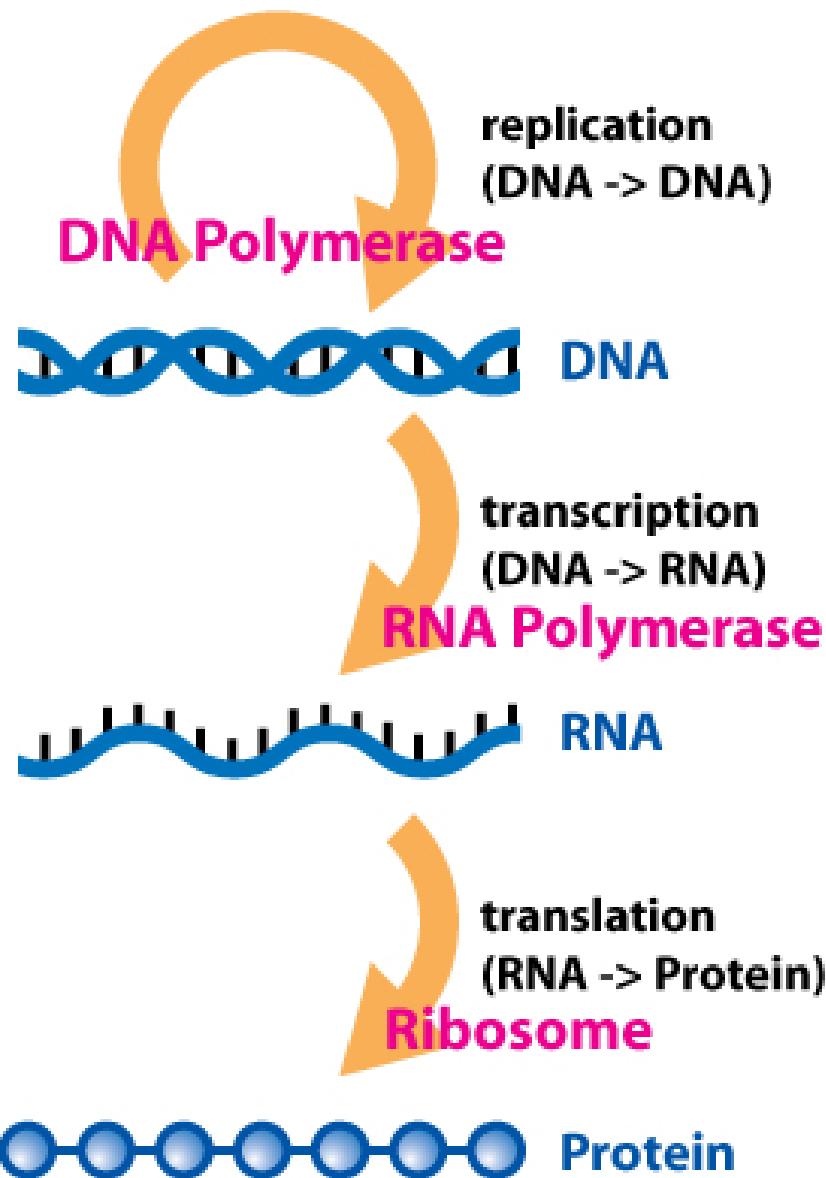


RNA sequence analysis using profiles

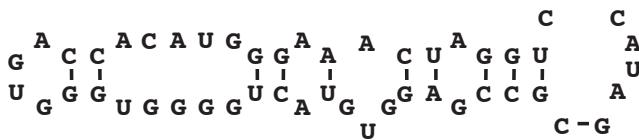
Eric Nawrocki



Central dogma of molecular biology



5S ribosomal RNA (119 nt)



3'
5'

A

G-C

C-G

C-G

U-A

C-G

A-A

U-A

G-C

C-G

C-G

U-A

G-C

U-A

U-A

G-C

G-C

A-U

C-A

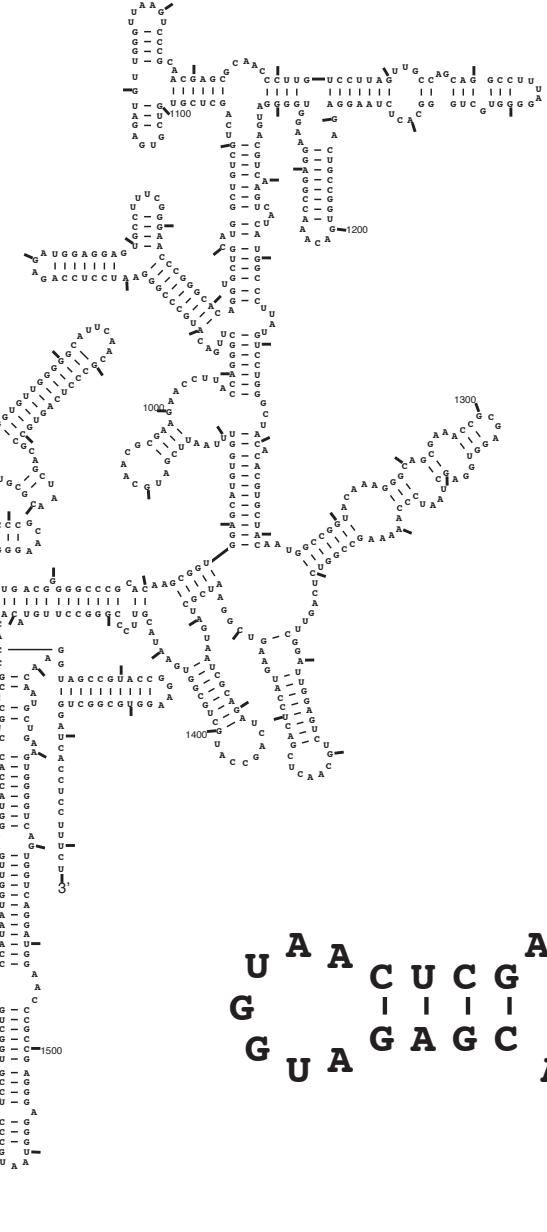
U-A

U-A

U-A

U-A

transfer
RNA (71 nt)



small subunit
ribosomal RNA
(SSU rRNA, 1582 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
splicing	U1, U2, U4, U5, U6			x
other	telomerase RNA			x
	Y RNA			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
splicing	U1, U2, U4, U5, U6			x
other	telomerase RNA			x
	Y RNA			x
	Vault RNA			x
	many more...			



database of more than 3000 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

Structure of a Ribonucleic Acid

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.

STRUCTURE OF AN ALANINE RNA

ROBERT W. HOLLEY, JEAN APgar

GEORGE A. EVERETT

JAMES T. MADISON

MARK MARQUISEE, SUSAN H. MERRILL

JOHN ROBERT PENSWICK, ADA ZAMIR

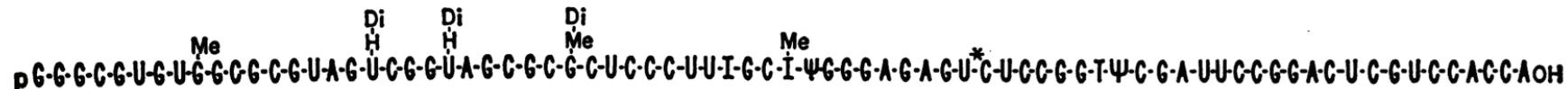
U.S. Plant, Soil, and Nutrition

Laboratory, U.S. Department of

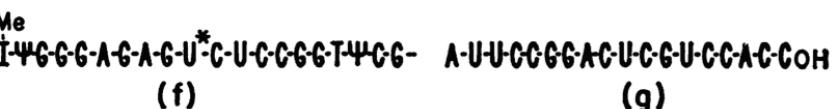
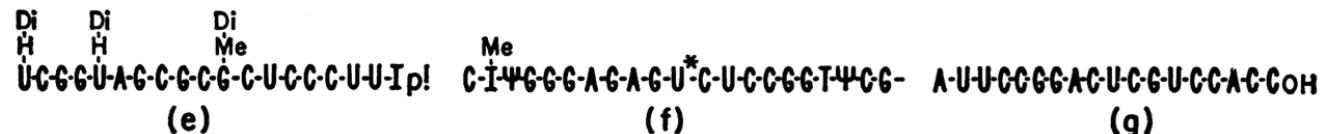
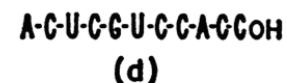
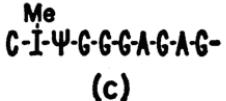
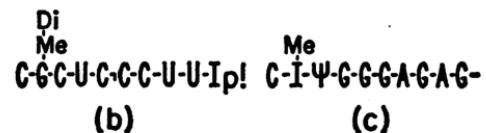
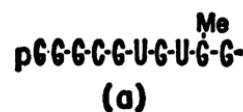
Agriculture, and

Department of Biochemistry,

Cornell University, Ithaca, New York



LARGE OLIGONUCLEOTIDE FRAGMENTS



(g)

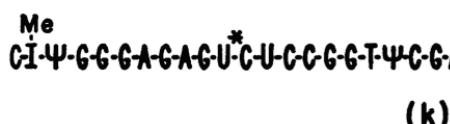
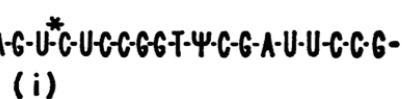
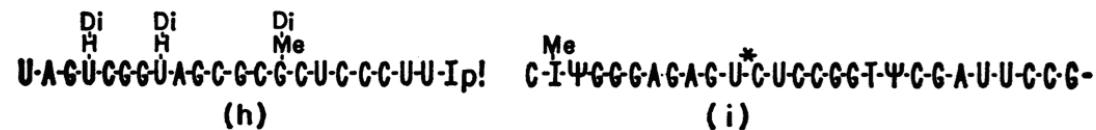


Fig. 1. The structure of an alanine transfer RNA, isolated from yeast, is shown at the top. Large oligonucleotide fragments that were crucial in the proof of structure are shown below.

