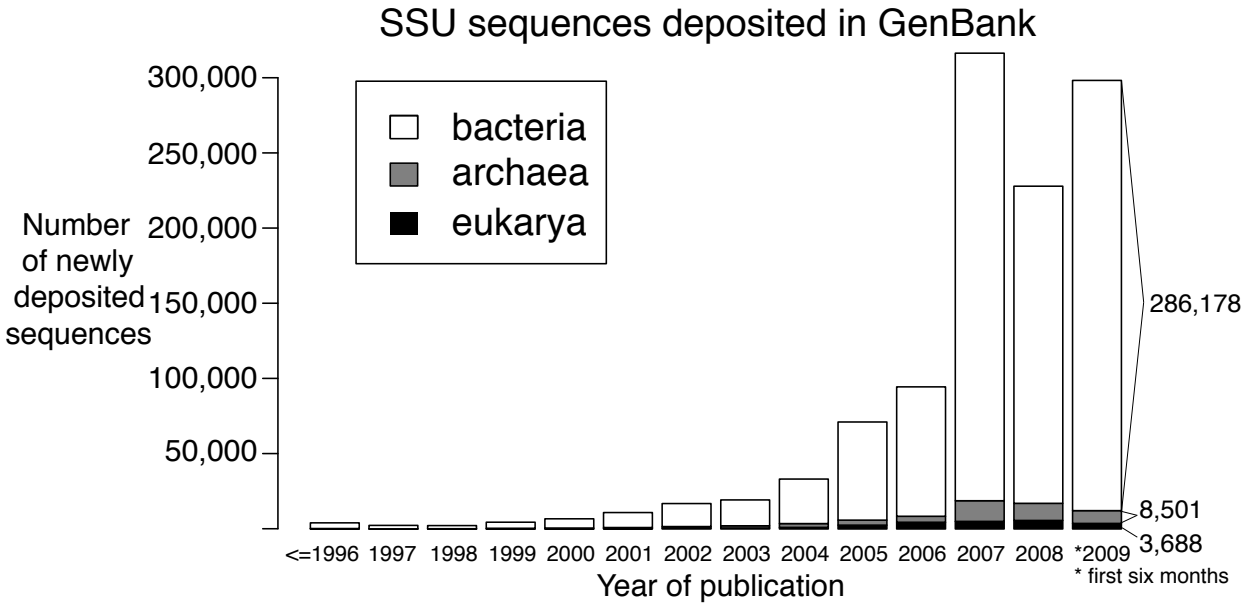
eukarya

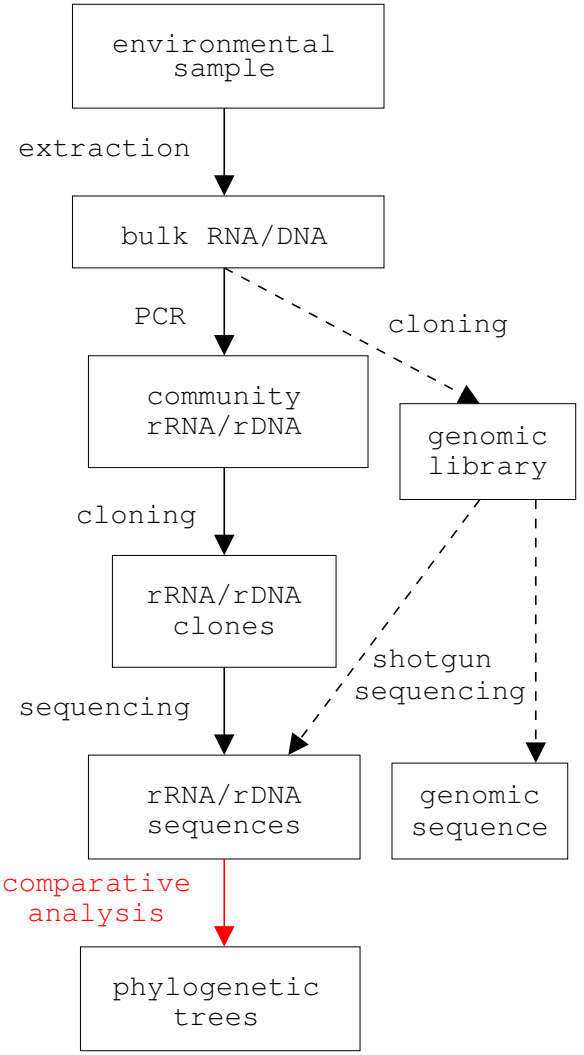


Secondary structure diagrams created based alignments and diagrams from:
URL:http://www.rna.ccbb.utexas.edu/

Environmental surveys target SSU

- mid 1980s - Norman Pace develops methodology for determination of SSU sequences without cultivation
- "the great plate-count anomaly" - vast majority of microbial species cannot be cultivated
- environmental surveys have become common
  - many different environments have been studied
  - commonly expand known biodiversity
    - recognized bacterial phyla: 11 in 1987, 36 in 1998, 52 in 2003, 67 in 2006...



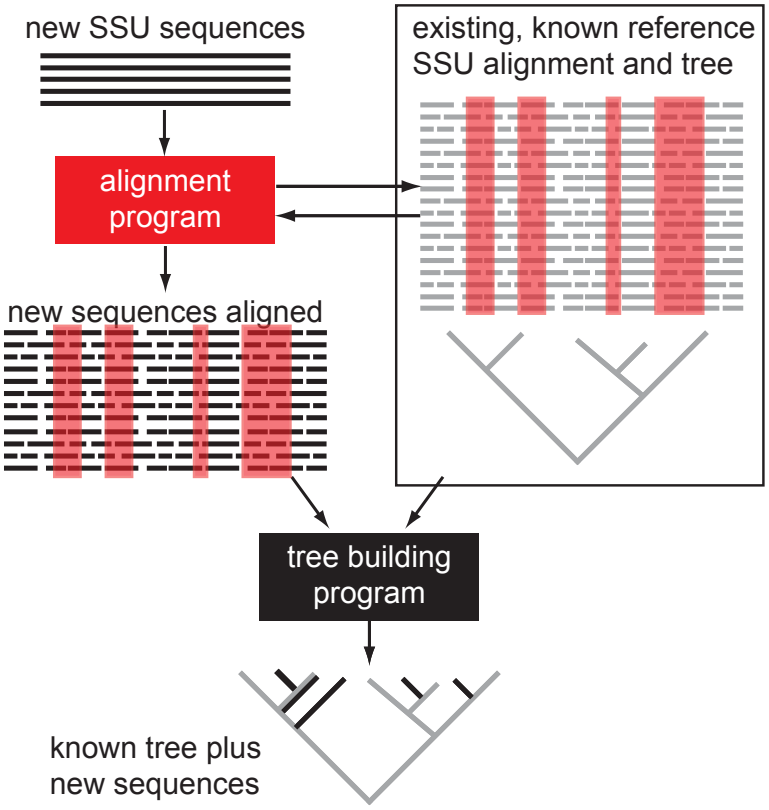


adapted from: Hugenholtz, Genome Biology:2002 3(2)

Environmental surveys target SSU rRNA

Two types of questions:
What organisms (known/unknown) are in my sample?
What is the phylogeny of a set of organisms?

Main assumption:
SSU gene tree approximates organismal tree.



Goals of the alignment program:

accurate: b/c alignment errors confound phylogenetic inference

fast and scalable: to handle up to millions of seqs

flexible: to be useful for all 3 domains

Sampling of recent SSU studies

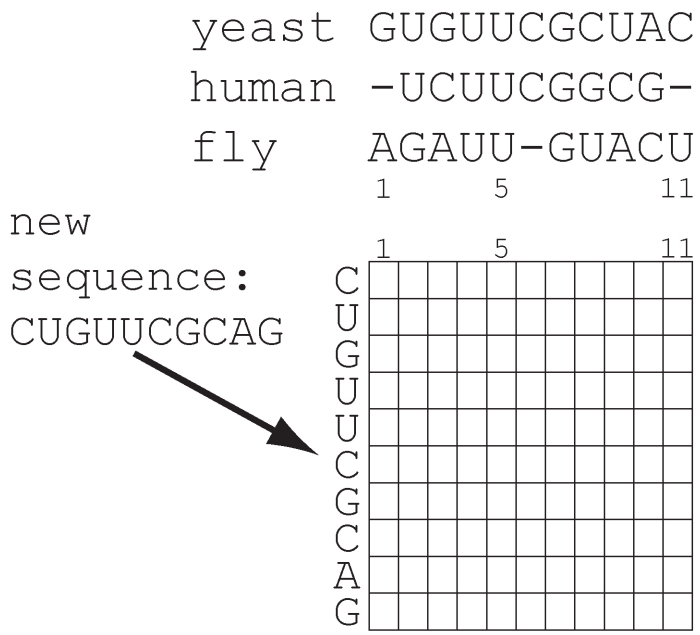
environment/ phylogeny	domain(s)	#seqs	year
soil, many others	bacteria	21,752	2007
cecal microbiota of mice	bacteria	5,088 4,157	2005 2006
Sargasso sea	all 3	1,164	2004
hydrothermal vents	eukarya	374	2002
endolithic environment (pore space of rocks)	archaea, bacteria	342 588	2005 2007
oxidized iron deposits, marine tidal mat, microbial steamers	bacteria	308	2004
soil & burrow casts of earthworms	archaea, bacteria	204	2002
tidal flat sediment	archaea	90	2005
salt marsh	eukarya	79	2003
dipteran hindgut	bacteria	59	2007
bumble bee phylogeny	eukarya	~200	2007
anaplasma phylogeny	bacteria	21	2003
protostome phylogeny	eukarya	20	2002

Accelerating CM alignment using HMMs

- **main idea:** use fast HMM when it's accurate, appealing to CM when it's not
- need some type of measure of confidence in regions of the HMM alignment

HMM alignment

- each column of the grid corresponds to a column of the seed alignment
- each row of the grid corresponds to a position of the new sequence

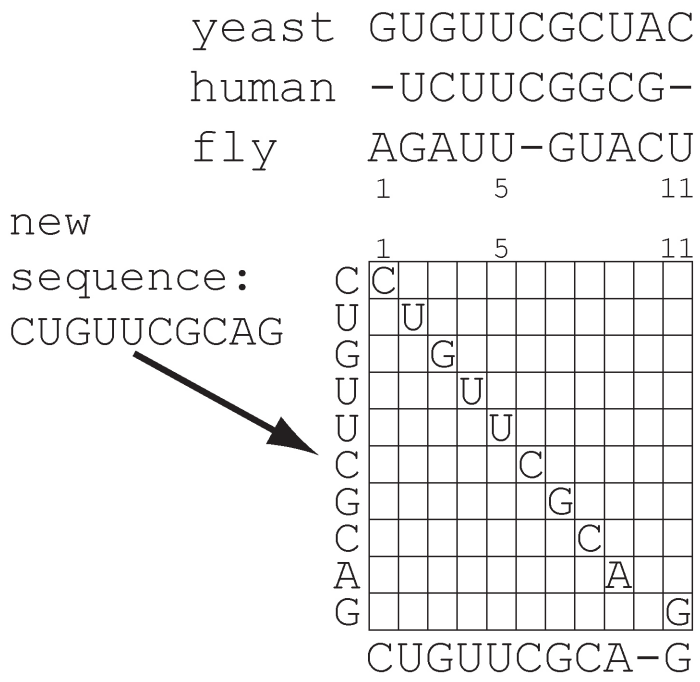


Accelerating CM alignment using HMMs

- **main idea:** use fast HMM when it's accurate, appealing to CM when it's not
- need some type of measure of confidence in regions of the HMM alignment

HMM alignment

- each column of the grid corresponds to a column of the seed alignment
- each row of the grid corresponds to a position of the new sequence



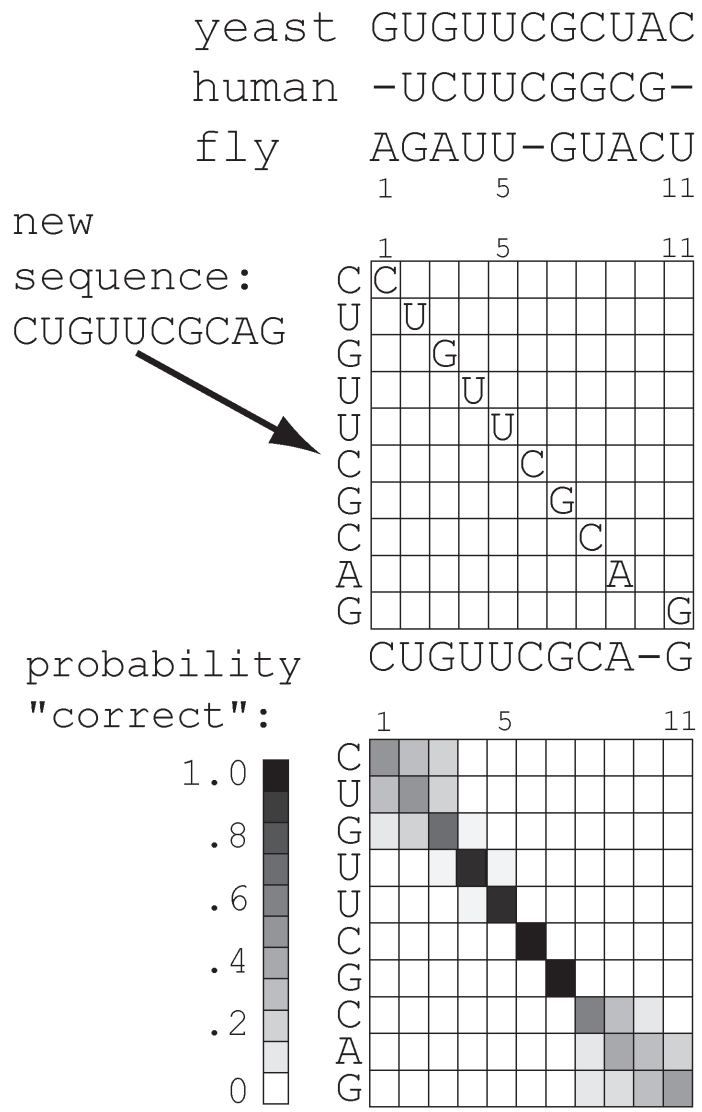
Accelerating CM alignment using HMMs

- **main idea:** use fast HMM when it's accurate, appealing to CM when it's not
- need some type of measure of confidence in regions of the HMM alignment

HMM alignment

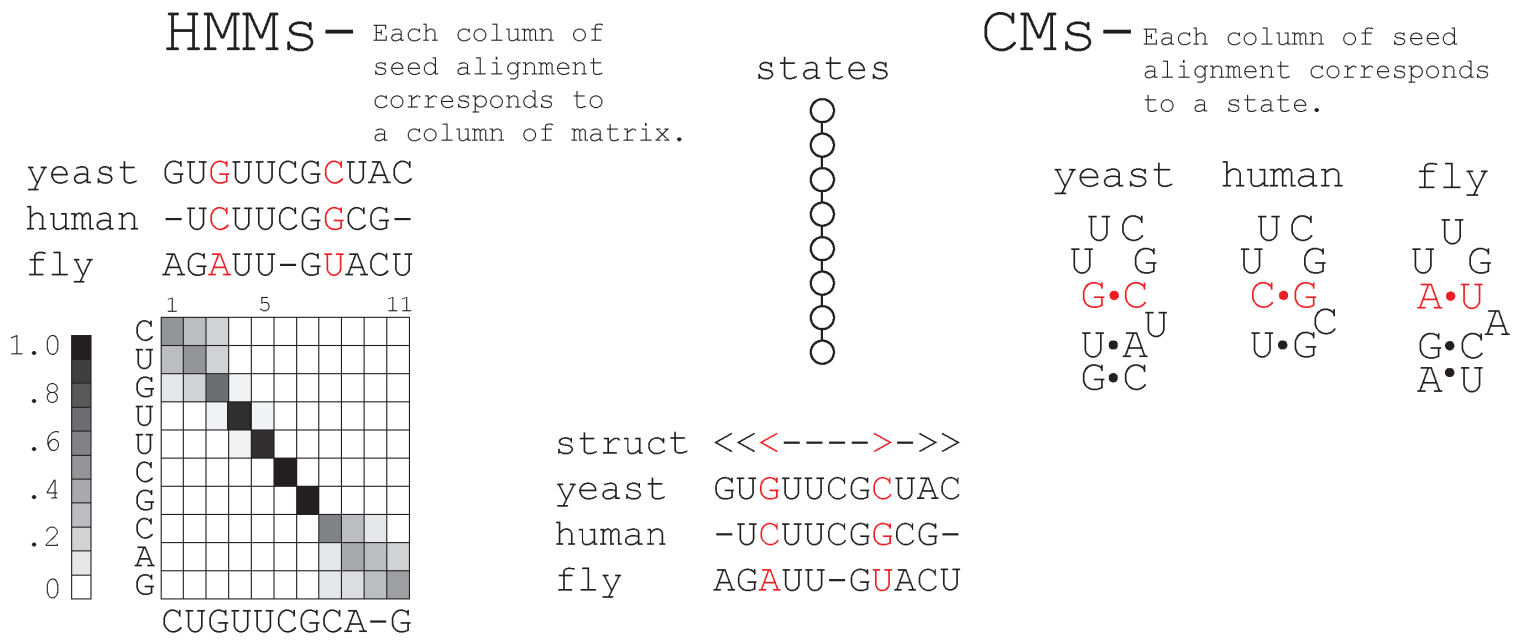
- each column of the grid corresponds to a column of the seed alignment
- each row of the grid corresponds to a position of the new sequence

How can we use this information during CM alignment?



