Multiple alignment using sequence family profiles

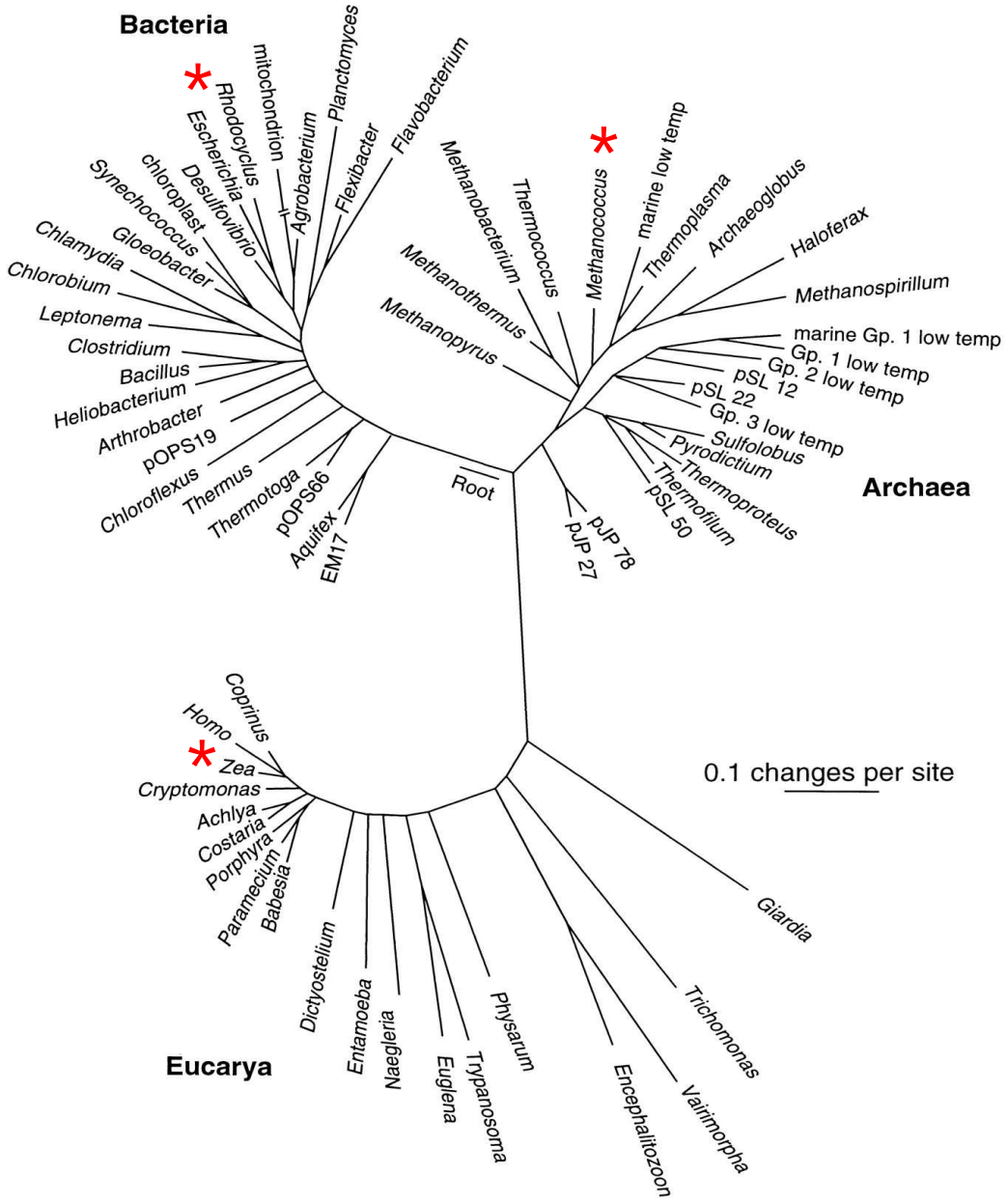
Eric Nawrocki

Sean Eddy's Lab

Howard Hughes Medical Institute
Janelia Farm Research Campus

Small subunit ribosomal RNA and the tree of life

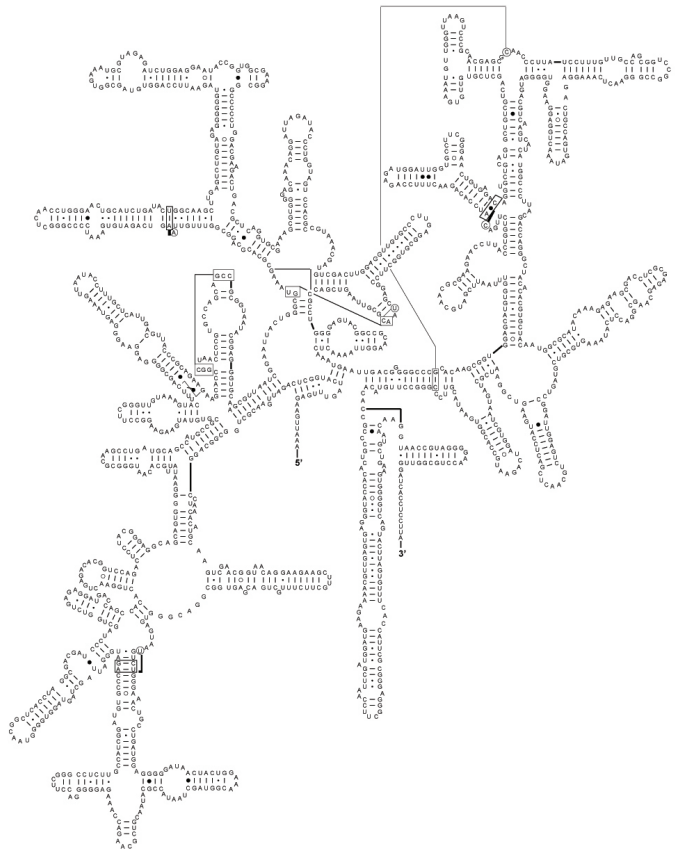
- 1977 - Carl Woese decided to classify all living things phylogenetically
- needed "*a molecule of appropriately broad distribution*" for comparative analysis
- SSU rRNA was chosen
  - universally distributed
  - highly conserved
  - large enough to provide sufficient data (1500-1800 nt)
  - readily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



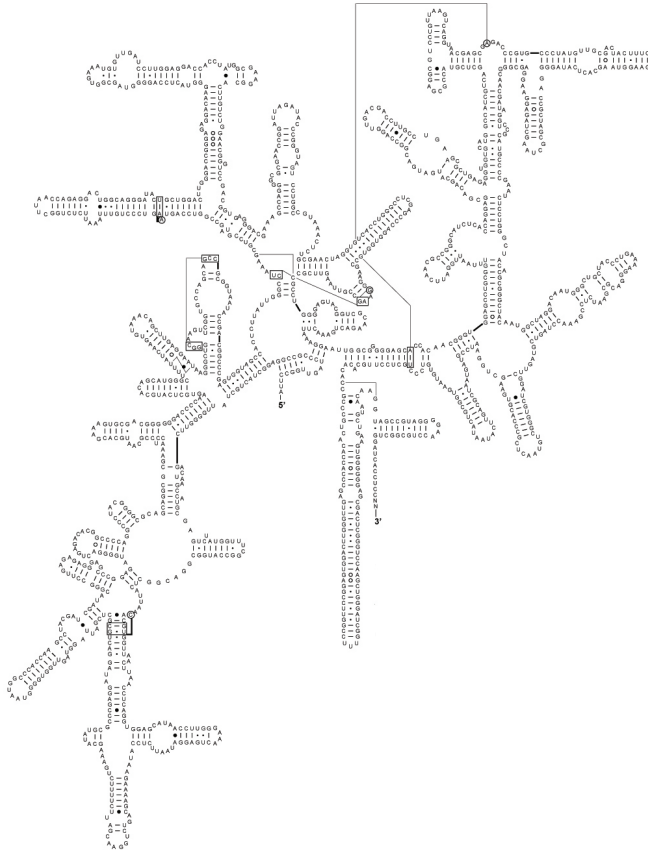
Pace, Science:276; 1997

Universal conservation of SSU rRNA

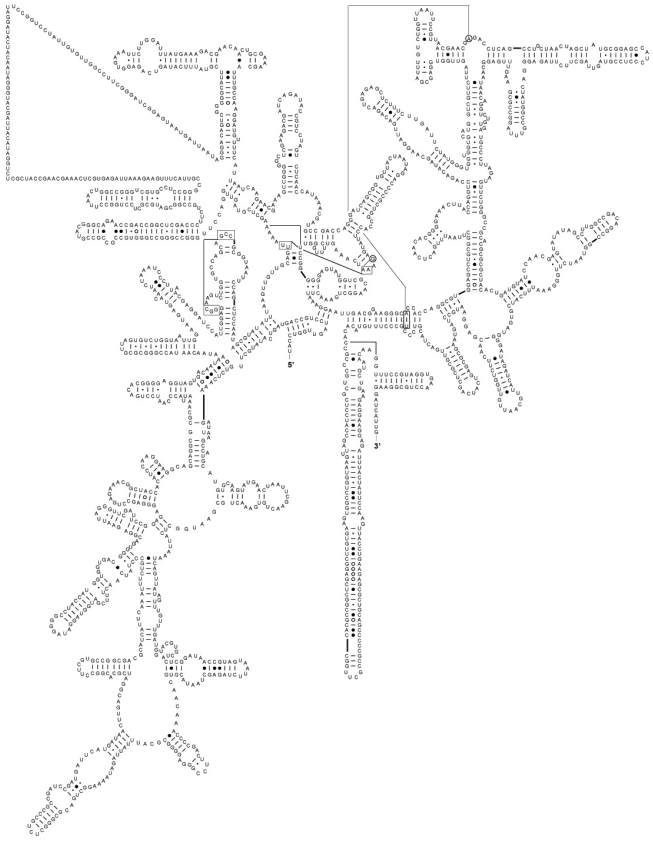
Escherichia coli



Methanococcus vannielii



Zea mays



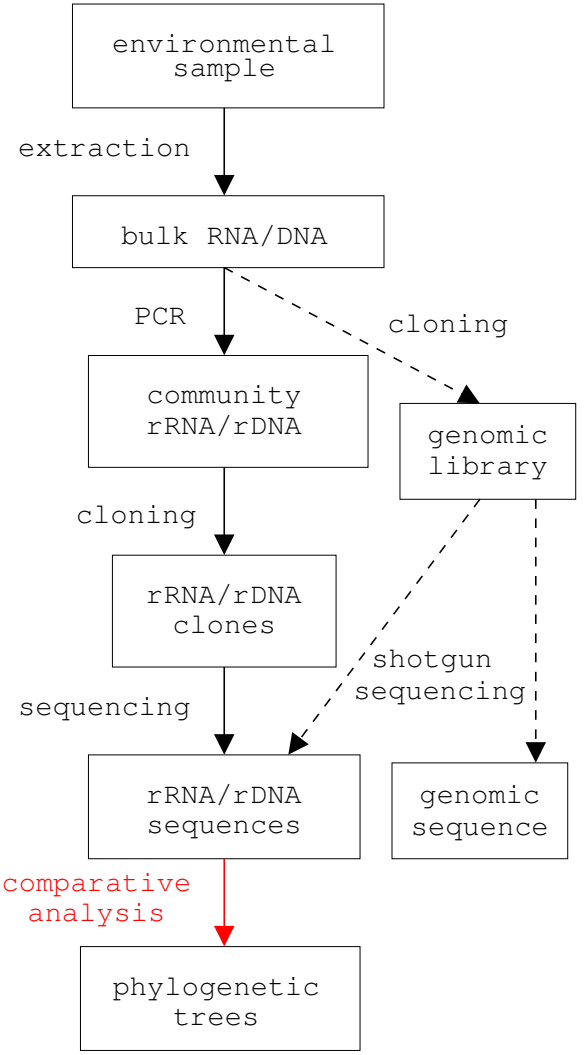
Secondary structure diagrams from:
URL:http://www.rna.ccbb.utexas.edu/

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:

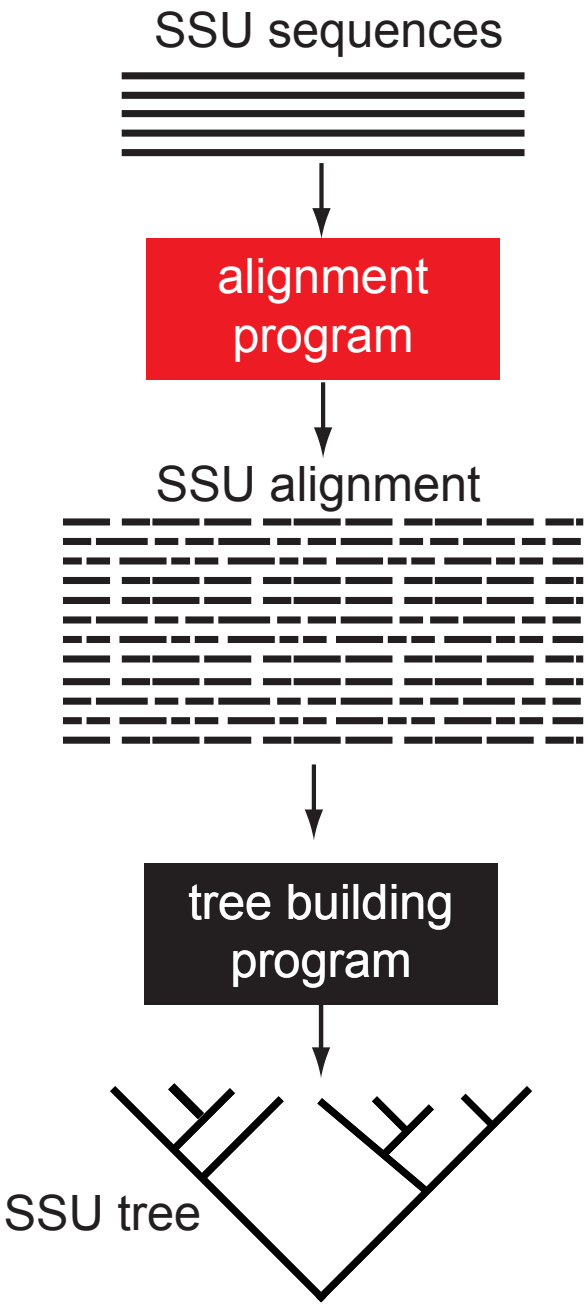
name	# seqs	# citations
Silva	3.2M	1125
RDP	2.6M	1170
Greengenes	1.0M	1012

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



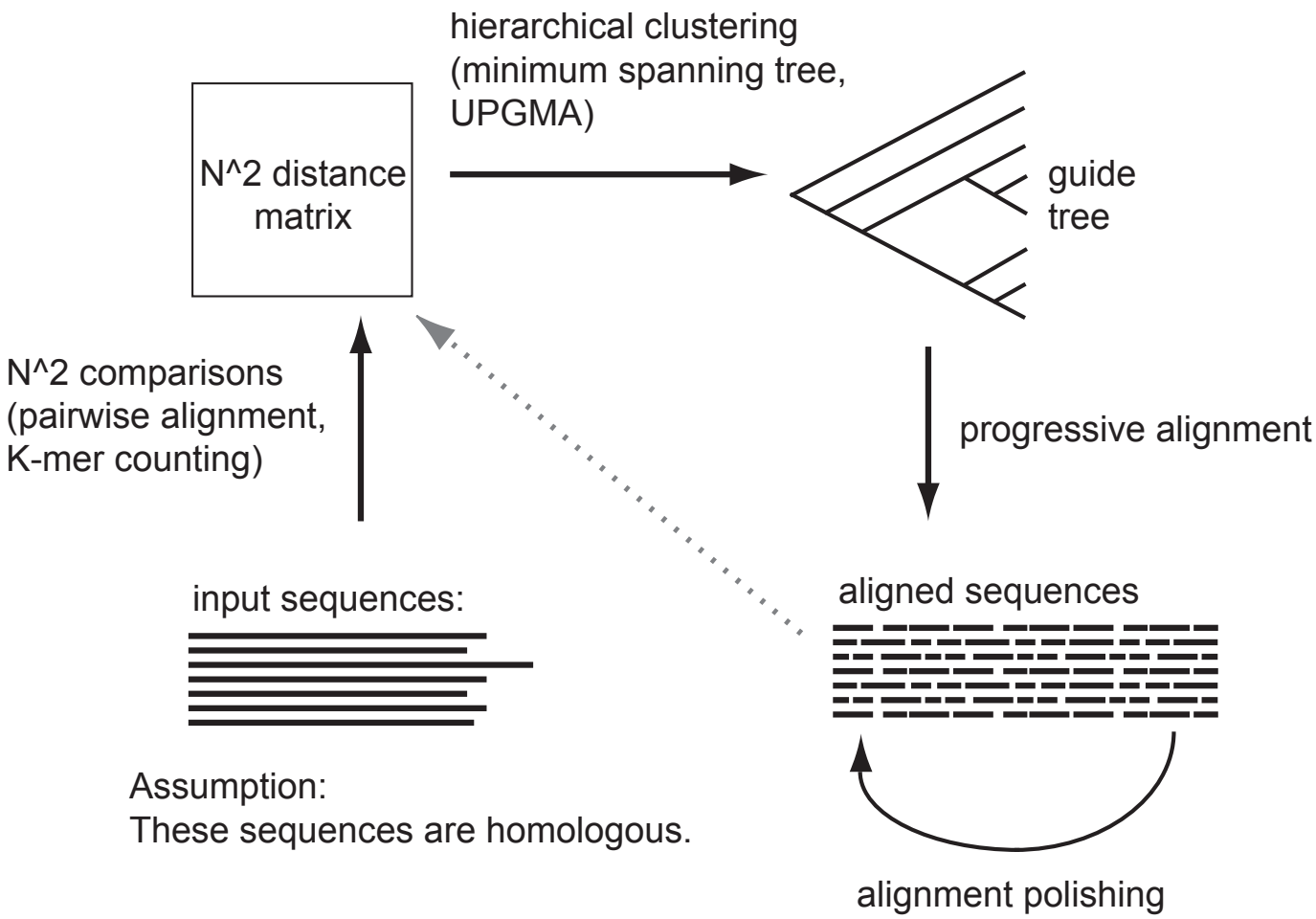
Goals of the alignment program:

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

Information that can help:

- many known sequences
- manually curated SSU alignments
- known secondary structure of SSU

