# Modeling structural RNA families with Infernal

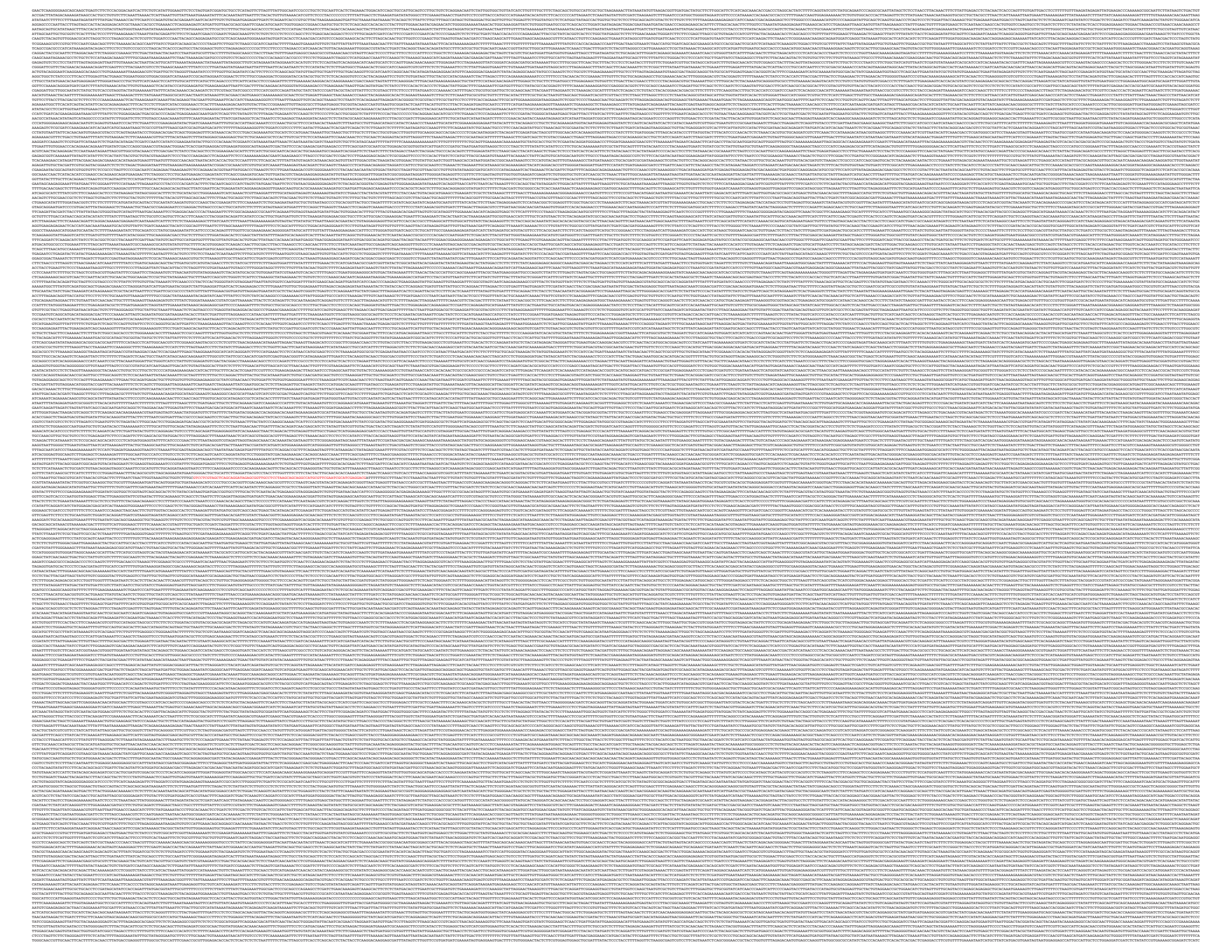
Eric Nawrocki

Sean Eddy's Lab

Howard Hughes Medical Institute
Janelia Farm Research Campus

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 → mRNA → protein

Functional RNA genes: DNA → RNA

## How can we find genes?

**By searching for similarity with known genes indicative of shared evolutionary history**

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**

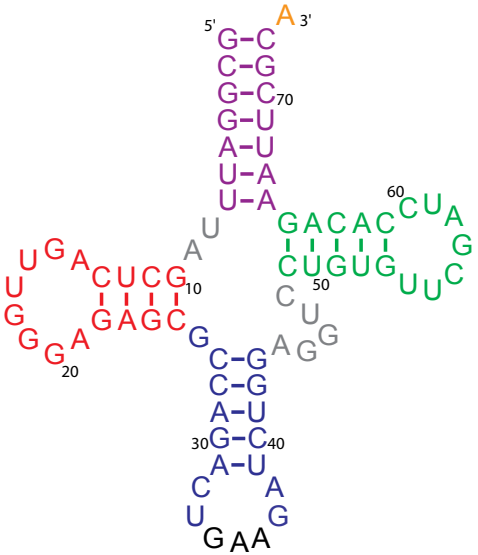
# Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell

Three representations of a transfer RNA:

## Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

## Secondary structure




## 3-dimensional structure

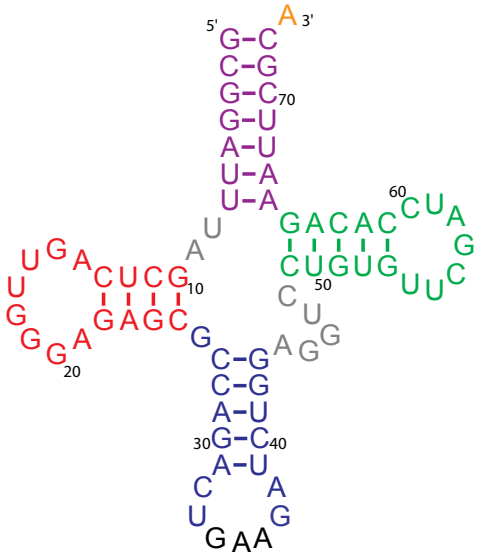
**Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell**

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



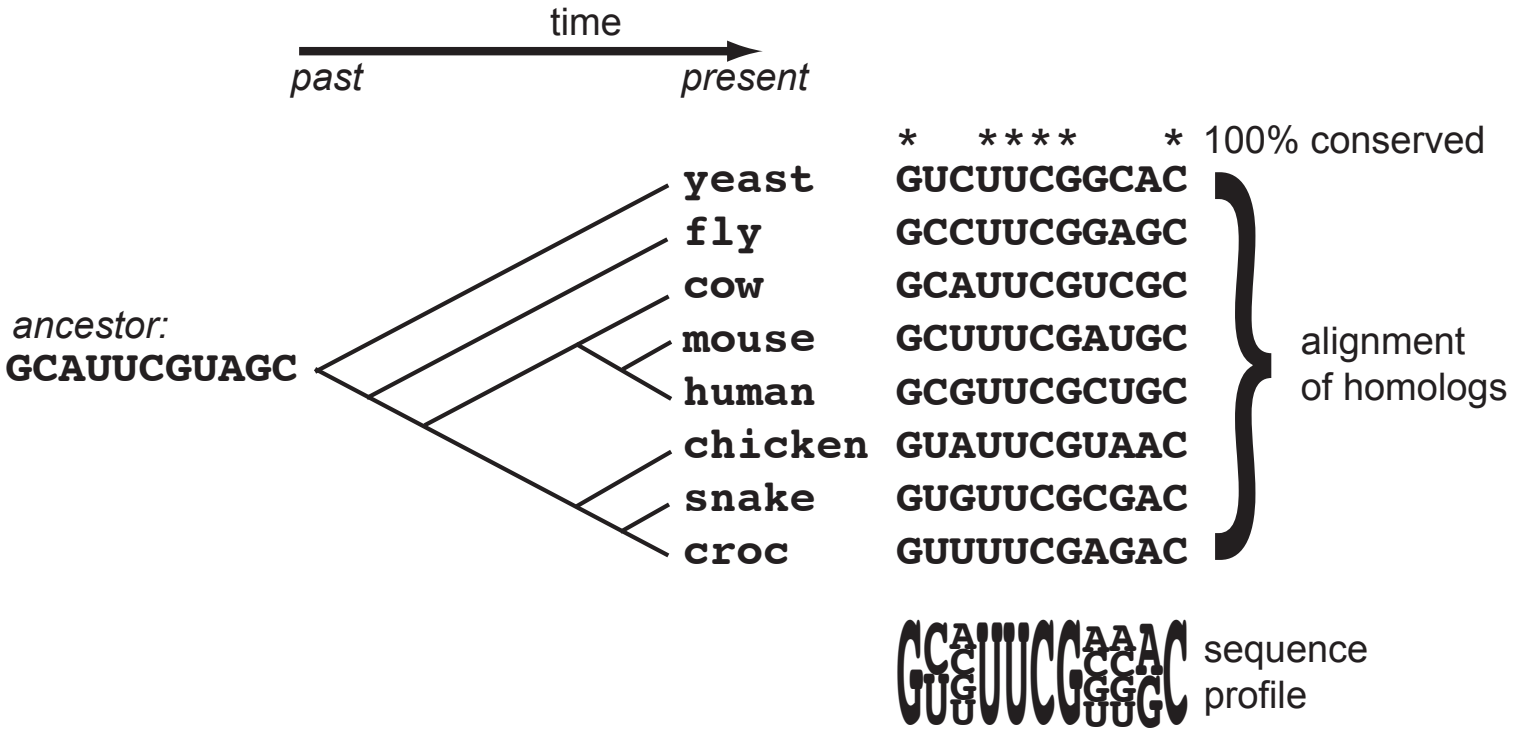
3-dimensional structure

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

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

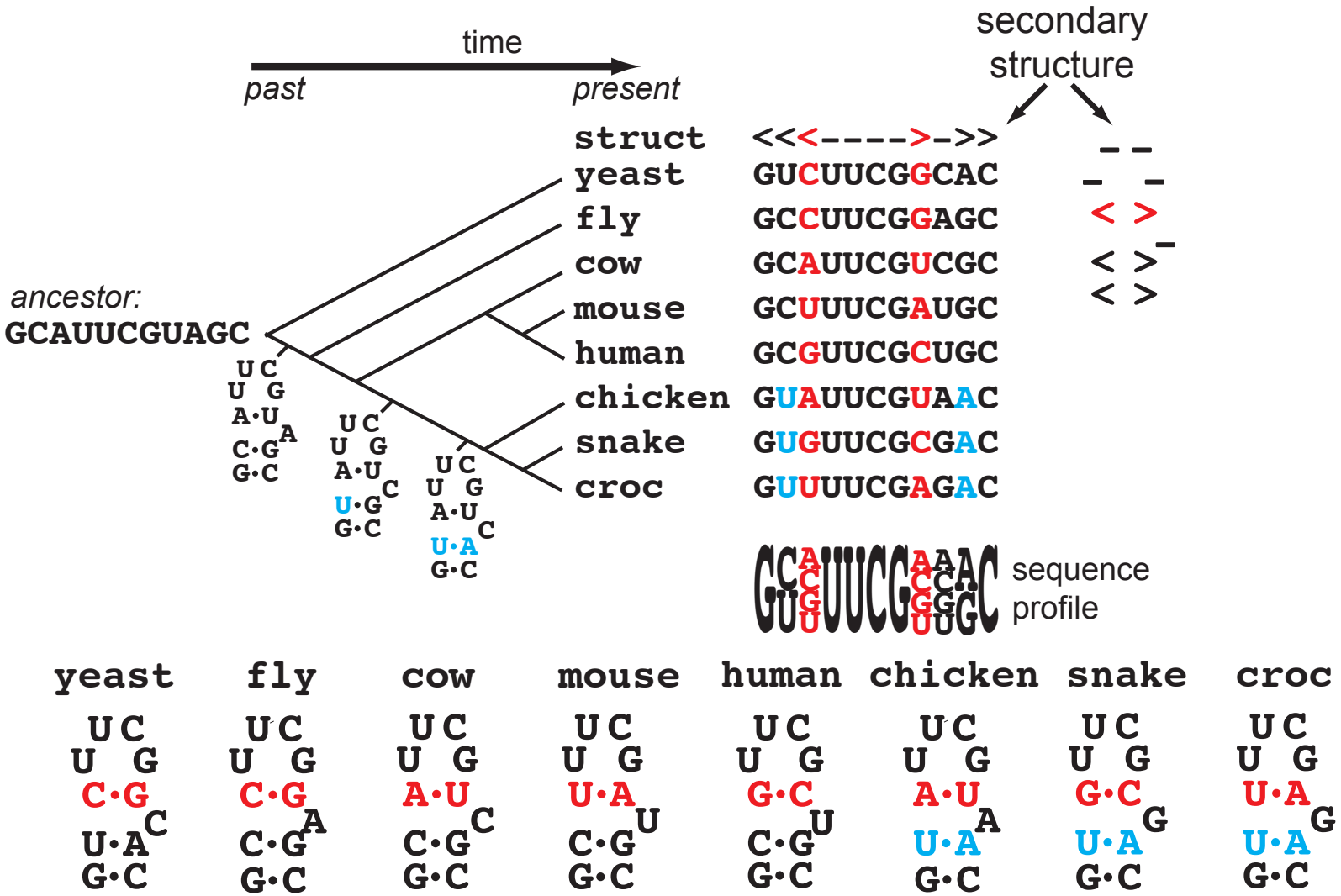
# Sequence conservation provides information for homology searches

Conservation levels vary across alignment columns.

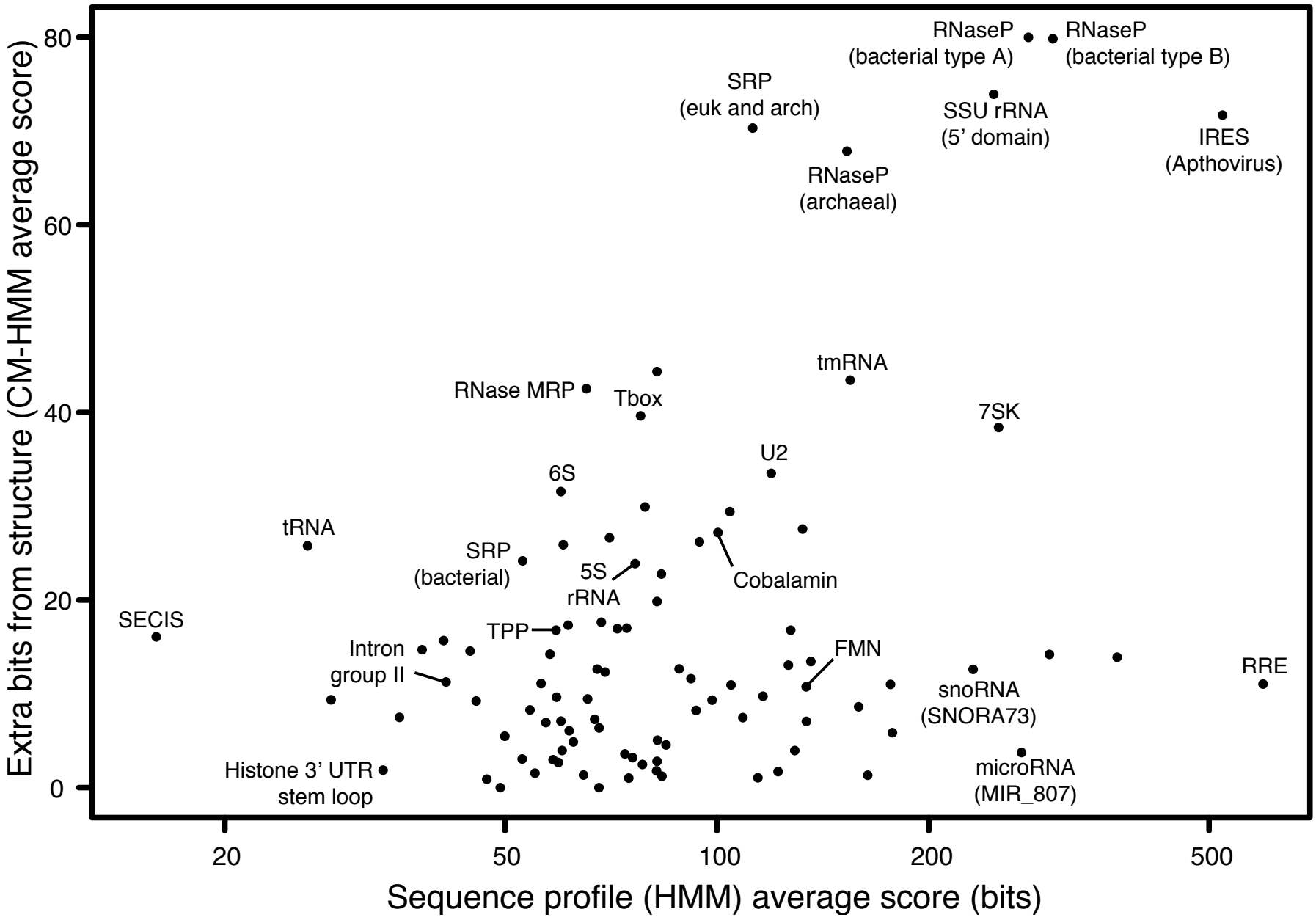


# Structure conservation provides additional information

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 |
| main use | proteins | RNAs |
| database | Pfam (14831 families) | Rfam (2208 families) |
| performance for RNAs | faster but less accurate | slower but more accurate |


http://hmmer.janelia.org

Eddy, SR. PLoS Comp. Biol., 7:e1002195, 2011.

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

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


http://infernal.janelia.org

Nawrocki EP, Kolbe DL, Eddy SR Bioinformatics, 25 (10):1335-1337, 2009.

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

# Profile HMMs: sequence family models built from alignments

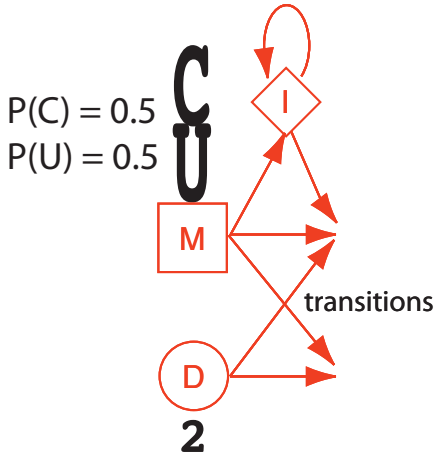

| | |
|---|---|
| yeast | GUCUUCGGCAC |
| fly | GCCUUCGGAGC |
| cow | GCAUUCGUCGC |
| mouse | GCUUUCGAUGC |
| human | GCGUUCGCUGC |
| chicken | GUAUUCGUAAC |
| snake | GUGUUCGCGAC |
| croc | GUUUUCGAGAC |

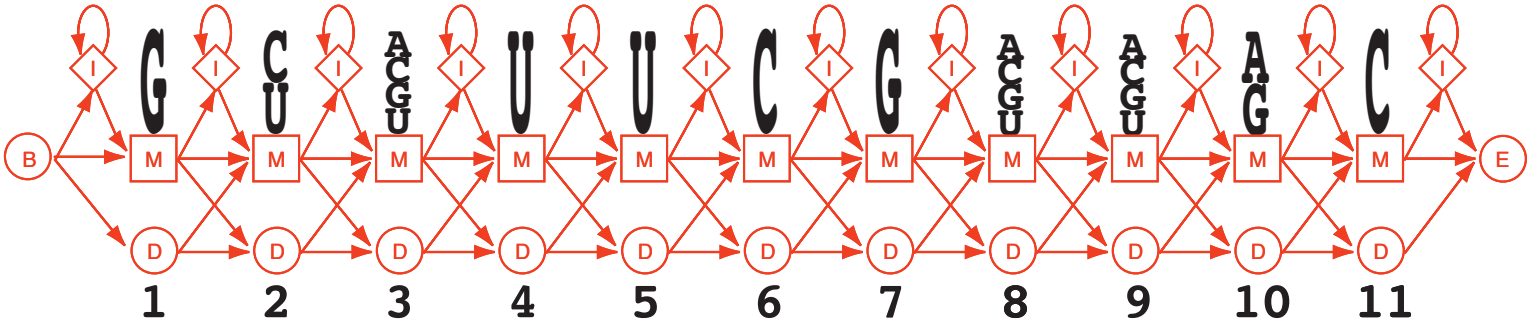
One HMM node per alignment column

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 |

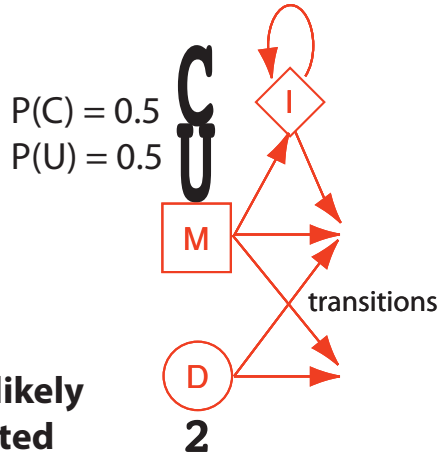
One HMM node per alignment column

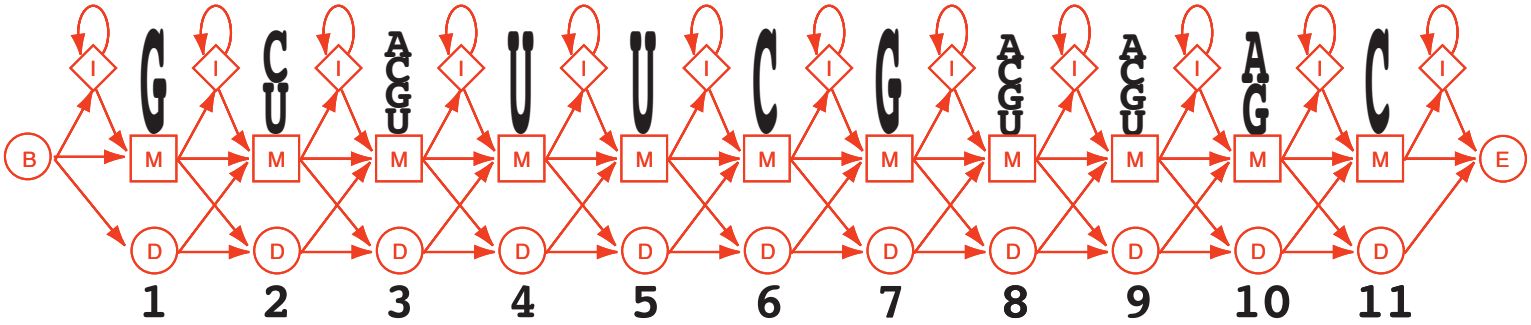
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

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 |

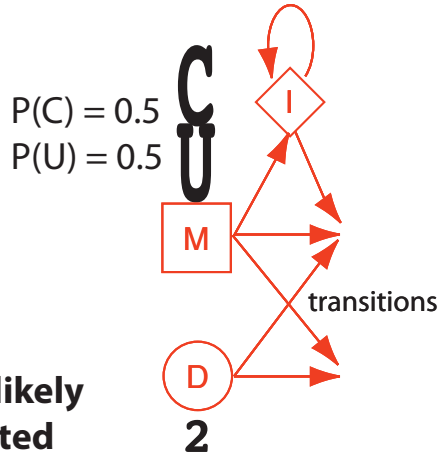
One HMM node per alignment column

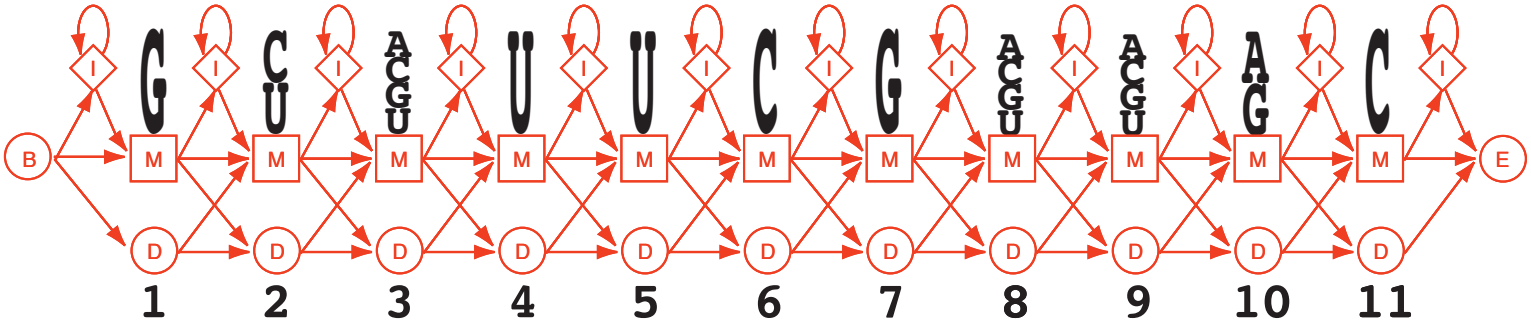
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

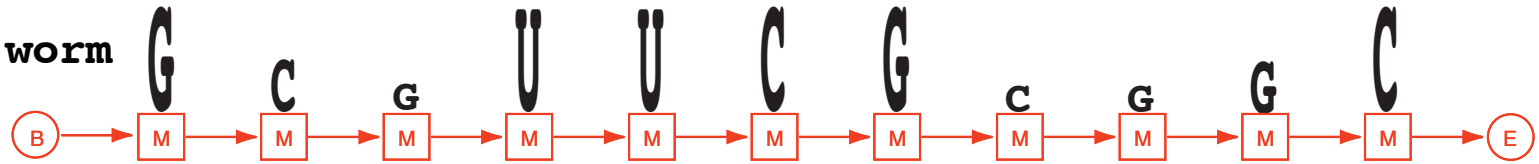
HMMs generate homologous sequences.