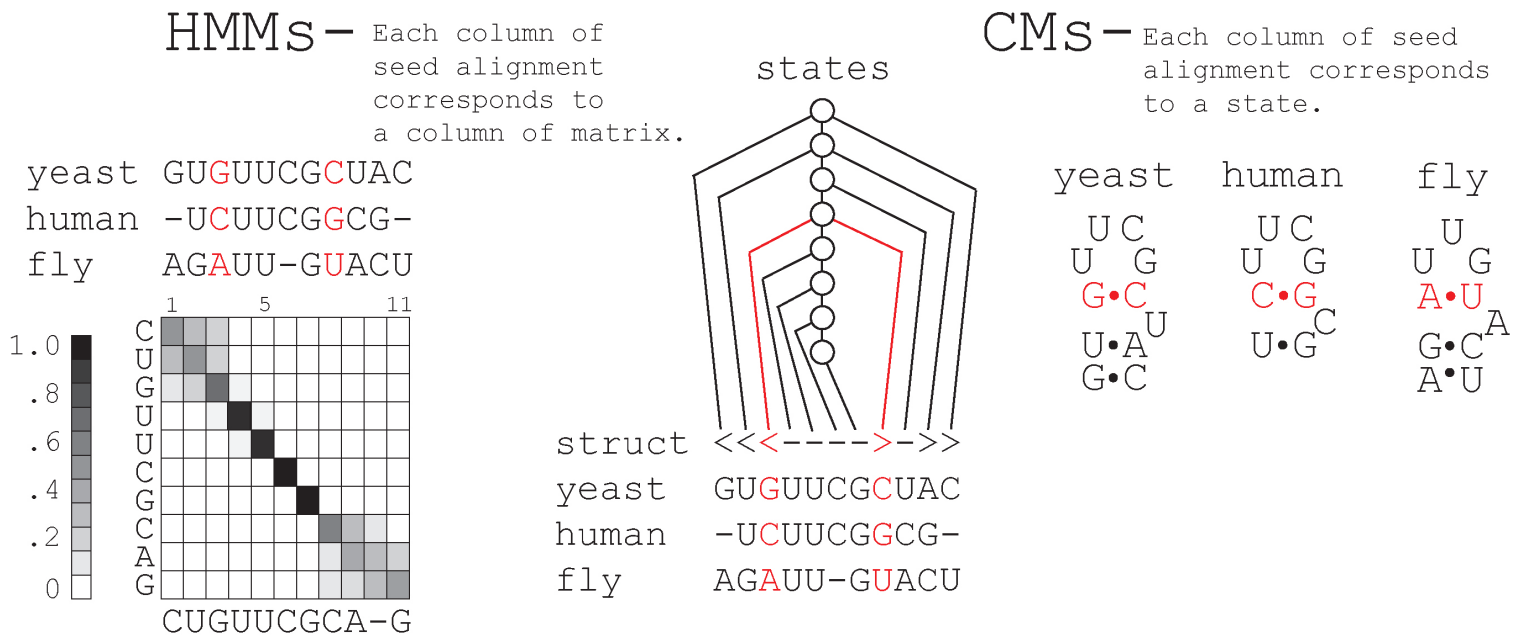
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



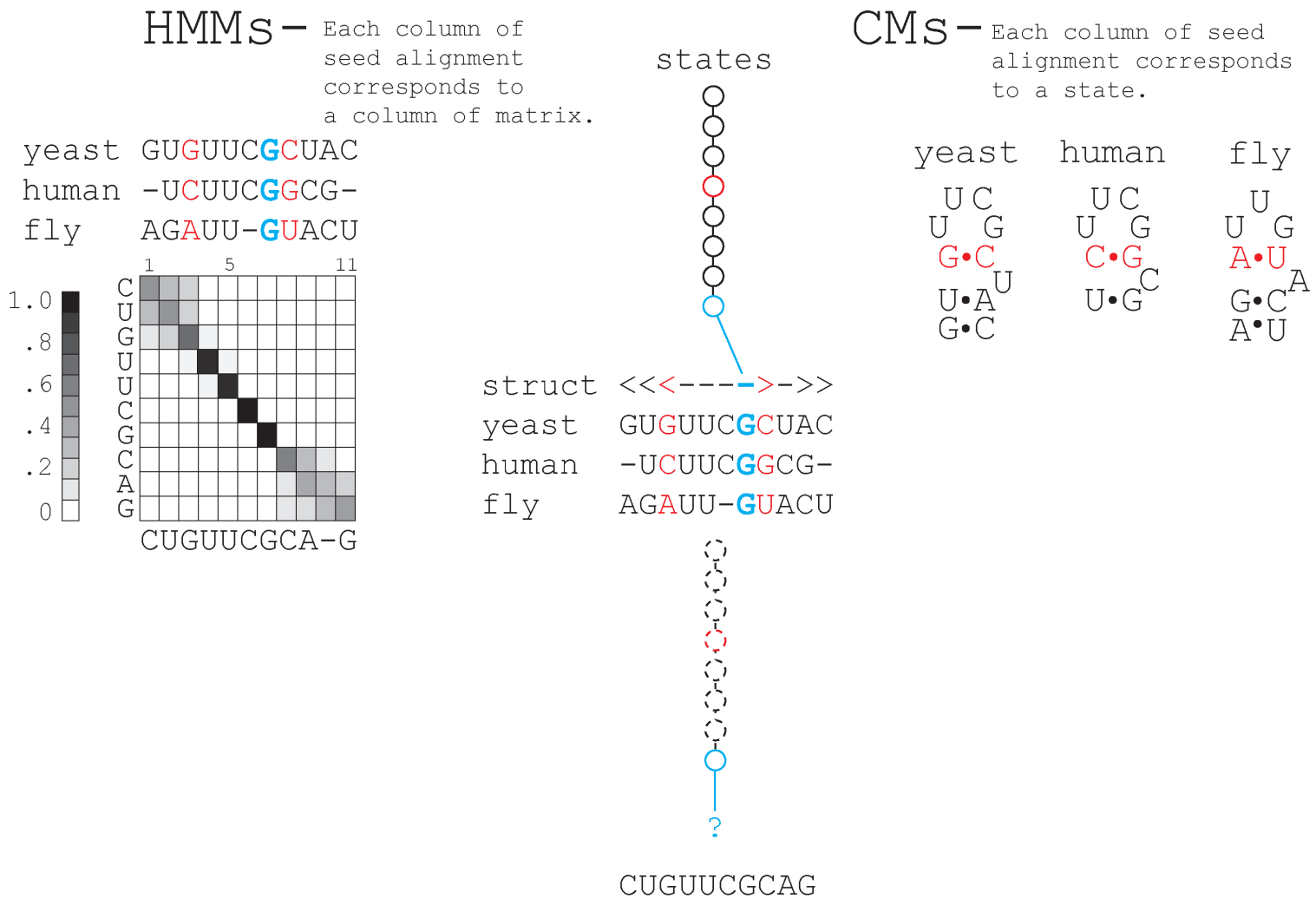
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



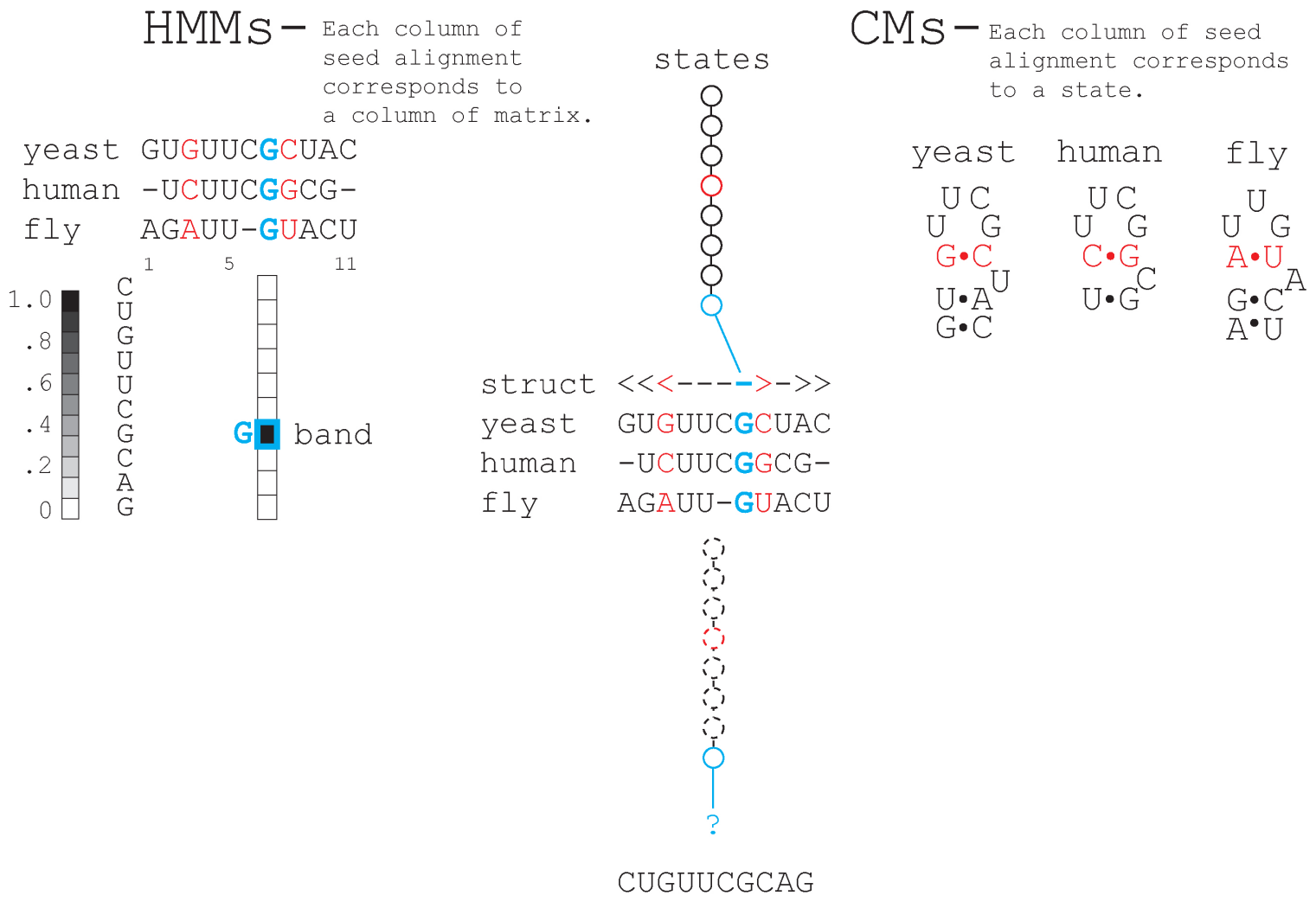
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



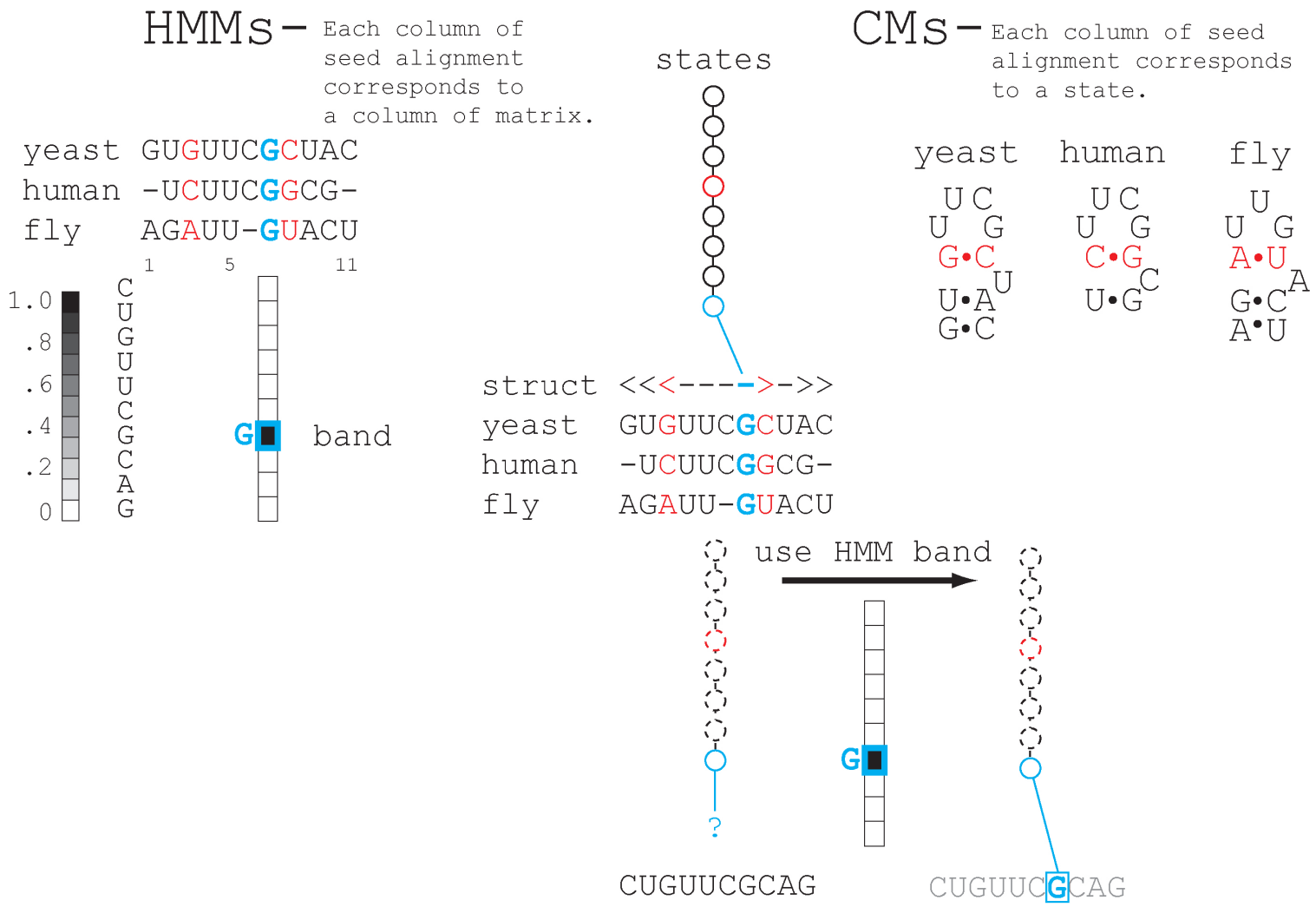
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



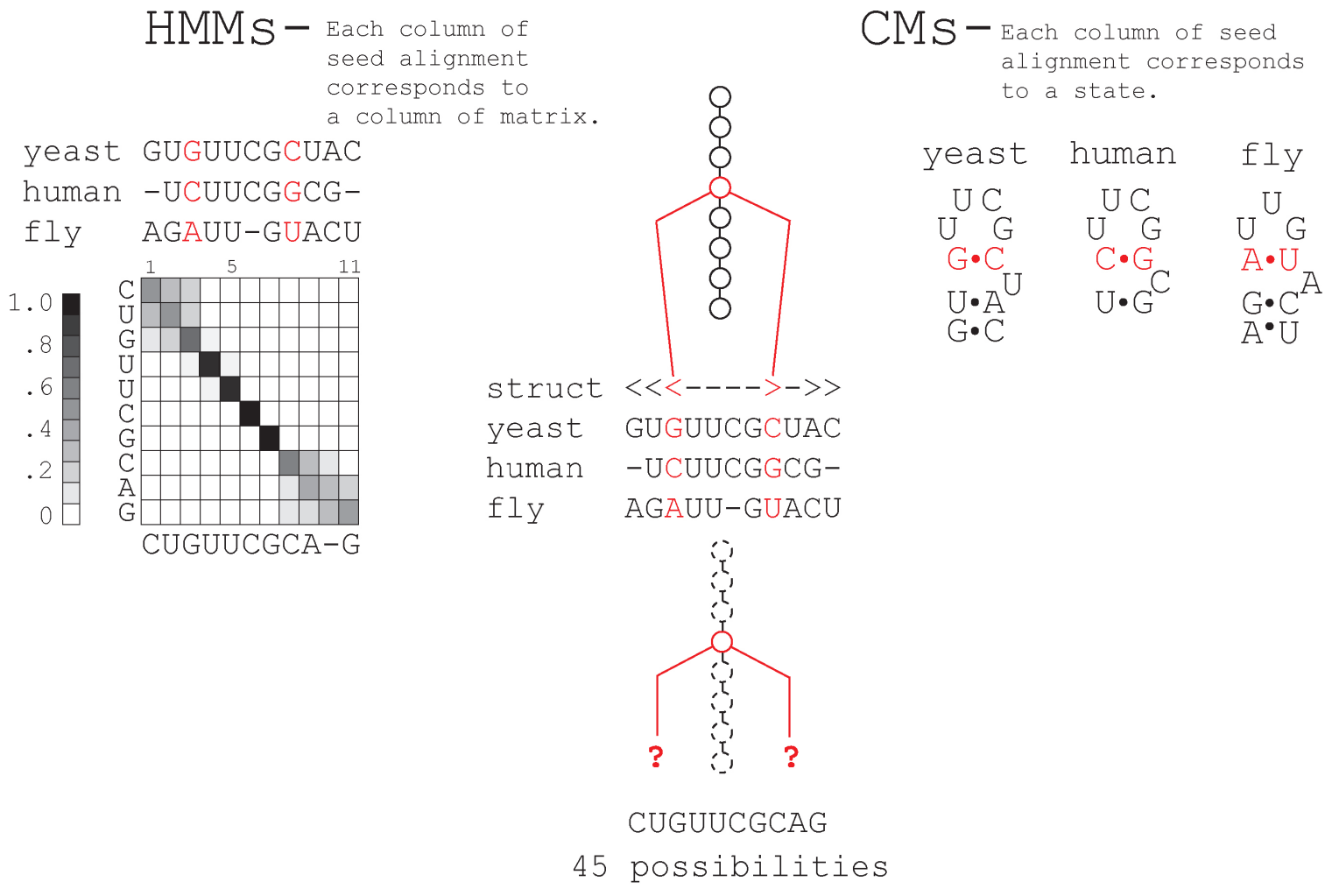
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