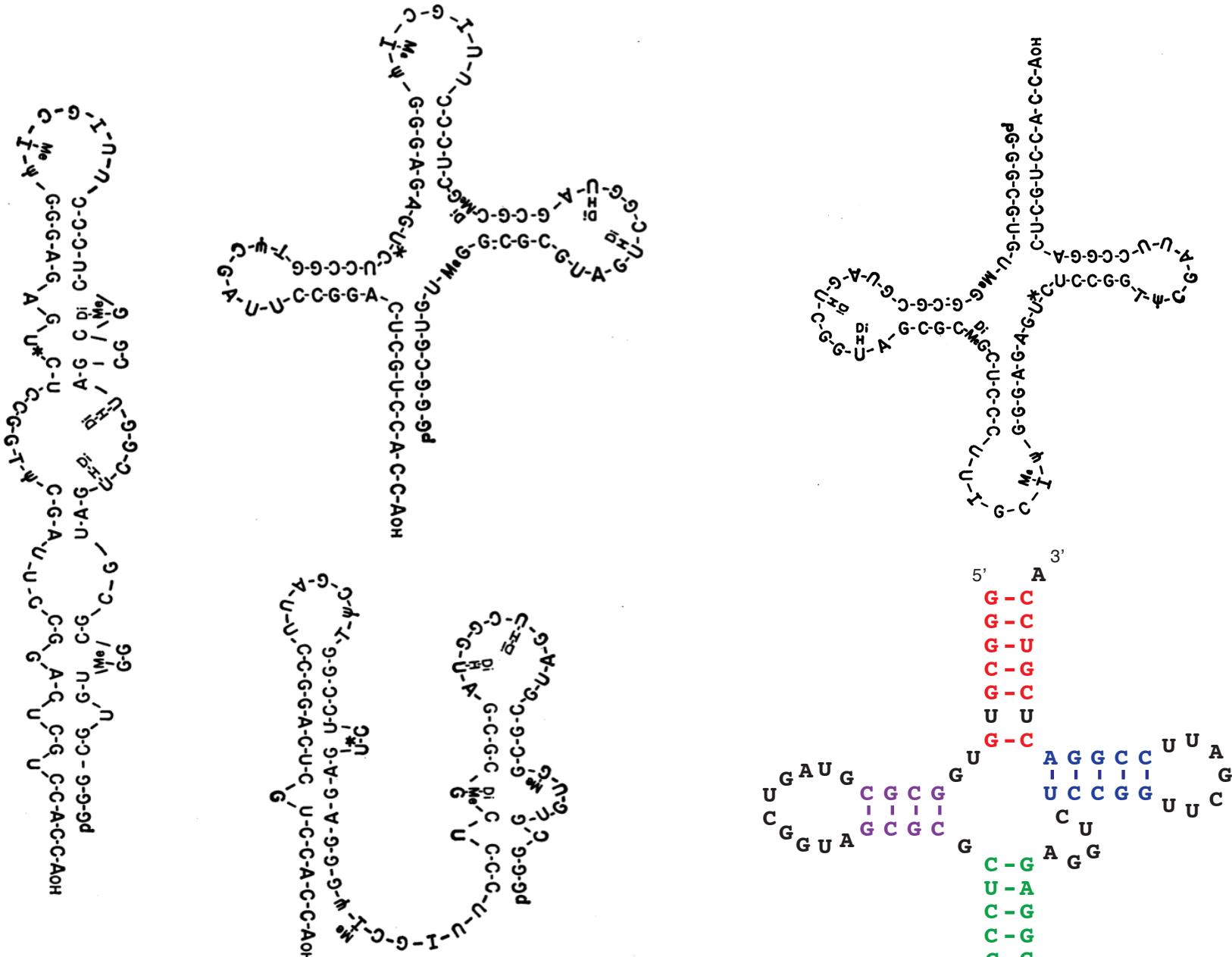


Fig. 2. Schematic representation of three conformations of the alanine RNA with short, double-stranded regions.

```

struct (((((((..<<<.....>>>.<<<<.....>>>>....<<<<.....>>>>))))).
Ala   GGGCGUGUGGCGCGUAGUCGGUAGCGCGCUCCCUUAGCAUGGGAGAGGUCCCGGUUCGAUUCGGACUCGUCCA
  
```

```

struct (((((((..<<<.....>>>. <<<<.....>>>>....<<<<.....>>>>) ))))).  

Ala GGGCGUGUGGCGCGGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCGGACUCGUCCA  

Tyr CUCUCGGUAGCA..AGUUGGUUUAAGGCAAGACUGUA ..UCUUGAGAUCGGCGUUCGACUCGCCCCCCGGGAGA  

Val GGUUUUCGUGGGCUU..AGUCGGU.UAUGGCAUCUGCUUAACACGGCAGAACGUCCCAGUUCGAUCCUGGGCGAAAUCA  

Iln GGUCUCUUGGCCCC..AGUUGGU.UAAGGCACCGUGCUAAUAACGGGGAUCAGCGGUUCGAUCCCGCUAGAGAC  

Glu UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCGGUAUCGG  

identical * * * * *** * * * ** * ***** *
```

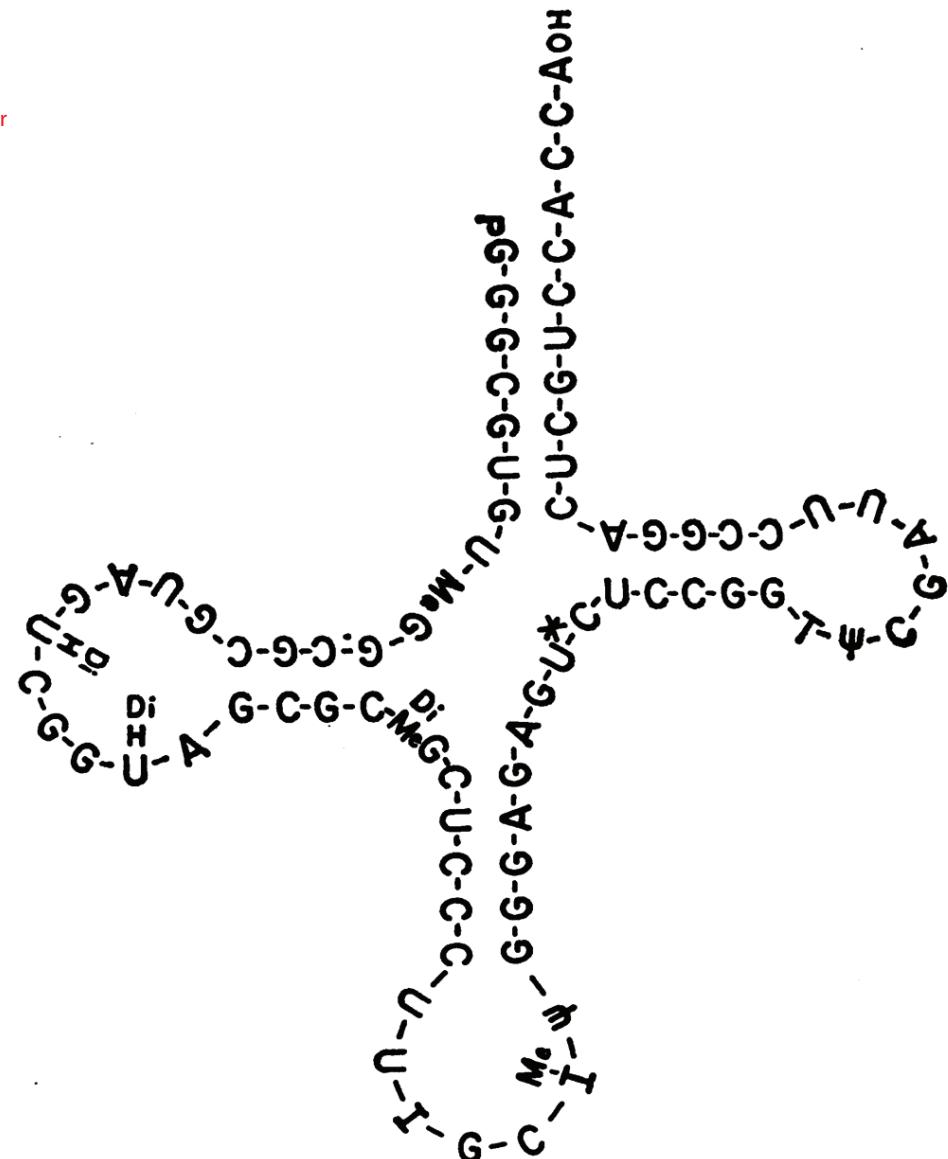
>0 non-WC ((((((..<<<.....>>>. <<<<.....>>>>....<<<<.....>>>>)))))).

Alignment color legend:

Black: Watson-Crick or GU/UG basepair

Red: non-Watson-Crick and non-GU/UG basepair

Grey: not basepaired



```

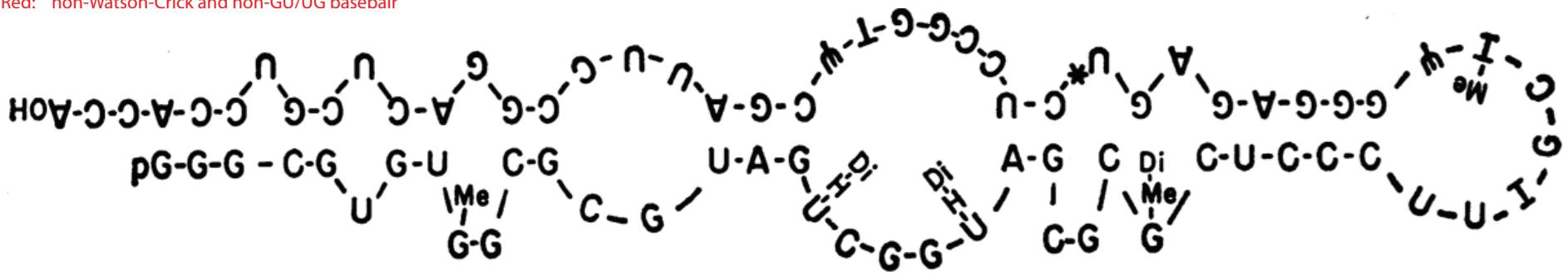
struct <<<<.<<..<<..<<<.....<<..<.<<<<.....>>>>.>..>>....>>...>>.>>.>>
Ala GGGCGUGUGGC CGCUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCCGGACUCGUCCA
Tyr CUCUCGGUAGCCA..AGUUGGUUUAGGC GCAAGACUGUA..UCUUGAGAU CGGGCGUUCGACUCGCCCCCGGGAGA
Val GUUUUCGUGGU CU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCAGUUCGAUCCUGGGCGAAAUC
Iln GGUCUCUUGGCC..AGUUGGU.UAAGGCACC GUGCUUUAACCGCGGGGAU CAGCGGUUCGAUCCCGCUAGAGACCA
Glu UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCAC GCUUUCACCGUGGGAGA.CC GGGGUUCGA CUC CCCGU AUCGGAG
identical * * * * * *** * * * * * * * * * * * * * * * * * *
>0 non-WC <<<<.<<..<<..<<<.....<<..<.<<<<.....>>>>.>..>>....>>...>>.>>.>>

```

Alignment color legend:

Black: Watson-Crick or GU/UG basepair

Red: non-Watson-Crick and non-GU/UG basepair

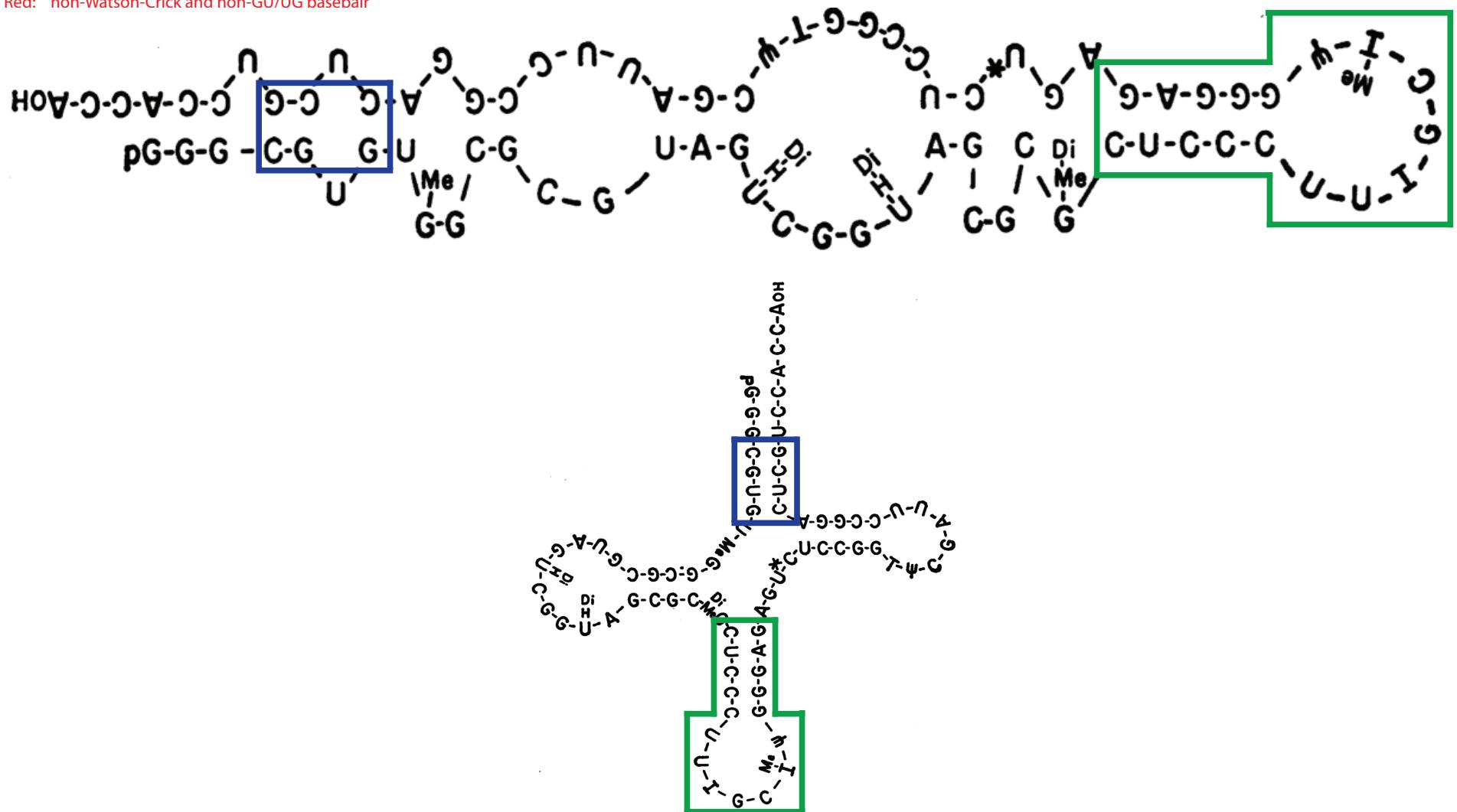


struct <<<<.<<..<<..<<<.....<<..<.<<<<.....>>>>.>..>>.....>>...>>.>>.>>
Ala GGGCGUGUGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCGGACUCGUCCA
Tyr CUCUCGGUAGCCA..AGUUGGUUUAGGCGCAAGACUGUA..UCUUGAGAU CGGGCGUUCGACUCGCCCGGGAGA
Val GGUUUCGUGGUUCU..AGUCGGU.UAUGGCAUCUGCUEAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUC
Iln GGUCUCUUGGCC..AGUUGGU.UAAGGCACCGUGCUUAAACGCGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu UCCGAUAUAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGGAGA.CCAGGGGUUCGACUCCCCCGUUAUCGGAG
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.UCUCCGGUUCGAUUCGGACUCGUCCA
Tyr       CUCUCGGUAGCCA..A GUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGGCGUUCGACUCGCCCCCGGGAGA
Val       GGUUUCGUGGGCU..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln       GGUCUCUUGGCCCC..AGUUGGU.UAAGGCACCGUGCCUAAAACGCGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA
Glu       UCCGAUUAAGUGU..AAC.GGC.UAUCACAUCACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCGUAUCGGAG
identical *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *
>0 non-WC <<<.<....<<<.....>>>.>.>>>.....<<<<<...<<<<.....>>>>>>.>>>.

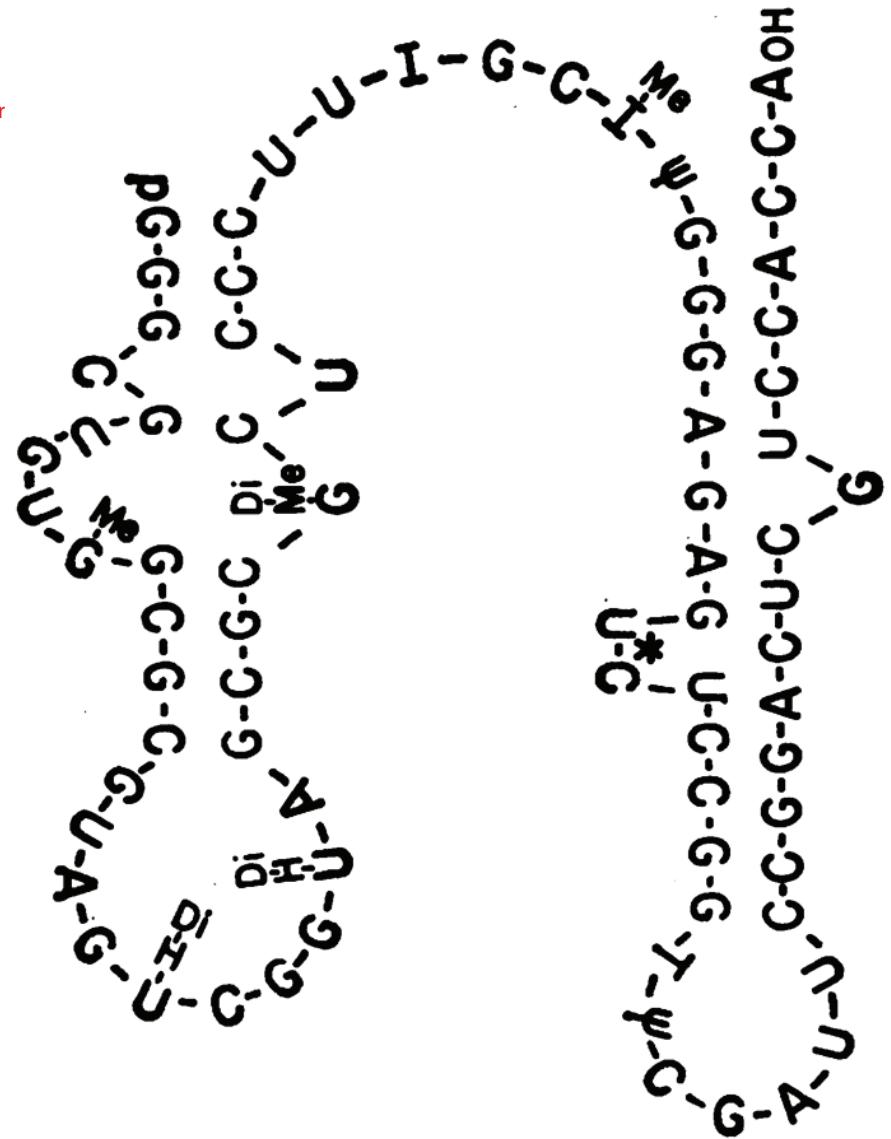
```

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.UCUCCGGUUCGAUUCGGACUCGUCCA
Tyr CUCUCGGUAGCA..A GUUGGUUUAAGGCGCAAGACUGUA..UCUUGAGAUCGGCGUUCGACUCGCCCGGGAGA
Val GGUUUUCGUGGGUC..AGUCGGU.UAUGGCAUCUGCUUAACACGCAGACGUCCCCAGUUCGAUCCUGGGCGAAAUCA
Iln GGUCUCUUGGCC..AGUUGGU.UAAGGCACCGUGCUAAAACGCGGGGAUCAGCGGUCGAGACA
Glu UCCGAUUAAGUGU..AAC.GGC.UAUCACAUCAGCUUCACGGUCGAUCGGA
identical * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
>0 non-WC <<<.<....<<<.....>>>.>.>>>.....<<<<<...<<<<.....>>>>>>.>>>.
clover <<<.....>>>
overlap <<<.....>>>

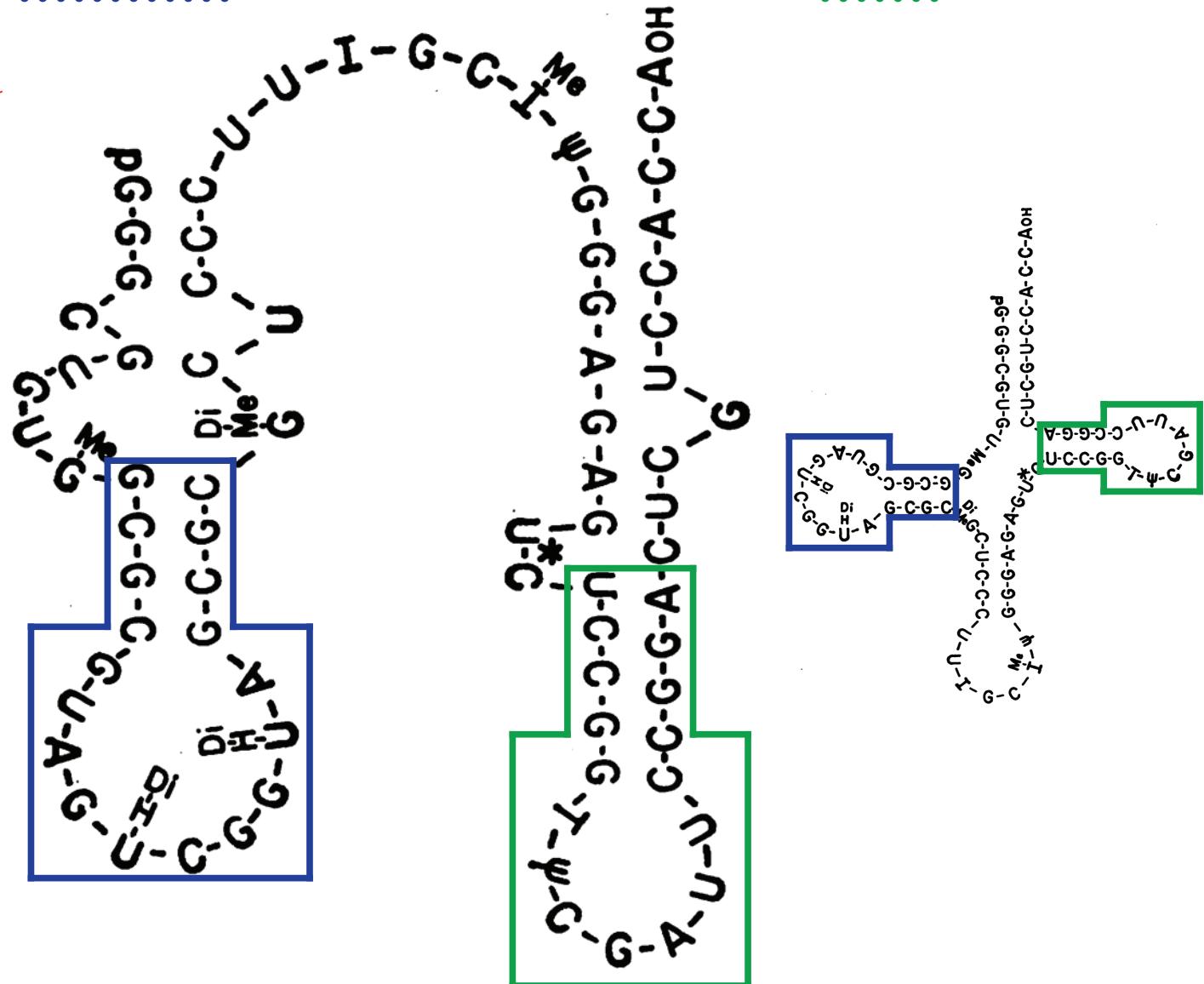
```

Alignment color legend:

Black: Watson-Crick or GU/UG basepair

Red: non-Watson-Crick and non-GU/UG basepair

Grey: not basepaired



```

struct ((((((..<<<.....>>>.<<<<.....>>>>.....<<<<.....>>>>) )))).  

Ala GGGCGUUGGGCGCGUAGUCGGU..AGCGCGCUCCCUUAGCAUGGGAGAG.UCUCCGGUUCGAUUCGGACUCGUCCA  

Tyr CUCUCGGUAGGCCA..AGUUGGUUAAGGCAGAACUGUA..UCUUGAGAUCGGCGUUCGACUCGCCCGGGAGA  

Val GGUUUUCGUGGUCU..AGUCGGU.UAUGGCAUCUGCUUAACACCGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAA  

Iln GGUCUCUUGGCC..AGUUGGU.UAAGGCACCGUGCUAAUAAACGCGGGAUCAGCGGUUCGAUCCCGCUAGAGACCA  

Glu UCCGAUAUAGUGU..AAC.GGC.UAUCACAUACGCUUUCACCGUGGAGA.CCGGGGUUCGACUCCCCGUAUCGGAG  

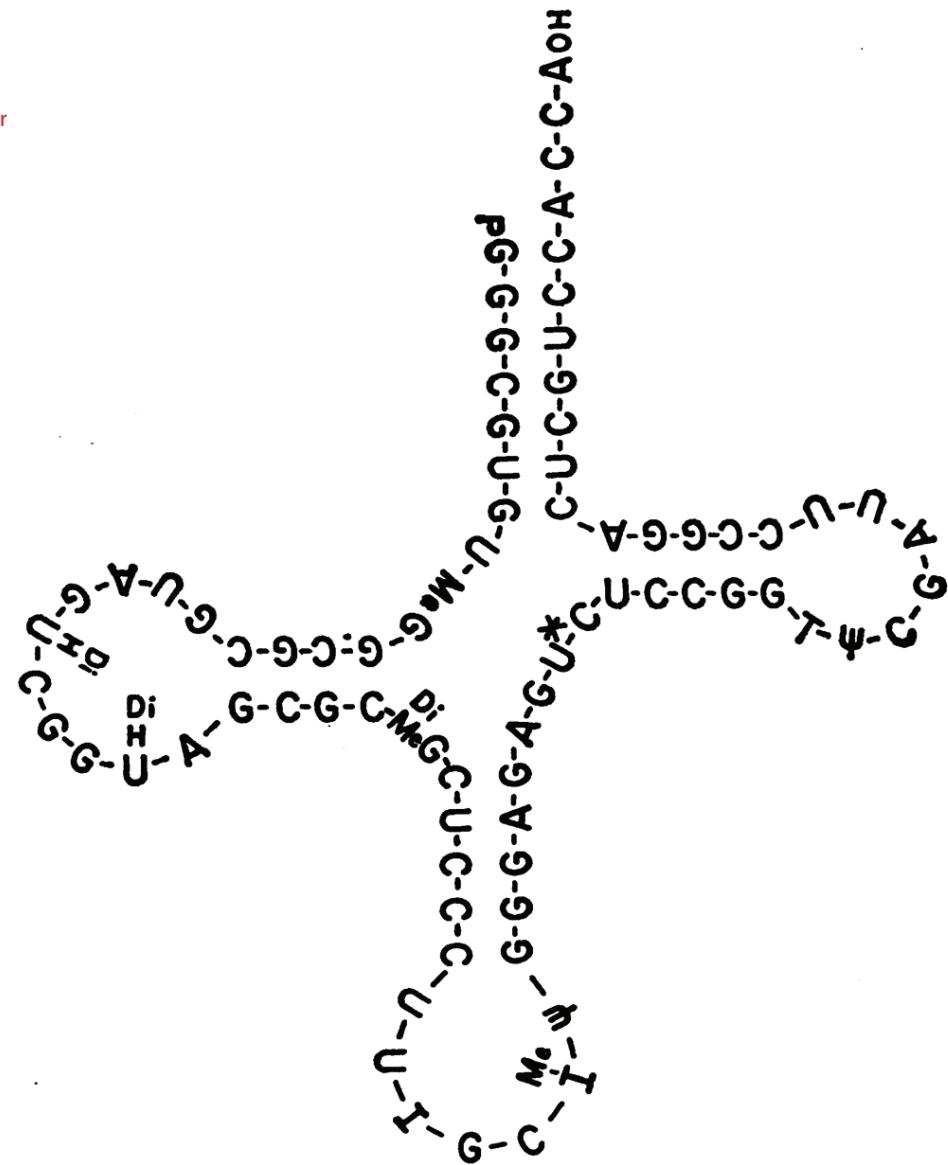
identical * * * * *** * * * * ** * ***** *
>0 non-WC (((((..<<<.....>>>.<<<<.....>>>>.....<<<<.....>>>>) )))).
```

Alignment color legend:

Black: Watson-Crick or GU/UG basepair

Red: non-Watson-Crick and non-GU/UG basepair

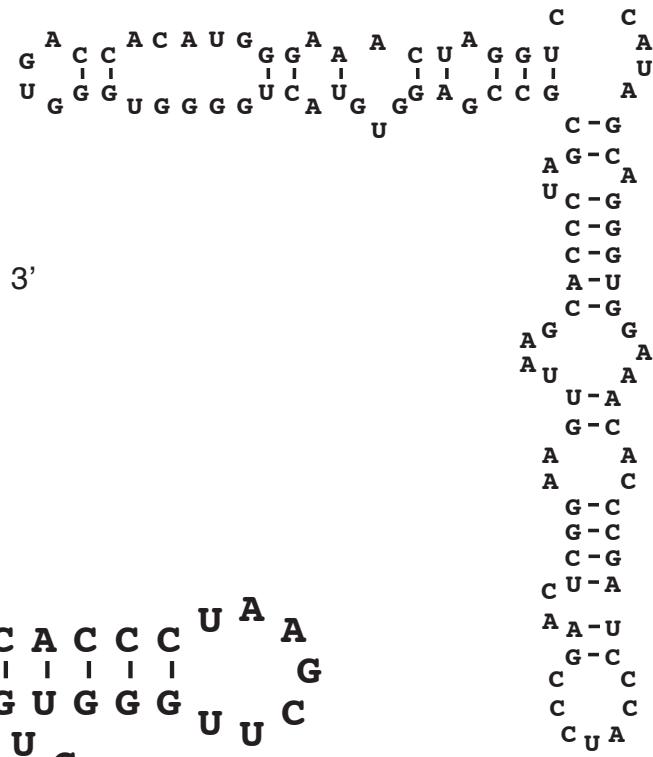
Grey: not basepaired



3'
U
C - G
G - C
A - U
C - G
C - G
C - G
C - G
C - G
C - G
5'

5S rRNA: 1975

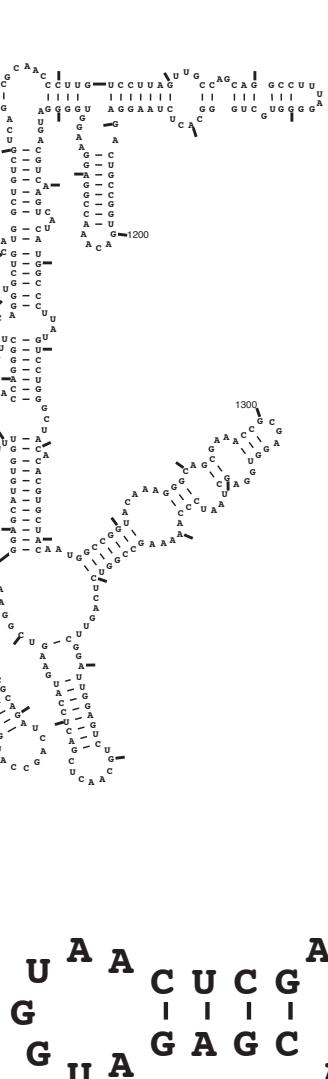
Fox, George E, and Carl R. Woese.
"5S RNA secondary structure."
Nature 256.5517 (1975): 505-507.



tRNA: ~1966

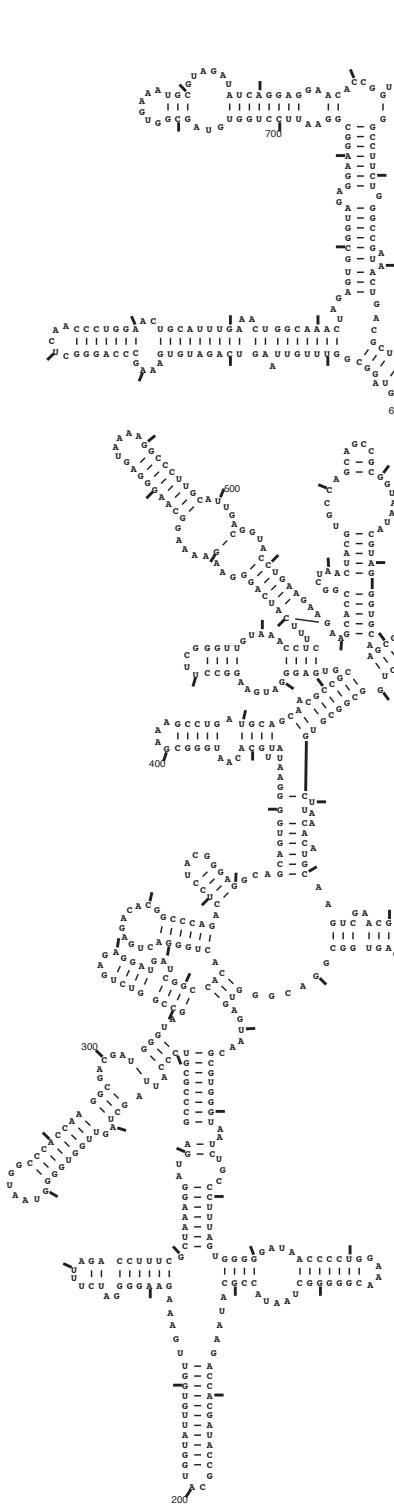
Holley, Robert W., et al. "Structure of a ribonucleic acid." Science 147.3664 (1965): 1462-1465.

3'
A
G - C
G - C
C - G
C - G
U - A
G - C
A - A
A - U
G - C
C - G
C - G
U - A
G - C
C - G
C - G
A - U
C - G
A - A
G - C
G - C
C - G
C - G
C - G
C - G
C - G
C - G
5'



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.