**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
```

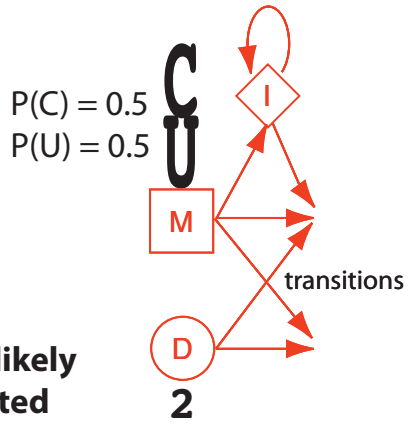
One HMM node per alignment column

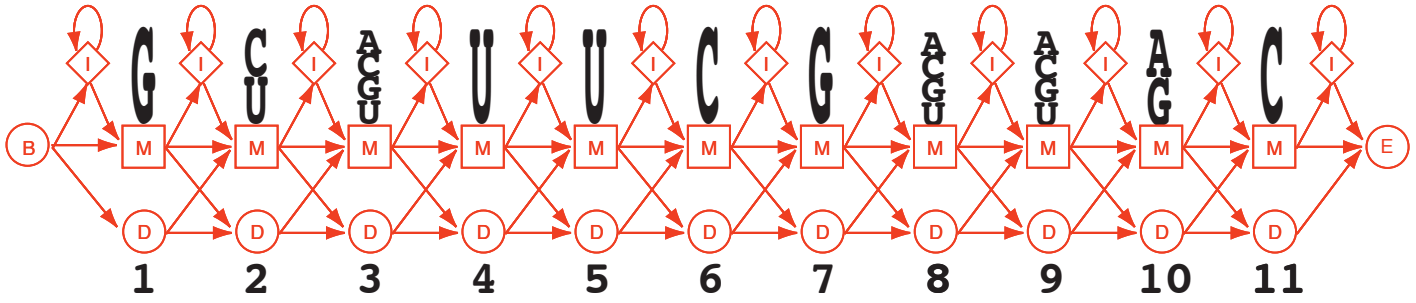
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

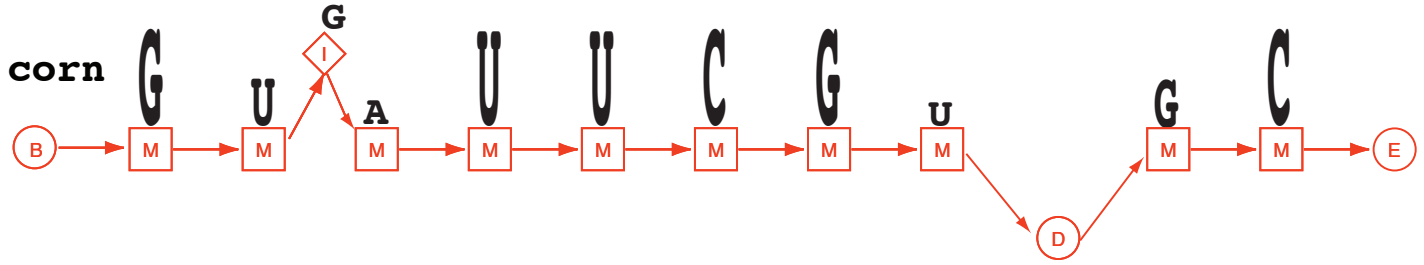
HMMs generate homologous sequences.

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

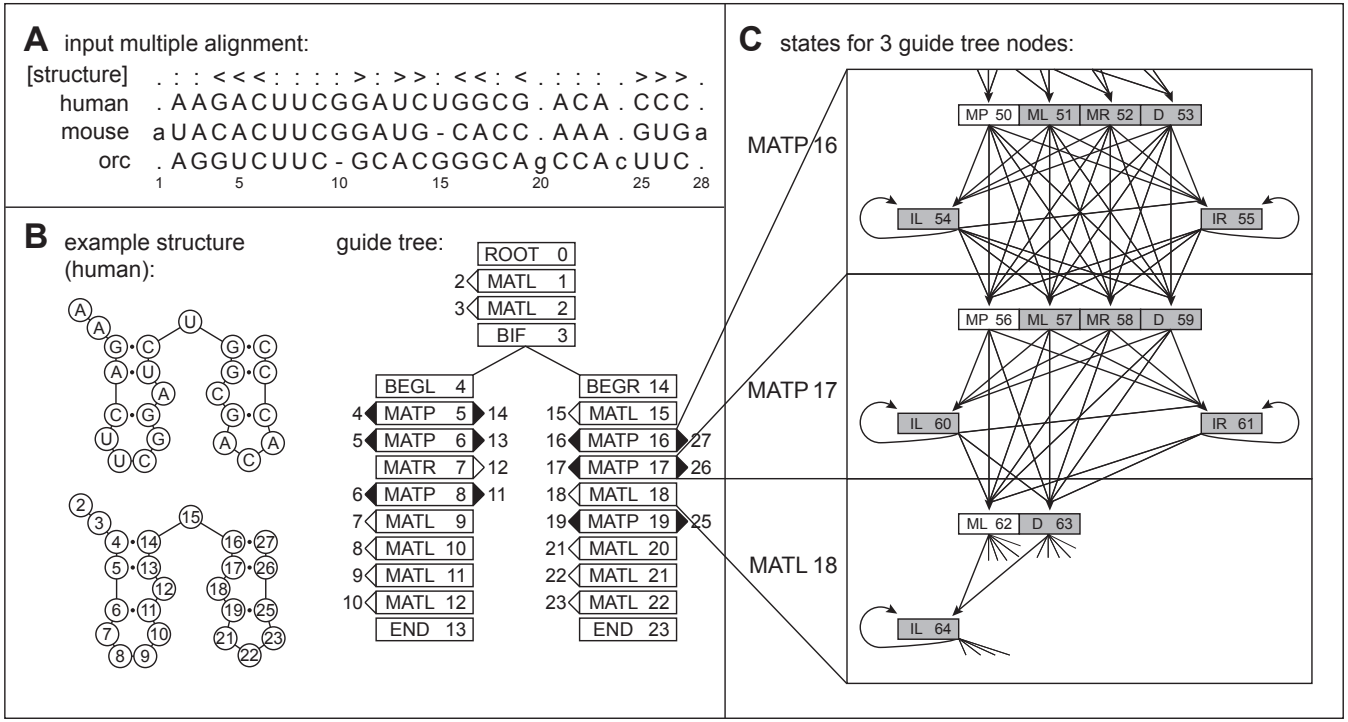
Node for column 2:

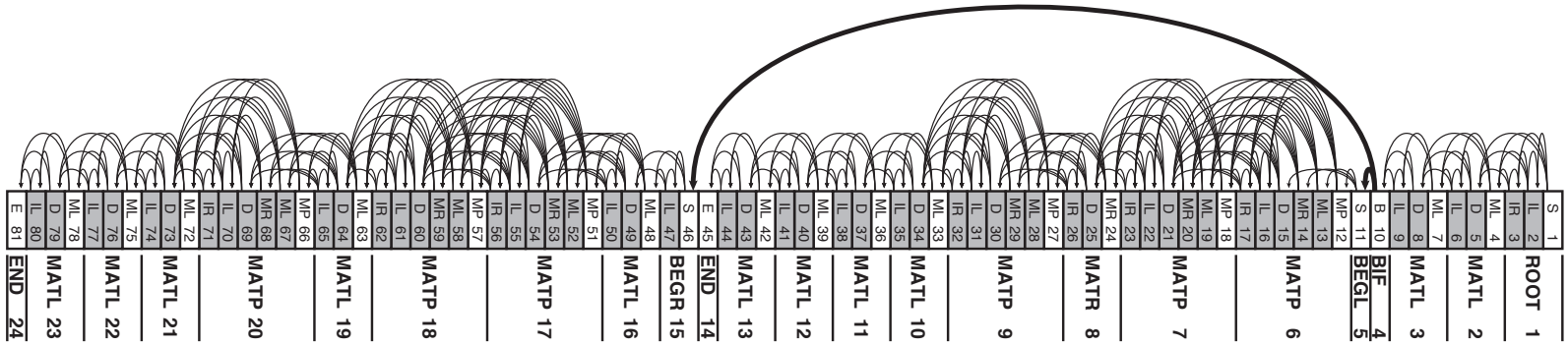


# Covariance models (CMs) are built from structure-annotated alignments

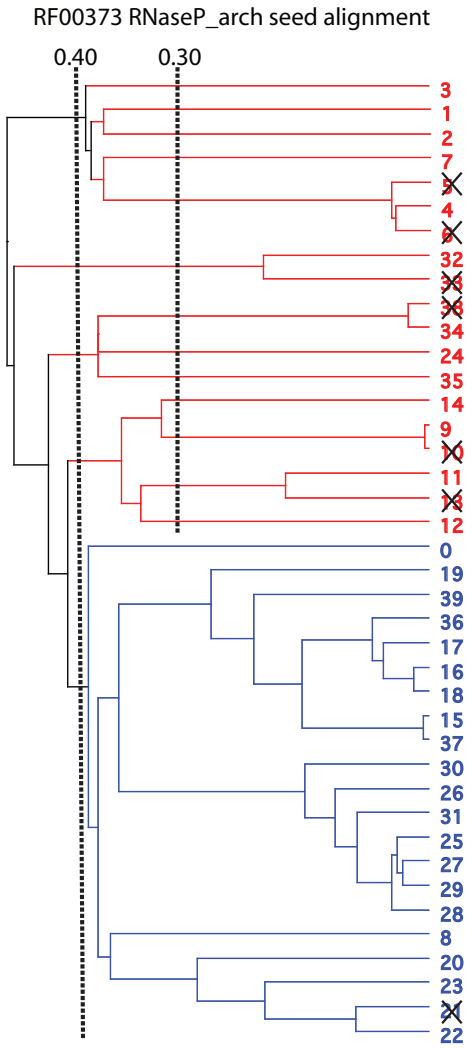


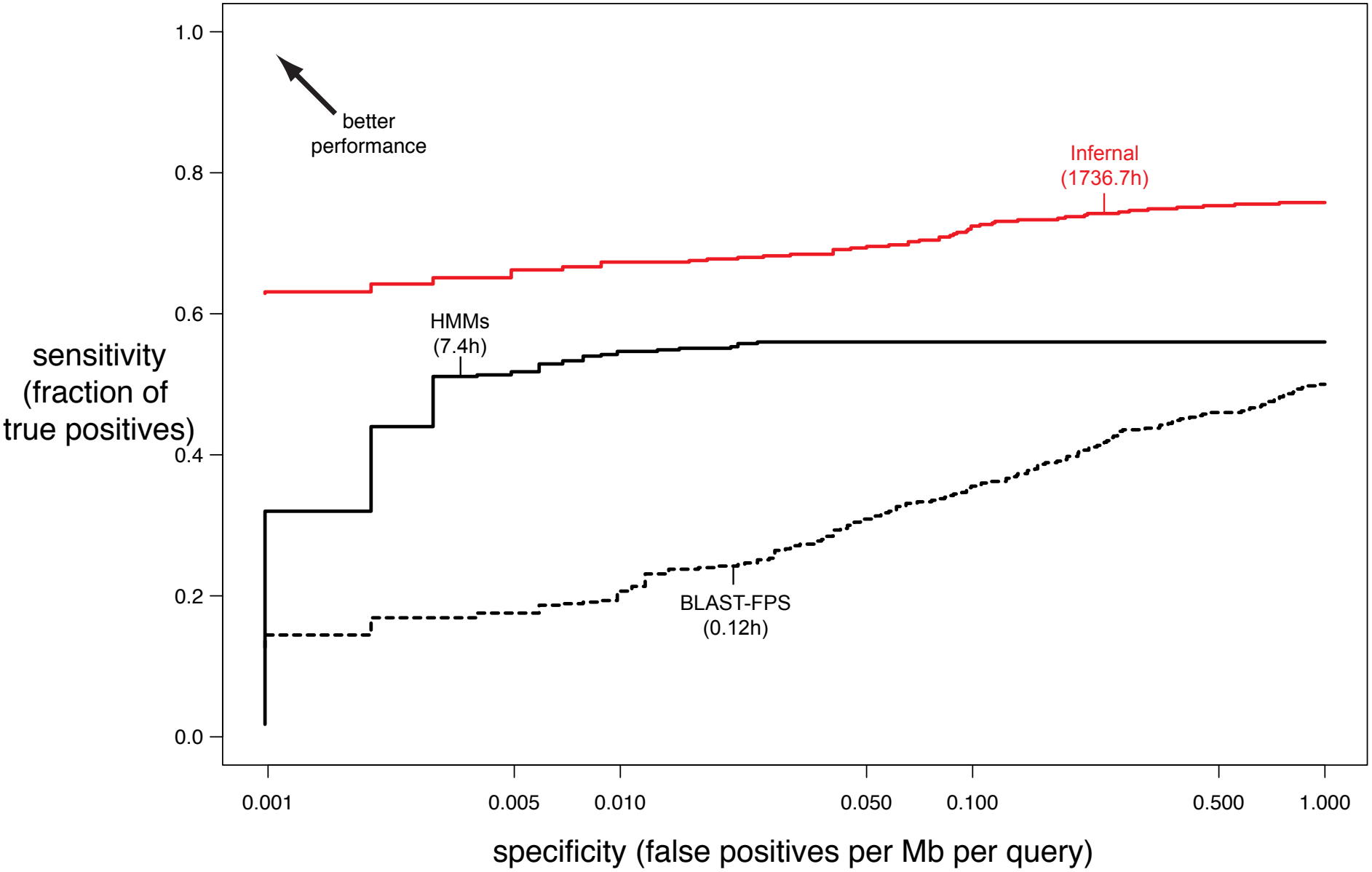
# Is the added complexity worth it?
# RMARK: a challenging internal RNA homology search benchmark for use during Infernal development

- RMARK construction - for each of the 1446 Rfam 10 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
  - 106 families qualify, with 780 test sequences
  - test seqs are embedded in a 10 Mb pseudo-genome of "realistic" base composition

Example:



# Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)




*Freyhult EK, Bollback JP, Gardner PP. Genome Res. 2007 17: 117-125.

# CM searches are especially slow for large RNAs

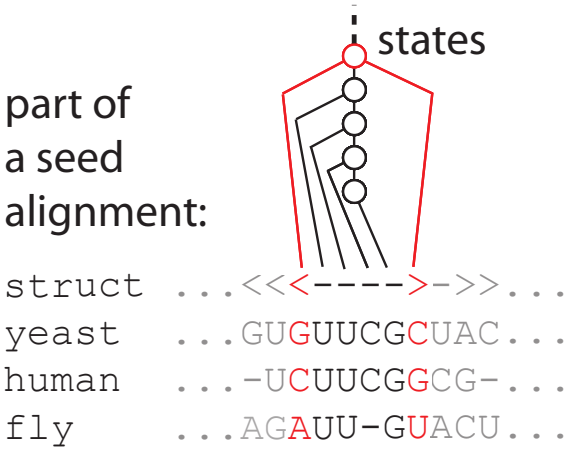
| | | search (min/Mb) | | |
|---|---|---|---|---|
| family | length | HMM | CM | CM/HMM |
| tRNA | 71 | 0.34 | 27.0 | 79.4 |
| Lysine riboswitch | 183 | 0.80 | 133.2 | 166.7 |
| SRP RNA | 304 | 1.32 | 276.4 | 214.4 |
| RNaseP RNA | 365 | 1.56 | 733.4 | 470.3 |

# 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

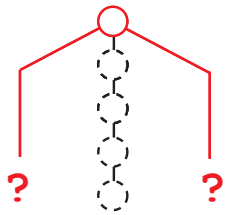
Example: Finding a hairpin loop

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.**

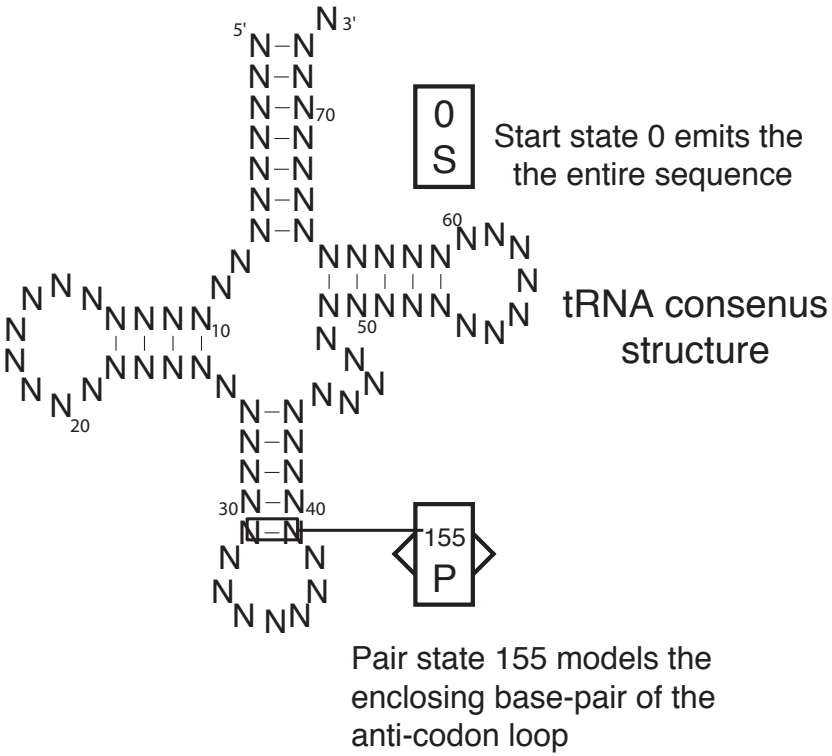
# Query-dependent banding (QDB) strategy

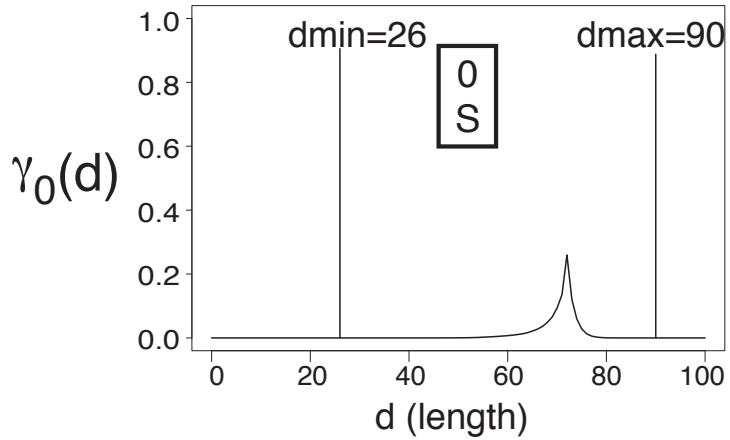
- Calculate $\gamma_v(d)$ probability each state $v$ will emit/align to subsequences of length $d$, for $d = 0..Z$

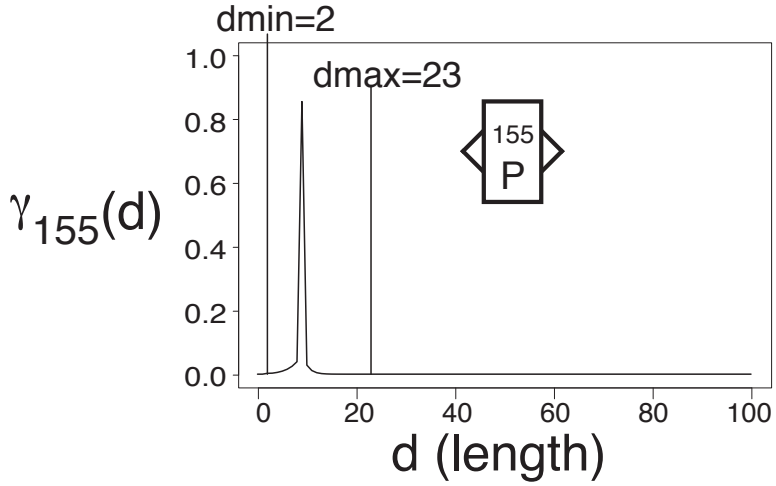
for states $v = M-1$ down to 0:

| | | |
|---|---|---|
| $v$ = end state ($E$): | $\gamma_v(0) = 1$<br>$\gamma_v(d) = 0$ | 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 ($v = S, P, L, R$): | $\gamma_v(d) = 0$<br>$\gamma_v(d) = \sum_{y \in C_v} \gamma_y(d - (\Delta_v^L + \Delta_v^R)) * t_v(y)$ | for $d = 0$ to $(\Delta_v^L + \Delta_v^R - 1)$<br>for $d = (\Delta_v^L + \Delta_v^R)$ to $Z$ |