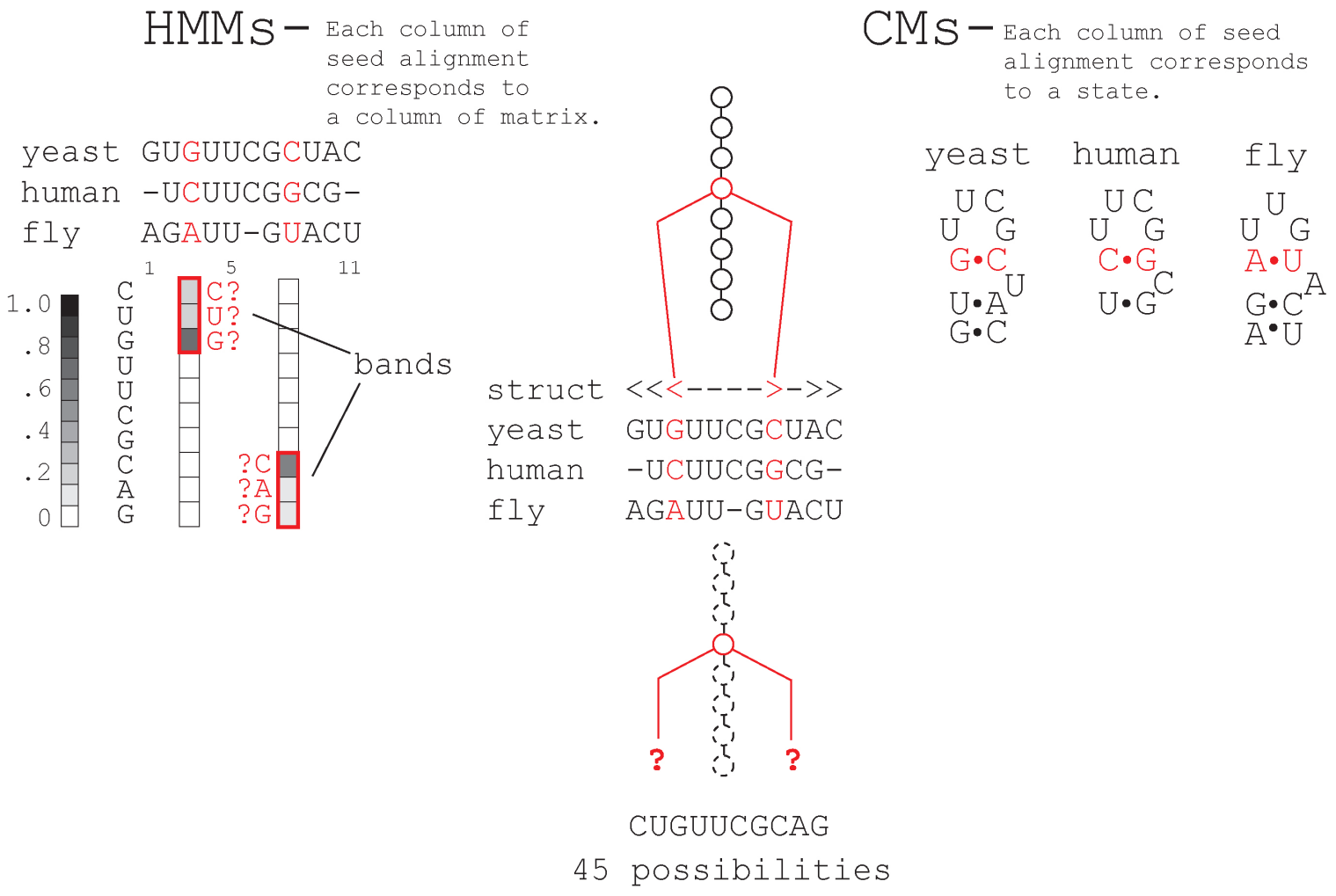
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



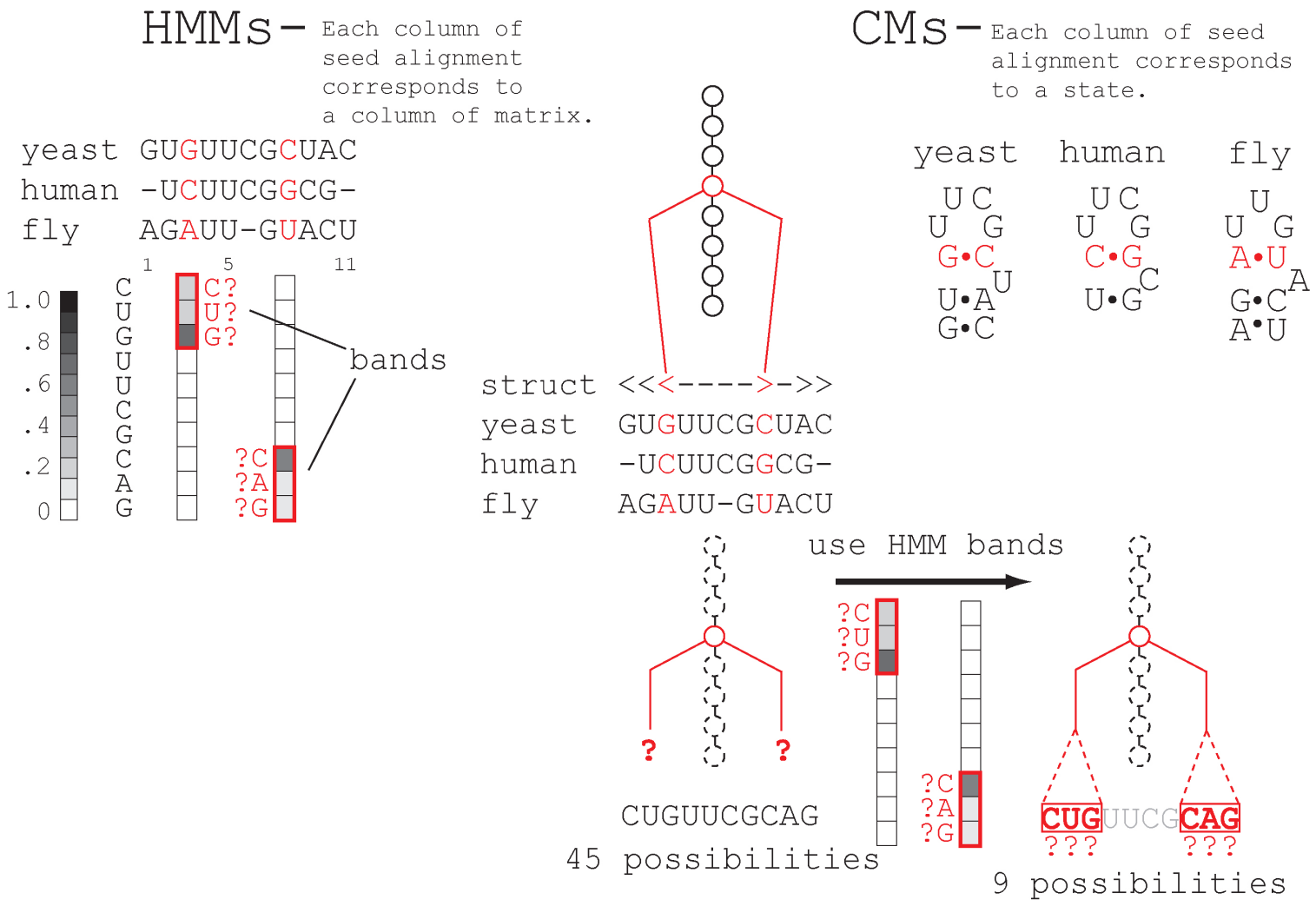
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



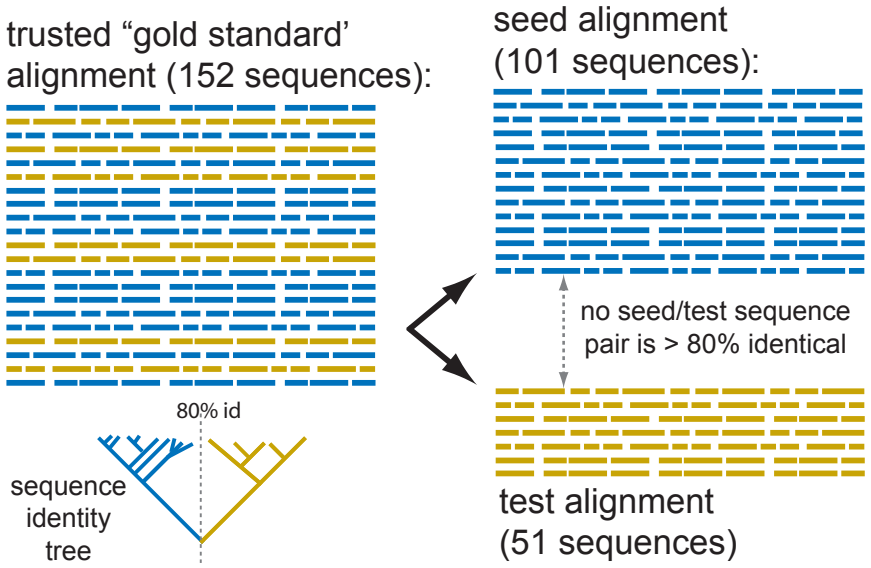
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



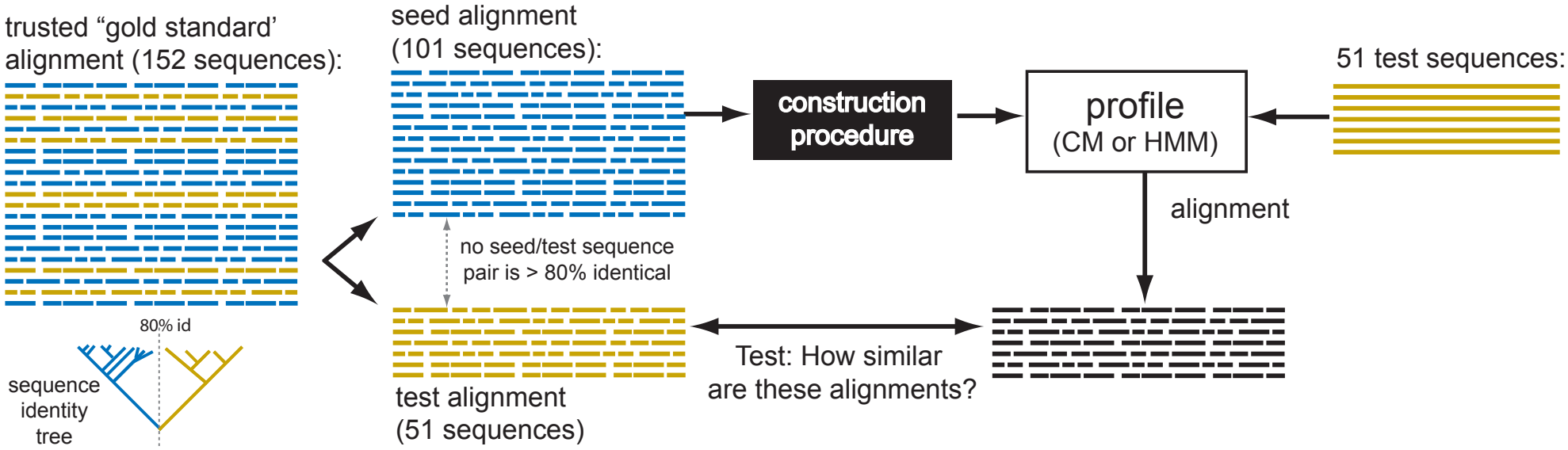
Benchmarking SSU alignment

- Does the banded CM approach sacrifice accuracy relative to non-banded CM alignment?
- 'Gold standard' testing dataset
  - structural alignment of 152 bacterial SSU sequences from Robin Gutell's database
  - this is the CRW bacterial seed alignment filtered to 92% identity
  - determined by 'manual' comparative analysis



Benchmarking SSU alignment

- Does the banded CM approach sacrifice accuracy relative to non-banded CM alignment?
- 'Gold standard' testing dataset
  - structural alignment of 152 bacterial SSU sequences from Robin Gutell's database
  - this is the CRW bacterial seed alignment filtered to 92% identity
  - determined by 'manual' comparative analysis



CMs are (slightly) more accurate, but much slower than HMMs

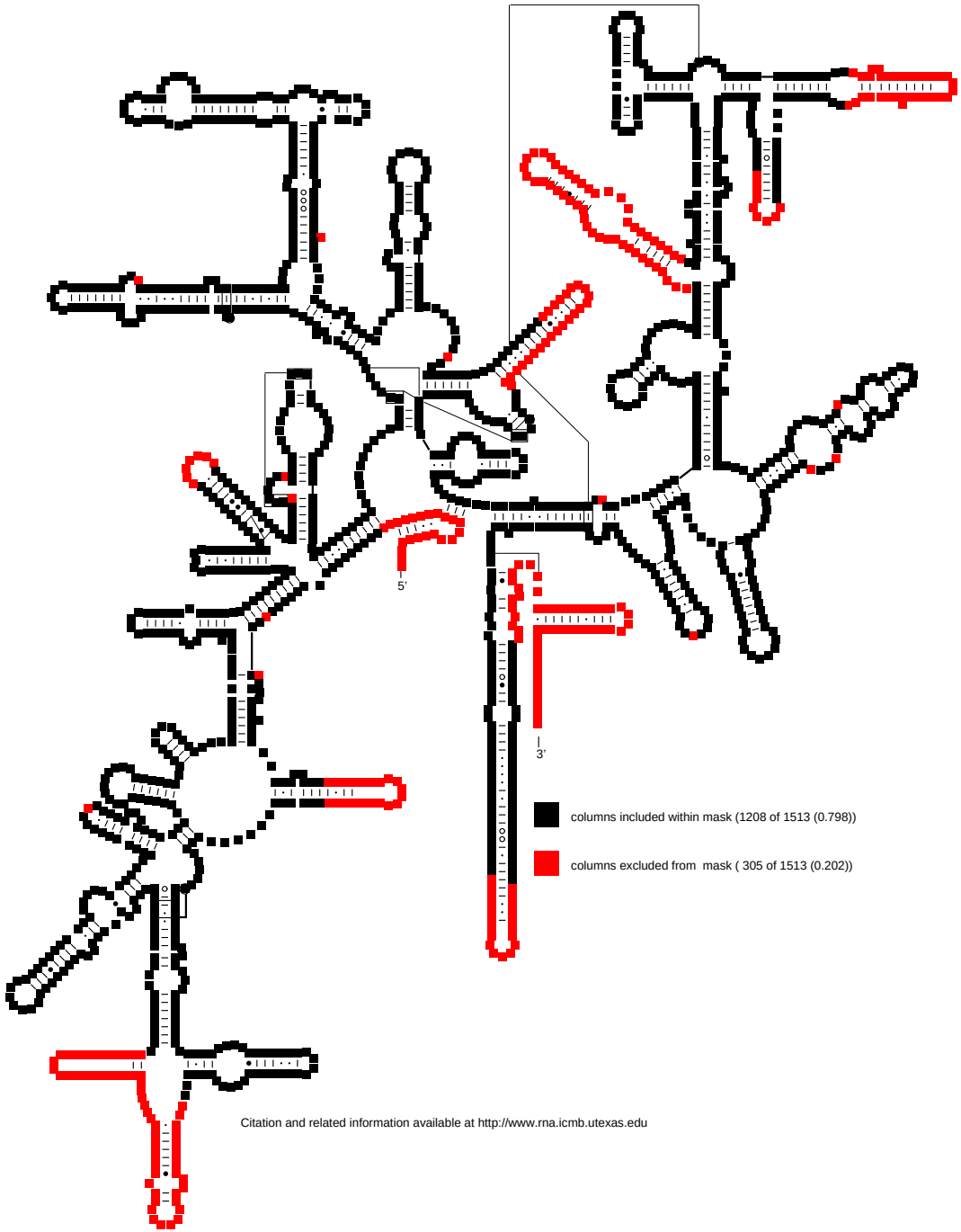
	alignment accuracy	time (sec/seq)
clustalw	92.2%	30.0
HMMs	96.6%	0.08
non-banded CMs	98.1%	1321.5

