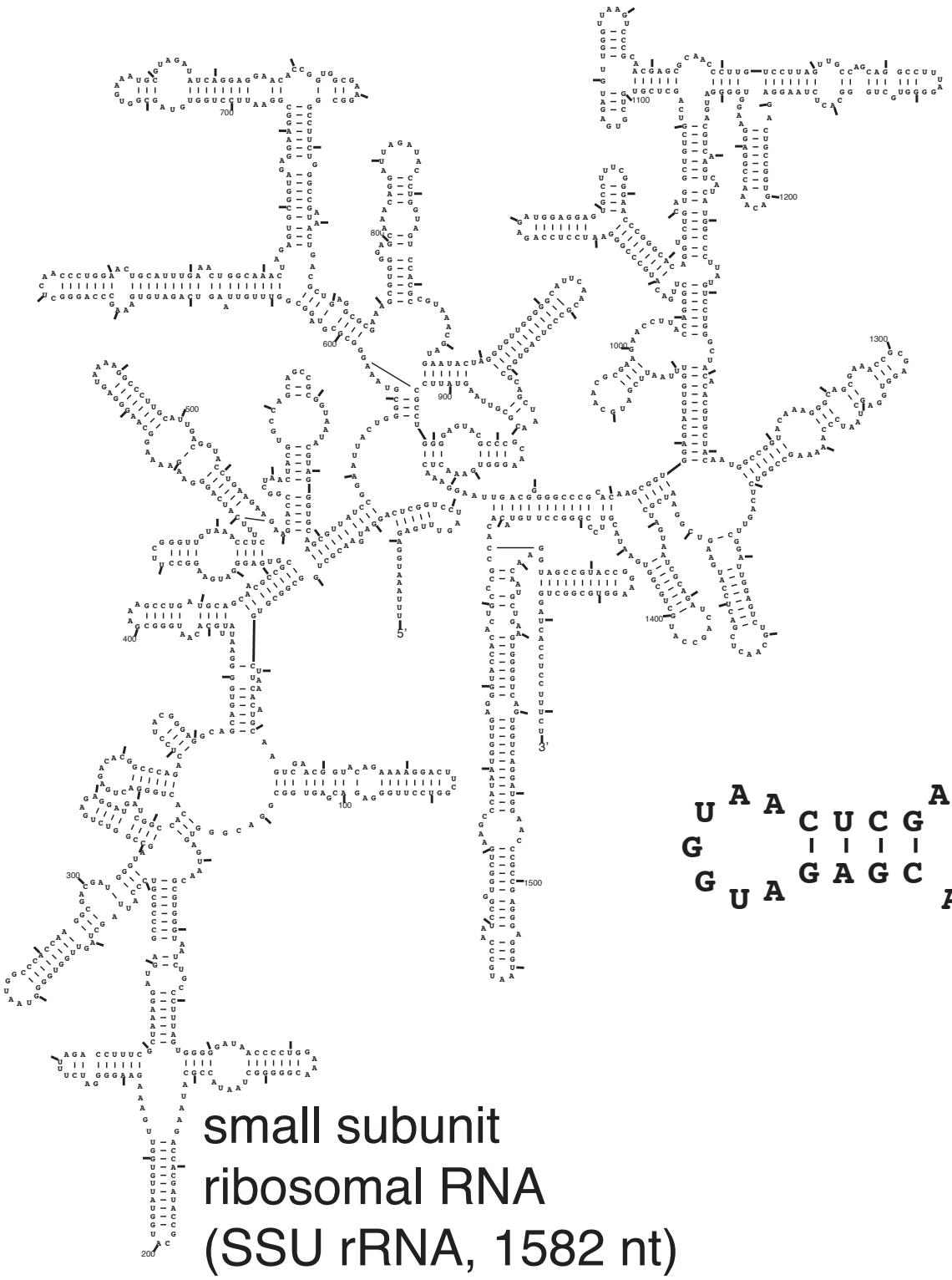
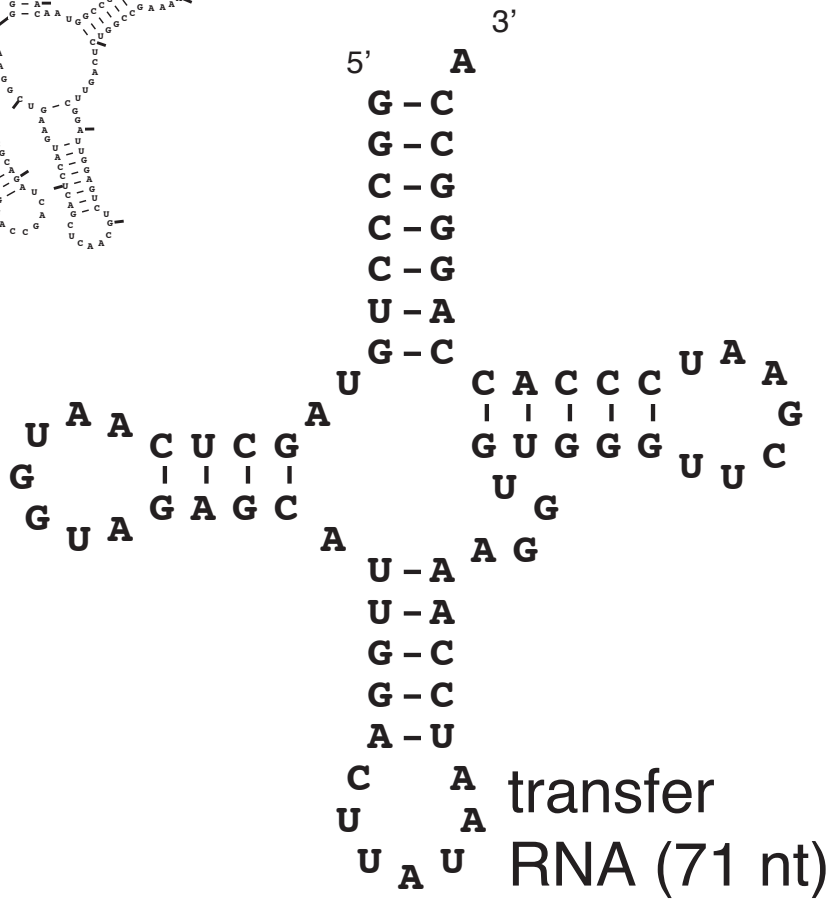
Computational identification of structural RNA homologs using Infernal and Rfam

Eric Nawrocki

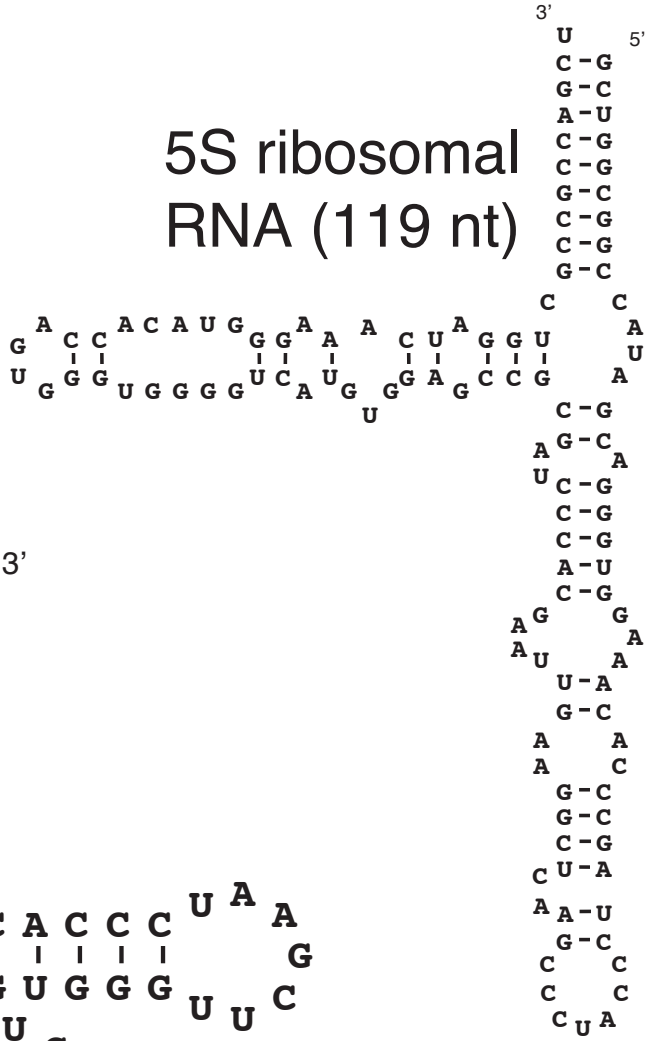
National Center for Biotechnology Information
National Institutes of Health
Bethesda, MD, USA



small subunit ribosomal RNA (SSU rRNA, 1582 nt)



transfer RNA (71 nt)



5S ribosomal RNA (119 nt)

Functional RNAs play many vital roles in the cell

	key RNAs involved	archaea	bacteria	eukarya
translation	ribosomal RNAs	x	x	x
	transfer RNAs	x	x	x
	RNase P RNA	x	x	x
	snoRNAs	x		x
	SRP RNA	x	x	x
	tmRNA		x	
	RNaseMRP			x
gene expression	riboswitches	?	x	?
	microRNAs			x
	6S RNA		x	x
splicing	U1, U2, U4, U5, U6			x
other	telomerase RNA			x
	Y RNA		x	x
	Vault RNA			x
	many more...			

Functional RNAs play many vital roles in the cell

	key RNAs involved	archaea	bacteria	eukarya
translation	ribosomal RNAs	x	x	x
	transfer RNAs	x	x	x
	RNase P RNA	x	x	x
	snoRNAs	x		x
	SRP RNA	x	x	x
	tmRNA		x	
	RNaseMRP			x
gene expression	riboswitches	?	x	?
	microRNAs			x
	6S RNA		x	x
splicing	U1, U2, U4, U5, U6			x
other	telomerase RNA			x
	Y RNA			x
	Vault RNA			x
	many more...			

database of more than 2600 non-coding RNA families
each represented by a secondary structure, alignment, and covariance model.

Outline of talk

1. Motivation: collecting homologs facilitates comparative sequence analysis.
1965: Secondary structure determination of transfer RNA.

2. Sequence and sequence+structure profiles

3. Accelerating RNA homology search

4. Implications for Rfam

5. Y RNAs

Structure of a Ribonucleic Acid

ROBERT W. HOLLEY, JEAN APGAR
GEORGE A. EVERETT
JAMES T. MADISON
MARK MARQUISEE, SUSAN H. MERRILL
JOHN ROBERT PENSWICK, ADA ZAMIR
U.S. Plant, Soil, and Nutrition Laboratory, U.S. Department of Agriculture, and
Department of Biochemistry, Cornell University, Ithaca, New York

Abstract. *The complete nucleotide sequence of an alanine transfer RNA, isolated from yeast, has been determined. This is the first nucleic acid for which the structure is known.*

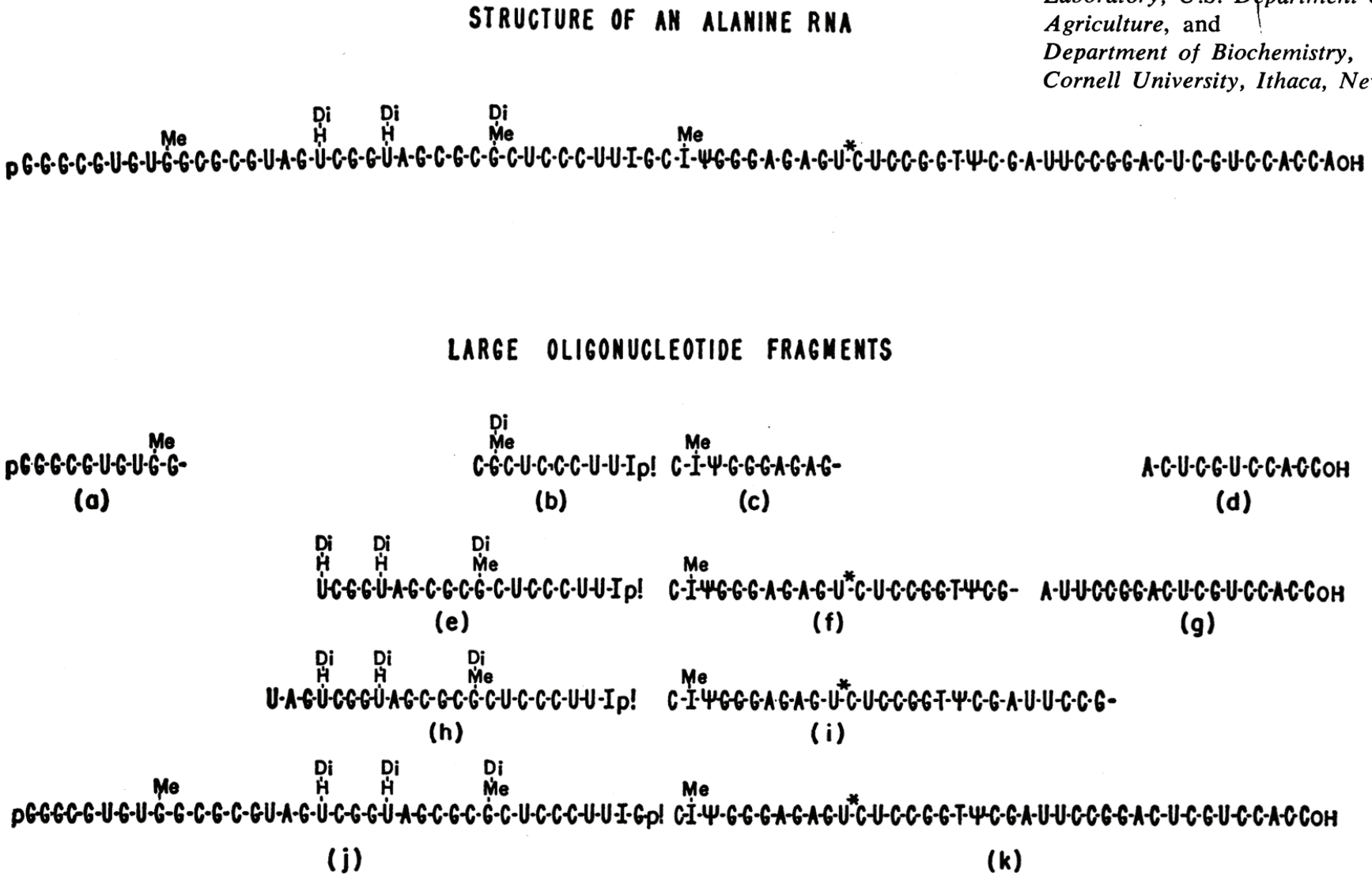


Fig. 1. The structure of an alanine transfer RNA, isolated from yeast, is shown at the top. Large oligonucleotide fragments that were crucial in the proof of structure are shown below.

Holley, Robert W., et al. "Structure of a ribonucleic acid." Science 147.3664 (1965): 1462-1465.

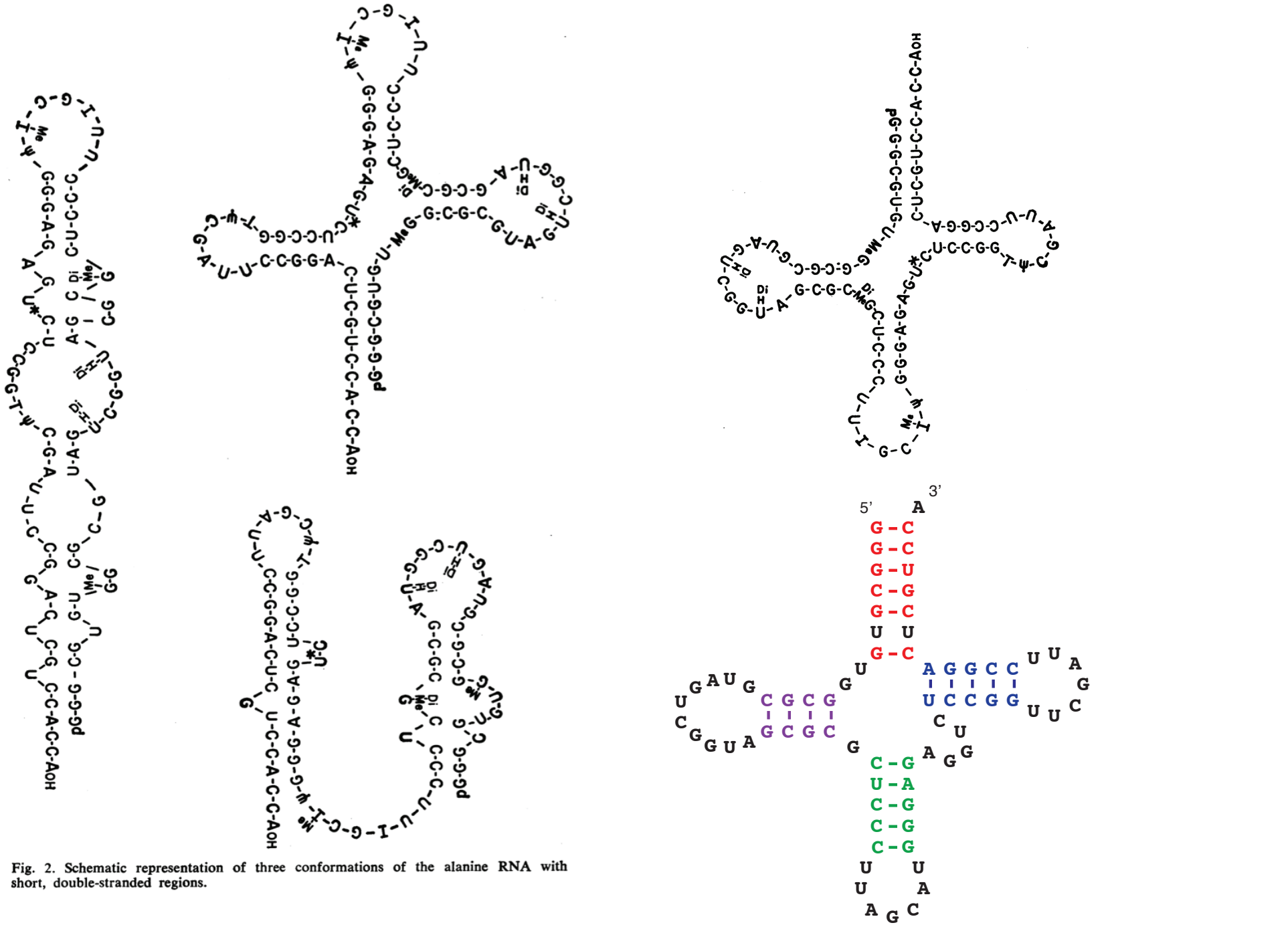
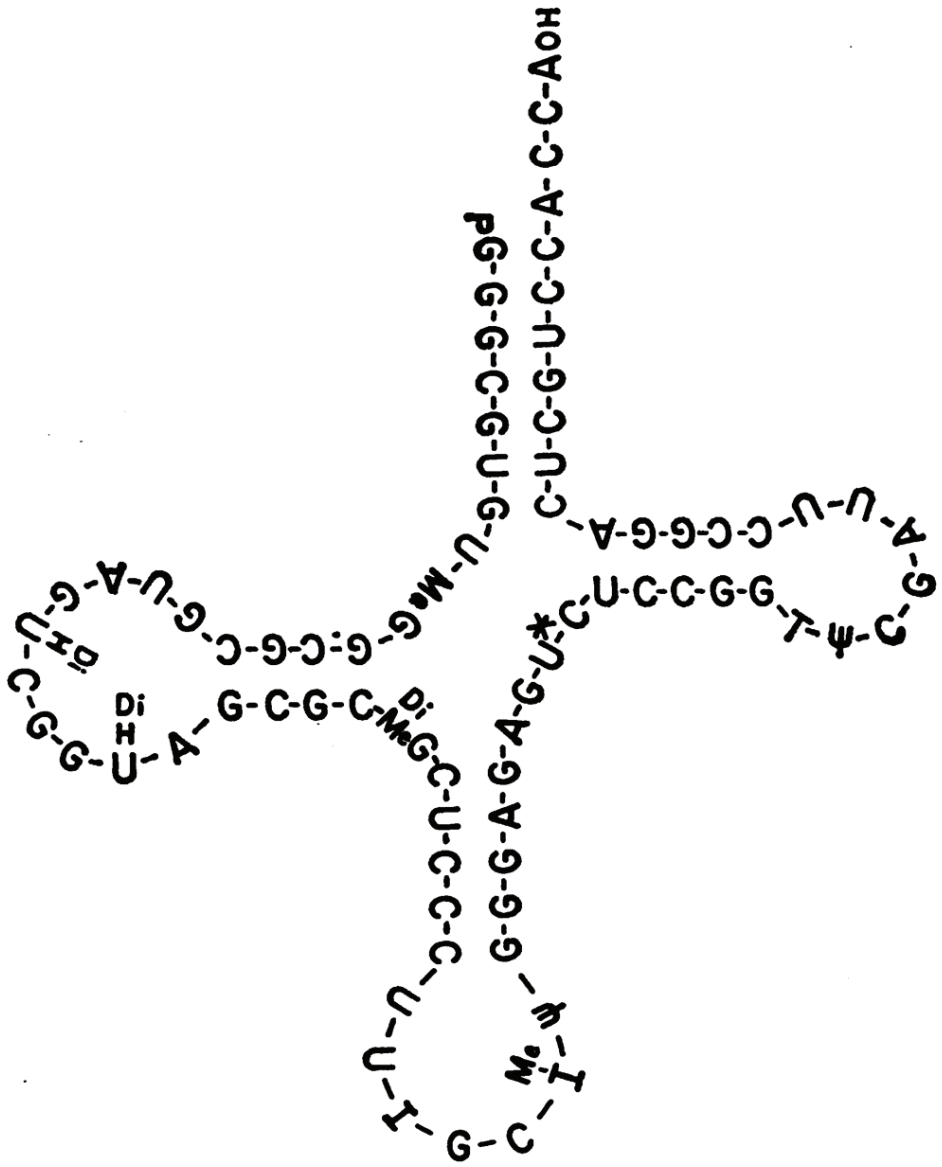


Fig. 2. Schematic representation of three conformations of the alanine RNA with short, double-stranded regions.

Holley, Robert W., et al. "Structure of a ribonucleic acid." Science 147.3664 (1965): 1462-1465.