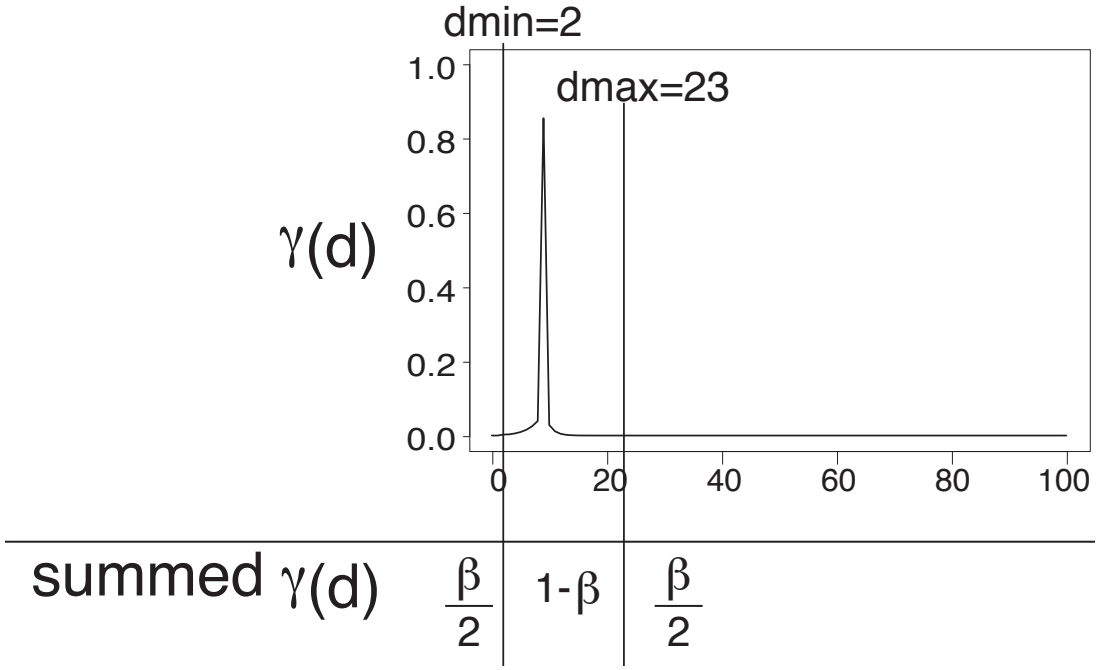
## QDBs for a tRNA CM:



# The $\beta$ parameter controls amount of probability loss



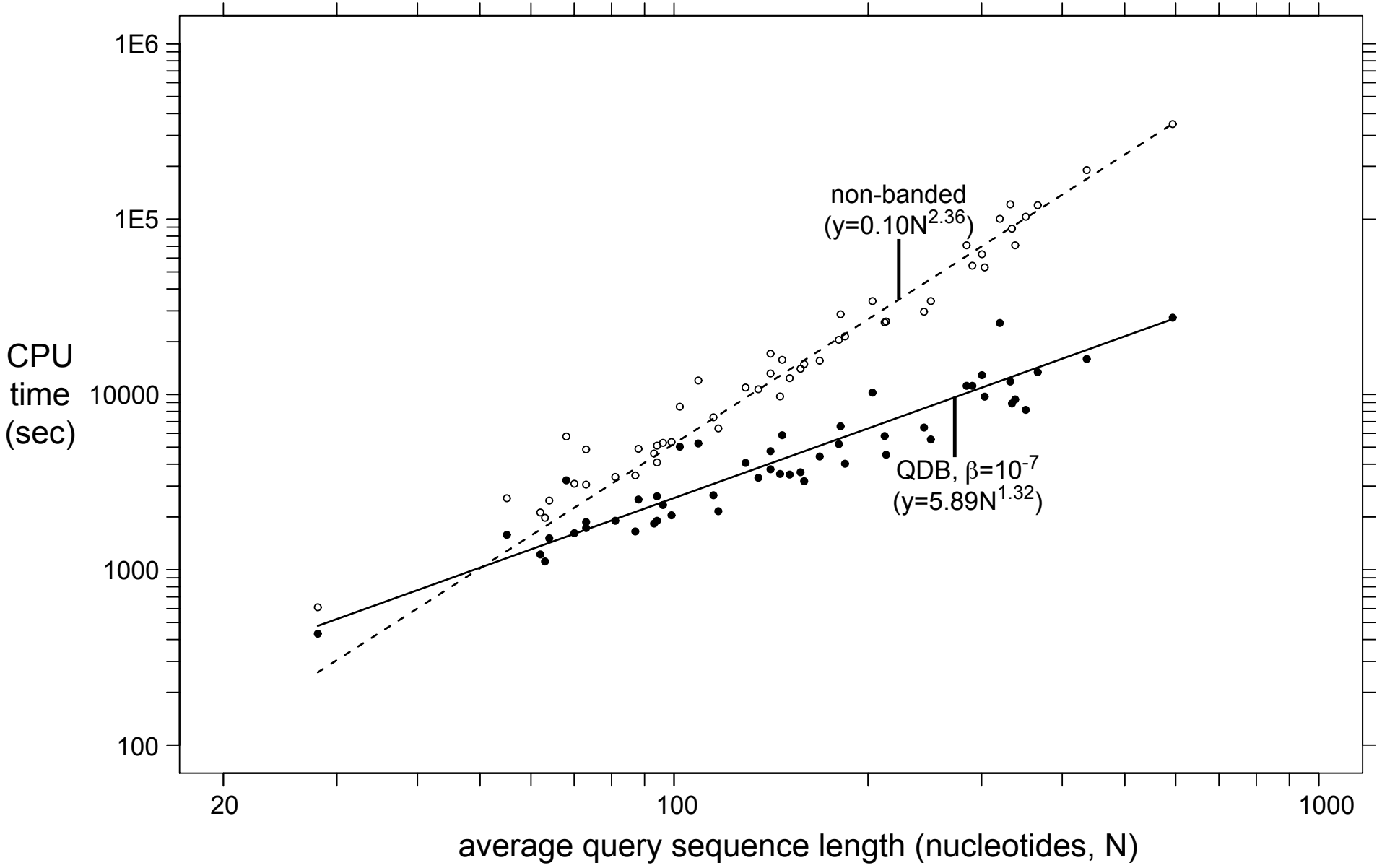

$$\sum_{d=0}^{\text{dmin}-1} \gamma(d) < \frac{\beta}{2}$$

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

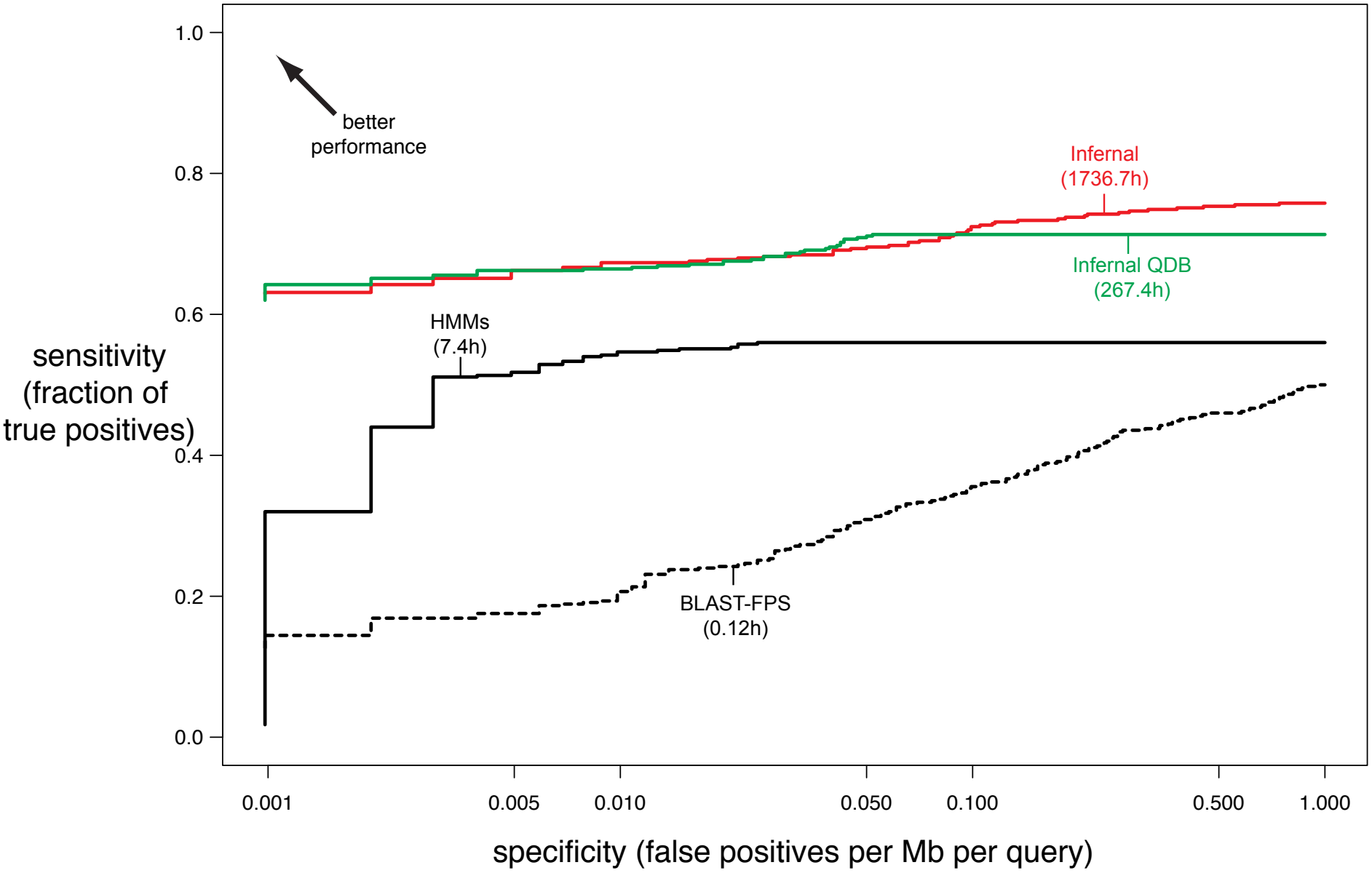
$$\sum_{d=\text{dmax}+1}^{Z} \gamma(d) < \frac{\beta}{2}$$

- $\beta$ is typically very small
  for example: $0.0000001(10^{-7})$
- Higher $\beta$ 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

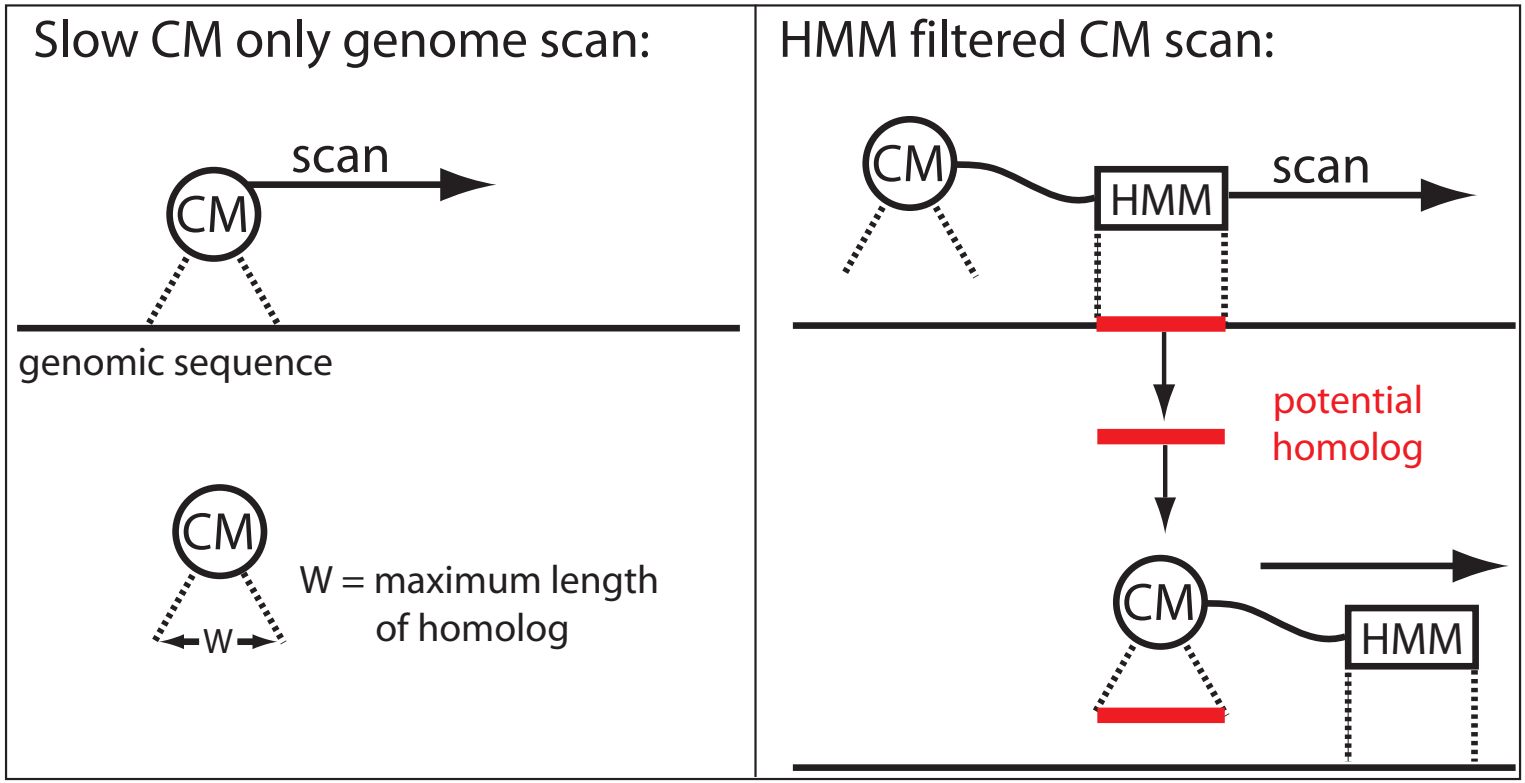


# CM homology searches are still slow

| | | 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 and Buhler †, Zhang and Bafna‡)



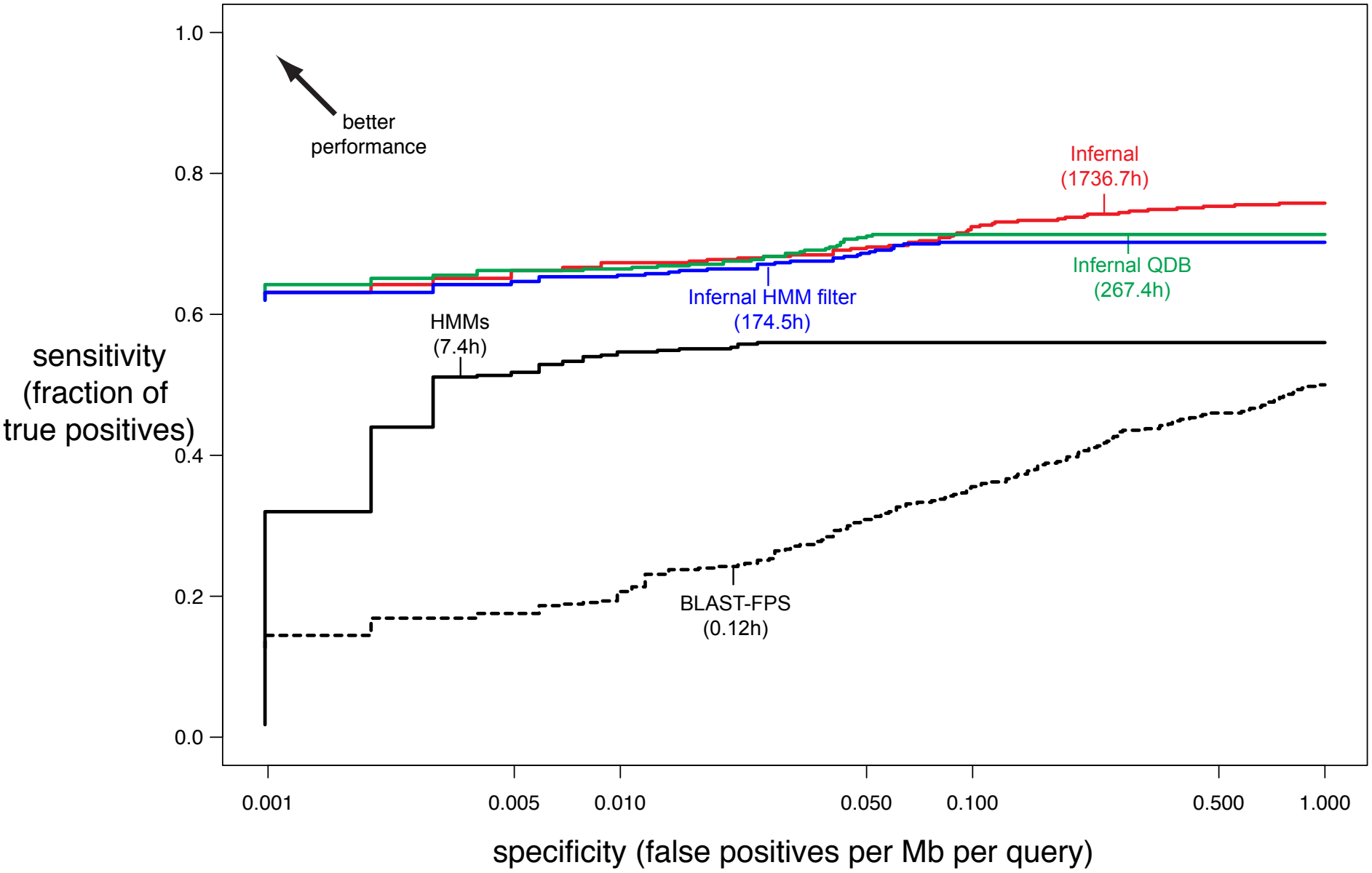


*Weinberg Z, Ruzzo WL. 22(1):3539, 2006.

†Sun Y, Buhler J, Comput. Systems Bioinf., p145-156, 2008.

‡Zhang S et al., Bioinformatics. 22(14):e557-e565, 2006.

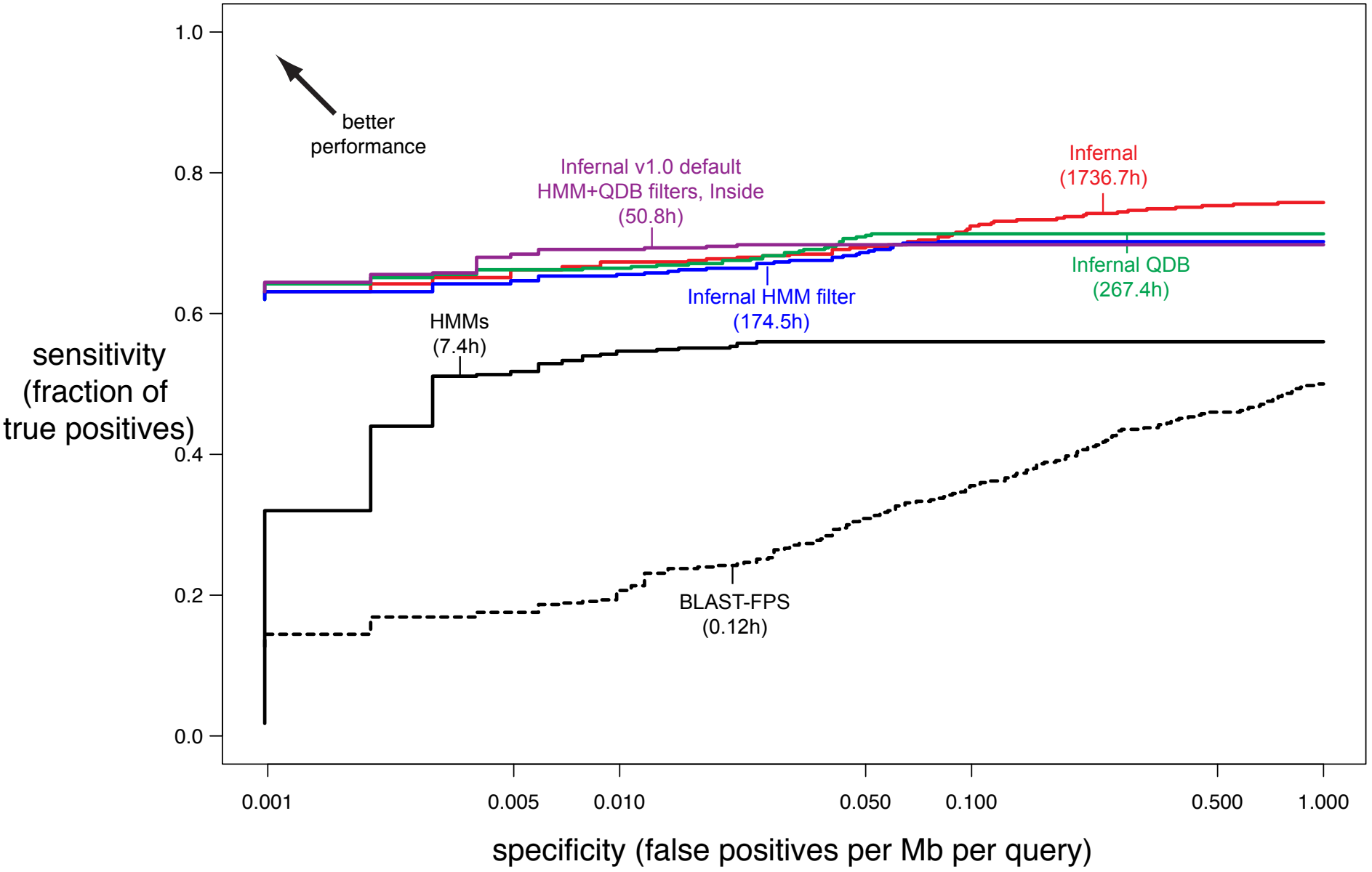
# HMM filters achieve 10-fold speedup at very small cost to accuracy



# Combining QDB and HMM filters yields greater acceleration

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

Infernal is now 30-fold faster and slightly more sensitive.

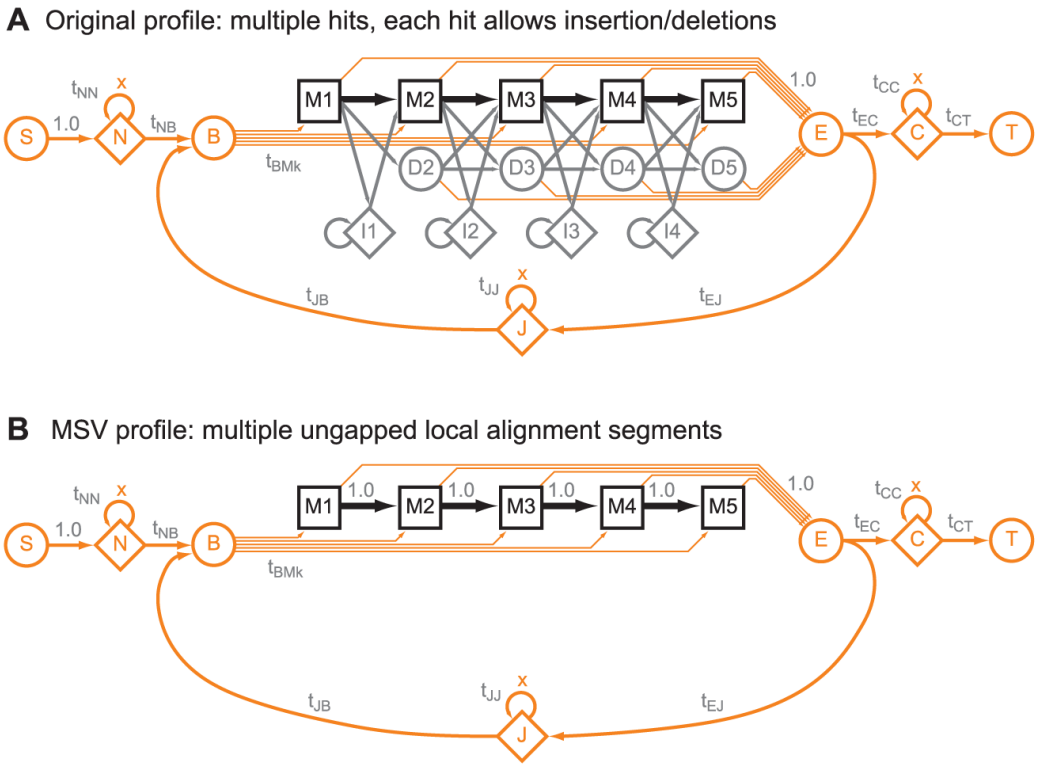


# CMs are nearly as fast as HMMs (usually)

| | | 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 |

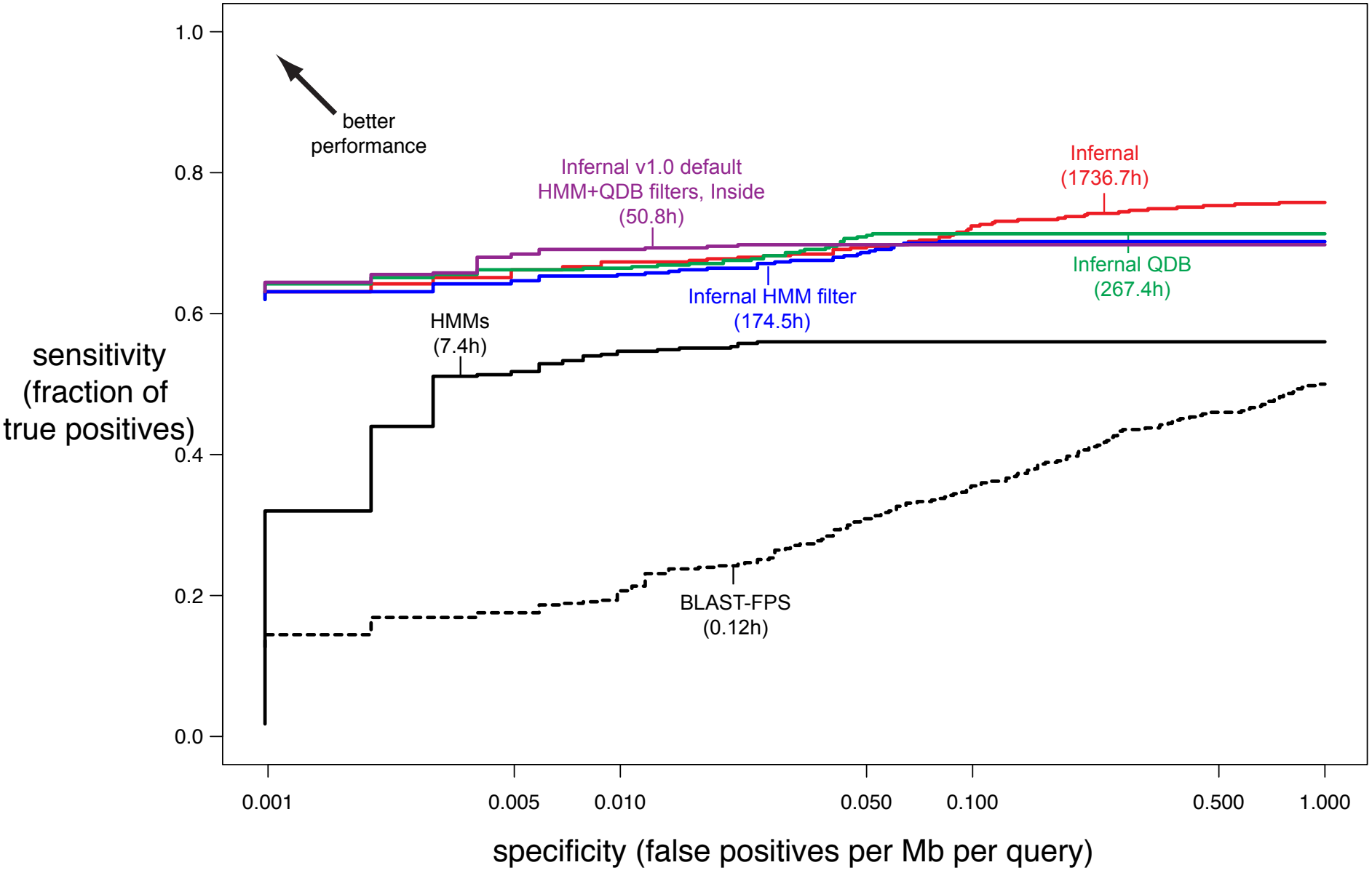
# With HMMER3, HMMs got about 100x faster using an MSV filter

- MSV (Multiple Segment Viterbi) algorithm, which disallows inserts and deletes, can be computed using 16-fold vector parallelism with low-precision (8-bit) scores*.



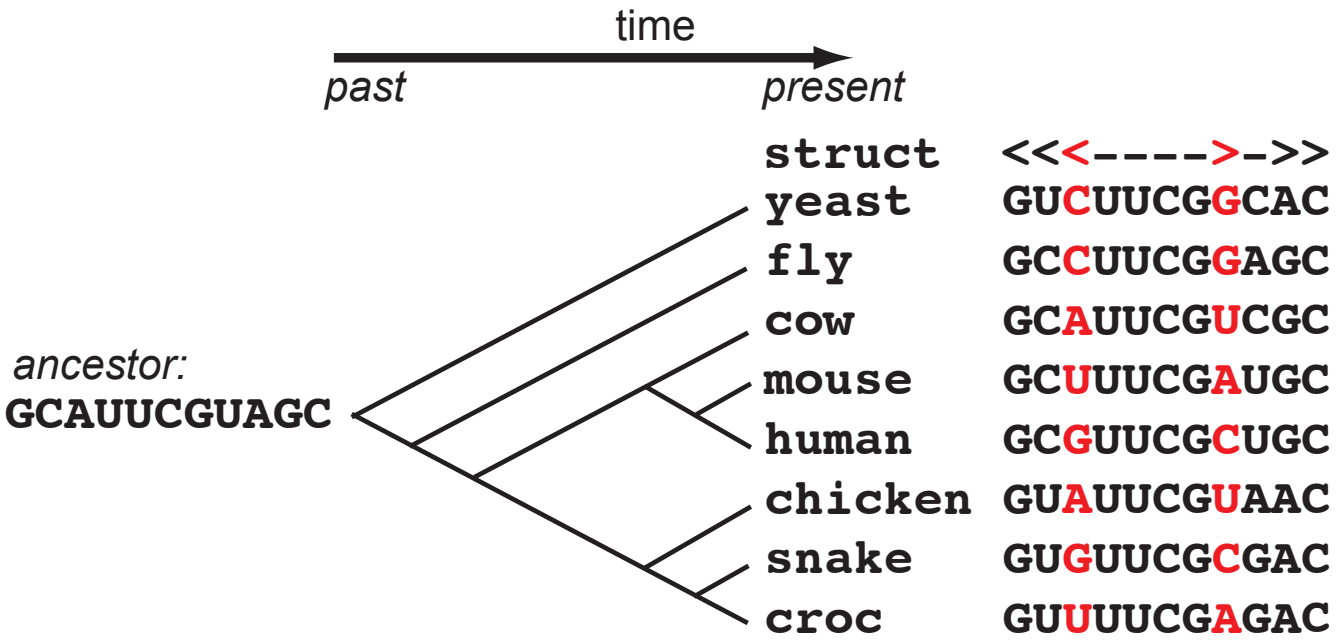

*Eddy SR. PLoS Comp. Biol., 7:e1002195, 2011

# Using faster HMM filters could only accelerate Infernal by about 15% (7.4/50.8)



**CMs can also be used to create structural alignments of homologous RNAs.**

- Given known homologs, place homologous residues in the same columns.

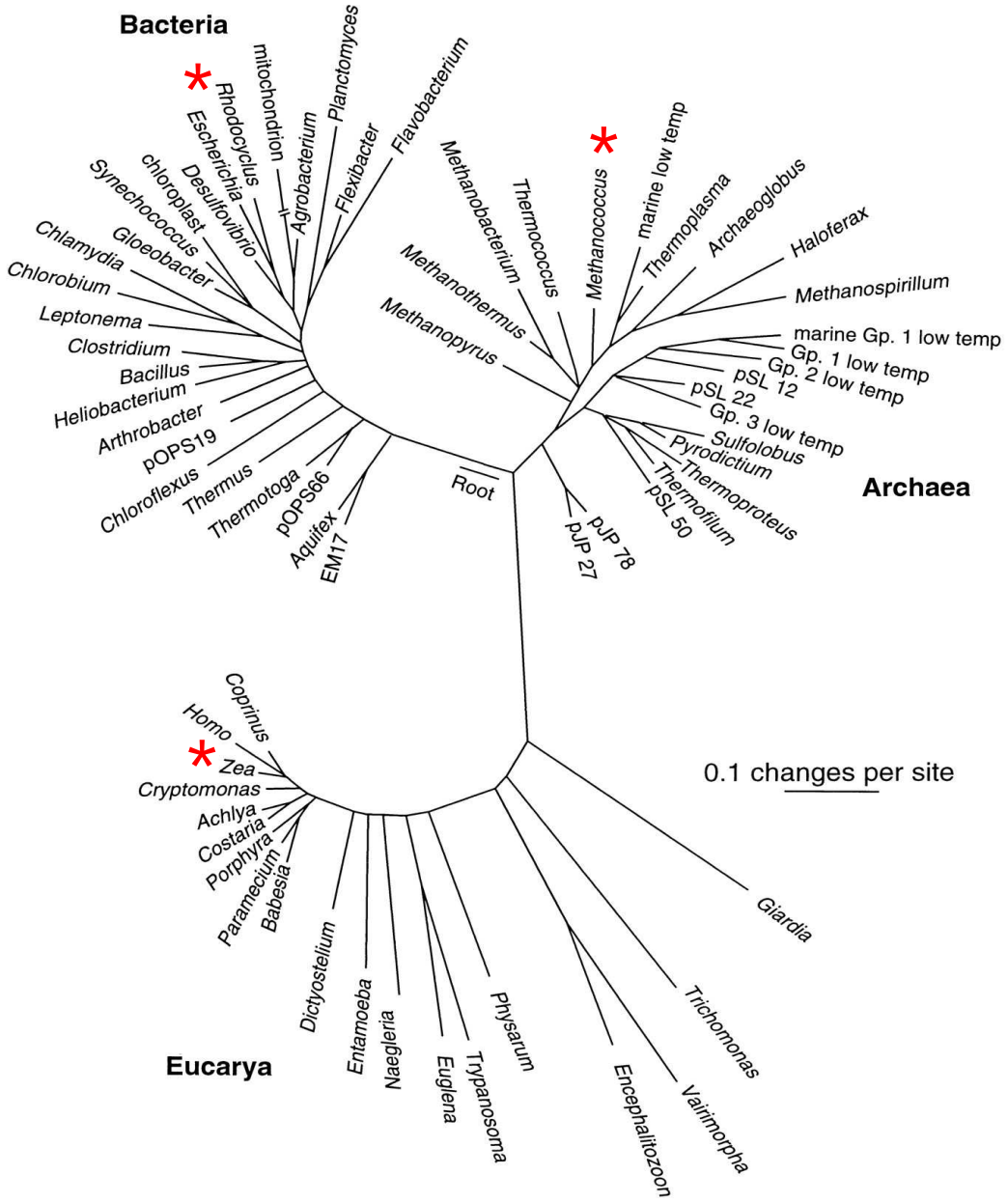



- Alignments of SSU rRNA have commonly been used for phylogenetic inference.

- However, CM alignment is too slow for SSU alignment.
  Aligning a single SSU sequence takes more than 10 minutes.

# Small subunit ribosomal RNA and the tree of life

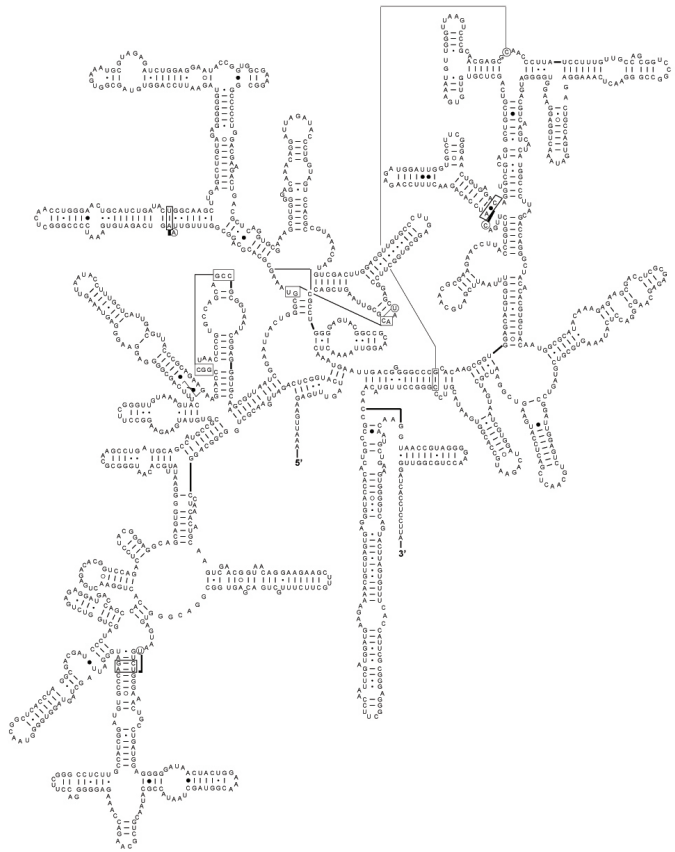
- 1977 - Carl Woese decided to classify all living things phylogenetically