De novo multiple sequence alignment methods assume very little

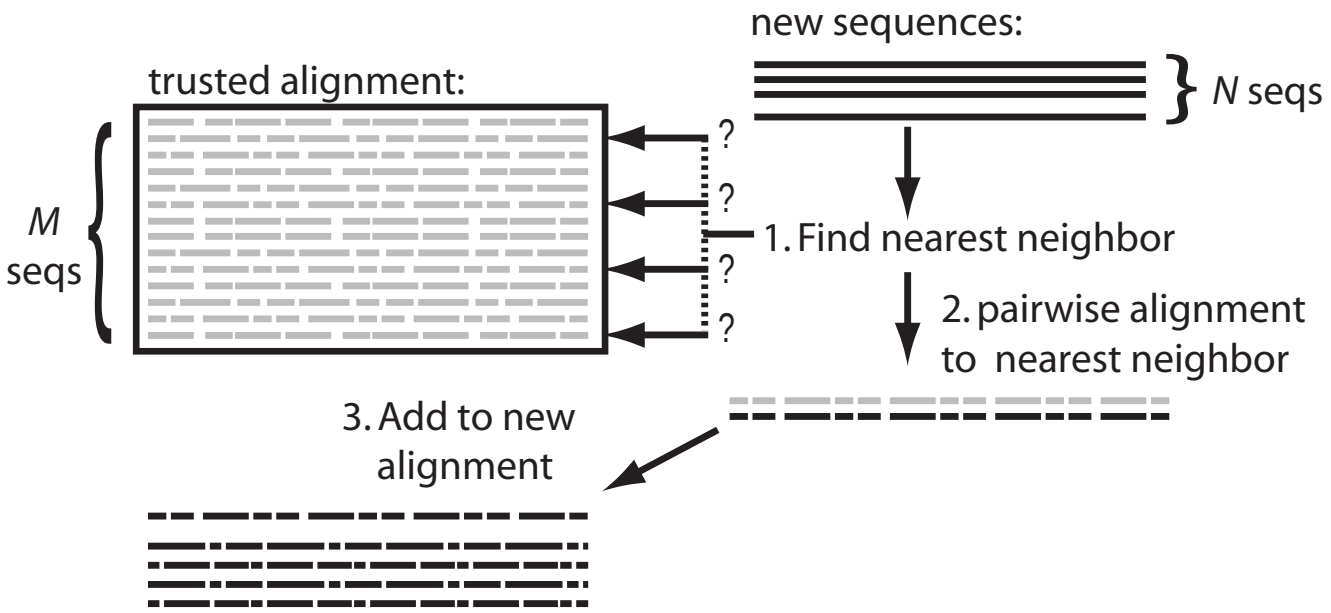


Drawbacks for SSU alignment:

1. Won't scale to millions of seqs (O(N^2))
2. Ignores previous knowledge of SSU

A trusted (probably manually curated) reference alignment can help

"nearest neighbor" alignment:



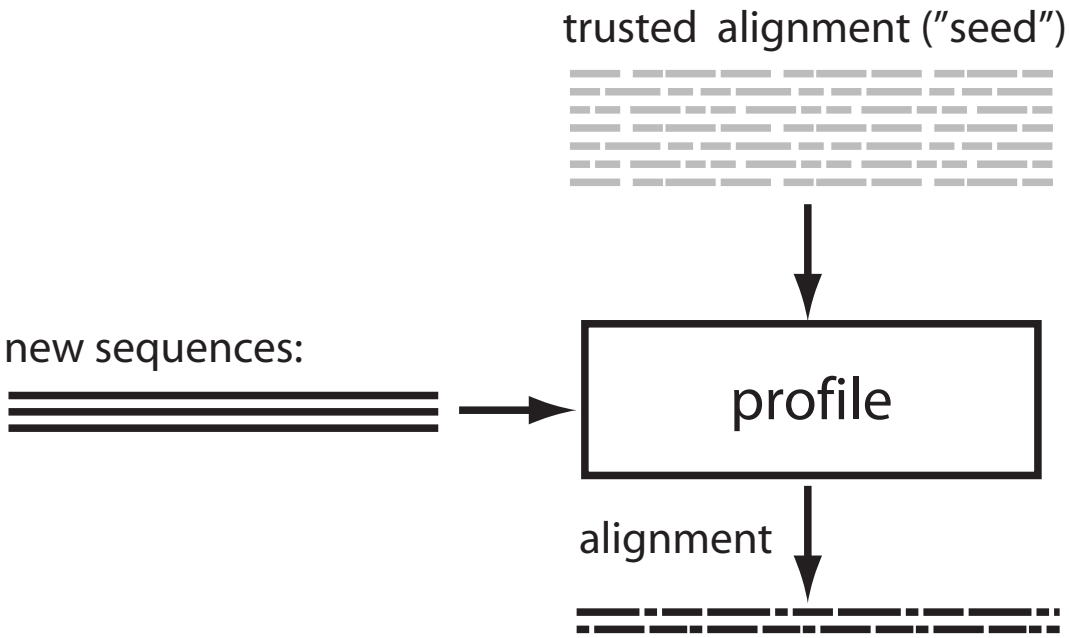
Advantages over *de novo*:

1. More scalable (O(MN))
2. Uses existing, trusted SSU alignment

Drawbacks:

1. Still slow if M is large (~5000 for Greengenes)
2. Pairwise alignment ignores varying conservation across alignment

Profile-based alignment is $O(N)$



Advantages over *de novo* and nearest-neighbor:

1. Scalable (O(N))
2. Uses existing, trusted SSU alignment
3. Uses position-specific scores

Drawbacks:

1. Only *consensus* positions are aligned, other nucleotides are *inserted*
2. Ignorant of phylogeny

Profiles have position-specific scores (substitutions, gap open, gap extend)

sequence profile

```
yeast     GUCaUUCGGC...AC
fly       GCC.UU-GGA...GC
cow       GCA.UUCGUC...-C
mouse     GCA.UU-GAU...GC
human     GCGaUUCGCU...GC
chicken   GUA.UUCGUA...AC
snake     GUGaUUCGCG...AC
croc      GUU.UU-GAG...AC
worm      G-G.UUCGCGccaAC
starfish  G-U.UUCGAU...-C
```

Profiles HMMs are probabilistic profiles built from alignments


```
yeast     GUCaUUCGGC...AC
fly       GCC.UU-GGA...GC
cow       GCA.UUCGUC...-C
mouse     GCA.UU-GAU...GC
human     GCGaUUCGCU...GC
chicken   GUA.UUCGUA...AC
snake     GUGaUUCGCG...AC
croc      GUU.UU-GAG...AC
worm      G-G.UUCGCGccaAC
starfish  G-U.UUCGAU...-C
```

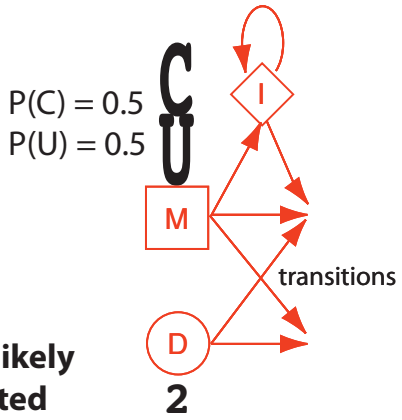
One HMM node per alignment column

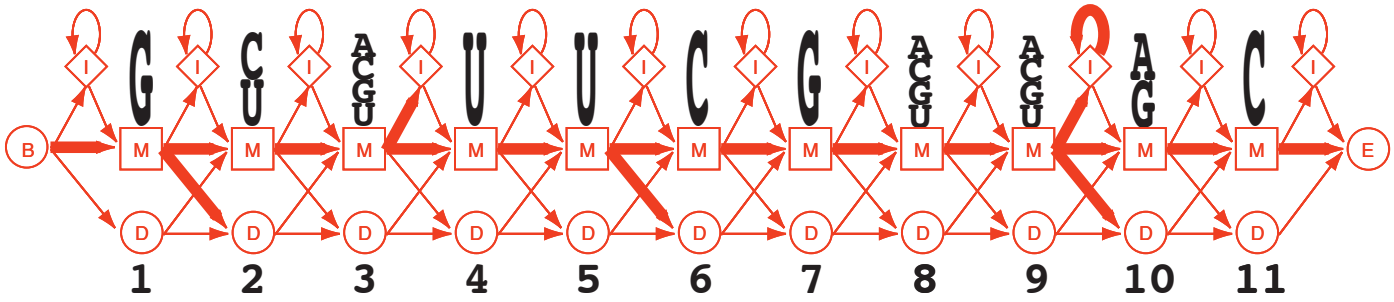
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

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

Node for column 2:





Sequences are aligned to profiles HMMs using dynamic programming algorithms very similar to Smith-Waterman


```
yeast     GUCaUUCGGC...AC
fly       GCC.UU-GGA...GC
cow       GCA.UUCGUC...-C
mouse     GCA.UU-GAU...GC
human     GCGaUUCGCU...GC
chicken   GUA.UUCGUA...AC
snake     GUGaUUCGCG...AC
croc      GUU.UU-GAG...AC
worm      G-G.UUCGCGccaAC
starfish  G-U.UUCGAU...-C
urchin    GUU.UUC-AA...AC
```

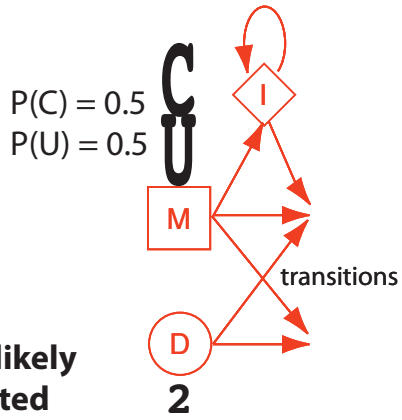
One HMM node per alignment column

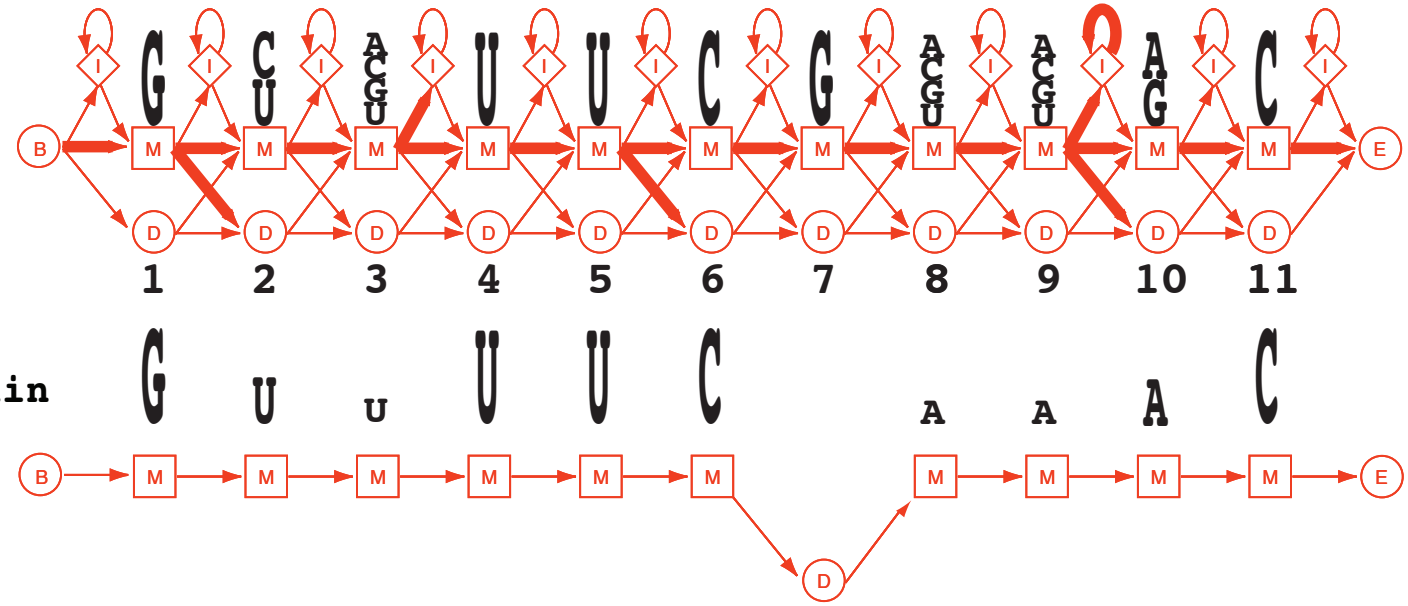
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

HMMs generate homologous sequences.

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

Node for column 2:





Profile alignment can differ from pairwise alignment


```
yeast     GUCaUUCGGC...AC
fly       GCC.UU-GGA...GC
cow       GCA.UUCGUC...-C
mouse     GCA.UU-GAU...GC
human     GCGaUUCGCU...GC
chicken   GUA.UUCGUA...AC
snake     GUGaUUCGCG...AC
croc      GUU.UU-GAG...AC
worm      G-G.UUCGCGccaAC
starfish  G-U.UUCGAU...-C
urchin    GUU.UUC-AA...AC
```

```
starfish GUUUCGAUC
urchin   GUUUUCAAAC
```

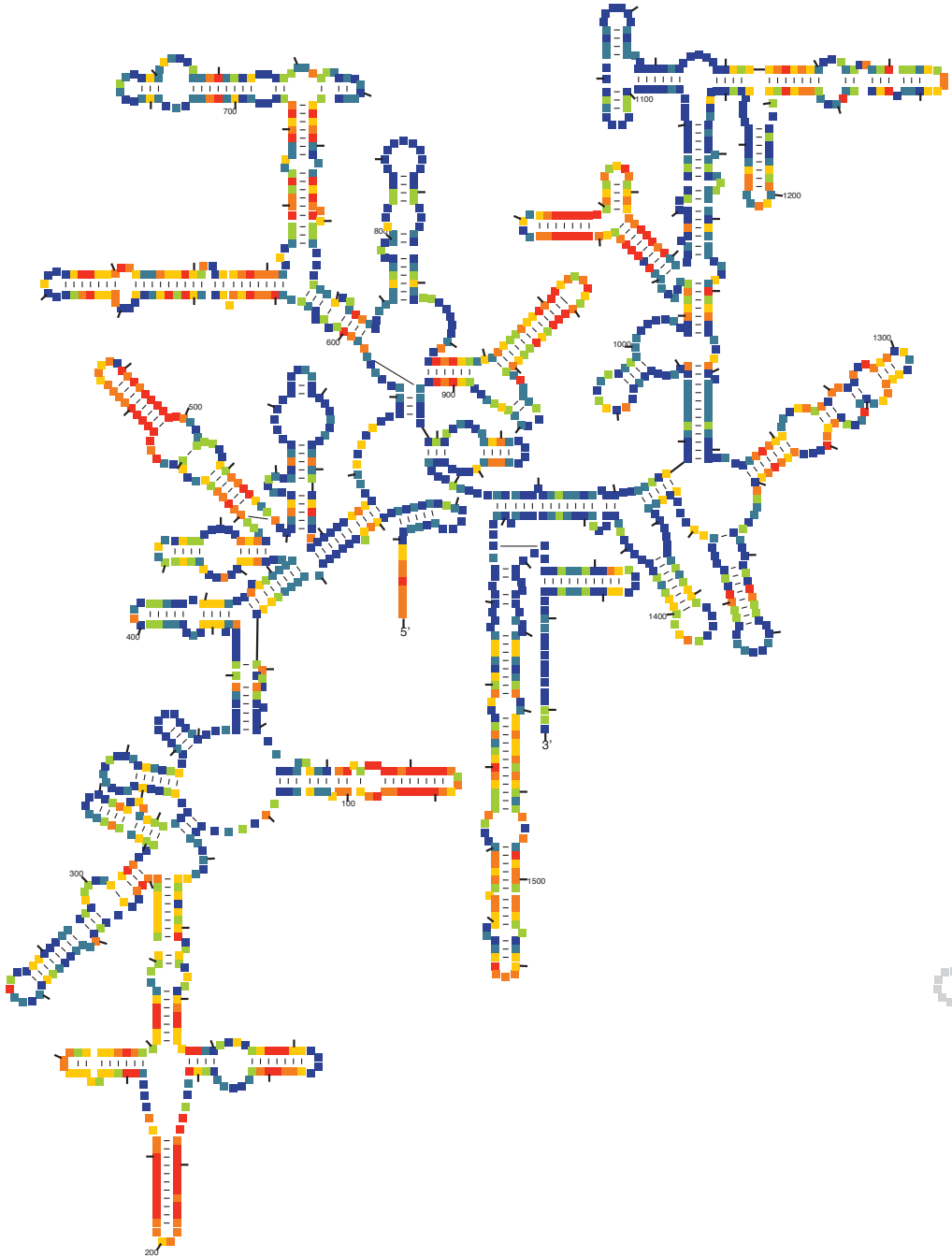
Pairwise alignment to nearest neighbor

```
starfish G-UUUCGAUC
urchin   GUUUUCAAAC
         * **** * *
```


```
yeast     GUCaUUCGGC...AC
fly       GCC.UU-GGA...GC
cow       GCA.UUCGUC...-C
mouse     GCA.UU-GAU...GC
human     GCGaUUCGCU...GC
chicken   GUA.UUCGUA...AC
snake     GUGaUUCGCG...AC
croc      GUU.UU-GAG...AC
worm      G-G.UUCGCGccaAC
starfish  G-U.UUCGAU...-C
urchin    GUU.UUCAAA...-C
```