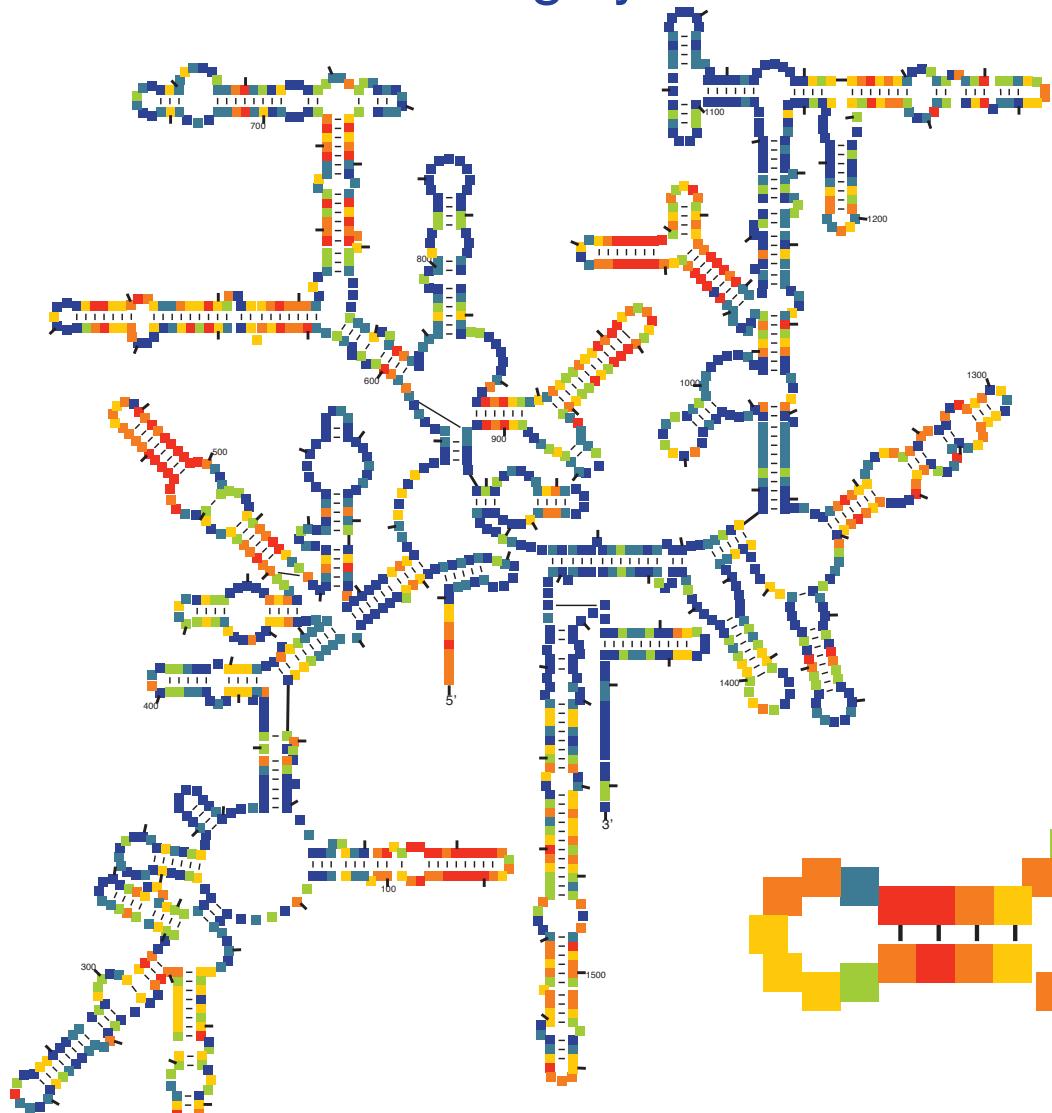


Comparative sequence analysis of homologs informs biologists

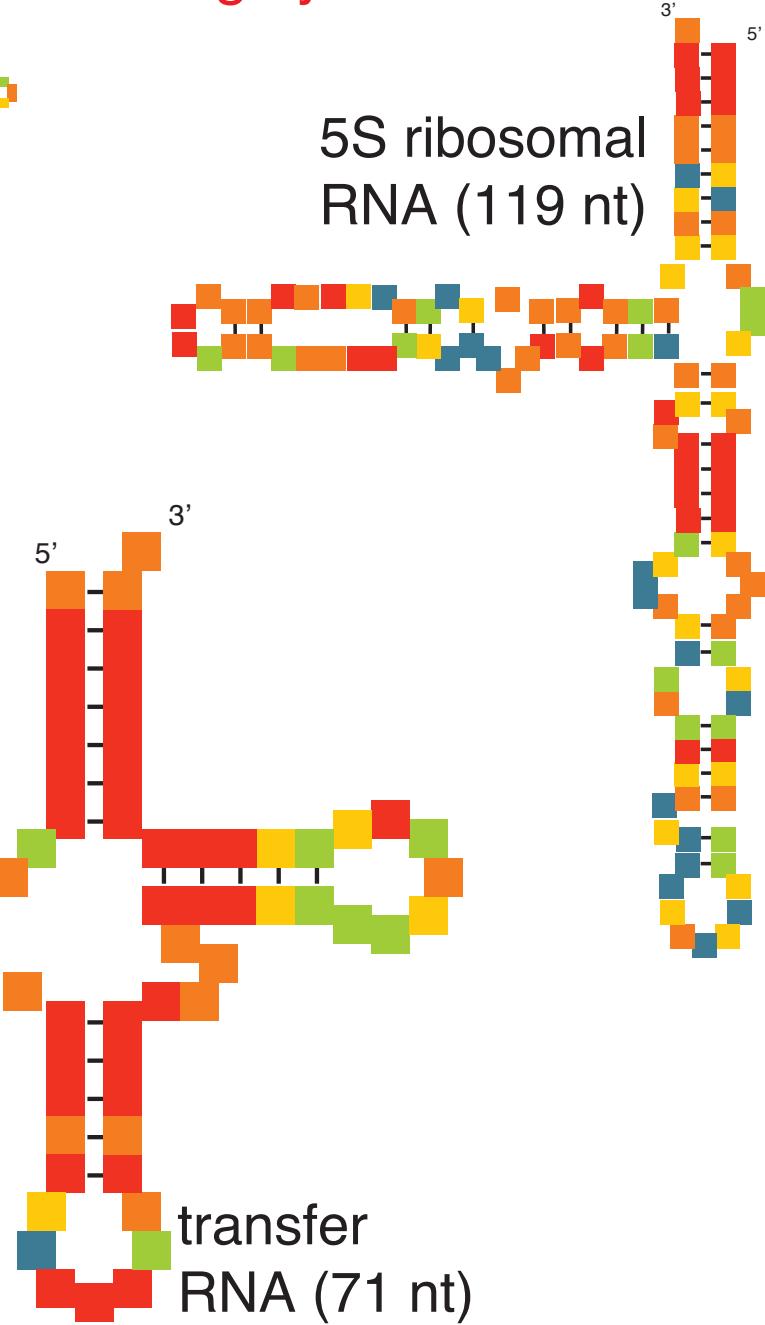
- Inference of structure
- Inference of phylogeny of organisms
- Inference of functional regions based on conservation levels

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

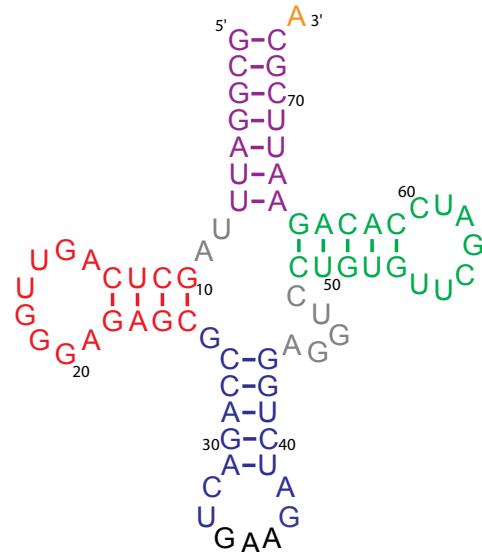
**Computational homology search methods use
one or more known family members to find additional homologs.**

How do we find structural RNA homologs?

Primary sequence

GC₁GGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUUCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



3-dimensional structure



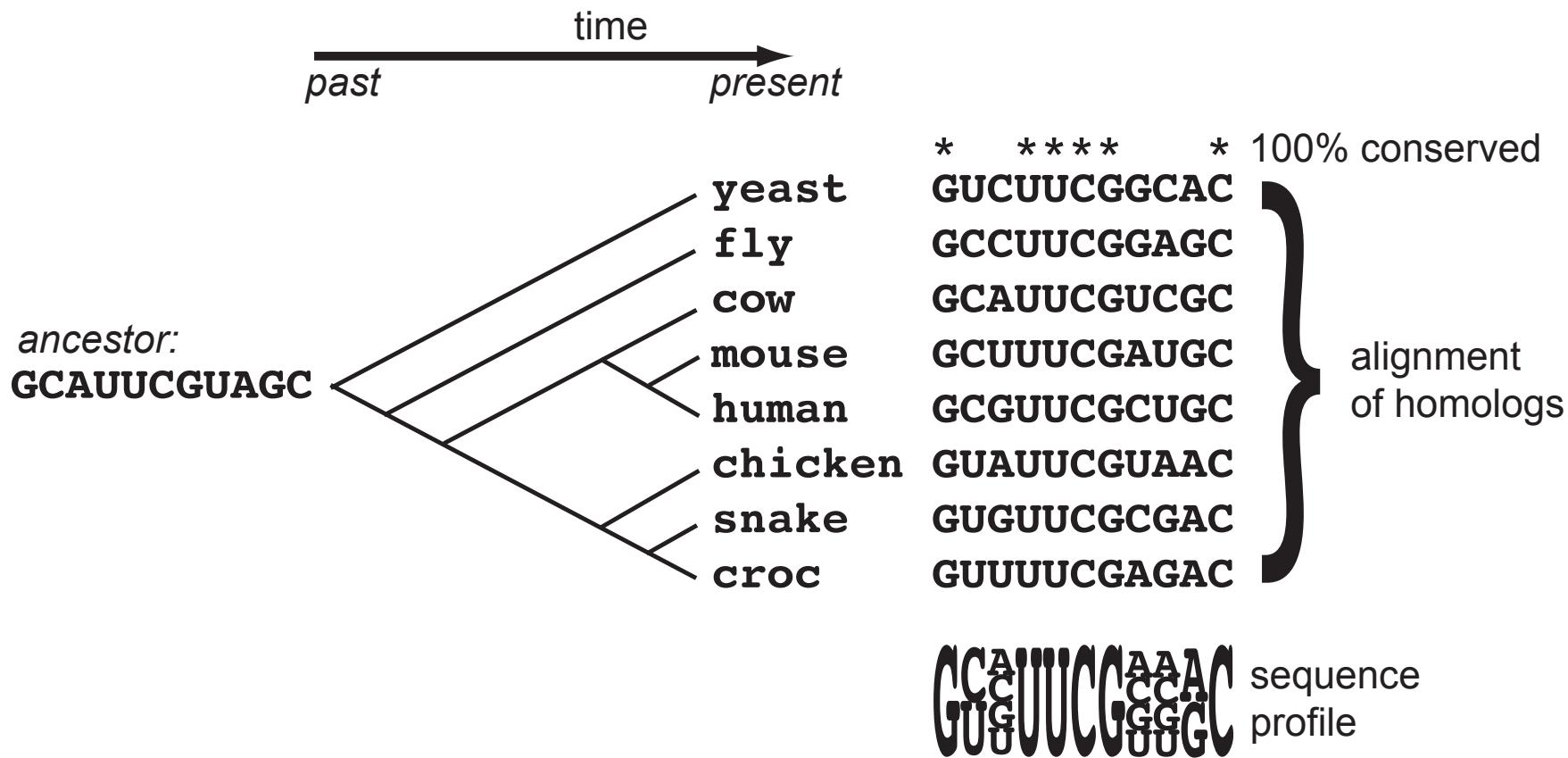
- BLAST: given a single sequence, search genomes for similar sequences.
- Structural RNAs are difficult to find
 - short (~ 100 nt) and evolve rapidly at sequence level
 - lack open reading frames
 - small, 4 letter alphabet
- BLAST cannot take advantage of:
 - sequence conservation, which varies across the gene
 - secondary structure