- needed "*a molecule of appropriately broad distribution*" for comparative analysis
- SSU rRNA was chosen
  - universally distributed
  - highly conserved
  - large enough to provide sufficient data (1500-1800 nt)
  - readily isolated



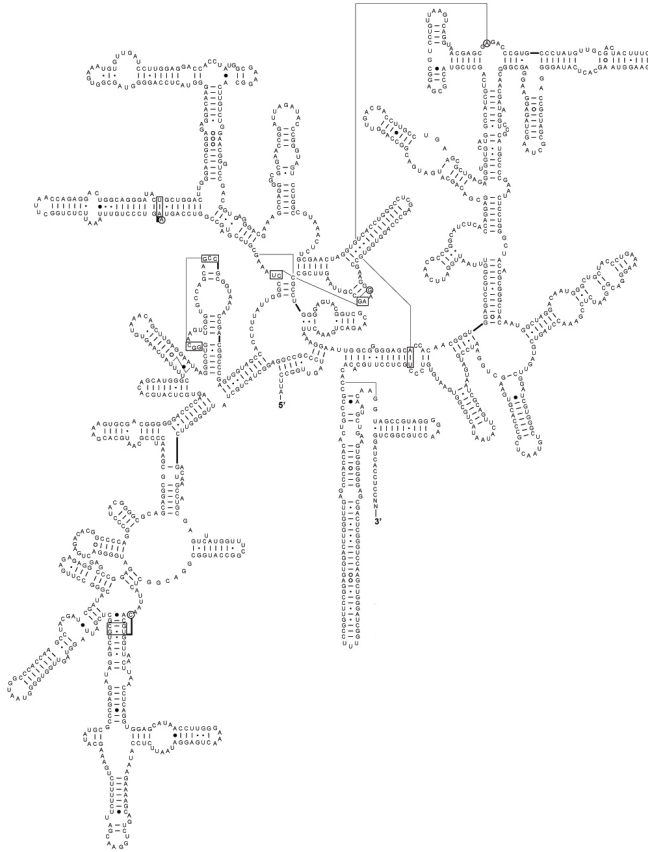

Pace, Science:276; 1997

# Universal conservation of SSU rRNA

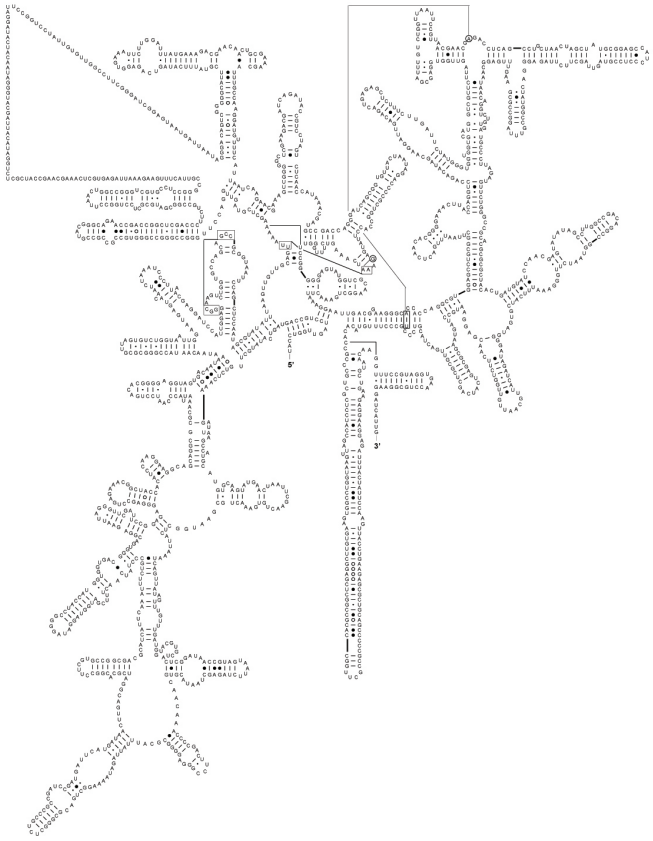
*Escherichia coli*



*Methanococcus vannielii*



*Zea mays*



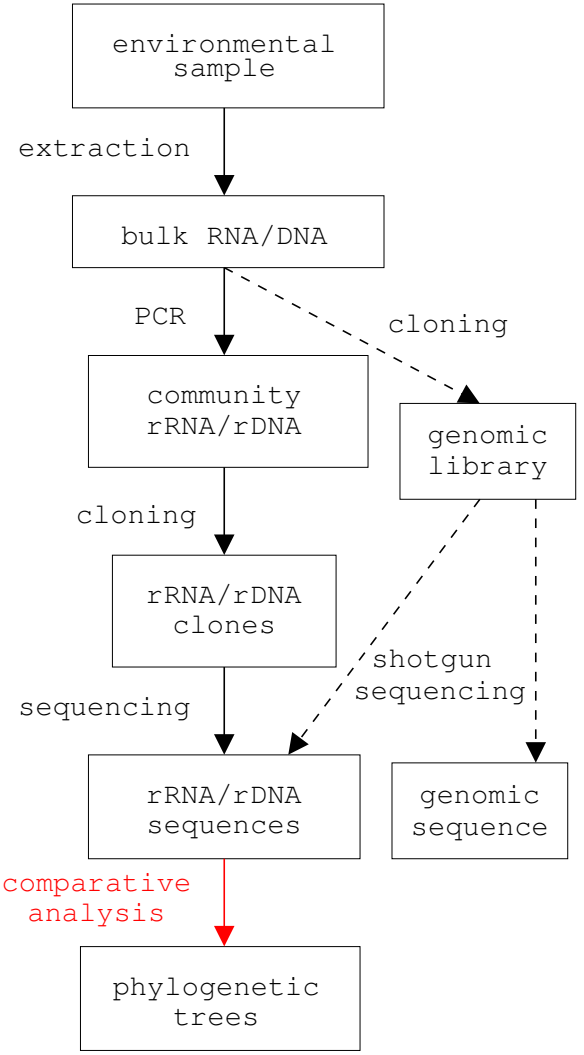
Secondary structure diagrams from:
URL:http://www.rna.ccbb.utexas.edu/

# Environmental surveys target SSU

- mid 1980s - Norman Pace develops methodology for determination of SSU sequences without cultivation
- many different environments have been surveyed
- known biodiversity has been greatly expanded:
  - recognized bacterial phyla:
    11 in 1987, 36 in 1998, 52 in 2003, 67 in 2006...
- SSU databases contain millions of sequences:

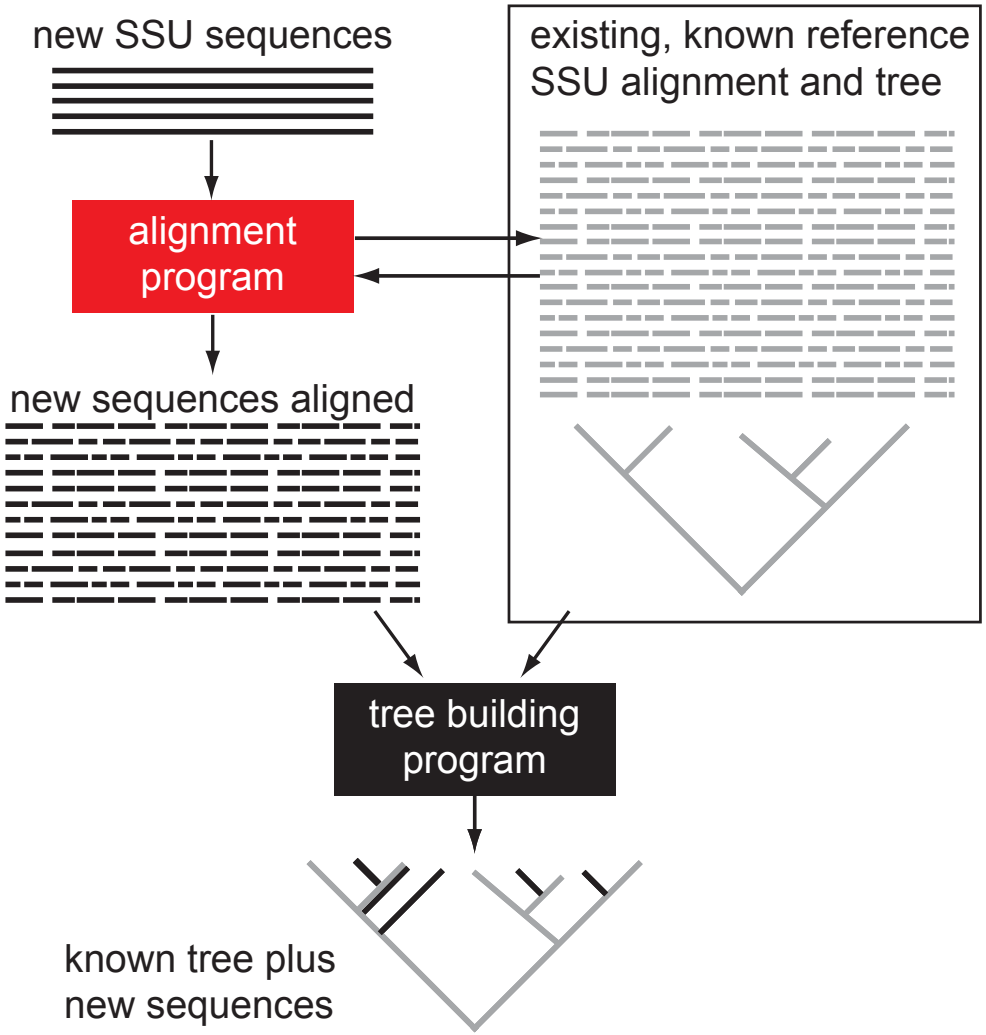
| name | # seqs | # citations |
|---|---|---|
| Silva | 3.2M | 1125 |
| RDP | 2.6M | 1170 |
| Greengenes | 1.0M | 1012 |

Silva: Pruesse et al., 2007 NAR 35.21:7188-96
RDP: Cole et al., 2009 NAR 37:D141-45
Greengenes: DeSantis et al., 2006 AEM 72:5069-72




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

# The comparative analysis step:
# Alignment and Phylogenetic Inference




Goals of the alignment program:

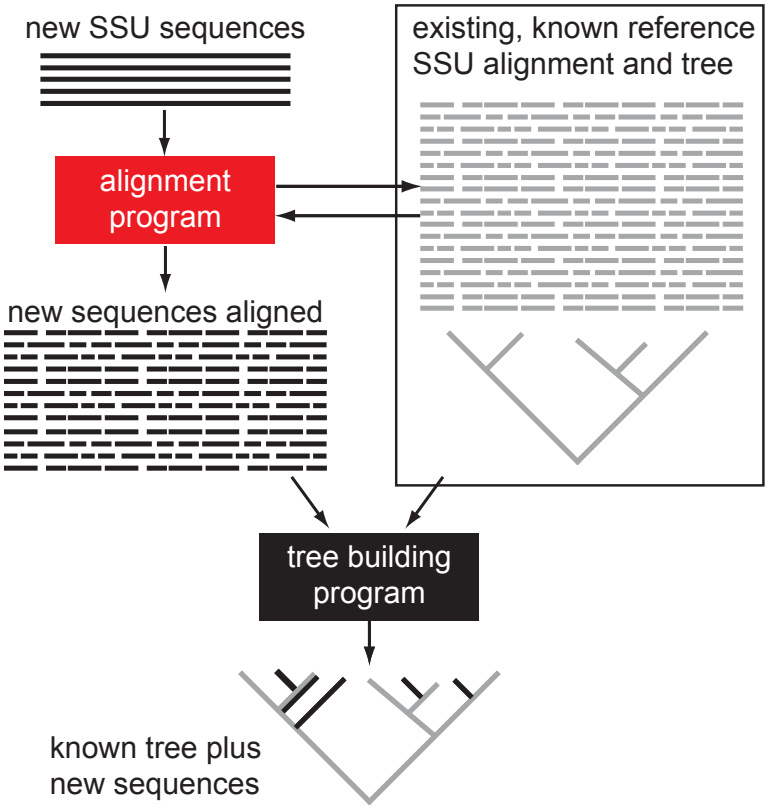
accurate: because alignment errors confound phylogenetic inference

fast: to handle up to millions of seqs

# 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

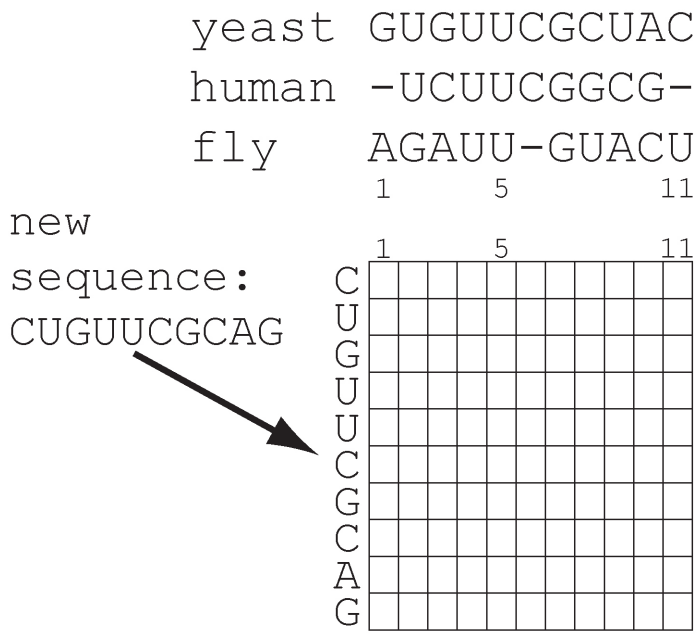
| environment/ phylogeny | domain(s) | #seqs | year |
|---|---|---|---|
| soil, many others | bacteria | 21,752 | 2007 |
| cecal microbiota of mice | bacteria | 5,088<br>4,157 | 2005<br>2006 |
| Sargasso sea | all 3 | 1,164 | 2004 |
| hydrothermal vents | eukarya | 374 | 2002 |
| endolithic environment (pore space of rocks) | archaea, bacteria | 342<br>588 | 2005<br>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

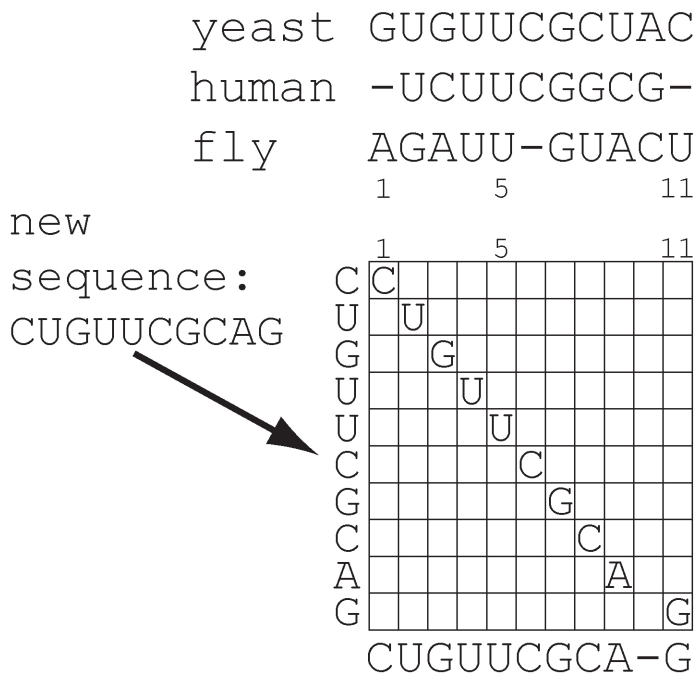


# Accelerating CM alignment using HMMs

- **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



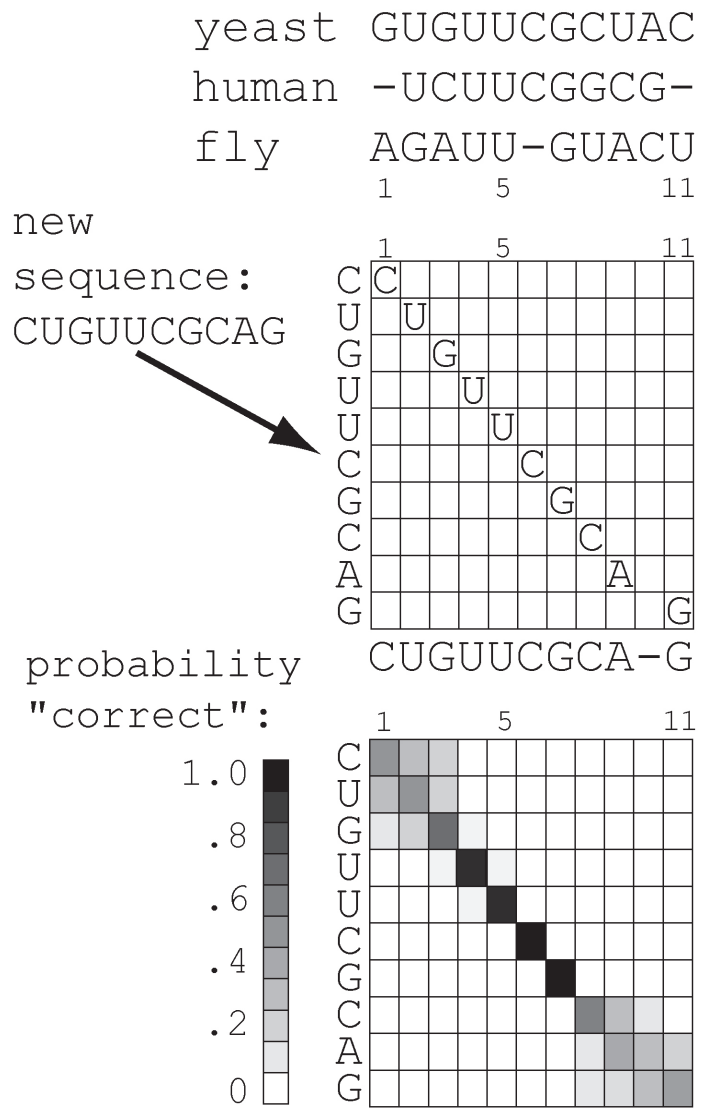
# Accelerating CM alignment using HMMs

- **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