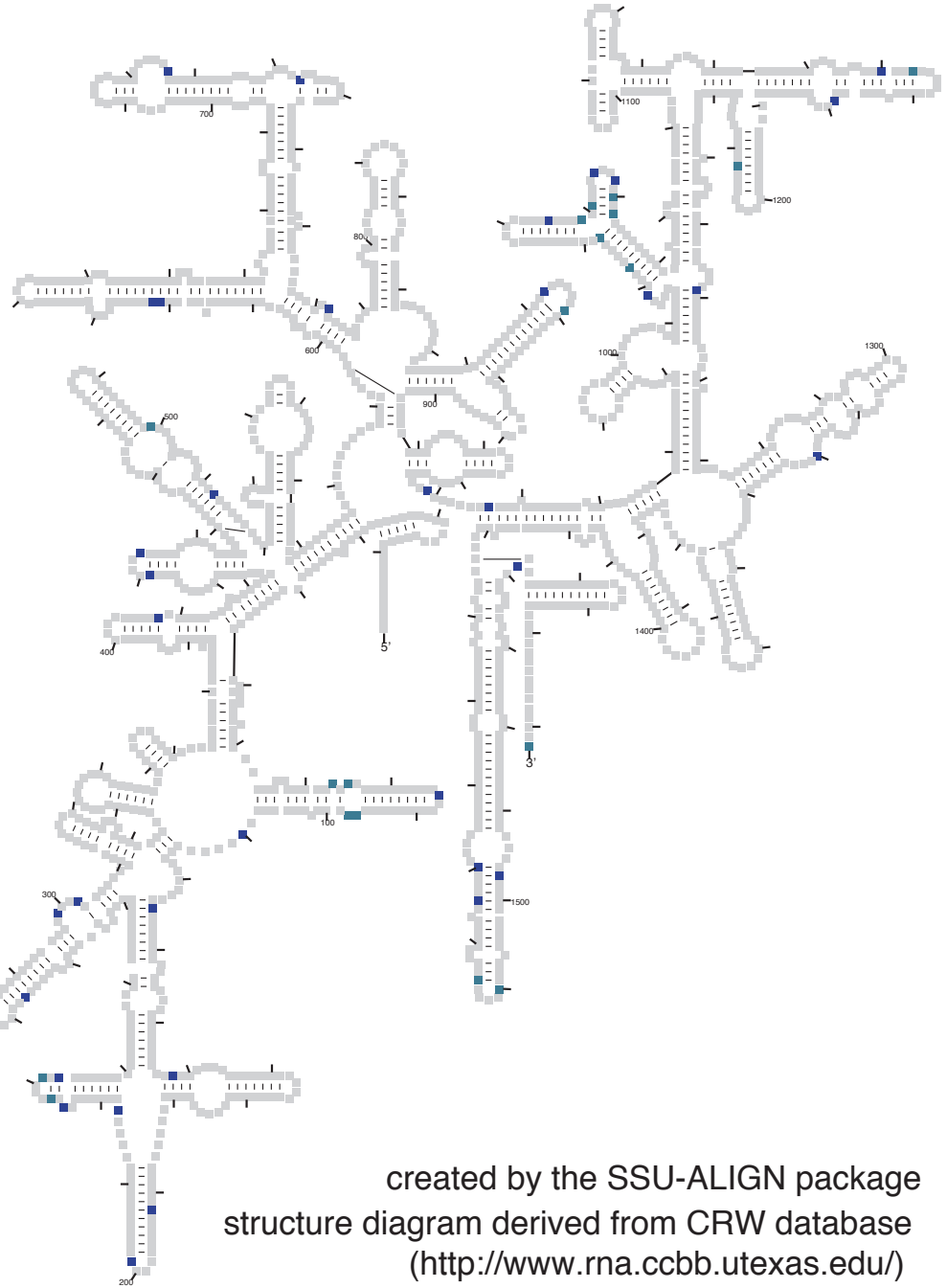
Sequence conservation per position

blue: highly conserved ..... red: highly variable



Frequency of insertions after each position

grey: zero to very few inserts ..... teal: 1-2% inserts

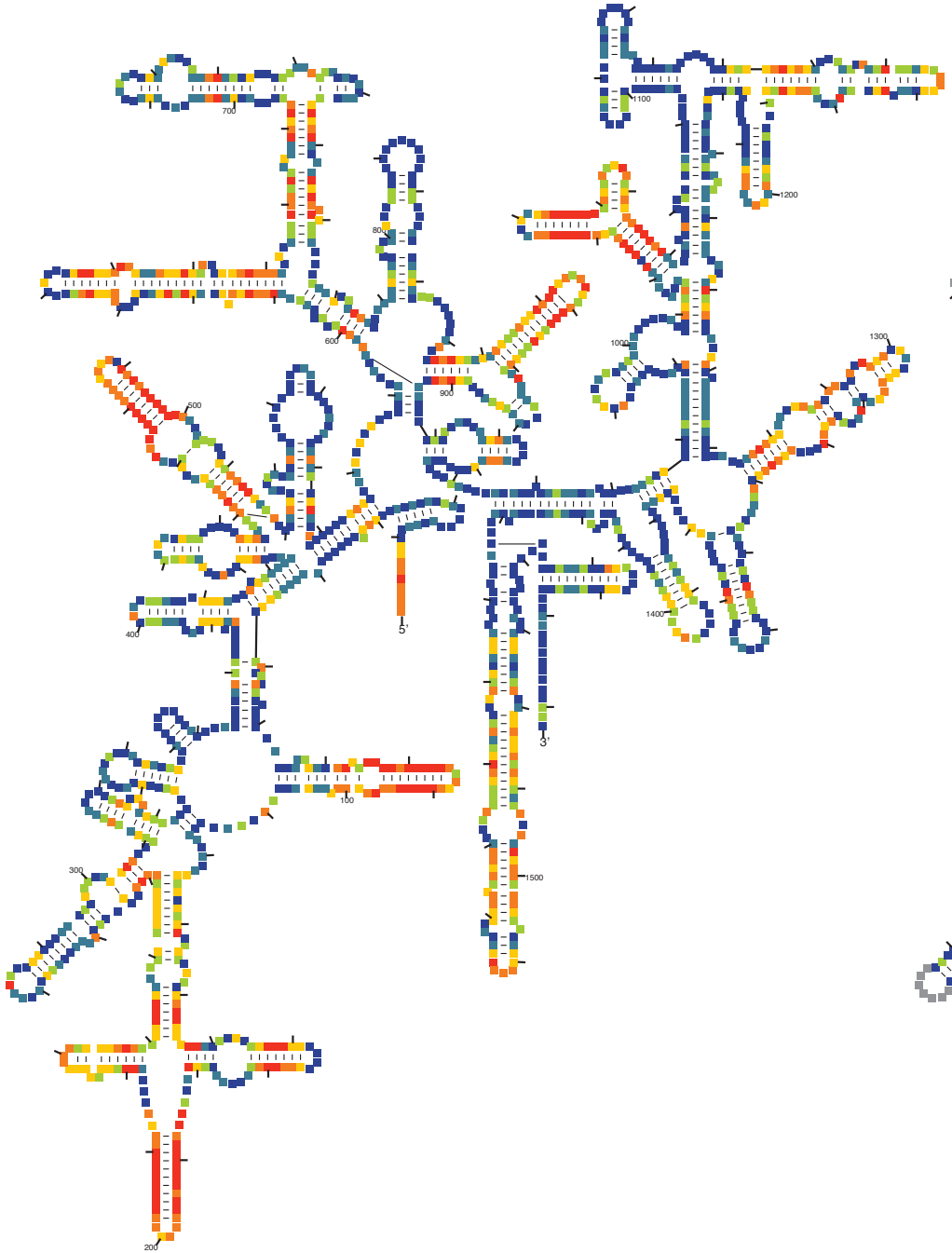


created by the SSU-ALIGN package

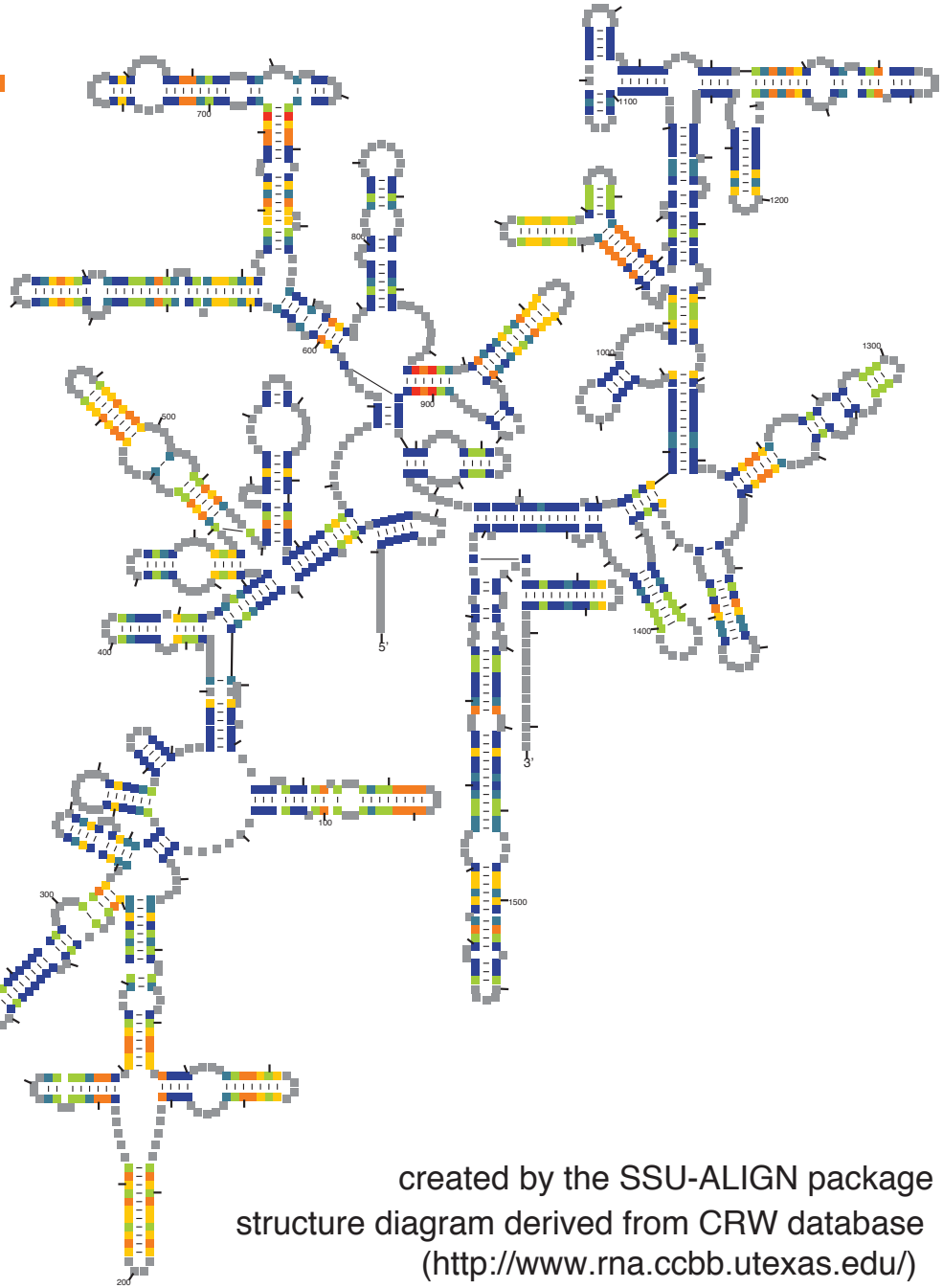
structure diagram derived from CRW database

(http://www.rna.ccbb.utexas.edu/)

Sequence conservation per position
blue: highly conserved ..... red: highly variable

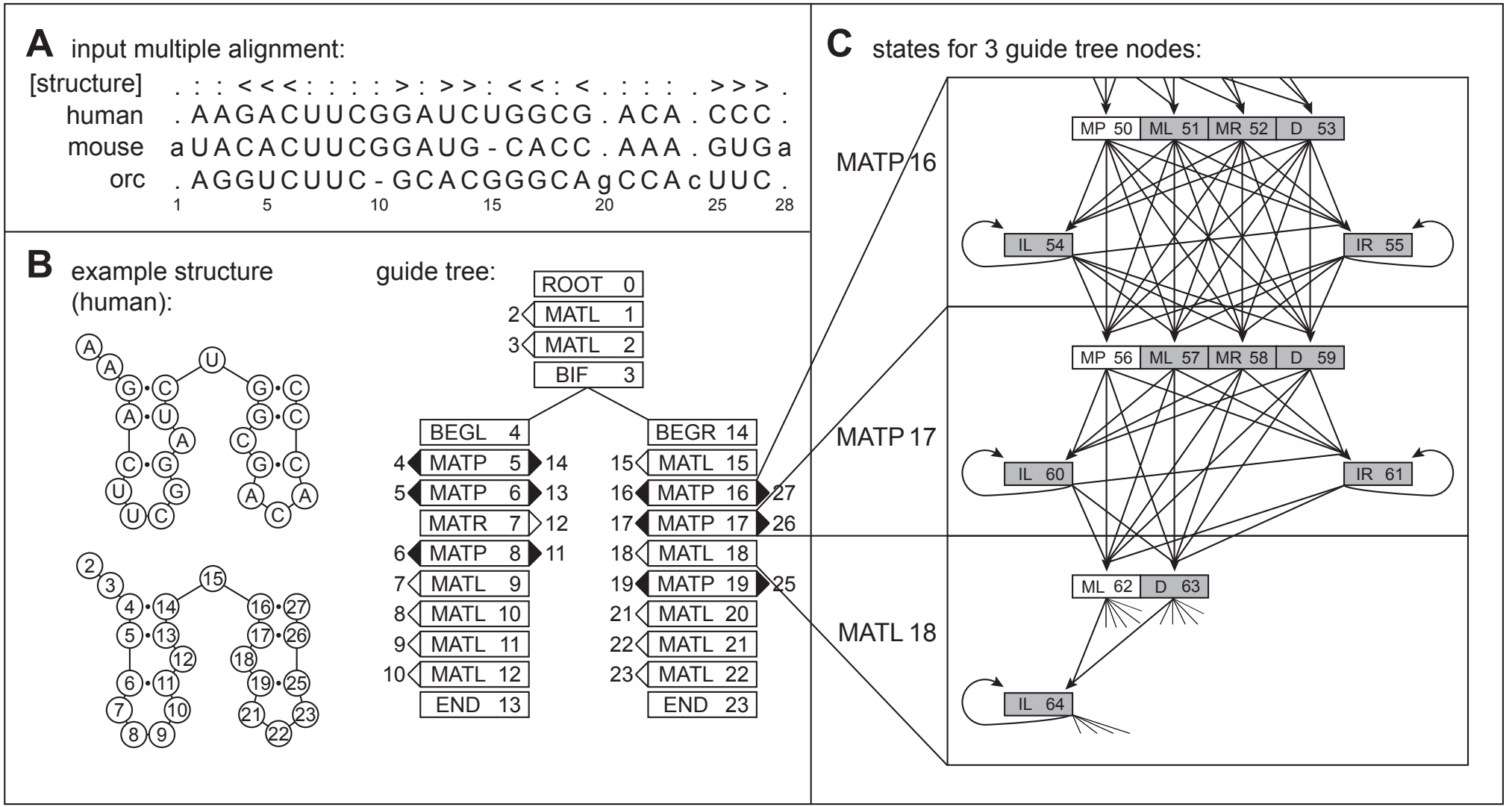


Secondary structure (mutual) information per position
blue: low information ..... red: high information



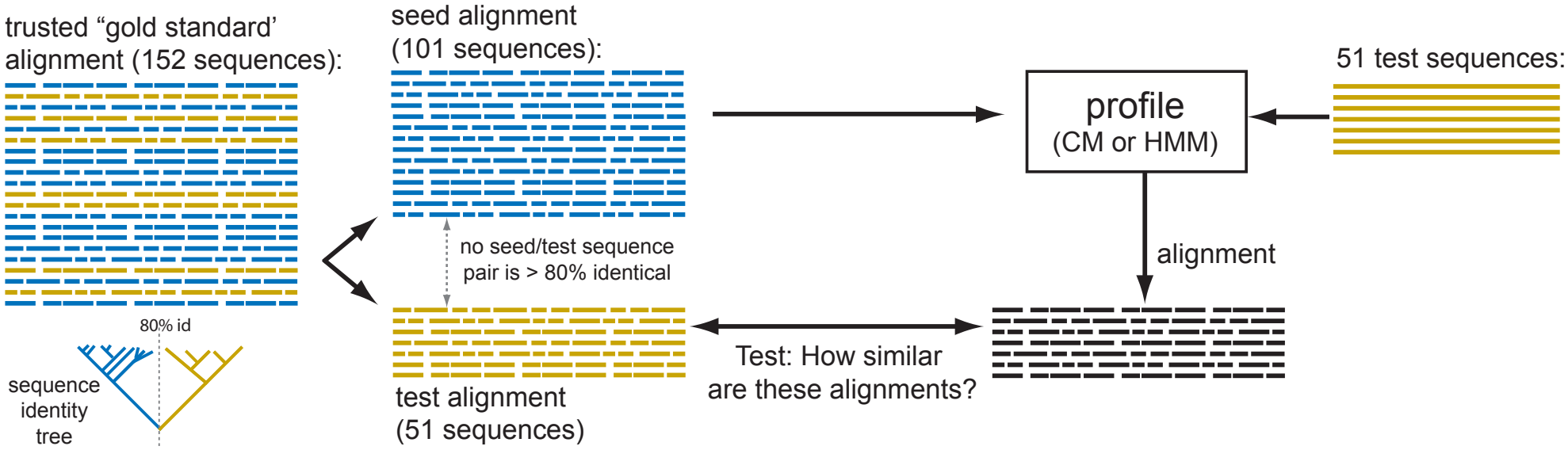
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

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



Benchmark of SSU alignment

- How accurate are profile-based alignments?
- '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



Profiles produce accurate SSU alignments

	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

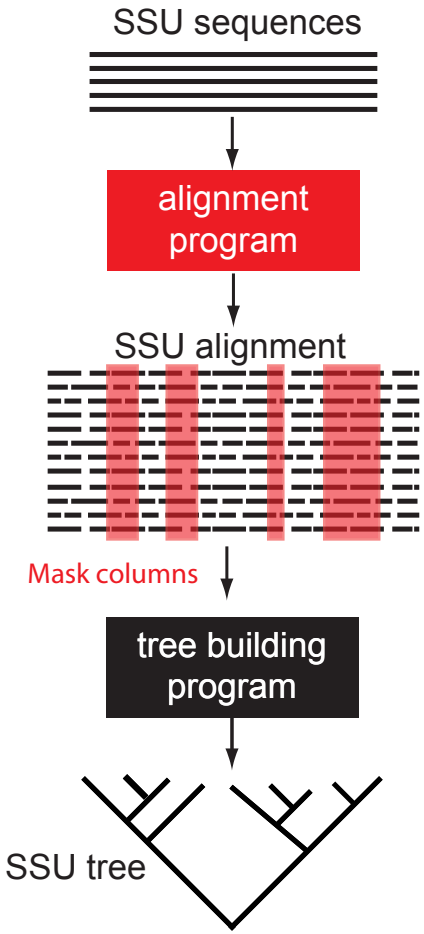
Muscle: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