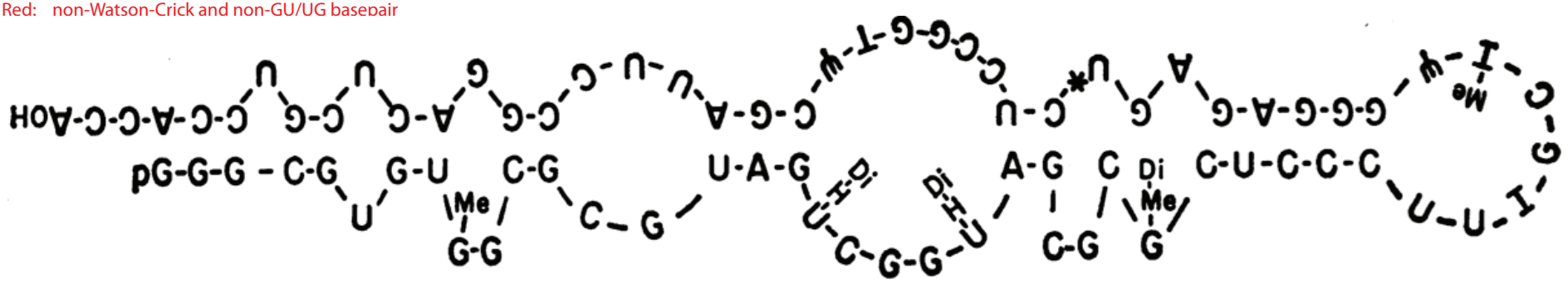
```
struct     (((((((..<<<<............>>>>.<<<<<.......>>>>>.....<<<<<.......>>>>>)))))))
Ala        GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr        CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGA
Val        GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln        GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu        UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CGGGGUUCGACUCCCCGUAUCGGAG
identical          * *     * * ** ***   *       *   *          **     ******  *
>0 non-WC  ((((((..<<<<............>>>>.<<<<<.......>>>>>.....<<<<<.......>>>>>))))))
```

Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair
Grey: not basepaired



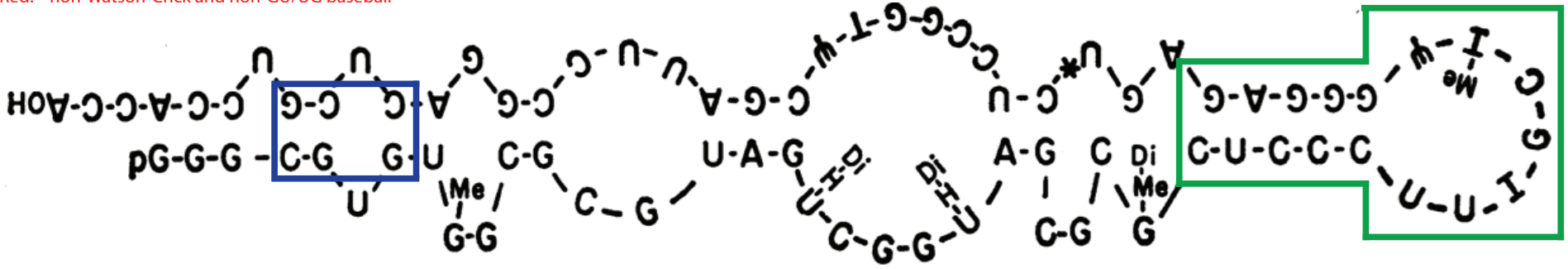
```
struct      <<<<<.<<..<<..<<<.......<<..<.<<<<<.......>>>>>.>..>>......>>>...>>.>>.>>.>>>
Ala         GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr         CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCGGGAGA
Val         GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln         GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu         UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGAUCUCCCGUAUCGGAG
identical           * *     * * ** ***   *        *   *          **    *******  *
>0 non-WC   <<<<<.<<..<<..<<<.......<<..<.<<<<<.......>>>>>.>..>>......>>>...>>.>>.>>.>>>
```

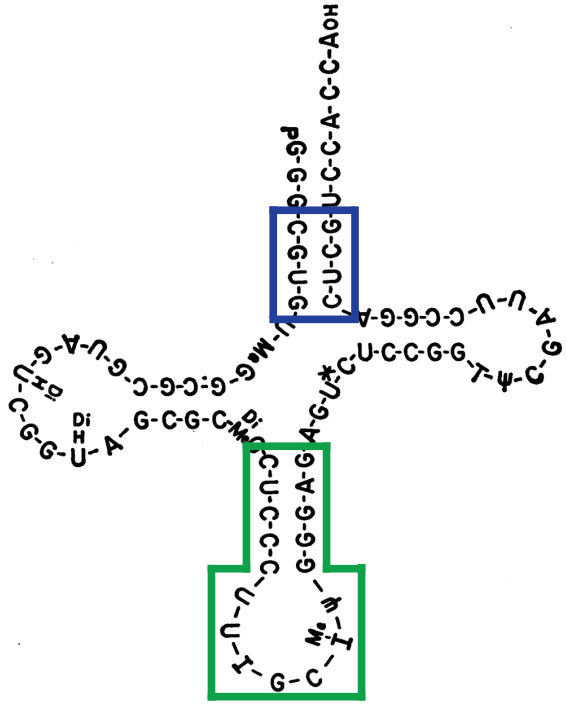
Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair



```
struct      <<<<<.<<..<<..<<<.......<<..<.<<<<<.......>>>>>.>..>>......>>>...>>.>>.>>.>>>
Ala         GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr         CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCGGGAGA
Val         GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln         GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu         UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGAUCUCCCGUAUCGGAG
identical         * *      * * ** ***   *        *   *           **     *******   *
>0 non-WC   <<<<<.<<..<<..<<<.......<<..<.<<<<<.......>>>>>.>..>>......>>>...>>.>>.>>.>>>
clover      << <                      <<<<<.......>>>>>                        > >>
overlap
```

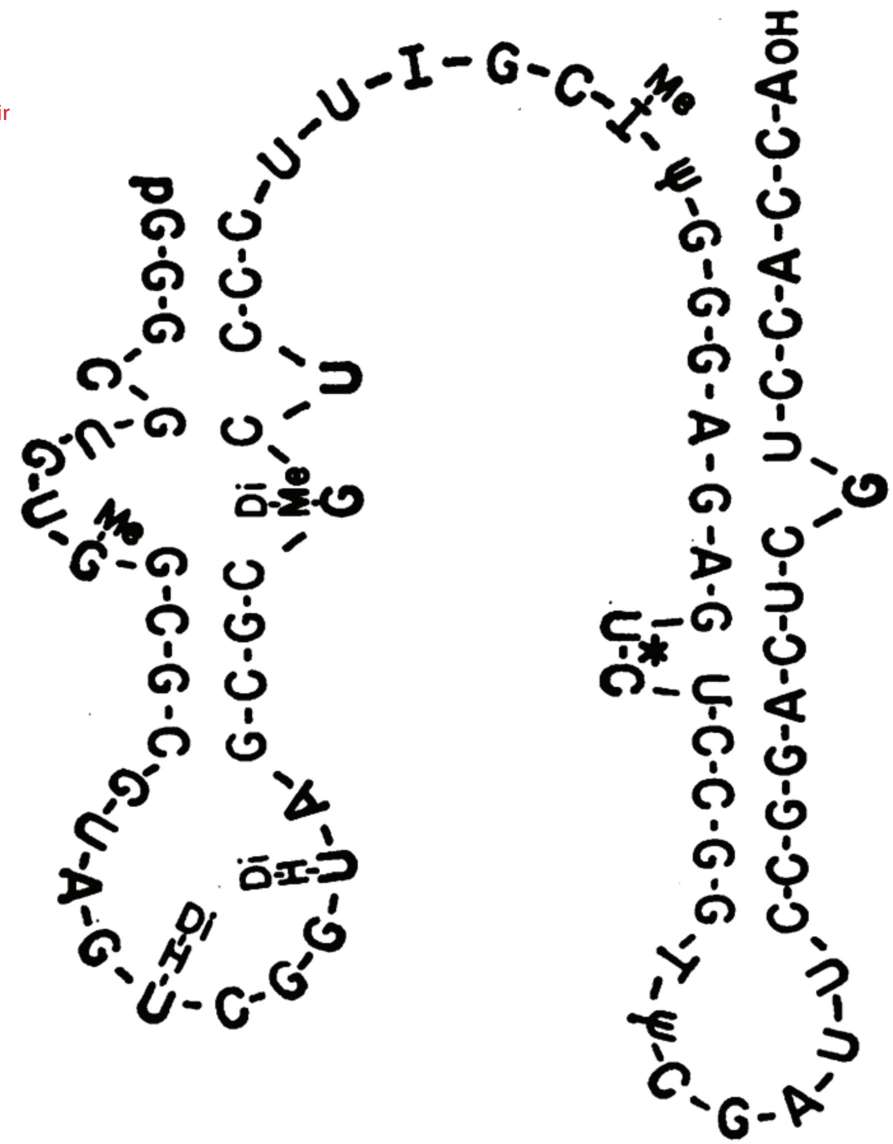
Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair





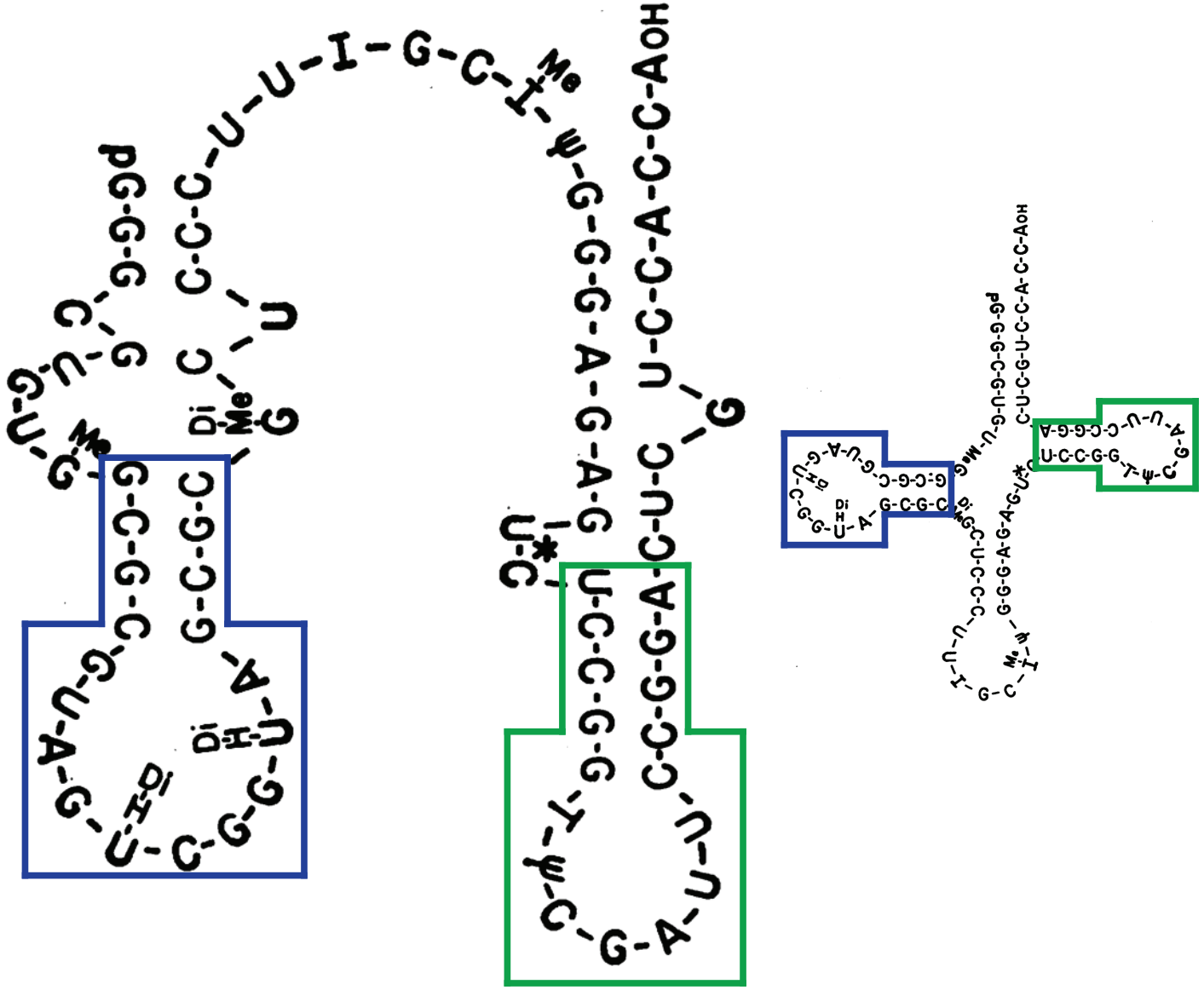
```
struct      <<<.<....<<<<............>>>>.>.>>>........<<<<<<...<<<<<.......>>>>>>>>.>>>.
Ala         GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr         CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGA
Val         GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln         GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu         UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCCGUAUCGGAG
identical          * *     * * ** ***   *       *   *         **    ******  *

>0 non-WC   <<<.<....<<<<............>>>>.>.>>>........<<<<<<...<<<<<.......>>>>>>>>.>>>.
```

Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair
Grey: not basepaired



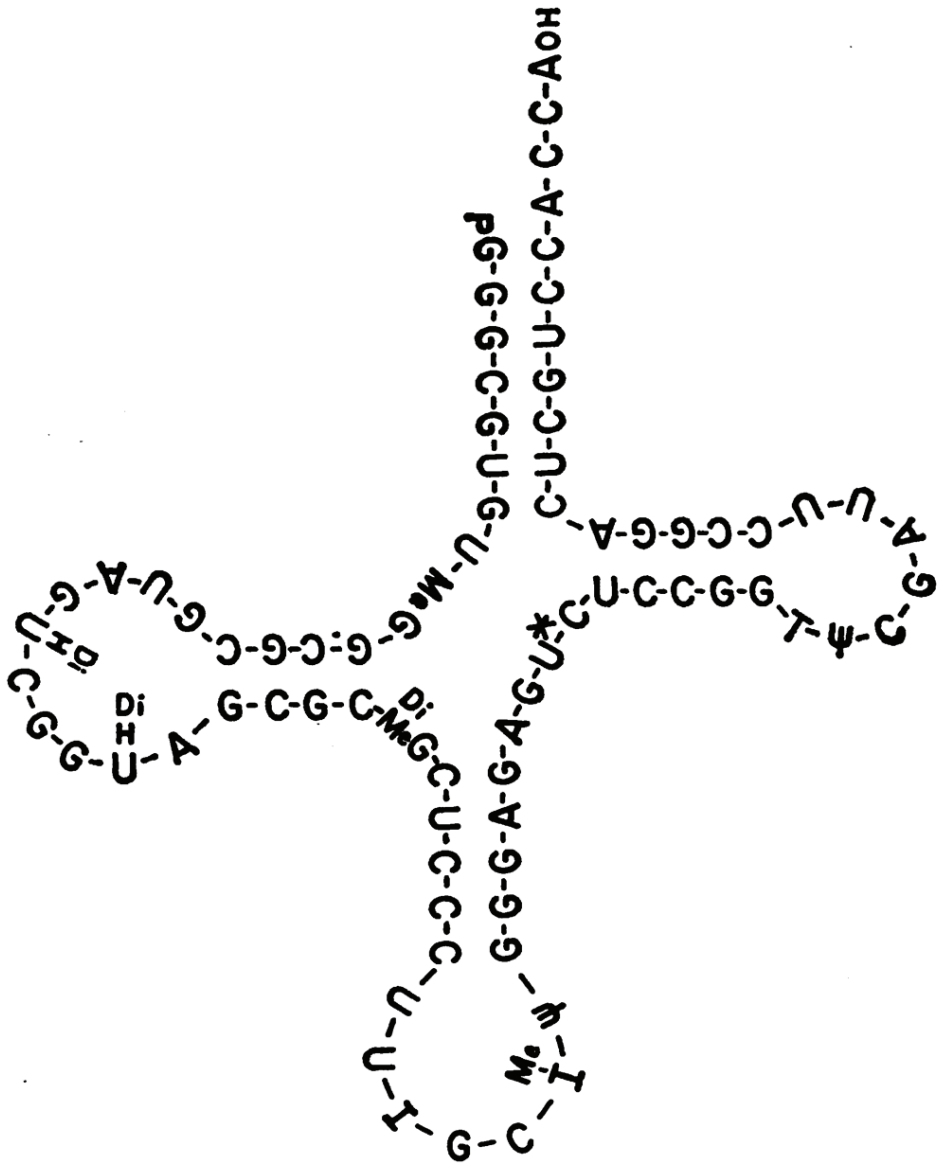
```
struct      <<<.<....<<<<.............>>>>.>.>>>........<<<<<...<<<<.......>>>>>>>>.>>>.
Ala         GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr         CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGA
Val         GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln         GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu         UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCCGUAUCGGAG
identical           * *     * * ** ***   *       *   *           **     ******   *
>0 non-WC   <<<.<....<<<<.............>>>>.>.>>>........<<<<<...<<<<.......>>>>>>>>.>>>.
clover                <<<<............>>>>                    <<<<.......>>>>>
overlap
```

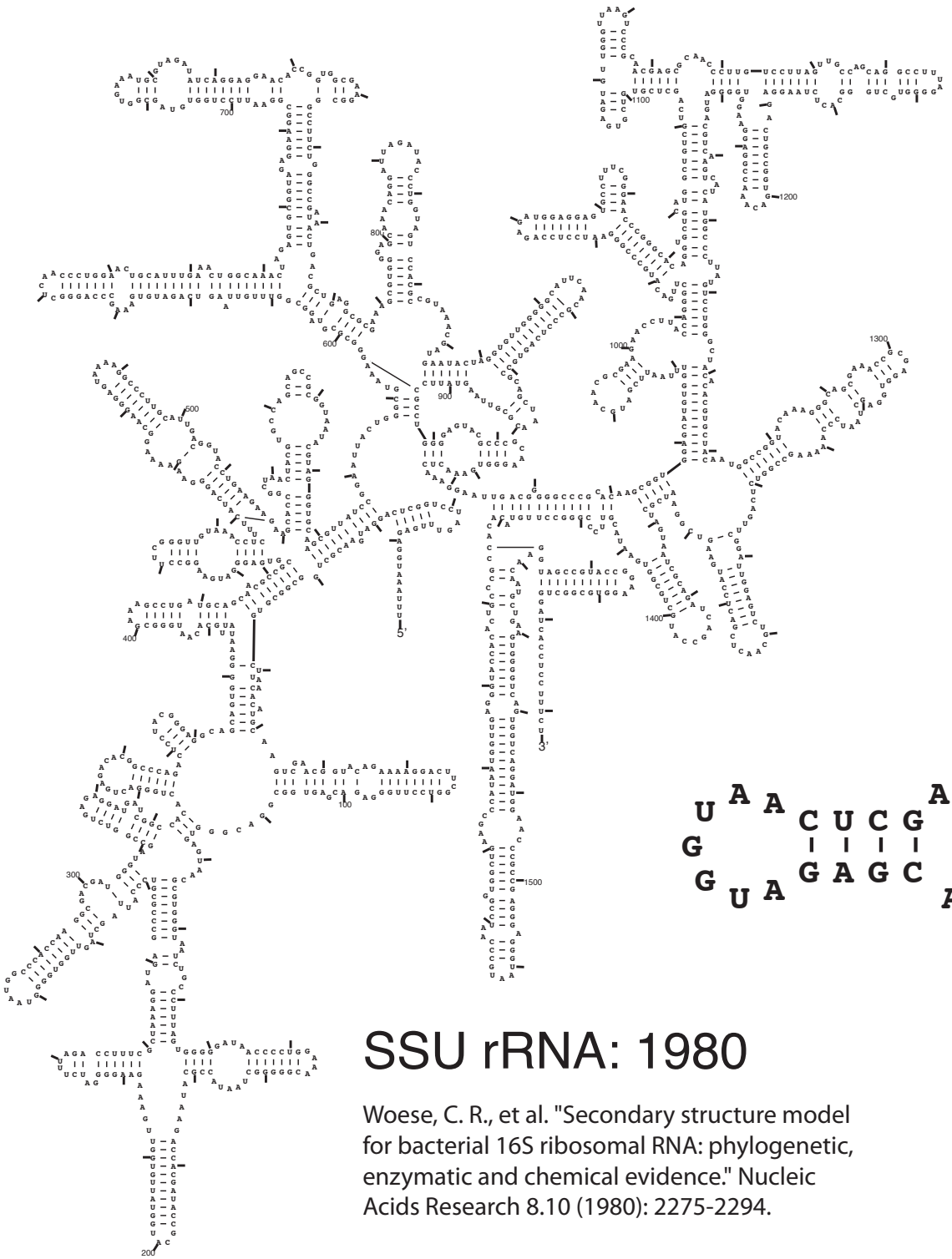
Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair
Grey: not basepaired



```
struct     (((((((..<<<<............>>>>.<<<<<.......>>>>>.....<<<<<.......>>>>>)))))))
Ala        GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr        CUCUCGGUAGCCA..AGUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGA
Val        GGUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln        GGUCUCUUGGCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu        UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CGGGGUUCGACUCCCCGUAUCGGAG
identical          * *     * * ** ***   *       *   *          **     ******  *
>0 non-WC  ((((((..<<<<............>>>>.<<<<<.......>>>>>.....<<<<<.......>>>>>))))))
```

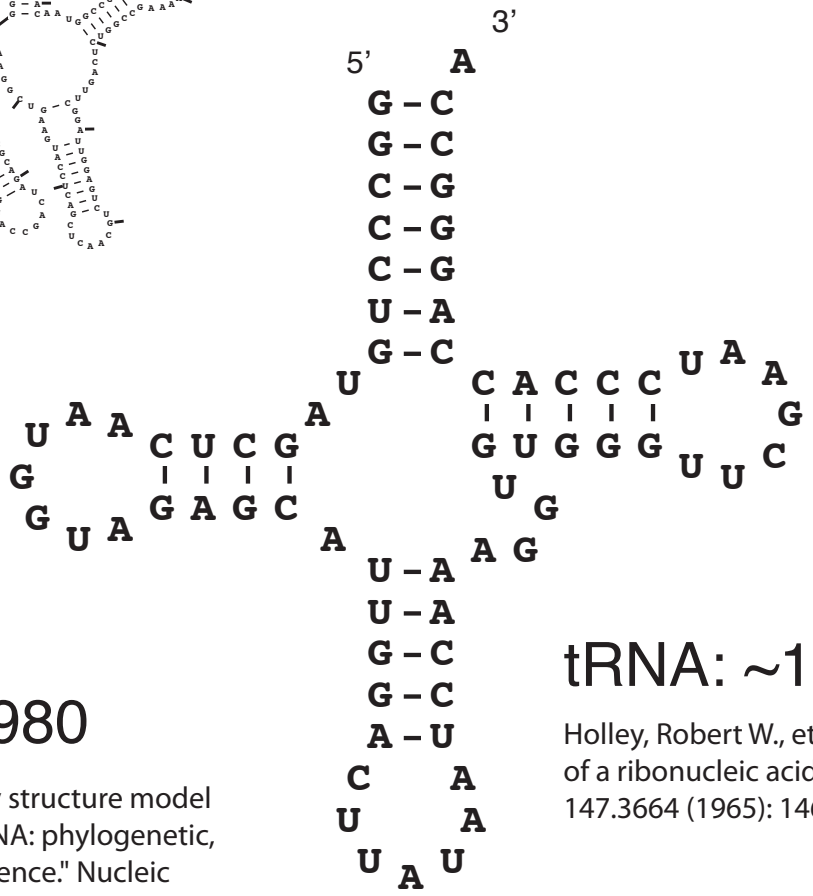
Alignment color legend:
Black: Watson-Crick or GU/UG basepair
Red: non-Watson-Crick and non-GU/UG basepair
Grey: not basepaired





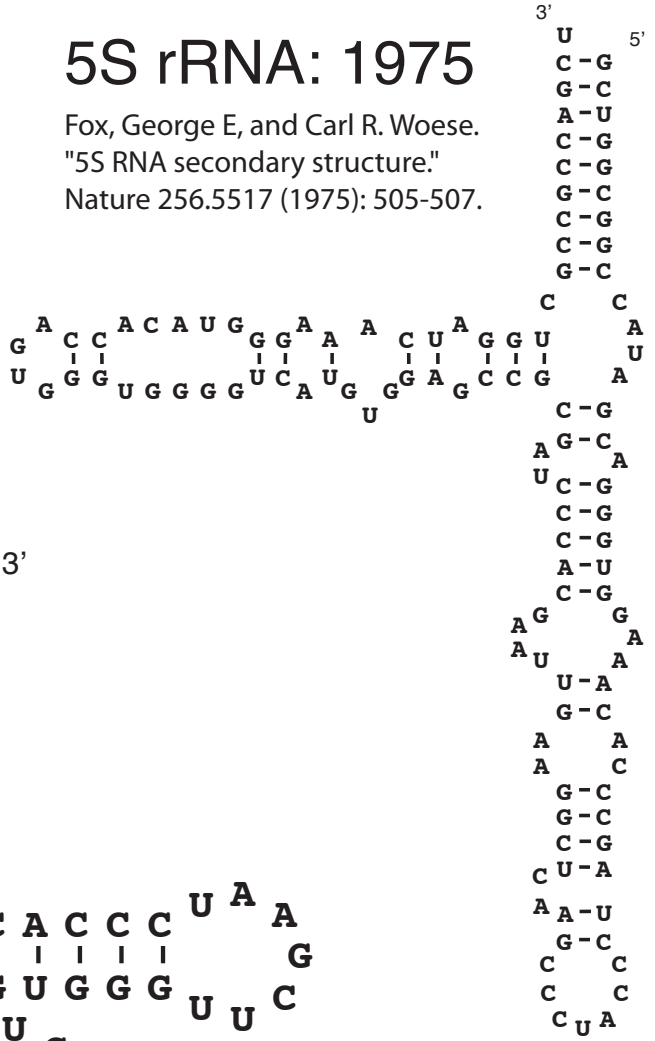
SSU rRNA: 1980

Woese, C. R., et al. "Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence." Nucleic Acids Research 8.10 (1980): 2275-2294.



tRNA: ~1966

Holley, Robert W., et al. "Structure of a ribonucleic acid." Science 147.3664 (1965): 1462-1465.