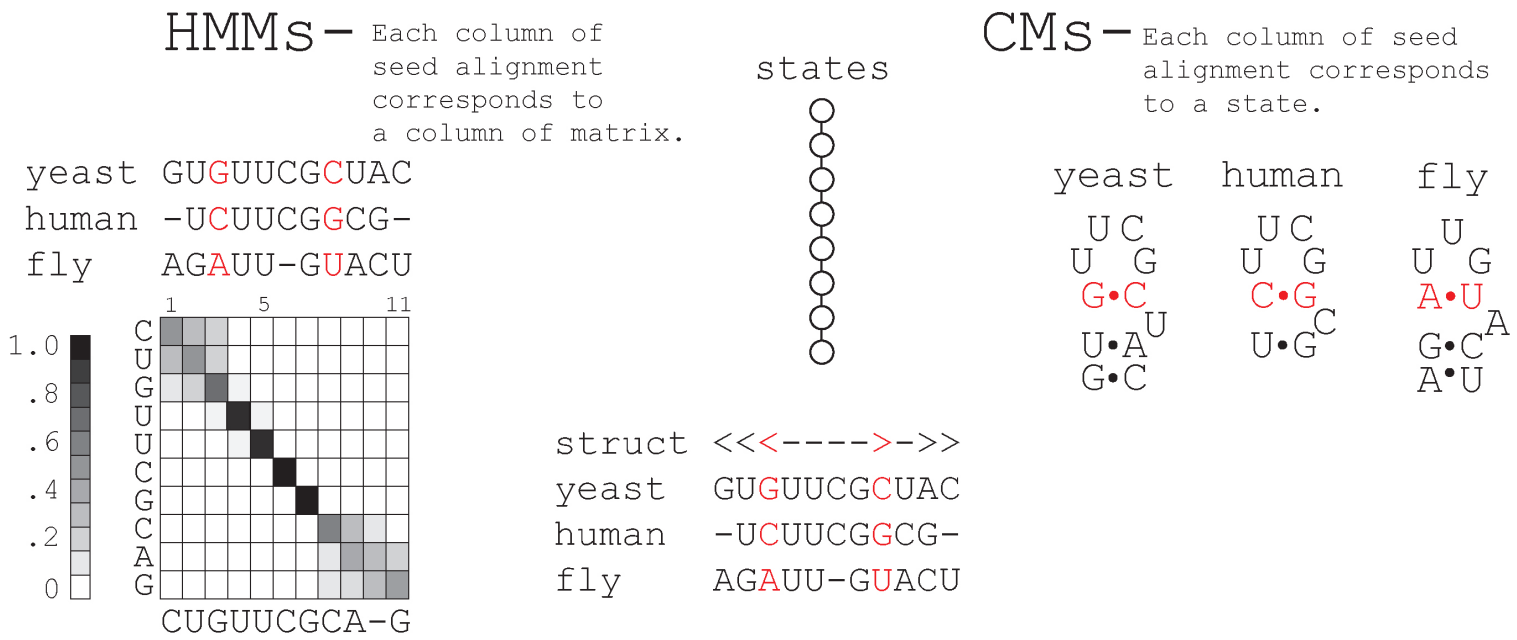
- each column of the grid corresponds to a column of the seed alignment
- each row of the grid corresponds to a position of the new sequence

**How can we use this information during CM alignment?**



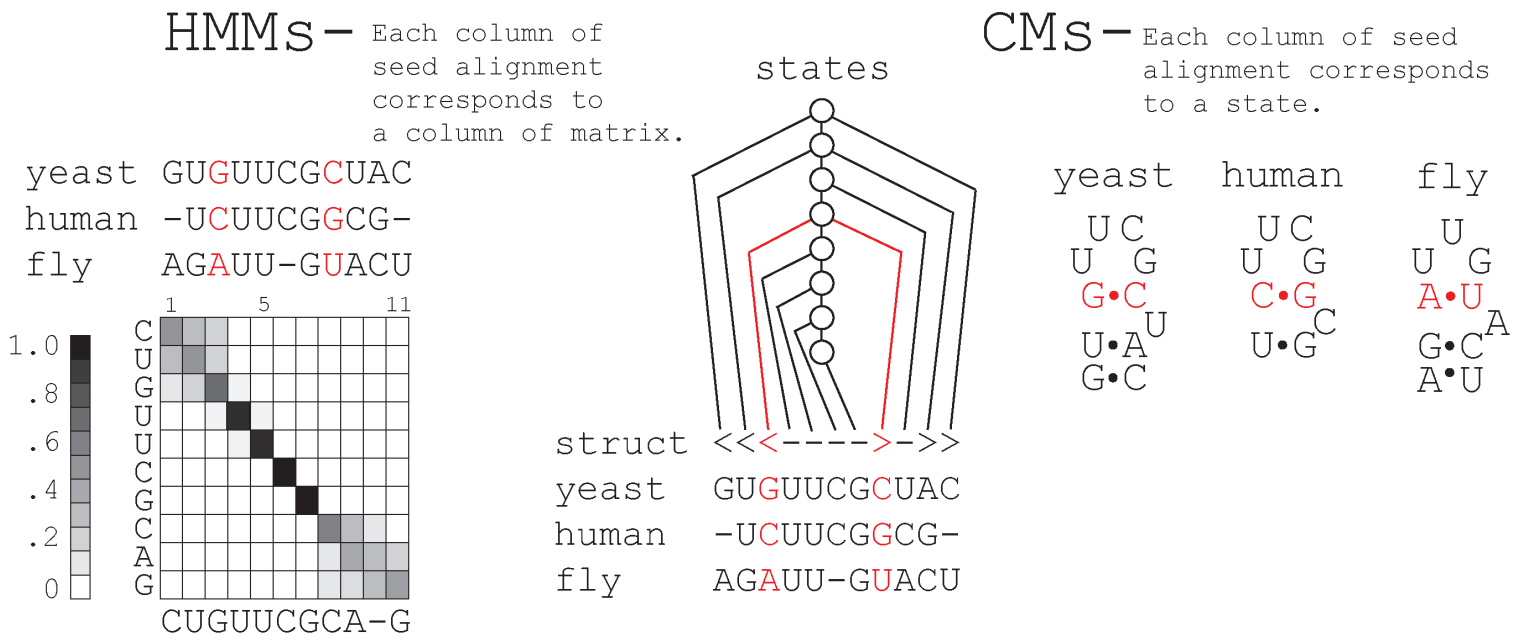
# HMM bands accelerate CM alignment

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



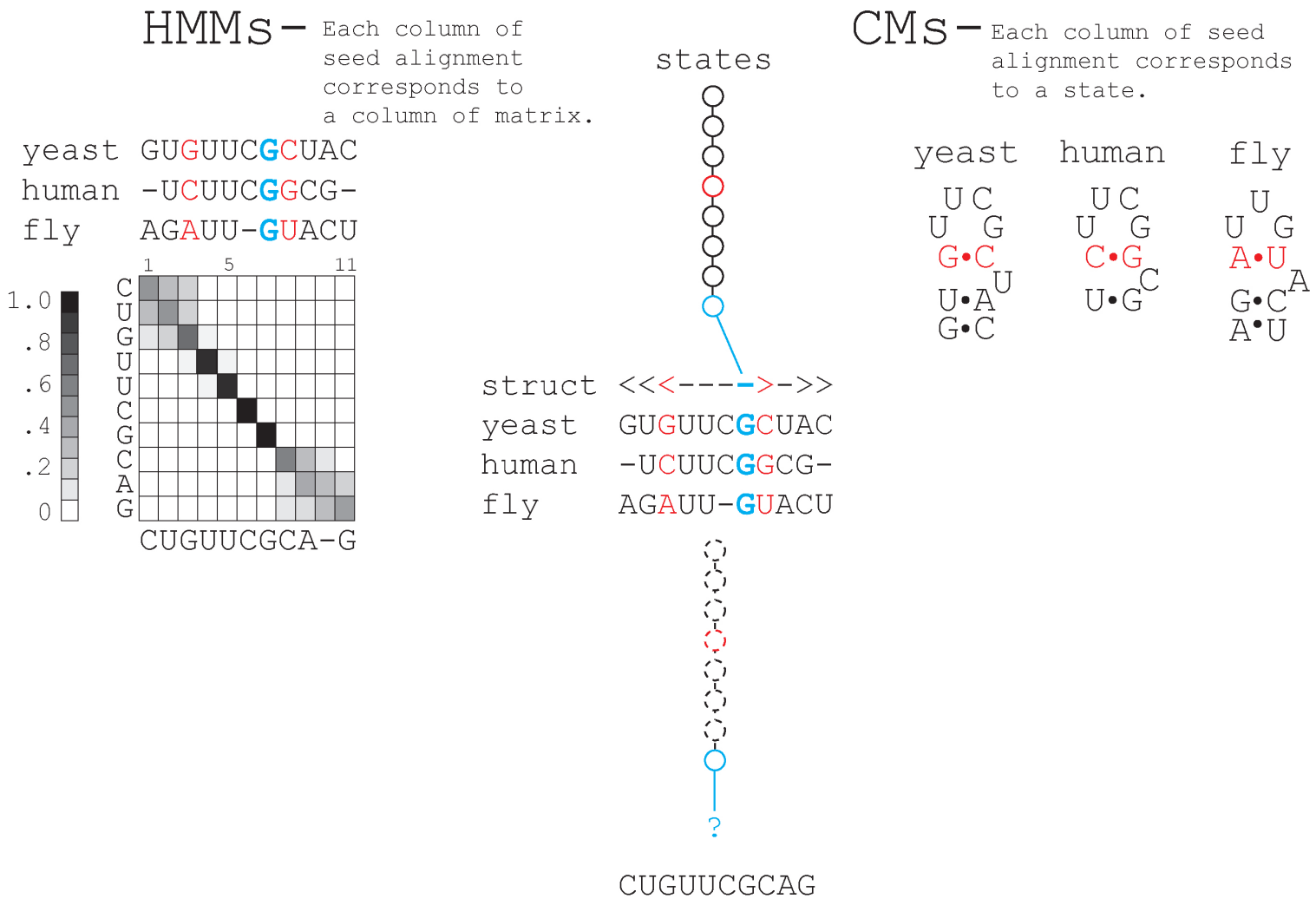
# HMM bands accelerate CM alignment

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



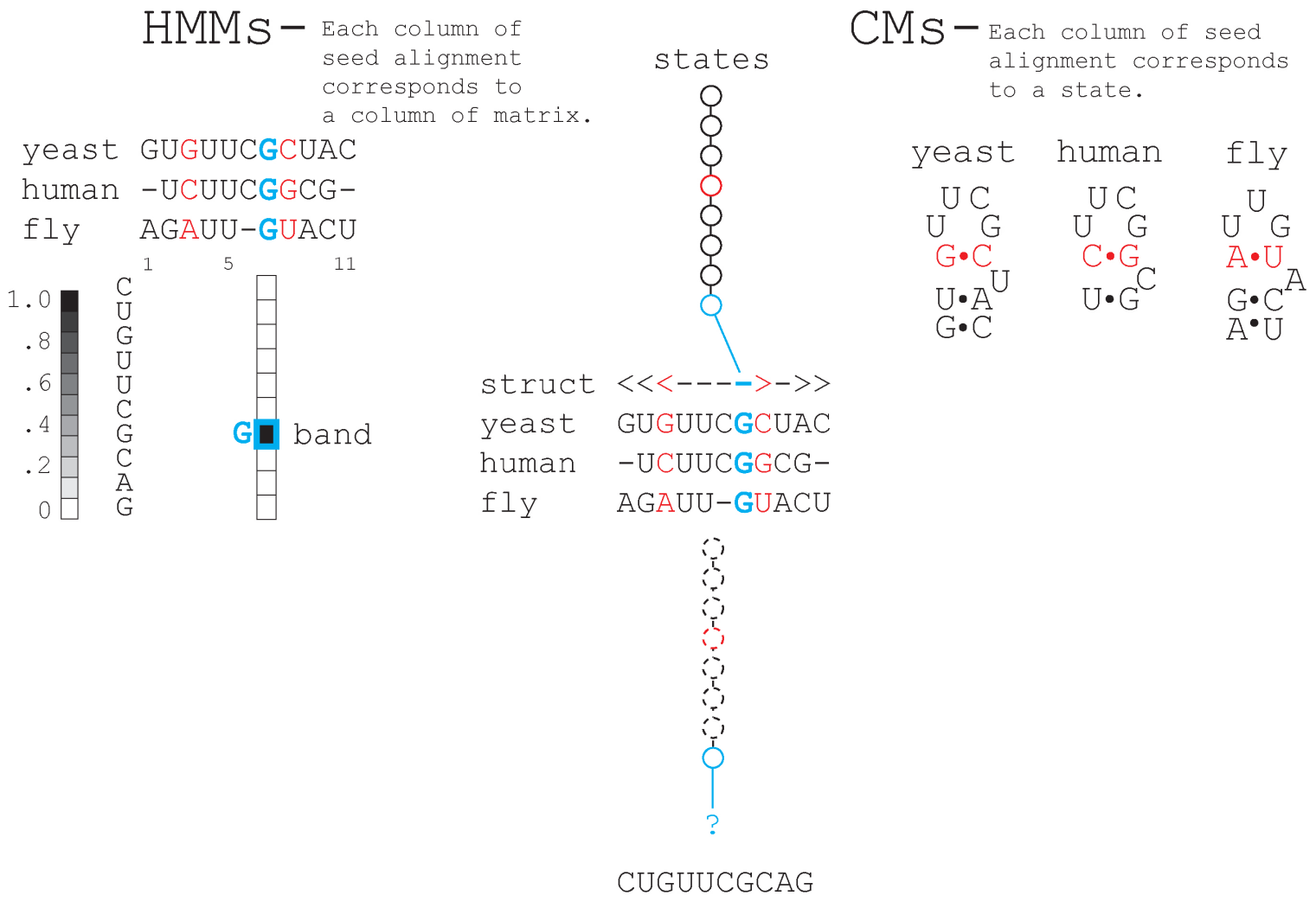
# HMM bands accelerate CM alignment

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



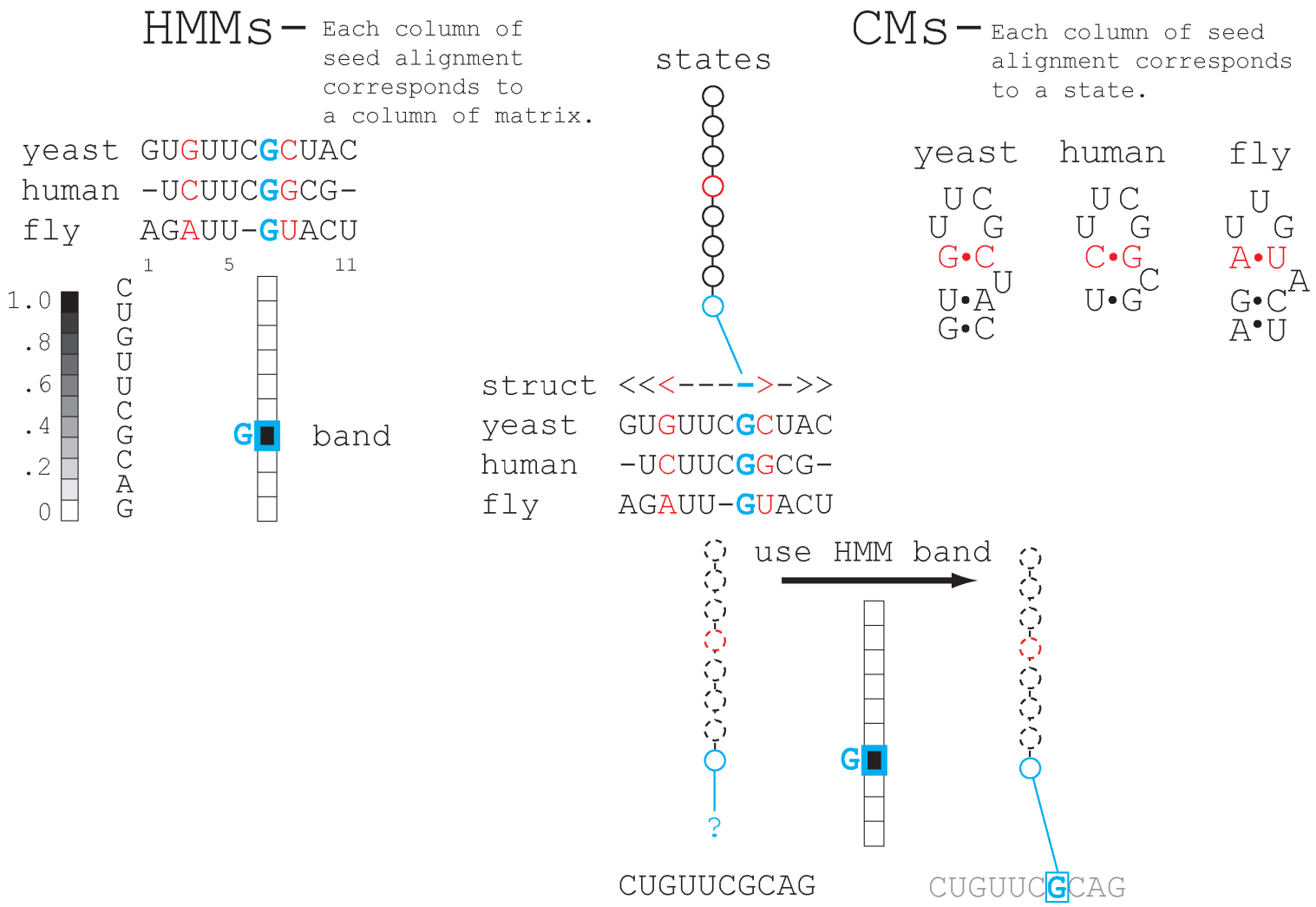
# HMM bands accelerate CM alignment

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



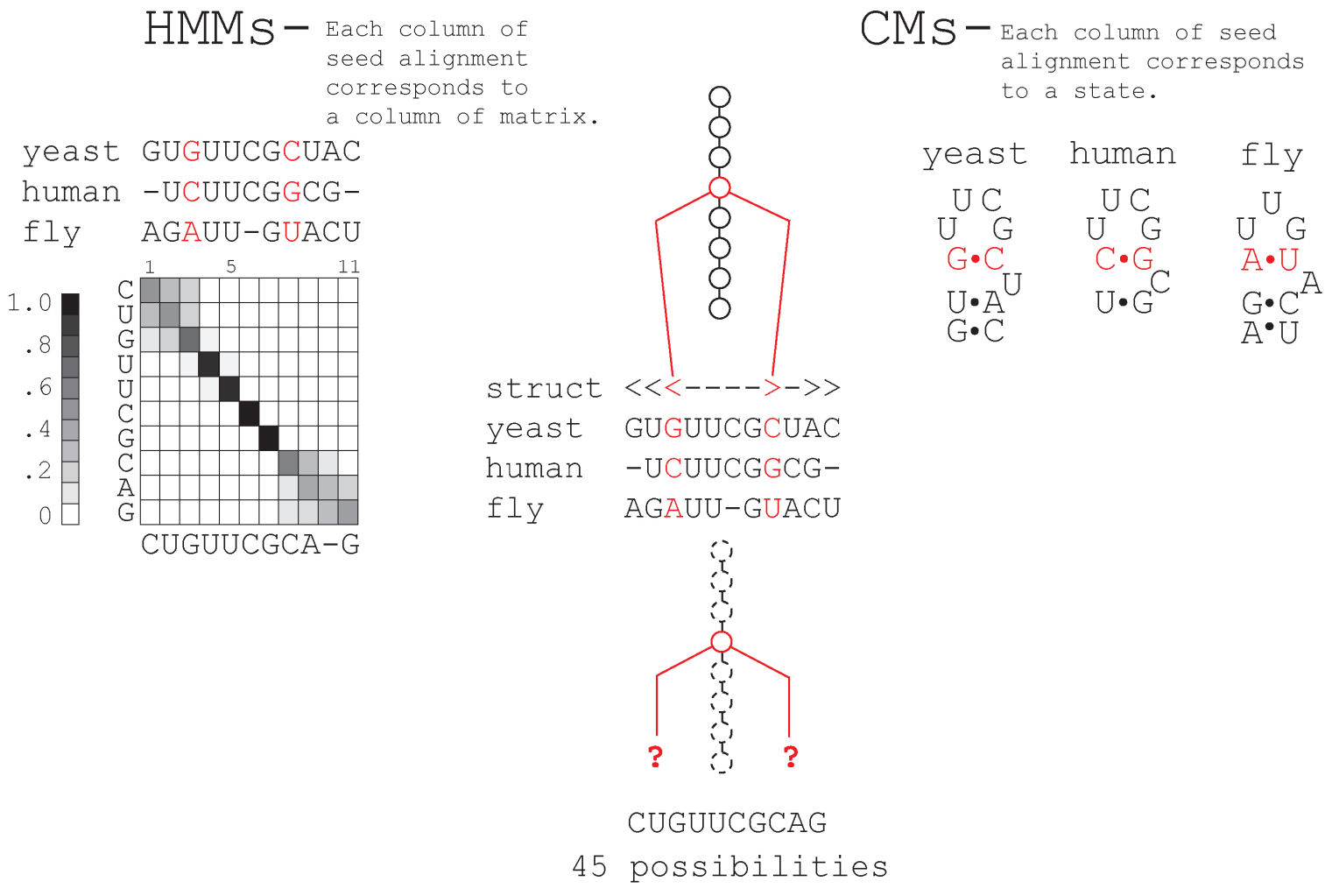
# HMM bands accelerate CM alignment

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



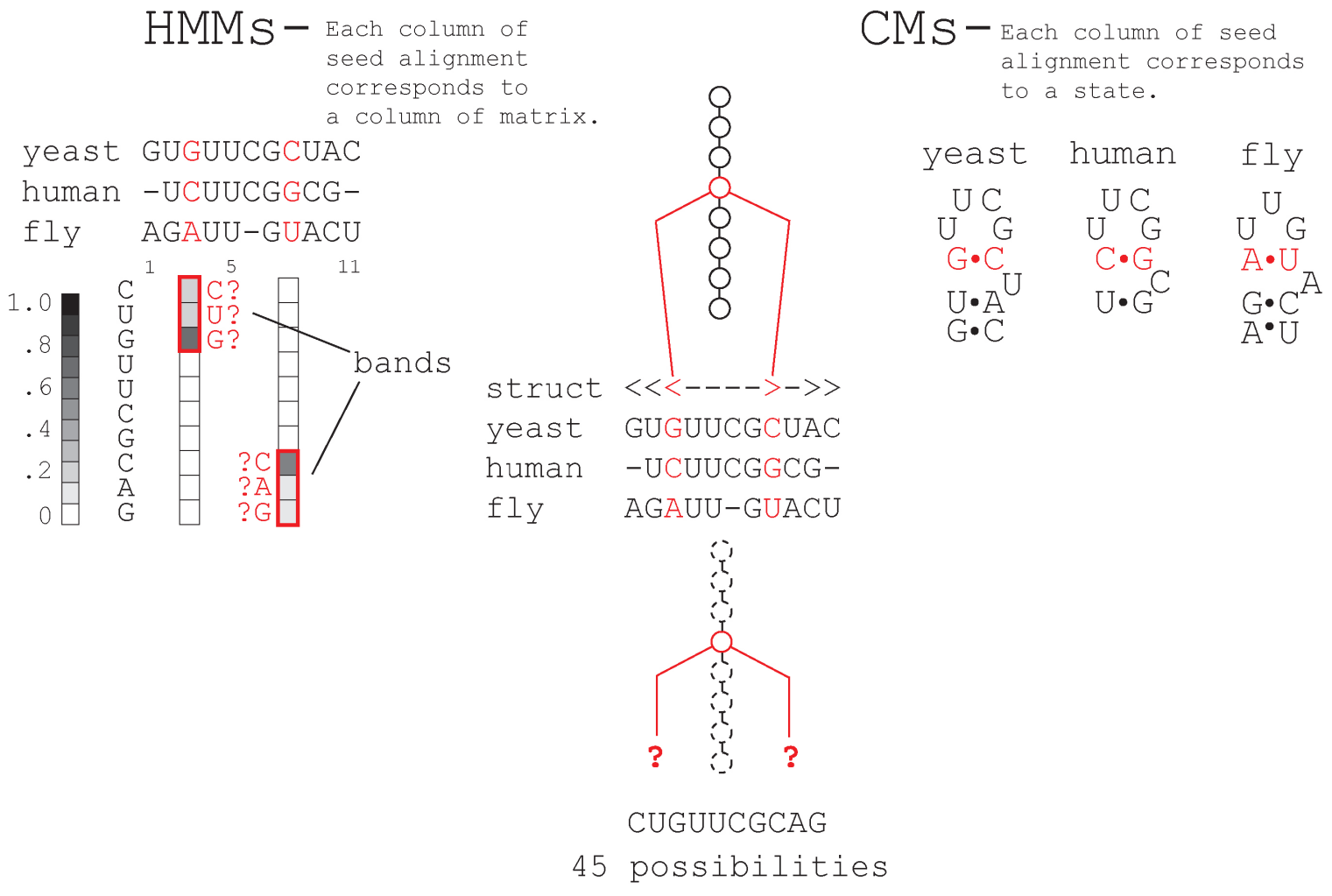
# HMM bands accelerate CM alignment

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



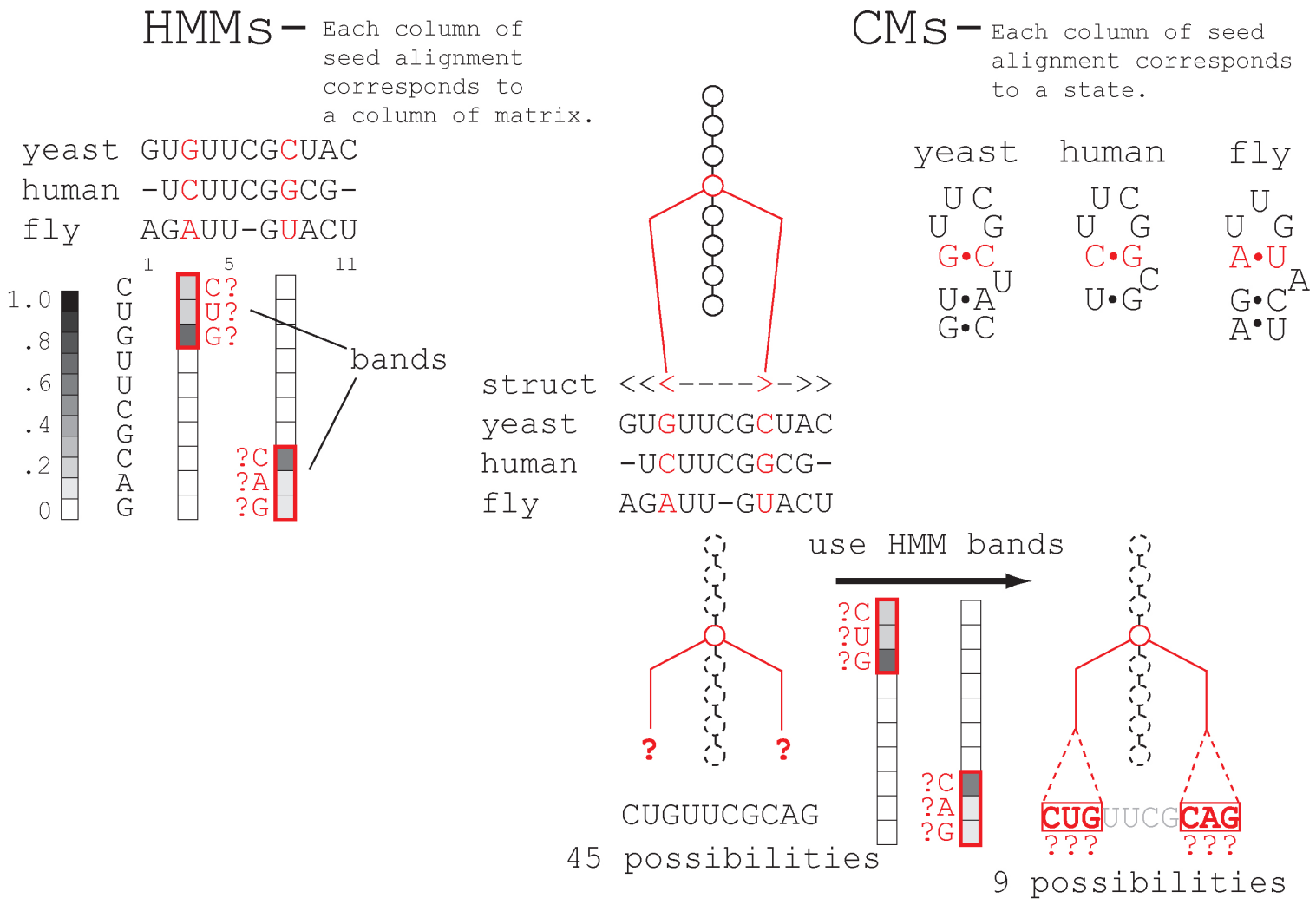
# HMM bands accelerate CM alignment

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



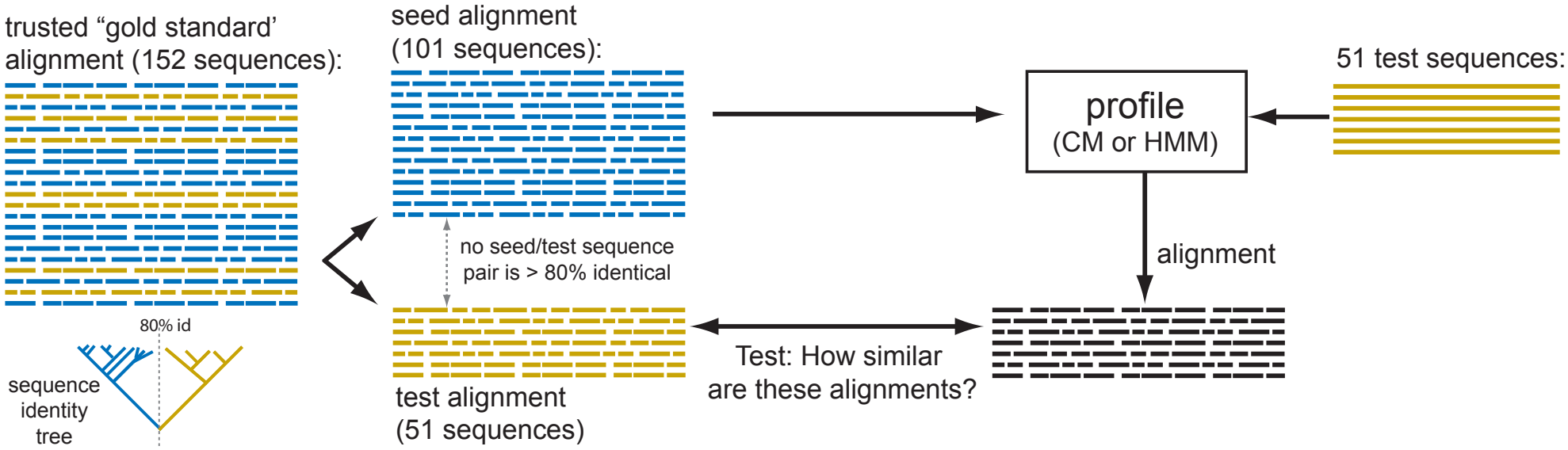
# HMM bands accelerate CM alignment

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



# Determining the impact on speed and accuracy

- 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) |
|---|---|---|
| Muscle-3.8.31 (*de novo*) | 95.4% | 0.49 |
| HMMER3 (HMMs) | 96.8% | 0.04 |
| non-banded CMs | 98.1% | ? |

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

| | alignment accuracy | time (sec/seq) |
|---|---|---|
| Muscle-3.8.31 (*de novo*) | 95.4% | 0.49 |
| HMMER3 (HMMs) | 96.8% | 0.04 |
| non-banded CMs | 98.1% | ? |
| HMM banded CMs | 98.1% | 0.50 |

# HMM banding accelerates CM alignment 1000-fold

| | alignment accuracy | time (sec/seq) |
|---|---|---|
| Muscle-3.8.31 (*de novo*) | 95.4% | 0.49 |
| HMMER3 (HMMs) | 96.8% | 0.04 |
| Infernal 1.1 (CMs) | 98.1% | 0.50 |

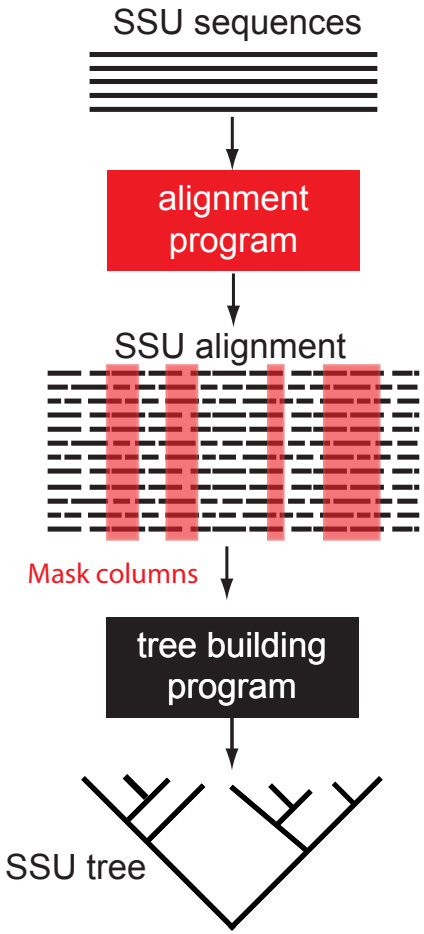
Muscle: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

HMMER: hmmer.janelia.org

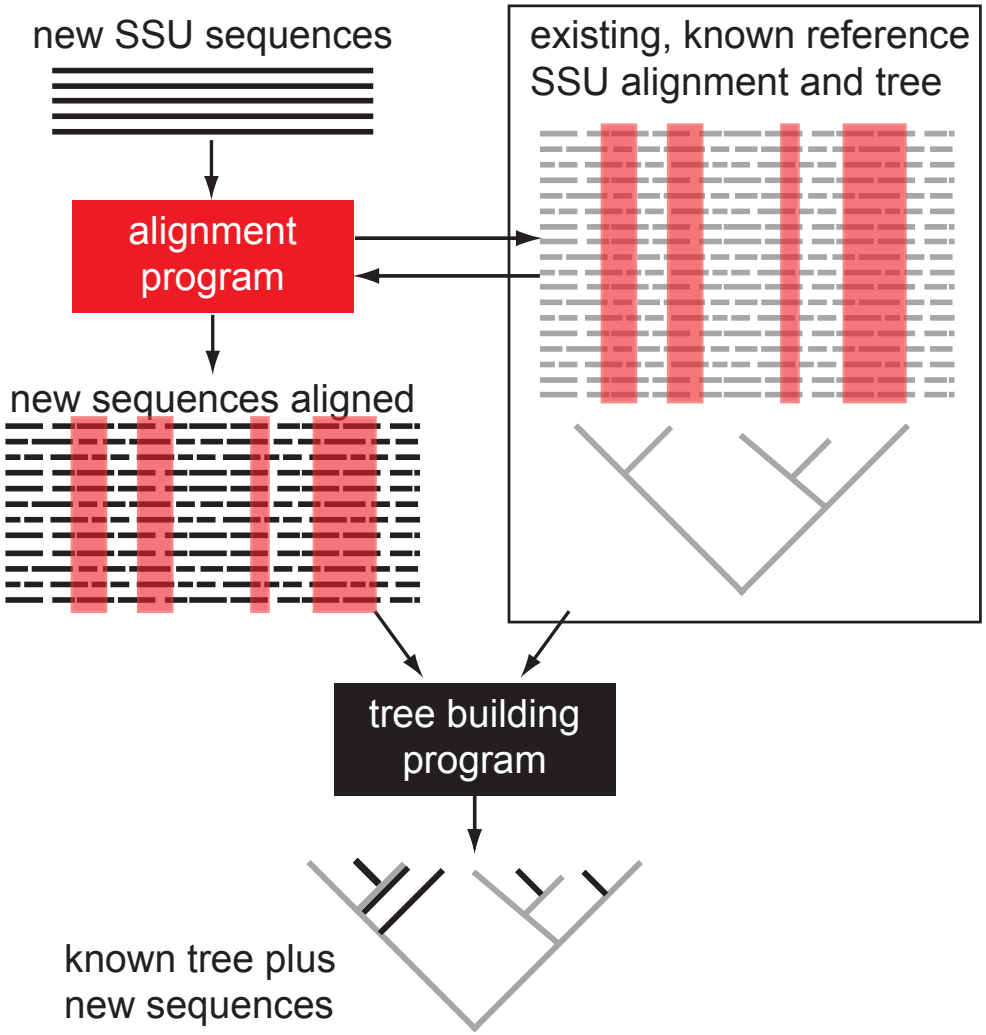
Infernal: infernal.janelia.org