HMM banding accelerates CM alignment 2000-fold

	alignment accuracy	time (sec/seq)
clustalw	92.2%	30.0
HMMs	96.6%	0.08
non-banded CMs	98.1%	1321.5
HMM banded CMs	98.1%	0.7

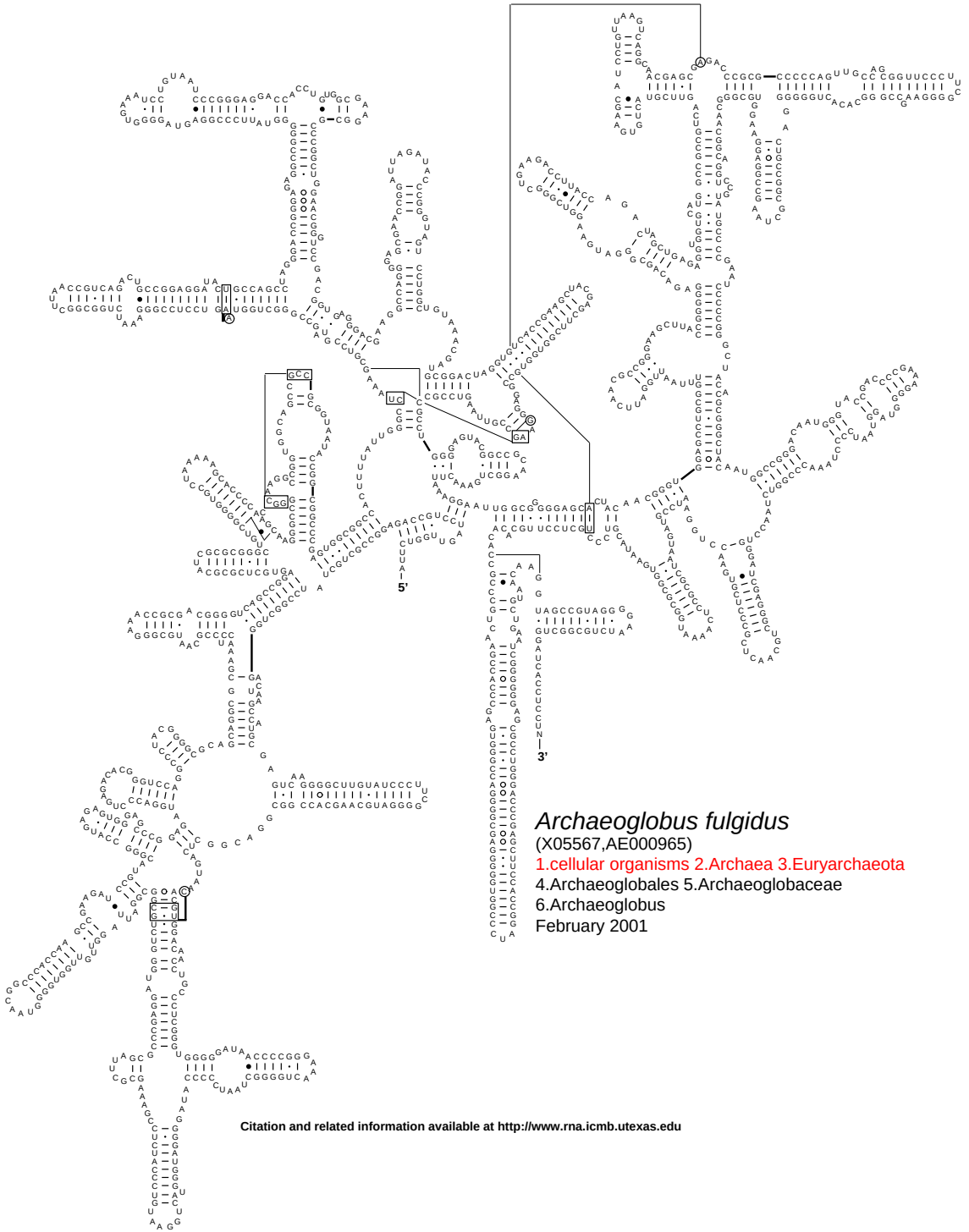
Phil Hugenholtz's manually created mask

"esl-ssudraw -q --mask-col inf.mask.from.nst.1208-1s.1513c.mask 1513.c.stk 1513.ps lmph_on_1513.ps" page 1/1



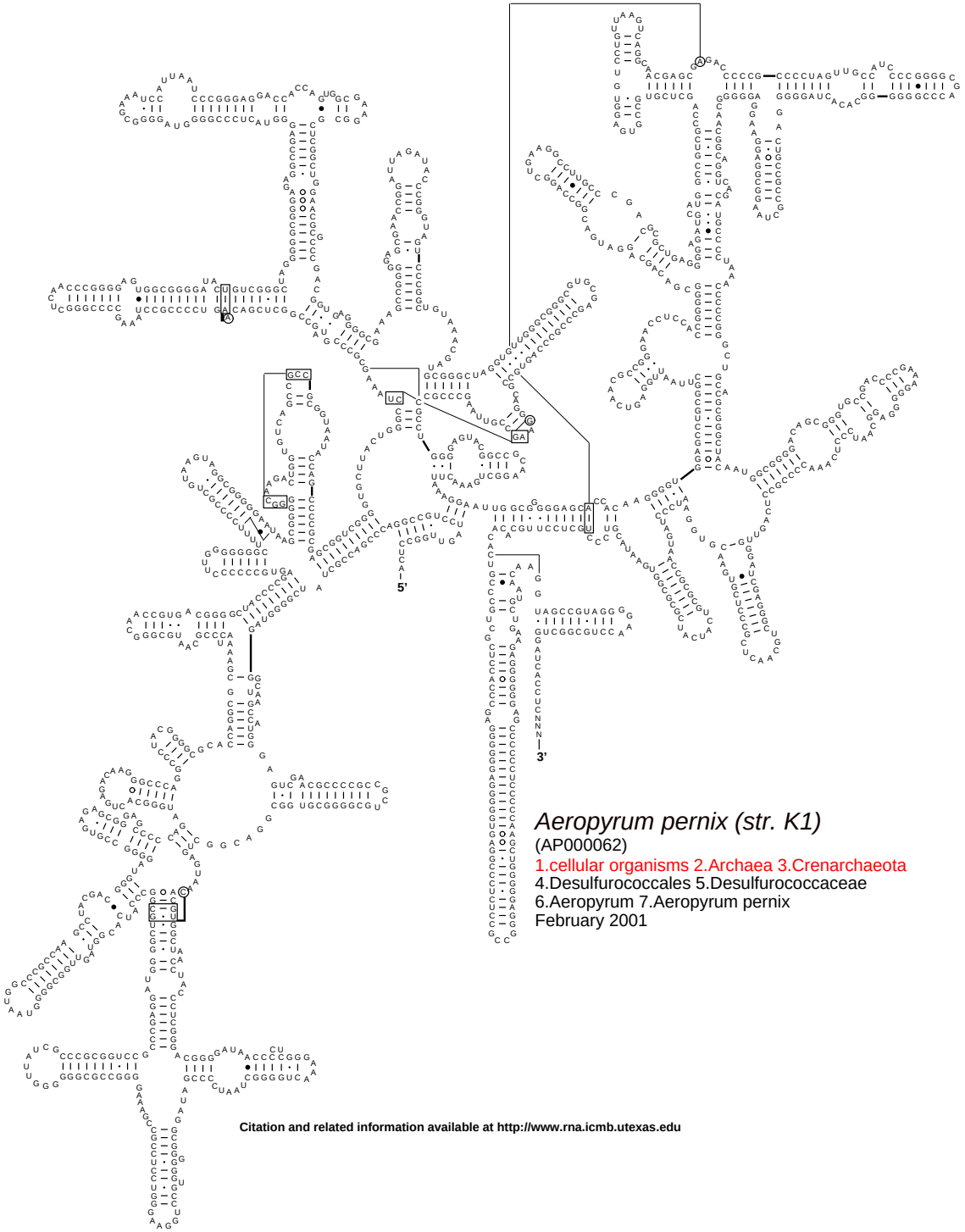
Citation and related information available at http://www.rna.icmb.utexas.edu

Secondary Structure: small subunit ribosomal RNA



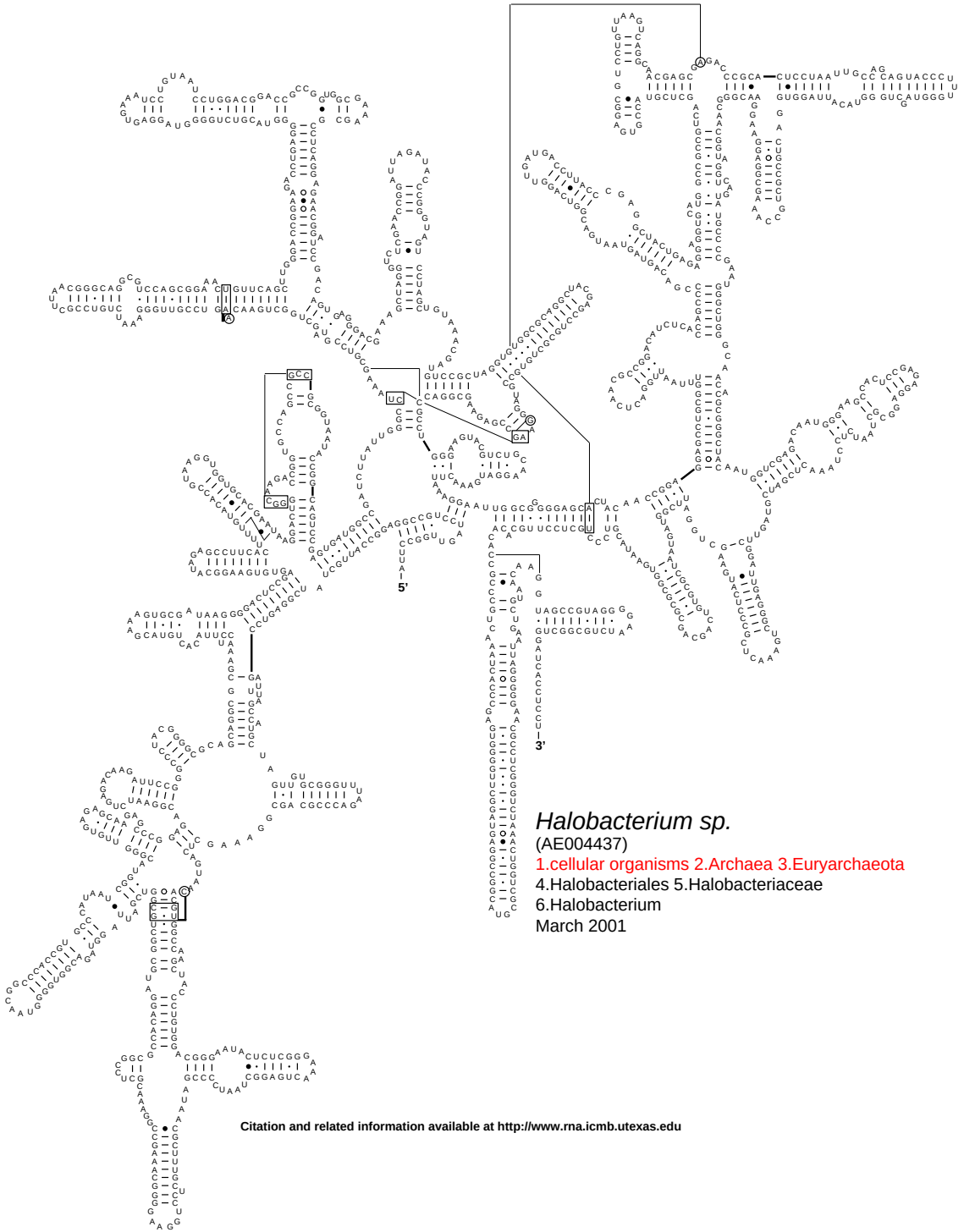
Archaeoglobus fulgidus
(X05567,AE000965)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Archaeoglobales 5.Archaeoglobaceae
6.Archaeoglobus
February 2001

Secondary Structure: small subunit ribosomal RNA



Aeropyrum pernix (str. K1)
(AP000062)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Desulfurococcales 5.Desulfurococcaceae
6.Aeropyrum 7.Aeropyrum pernix
February 2001

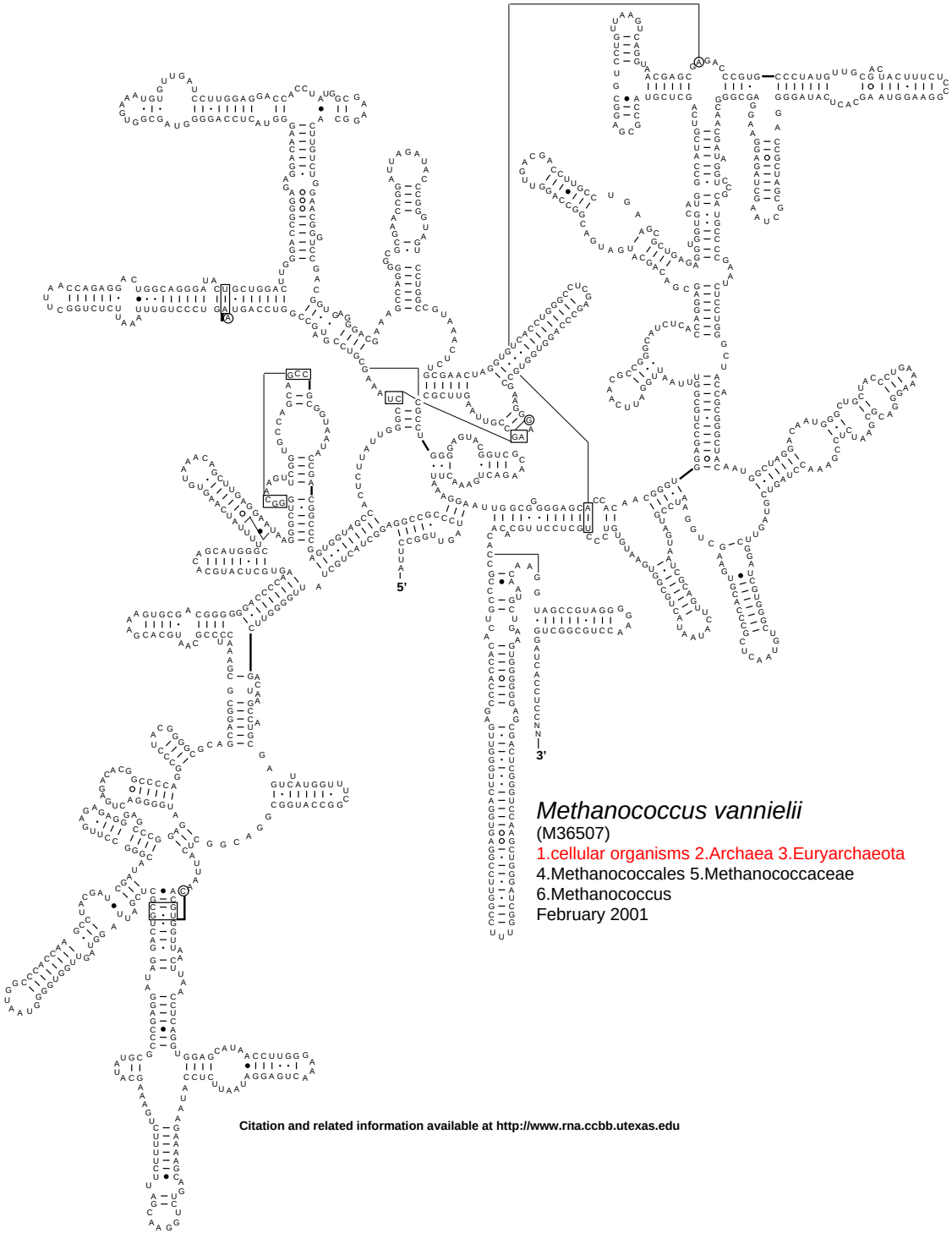
Secondary Structure: small subunit ribosomal RNA



Halobacterium sp.
(AE004437)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Halobacterium
March 2001