5S rRNA: 1975

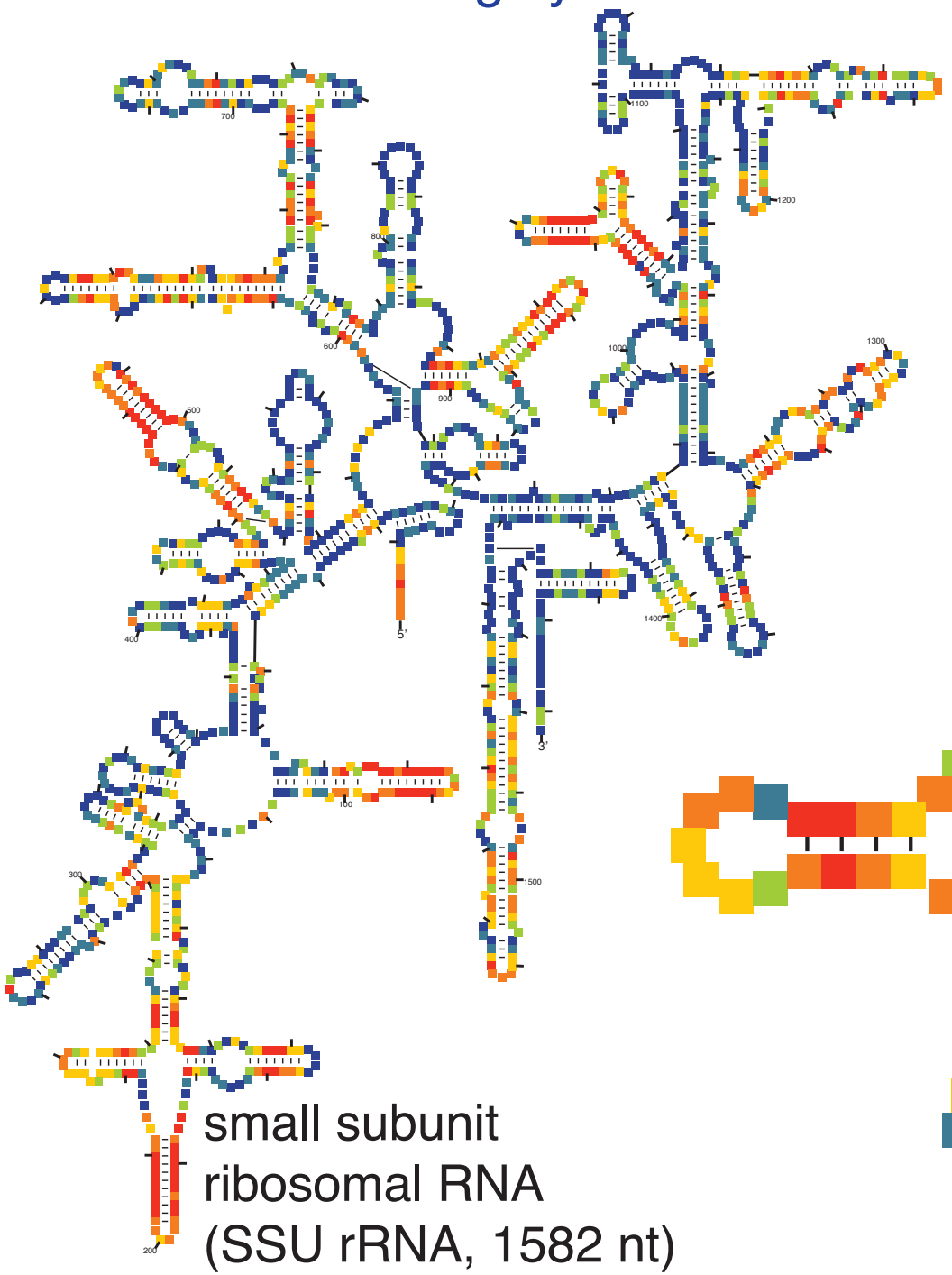
Fox, George E, and Carl R. Woese. "5S RNA secondary structure." Nature 256.5517 (1975): 505-507.

Comparative sequence analysis of homologs informs biologists

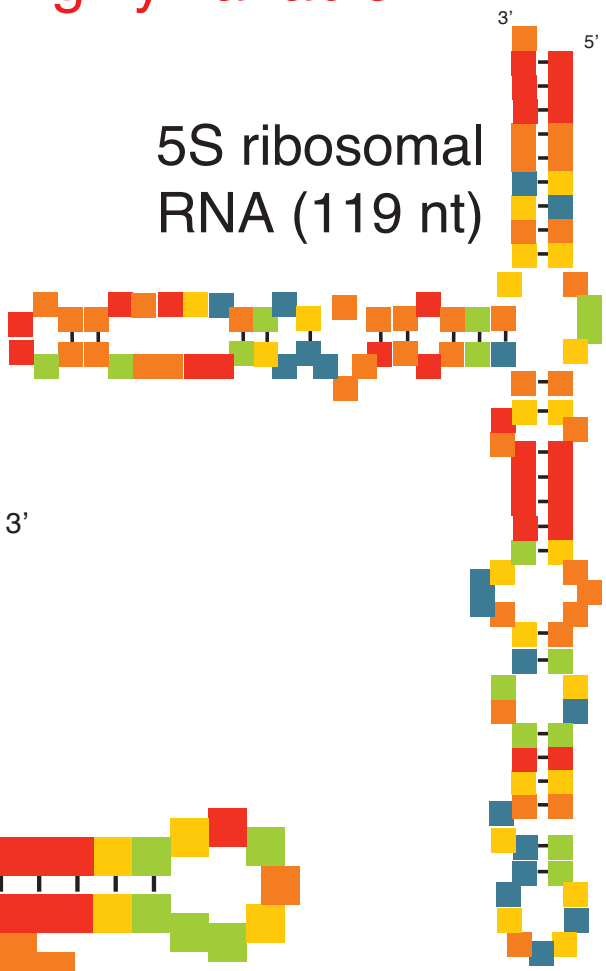
- Inference of structure
- Inference of phylogeny of organisms
- Inference of functional regions based on conservation levels

Sequence conservation per position

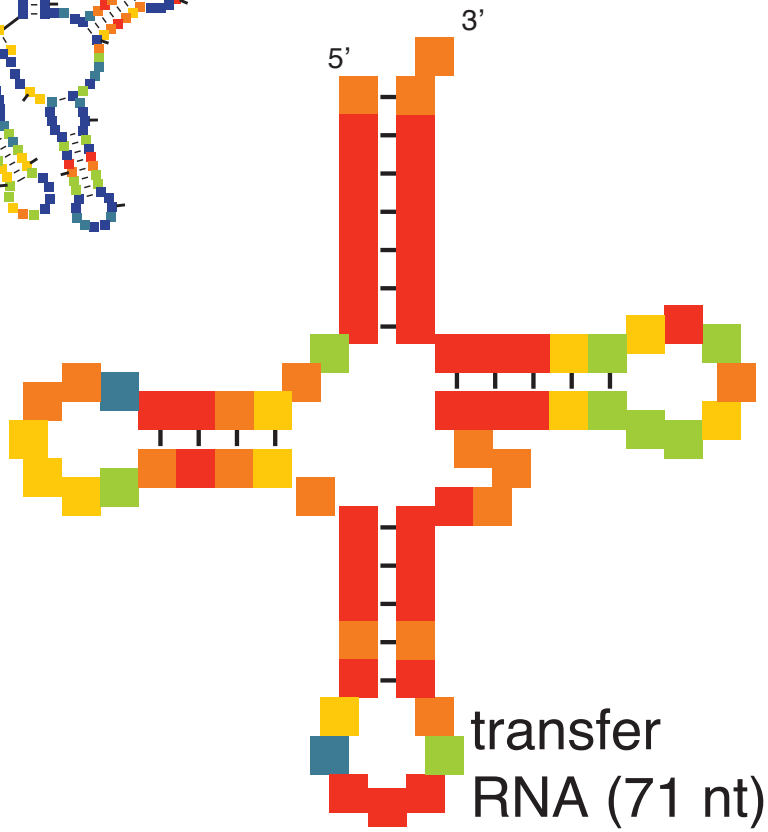
blue:highly conserved ..... red: highly variable



small subunit ribosomal RNA (SSU rRNA, 1582 nt)



5S ribosomal RNA (119 nt)



transfer RNA (71 nt)

Comparative sequence analysis of homologs informs biologists

- Inference of structure
- Inference of phylogeny of organisms
- Inference of functional regions based on conservation levels

Computational homology search methods use one or more known family members to find additional homologs.

Outline of talk

1. Motivation: collecting homologs facilitates comparative sequence analysis.
1965: Secondary structure determination of transfer RNA.

2. Sequence and sequence+structure profiles

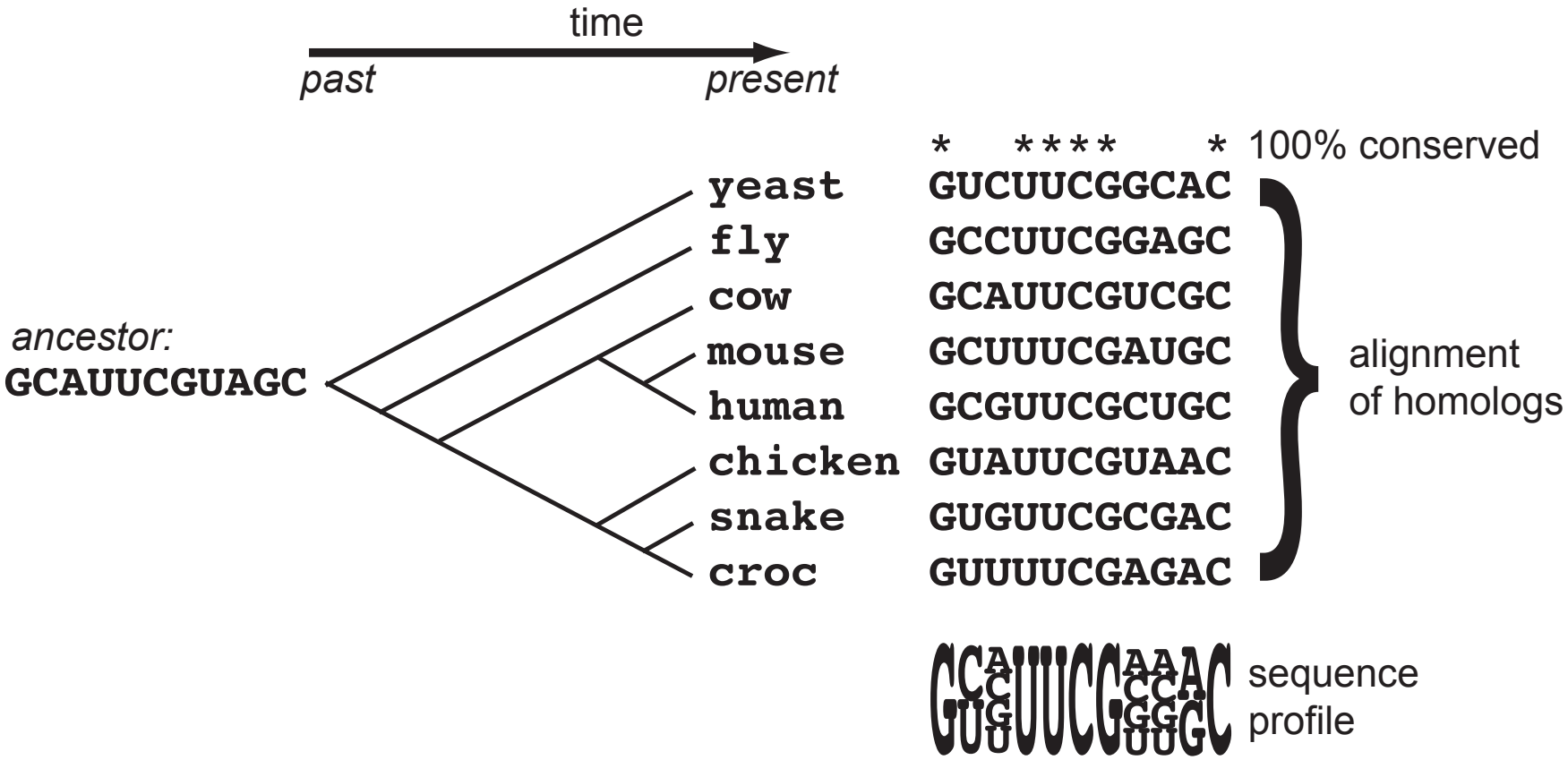
3. Accelerating RNA homology search

4. Implications for Rfam

5. Y RNAs

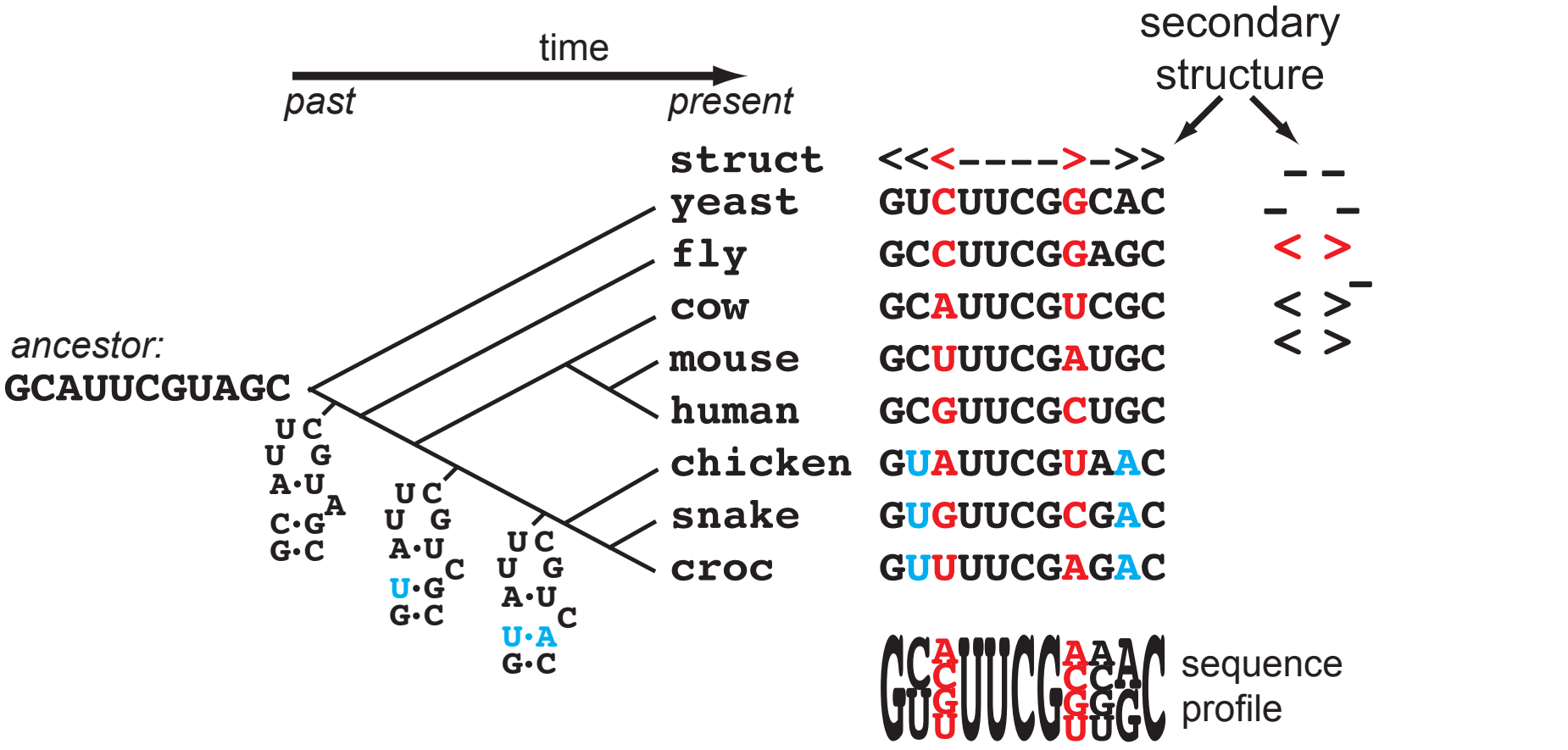
Sequence conservation provides information for homology searches

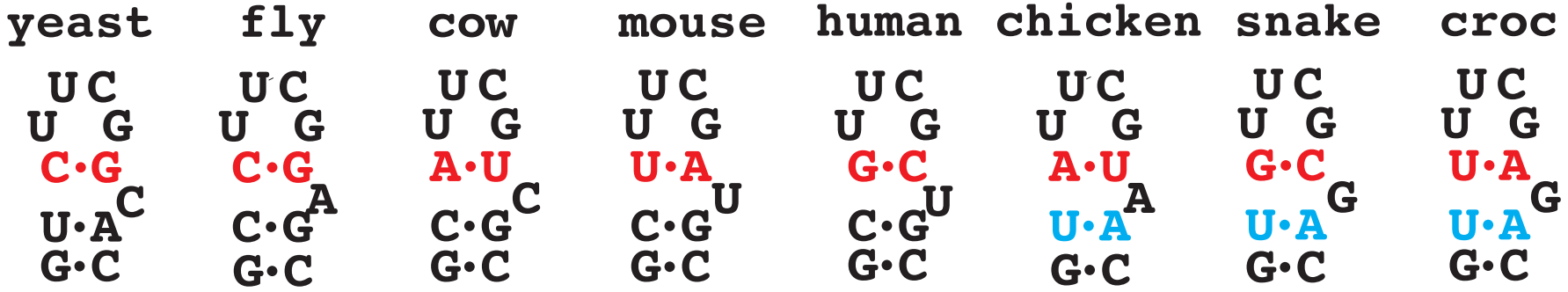
Conservation levels vary across alignment columns.