# Probabilistic models allow direction calculation of useful quantities

- Posterior decoding algorithm computes the posterior probability that each nucleotide is correctly aligned given the model
  - allows HMM banding for CM alignment ($O(N^3 log(N))$ reduced to close to $O(N^2)$
  - useful for identifying and removing (masking) columns that are not reliably aligned prior to phylogenetic inference



# The comparative analysis step:
# Alignment and Phylogenetic Inference

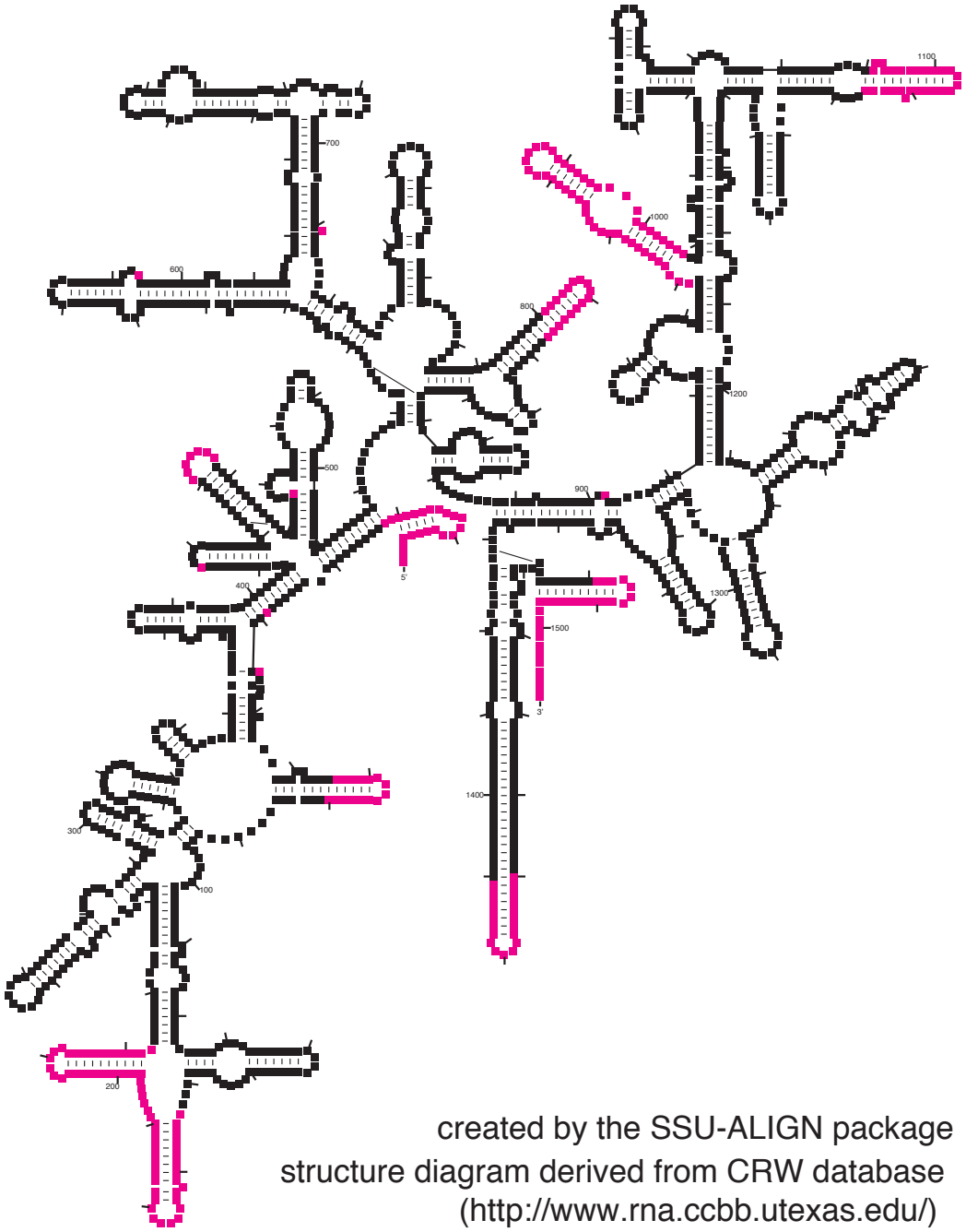



Goals of the alignment program:

accurate: because alignment errors confound phylogenetic inference

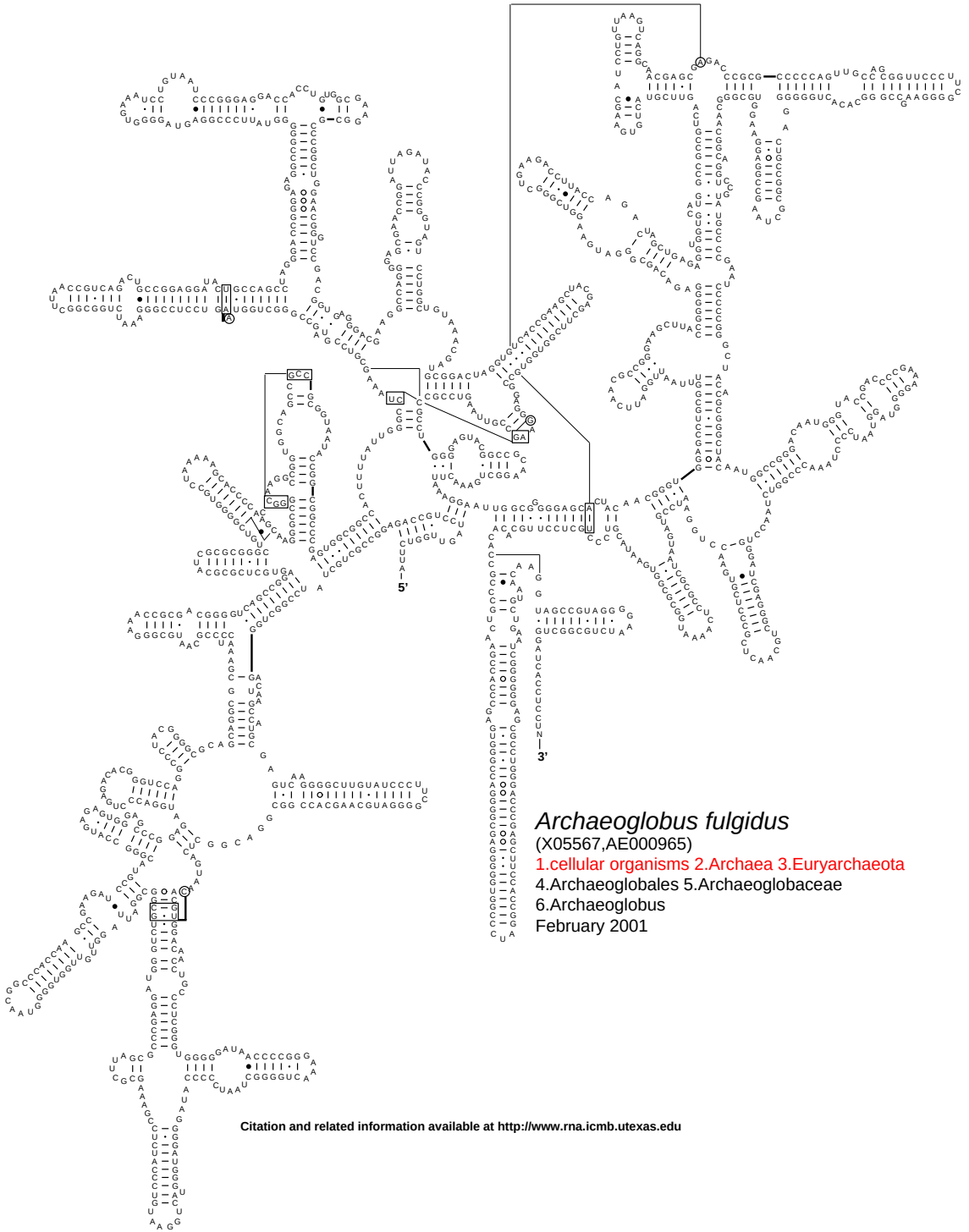
fast: to handle up to millions of seqs

Phil Hugenholtz's manually created mask imposed on archaeal SSU
black: included in alignment (1257)
pink: excluded from alignment (251)



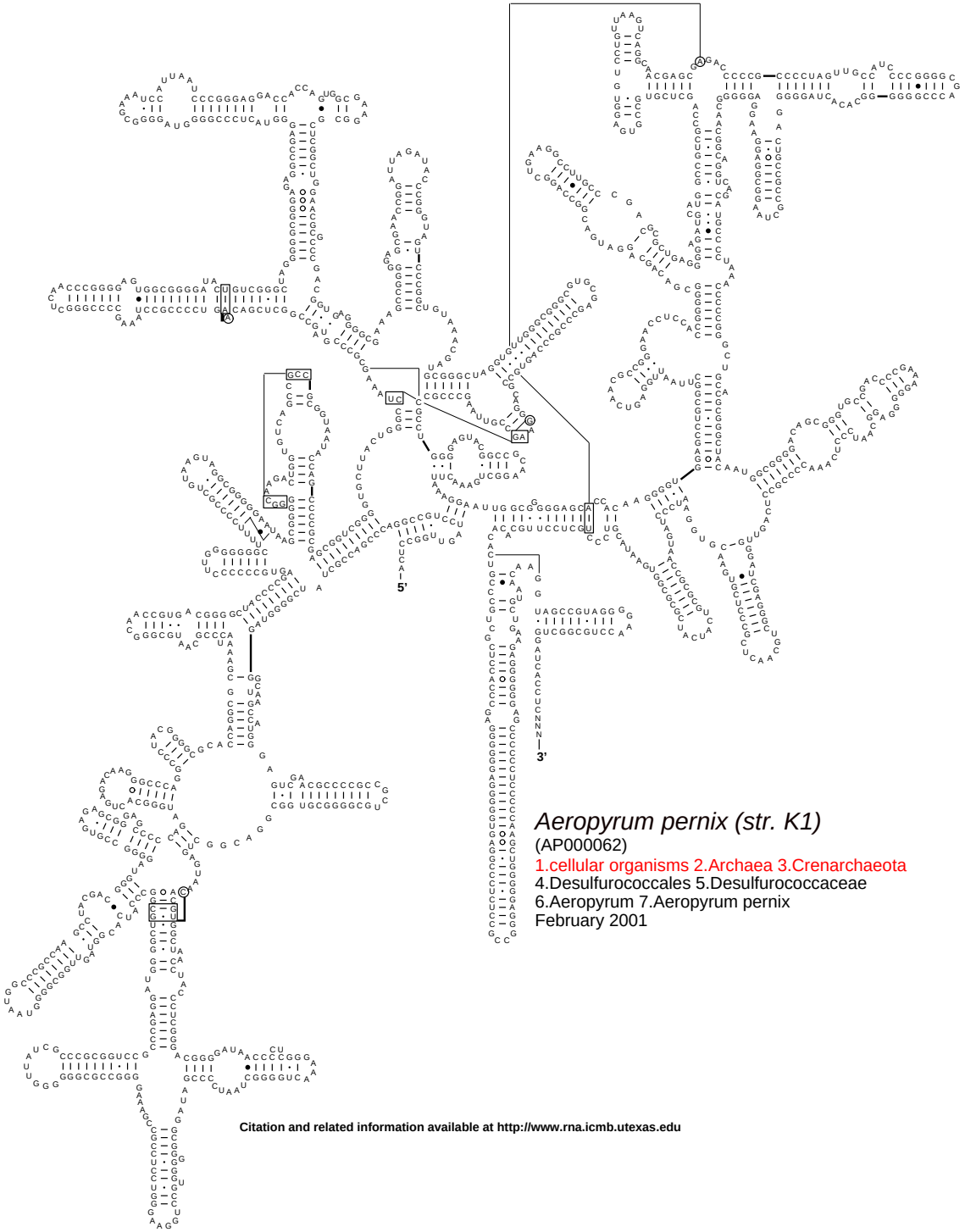

created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

# Secondary Structure: small subunit ribosomal RNA



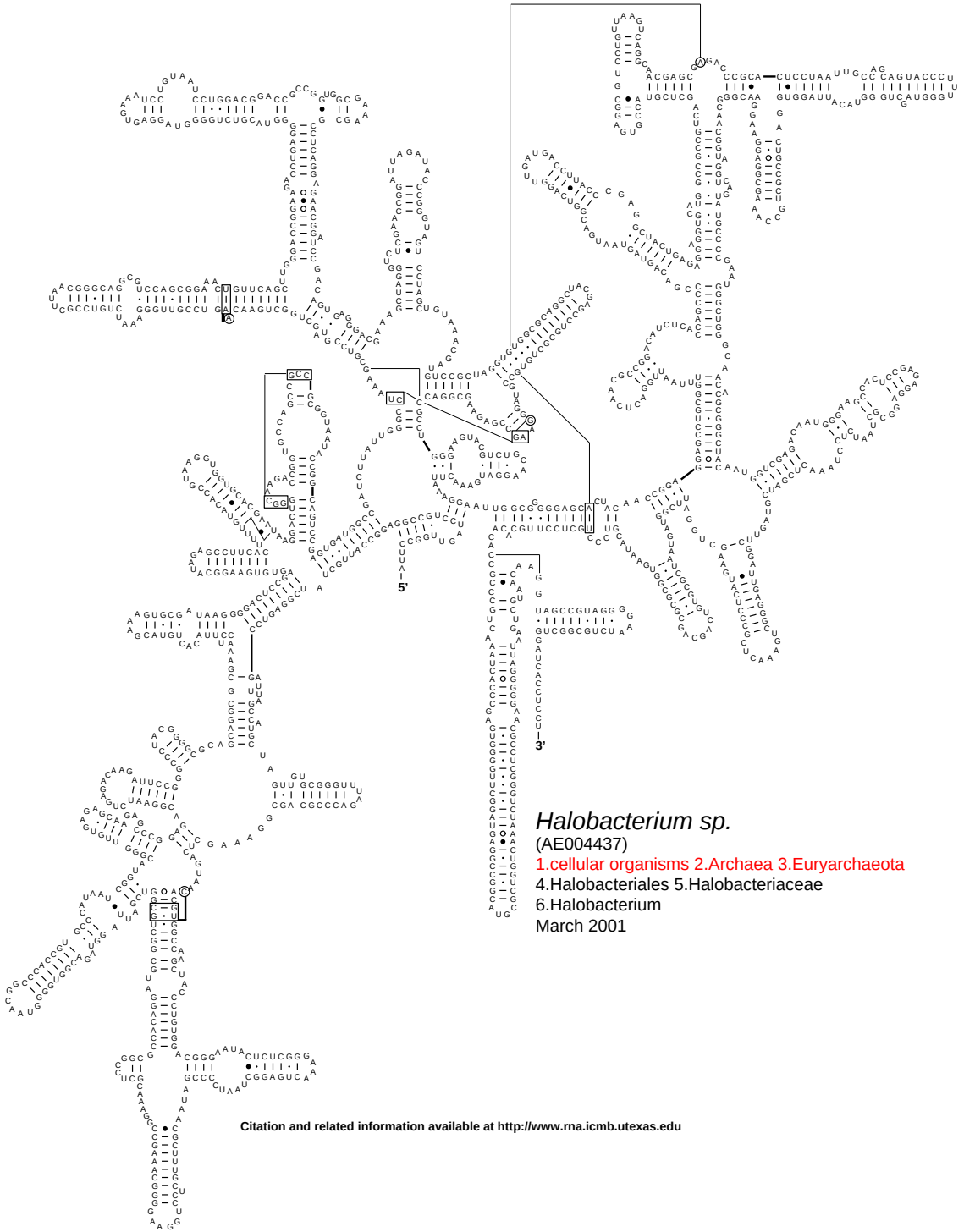

*Archaeoglobus fulgidus*
(X05567,AE000965)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Archaeoglobales 5.Archaeoglobaceae
6.Archaeoglobus
February 2001

# Secondary Structure: small subunit ribosomal RNA



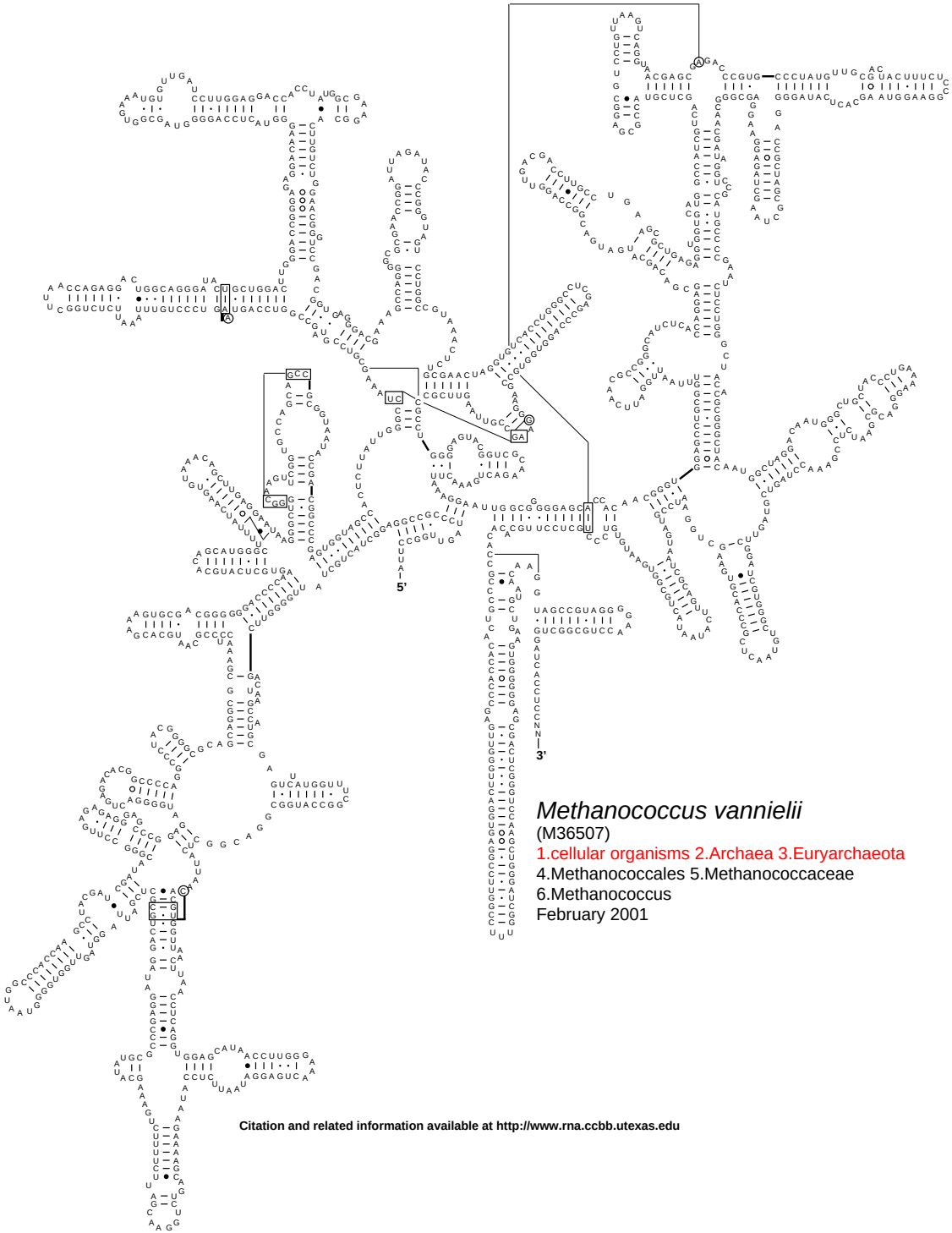

*Aeropyrum pernix (str. K1)*
(AP000062)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Desulfurococcales 5.Desulfurococcaceae
6.Aeropyrum 7.Aeropyrum pernix
February 2001

# Secondary Structure: small subunit ribosomal RNA



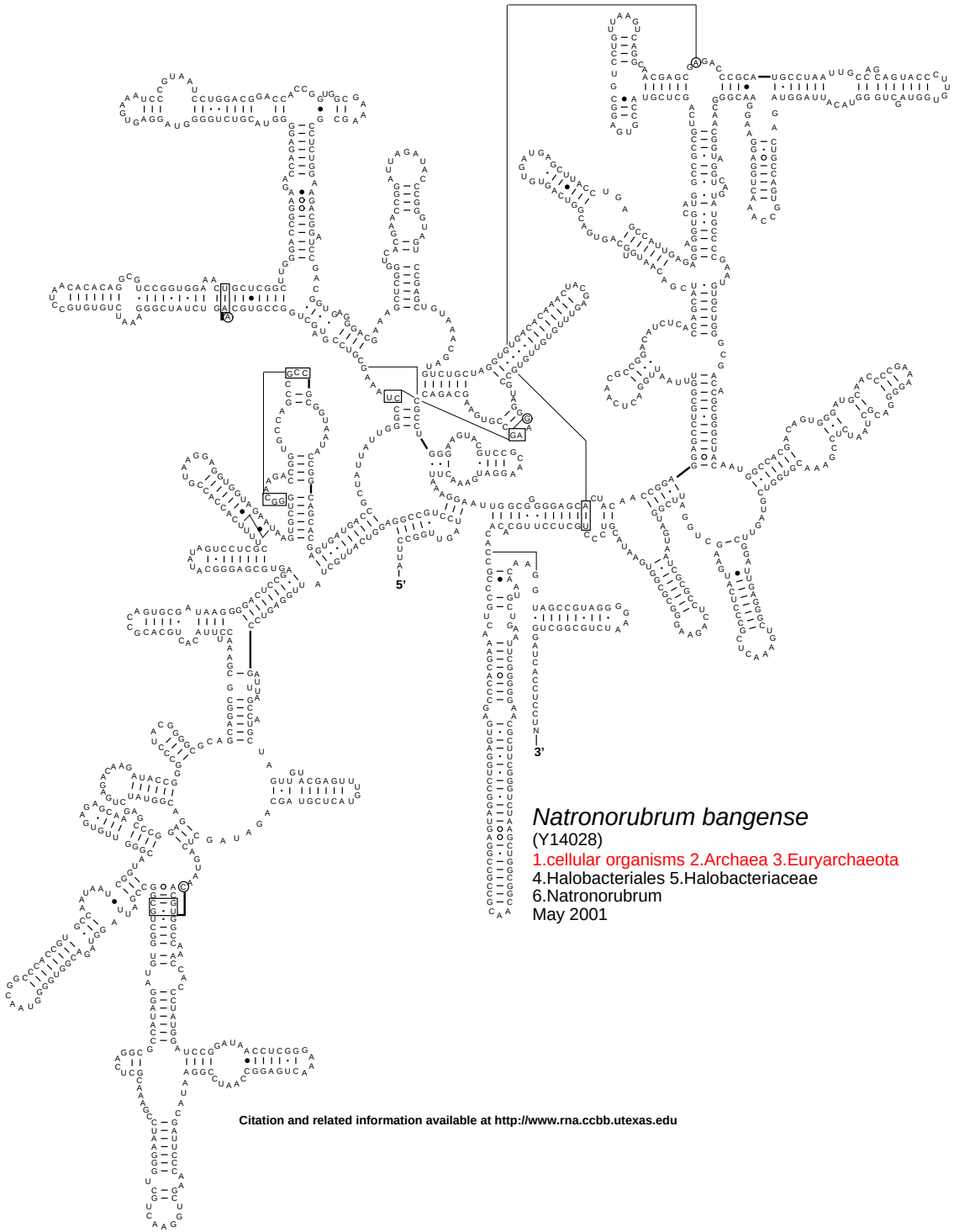

*Halobacterium sp.*
(AE004437)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Halobacterium
March 2001

# Secondary Structure: small subunit ribosomal RNA



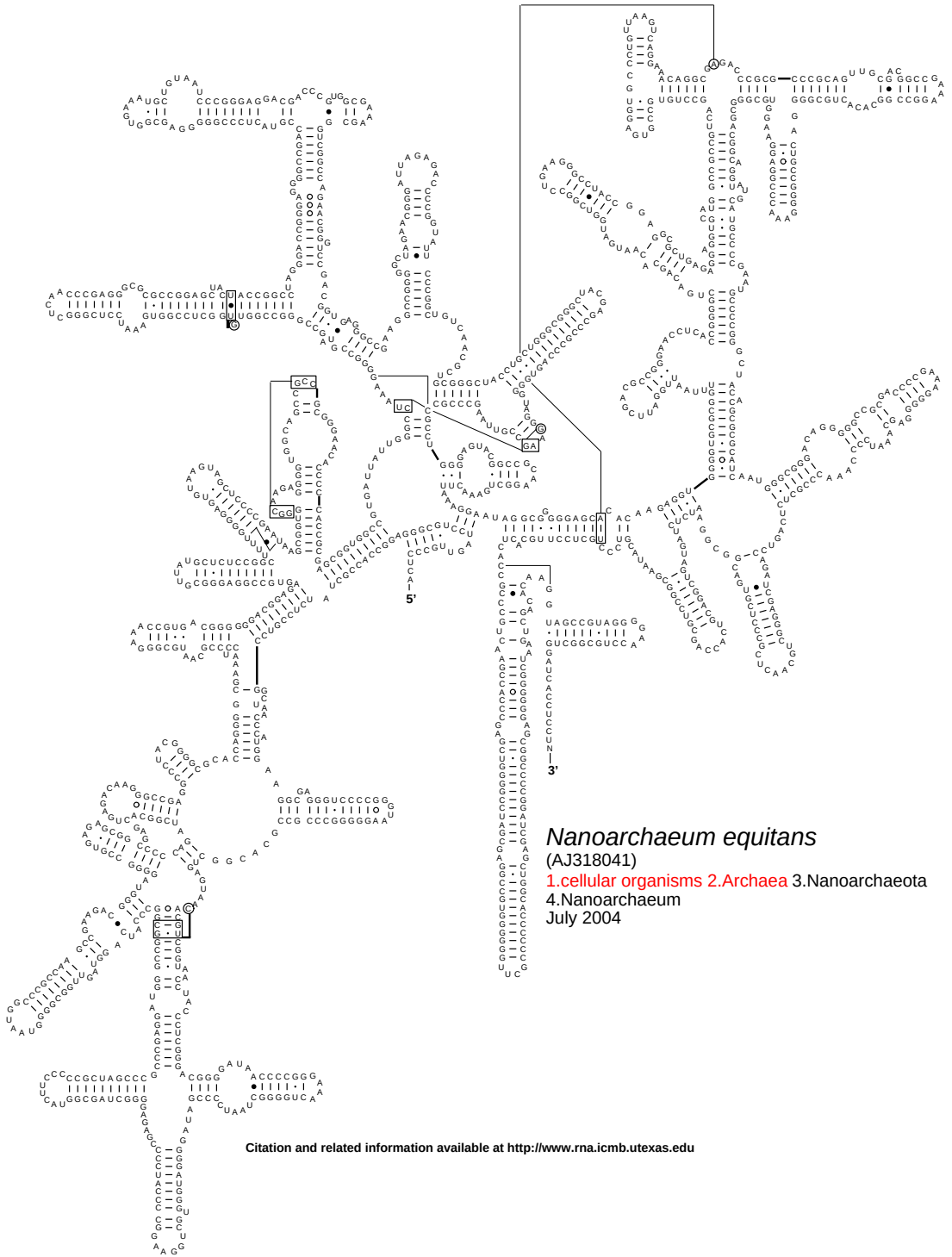

*Methanococcus vannielii*
(M36507)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Methanococcales 5.Methanococcaceae
6.Methanococcus
February 2001

# Secondary Structure: small subunit ribosomal RNA



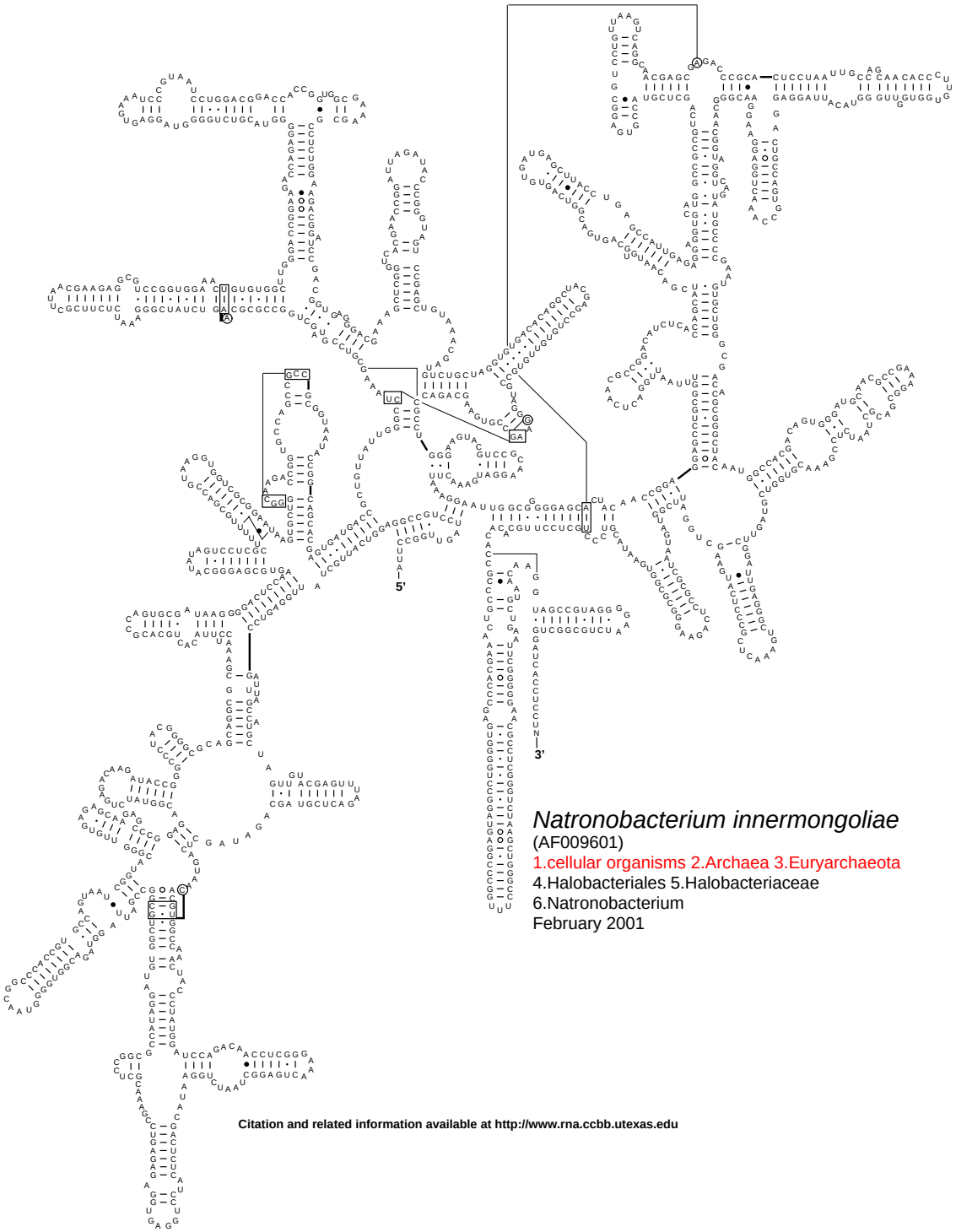

*Natronorubrum bangense*
(Y14028)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronorubrum
May 2001

# Secondary Structure: small subunit ribosomal RNA



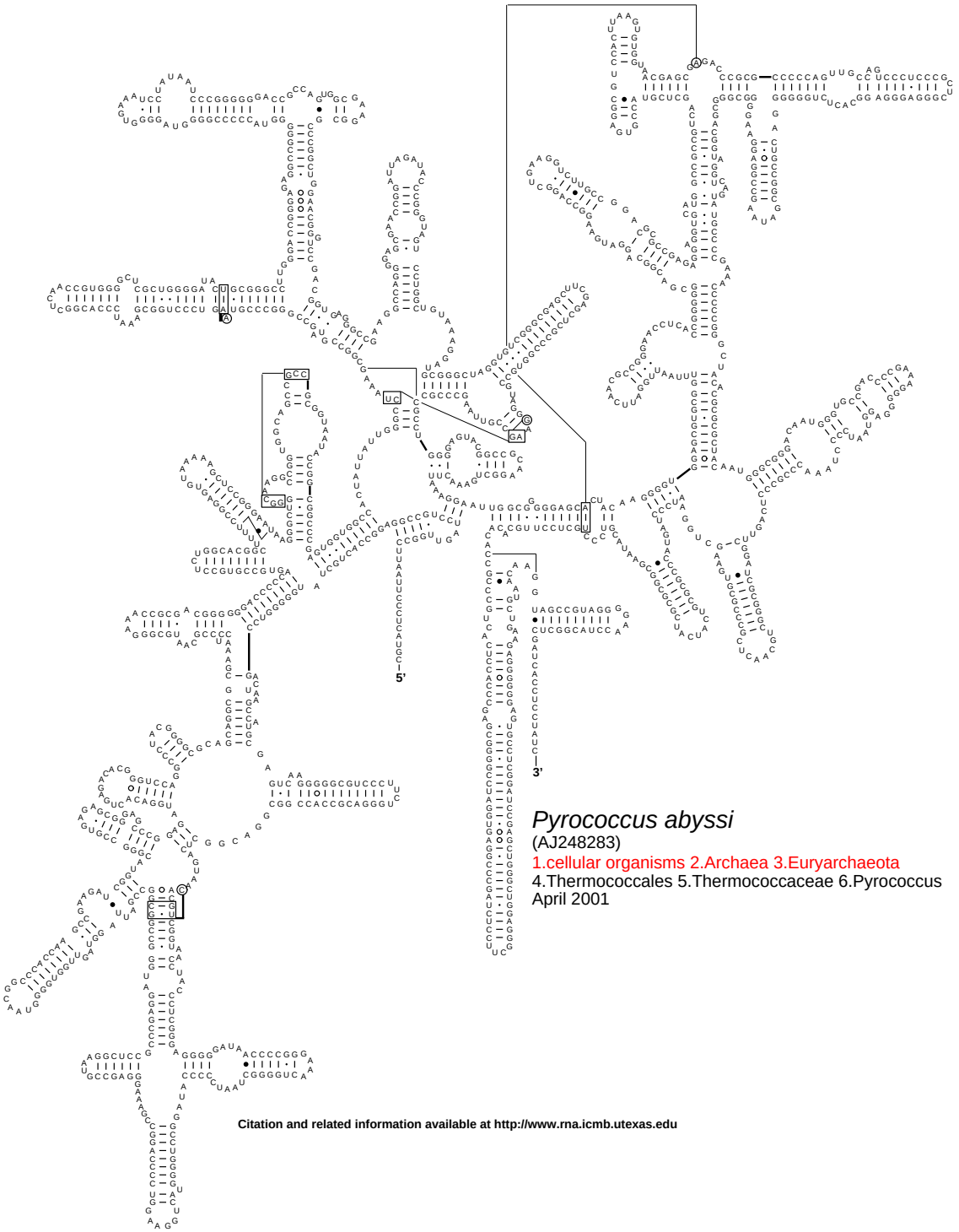

*Nanoarchaeum equitans*
(AJ318041)
1.cellular organisms 2.Archaea 3.Nanoarchaeota
4.Nanoarchaeum
July 2004

# Secondary Structure: small subunit ribosomal RNA



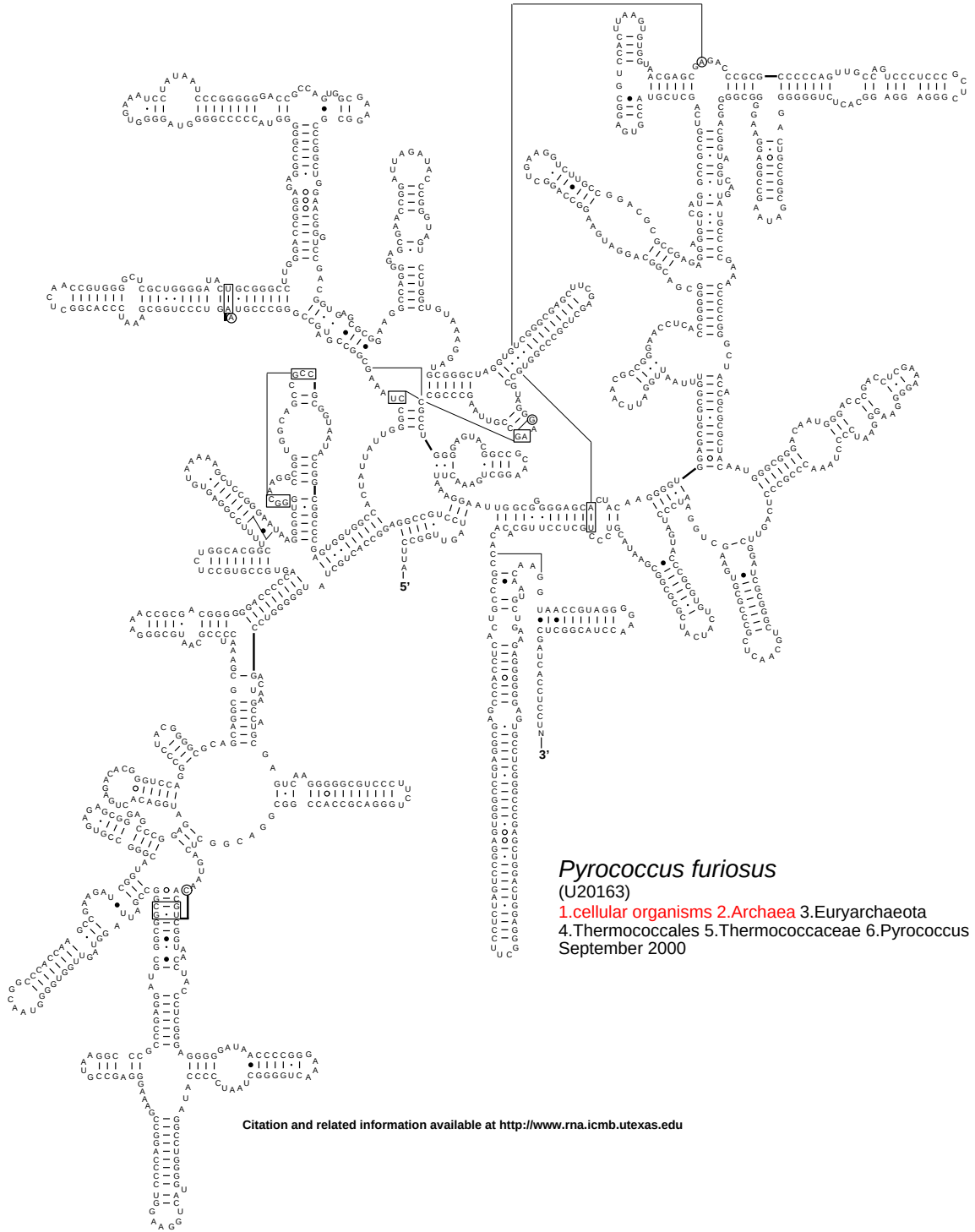