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

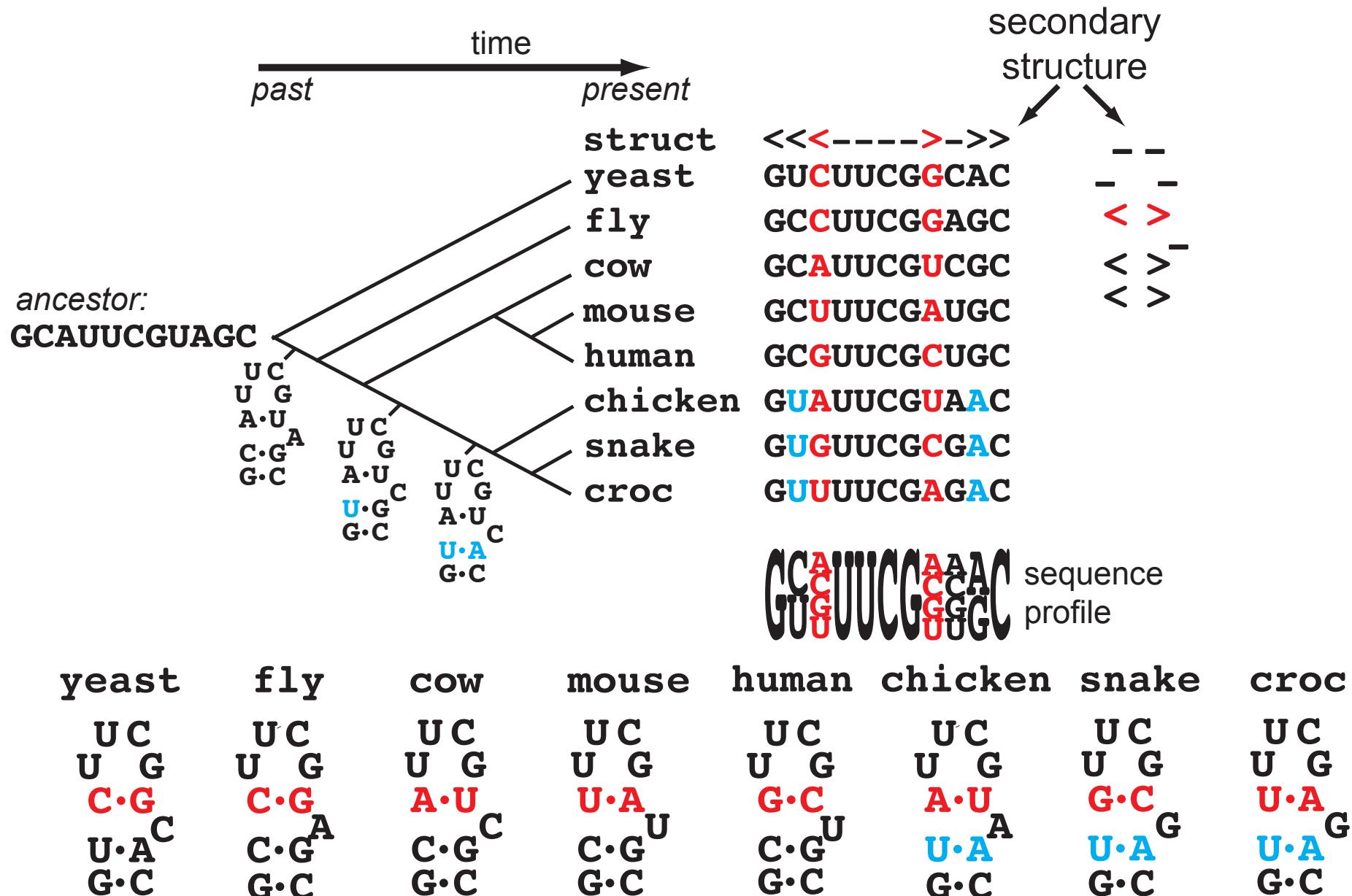
Sequence conservation provides information for homology searches

Conservation levels vary across alignment columns.

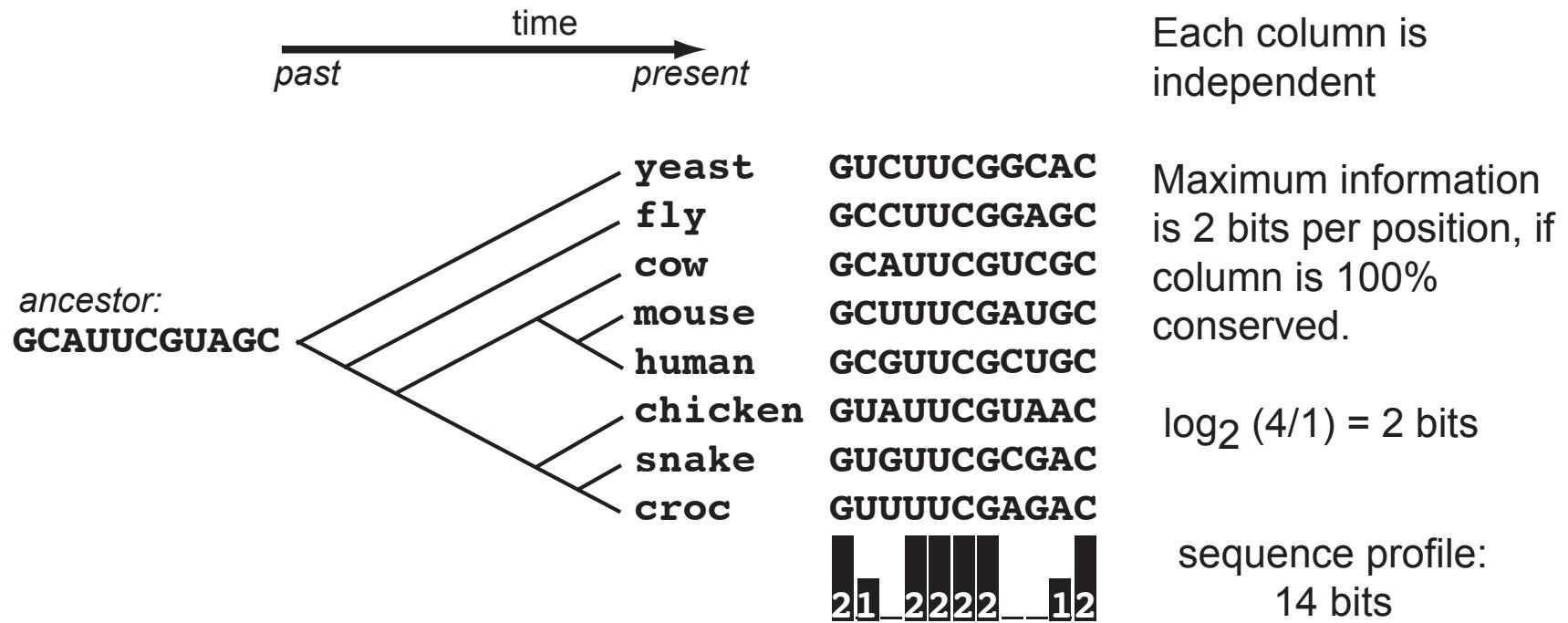


Structure conservation provides additional information

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

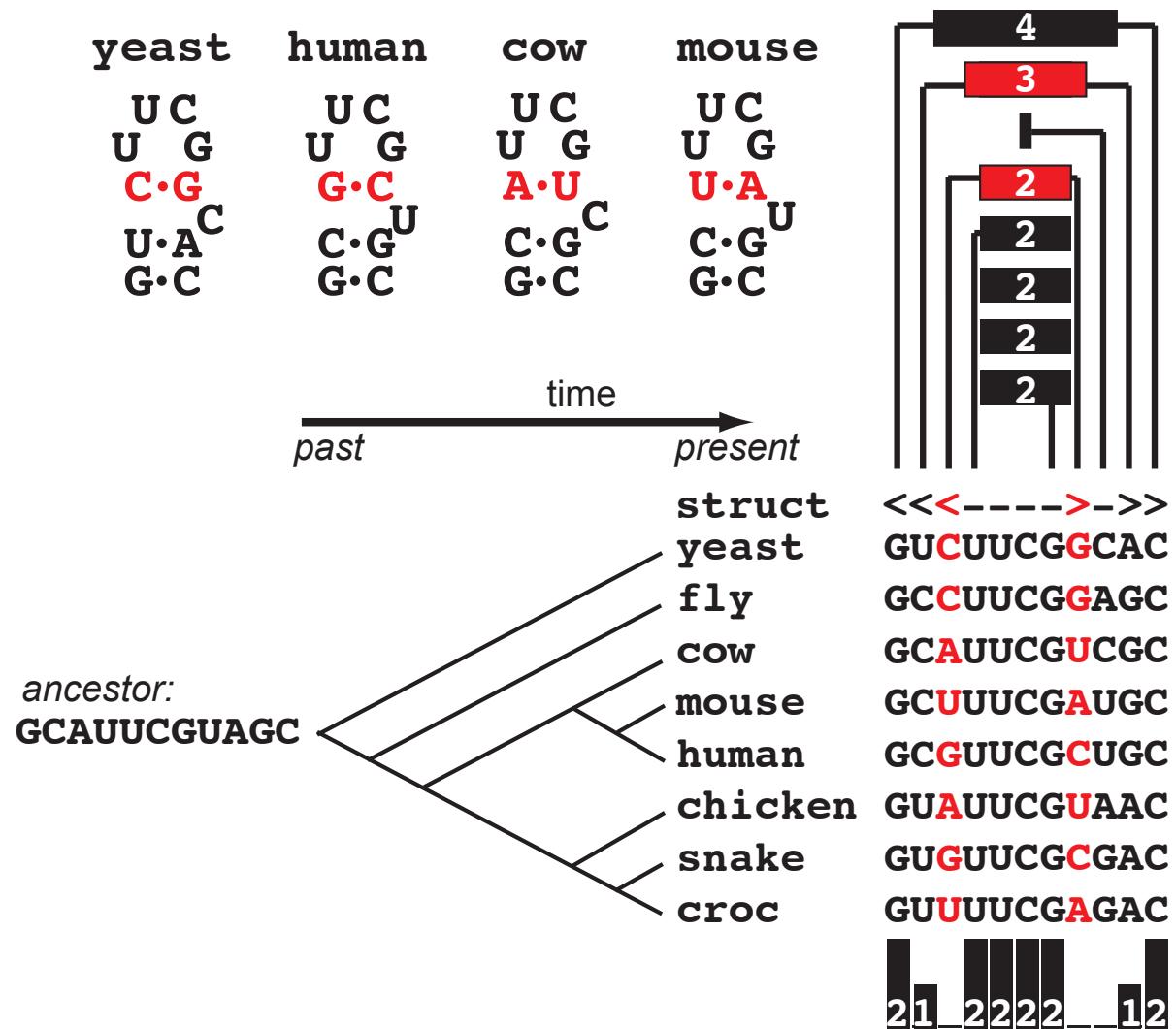


Amount of information in a profile can be measured in bits



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

Structure contributes additional information from covariation



sequence + **structure**
profile: 17 bits

Base-paired columns
are not independent

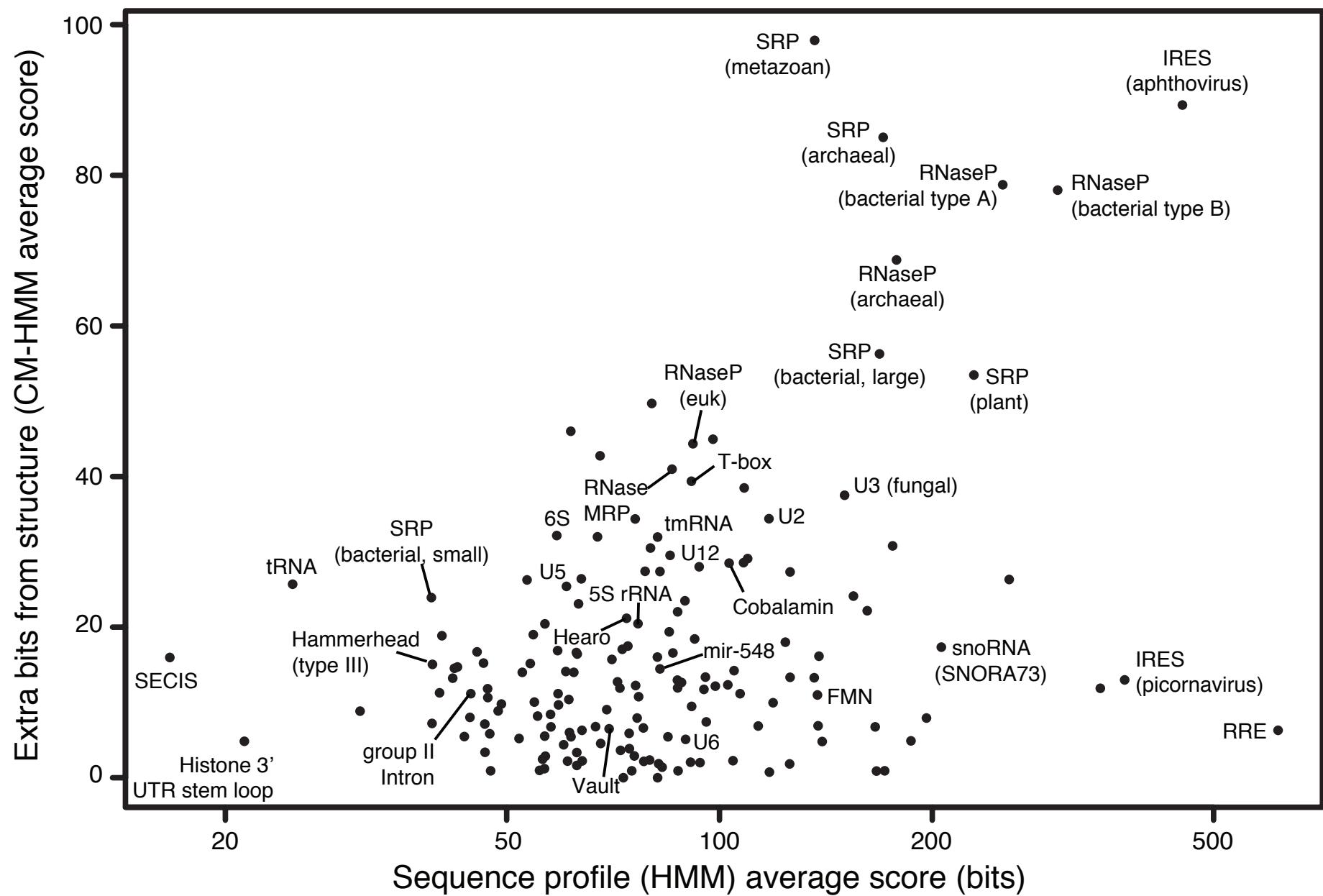
Maximum **extra** info:
2 bits per base pair