HMMER: hmmer.janelia.org

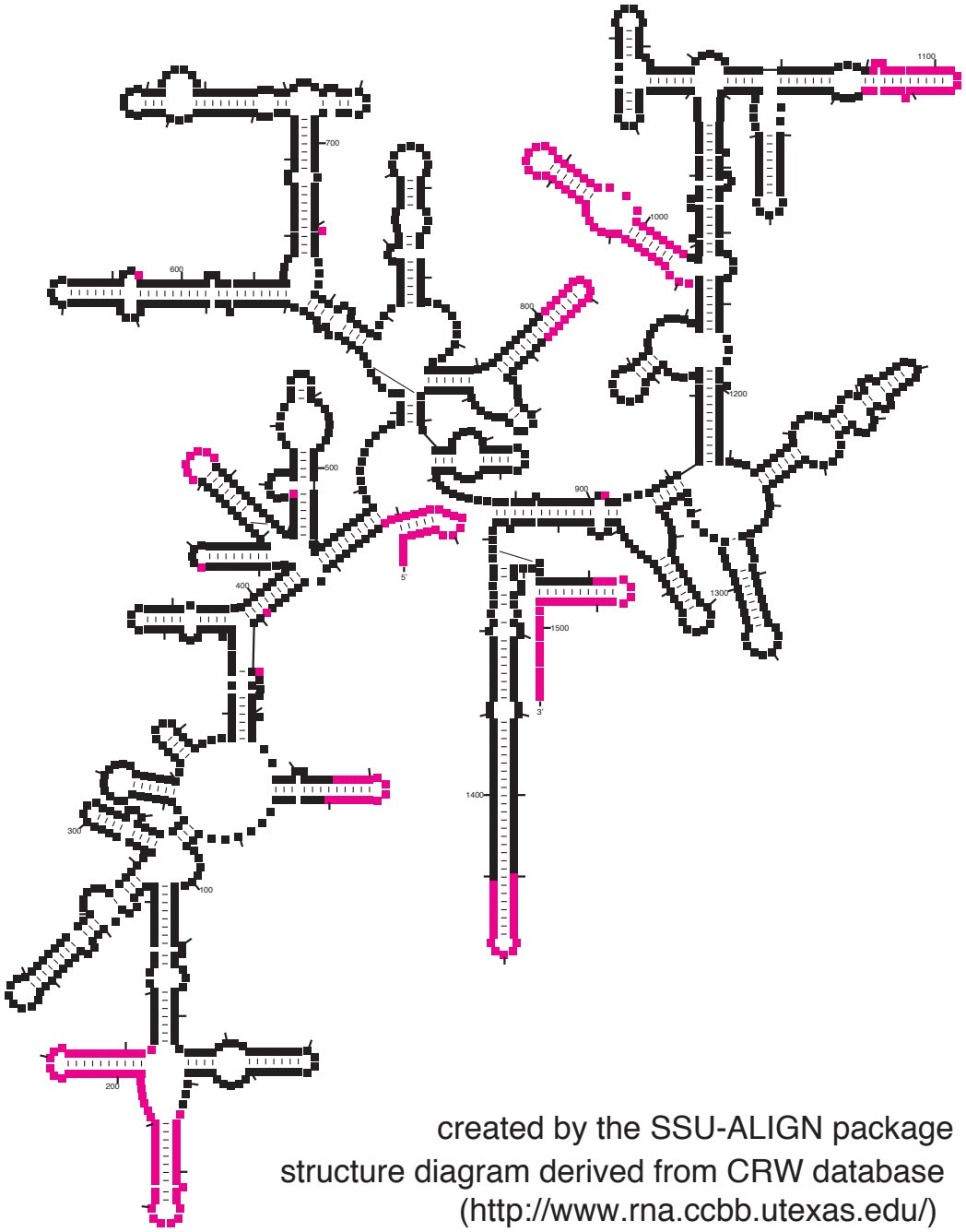
Infernal: infernal.janelia.org

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

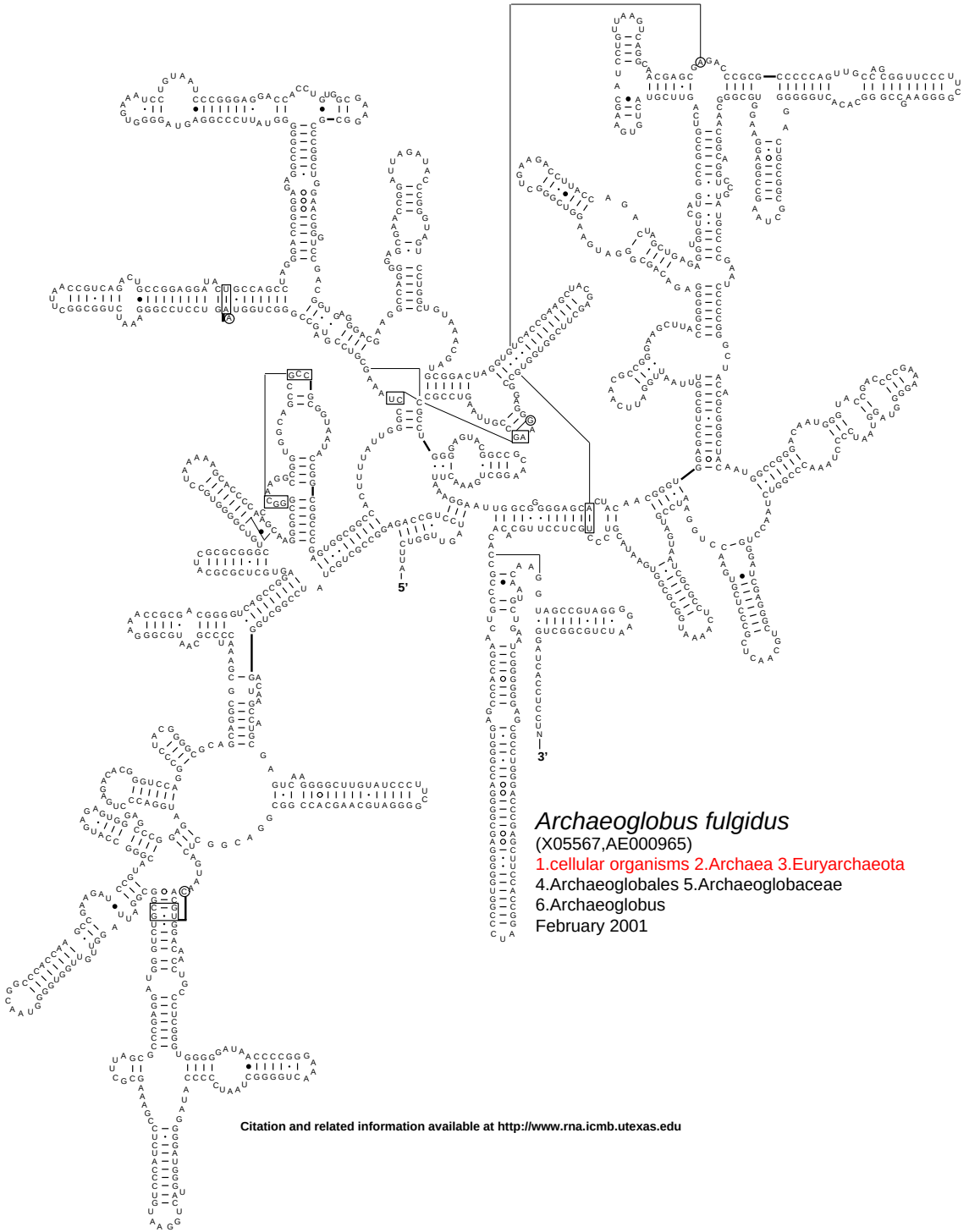


Phil Hugenholtz's manually created mask imposed on archaeal SSU
black: included in alignment (1257)
pink: excluded from alignment (251)



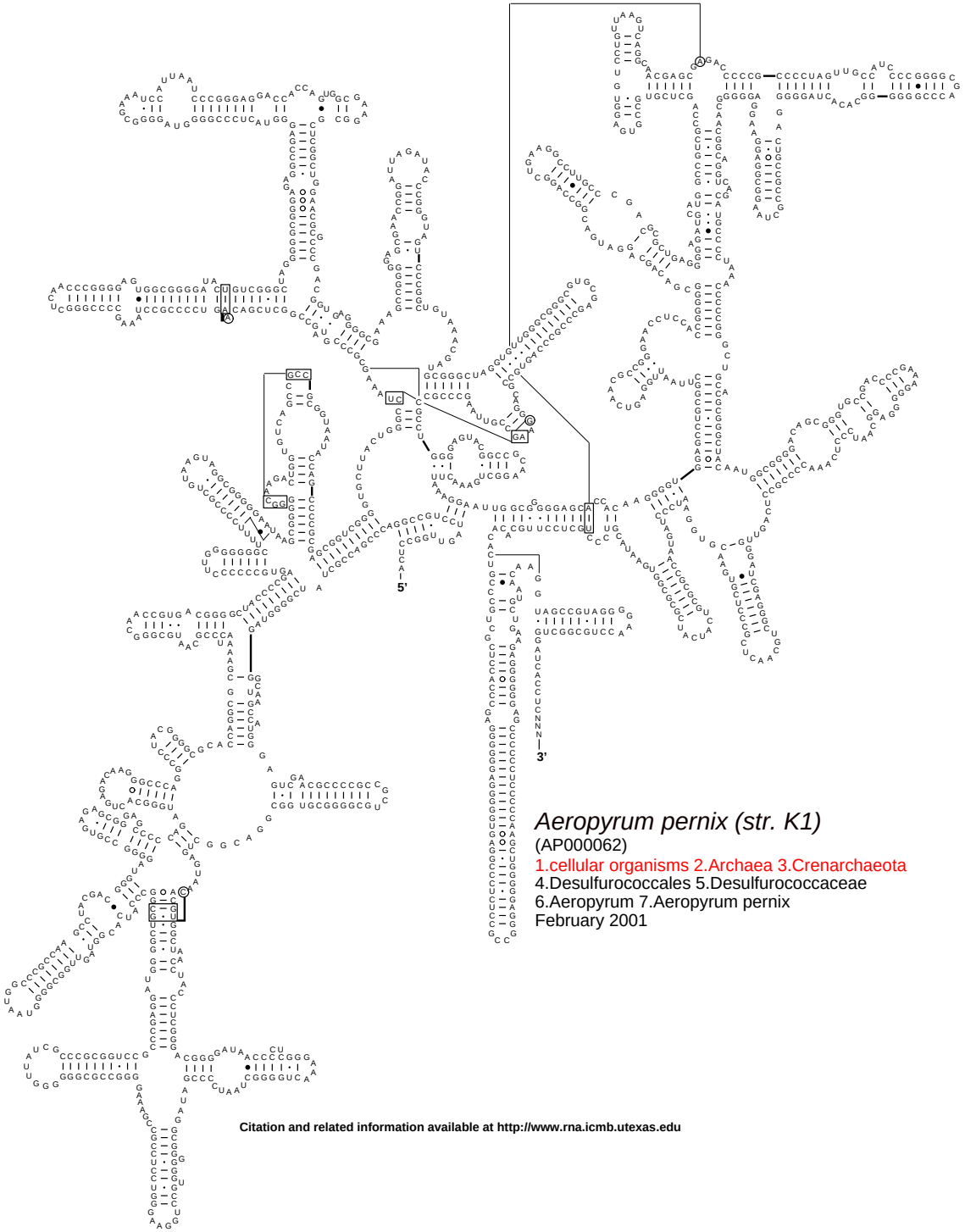
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Secondary Structure: small subunit ribosomal RNA



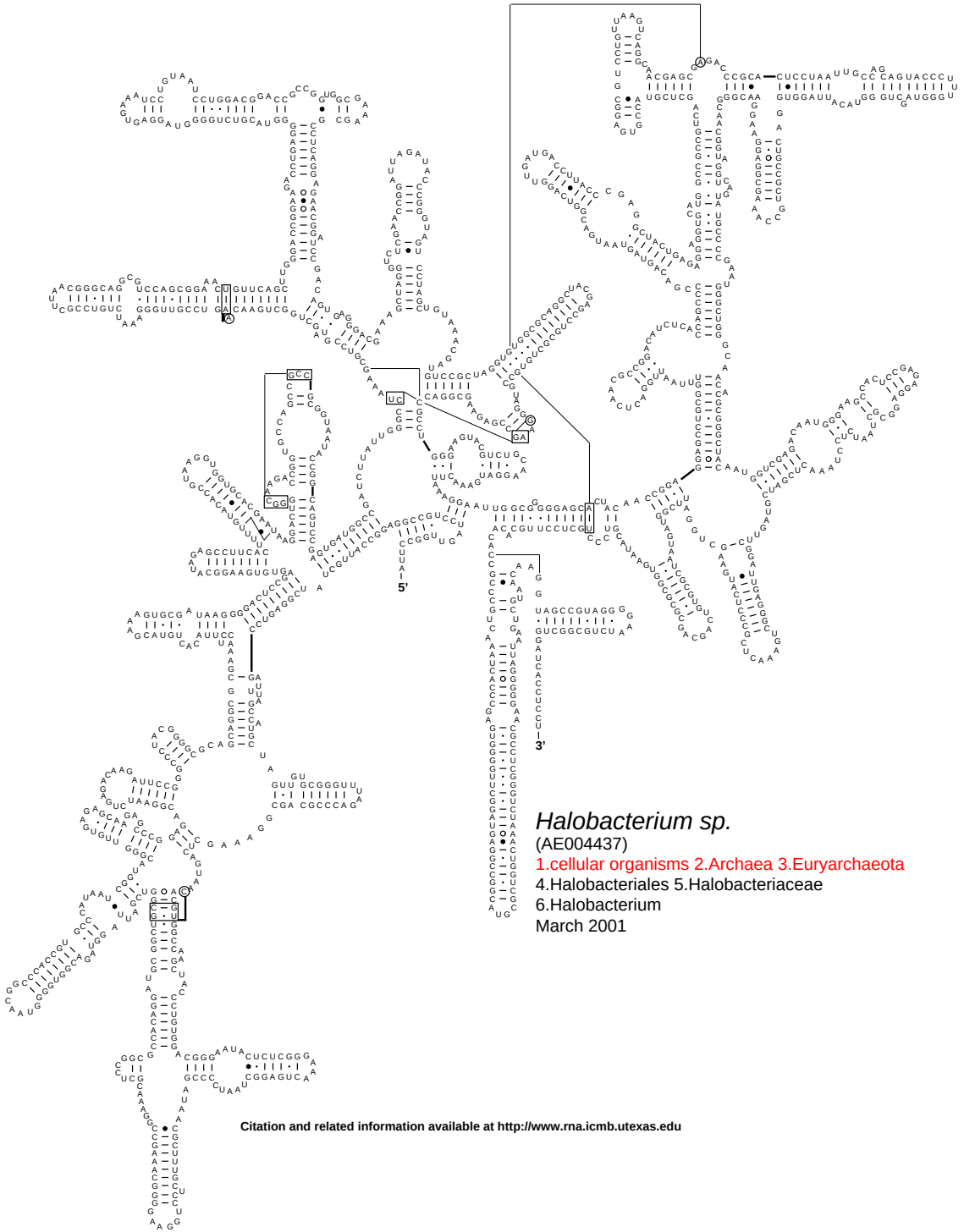
Archaeoglobus fulgidus
(X05567,AE000965)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Archaeoglobales 5.Archaeoglobaceae
6.Archaeoglobus
February 2001

Secondary Structure: small subunit ribosomal RNA



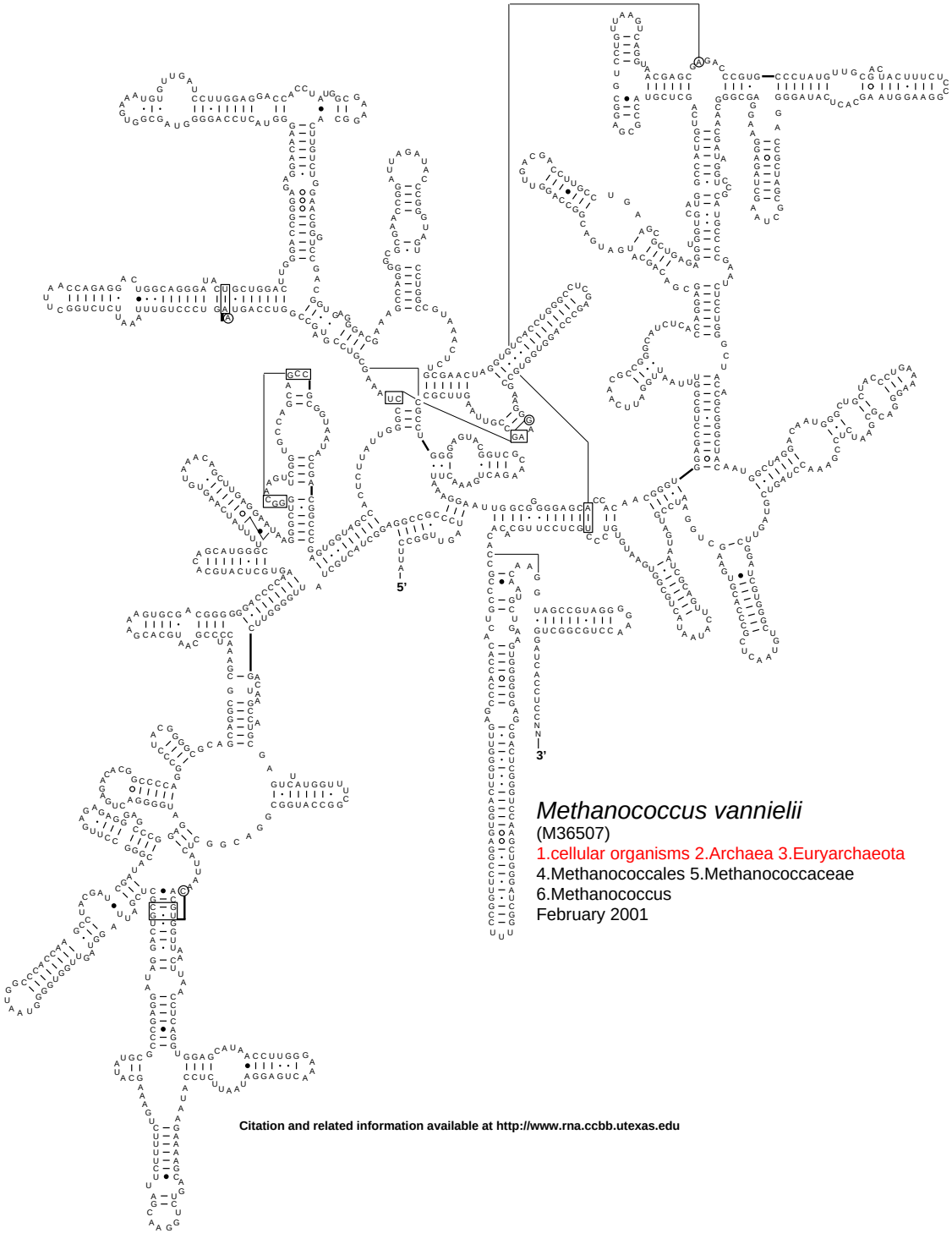
Aeropyrum pernix (str. K1)
(AP000062)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Desulfurococcales 5.Desulfurococcaceae
6.Aeropyrum 7.Aeropyrum pernix
February 2001