*Natronobacterium innermongoliae*
(AF009601)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronobacterium
February 2001

# Secondary Structure: small subunit ribosomal RNA



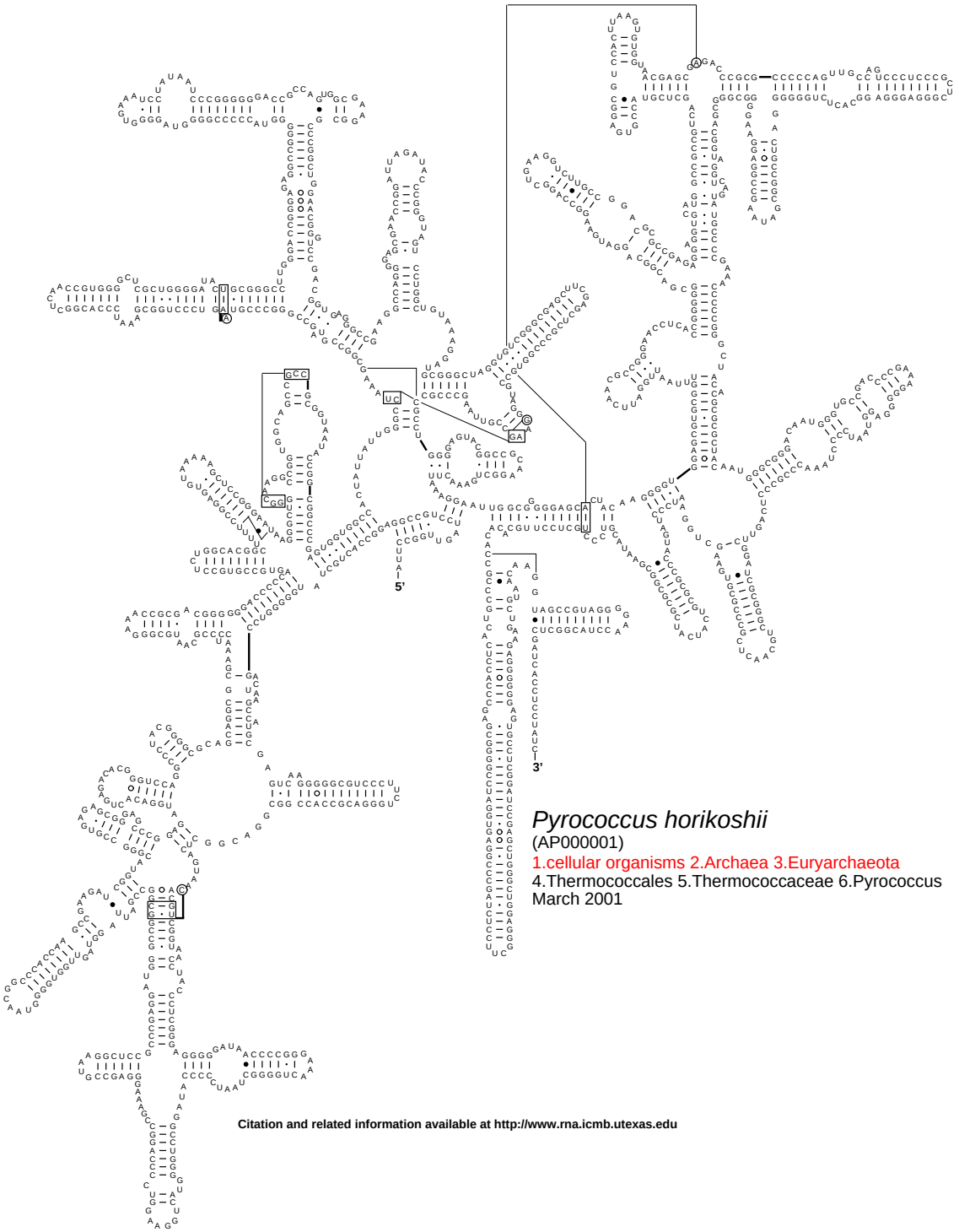

*Pyrococcus abyssi*
(AJ248283)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
April 2001

# Secondary Structure: small subunit ribosomal RNA



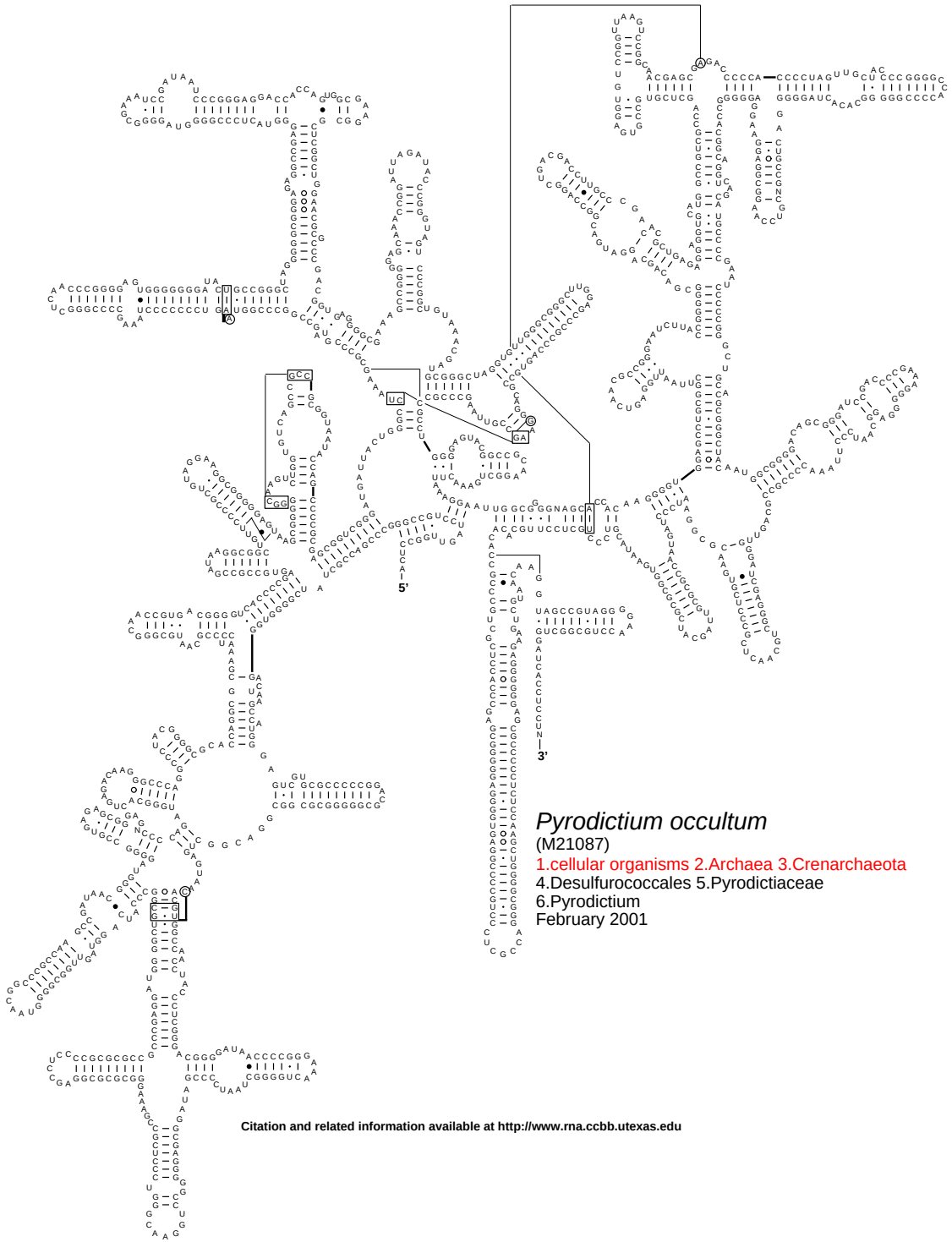

*Pyrococcus furiosus*
(U20163)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
September 2000

# Secondary Structure: small subunit ribosomal RNA



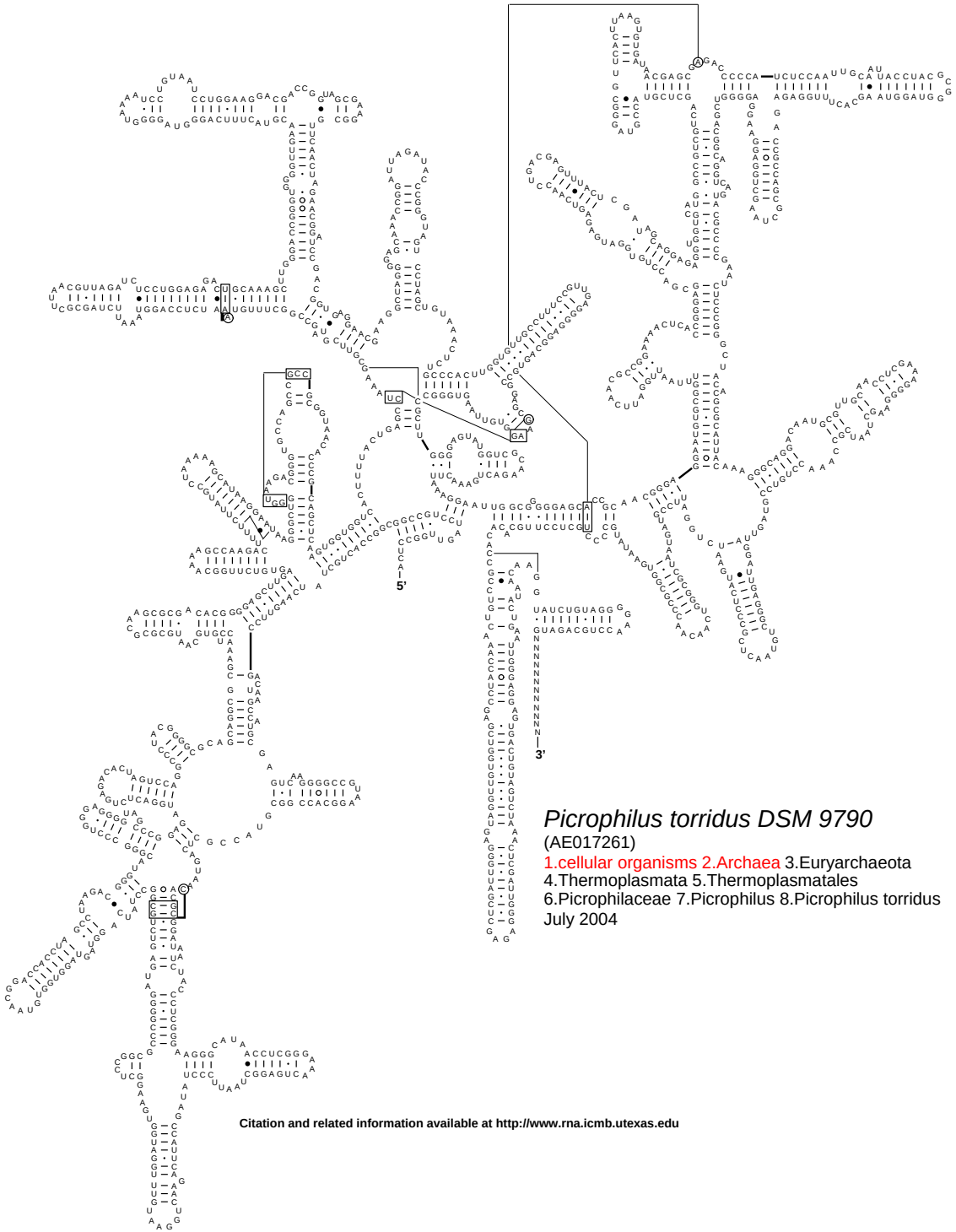

*Pyrococcus horikoshii*
(AP000001)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
March 2001

# 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

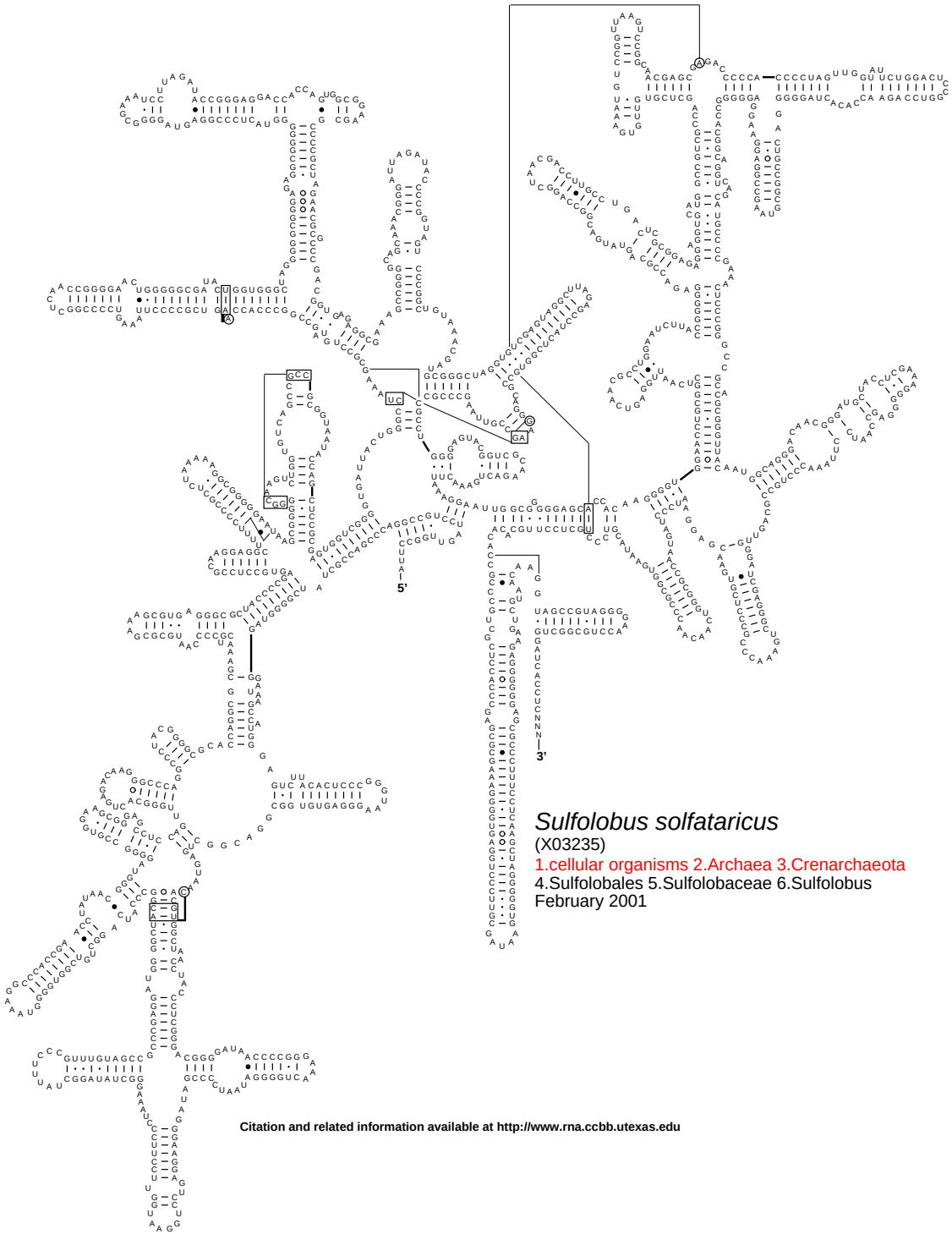



*Picrophilus torridus DSM 9790*

(AE017261)

1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermoplasmata 5.Thermoplasmatales
6.Picrophilaceae 7.Picrophilus 8.Picrophilus torridus
July 2004

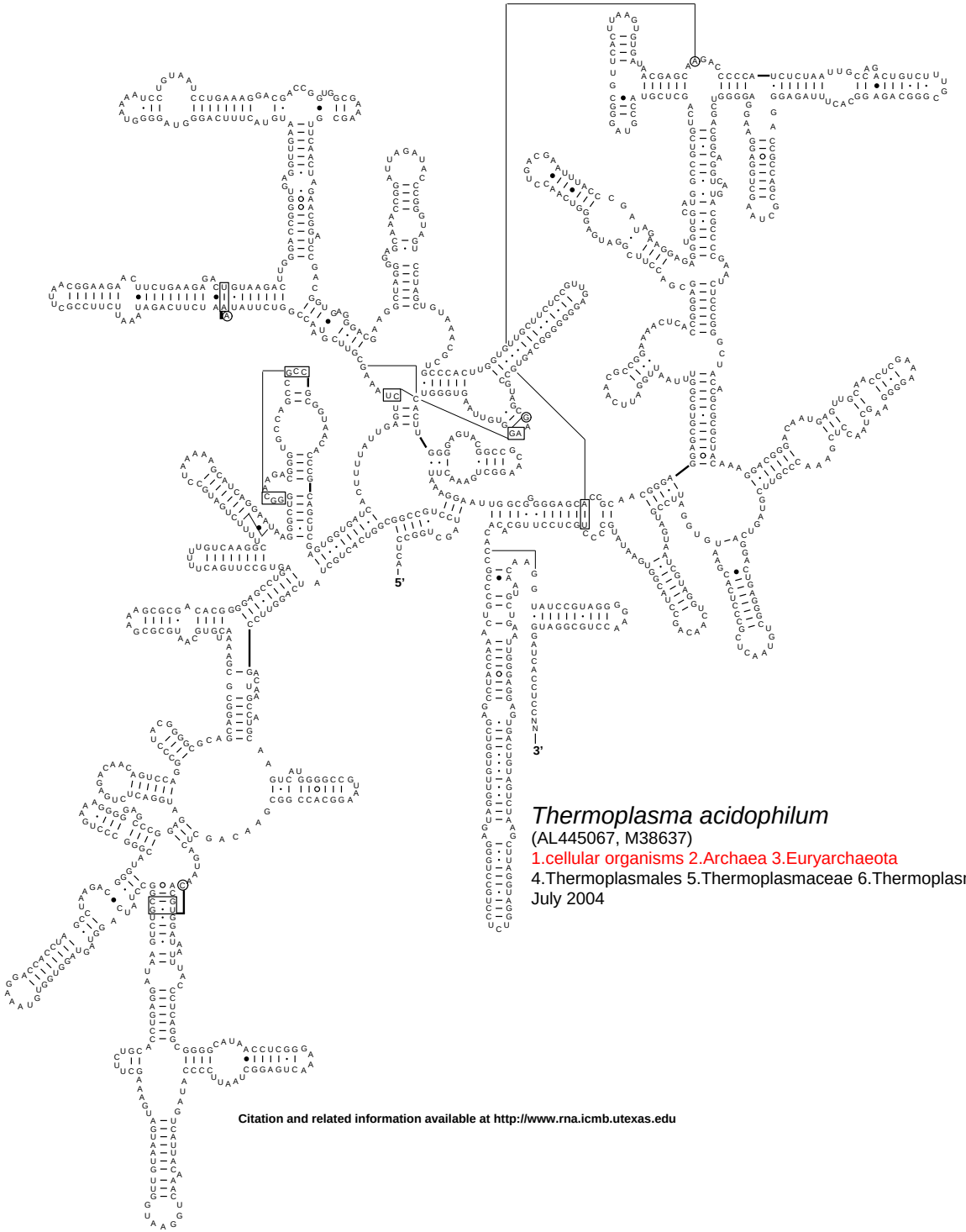
# Secondary Structure: small subunit ribosomal RNA




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

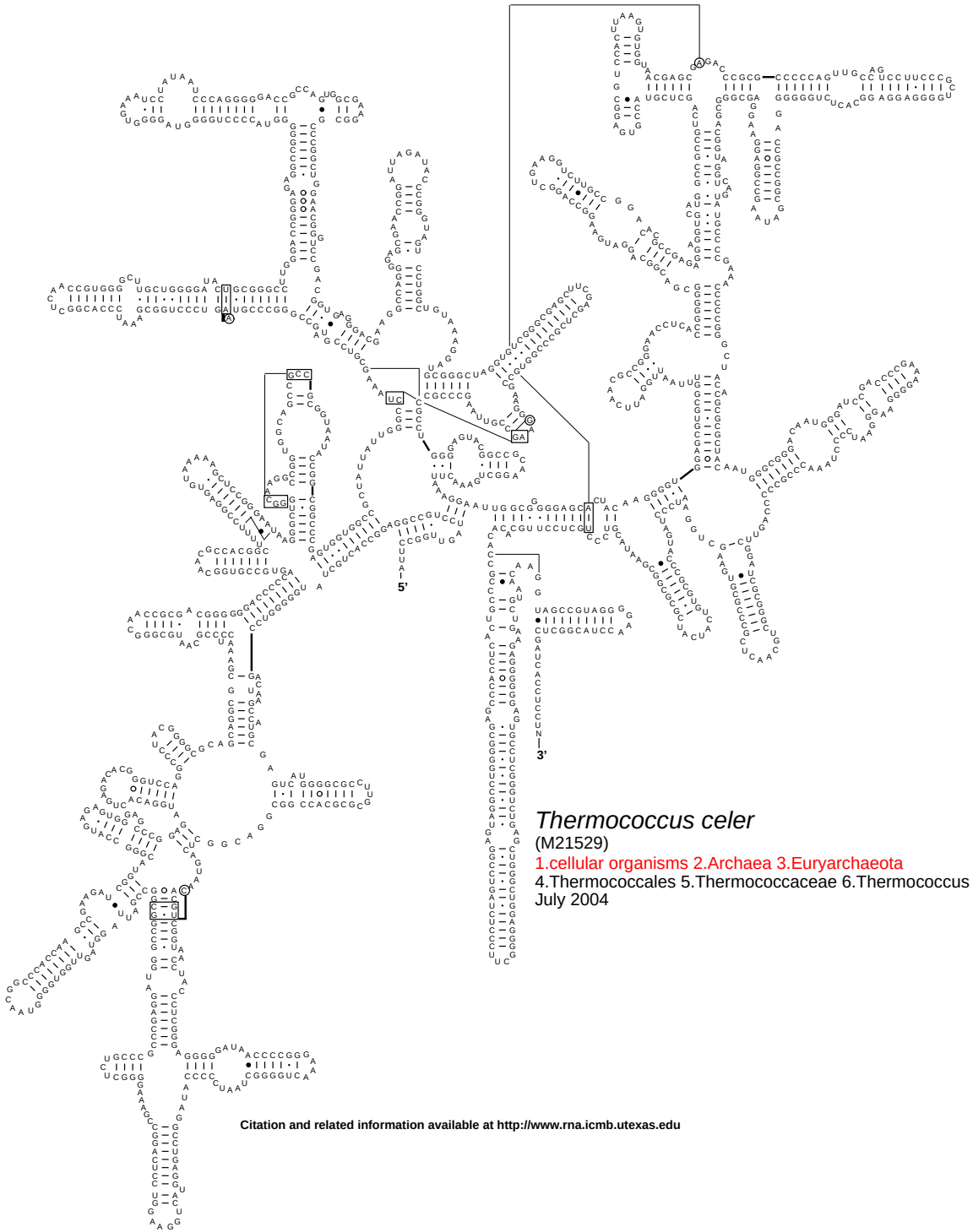
**Citation and related information available at http://www.rna.ccbb.utexas.edu**

# 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



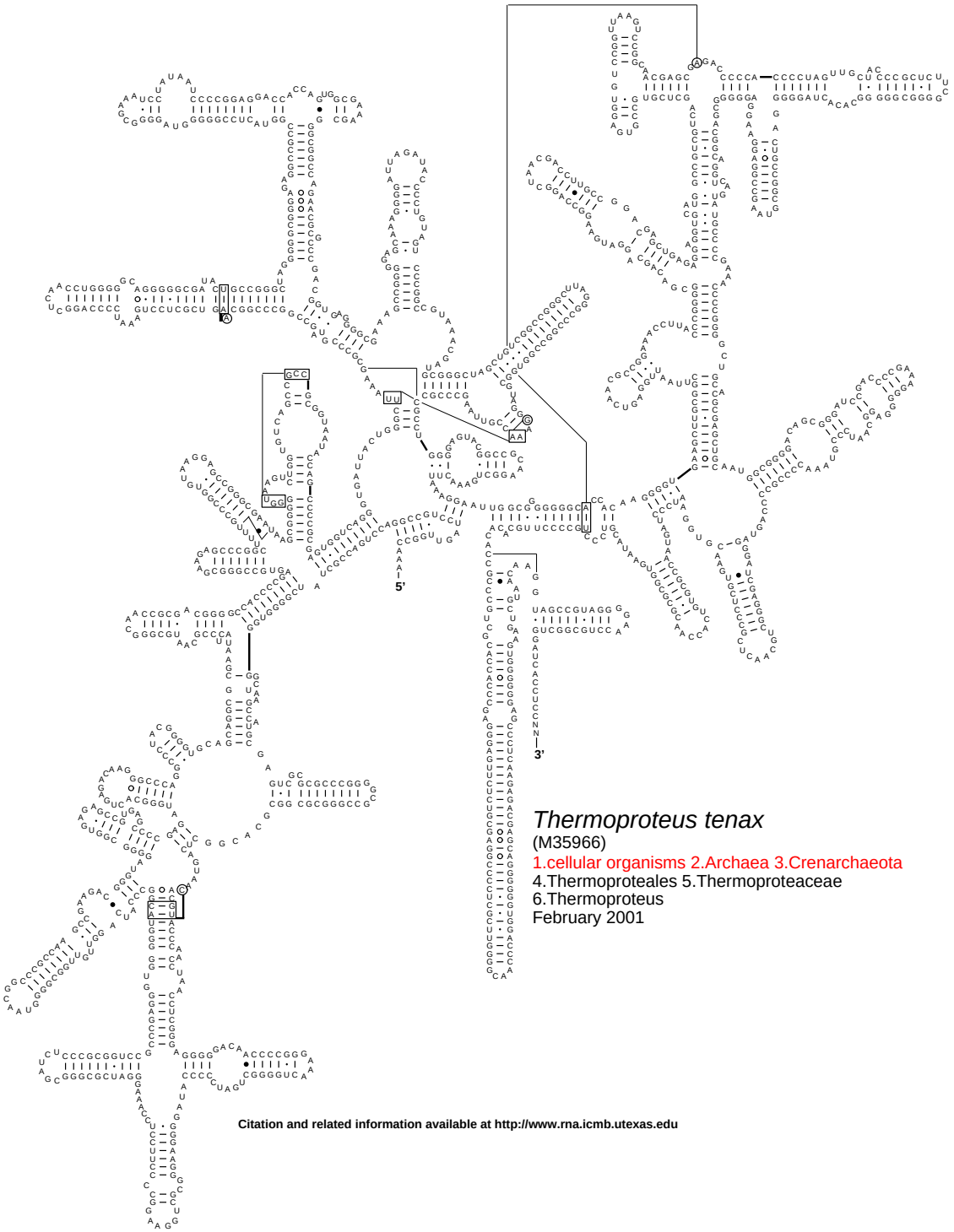

*Thermococcus celer*

(M21529)

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

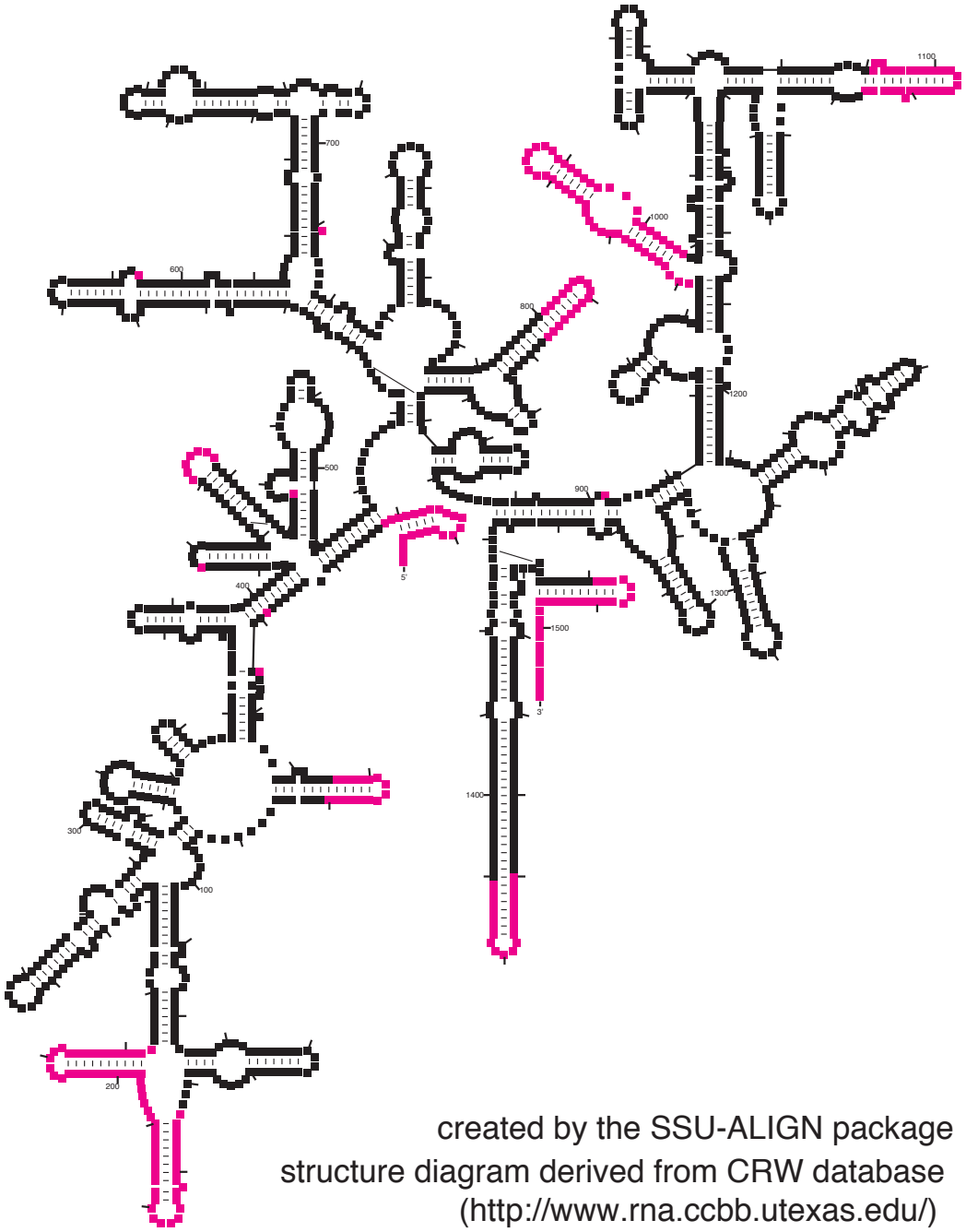
July 2004

# Secondary Structure: small subunit ribosomal RNA



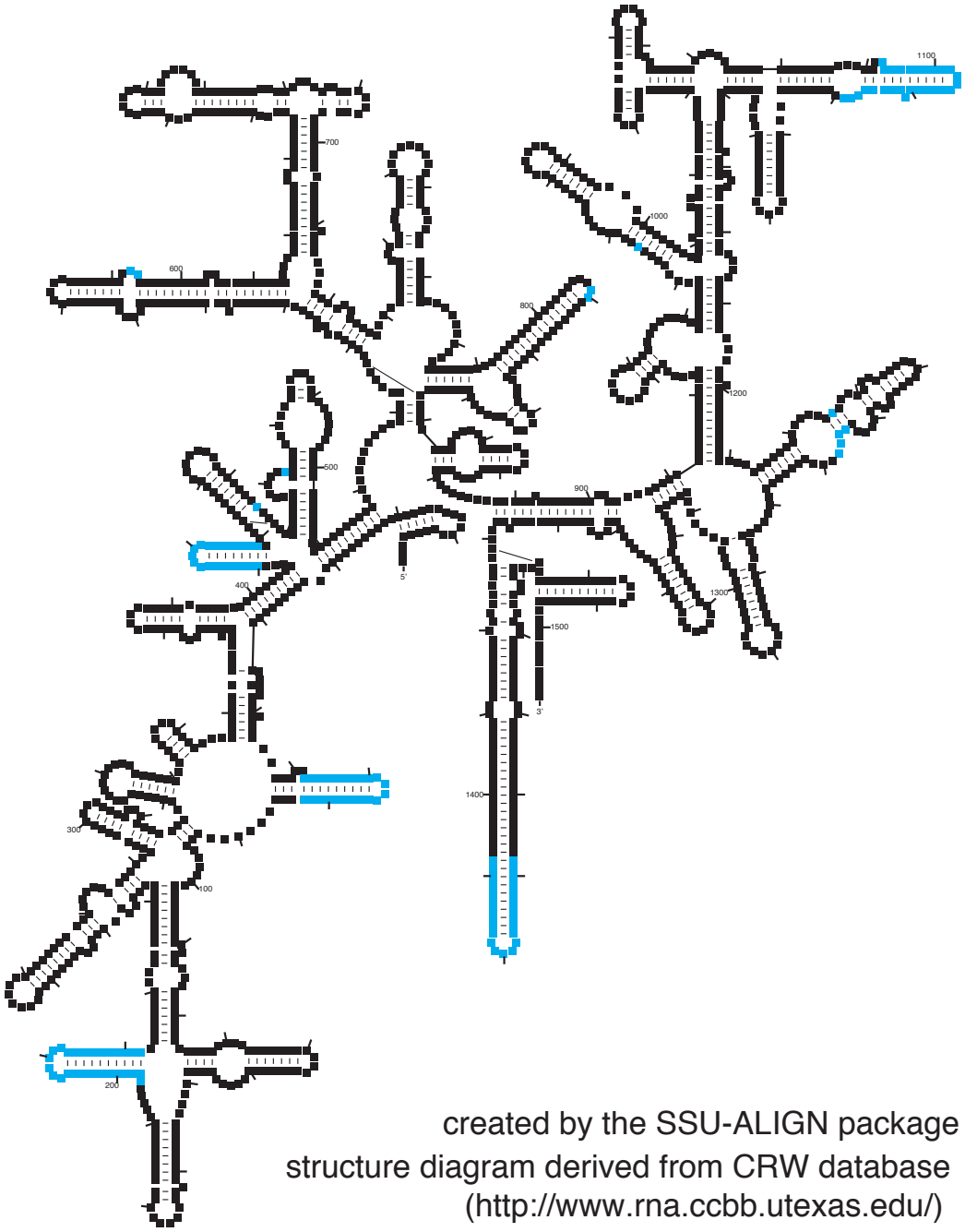

*Thermoproteus tenax*
(M35966)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Thermoproteales 5.Thermoproteaceae
6.Thermoproteus
February 2001

Phil Hugenholtz's manually created mask imposed on archaeal SSU
black: included in alignment (1257)
pink: excluded from alignment (251)



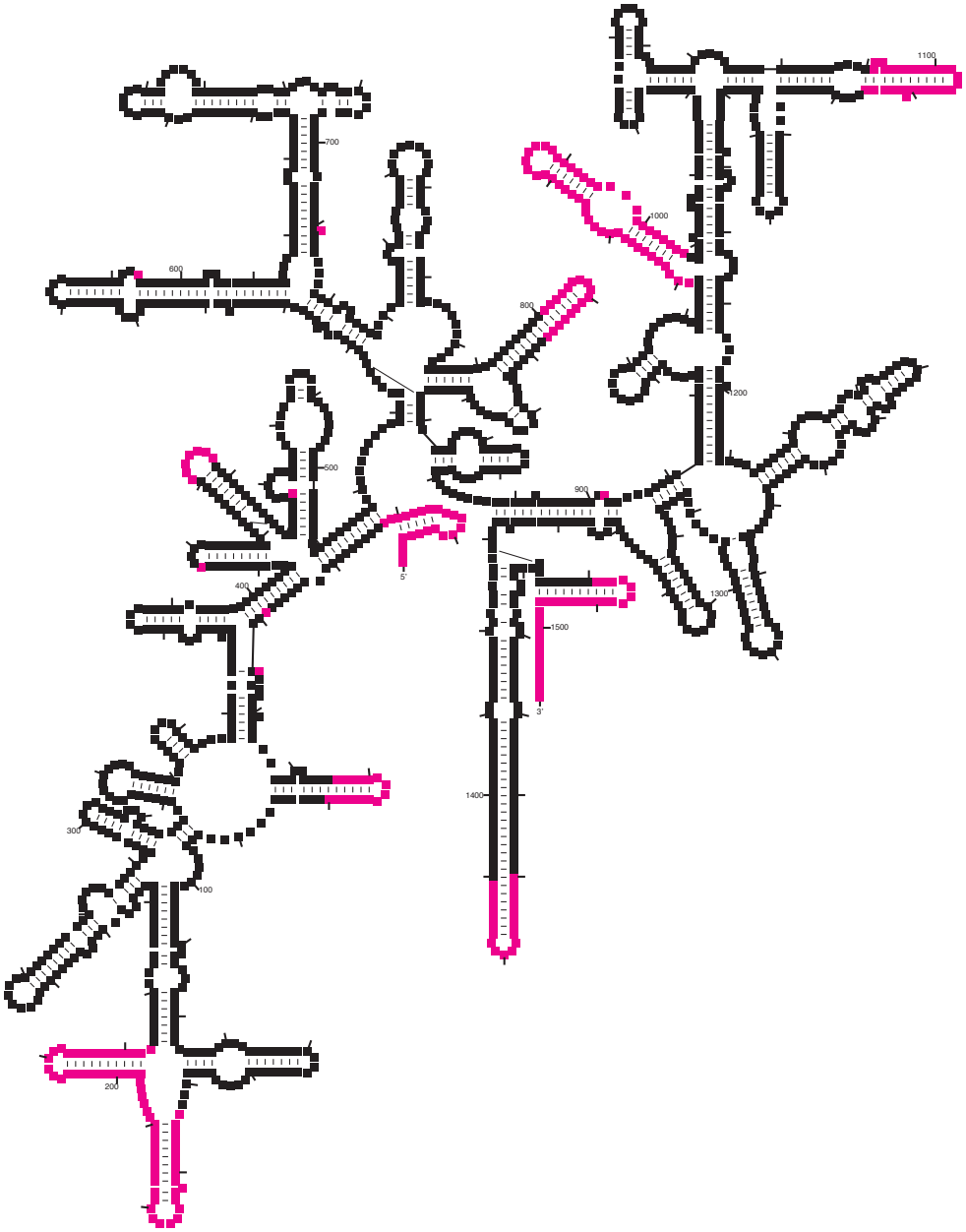

created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Posterior probability based archaeal SSU mask
black: included in alignment (1376)
blue: excluded from alignment (132)

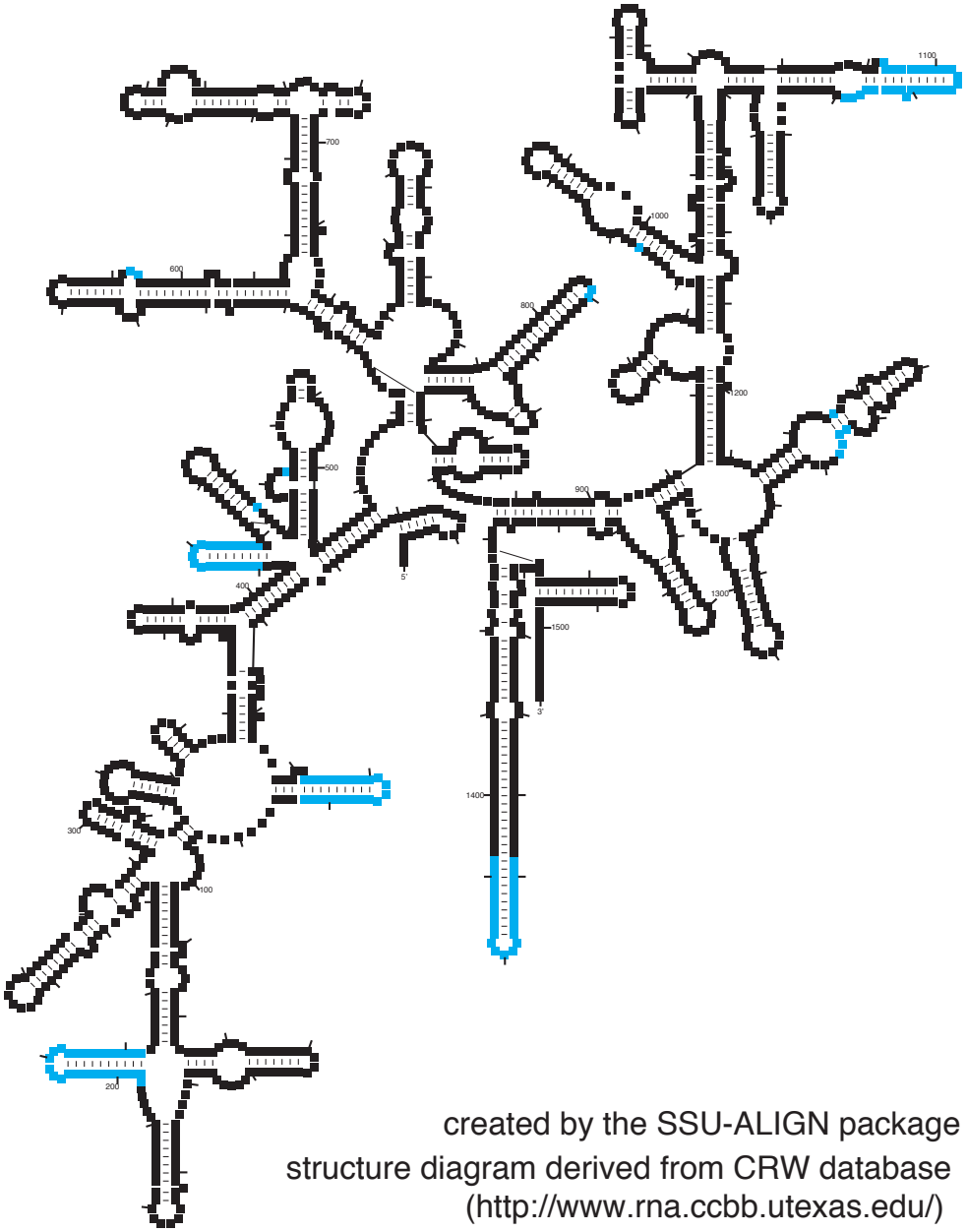



created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Phil Hugenholtz's manually created mask imposed on archaeal SSU
black: included in alignment (1257)