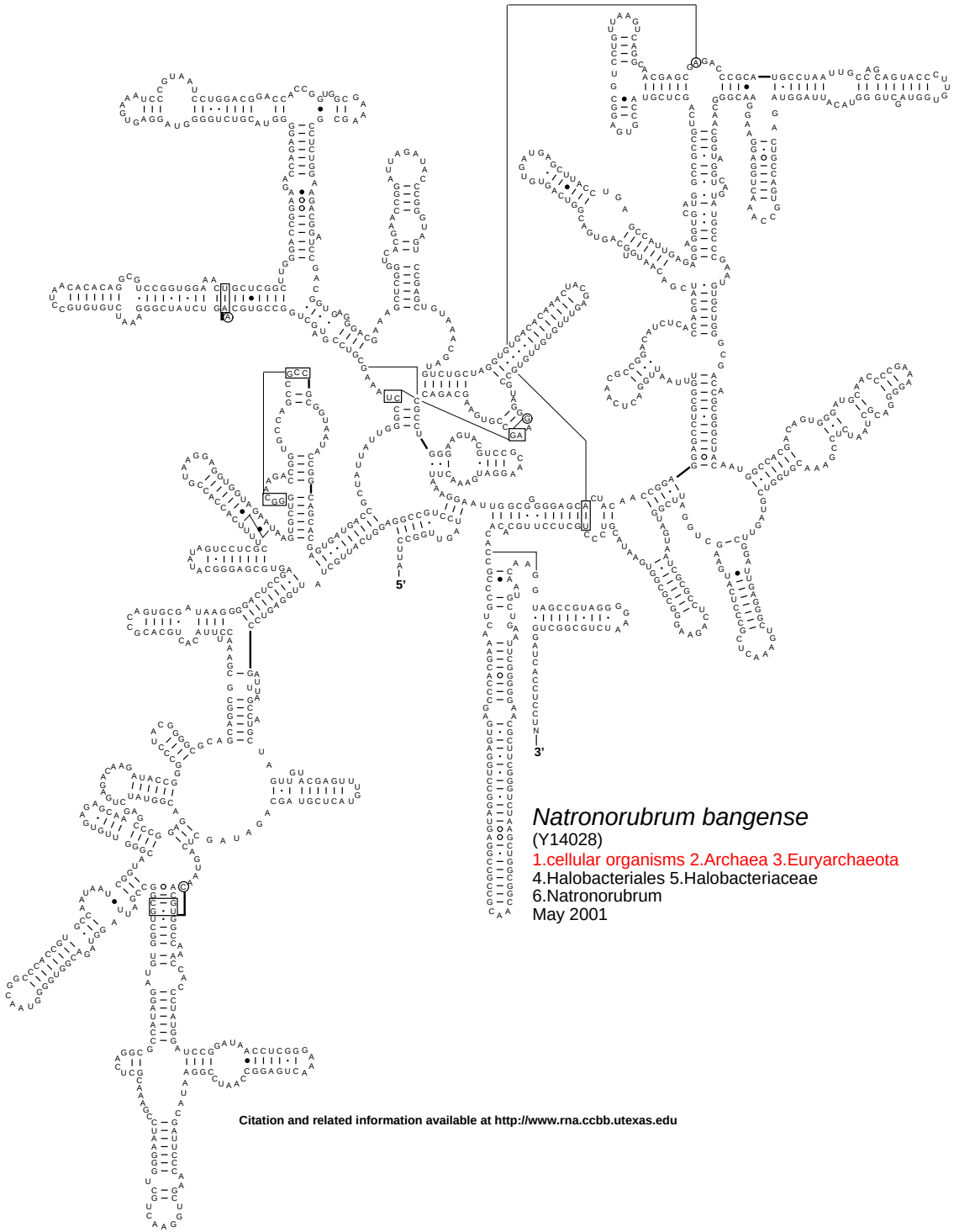
Citation and related information available at http://www.rna.icmb.utexas.edu

Secondary Structure: small subunit ribosomal RNA



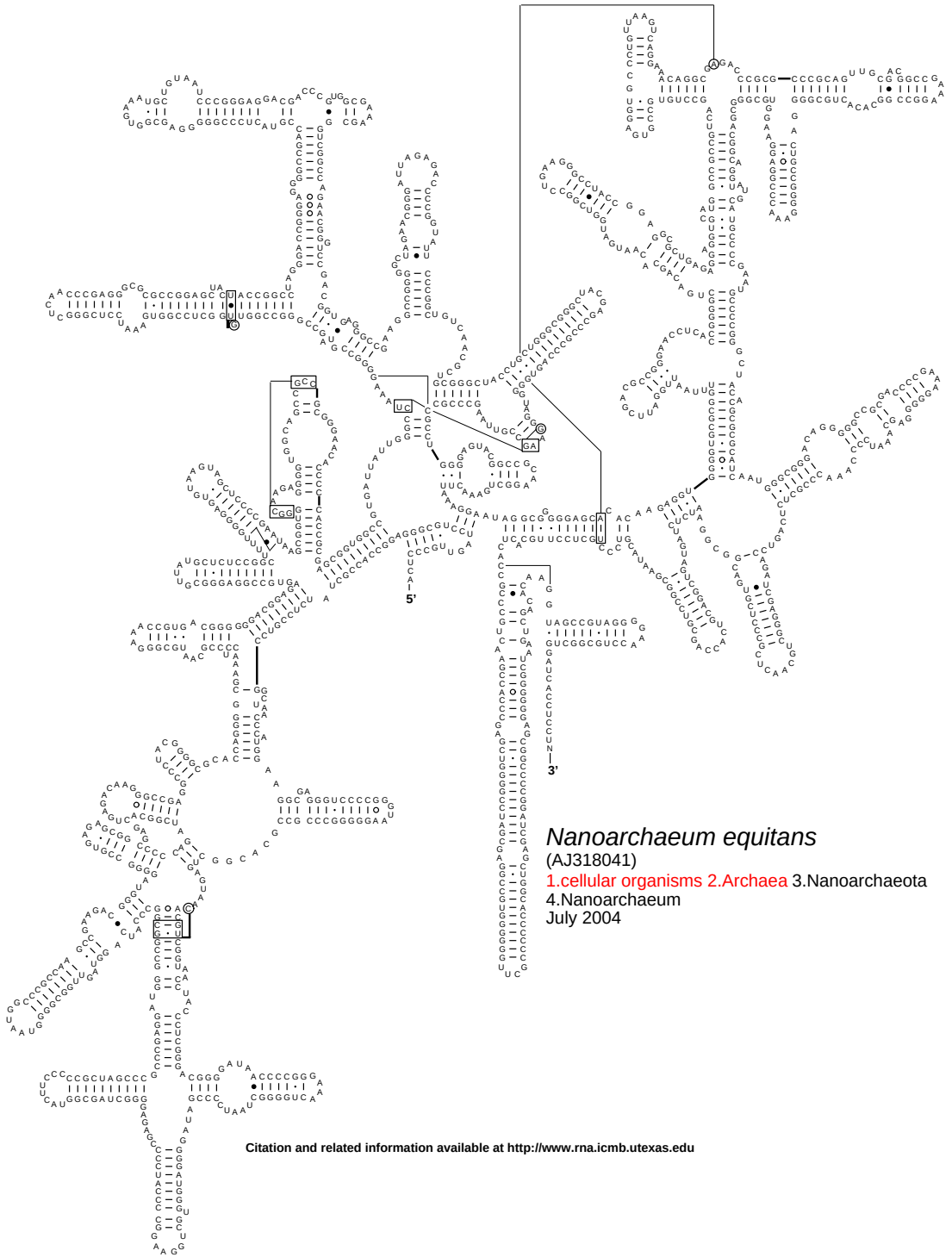
Methanococcus vannielii
(M36507)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Methanococcales 5.Methanococcaceae
6.Methanococcus
February 2001

Secondary Structure: small subunit ribosomal RNA



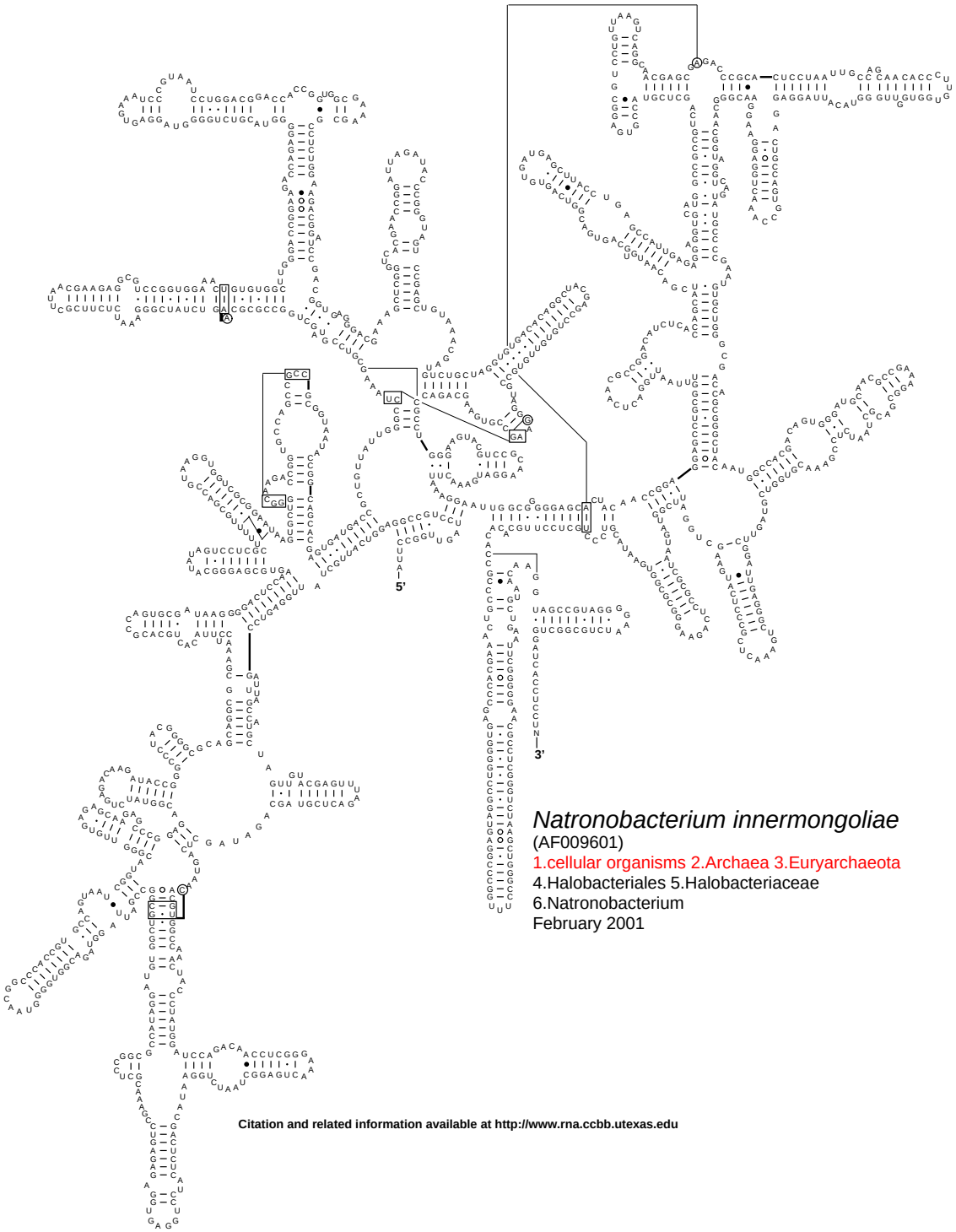
Natronorubrum bangense
(Y14028)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronorubrum
May 2001

Secondary Structure: small subunit ribosomal RNA



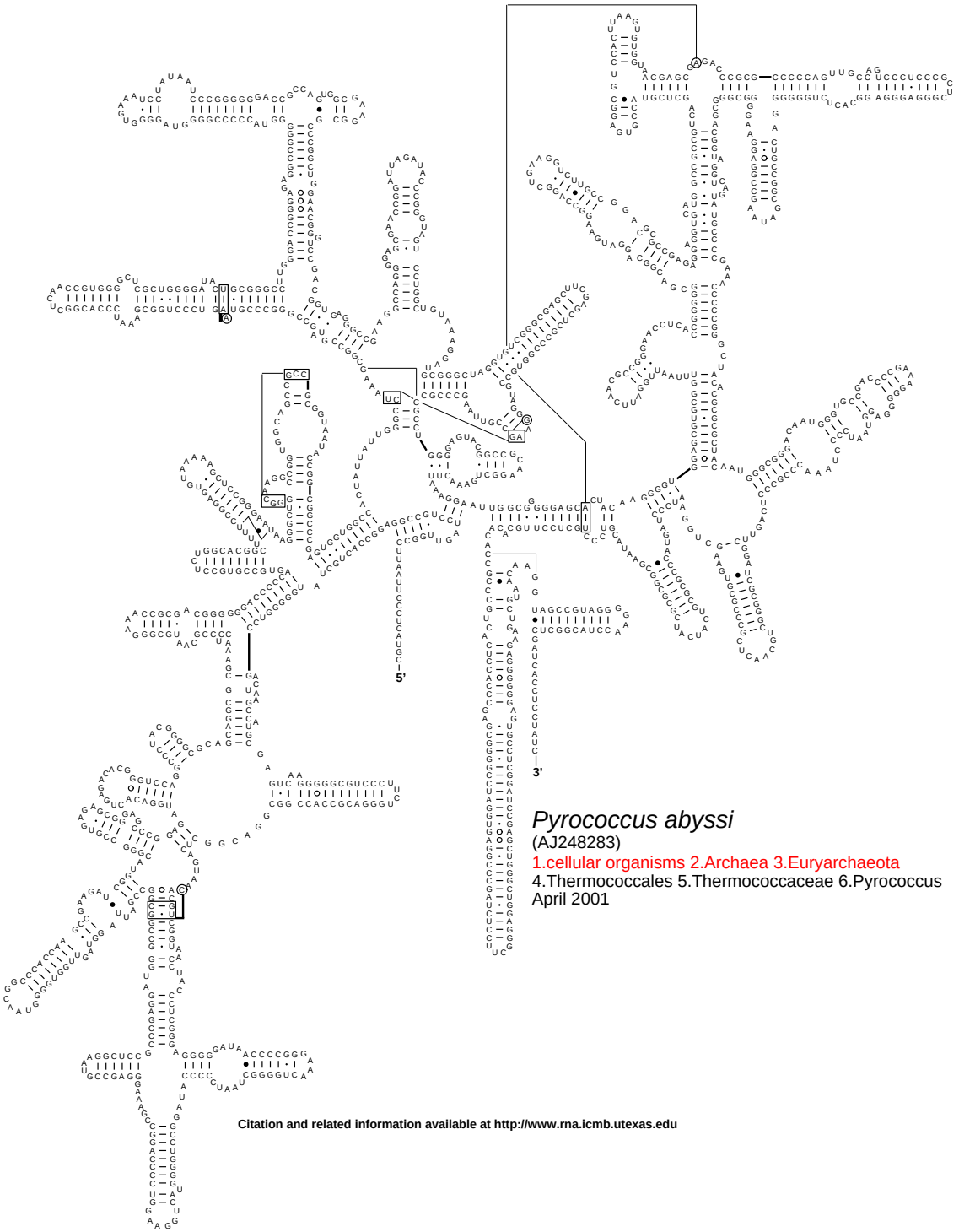
Nanoarchaeum equitans
(AJ318041)
1.cellular organisms 2.Archaea 3.Nanoarchaeota
4.Nanoarchaeum
July 2004