Secondary Structure: small subunit ribosomal RNA



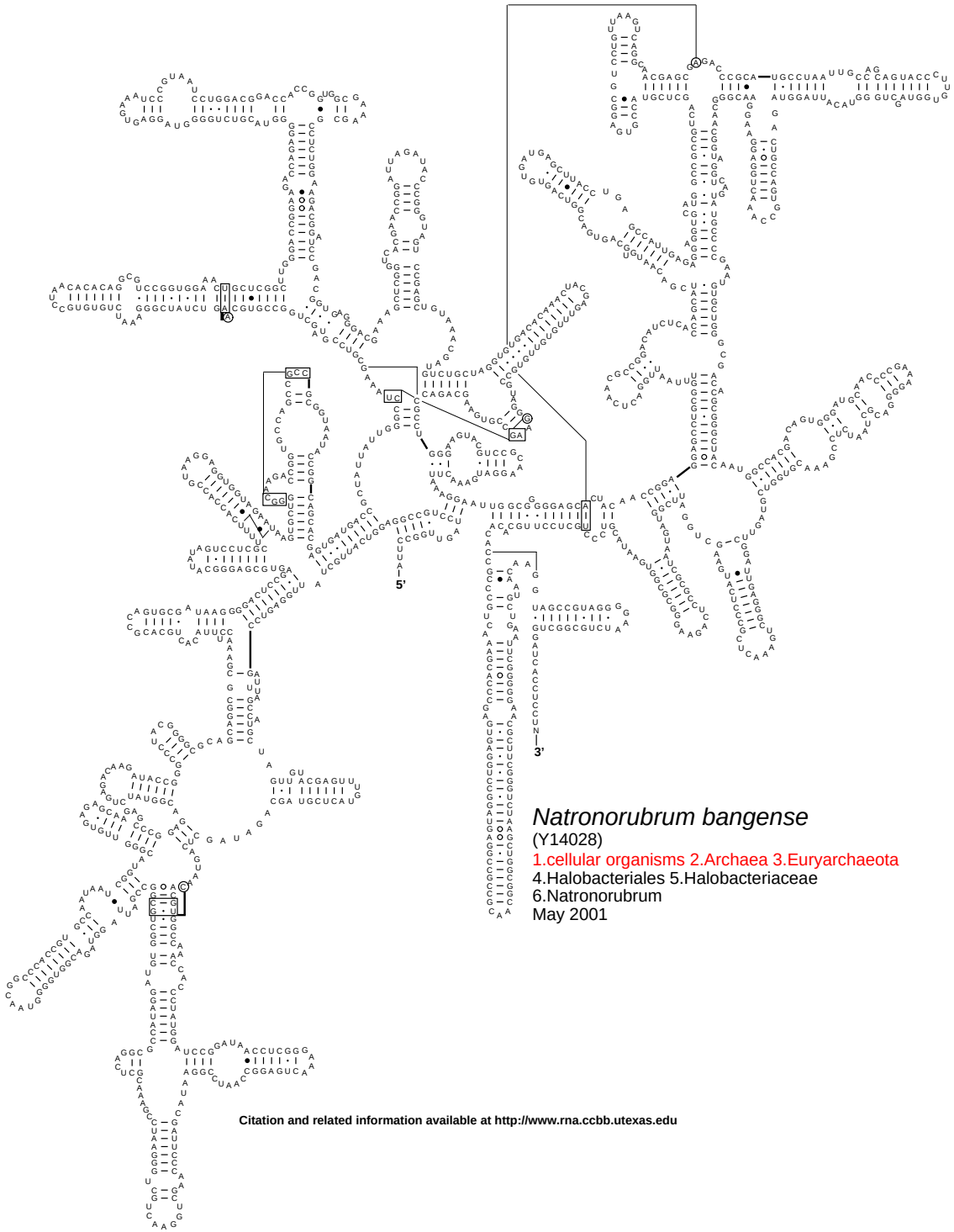
Halobacterium sp.
(AE004437)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Halobacterium
March 2001

Secondary Structure: small subunit ribosomal RNA



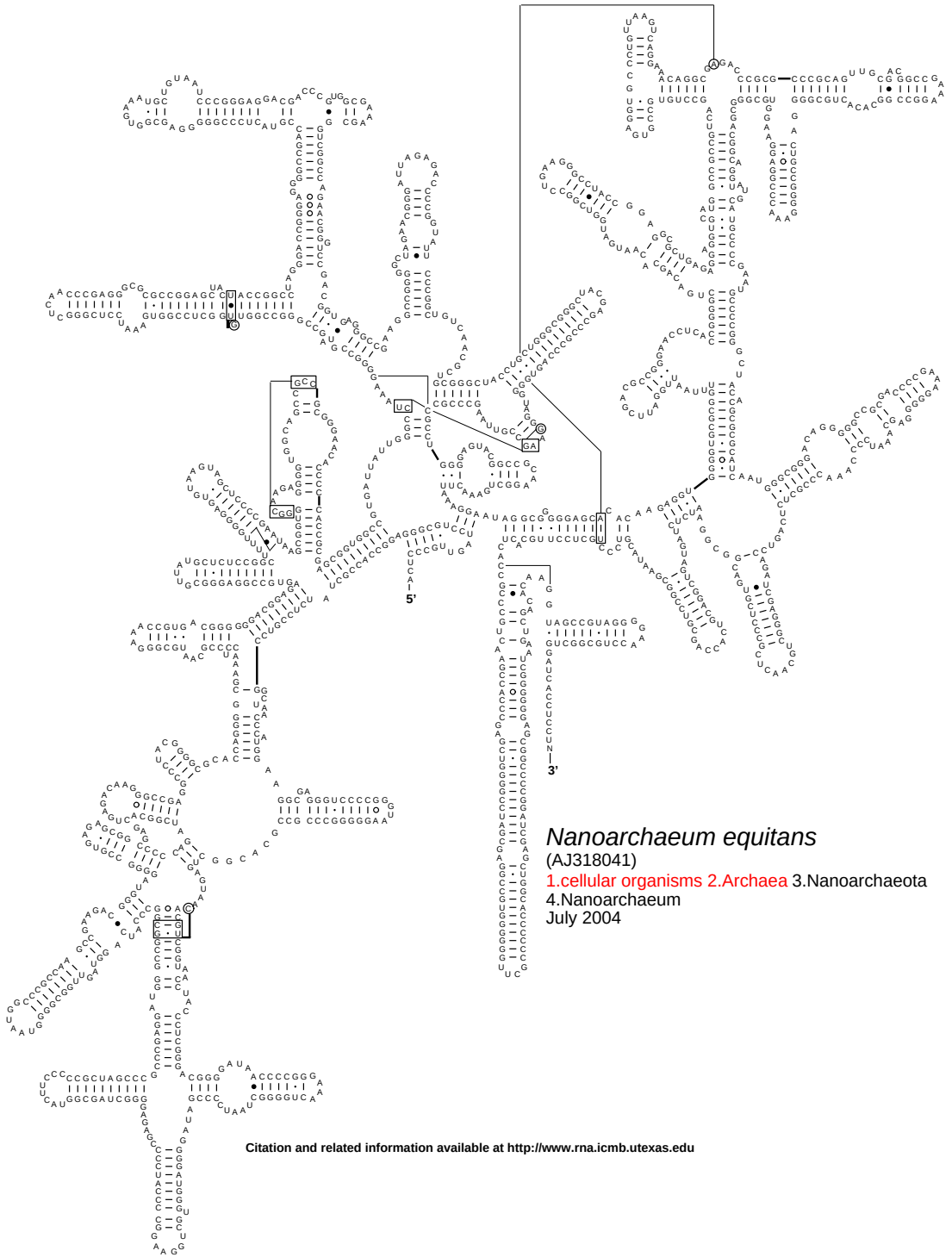
Methanococcus vannielii
(M36507)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Methanococcales 5.Methanococcaceae
6.Methanococcus
February 2001

Secondary Structure: small subunit ribosomal RNA



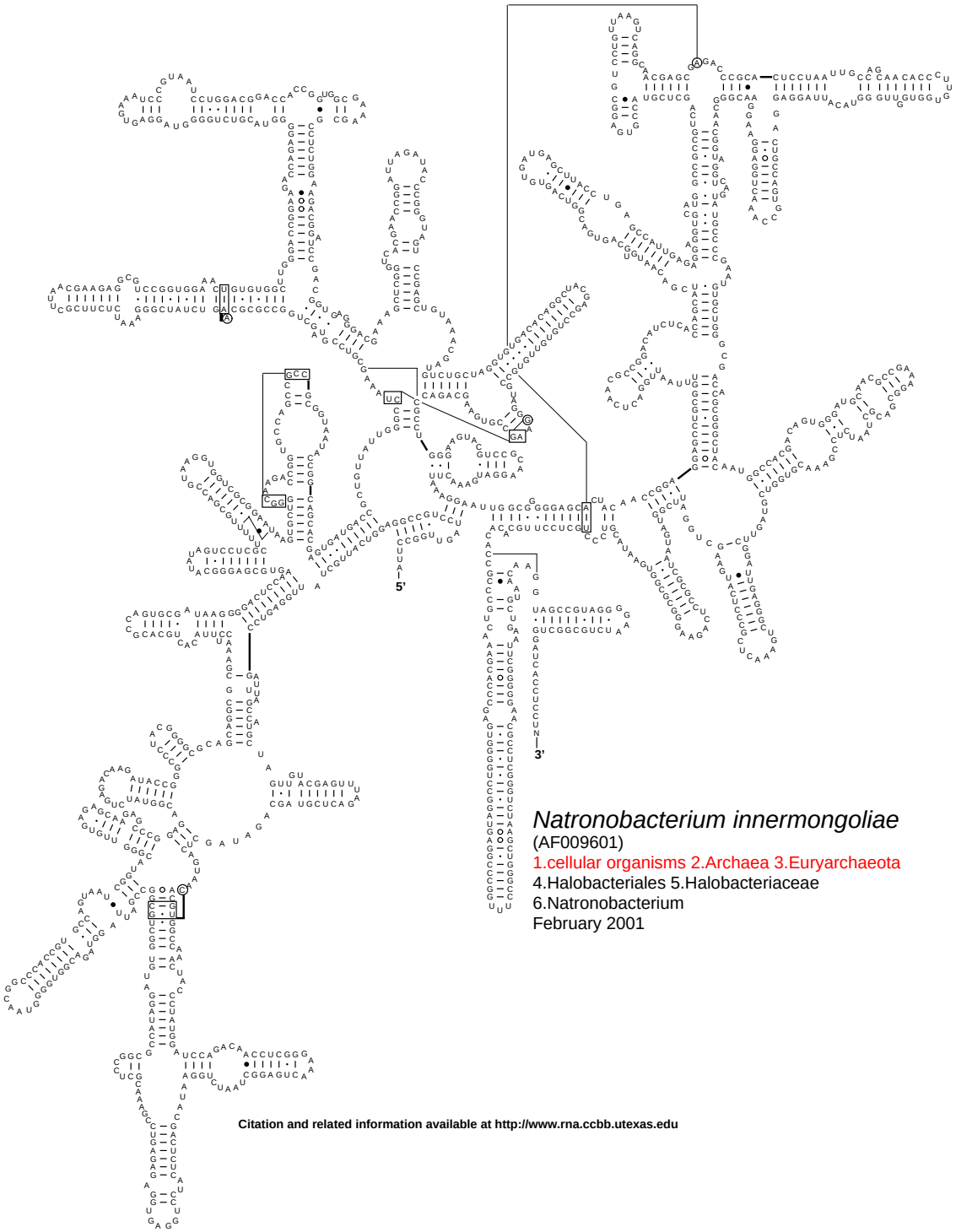
Natronorubrum bangense
(Y14028)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronorubrum
May 2001

Secondary Structure: small subunit ribosomal RNA



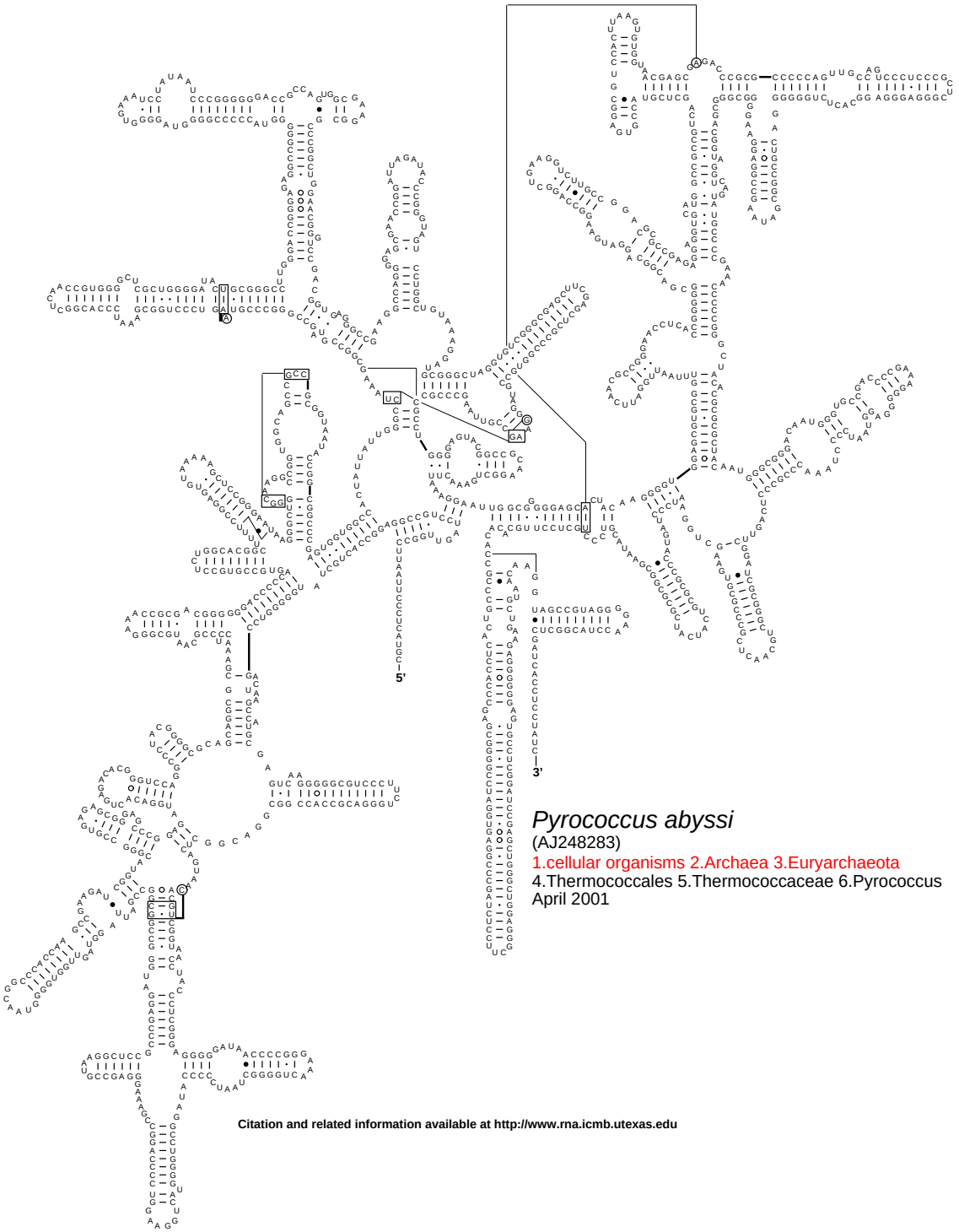
Nanoarchaeum equitans
(AJ318041)
1.cellular organisms 2.Archaea 3.Nanoarchaeota
4.Nanoarchaeum
July 2004

Secondary Structure: small subunit ribosomal RNA



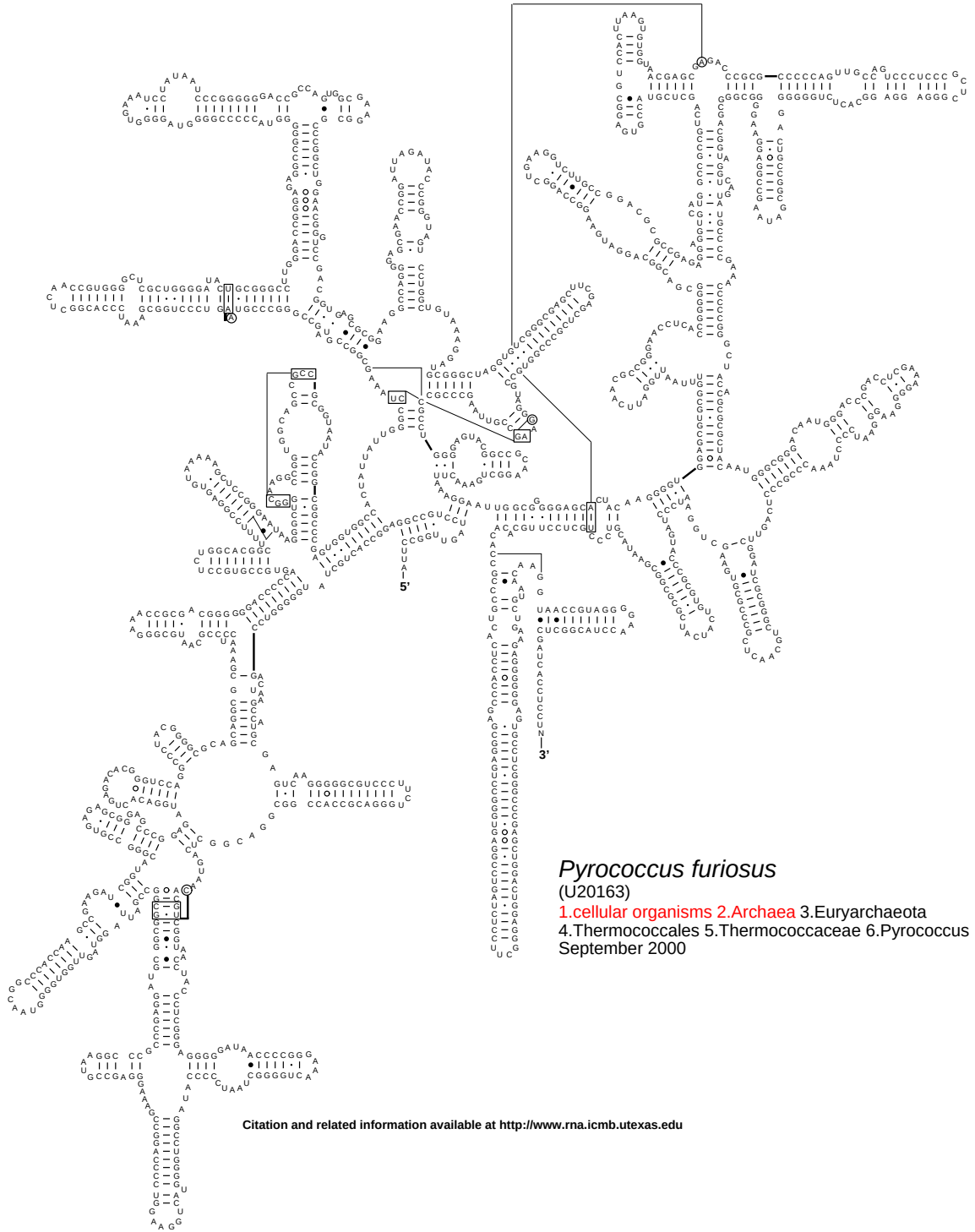
Natronobacterium innermongoliae
(AF009601)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Halobacteriales 5.Halobacteriaceae
6.Natronobacterium
February 2001