pink: excluded from alignment (251)

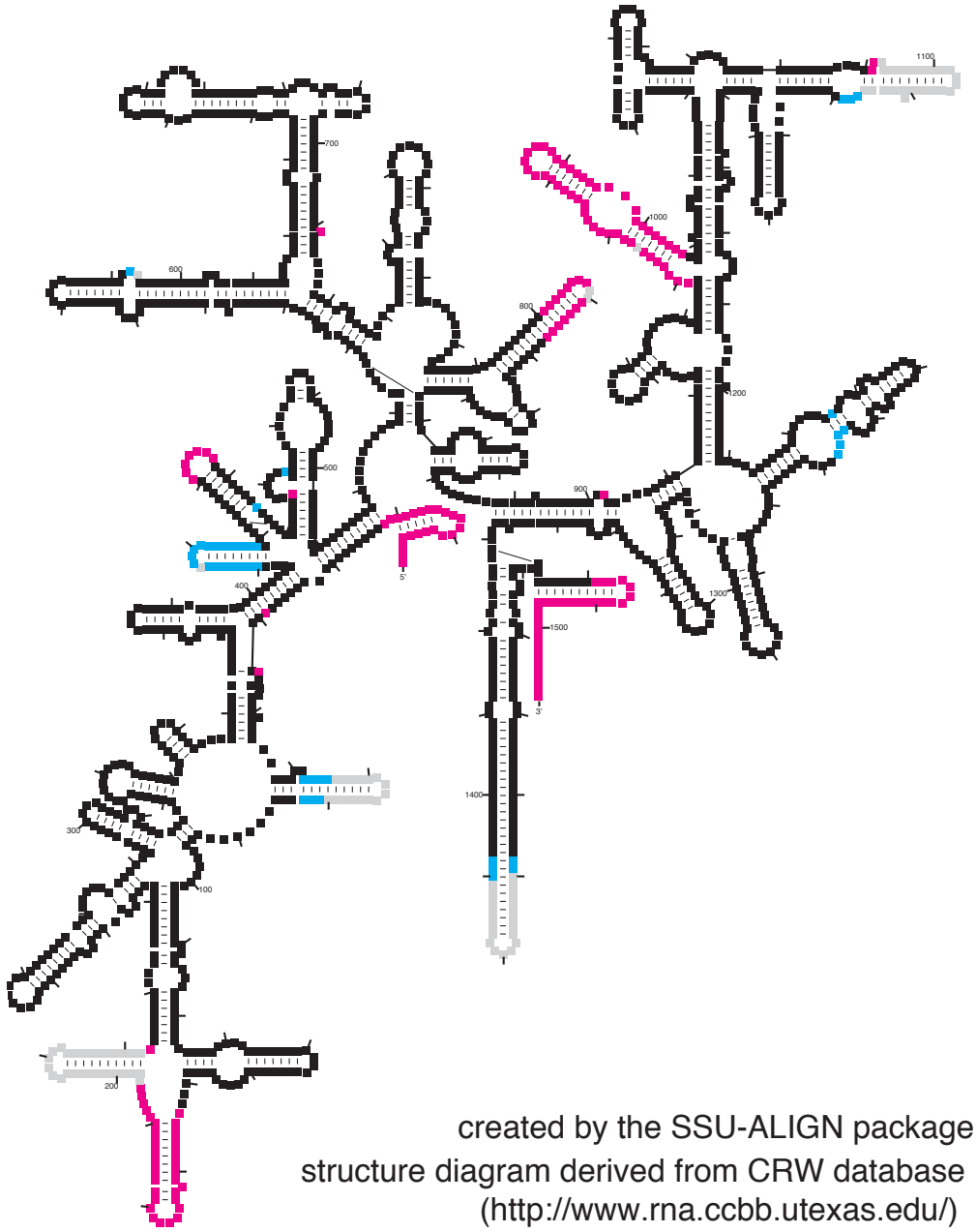


Posterior probability based archaeal SSU mask
black: included in alignment (1376)
blue: excluded from alignment (132)



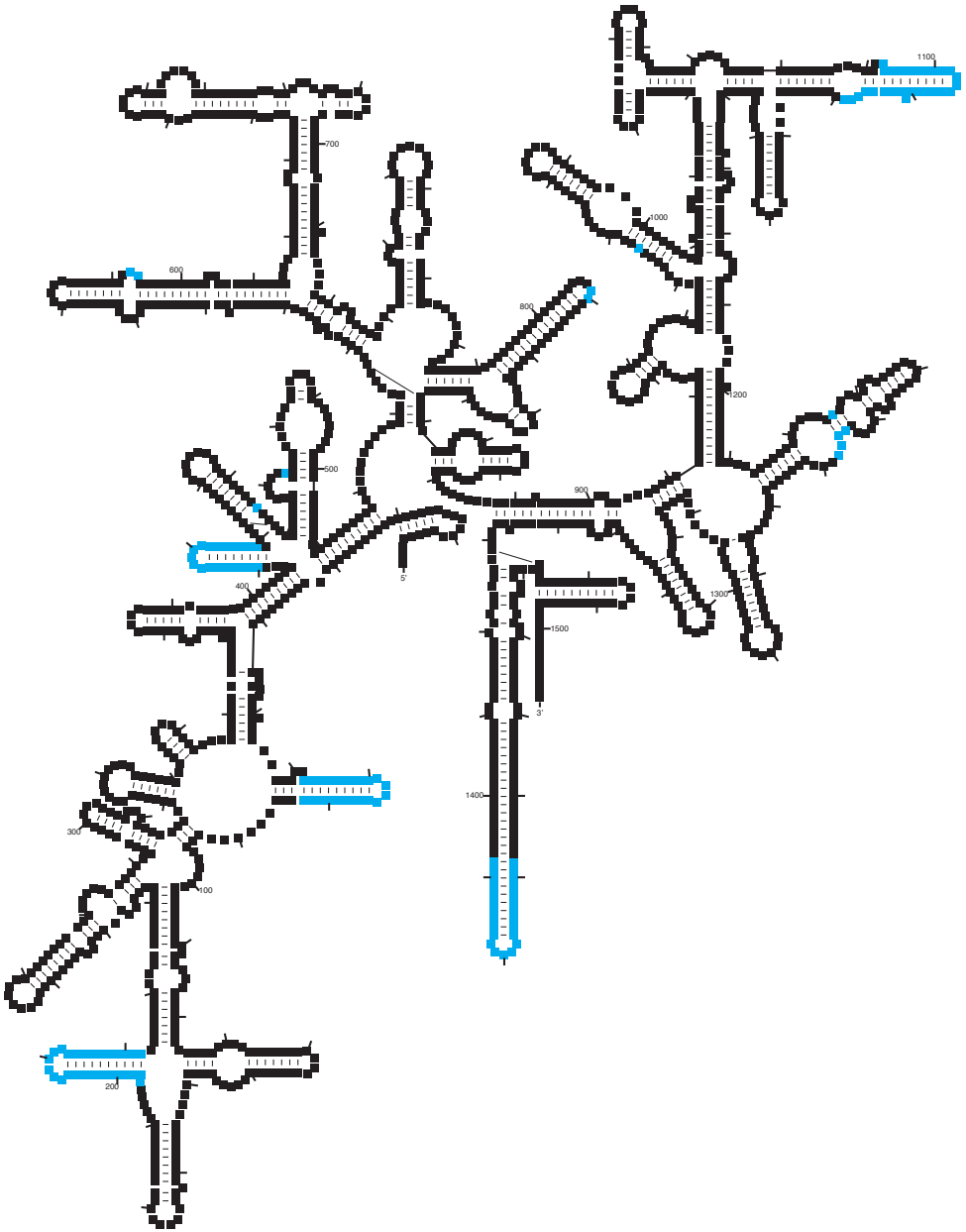
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Comparison of manual and posterior-probability-based masks
black: included in both alignment (1216)
pink: excluded only from manual mask (160)
blue: excluded only from PP mask (41)
grey: excluded from both masks (91)

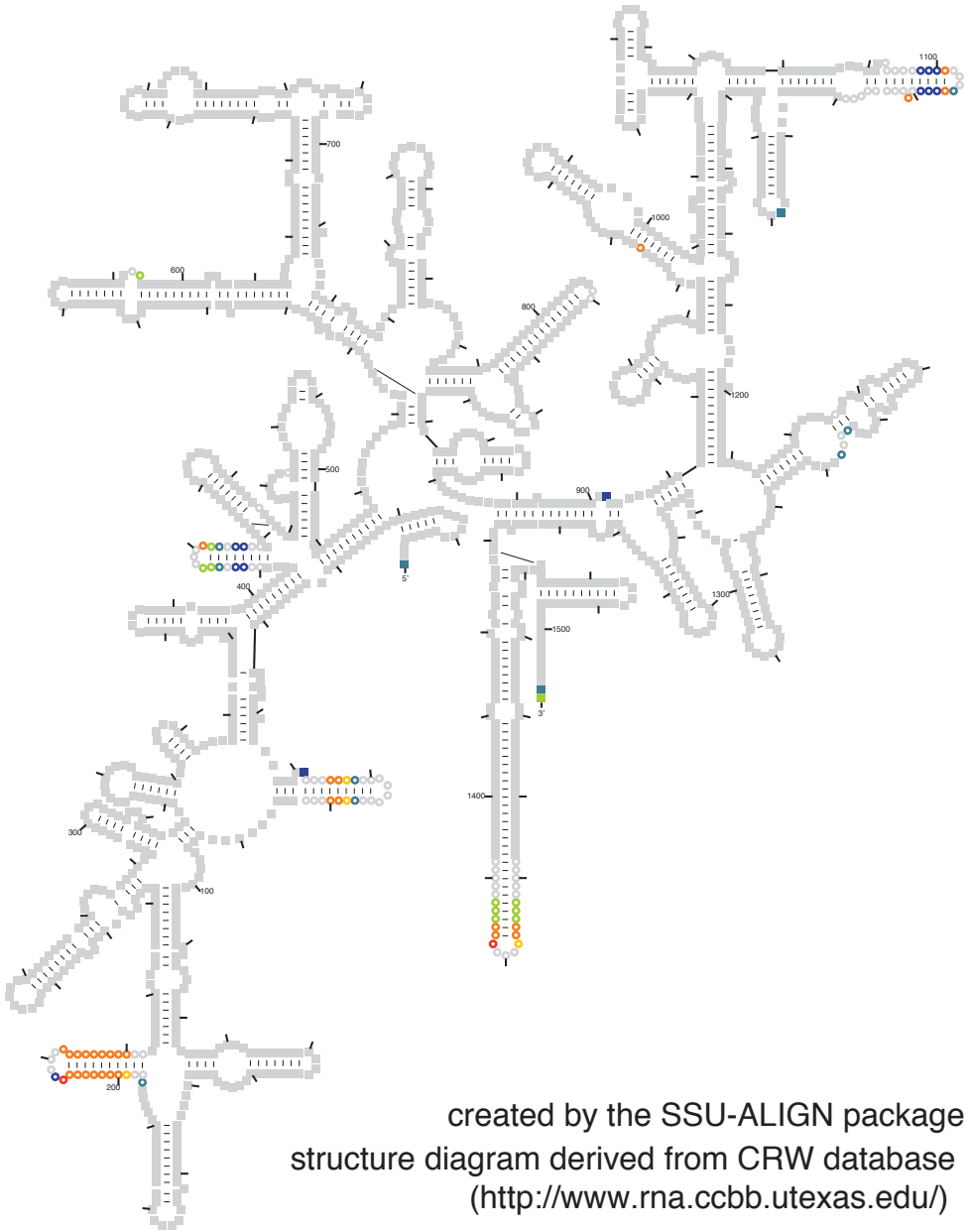



created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Posterior probability based archaeal SSU mask
black: included in alignment (1376)
blue: excluded from alignment (132)




Probability-based mask colored by deletion frequency
grey: zero to very few deletions (1370/1376)
blue: few deletions ..... red: many deletions
circles indicate excluded positions



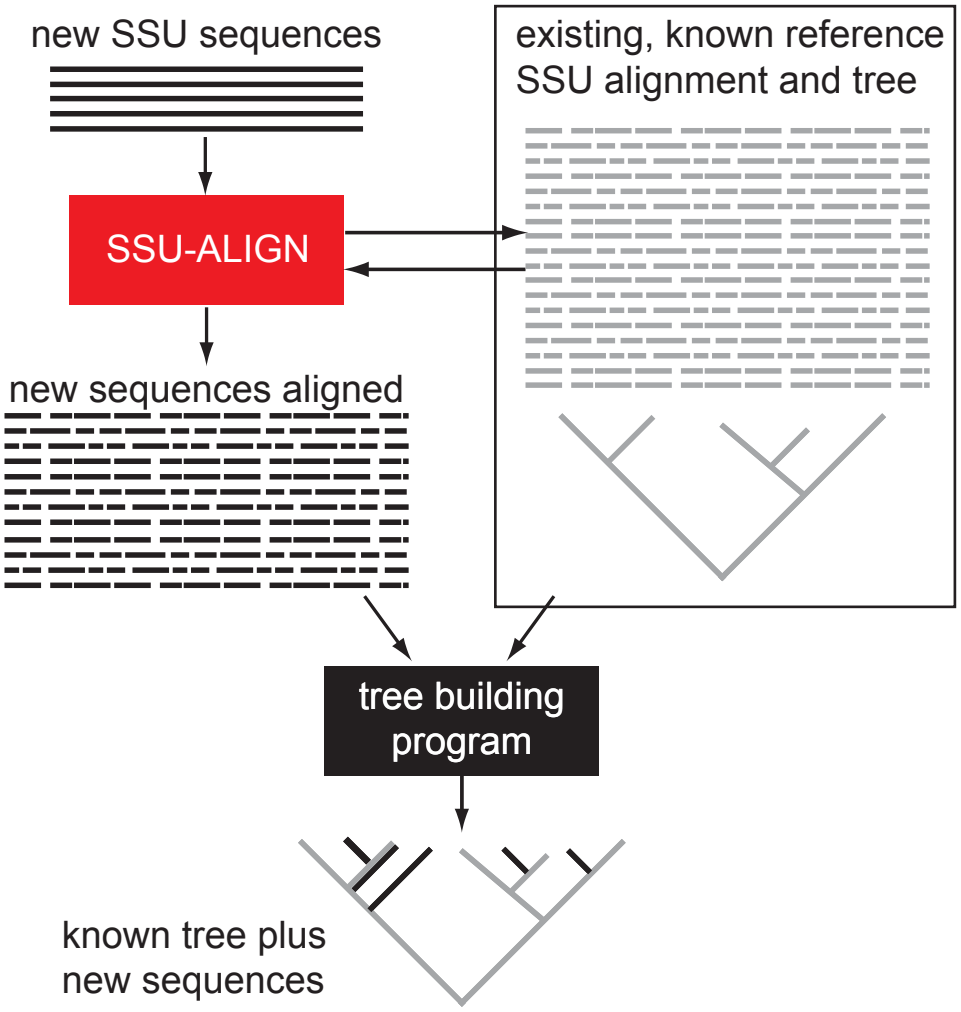

created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

## HMM banding accelerates CM alignment 1000-fold

| | alignment accuracy | time (sec/seq) |
|---|---|---|
| Muscle-3.8.31 (*de novo*) | 95.4% | 0.49 |
| HMMER3 (HMMs) | 96.8% | 0.04 |
| Infernal 1.1 (CMs) | 98.1% | 0.50 |
| Infernal 1.1 (CMs) posterior probability masked (1302/1530 columns) | 99.5% | 0.50 |

**Infernal produces alignment that are very similar to manually refined alignments.**

# SSU-ALIGN: structural alignment of SSU rRNAs using CMs




**Goals of the alignment program:**

- accurate: because alignment errors confound phylogenetic inference
- scalable to handle up to millions of seqs and fast:

**SSU-ALIGN**

Includes Infernal CMs for archaeal, bacterial and eukaryotic SSU rRNA

accurate:

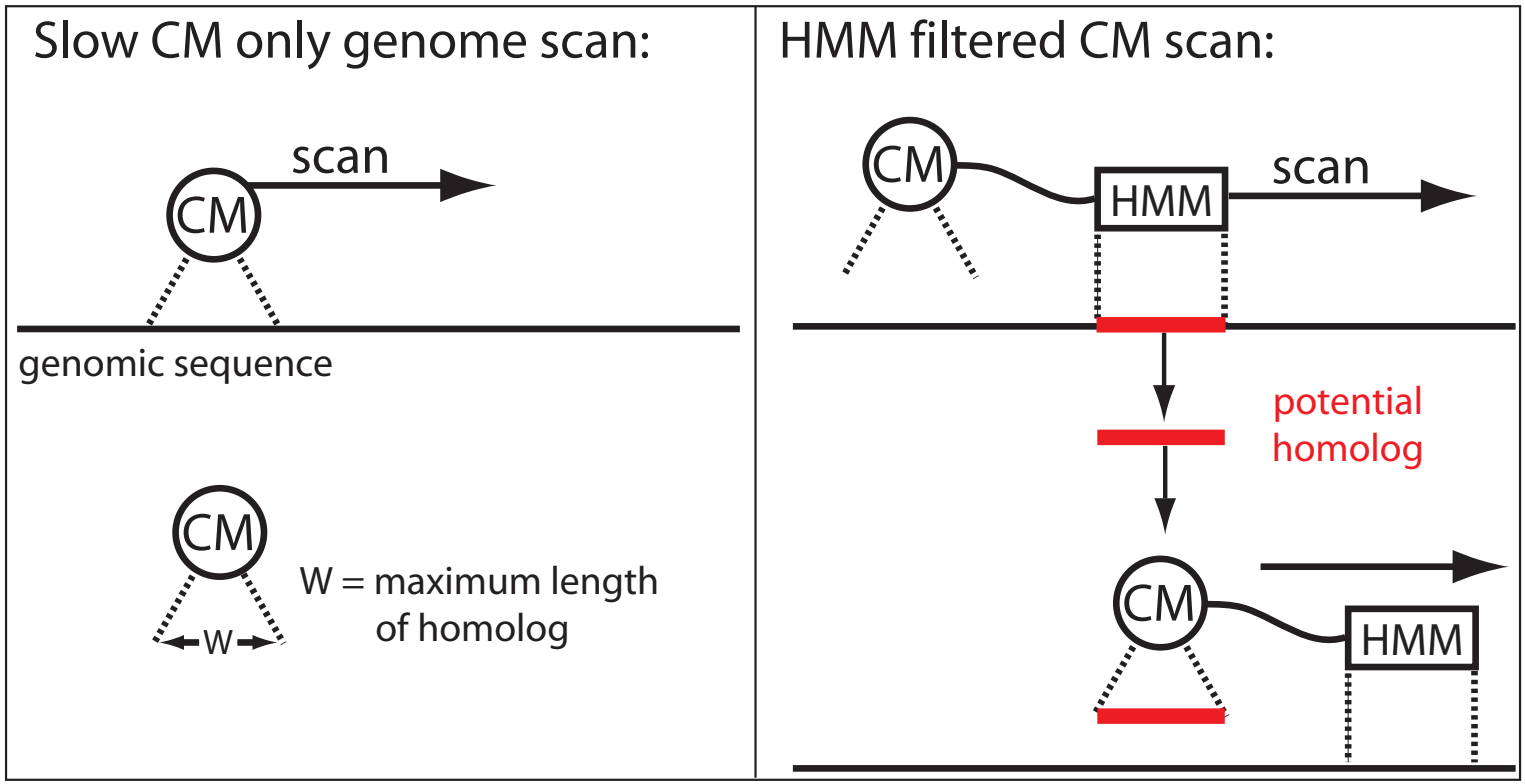
- structural alignment of sequences
- probabilistic masking of ambiguous columns

scalable and fast:

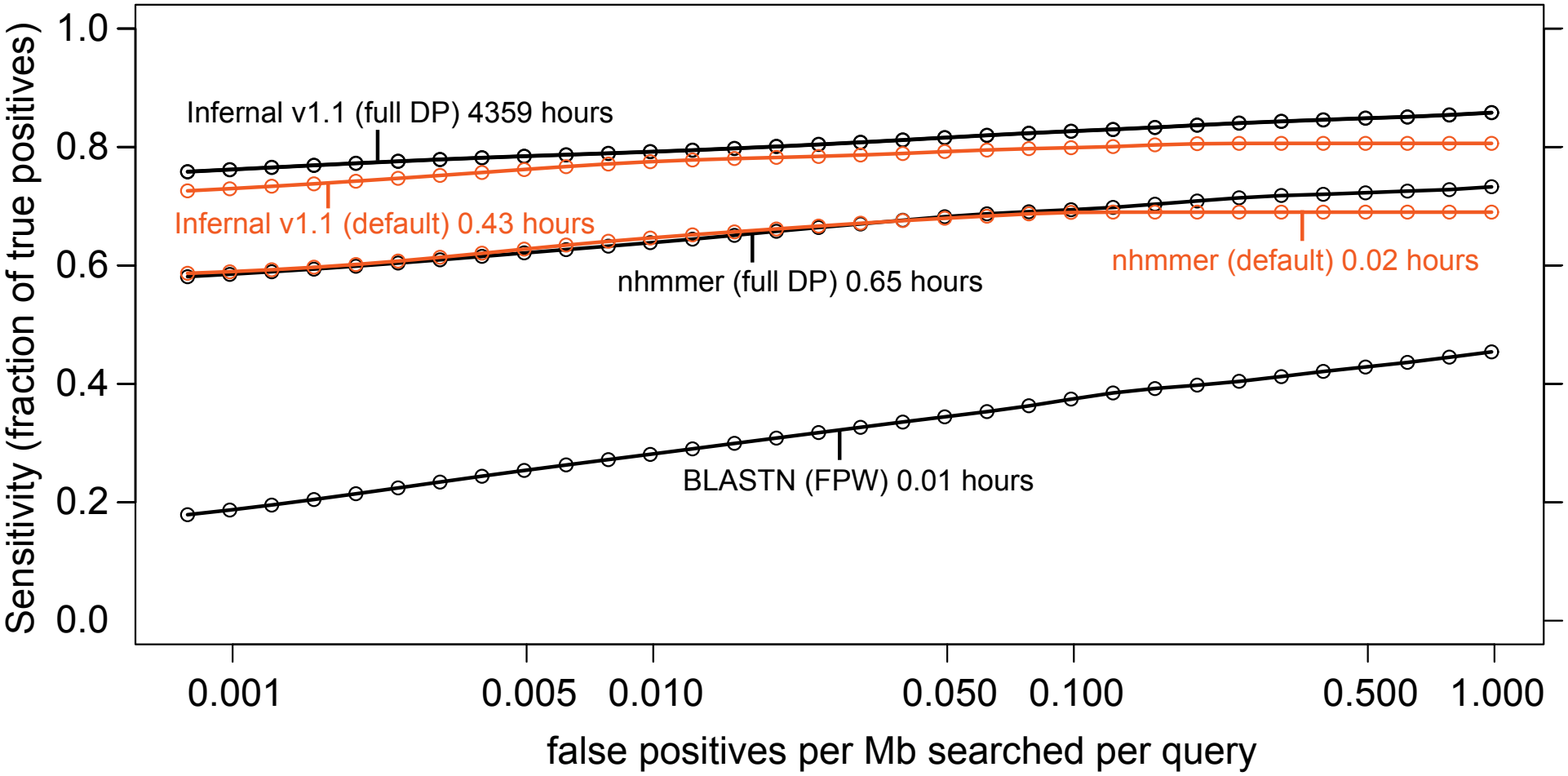
- can generate alignment of millions of seqs
- speed is about 1 second/full length sequence
- easily parallelized on clusters

SSU-ALIGN tutorial tomorrow at 10:30

# HMM banding can be used for homology search too



# Infernal 1.1: RNA homology searches 10,000 times faster



# Acknowledgements

Sean Eddy
Elena Rivas
Travis Wheeler
Tom Jones
Diana Kolbe
Seolkyoung Jung
Rob Finn
Jody Clements
Fred Davis
Lee Henry
Michael Farrar