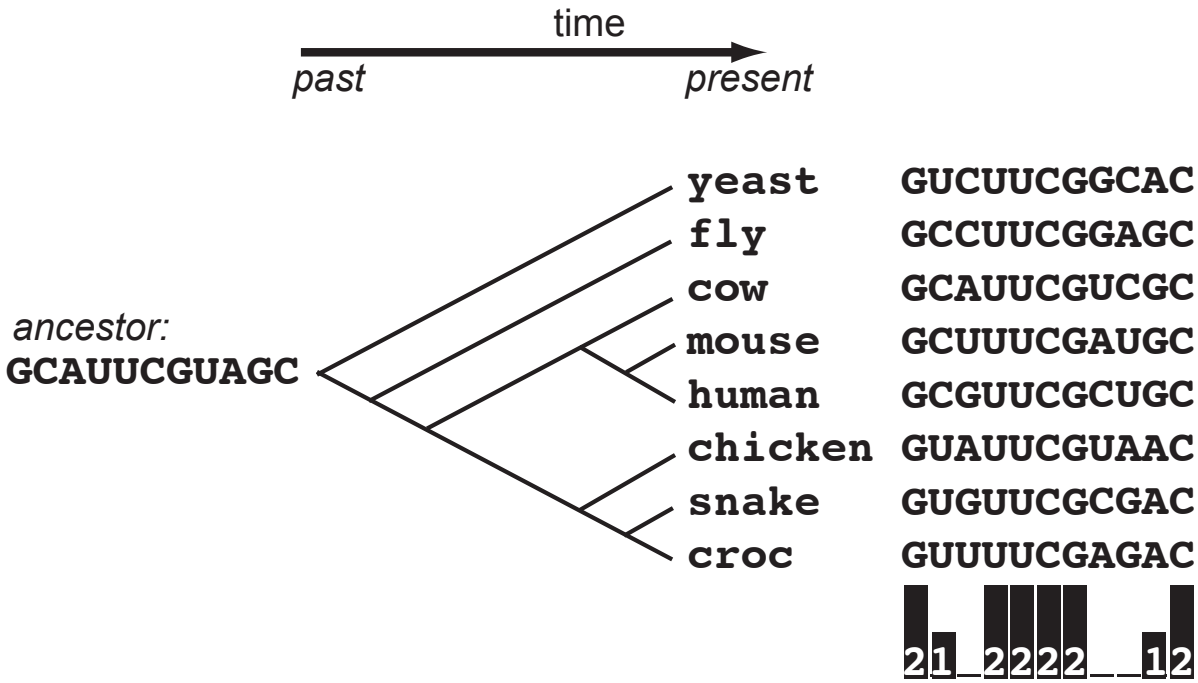
Structure conservation provides additional information

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





Amount of information in a profile can be measured in bits



Each column is independent

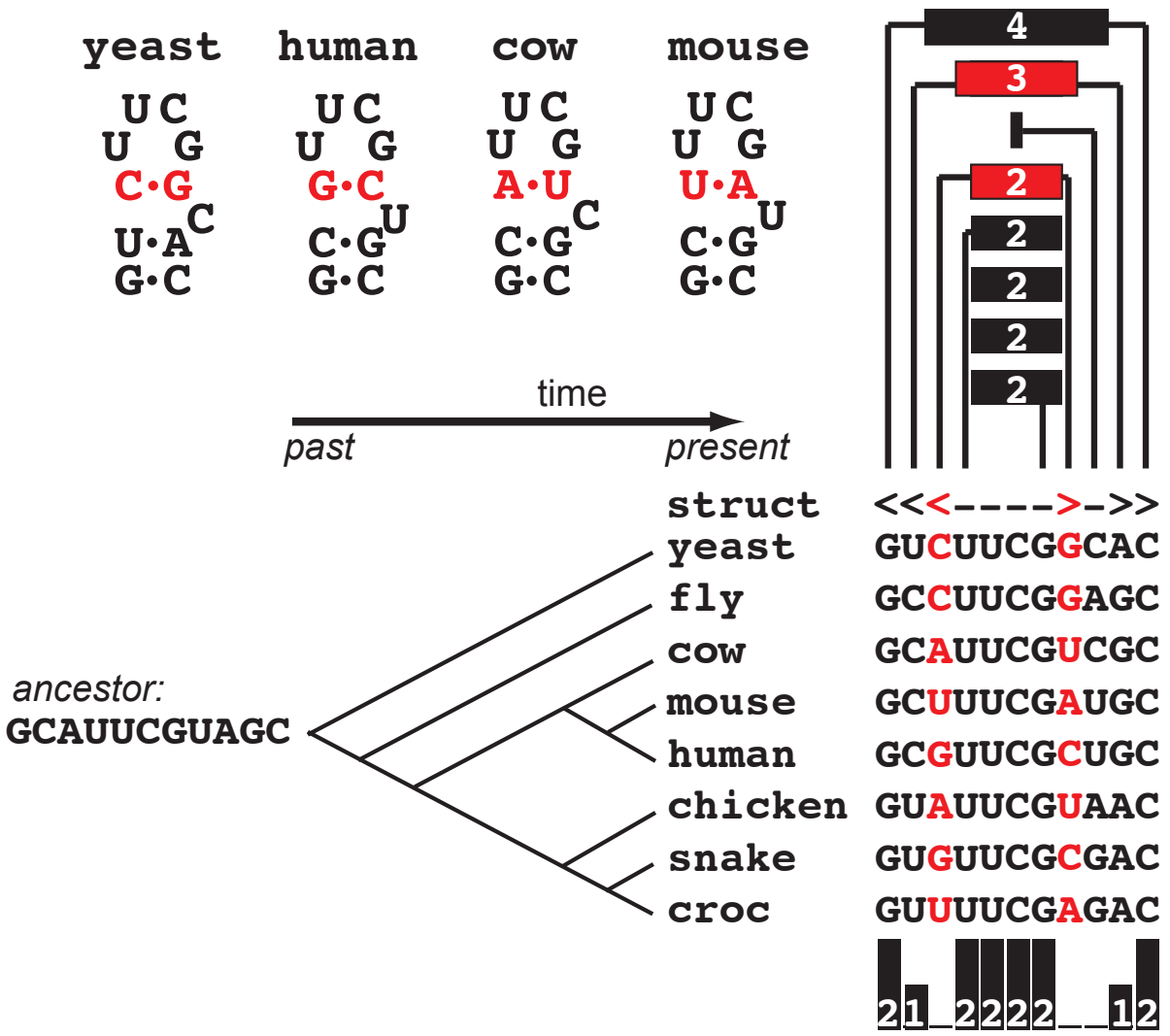
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ bits

sequence profile: 14 bits

expect a match by chance: 1 in 2^{14} nt =~ 16 Kb

Structure contributes additional information from covariation



sequence + structure profile: 17 bits

Base-paired columns are not independent

Maximum ***extra*** info: 2 bits per base pair

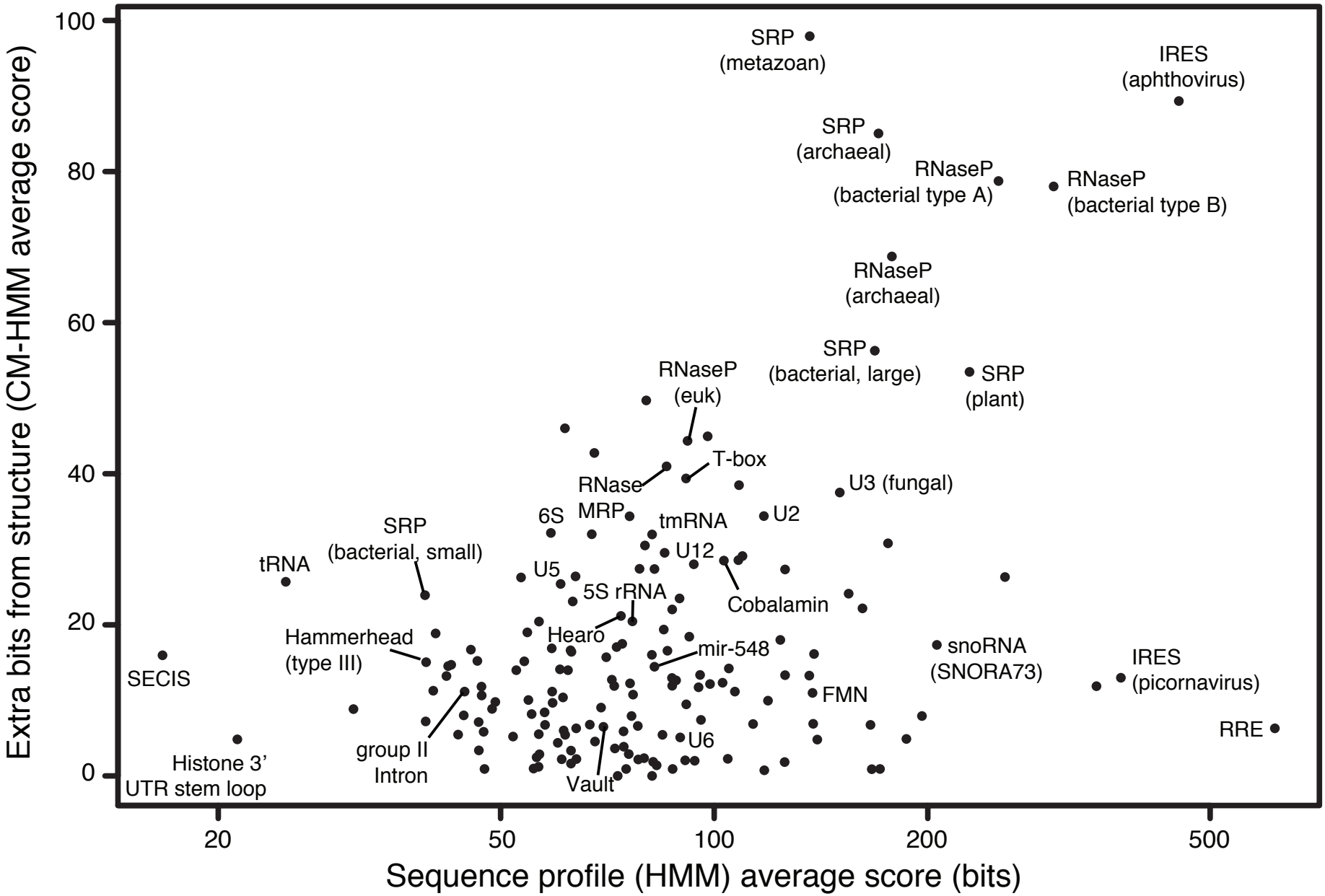
$\log_2 (16/4) = 2$ bits

sequence profile: 14 bits

expect a match by chance: 1 in 2^{17} nt =~ 130 Kb

reducing expected false positives by $2^3 = 8$-fold

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, repetitive DNA elements	structural RNAs
databases	Pfam and Dfam (16306 and 4150 entries)	Rfam (2474 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, Eddy SR. Bioinformatics, 29: 2487-2489, 2013.

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

Profile HMMs: sequence family models built from alignments

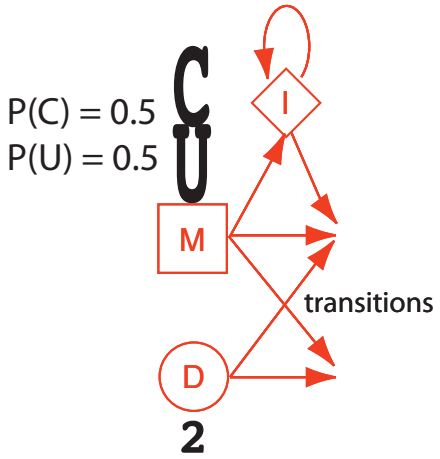
yeast	GUCUUCGGCAC
fly	GCCUUCGGAGC
cow	GCAUUCGUCGC
mouse	GCUUUCGAUGC
human	GCGUUCGCUGC
chicken	GUAUUCGUAAC
snake	GUGUUCGCGAC
croc	GUUUUCGAGAC

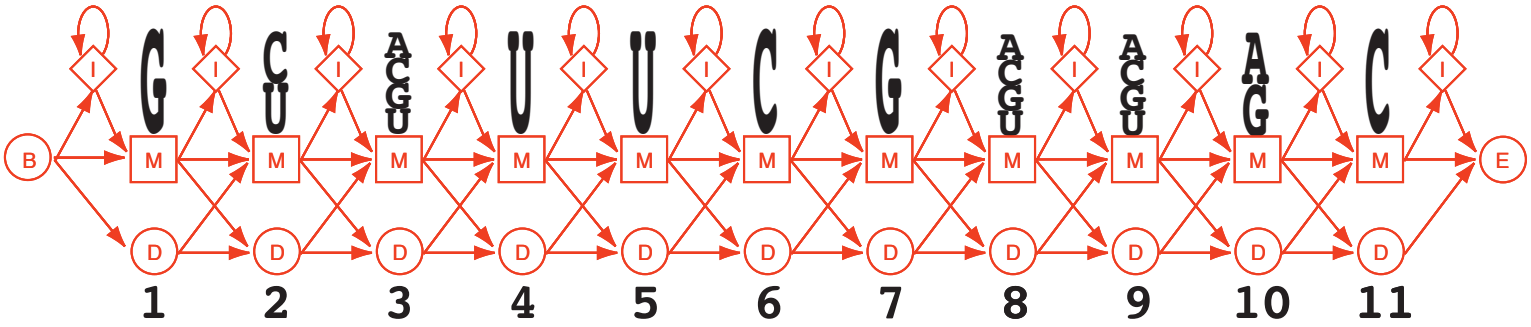
One HMM node per alignment column

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

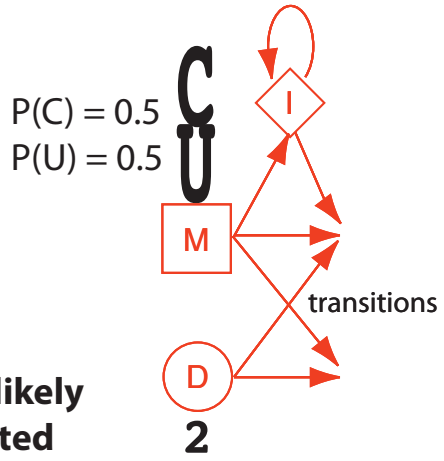
One HMM node per alignment column

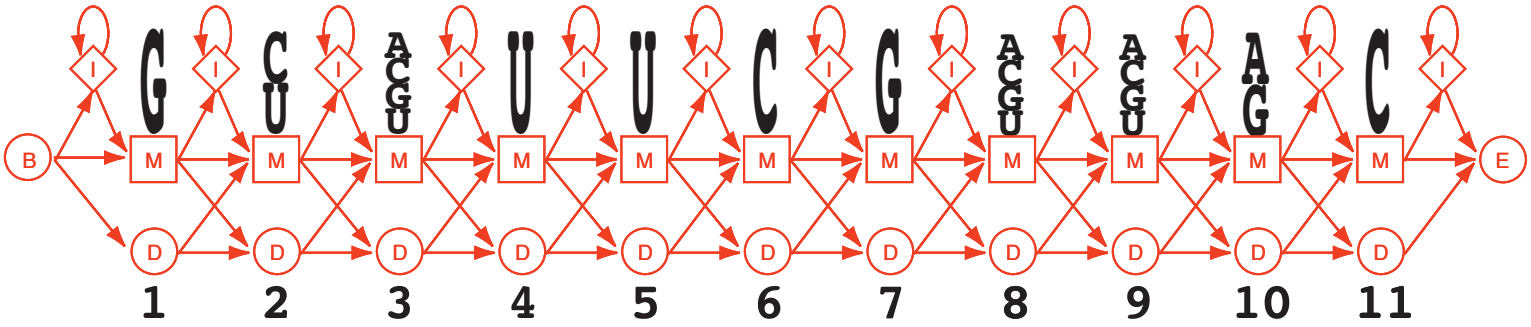
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

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

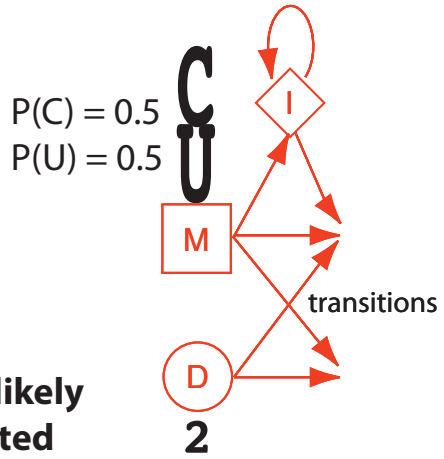
One HMM node per alignment column

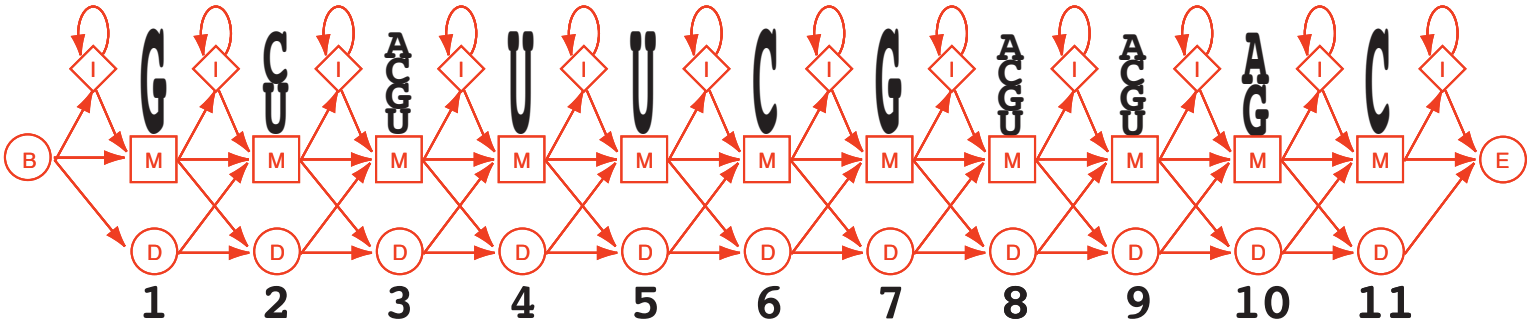
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

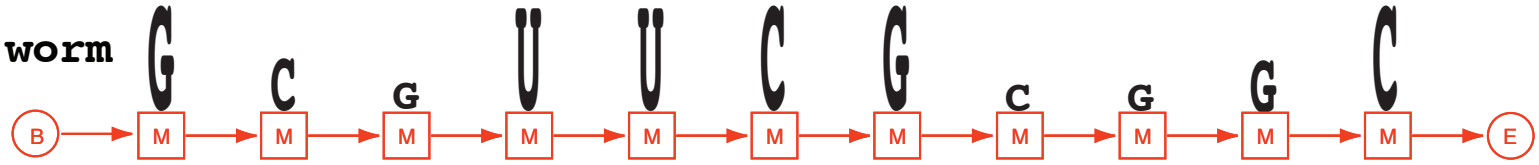
HMMs generate homologous sequences.