$$\log_2 (16/4) = 2 \text{ bits}$$

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 (17929 and 6915 entries)	Rfam (3016 families)
performance for RNAs	faster but less accurate	slower but more accurate

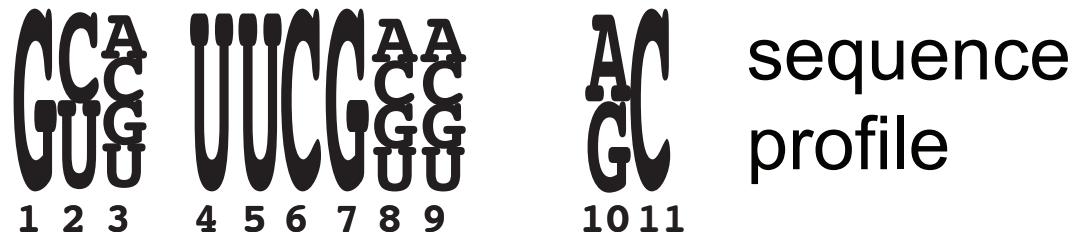


<http://hmmer.janelia.org>
Eddy, SR. PLoS Comp. Biol.,
7:e1002195, 2011.
Eddy, SR. PLoS Comp. Biol.,
4:e1000069, 2008.
Eddy, SR. Bioinformatics,
14:755-763, 1998.



<http://infernal.janelia.org>
Nawrocki EP, Eddy SR.
Bioinformatics, 29:
2487-2489, 2013.
Eddy SR, Durbin R.
Nucleic Acids Research,
22:2079-2088, 1994.

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



yeast	GU <u>C</u> aUU <u>C</u> G <u>G</u> C...AC
fly	GCC.UU-GGA...GC
cow	GCA.UUCGUC...-C
mouse	GCA.UU-GAU...GC
human	GCGaUU <u>C</u> GU...GC
chicken	GUA.UUCGUA...AC
snake	GUGaUU <u>C</u> CG...AC
croc	GUU.UU-GAG...AC
worm	G-G.UUCGCGcc <u>A</u> C
starfish	G-U.UUCGAU...-C

Profiles HMMs are probabilistic profiles built from alignments

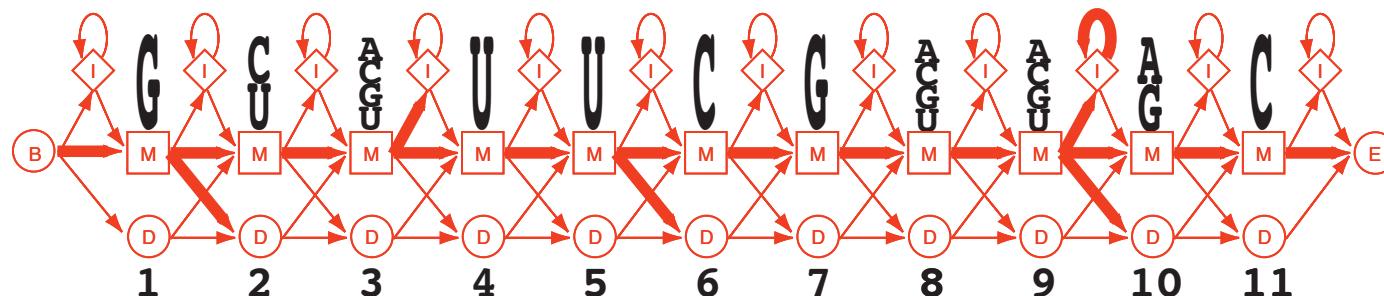
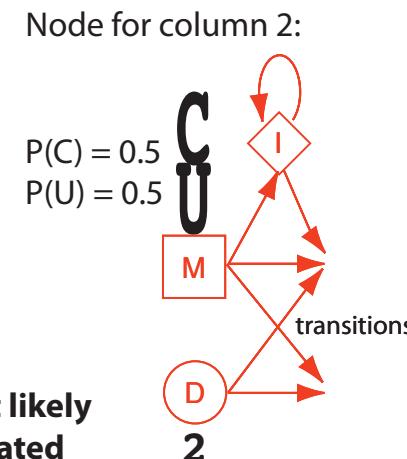
	GCA	UUCG	AA	AC
	G	U	C	G
	C	U	C	C
	A	G	G	A
	1 2 3	4 5 6 7 8 9	10 11	
yeast	GU	CaUUUCGGC...	AC	
fly	GCC.	UU-GGA...	GC	
cow	GCA.	UUCGUC...	-C	
mouse	GCA.	UU-GAU...	GC	
human	GC	GauUUCGCU...	GC	
chicken	GU	A.UUCGUA...	AC	
snake	GU	GauUUCGCG...	AC	
croc	UU.	UU-GAG...	AC	
worm	G-	G.UUCGCGGccaAC		
starfish	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.



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

	GCA	UU	UUCG	AA		AC					
	1	2	3	4	5	6	7	8	9	10	11
yeast	GU	C	A	U	U	C	G	G	U		
fly	GC	C	.	U	U	-	G	G	A	.	AC
cow	GC	A	.	U	U	C	G	U	C	.	-C
mouse	GC	A	.	U	U	-	G	A	U	.	GC
human	GC	G	a	U	U	C	G	C	U	.	GC
chicken	GU	A	.	U	U	C	G	A	U	.	AC
snake	G	U	G	A	U	U	C	G	C	G	.
croc	GU	U	U	-	G	A	G	.	A	C	.
worm	G	-	G	.	U	U	C	G	C	g	ca
starfish	G	-	U	.	U	U	C	G	A	U	.
urchin	GU	U	U	C	-	A	A	AA	C	AA	.

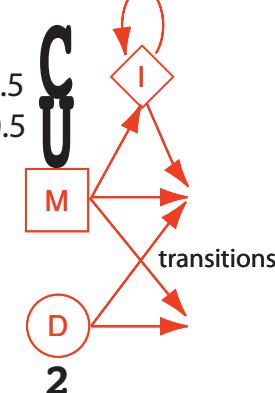
One HMM node per alignment column

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

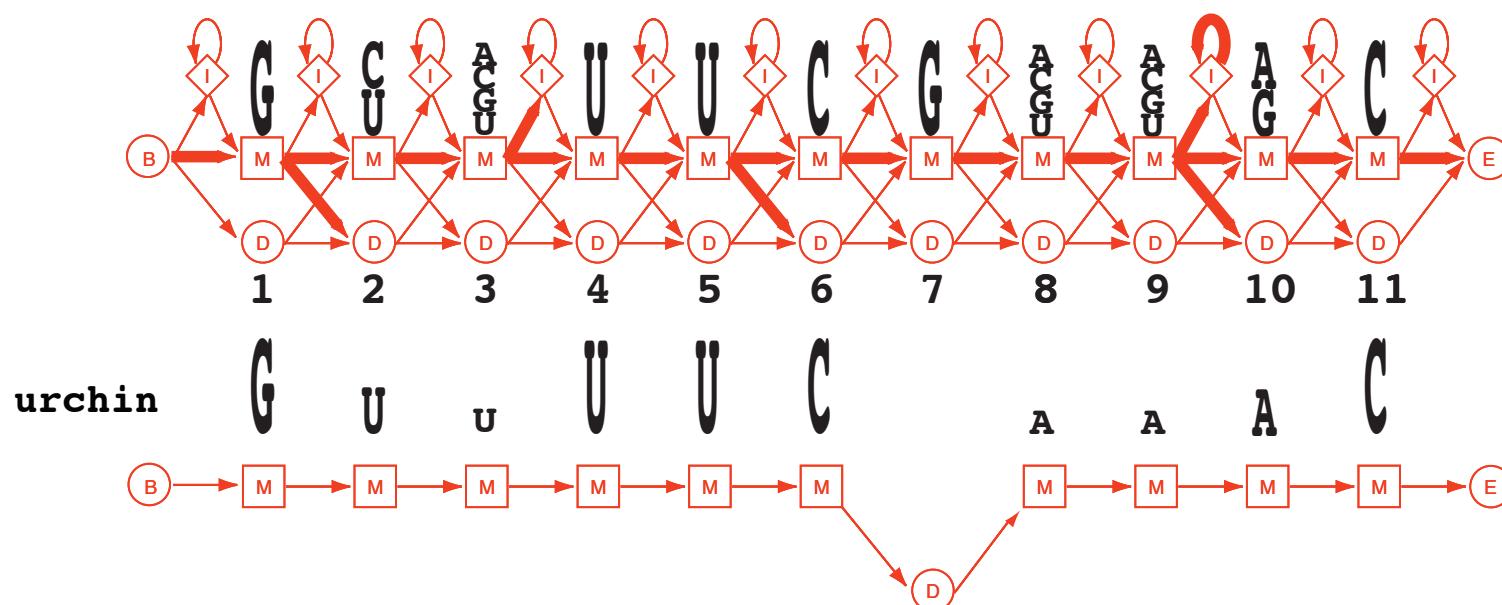
HMMs generate homologous sequences.