Secondary Structure: small subunit ribosomal RNA



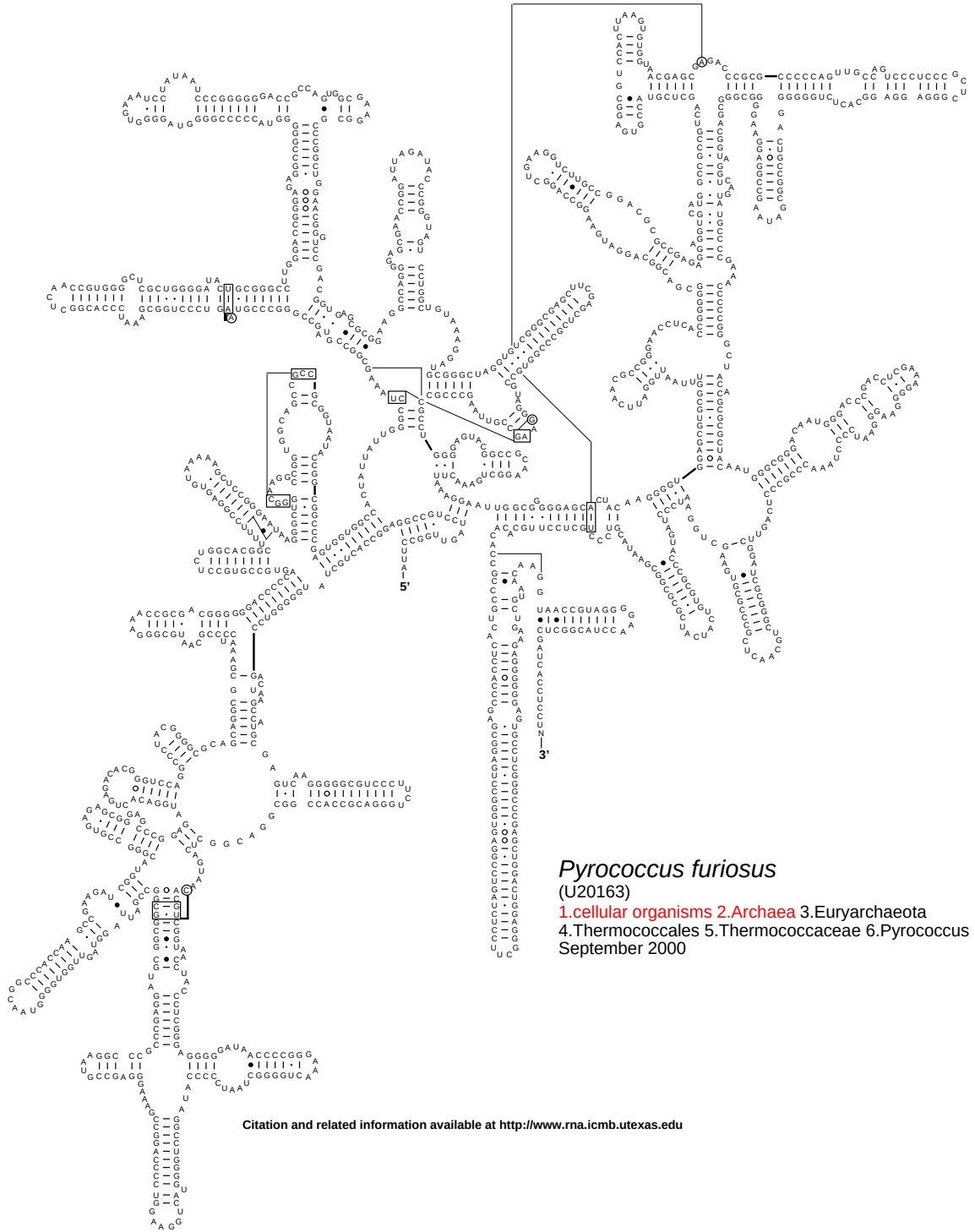
Natronobacterium innermongoliae
(AF009601)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronobacterium
February 2001

Secondary Structure: small subunit ribosomal RNA



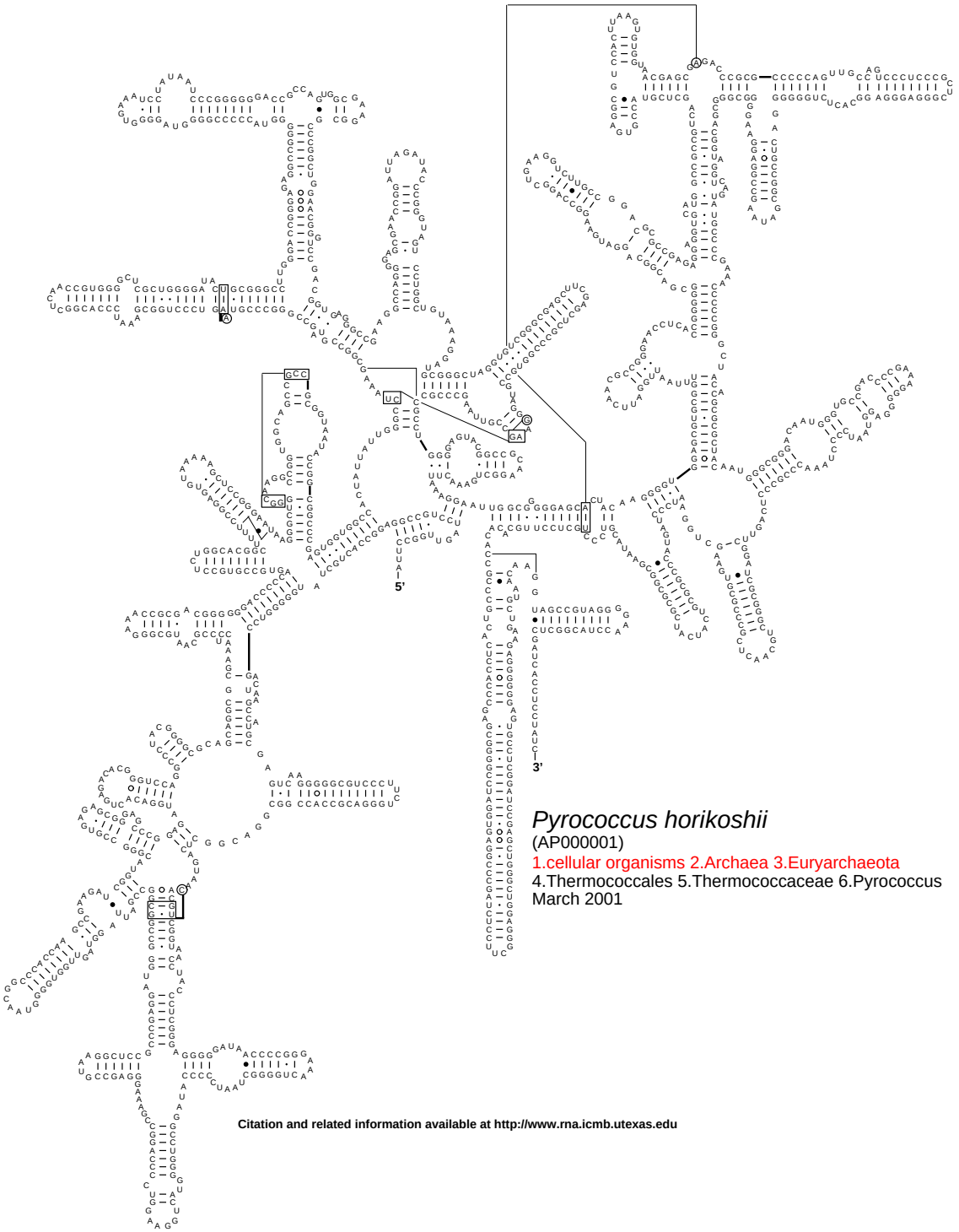
Pyrococcus abyssi
(AJ248283)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
April 2001

Secondary Structure: small subunit ribosomal RNA



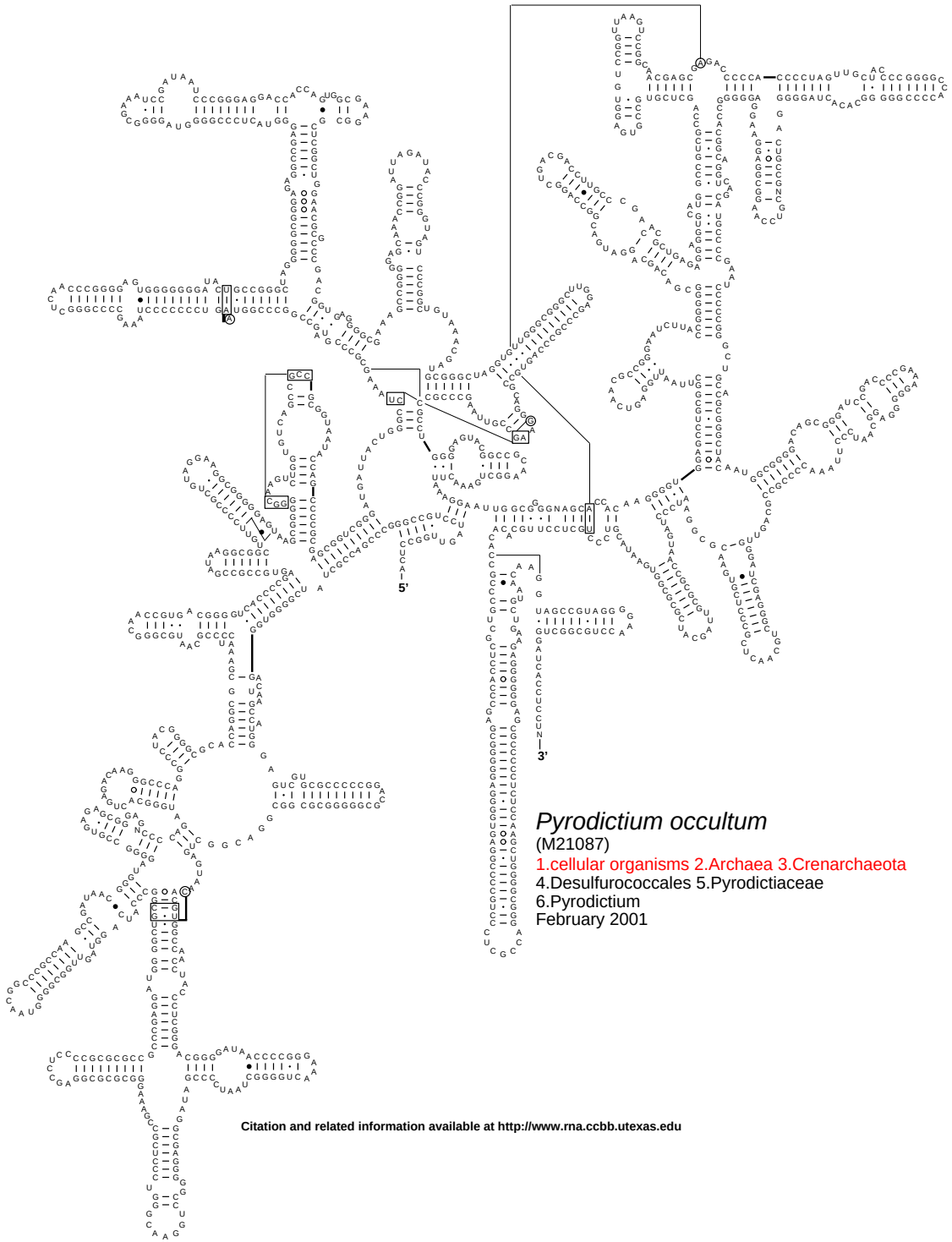
Pyrococcus furiosus
(U20163)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
September 2000

Secondary Structure: small subunit ribosomal RNA



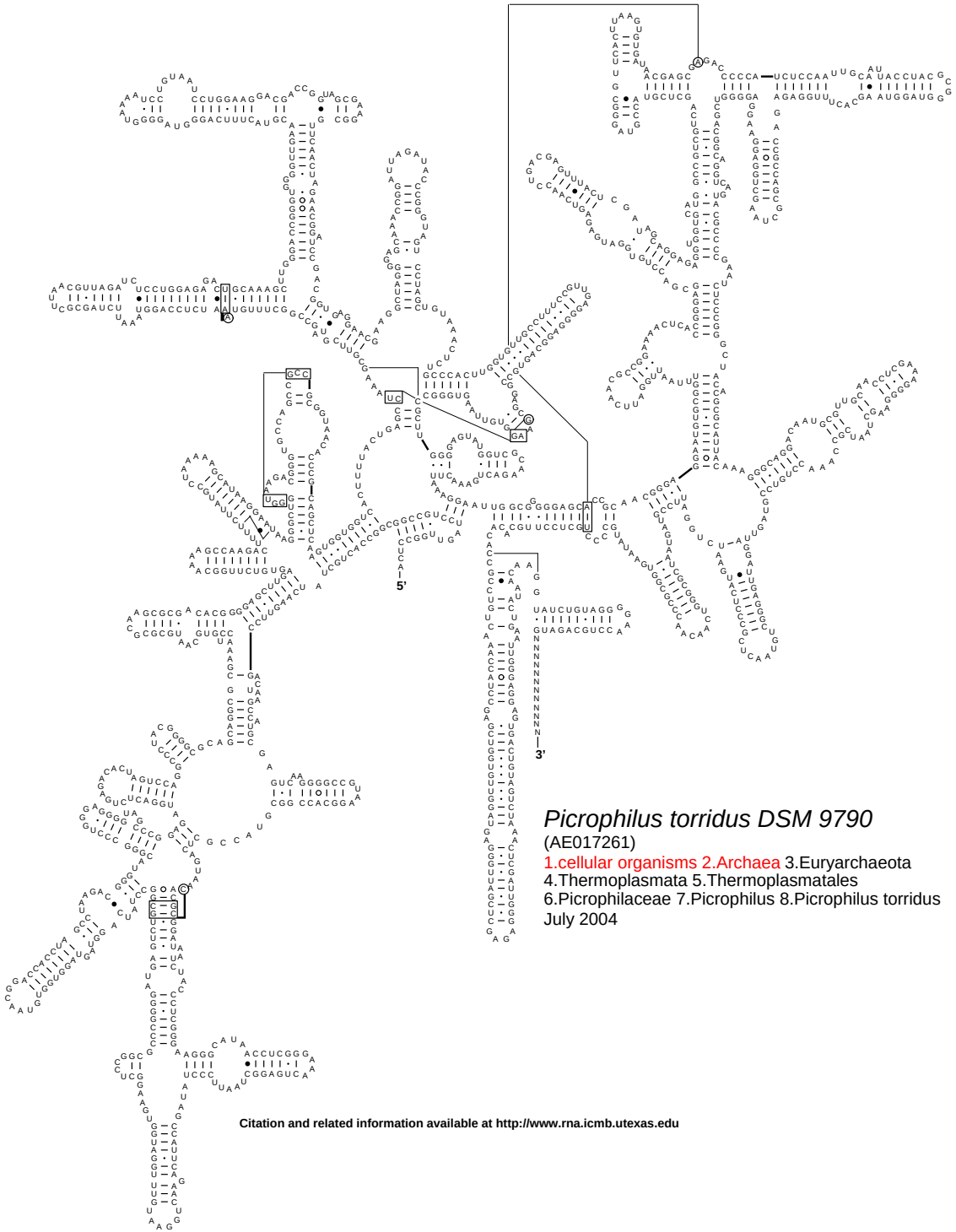
Pyrococcus horikoshii
(AP000001)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
March 2001

Secondary Structure: small subunit ribosomal RNA



Pyrodictium occultum
(M21087)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Desulfurococcales 5.Pyrodictiaceae
6.Pyrodictium
February 2001

Secondary Structure: small subunit ribosomal RNA



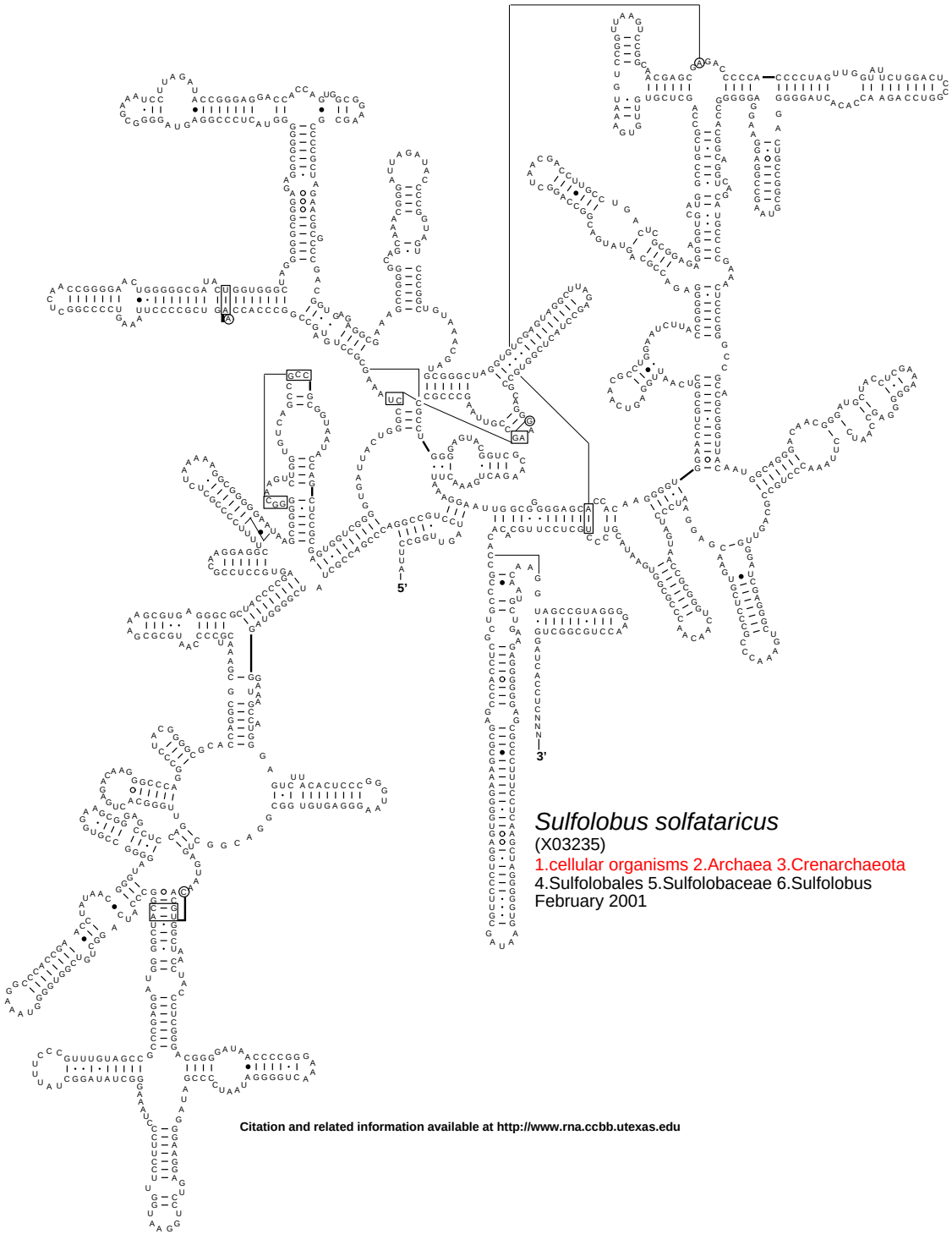
Picrophilus torridus DSM 9790

(AE017261)