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
```

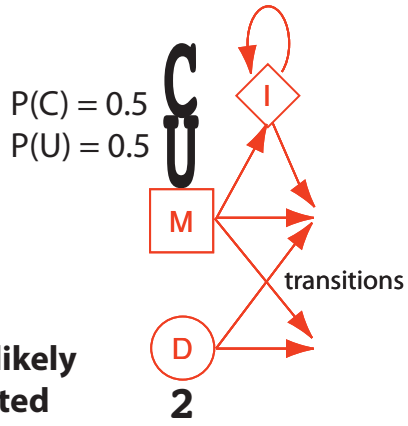
One HMM node per alignment column

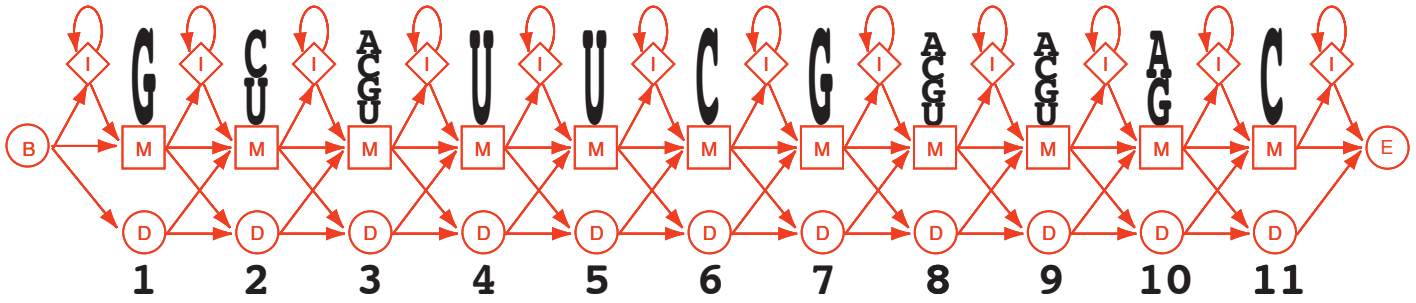
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

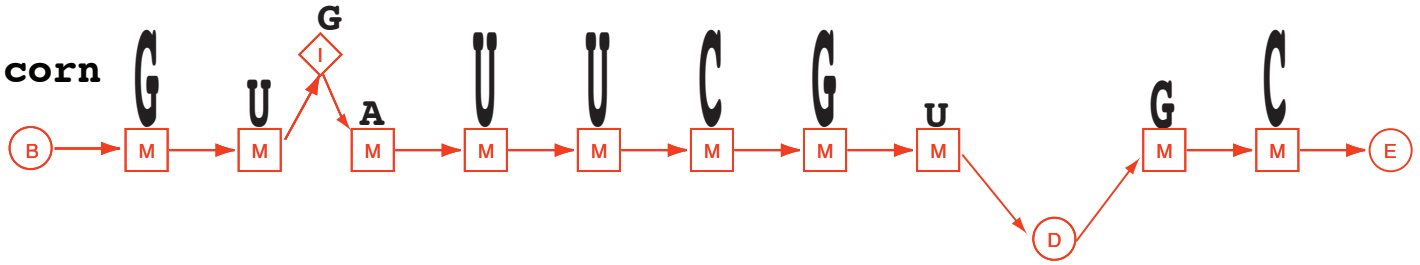
HMMs generate homologous sequences.

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

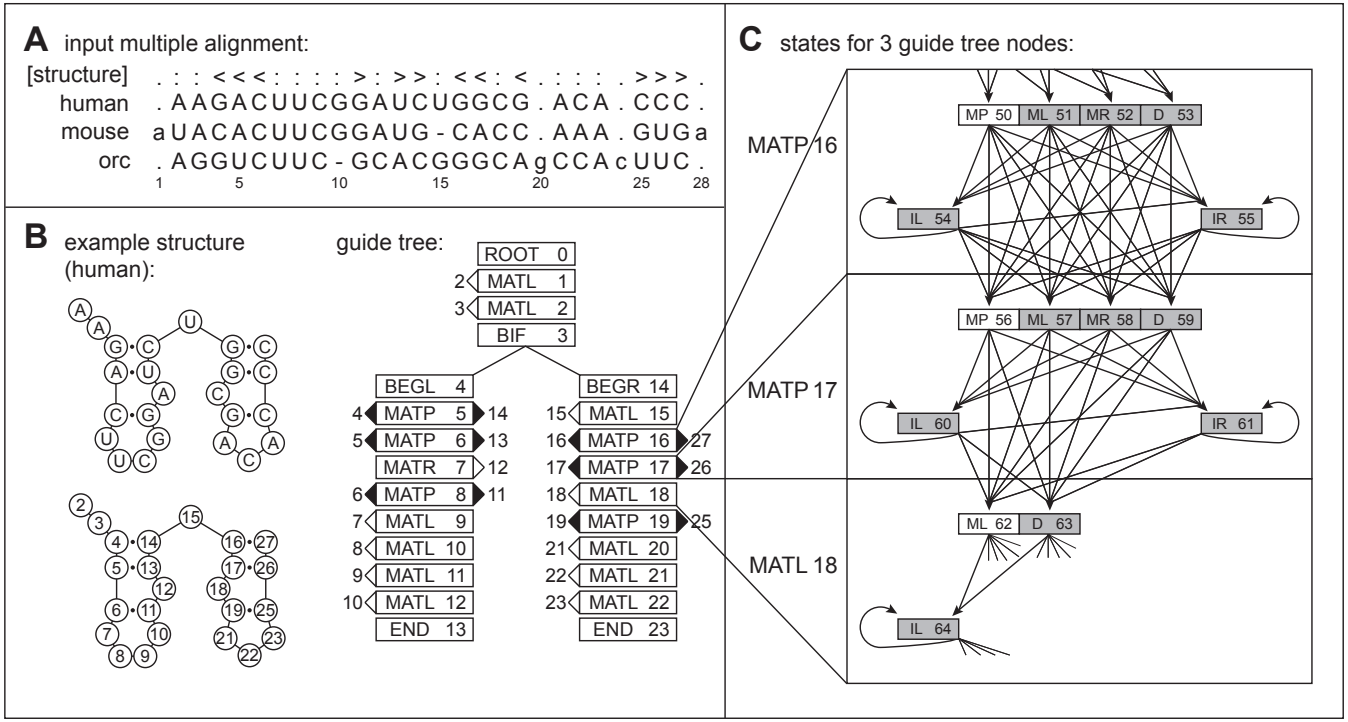
Node for column 2:

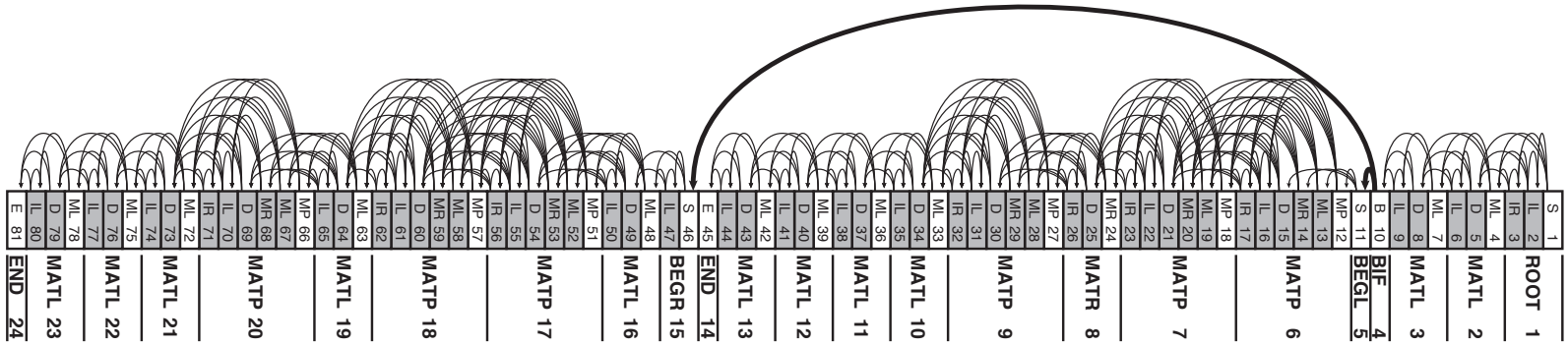






Covariance models (CMs) are built from structure-annotated alignments

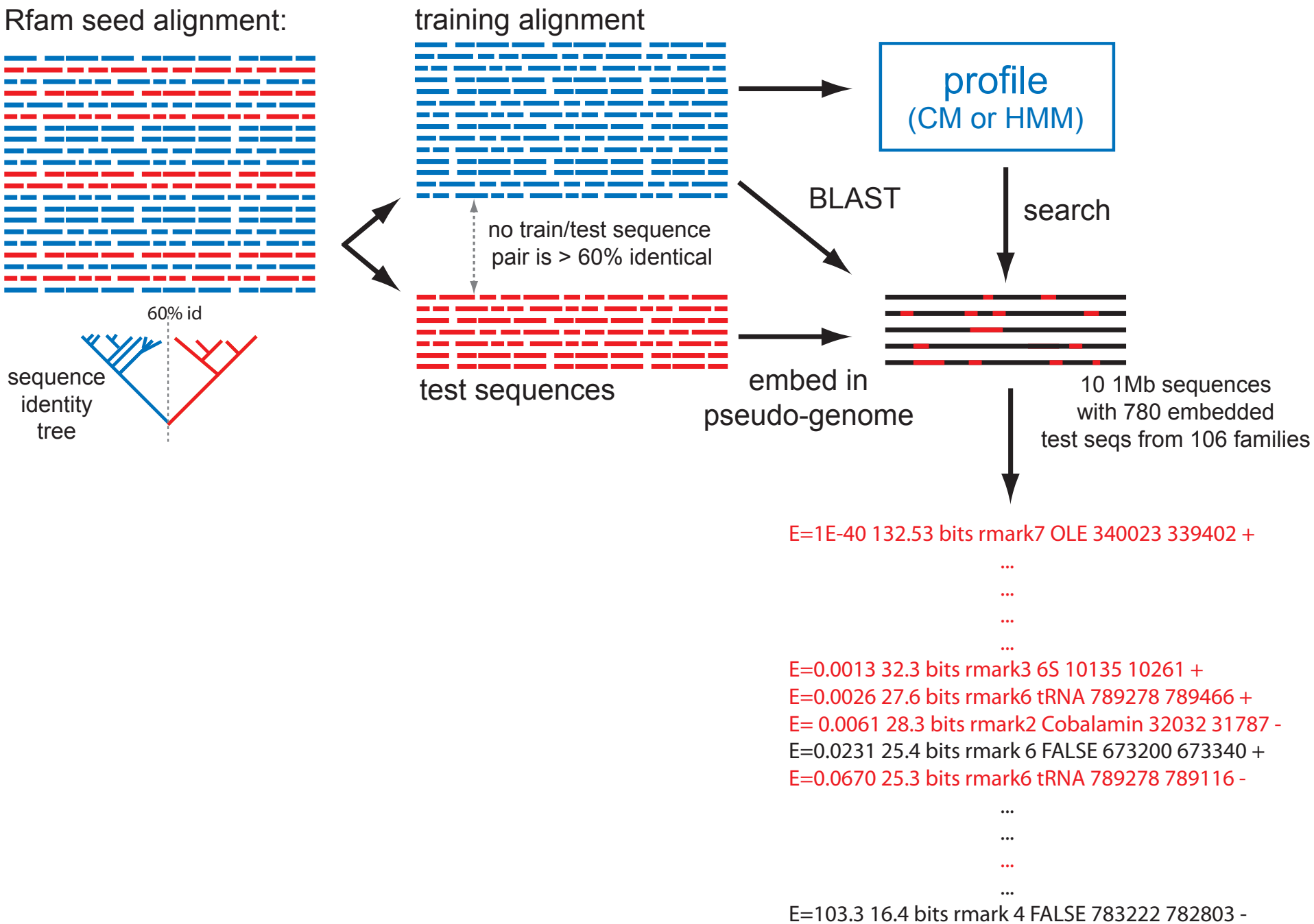




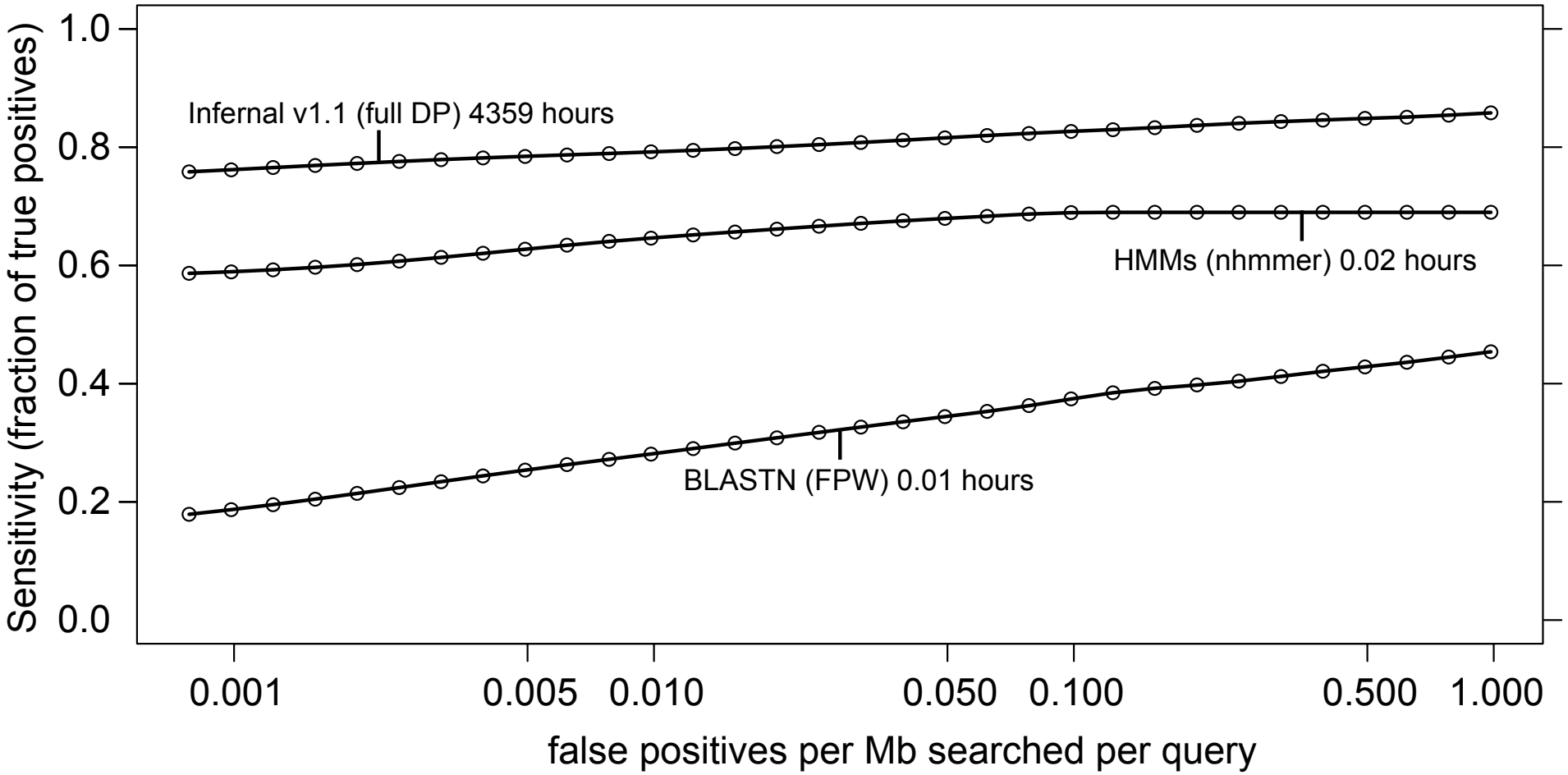
Is the added complexity worth it?

RMARK: a challenging internal RNA homology search benchmark

Is the added complexity worth it?
RMARK: a challenging internal RNA homology search benchmark



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



Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

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

Outline of talk

1. Motivation: collecting homologs facilitates comparative sequence analysis.
1965: Secondary structure determination of transfer RNA.

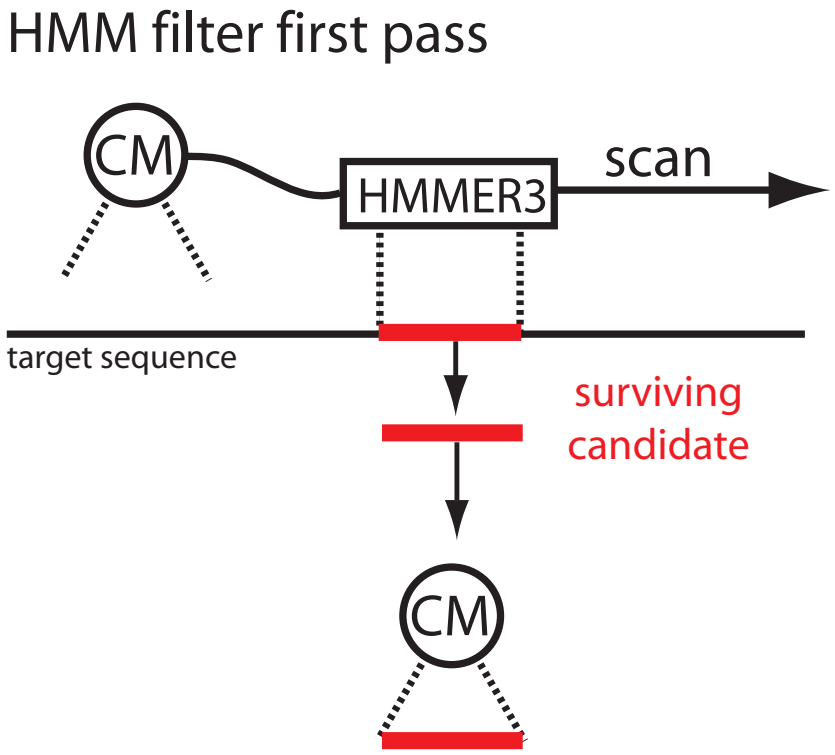
2. Sequence and sequence+structure profiles

3. Accelerating RNA homology search

4. Implications for Rfam

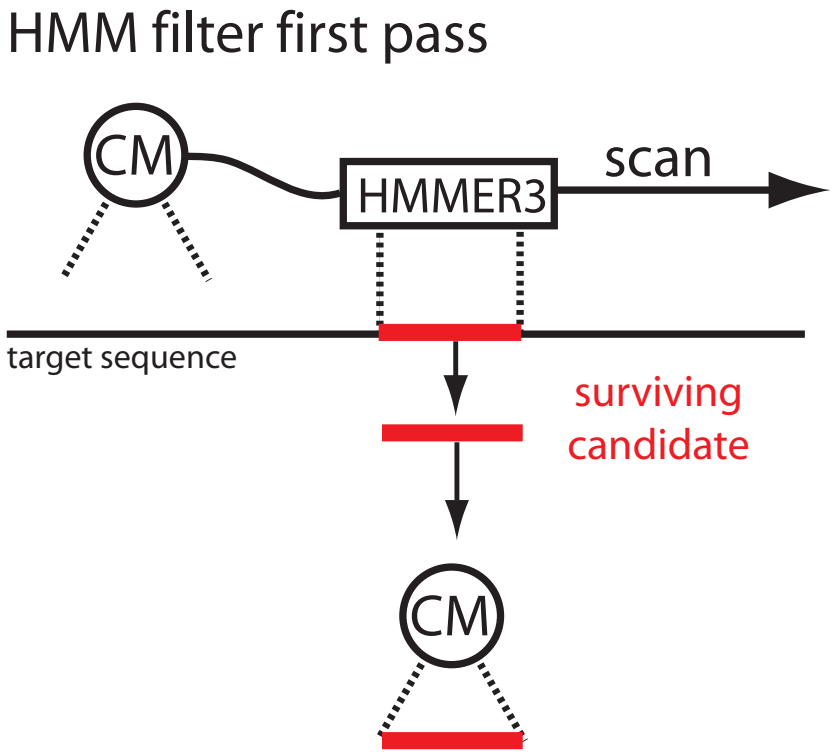
5. Y RNAs

Filter target database using profile HMMs*



*Weinberg, Ruzzo, RECOMB, 243-251, 2004; Weinberg, Ruzzo, Bioinformatics, 22(1) 35-39 2006.

Filter target database using profile HMMs*

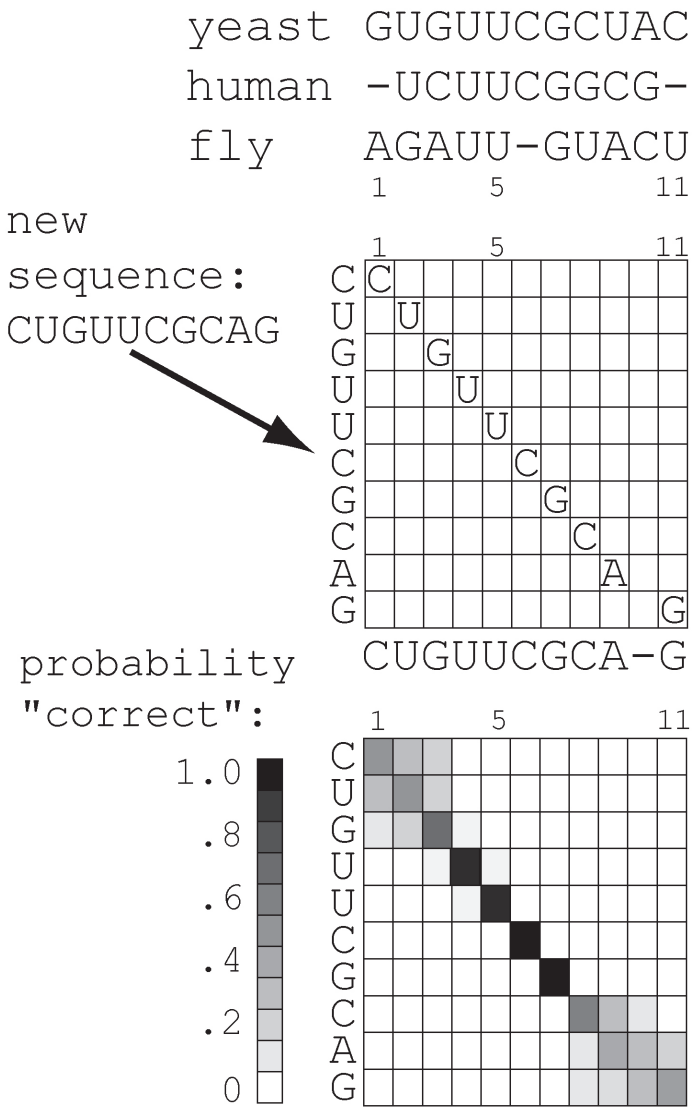


- Even if we filter out 99% of the database (for up to 100X acceleration), searches will still be too slow.

- CM step needs to be accelerated.

*Weinberg, Ruzzo, RECOMB, 243-251, 2004; Weinberg, Ruzzo, Bioinformatics, 22(1) 35-39 2006.

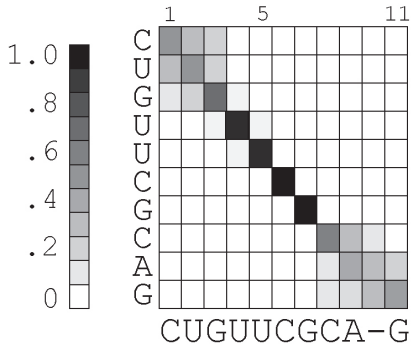
Accelerating CM alignment step 1: HMM posterior decoding to get confidence estimates

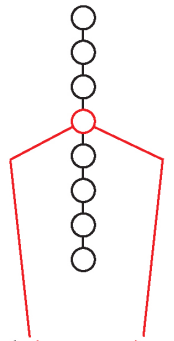


Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*

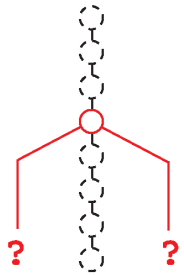
HMMs – Each column of seed alignment corresponds to a column of matrix.