Secondary Structure: small subunit ribosomal RNA



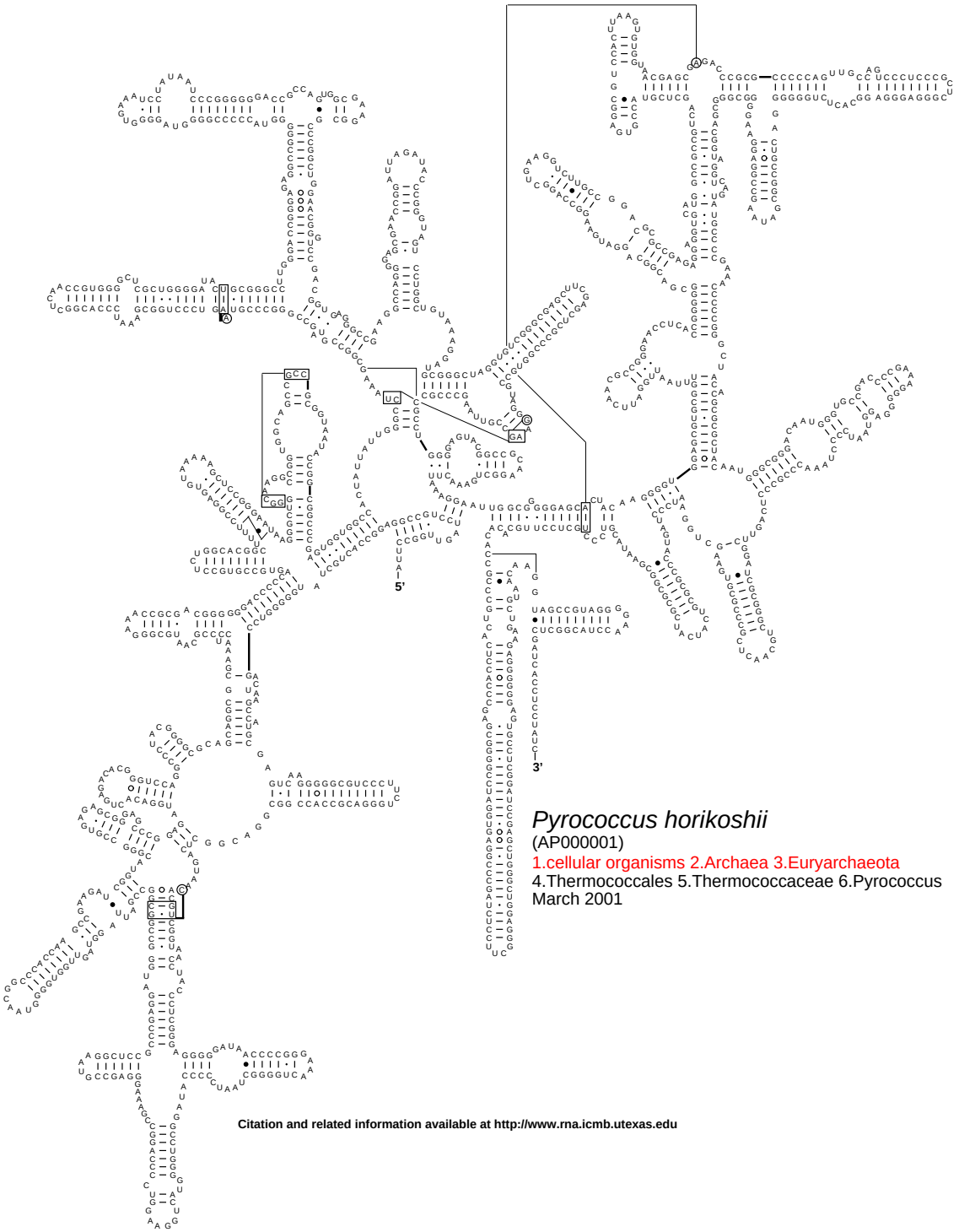
Pyrococcus abyssi
(AJ248283)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
April 2001

Secondary Structure: small subunit ribosomal RNA



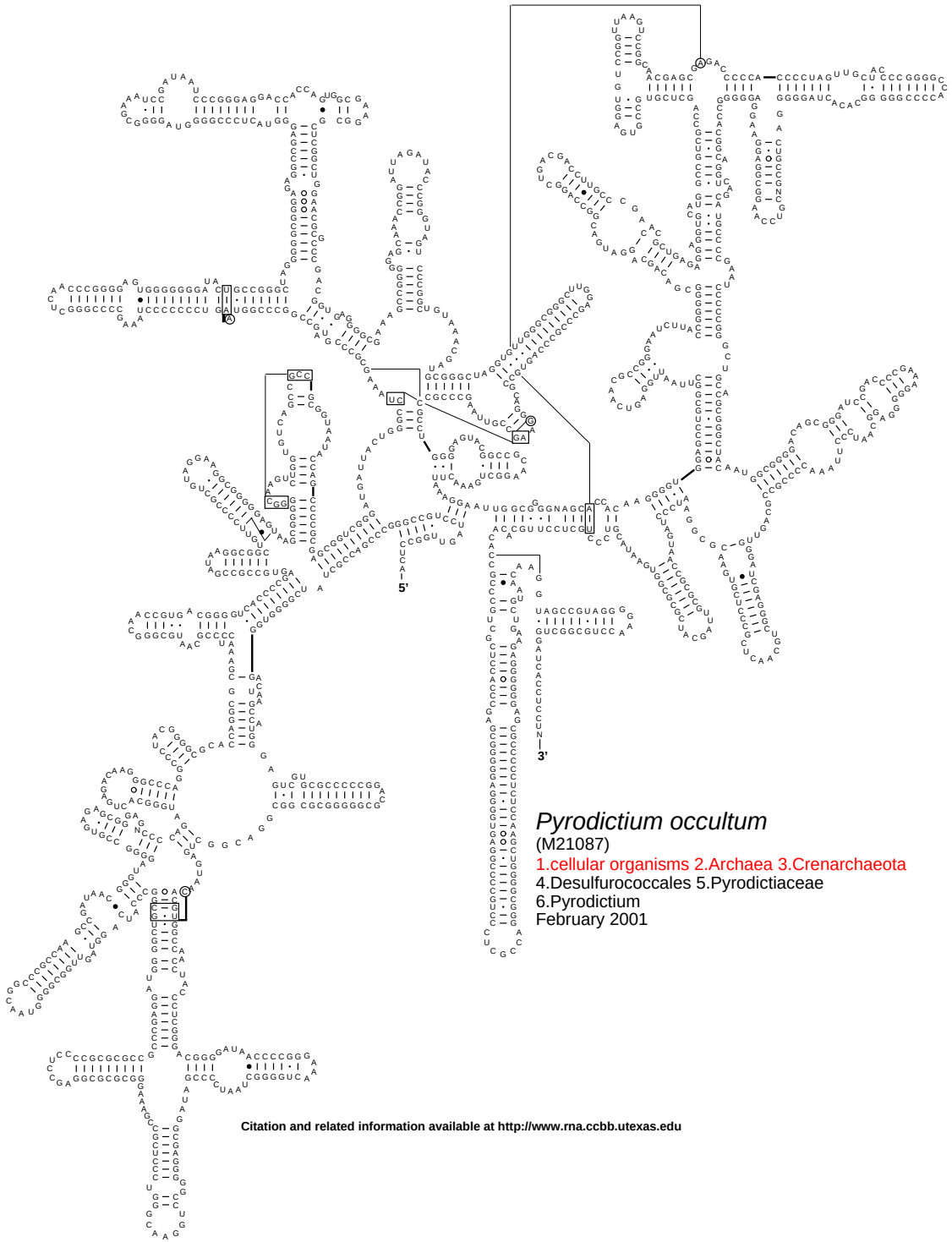
Pyrococcus furiosus
(U20163)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
September 2000

Secondary Structure: small subunit ribosomal RNA



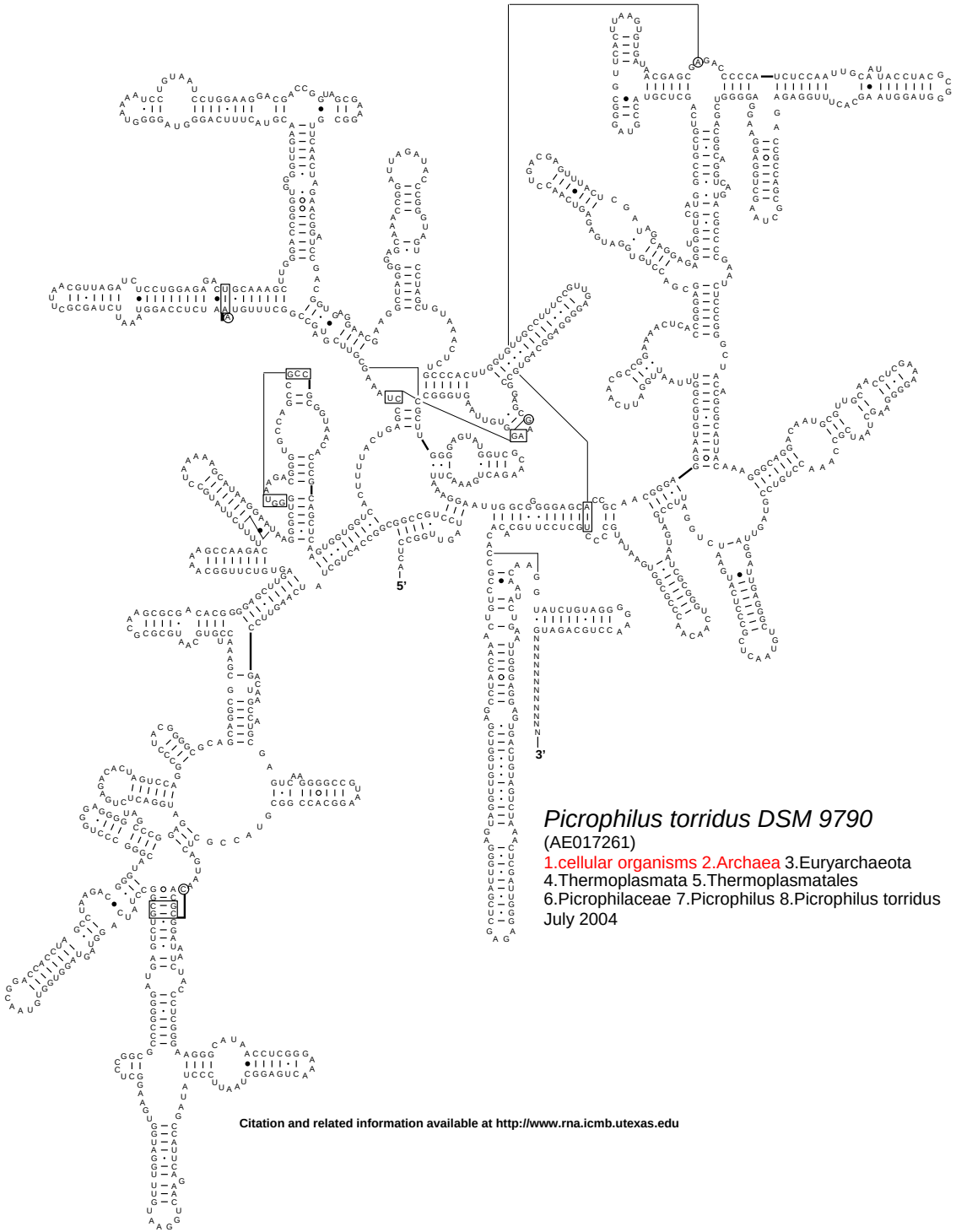
Pyrococcus horikoshii
(AP000001)
1.cellular organisms 2.Archaea 3.Euryarchaeota
4.Thermococcales 5.Thermococcaceae 6.Pyrococcus
March 2001

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

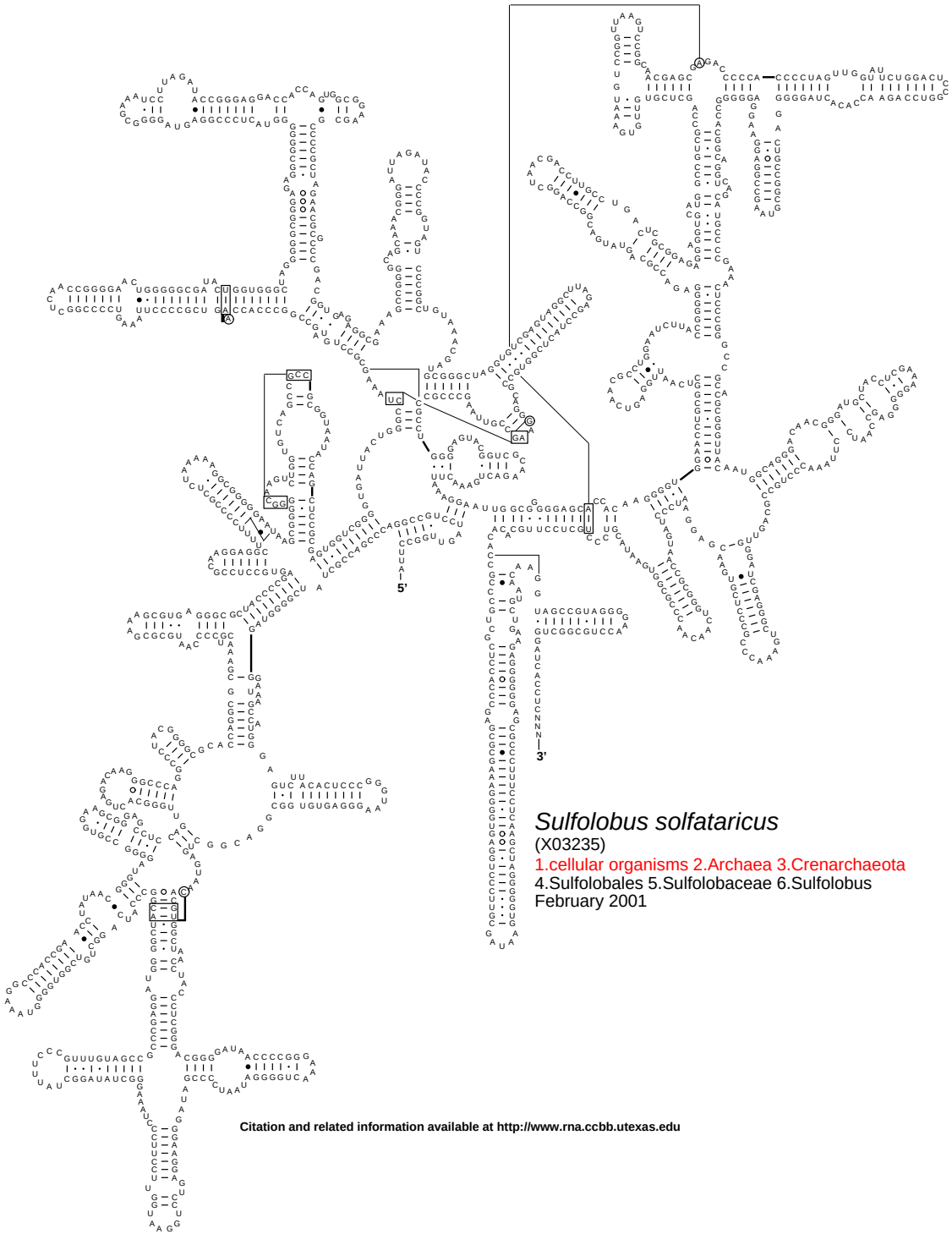


Picrophilus torridus DSM 9790

(AE017261)

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

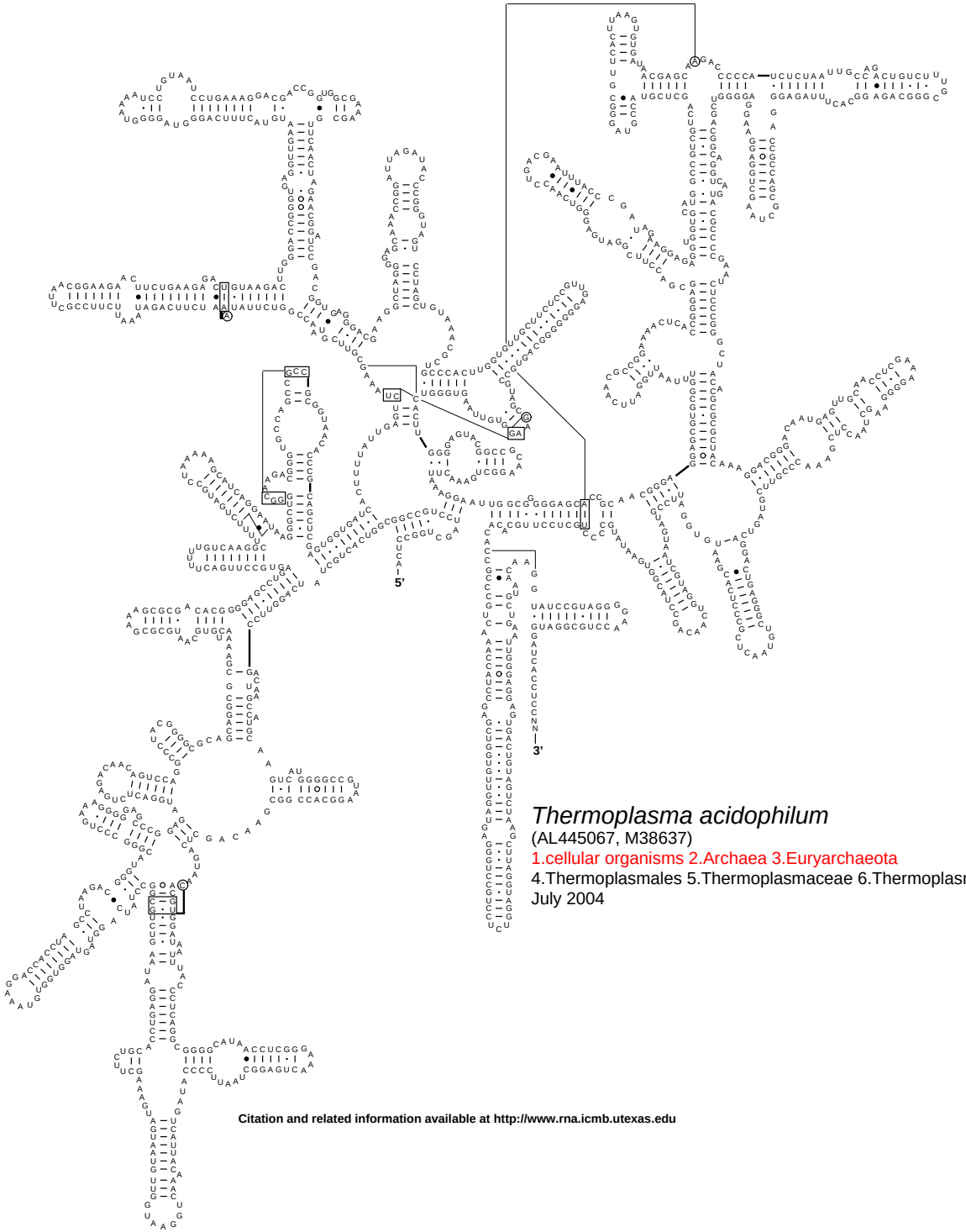
Secondary Structure: small subunit ribosomal RNA