1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermoplasmata 5.Thermoplasmatales
6.Picrophilaceae 7.Picrophilus 8.Picrophilus torridus

July 2004

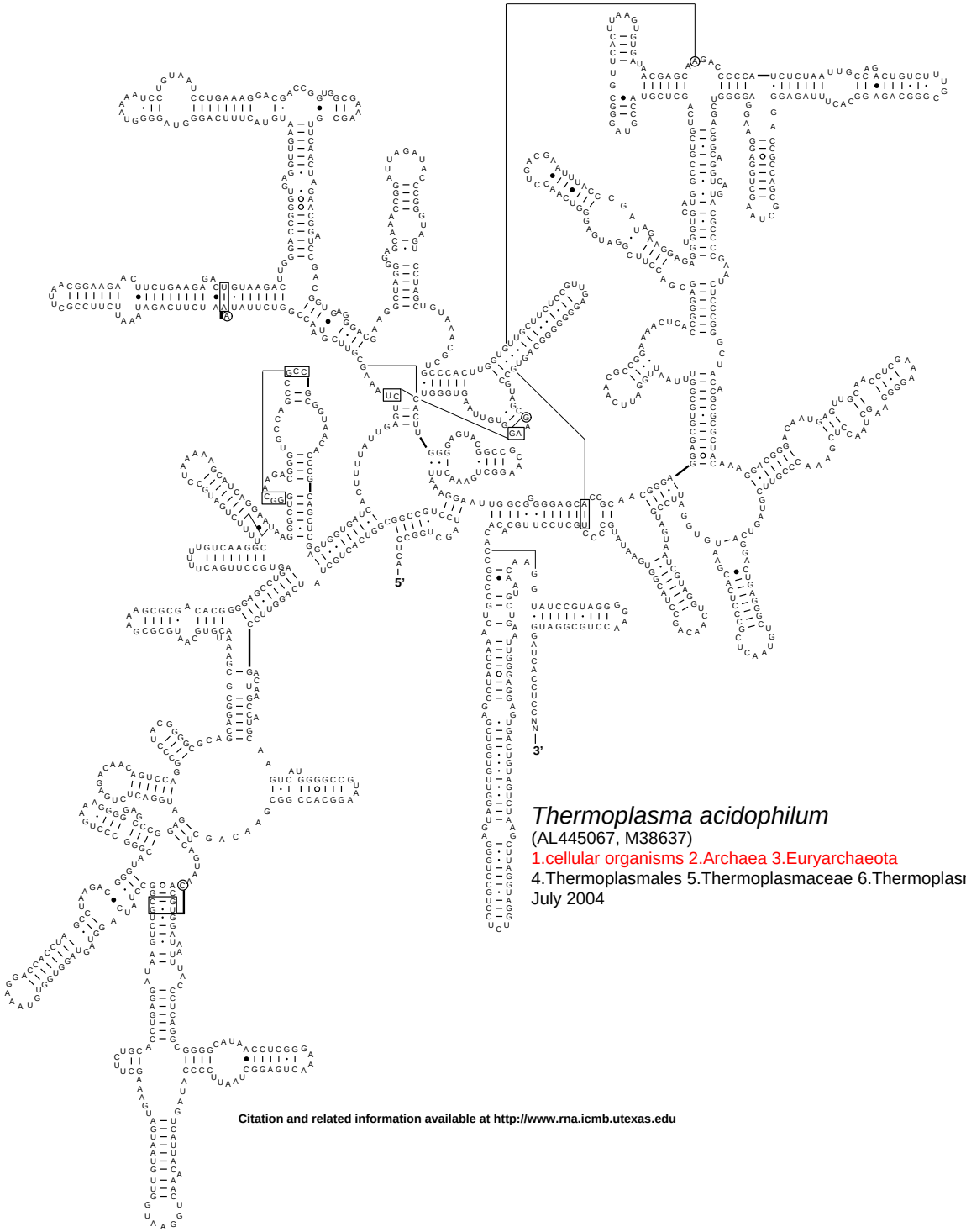
Secondary Structure: small subunit ribosomal RNA



Sulfolobus solfataricus
(X03235)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Sulfolobales 5.Sulfolobaceae 6.Sulfolobus
February 2001

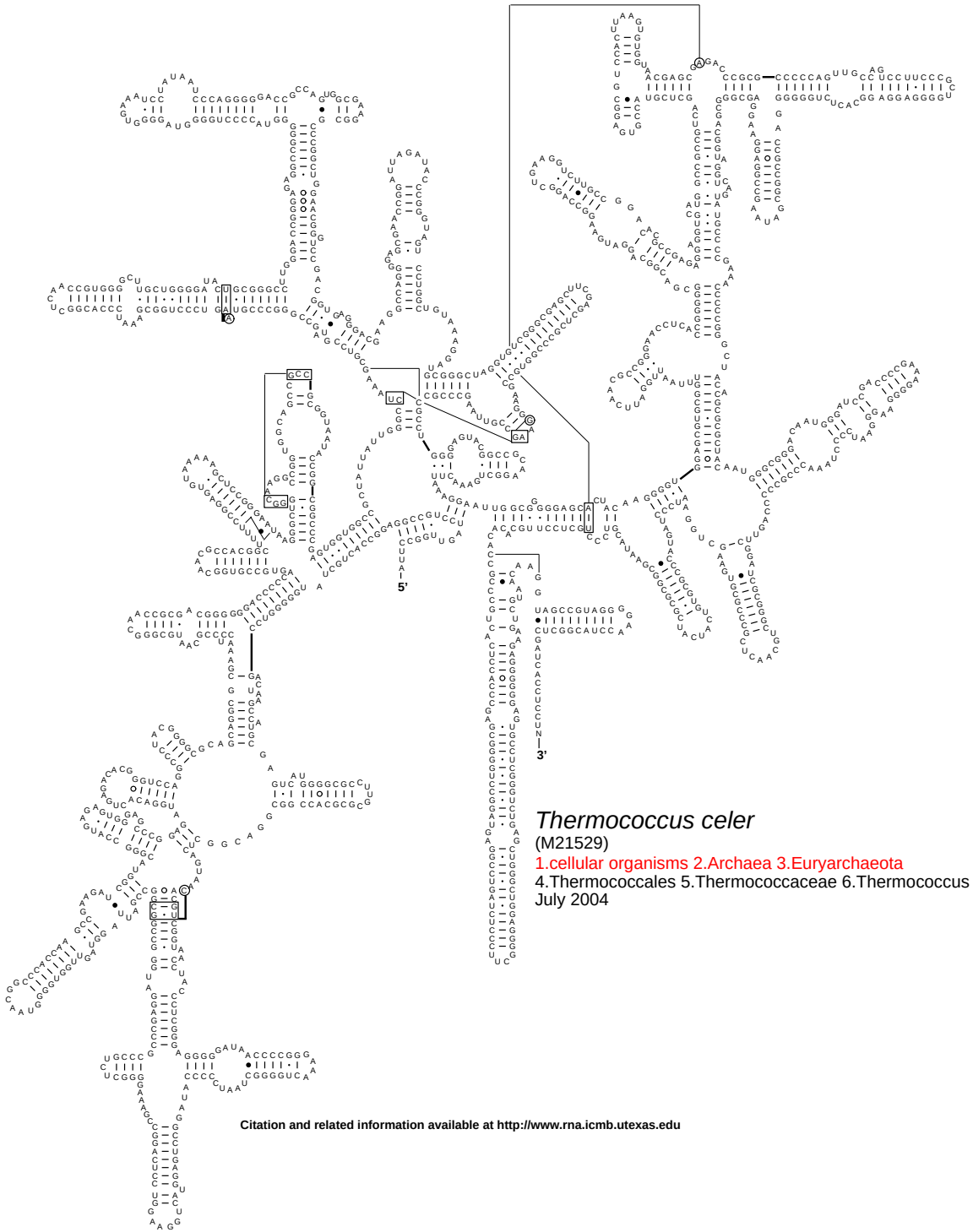
Citation and related information available at http://www.rna.ccbb.utexas.edu

Secondary Structure: small subunit ribosomal RNA



Thermoplasma acidophilum
(AL445067, M38637)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermoplasmales 5.Thermoplasmaceae 6.Thermoplasm
July 2004

Secondary Structure: small subunit ribosomal RNA



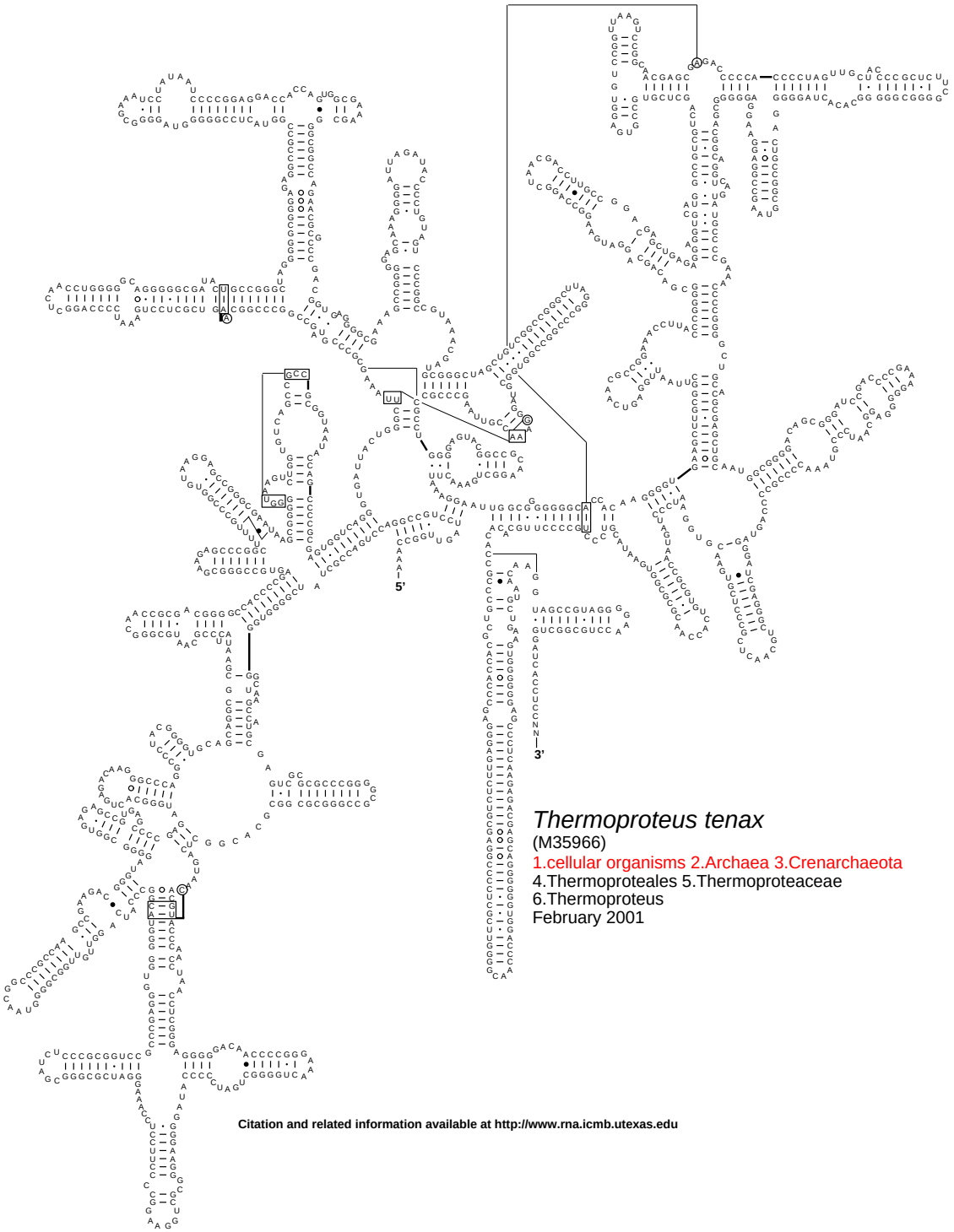
Thermococcus celer

(M21529)

1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Thermococcus

July 2004

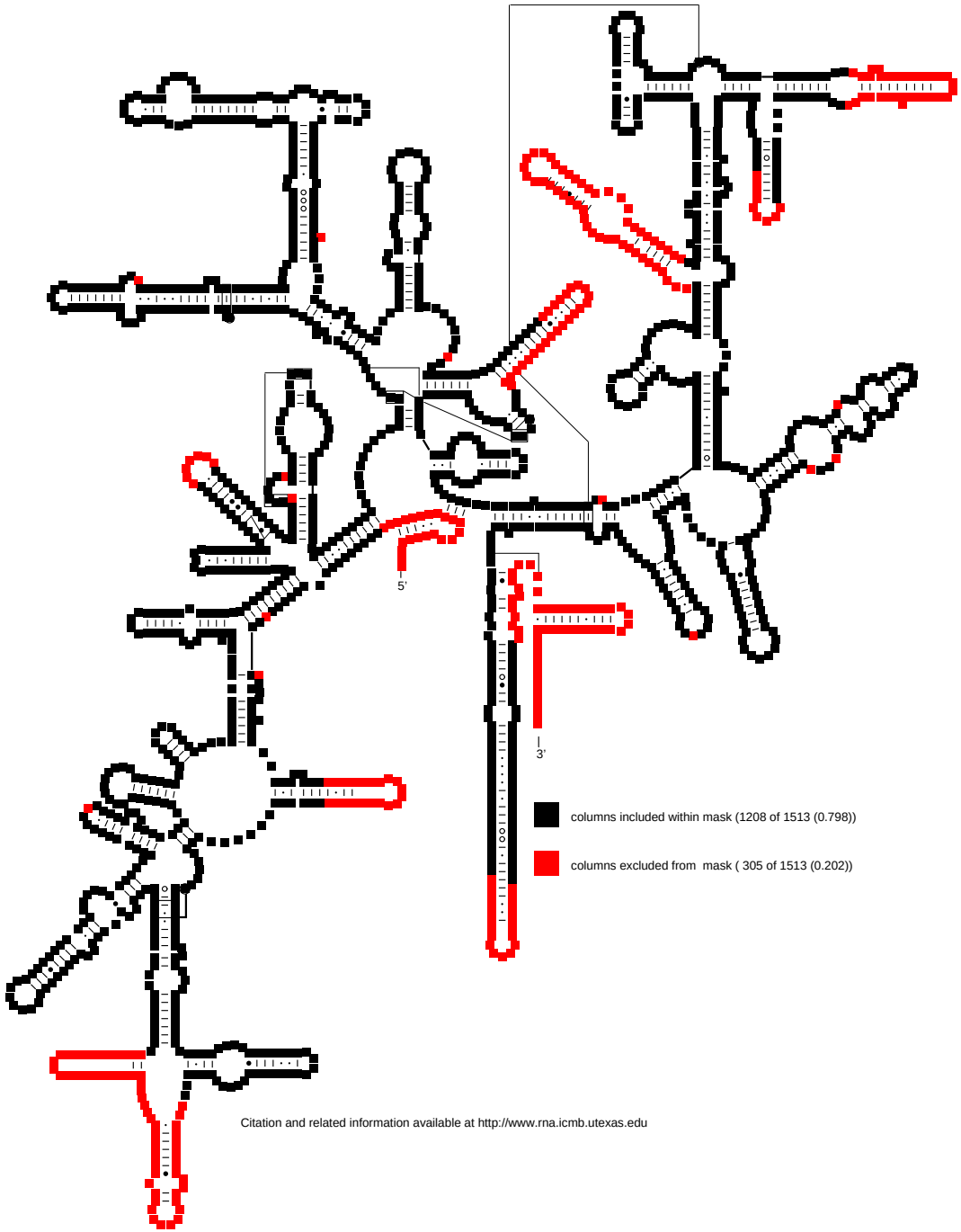
Secondary Structure: small subunit ribosomal RNA



Thermoproteus tenax
(M35966)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Thermoproteales 5.Thermoproteaceae
6.Thermoproteus
February 2001