```
yeast GUGUUCGCUAC
human -UCUUCGGCG-
fly   AGAUU-GUACU
```





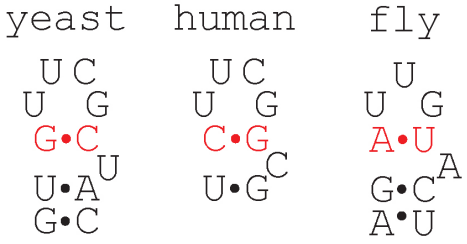
```
struct <<<<----->->>
yeast  GUGUUCGCUAC
human  -UCUUCGGCG-
fly    AGAUU-GUACU
```



CUGUUCGCAG

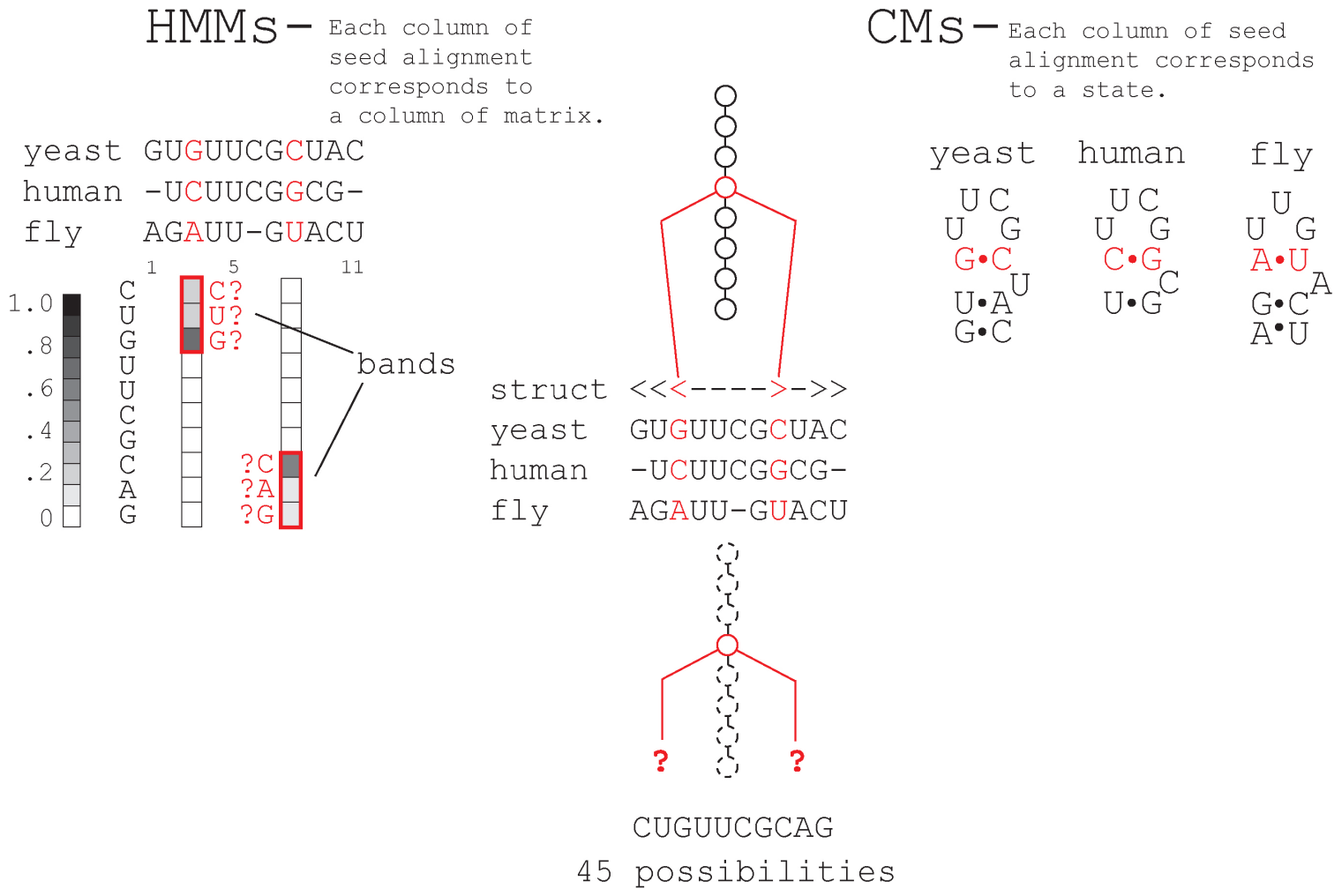
45 possibilities

CMs – Each column of seed alignment corresponds to a state.



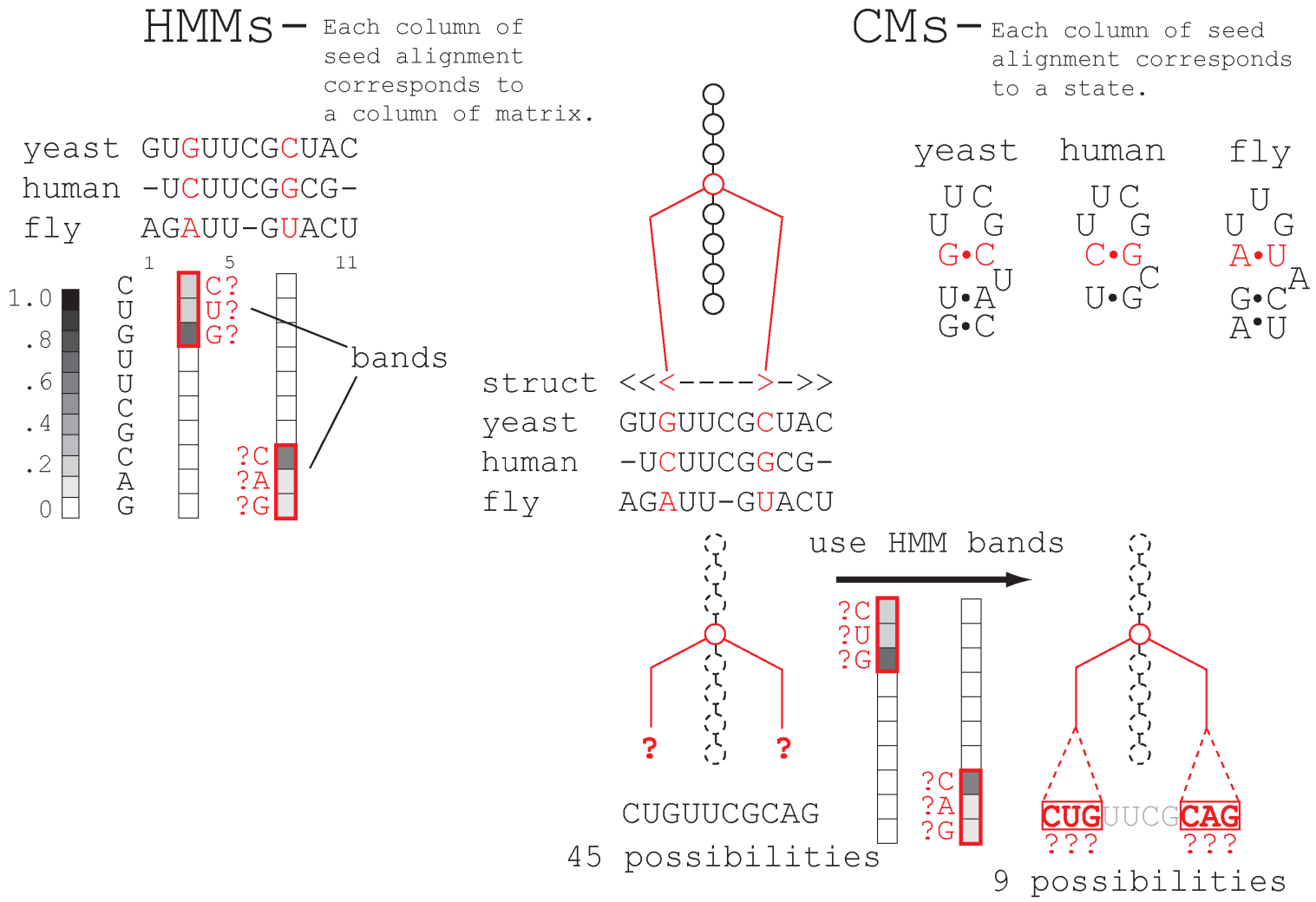
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*



*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

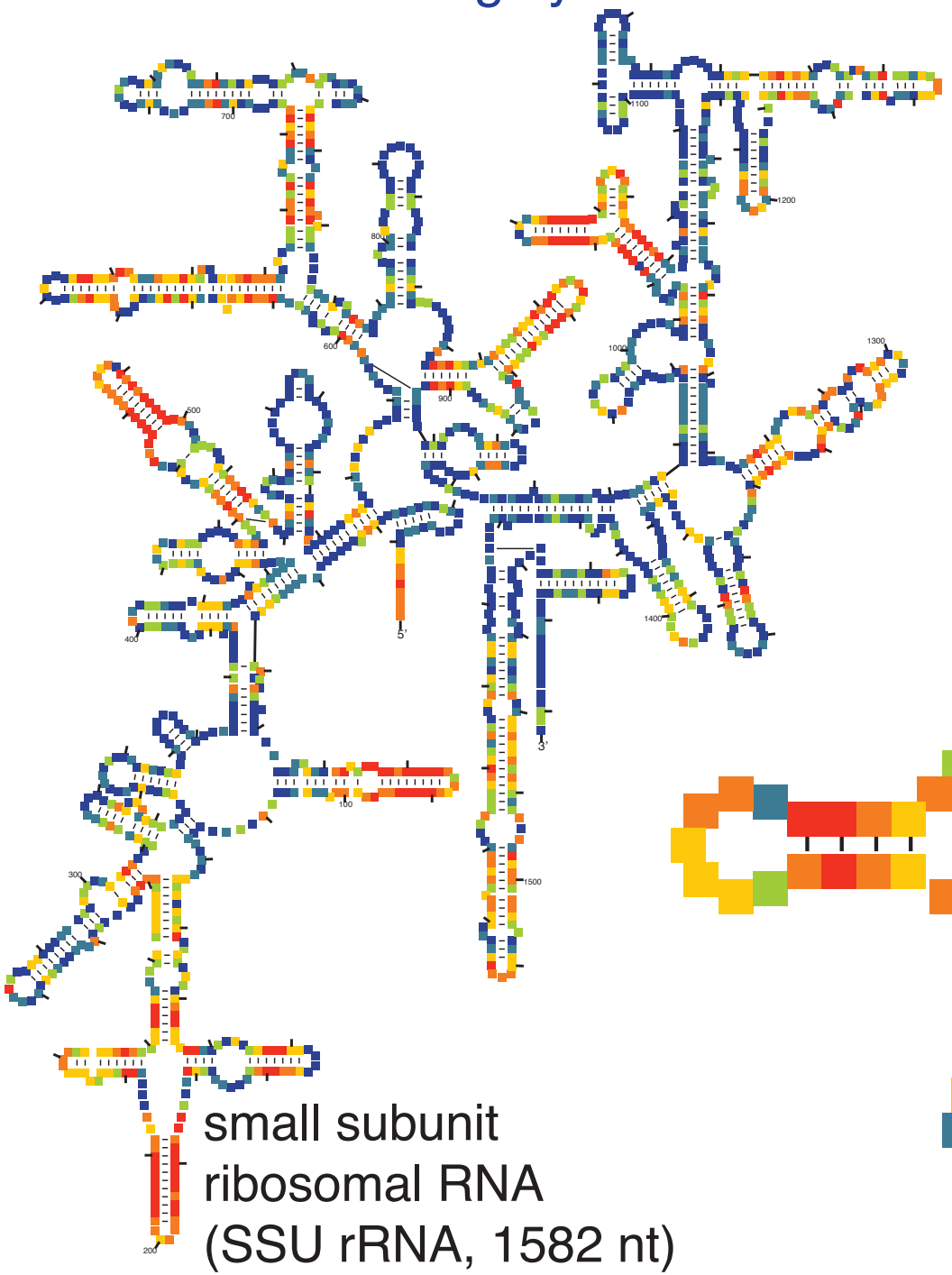
Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*



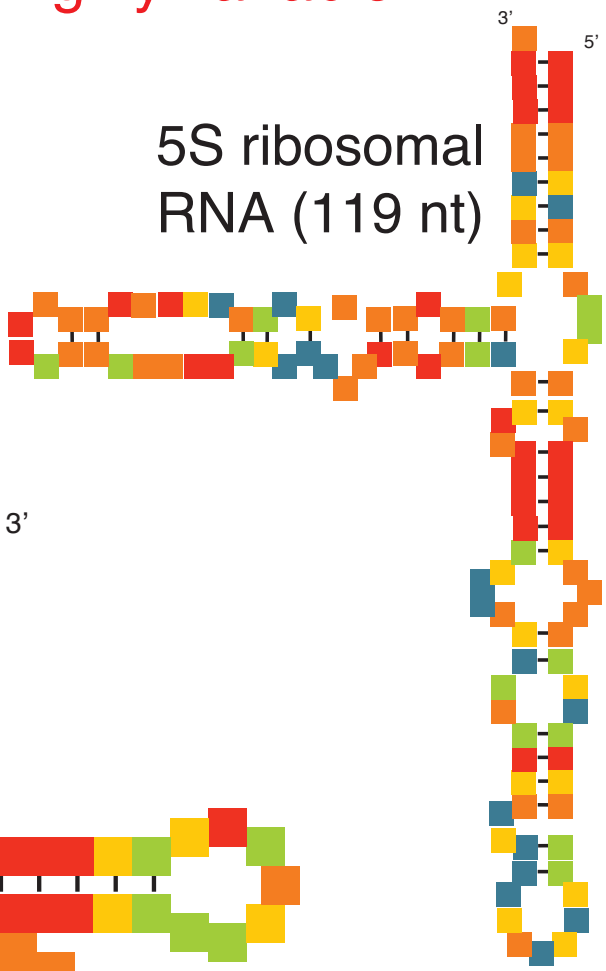
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

Sequence conservation per position

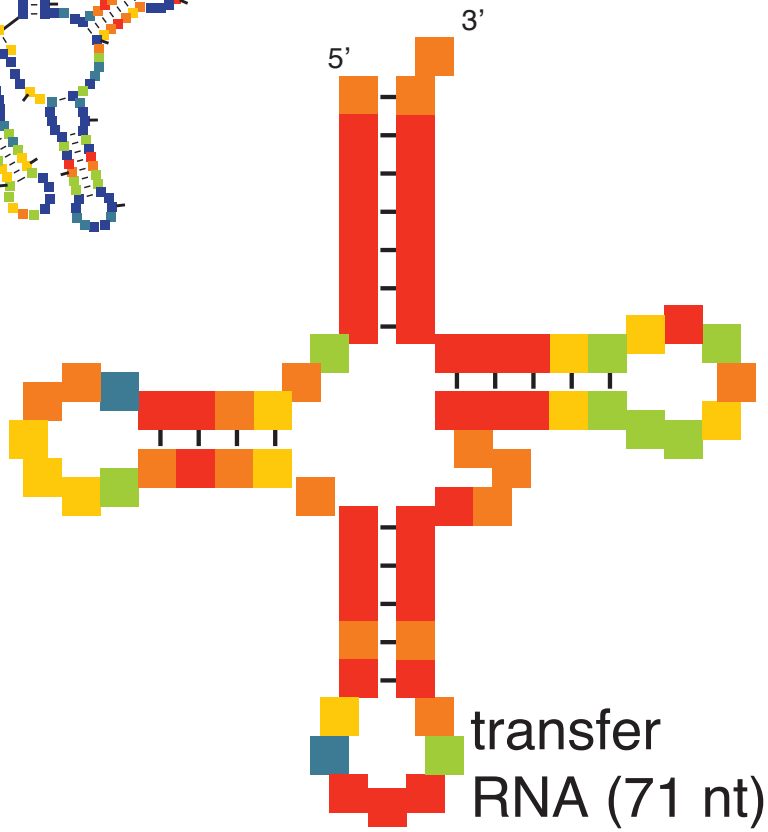
blue:highly conserved ..... red: highly variable



small subunit ribosomal RNA (SSU rRNA, 1582 nt)

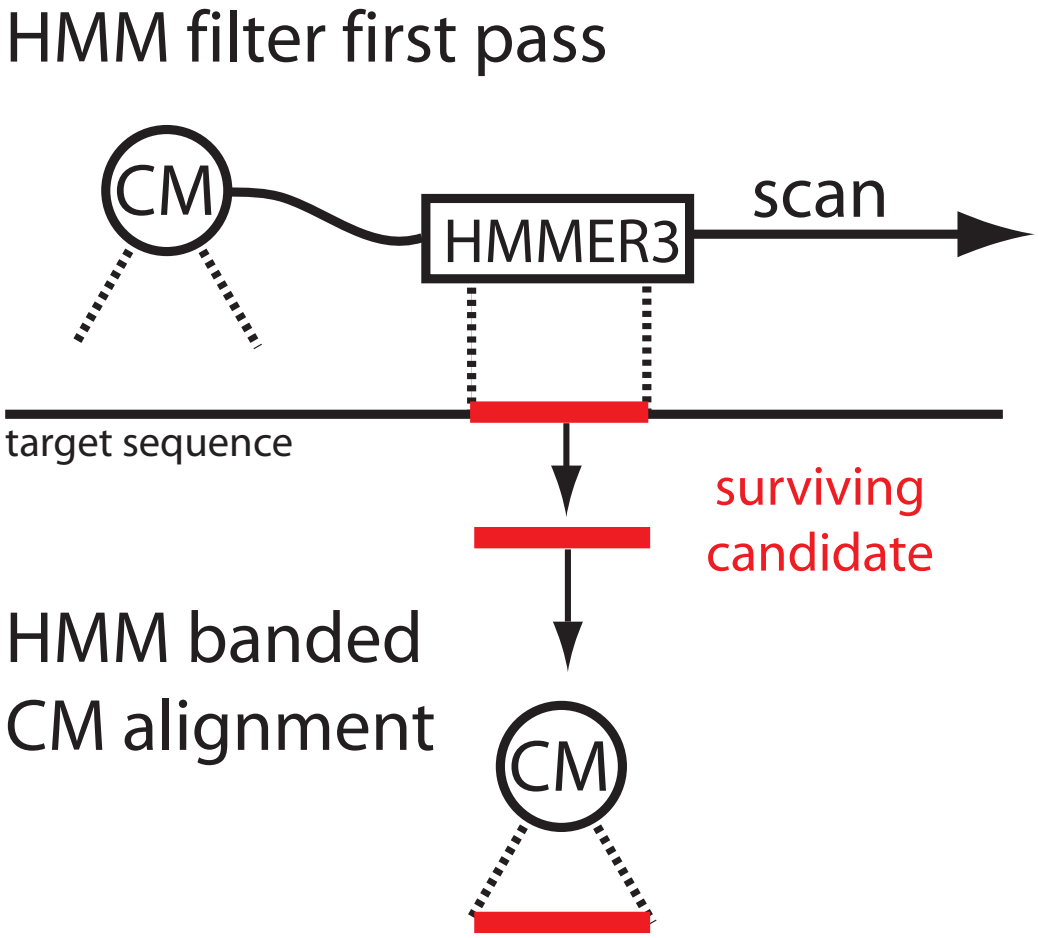


5S ribosomal RNA (119 nt)

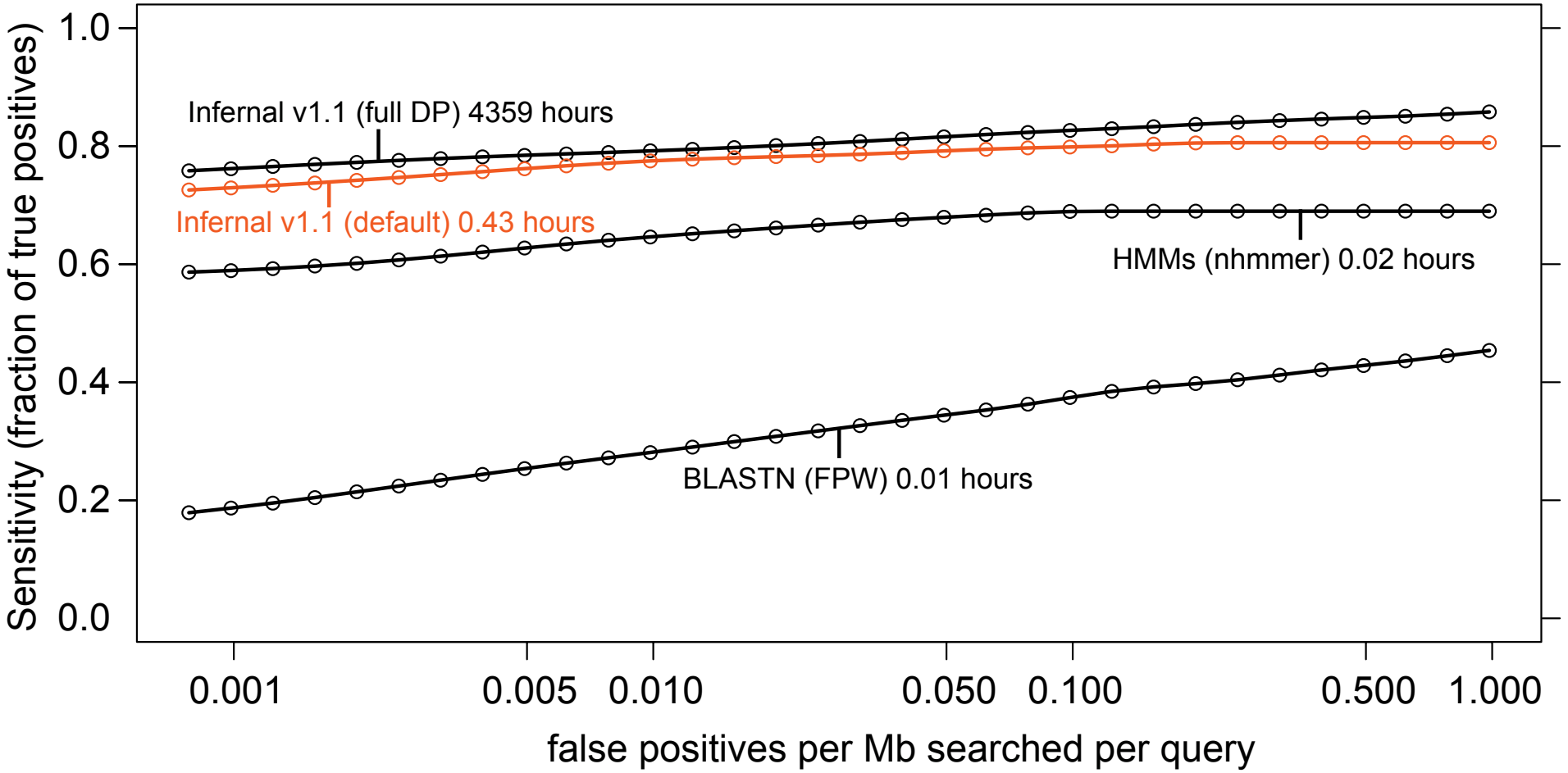


transfer RNA (71 nt)

Use HMMs as filters and to constrain CM alignment



HMM-based acceleration makes Infernal 10,000 times faster



Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

Outline of talk

1. Motivation: collecting homologs facilitates comparative sequence analysis.
1965: Secondary structure determination of transfer RNA.

2. Sequence and sequence+structure profiles

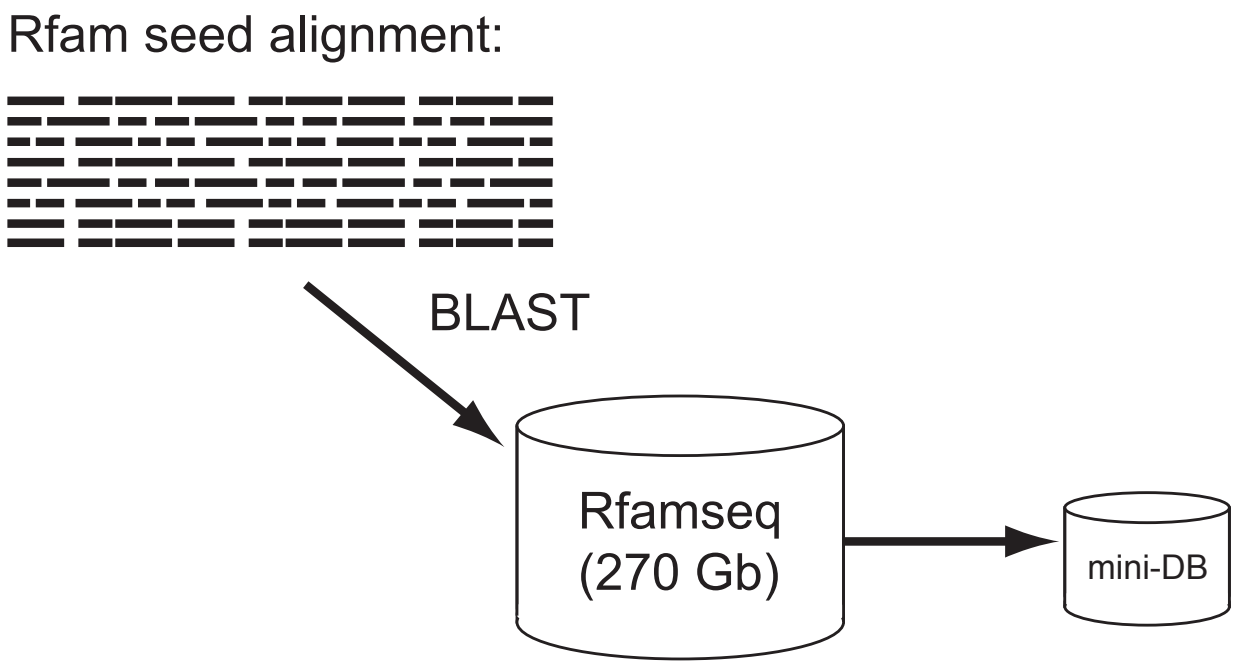
3. Accelerating RNA homology search

4. Implications for Rfam

5. Y RNAs

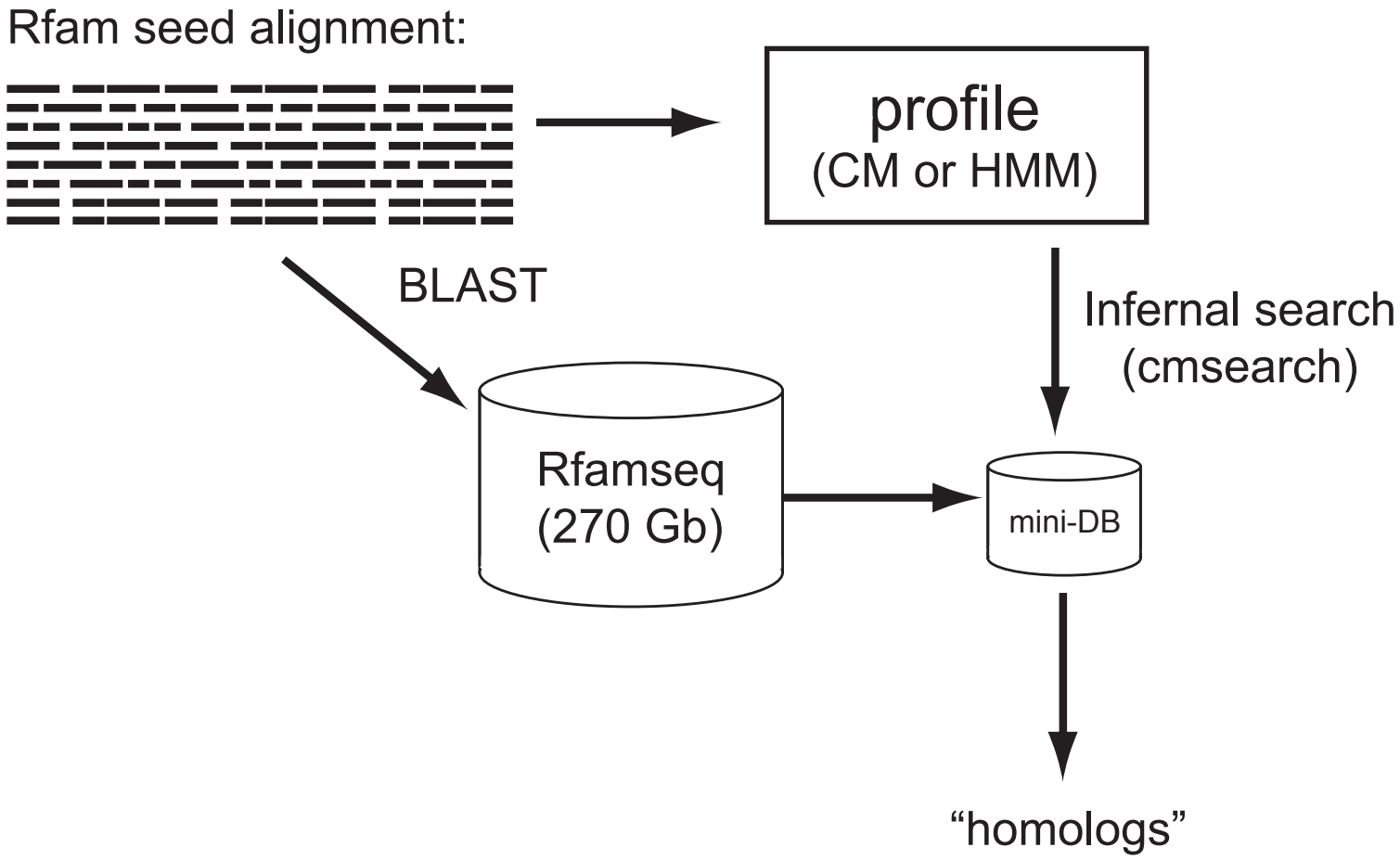
Rfam used BLAST filters from 2003 to 2012

- Rfam includes $>$ 2600 RNA families, each represented by an alignment, CM and set of predicted homologs in a large database (Rfamseq).

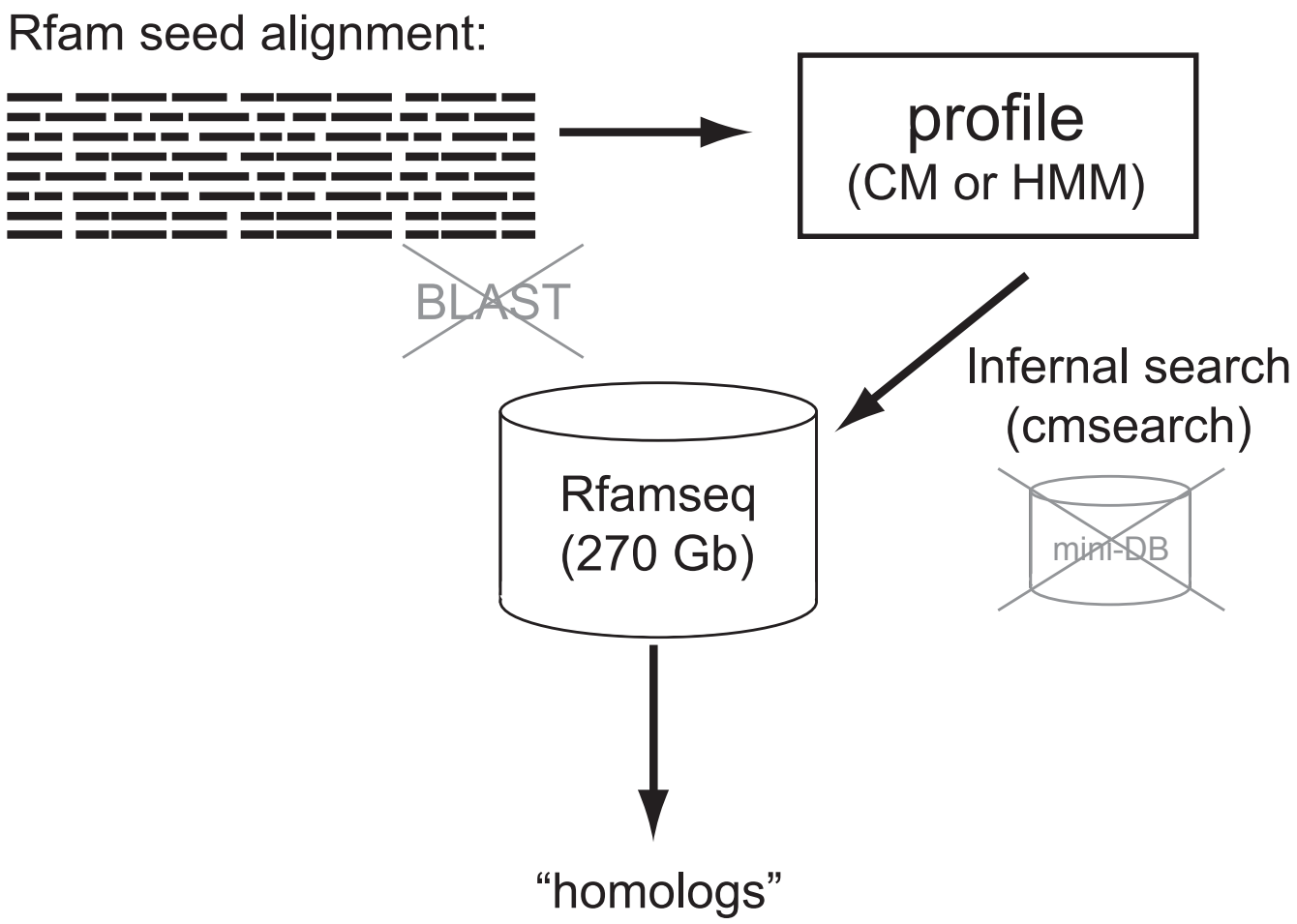


Rfam used BLAST filters from 2003 to 2012

- Rfam includes $>$ 2600 RNA families, each represented by an alignment, CM and set of predicted homologs in a large database (Rfamseq).



Rfam 12.0 (2014)*, first release without BLAST filtering



*Nawrocki, Burge et. al, NAR 43:D130-D137, 2015.

Rfam 12.0 (2014)*, first release without BLAST filtering

Search results against Rfamseq for 200 random families:

strategy	time (h)	# hits	# unique hits
Old (BLAST + Infernal 1.0)	4069.8	179,681	53
New (Infernal 1.1)	4222.2	201,814	22,312

*Nawrocki, Burge et. al, NAR 43:D130-D137, 2015.

Infernal 1.1 finds 11,000 new group I intron candidates

Table 1. Comparison of the old Rfam 11.0 BLAST and Infernal 1.0 search strategy versus the new Rfam 12.0 Infernal 1.1 search strategy for 15 of 200 randomly chosen families

Accession	Family ID	Length (nt)	#of seed seqs	Time new (h)	Time old (h)	Time (old/new)	New total hits	Old total hits	New unique hits	Old unique hits
Top five families										
RF00028	Intron_gpI	251	12	125.0	357.2	2.8	71 433	60 264	11 175	1
RF00026	U6	104	188	31.2	181.1	5.8	66 517	62 174	4367	14
RF00003	U1	166	100	11.6	64.0	5.5	15 770	14 867	904	1
RF00162	SAM	108	433	8.3	590.0	70.8	4905	4797	108	0
RF00050	FMN	140	144	17.1	169.9	23.9	4381	4306	76	1

Nawrocki, Burge et. al, NAR 43:D130-D137, 2015.

It is now easier to use Rfam/Infernal to annotate your own datasets.