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

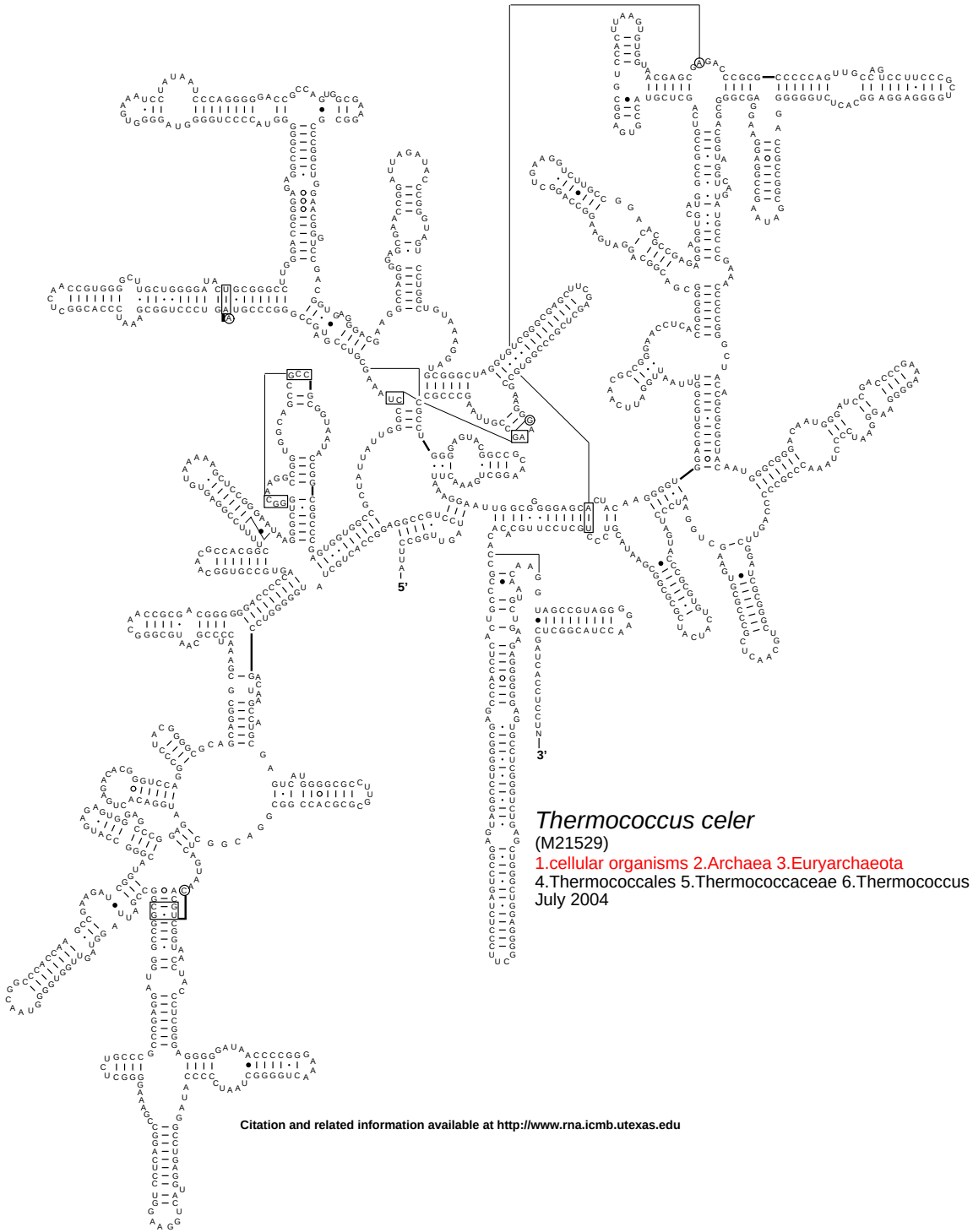
Citation and related information available at http://www.rna.ccbb.utexas.edu

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



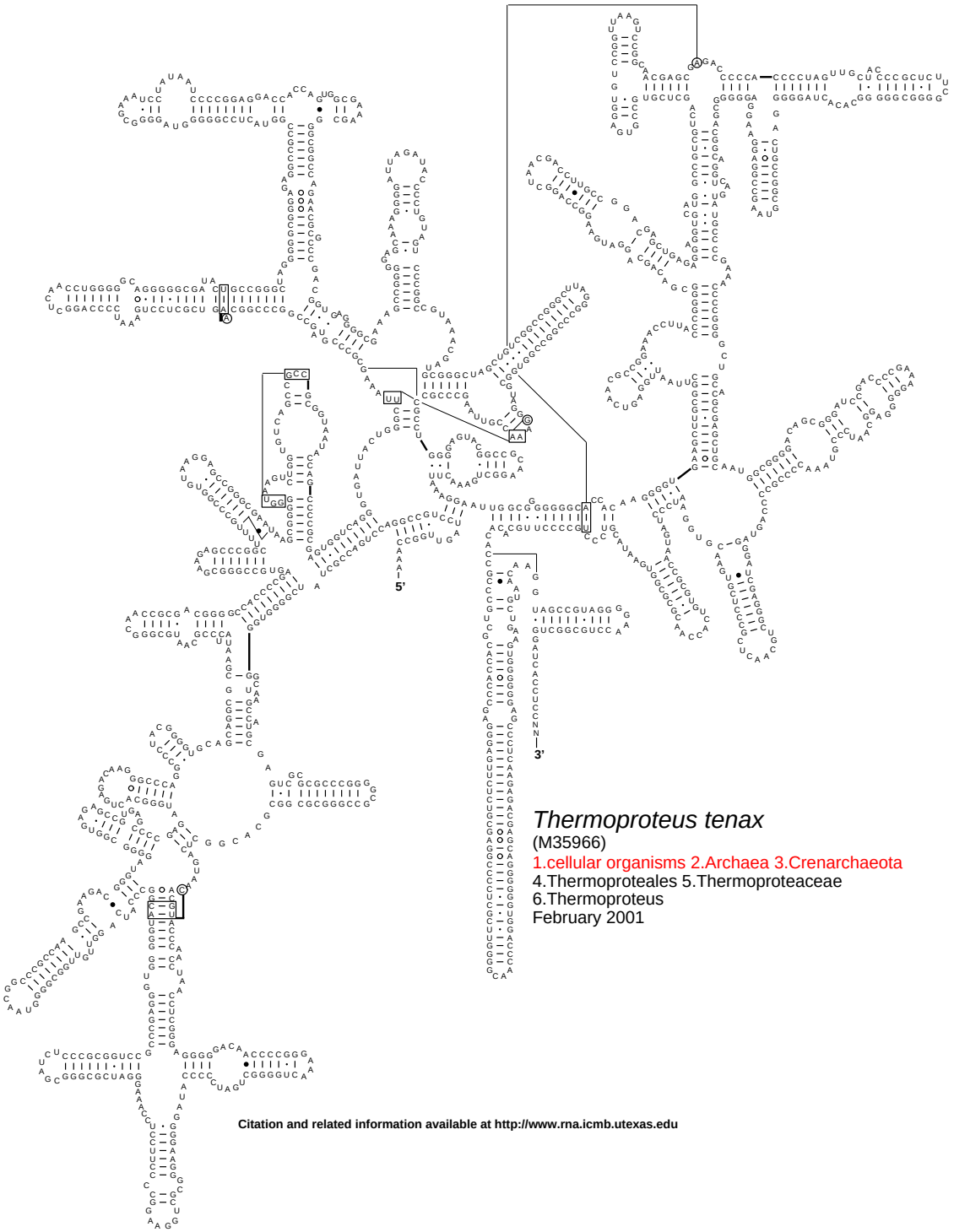
Thermococcus celer

(M21529)

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

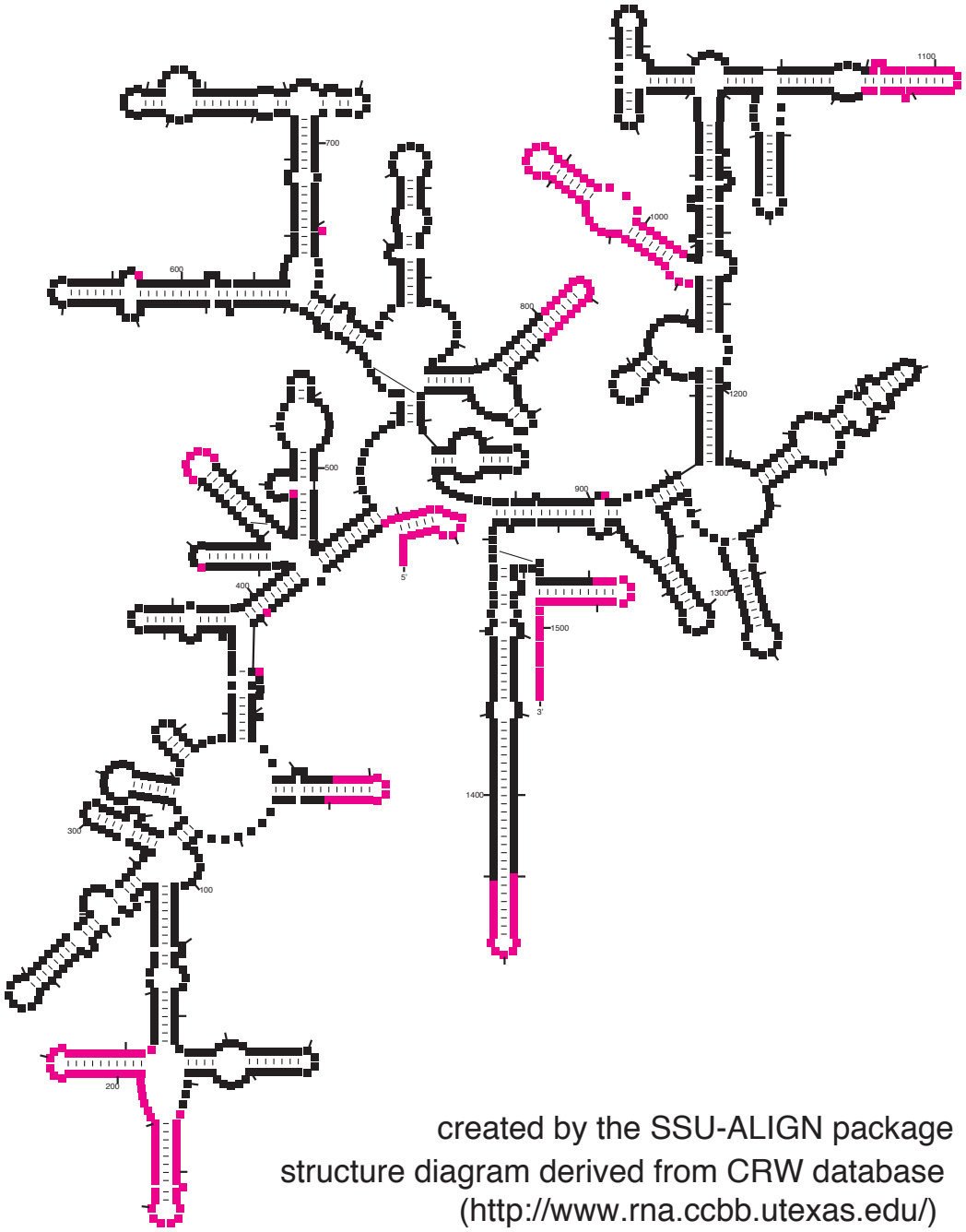
July 2004

Secondary Structure: small subunit ribosomal RNA



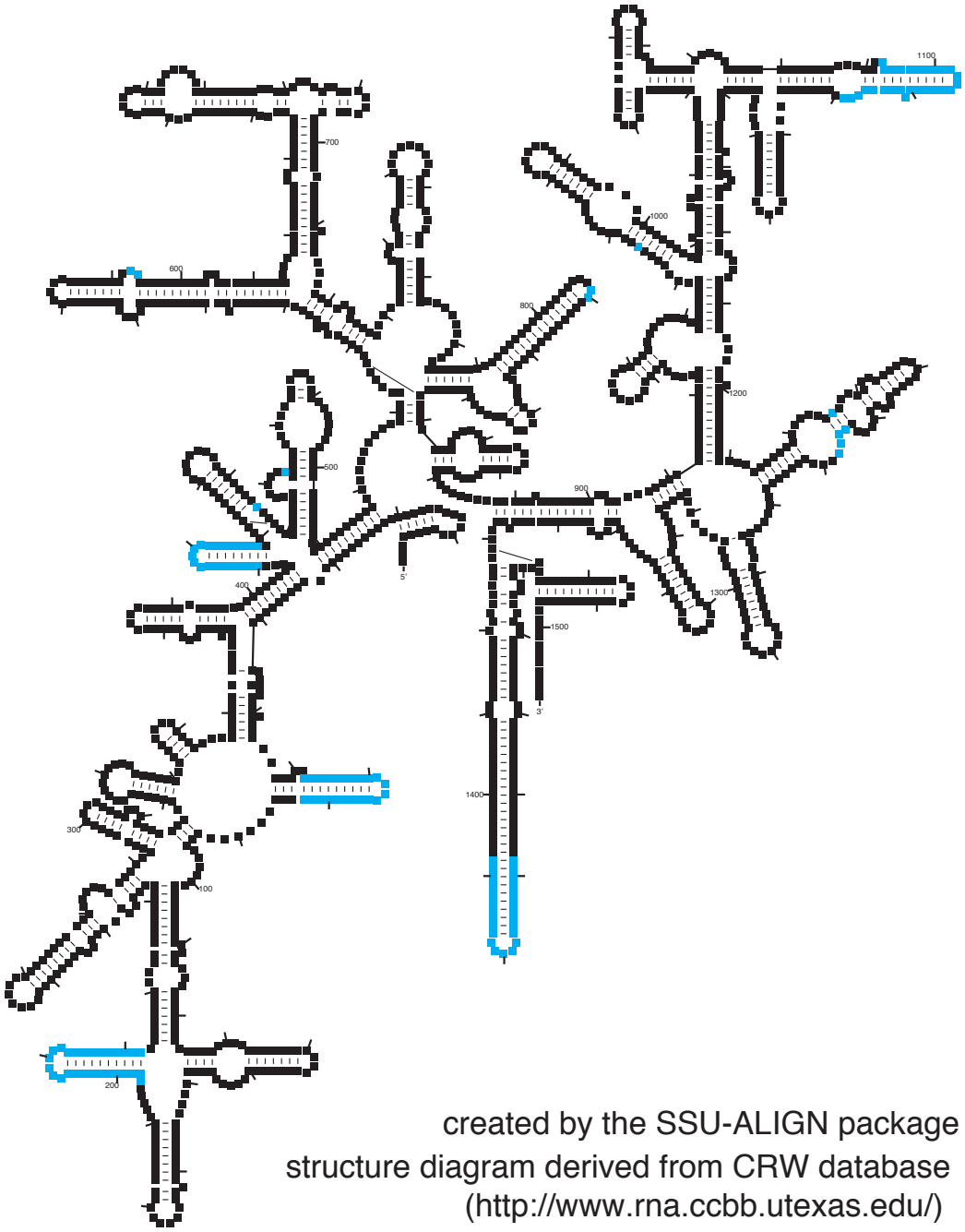
Thermoproteus tenax
(M35966)
1.cellular organisms 2.Archaea 3.Crenarchaeota
4.Thermoproteales 5.Thermoproteaceae
6.Thermoproteus
February 2001

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



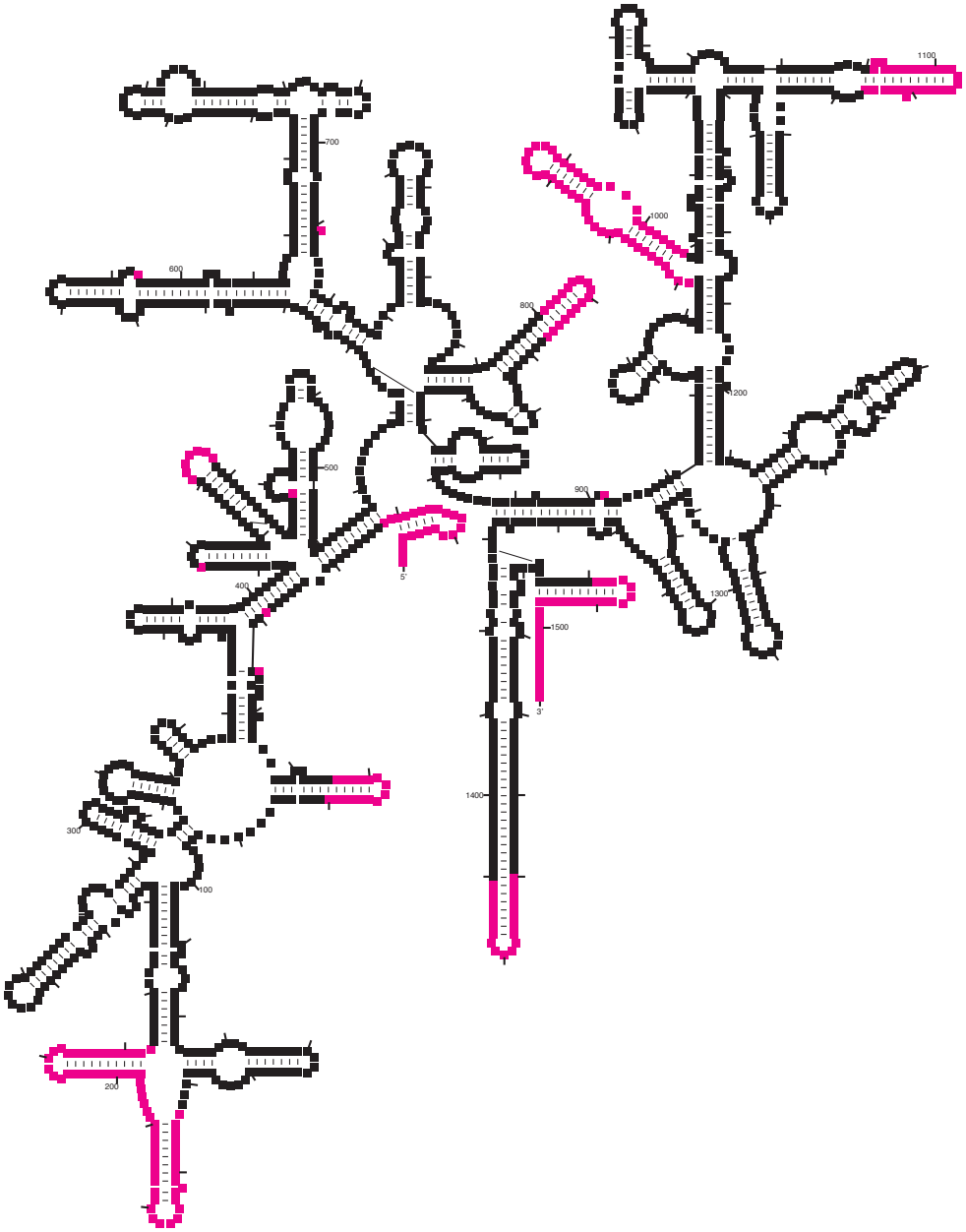
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