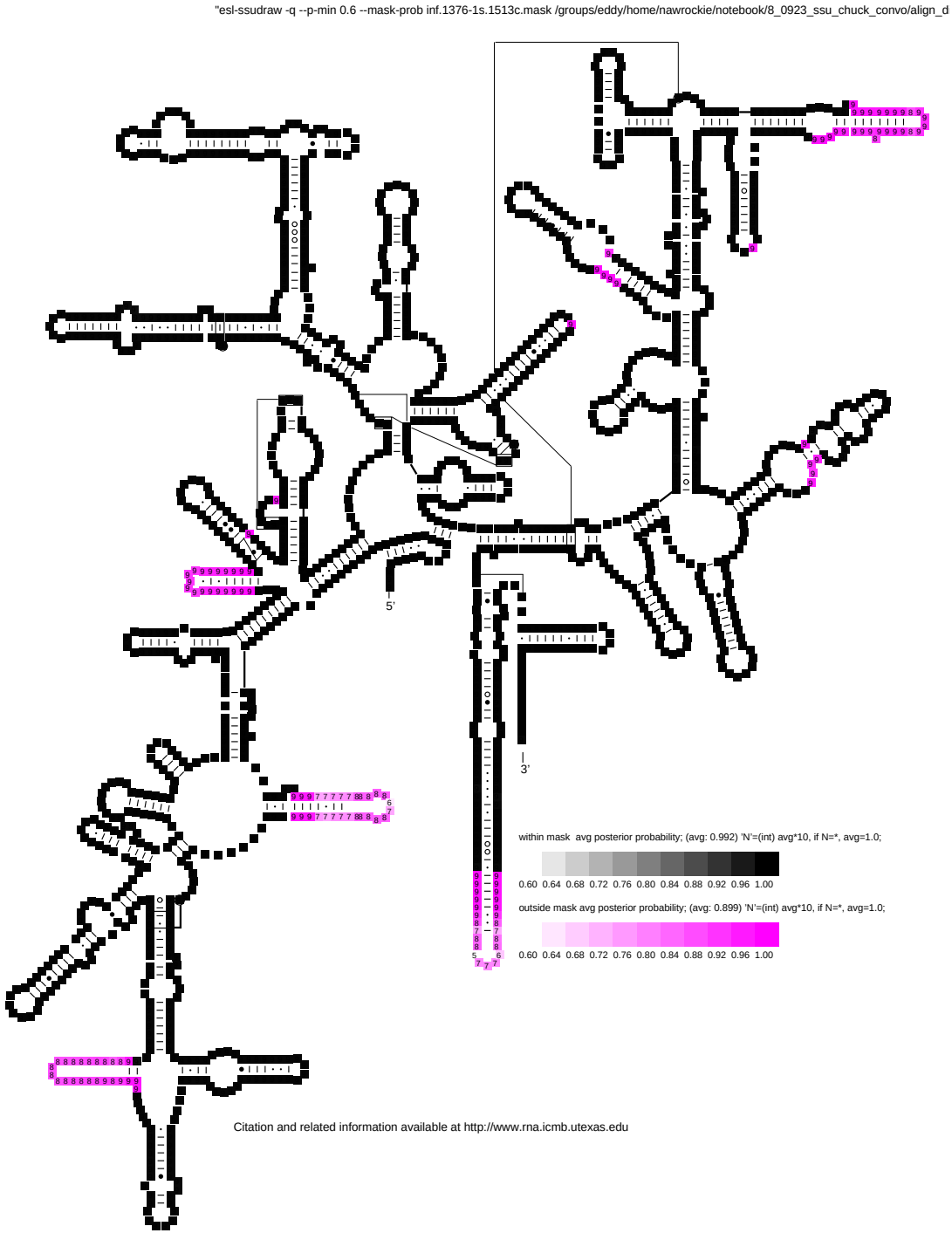
Phil Hugenholtz's manually created mask

"esl-ssudraw -q --mask-col inf.mask.from.nst.1208-1s.1513c.mask 1513.c.stk 1513.ps lmph_on_1513.ps" page 1/1

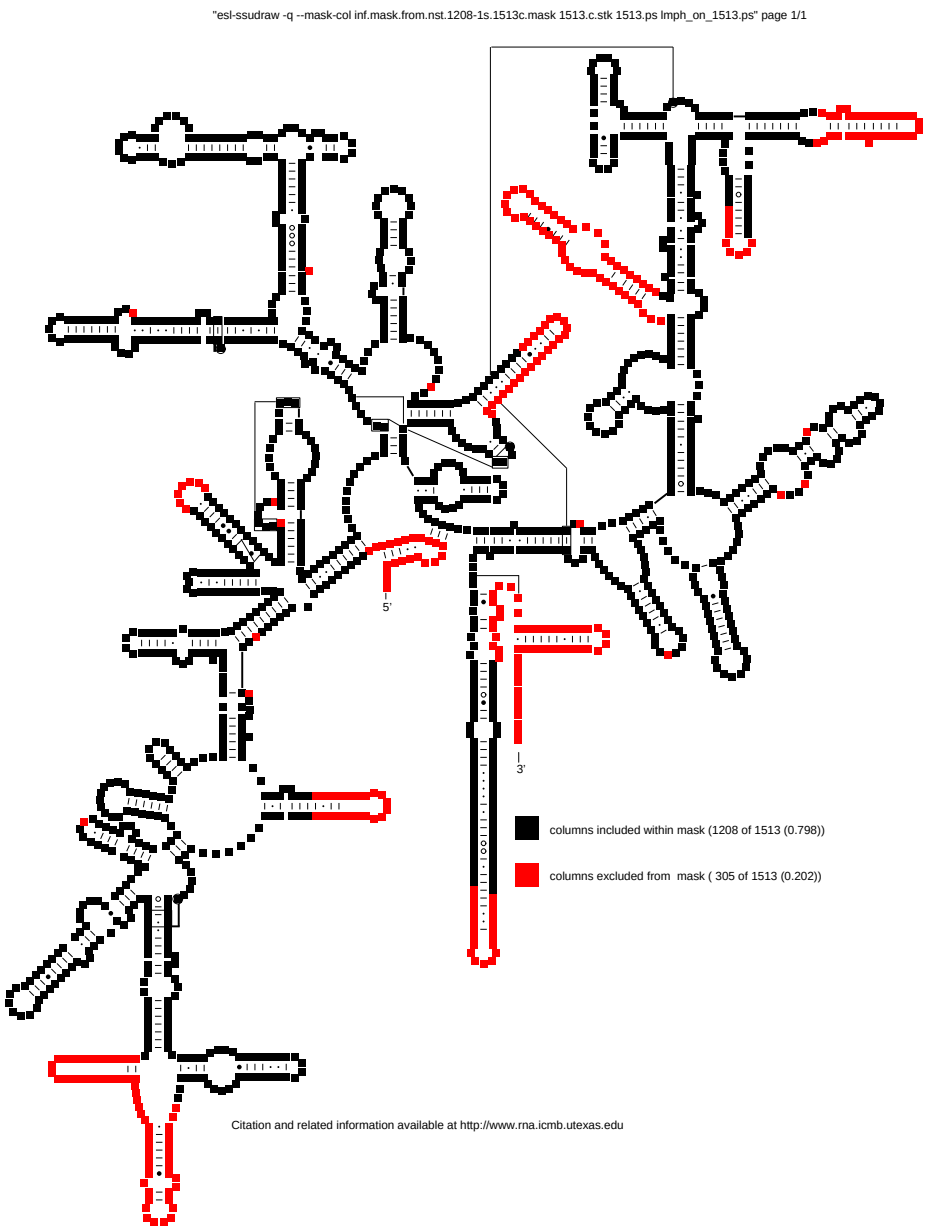


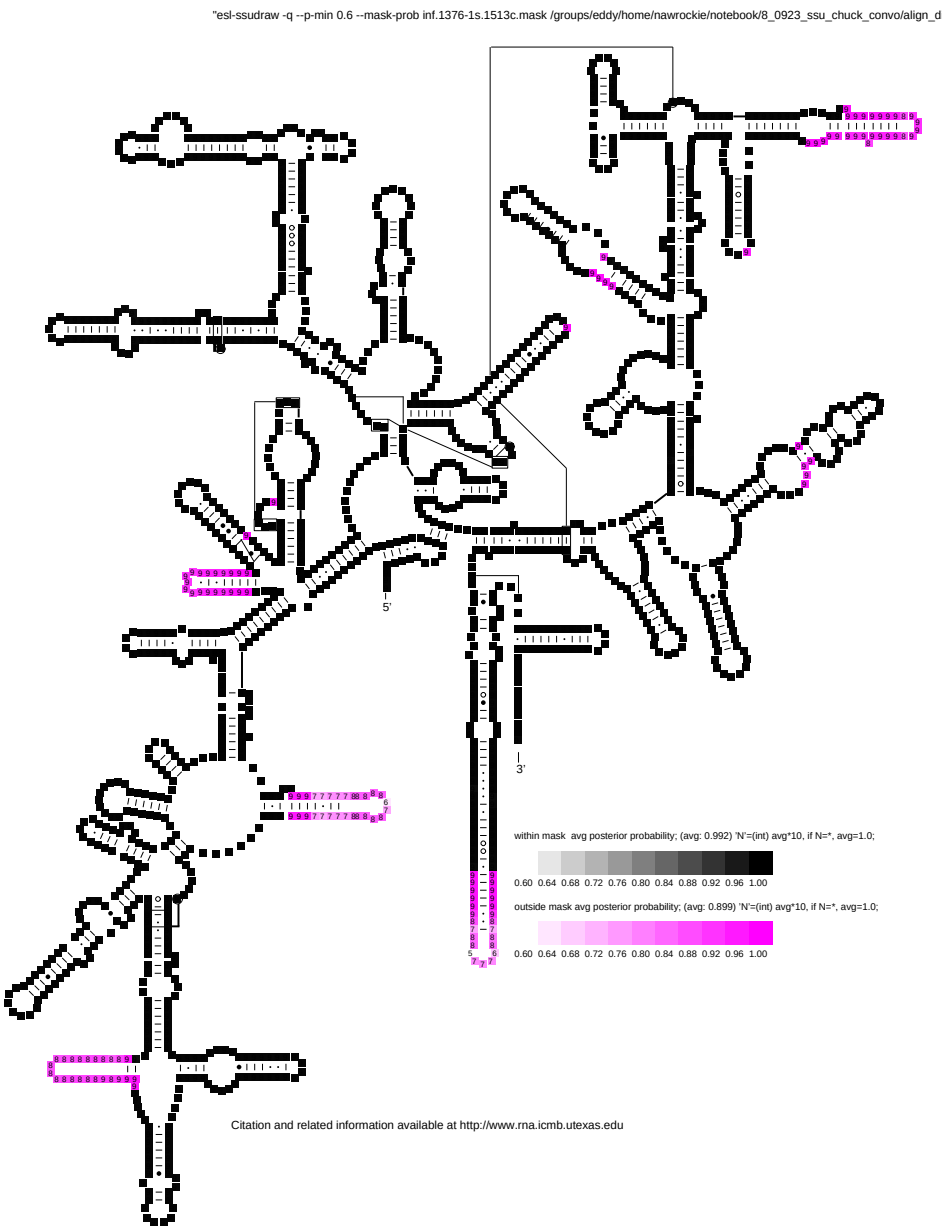
Citation and related information available at http://www.rna.icmb.utexas.edu

Infernal's automatically generated Archaeal mask



The manually created mask and Infernal's mask are similar





Automated masking removes the majority of alignment errors

	alignment accuracy	time (sec/seq)
clustalw	92.2%	30.0
HMMs	96.6%	0.08
non-banded CMs	98.1%	1321.5
HMM banded CMs	98.1%	0.7
probabilistically masked HMM banded CMs	99.7%	1.3

Infernal creates alignments that are very similar to manually refined alignments.

Large-scale SSU alignment with Infernal is now possible

- Infernal has been adopted as the alignment engine within the Ribosomal Database Project (RDP) which has more than 800,000 aligned SSU sequences.

- We are about to release **SSU-align**, an Infernal based SSU alignment program:
  - SSU models of archaea, bacteria, eukarya derived from Comparative RNA Website*
  - Automated probabilistic masking
  - User's guide with tutorial

*Cannone et.al., BMC Bioinformatics, 3:2, 2002.

Large-scale SSU alignment with Infernal is now possible

- Infernal has been adopted as the alignment engine within the Ribosomal Database Project (RDP) which has more than 800,000 aligned SSU sequences.
- We are about to release **SSU-align**, an Infernal based SSU alignment program:
  - SSU models of archaea, bacteria, eukarya derived from Comparative RNA Website*
  - Automated probabilistic masking
  - User's guide with tutorial

Genome-wide homology searches are faster and more powerful

- We (myself, Diana Kolbe, and Sean Eddy) have recently released version 1.0, the first "production" release of Infernal.†
- Infernal 1.0 was used by Shi et. al, 2009 (Nature) to identify known RNAs in a metagenomics transcriptome project

*Cannone et.al., BMC Bioinformatics, 3:2, 2002.

†Nawrocki EP. Kolbe DL. Eddy SR. Bioinformatics, 25:1335-1337, 2009.

Acknowledgements

I am incredibly lucky

NIH/NHGRI Inst. Training Grant in Genomic Science (T32-HG000045)
Howard Hughes Medical Institute
Washington University in St. Louis

Sean Eddy
Elena Rivas
Tom Jones

Sergi Castellano
Fred Davis
Michael Farrar
Lee Henry
Seolkyoung Jung
Diana Kolbe

Robin Dowell
Steve Johnson
John McCutcheon
Shawn Stricklin

Michael Brent
Jeremy Buhler
Justin Fay
Jeff Gordon
Rob Mitra
Gary Stormo

Goran Ceric

Mary Pichler
Margaret Jefferies

Jay Gertz
KT Varley
Scott Doniger
Justin Gerke

the WU-Blasters

Kathy Nawrocki
Ed Nawrocki
Kevin Nawrocki
Brigid Nawrocki

Don Hopkins
Linda Hopkins

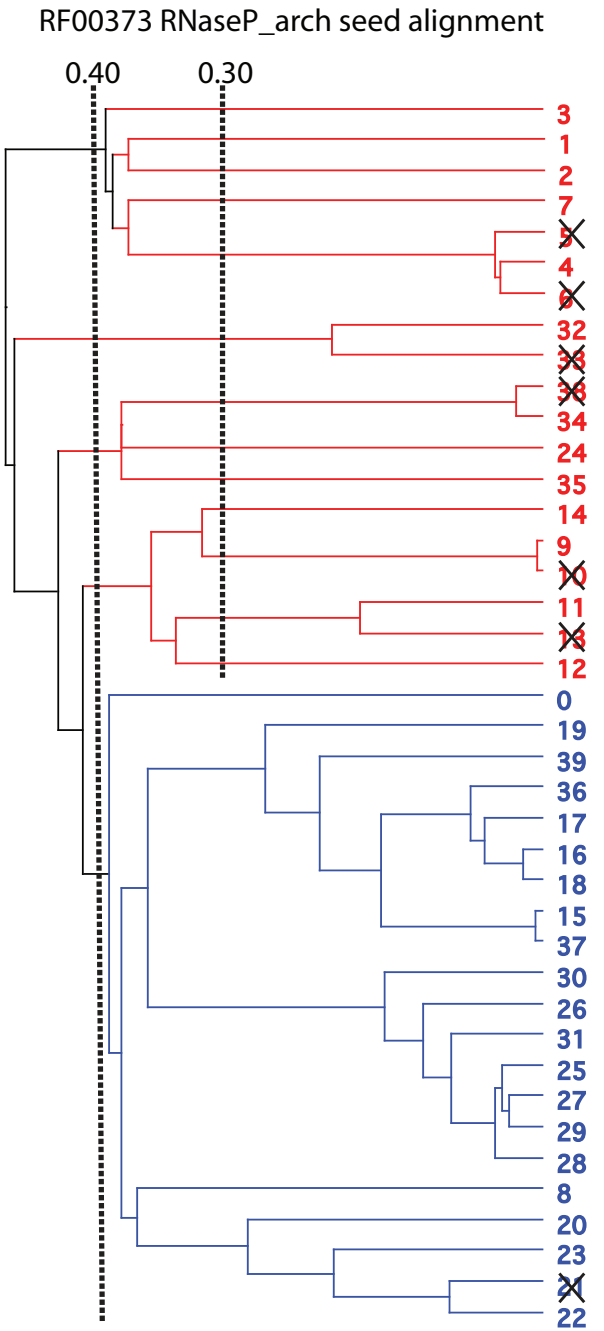
Michael Hopkins

John
Paul
George
Ringo

RMARK: an internal RNA homology search benchmark

- RMARK construction - for each of the 503 Rfam 7 seed alignments:
  - cluster sequences by sequence identity given the alignment
  - look for a training cluster and testing cluster such that:
    - no training/test sequence pair is $> 60\%$ identical
    - at least five sequences are in the training set
  - filter test set so no two test seqs $> 70\%$ identical
  - 51 families qualify, with 450 test sequences
  - test seqs are embedded in a 10 Mb pseudo-genome of "realistic" base composition

Example:



Empirical time complexity of CM alignment