- A bacterial genome takes about 30 minutes for all 2474 models.

Table 2. Summary statistics for Rfam-based annotation of RNAs in various genomes and metagenomics data sets

Genome/data set	Size (Mb)	# of hits	# of fams	CPU time (hours)	Mb/hour
Homo sapiens	3099.7	14 508	796	650	4.8
Sus scrofa (pig)	2808.5	6177	625	460	6.1
Drosophila melanogaster	168.7	4321	156	30	5.7
Caenorhabditis elegans	100.3	1022	175	20	5.2
Saccharomyces cerevisiae	12.2	376	96	1.7	7.3
Escherichia coli	4.6	256	112	0.46	10.2
Bacillus subtilis	4.1	211	52	0.57	7.2
Methanocaldococcus jannaschii	1.7	257	18	0.31	5.6
Aquifex aeolicus	1.6	52	7	0.22	7.3
Borrelia burgdorferi	0.9	44	7	0.22	4.1
Human immunodeficiency virus (HIV)	0.01	12	10	0.016	0.63
Human gut microbiome sample (sample ERS167139, 454 sequencing)	166.1	4342	54	22	7.7
Human gut microbiome sample (sample ERS235581, Illumina HiSeq sequencing) (28)	52.9	3159	47	8.5	6.2
Ocean metagenome (sample SRS580499, Illumina genome analyzer)	44.3	6692	59	13	3.5

Nawrocki, Burge et. al, NAR 43:D130-D137, 2015.

Outline of talk

1. Motivation: collecting homologs facilitates comparative sequence analysis.
1965: Secondary structure determination of transfer RNA.

2. Sequence and sequence+structure profiles

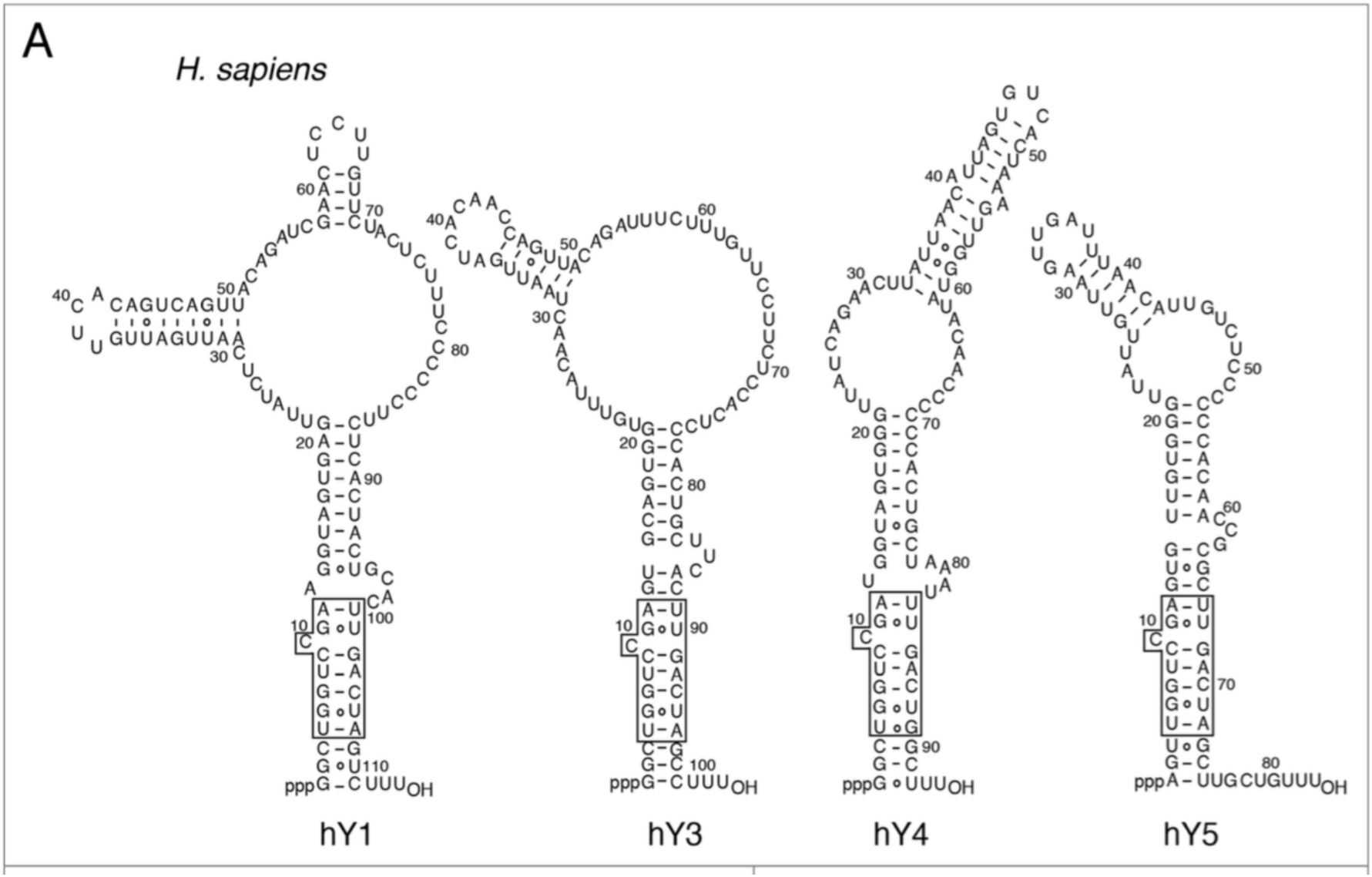
3. Accelerating RNA homology search

4. Implications for Rfam

5. Y RNAs

Y RNAs

- Originally discovered in 1981 in complex with Ro60 protein, a target of autoimmune antibodies in patients with systemic lupus erythematosus and Sjogren's synodrome*.
- Four distinct Y RNAs are encoded in the human genome



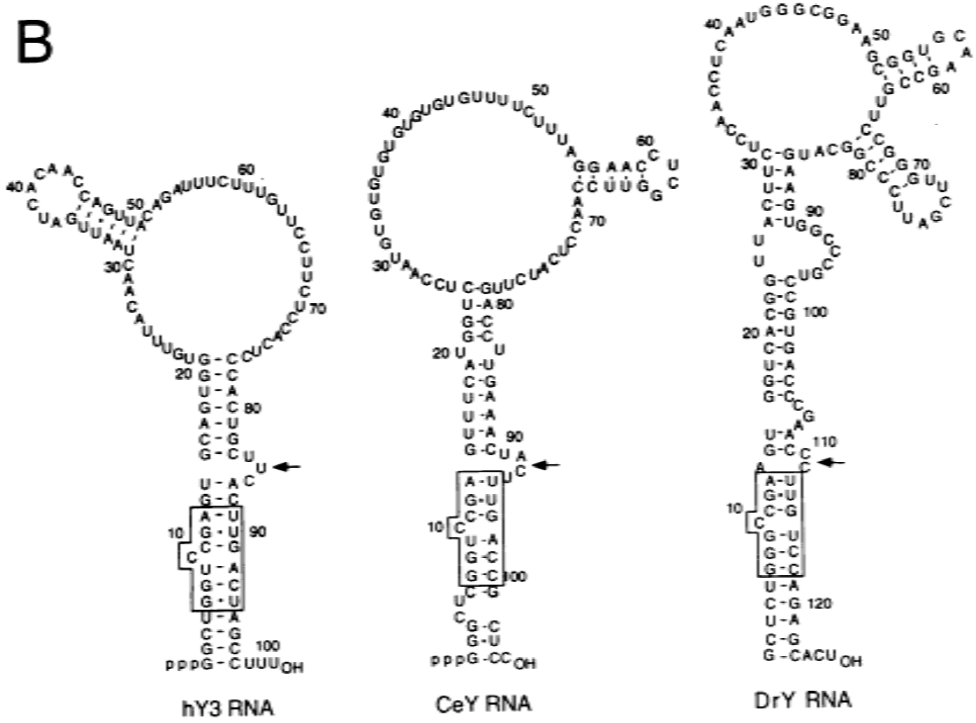
*Mol Cell Biol. 1981 Dec; 1(12): 11381149.

Ro ribonucleoproteins contribute to the resistance of *Deinococcus radiodurans* to ultraviolet irradiation

Xinguo Chen, Anne Marie Quinn, and Sandra L. Wolin[1]

Departments of Cell Biology and Molecular Biophysics and Biochemistry, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06536 USA

GENES & DEVELOPMENT 14:777–782



J Mol Evol (2010) 70:346–358
DOI 10.1007/s00239-010-9332-4

Nematode sbRNAs: Homologs of Vertebrate Y RNAs

Ilenia Boria · Andreas R. Gruber · Andrea Tanzer · Stephan H. Bernhart · Ronny Lorenz · Michael M. Mueller · Ivo L. Hofacker · Peter F. Stadler

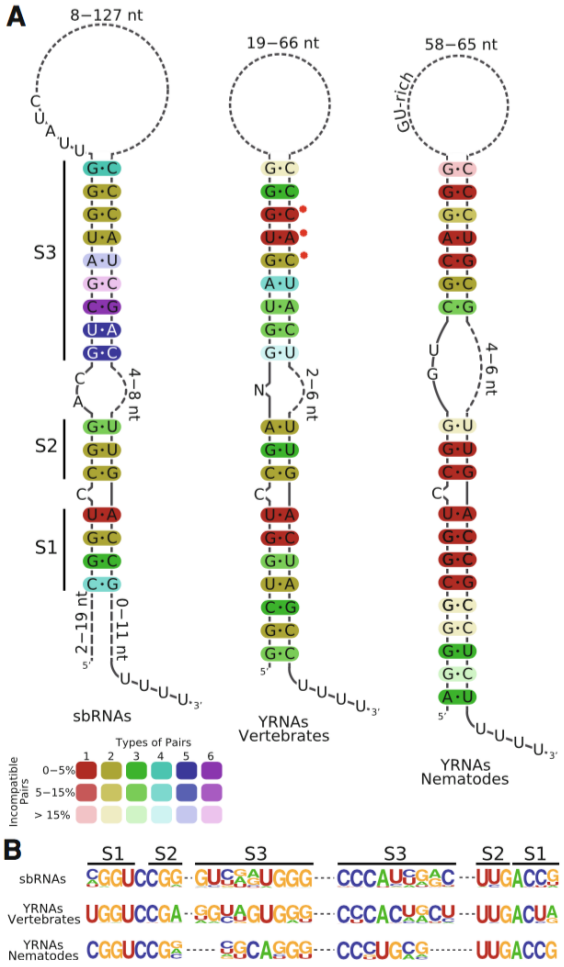
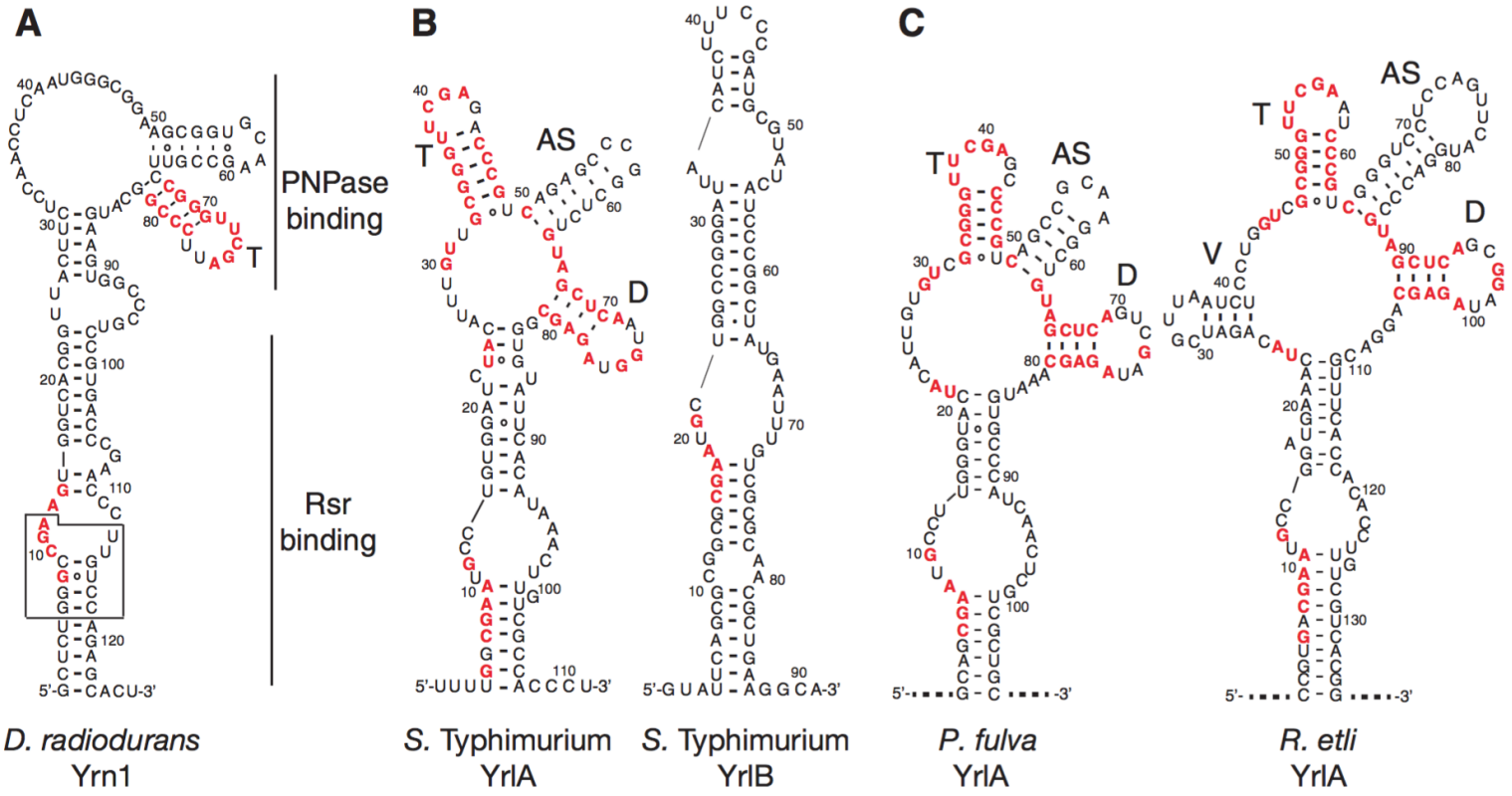


Fig. 4 **a** Comparison of secondary structures for nematode sbRNAs, vertebrate Y RNAs and the previously described Y RNA family in the genus *Caenorhabditis*. *Stars* denote the region identified by Gardiner et al. (2009) to be crucial for the function of Y RNA in DNA replication. **b** Sequence logos for helical regions S1, S2, and S3

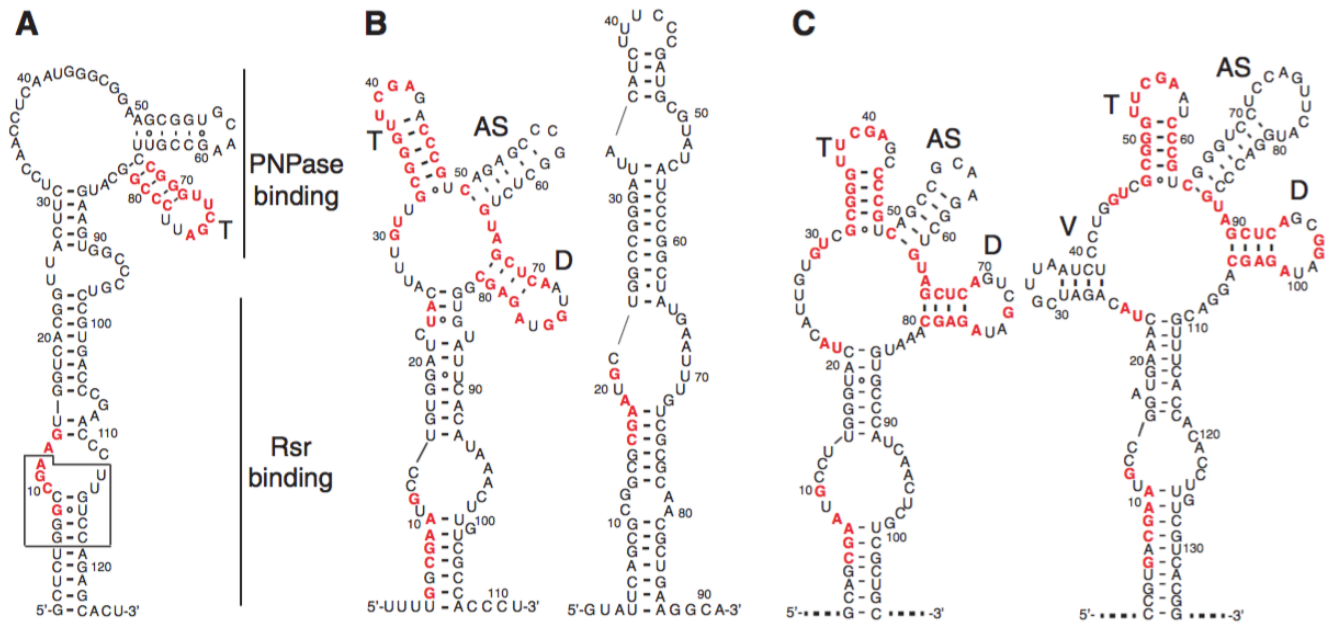
Bacterial noncoding Y RNAs are widespread and mimic tRNAs

XINGUO CHEN,[1,3] SOYEONG SIM,[1,3] ELISABETH J. WURTMANN,[1,4] ANN FEKE,[1] and SANDRA L. WOLIN[1,2]

[1]Department of Cell Biology, [2]Department of Molecular Biophysics and Biochemistry, Yale School of Medicine, New Haven, Connecticut 06510, USA



RNA 20:1715–1724; Published by Cold Spring Harbor Laboratory Press for the RNA Society



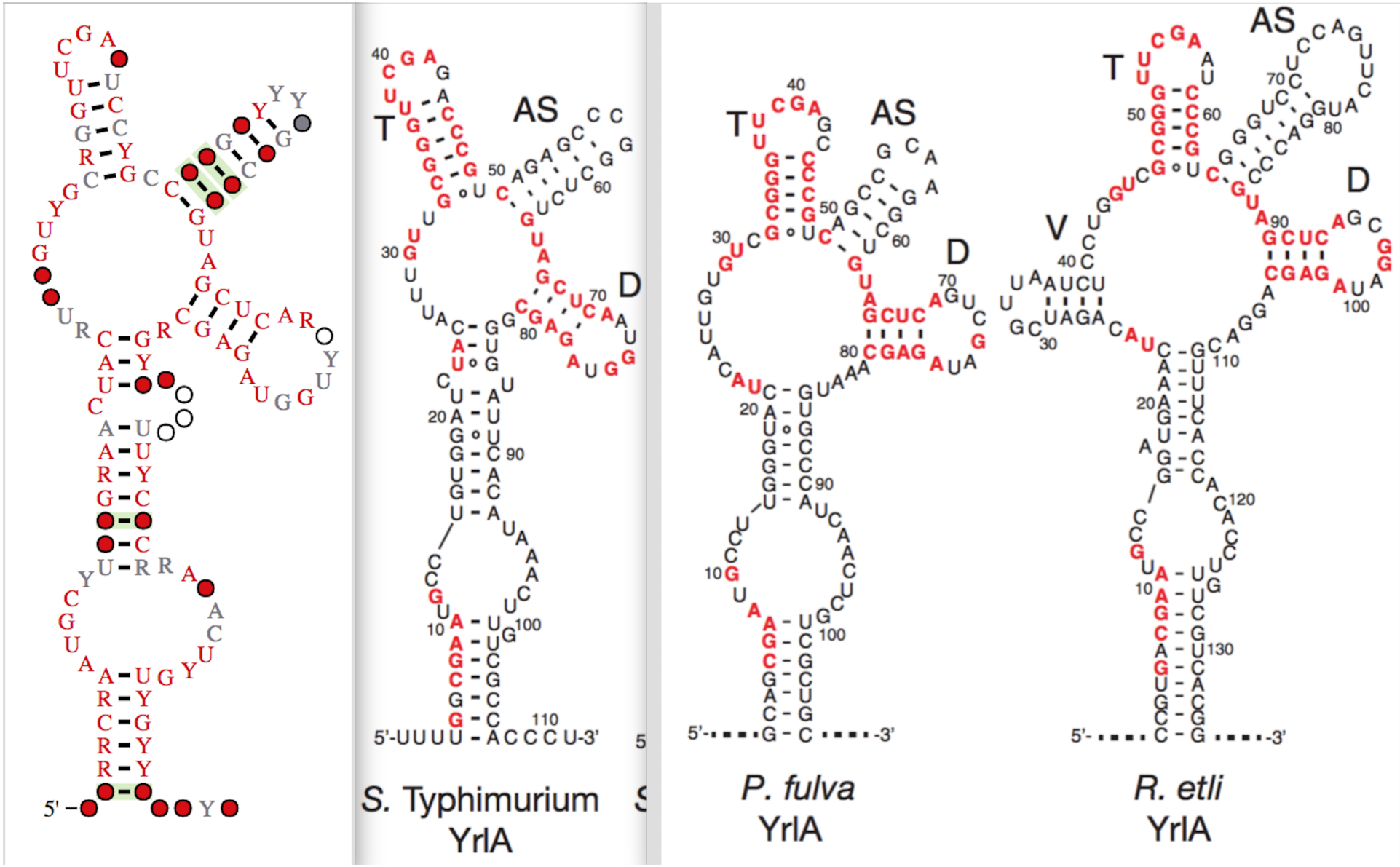
D