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

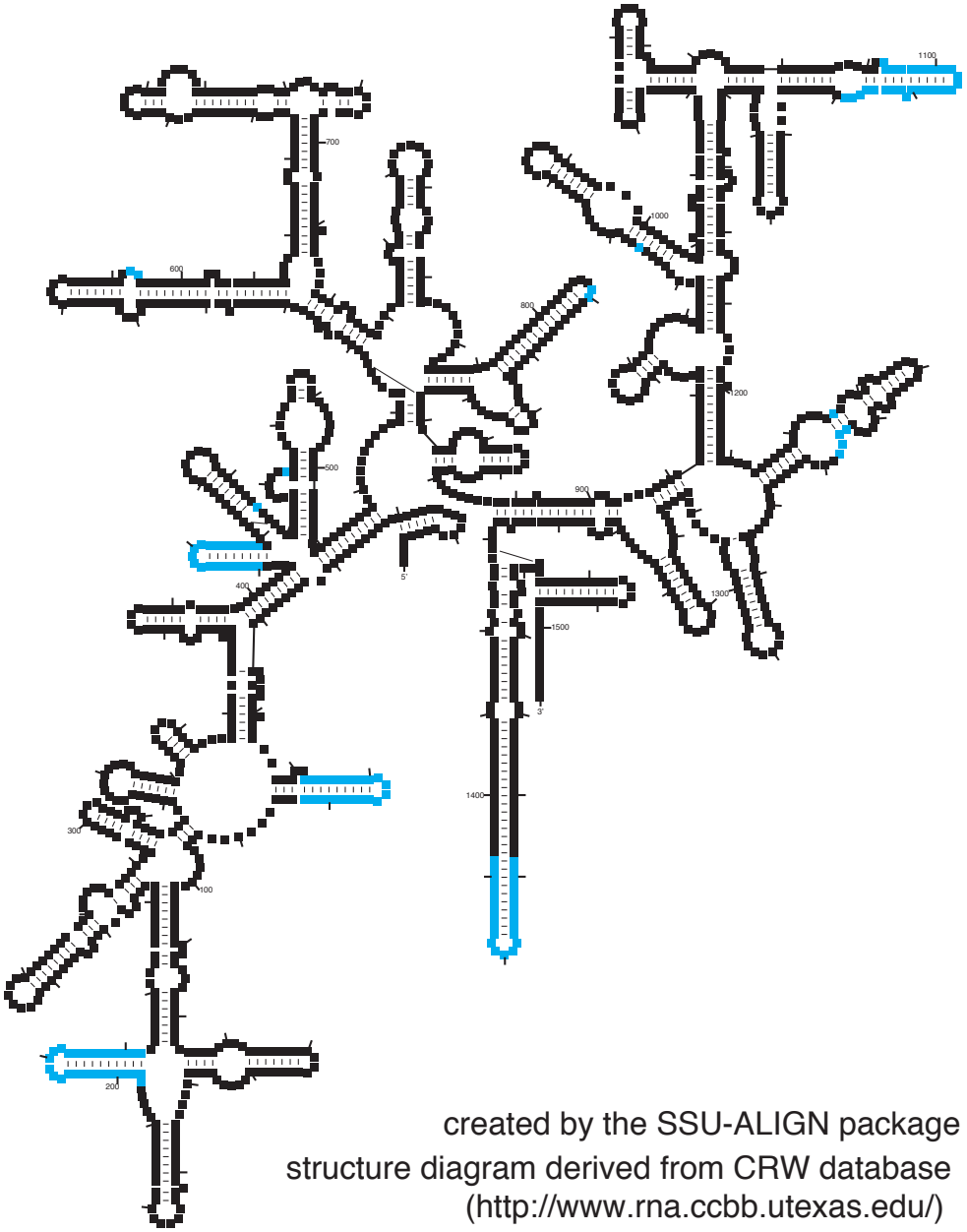


created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

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

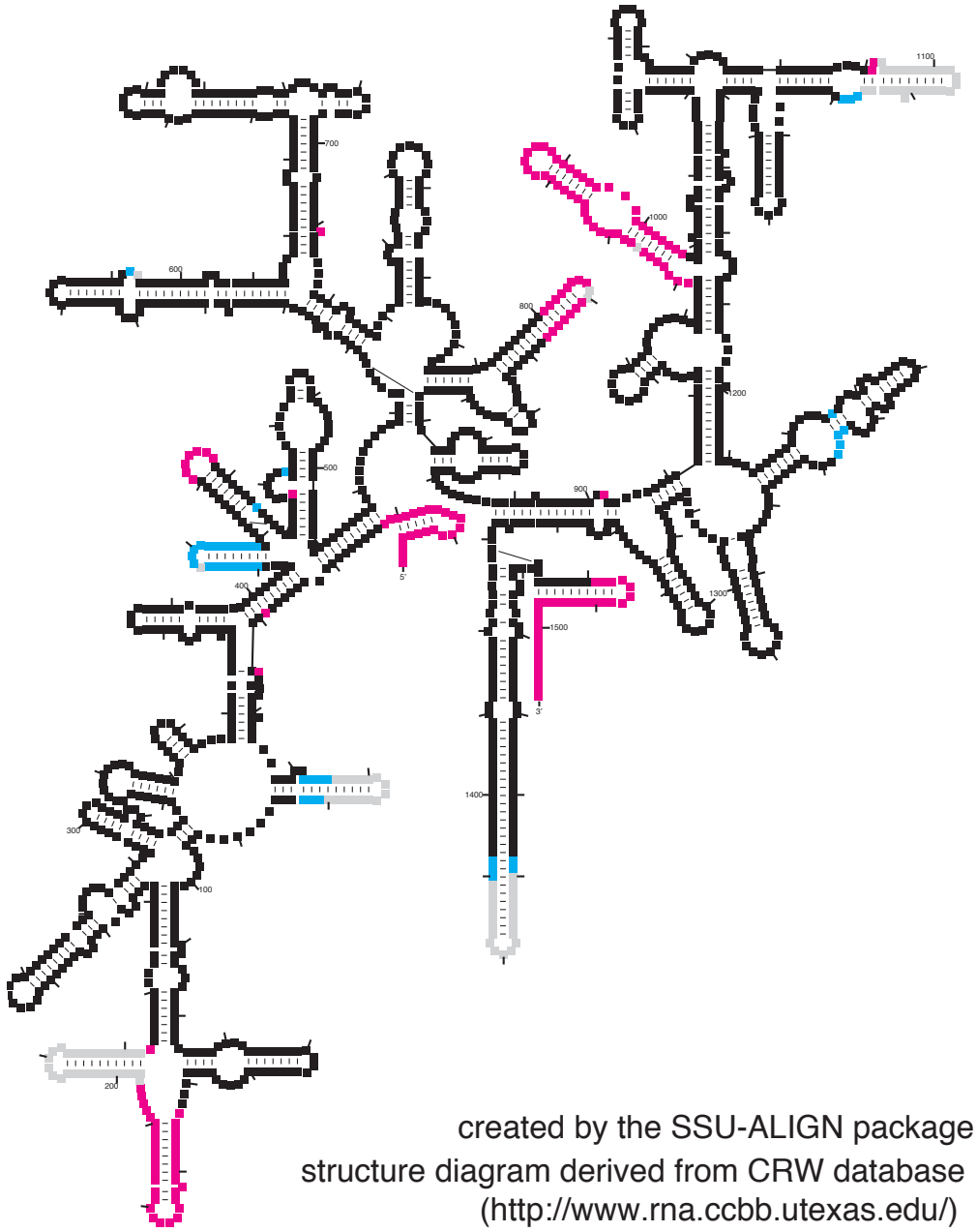


Posterior probability based archaeal SSU mask
black: included in alignment (1376)
blue: excluded from alignment (132)



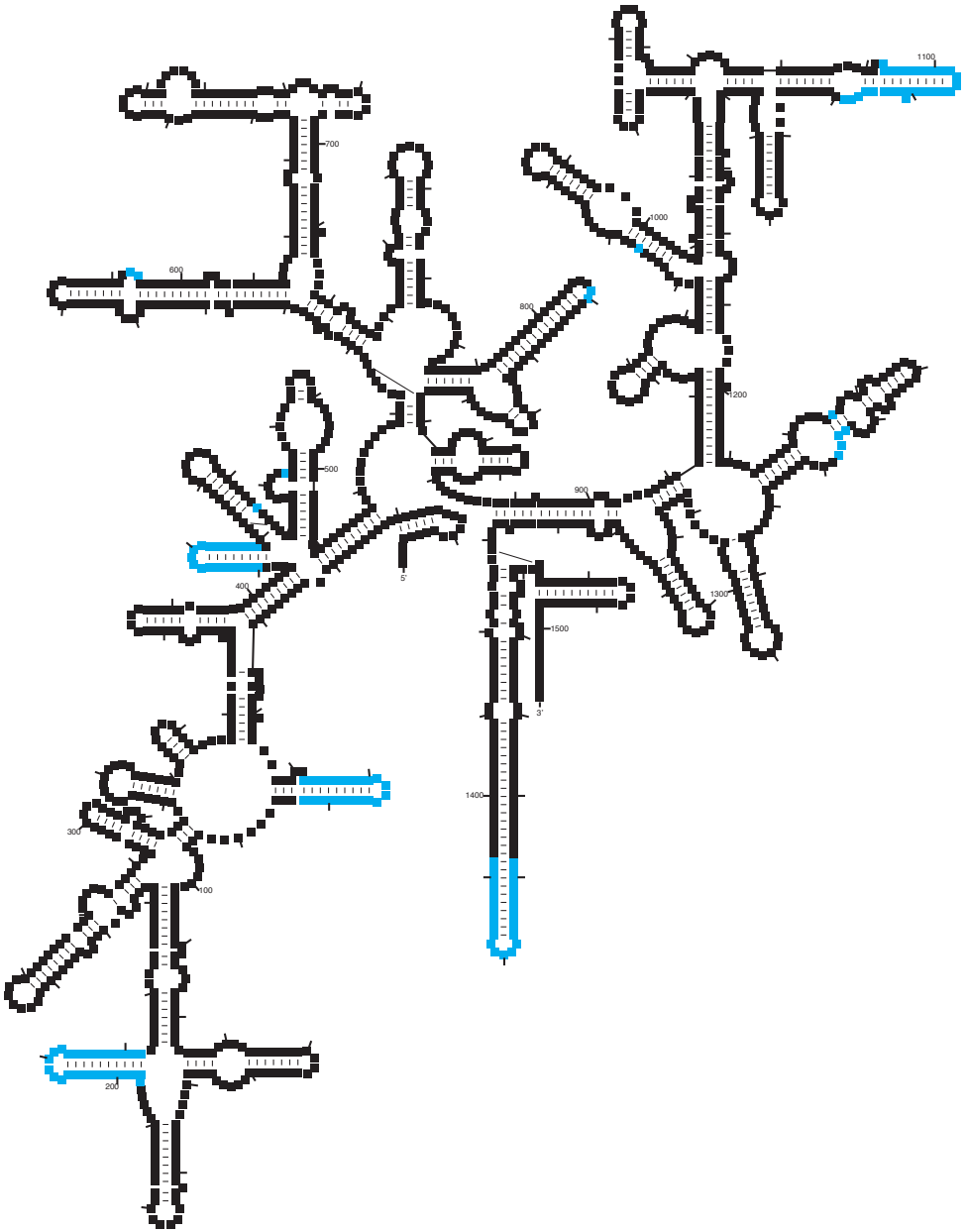
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

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)

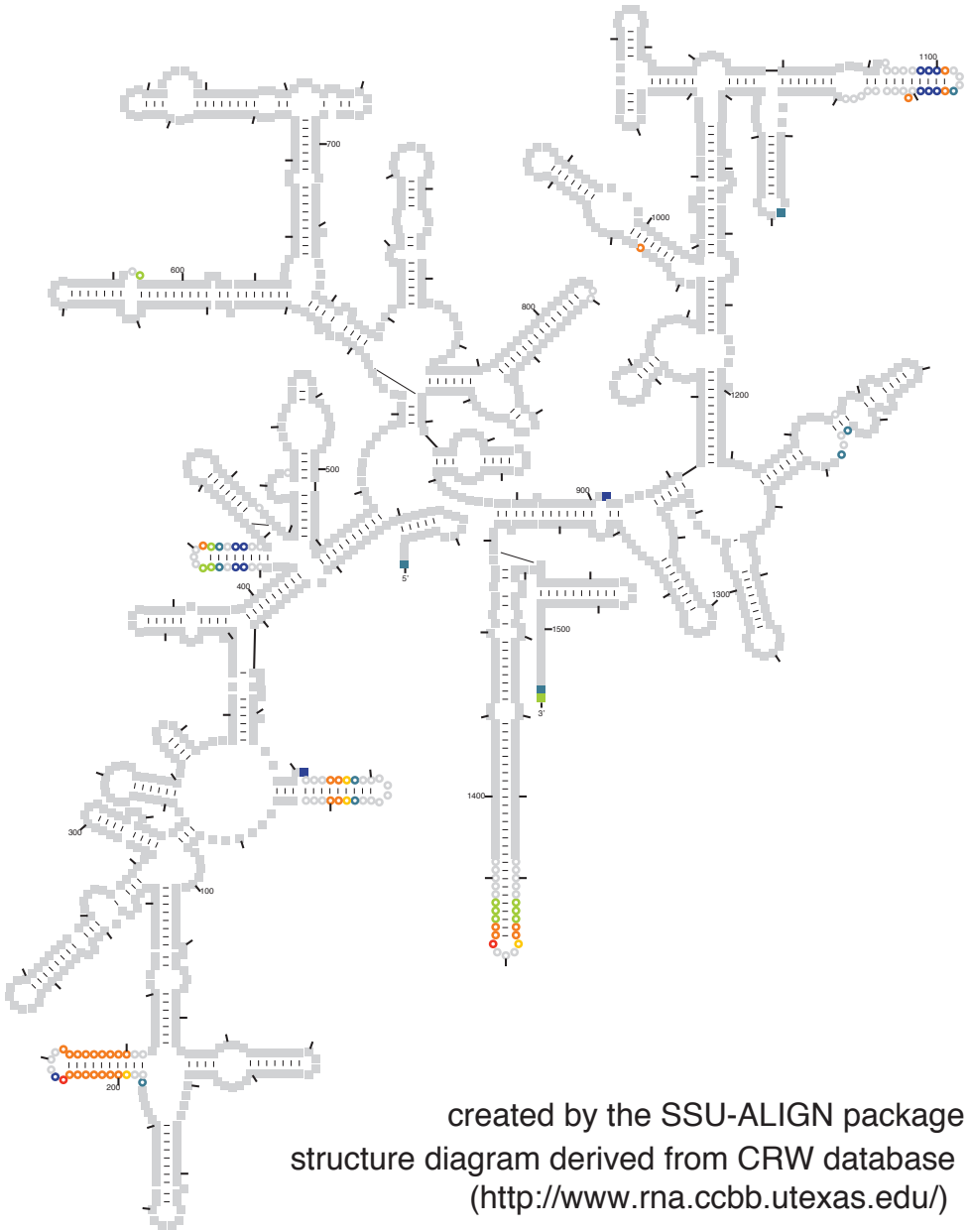


created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

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



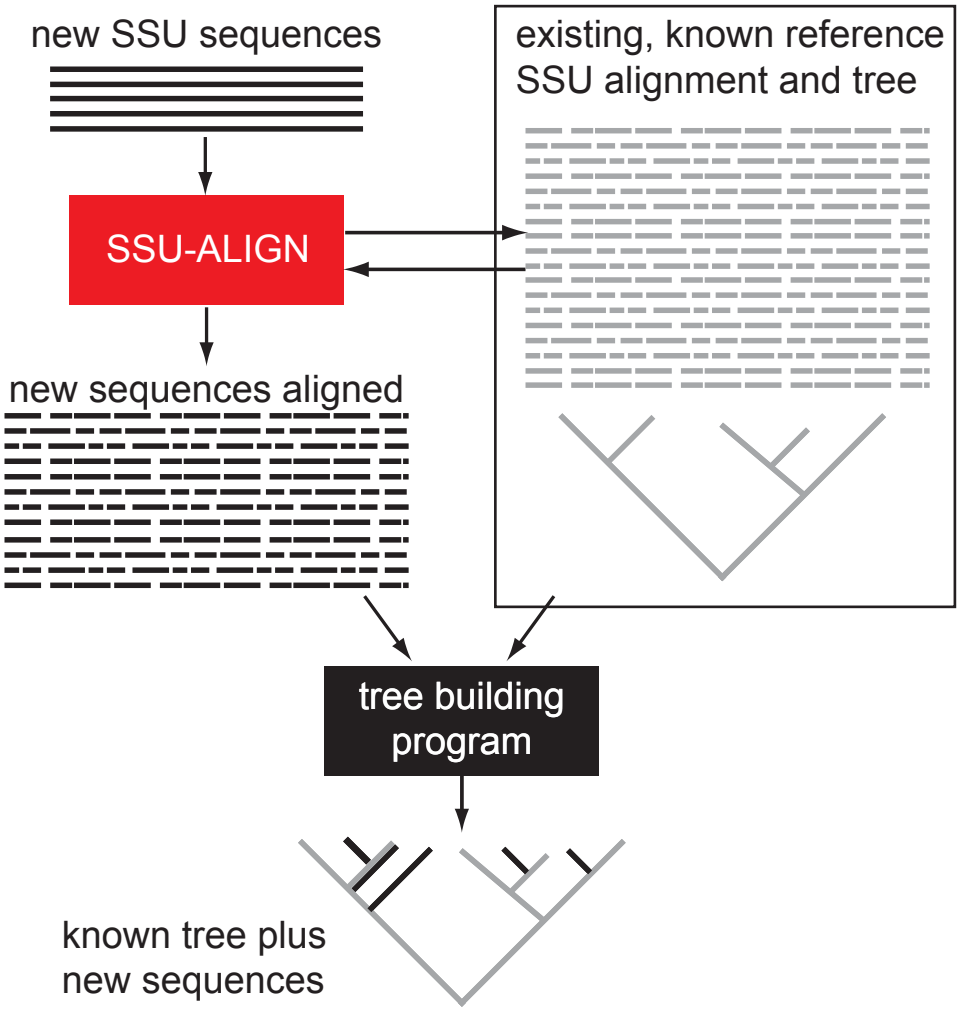
created by the SSU-ALIGN package
structure diagram derived from CRW database
(http://www.rna.ccbb.utexas.edu/)

Automated masking removes the majority of alignment errors

	alignment accuracy	time (sec/seq)
Muscle-3.8.31*	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

Acknowledgements

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