```
                                  --Rsr ---                        --------V-------------*
S.  Typhimurium YrlA     1 ----------UUUUGGCGAAUGCCUGUGGA-UCUACAU-U-----------UGUU
P.  fulva                1 ------------GCAGCGAAUGCCUUUGGGU-ACUACAU-U-----------GUGUC
R.  etli                 1 ----------CCGUGACGAAUGCCGGAUGAAACUACAGAUCGUUAAUCUCCUGGUC
G.  polyisoprenivorans   1 -----CGAGAGCGGGCCGAA-GUCGAUCG--GUUACCA-CUCAUAAUGGGCAGGUU
S.  roseum               1 ---UGUCGGUACGGGUCGAA-GCGCGUCG--GUUAUCGCCUUCUAAGCGGGUGGUC
M.  smegmatis            1 GCUUGGUGAUGUGAGCCGGACGUCGAUCG--GUUACCA-CUUCUAAUGGAAUGGUC
Mycobacteriophage        1 -------GAGGCAGGCCGAGAGUUUGUCG--GUUACCUCAUUGAAAGAGAGAGGUC
CONSENSUS                  .............G.CGAA.GY..R..G..RYUAY...Y.............GUY

                           <<<<<  T  >>>>><<<<<       AS            >>>>>  <<<< *
S.  Typhimurium YrlA     33 GCGGGUUCGAGACCCGUCAGAGCCC---------------GGCUCUGUAGCUCAAU
P.  fulva                32 GCGGGUUCGAGCCCCGUCAGCCGC---------------AAGGCUGUAGCUCAGU
R.  etli                 47 GCGGGUUCGAAUCCCGUCGGGUCCUCCAGUUCAU------GGACCCGUAGCUCAGC
G.  polyisoprenivorans   48 GCGGGUUCGAGUCCCGCCAUCACCUCCG-----------GGUGGUGUAGCUCAAC
S.  roseum               51 GCGGGUUCGAAUCCCGUCACAGAUUCGA-----------GUCUGUGUAGCUCAGU
M.  smegmatis            54 GCGGGUUCGAAUCCCGCCCGCCCUACGC-----------AGGGCGGUAGCUCAGC
Mycobacteriophage        48 GCGGGUUCGAAUCCCGUCAGGUGGUUACCGGUACGCCCGGACACCUGUAGCUCAGU
CONSENSUS                  GCGGGUUCGARYCCCGYCR.......................YGUAGCUCARY

                              D  >>>>
S.  Typhimurium YrlA     74 -GGU-AGAGCGGUGUAU-------------UCACAUAAACUUGUCGCCACCCU-
P.  fulva                72 CGAU-AGAGCA-AAUGU-------------GCCCAUCAACUCGUCGCUGCAC--
R.  etli                 97 -GGAUAGAGCAGGA------CGUU-----UCACCACACCUUGUCGUCACGG----
G.  polyisoprenivorans   92 GGGU-AGAGCAGUGAAAACCCGGUCGGCAUAACAUG-------CCCGCUCUCG--
S.  roseum               95 -GGCCAGAGCGCCUAACGGCCGGUGUGCGUCCCCACAUCCGUGCCGGCUC-----
M.  smegmatis            98 UGGC-AGAGCAUU-----CCCGGUUGAUACCUCACACGCCCACAUCACCGAGCUC
Mycobacteriophage       104 -GGU-AGAGCGCUUAAACGCCGGUGGACACCACCA-------UGUCUGCCUC---
CONSENSUS                  .GGY.AGAGCR.....................C............YY..Y.....
```

```
# STOCKHOLM 1.0

#=GF AC   RF02553
#=GF ID   YrlA
#=GF DE   Y RNA-like
#=GF AU   Argasinska J
#=GF SE   Argasinska J
#=GF SS   Published; PMID:23540697
#=GF GA   49.00
#=GF TC   49.20
#=GF NC   48.60
#=GF TP   Gene; sRNA;
#=GF BM   cmbuild -F CM SEED
#=GF CB   cmcalibrate --mpi CM
#=GF SM   cmsearch --cpu 4 --verbose --nohmmonly -T 30.00 -Z 549862.597050 --mxsize
#=GF SM   128 CM SEQDB
#=GF DR   SO; 0000405; Y_RNA;
#=GF RN   [1]
#=GF RM   23540697
#=GF RT   An RNA degradation machine sculpted by Ro autoantigen and noncoding RNA.
#=GF RA   Chen X, Taylor DW, Fowler CC, Galan JE, Wang HW, Wolin SL
#=GF RL   Cell. 2013;153:166-177.
#=GF RN   [2]
#=GF RM   25232022
#=GF RT   Bacterial noncoding Y RNAs are widespread and mimic tRNAs.
#=GF RA   Chen X, Sim S, Wurtmann EJ, Feke A, Wolin SL
#=GF RL   RNA. 2014;20:1715-1724.
#=GF CC   In Deinococcus radiodurans Rsr is tethered via Y RNA to the
#=GF CC   exoribonuclease PNPase and channels single-stranded RNA into the PNPase
#=GF CC   cavity. Rsr and Y RNA enhance degradation of structured RNAs by PNPase.
#=GF CC   This role could be conserved, as Rsr and ncRNAs called YrlA and YrlB (Y
#=GF CC   RNA like) also associate with PNPase in an evolutionary distant bacterium
#=GF CC   Salmonella Typhimurium.[1]
#=GF WK   Y_RNA
#=GF SQ   5


AEYZ01000396.1/7368-7480   AUGGCGAAUGCAUGGGGAACUACA-----UCUGUGUCGCGAGUUCGAUUCUCGCCAGGGCCUGGUCCUGUAGCUCAG-UUGGUAGAGCA-AGCUG----UUCCCCGGACACUUGUCGUCAUGAC
ACEA01000002.1/40396-40510 AAAACAAAUGCCUUGGAAACUAC-------AUUGGUCGCAGGUUCGAGUCCUGCCCCGCUCCAAGCGGGUAGCUCAGUCUGGUAGAGCG--GUUAAACGUUUCCCAAAAACUCGUUGUUUGCUG
FQ312003.1/3704232-3704124 UUGGCGAAUGCCUGUGGAUCUAC-------AUUUGUUGCGGGUUCGAGACCCGUCAGAGCCCGGCUCUGUAGCUCAA--UGGUAGAGCG--GUGU----AUUCACAUAAACUUGUCGCCACCCU
CP002727.1/4011288-4011407 CCGGCGAAUGCCGGUGGAACUACAGGACGUGAAGGUUGCGGGUUCGACUCCCGCCCCGGC---GACGGGUAGCUCAACUGGAUAGAGCAACGCGCA-UGUUCCACCAACGAUCGUCGCCGCACU
ADBF01000255.1/32505-32391 AAAACAAAUGCUUUGGAAACUAC-------AUUGGUCGUAGGUUCGAGUCCUGCCCCGCUUUCGGCGGGUAGCUCAGUCUGGUAGAGCG--GUUAAACGUUUCCCAAAAACUCGUUGUUUGCUG
#=GC SS_cons               :((((((-----(((((((-(((.....,,,,,,,,<<<<<_____>>>>>,<<<<<<___>>>>>>,,<<<<_________>>>>,..)))--.--)))))))--------))))))::::
#=GC RF                    acggCgaAUGCcuggGgaaCcaC.......auugGUuGcgGGcUCGAggCCcgcCccGcccuaggCggGUAGCUCAauuuGgUAGAGCa..GugaA.uGuucCccaaAaAcUcGucGccguauu
//
```

```
AEYZ01000396.1/7368-7480    AUGGCGAAUGCAUGGGGAACUACA-----UCUGUGUCGCGAGUUCGAUUCUCGCCAGGGCCUGGUCCUGUAGCUCAG-UUGGUAGAGCA-AGCUG----UUCCCCGGACACUUGUCGUCAUGAC
ACEA01000002.1/40396-40510  AAAACAAAUGCCUUGGAAACUAC-------AUUGGUCGCAGGUUCGAGUCCUGCCCCGCUCCAAGCGGGUAGCUCAGUCUGGUAGAGCG--GUUAAACGUUUCCCAAAAACUCGUUGUUUGCUG
FQ312003.1/3704232-3704124  UUGGCGAAUGCCUGUGGAUCUAC-------AUUUGUUGCGGGUUCGAGACCCGUCAGAGCCCGGCUCUGUAGCUCAA--UGGUAGAGCG--GUGU----AUUCACAUAAACUUGUCGCCACCCU
CP002727.1/4011288-4011407  CCGGCGAAUGCCGGUGGAACUACAGGACGUGAAGGUUGCGGGUUCGACUCCCGCCCCGGC---GACGGGUAGCUCAACUGGAUAGAGCAACGCGCA-UGUUCCACCAACGAUCGUCGCCGCACU
ADBF01000255.1/32505-32391  AAAACAAAUGCUUUGGAAACUAC-------AUUGGUCGUAGGUUCGAGUCCUGCCCCGCUUUCGGCGGGUAGCUCAGUCUGGUAGAGCG--GUUAAACGUUUCCCAAAAACUCGUUGUUUGCUG
#=GC SS_cons                :((((((-----((((((((-(((........,,,,,,,<<<<<_____>>>>>,<<<<<<___>>>>>>,,<<<<__________>>>>,..)))--.--)))))))--------))))))::::
#=GC RF                     acggCgaAUGCcuggGgaaCcaC.......auugGUuGcgGGcUCGAggCCcgcCccGcccuaggCggGUAGCUCAauuuGgUAGAGCa..GugaA.uGuucCccaaAaAcUcGucGccguauu
//
```

```
                                          --Rsr ---              --------V--------------*
S. Typhimurium YrlA        1   ----------UUUUGGCGAAUGCCUGUGGA-UCUACAU-U------------UGUU
P. fulva                   1   ------------GCAGCGAAUGCCUUGGGU-ACUACAU-U-----------GUGUC
R. etli                    1   ----------CCGUGACGAAUGCCGGAUGAAACUACAGAUCGUUAAUCUCCUGGUC
G. polyisoprenivorans      1   -----CGAGAGCGGGCCGAA-GUCGAUCG--GUUACCA-CUCAUAAUGGGCAGGUU
S. roseum                  1   ---UGUCGGUACGGGUCGAA-GCGCGUCG--GUUAUCGCCUUCUAAGCGGGUGGUC
M. smegmatis               1   GCUUGGUGAUGUGAGCCGGACGUCGAUCG--GUUACCA-CUUCUAAUGGAAUGGUC
Mycobacteriophage          1   -------GAGGCAGGCCGAGAGUUUGUCG--GUUACCUCAUUGAAAGAGAGAGGUC
CONSENSUS                      ..............G.CGAA.GY..R..G..RYUAY...Y.............GUY

                               <<<<<    T    >>>>><<<<<          AS          >>>>>  <<<<  *
S. Typhimurium YrlA       33   GCGGGUUCGAGACCCGUCAGAGCCC---------------GGCUCUGUAGCUCAAU
P. fulva                  32   GCGGGUUCGAGCCCCGUCAGCCGC----------------AAGGCUGUAGCUCAGU
R. etli                   47   GCGGGUUCGAAUCCCGUCGGGUCCUCCAGUUCAU------GGACCCGUAGCUCAGC
G. polyisoprenivorans     48   GCGGGUUCGAGUCCCGCCAUCACCUCCG------------GGUGGUGUAGCUCAAC
S. roseum                 51   GCGGGUUCGAAUCCCGUCACAGAUUCGA------------GUCUGUGUAGCUCAGU
M. smegmatis              54   GCGGGUUCGAAUCCCGCCCGCCCUACGC------------AGGGCGGUAGCUCAGC
Mycobacteriophage         48   GCGGGUUCGAAUCCCGUCAGGUGGUUACCGGUACGCCCGGACACCUGUAGCUCAGU
CONSENSUS                      GCGGGUUCGARYCCCGYCR.........................YGUAGCUCARY

                                  D   >>>>
S. Typhimurium YrlA       74   -GGU-AGAGCGGUGUAU-------------UCACAUAAACUUGUCGCCACCCU-
P. fulva                  72   CGAU-AGAGCA-AAUGU--------------GCCCAUCAACUCGUCGCUGCAC--
R. etli                   97   -GGAUAGAGCAGGA------CGUU-----UCACCACACCUUGUCGUCACGG----
G. polyisoprenivorans     92   GGGU-AGAGCAGUGAAAACCCGGUCGGCAUAACAUG-------CCCGCUCUCG--
S. roseum                 95   -GGCCAGAGCGCCUAACGGCCGGUGUGCGUCCCCACAUCCGUGCCGGCUC-----
M. smegmatis              98   UGGC-AGAGCAUU-----CCCGGUUGAUACCUCACACGCCCACAUCACCGAGCUC
Mycobacteriophage        104   -GGU-AGAGCGCUUAAACGCCGGUGGACACCACCA-------UGUCUGCCUC---
CONSENSUS                      .GGY.AGAGCR.....................C...........YY..Y.....
```


T
AS
D
V
S. Typhimurium
YrlA
P. fulva
YrlA
R. etli
YrlA

Attempting to start a collaboration with the Wolin lab: Can we identify more Y RNAs?

1. Downloaded UniProt's collection of *reference proteomes* and associated genomes.
   - 8586 bacteria
   - 5152 viruses
   - 465 archaea
   - 1064 Eukaryota

2. Searched bacterial proteomes for TROVE (Ro and Rsr have this domain) and PNPase domains using HMMER and Pfam.

3. Searched bacterial genomes with current Rfam YrlA model (RF02553) using Infernal and Rfam.

TROVE domains are scattered around in bacteria

TROVE results

taxonomy	#assembly	#asbly w-hit	minevalue	numhits
Actinobacteria;	1144	175	2.4e-93	190
Proteobacteria;	2924	124	2.7e-84	129
Bacteroidetes;	804	75	3.4e-28	80
Cyanobacteria;	153	40	2.5e-43	40
Firmicutes;	1454	30	1.7e-45	31
Planctomycetes;	85	15	3e-87	16
Deinococcus-Thermus;	25	6	4e-56	6
Verrucomicrobia;	57	5	3.7e-88	6
Armatimonadetes;	17	5	9.2e-88	5
Acidobacteria;	66	5	5.2e-12	5
unclassified_Parcubacteria_group;	144	1	4.1e-14	1
Synergistetes;	18	1	1.8e-11	1
Spirochaetes;	67	1	2.9e-12	1
Nitrospinae/Tectomicrobia_group;	11	1	2.3e-49	1
Chloroflexi;	145	1	1.5e-16	1
Chlamydiae;	24	1	8.6e-85	1
Candidatus_Woesebacteria;	52	1	3.2e-14	1
Candidatus_Uhrbacteria;	25	1	2.4e-15	1
Candidatus_Sungbacteria;	11	1	3e-70	1
Candidatus_Adlerbacteria;	7	1	8.7e-77	1
Balneolaeota;	2	1	4.2e-20	1

22 'phyla' have at least 1 TROVE hit;
123 'phyla' have 0 TROVE hits;

126 'phyla' have at least 1 PNPase hit;
18 'phyla' have 0 PNPase hits;

PNPase results

taxonomy	#assembly	#asbly w-hit	minevalue	numhits
Proteobacteria;	2924	2701	3.7e-25	2863
Firmicutes;	1454	1254	1.6e-17	1263
Actinobacteria;	1144	1077	7.2e-20	1083
Bacteroidetes;	804	745	5.3e-13	749
Cyanobacteria;	153	142	1.6e-20	146
Chloroflexi;	145	78	5.9e-15	79

Soyeong has found Y RNAs in many bacteria

Soyeong's results	taxonomy
119 of 140	Proteobacteria;
50 of 59	Actinobacteria;
24 of 32	Firmicutes;
22 of 29	Cyanobacteria;
19 of 27	Bacteroidetes;
8 of 8	Planctomycetes;
2 of 3	Deinococcus-Thermus;
2 of 2	Verrucomicrobia;
2 of 2	Acidobacteria;
0 of 1	Chloroflexi;
1 of 1	Armatimonadetes;
0 of 1	Synergistetes;
0 of 1	Chlamydiae;

Rfam model needs improvement

		RF02553 results (only count hits with E-value < 0.1)						
			#assembly-w-hit				number of hits	
Soyeong's results	taxonomy	#assembly	TROVE	RNA	both	minEvalue	TROVE	RNA
119 of 140	Proteobacteria;	2924	124	128	104	1.9e-24	129	257
50 of 59	Actinobacteria;	1146	175	113	105	9.9e-09	190	129
24 of 32	Firmicutes;	1455	30	28	15	3.5e-05	31	31
22 of 29	Cyanobacteria;	153	40	1	0	0.024	40	1
19 of 27	Bacteroidetes;	804	75	60	34	1.8e-05	80	73
8 of 8	Planctomycetes;	85	15	12	10	1.7e-13	16	18
2 of 3	Deinococcus-Thermus;	25	6	2	1	0.00042	6	2
2 of 2	Verrucomicrobia;	57	5	5	5	3.4e-15	6	13
2 of 2	Acidobacteria;	66	5	5	5	0.00031	5	7
0 of 1	Chloroflexi;	145	1	2	0	0.079	1	2
1 of 1	Armatimonadetes;	17	5	2	2	7.7e-05	5	2
0 of 1	Synergistetes;	18	1	0	0	-	1	0
0 of 1	Chlamydiae;	24	1	0	0	-	1	0
-	unclassified_Parcubacteria_group;	144	1	0	0	-	1	0
-	Spirochaetes;	67	1	1	0	0.031	1	1
-	Candidatus_Woesebacteria;	52	1	0	0	-	1	0
-	Candidatus_Uhrbacteria;	25	1	0	0	-	1	0
-	Nitrospinae/Tectomicrobia_group;	11	1	0	0	-	1	0
-	Candidatus_Sungbacteria;	11	1	0	0	-	1	0
-	Candidatus_Adlerbacteria;	7	1	0	0	-	1	0
-	Balneolaeota;	2	1	1	1	0.057	1	1

Acknowledgements

NCBI

David Landsman
Alejandro Schäffer

Janelia	**EBI (Rfam)**
Sean Eddy	Alex Bateman
Elena Rivas	Rob Finn
Travis Wheeler	Anton Petrov
Tom Jones	Ioanna Kalvari
Diana Kolbe	Joanna Argasinska
Seolkyoung Jung	Paul Gardner
Rob Finn	Sarah Burge
Jody Clements	Evan Floden
Fred Davis	John Tate
Lee Henry	Jen Daub
Michael Farrar	

Applications of CMs

- homology search/alignment: Infernal, COVE, Rfam[*], Alternal[†], RNATOPS[‡]
- RNA discovery: CMfinder[§], Zasha's pipeline(s)[¶]
- structure comparison: CMCompare[‖]
- family-specific programs:
  - tRNAscan-SE[**],
  - 16S/18S rRNA alignment: SSU-ALIGN[††]
  - bacterial terminator identification: RNIE[‡‡]

[*]E. P. Nawrocki, S. W. Burge et. al. NAR, 43:D130-D137, 2015.

[†]S. Janssen and R. Giegerich. BMC Bioinformatics 2015, 16:178

[‡]Z. Huang et. al, Bioinformatics, 24(20), 2281-2287, 2008.

[§]Z. Yao, Z. Weinberg, W. L. Ruzzo, Bioinformatics 2006, 22(4), 445-452.

[¶]Z. Weinberg, Z et. al. Nucleic acids research, 2007. 35(14), 4809-4819, Z. Weinberg et. al. Genome Biol, 2010. 11(3), R31.

[‖]C. H. zu Siederdissen, and I. L. Hofacker Bioinformatics, 2010. 26(18), i453-i459.

[**]T. M. Lowe, S. R. Eddy. NAR, 25:955-964, 1997.

[††]E. P. Nawrocki. PhD Thesis: 2009, Washington University School of Medicine

[‡‡]P.P. Gardner et. al. Nucleic acids research, 2011, 39(14), 5845-5852.