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

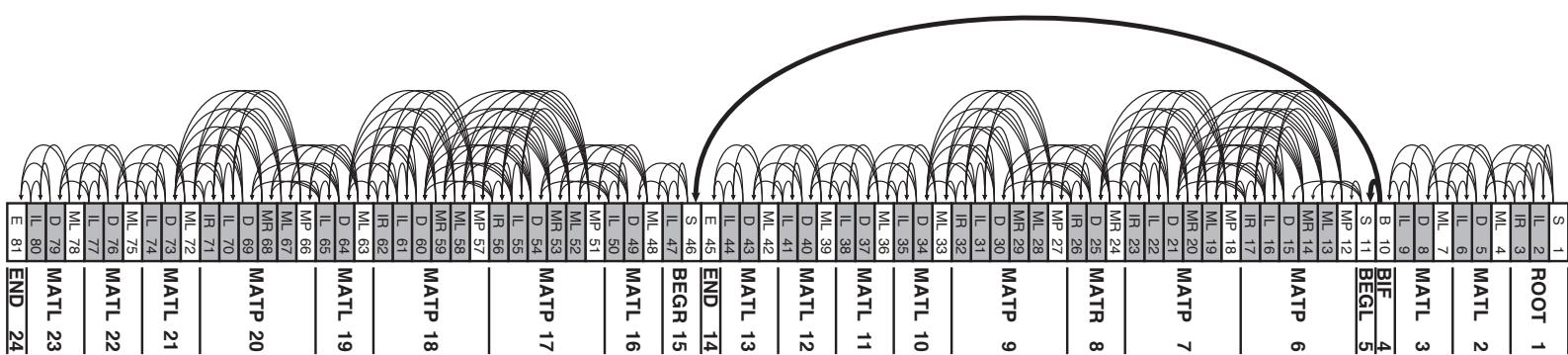
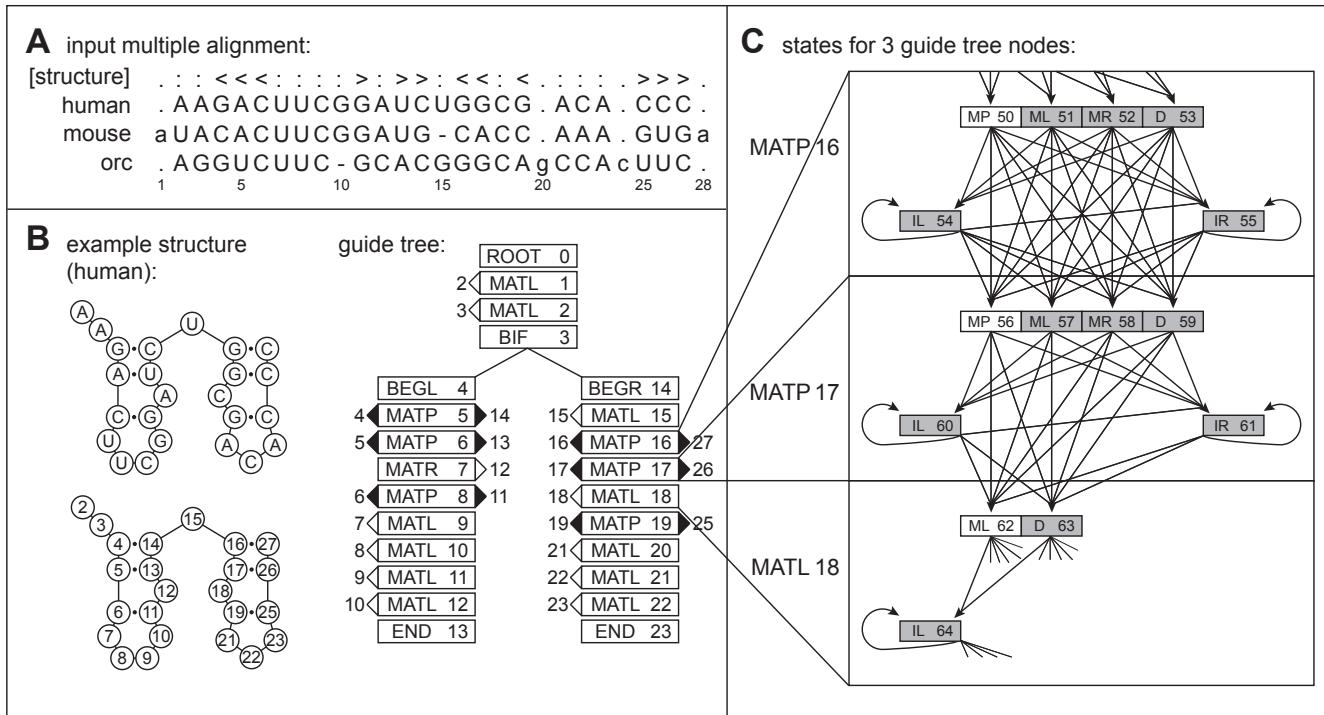
Node for column 2:



$$P(C) = 0.5 \\ P(U) = 0.5$$



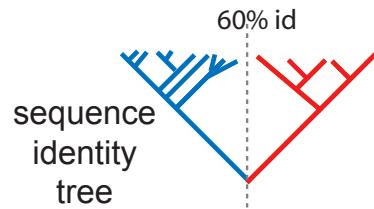
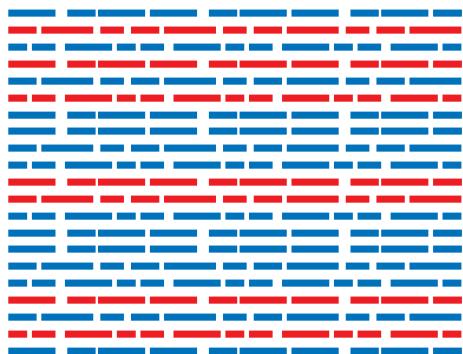
Covariance models (CMs) are built from structure-annotated alignments



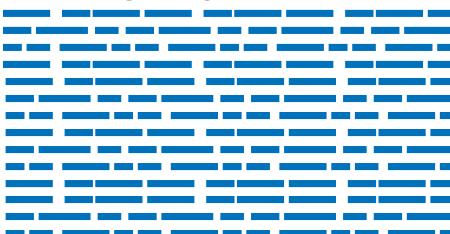
Is the added complexity worth it?

RMARK: a challenging internal RNA homology search benchmark

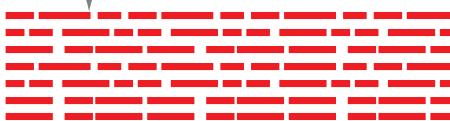
Rfam seed alignment:



training alignment

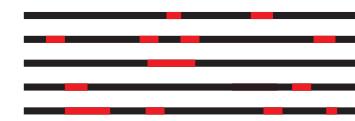


no train/test sequence pair is > 60% identical



test sequences

embed in
pseudo-genome

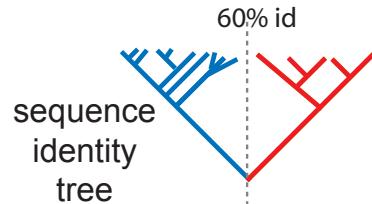
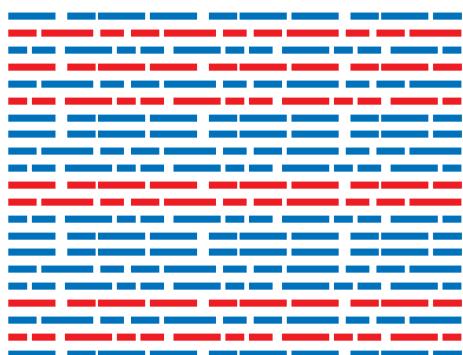


10 1Mb sequences
with 780 embedded
test seqs from 106 families

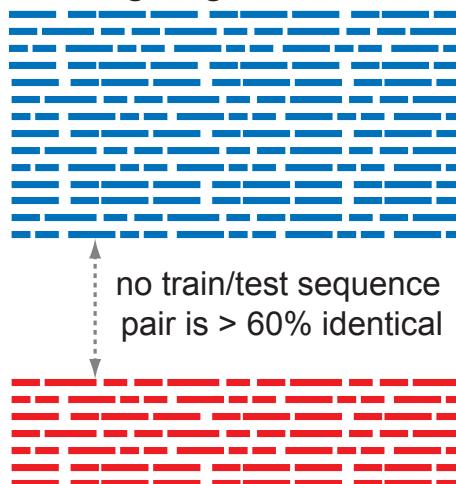
Is the added complexity worth it?

RMARK: a challenging internal RNA homology search benchmark

Rfam seed alignment:



training alignment



test sequences

profile
(CM or HMM)

BLAST

search

embed in
pseudo-genome



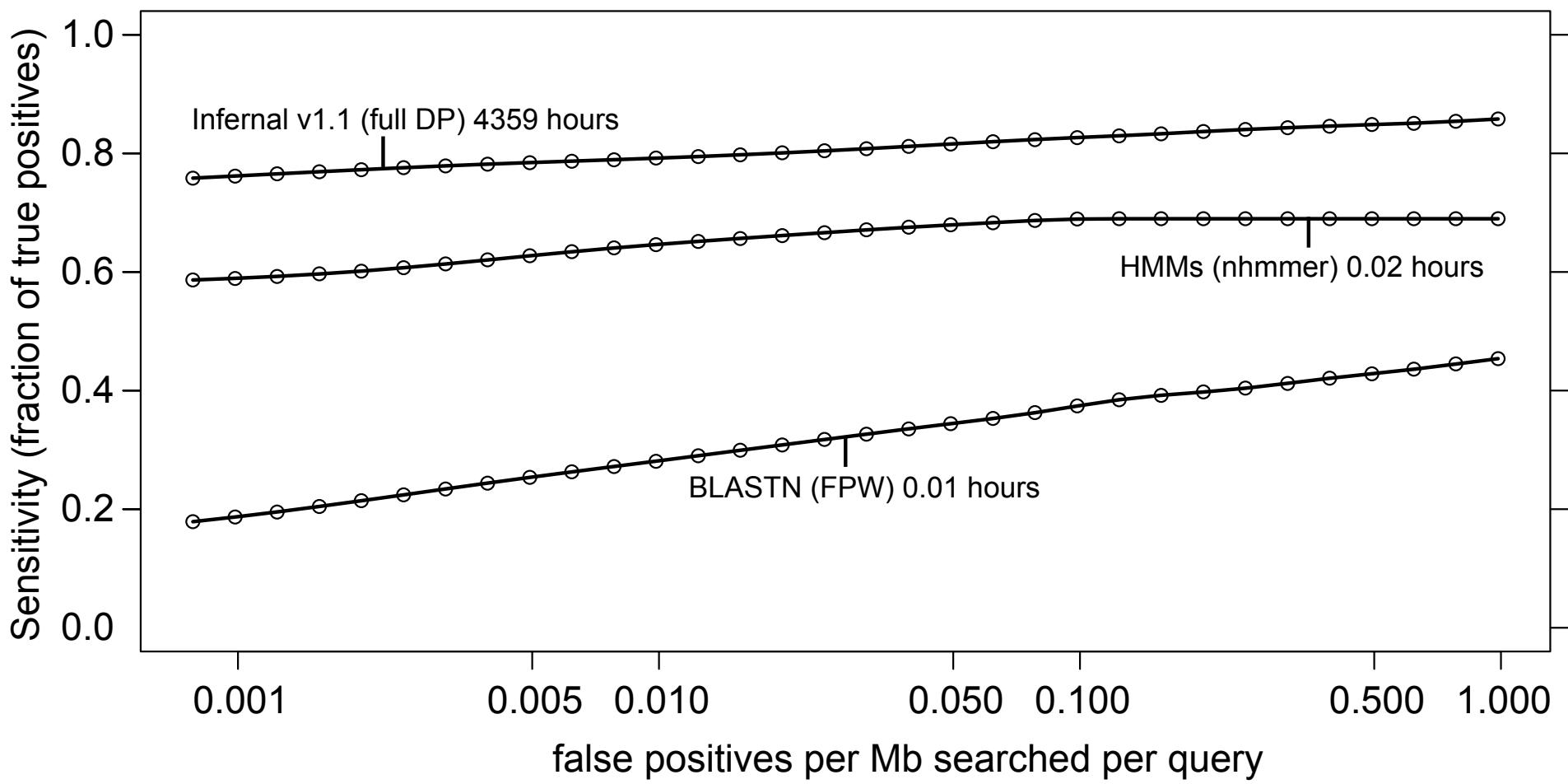
10 1Mb sequences
with 780 embedded
test seqs from 106 families

E=1E-40 132.53 bits rmark7 OLE 340023 339402 +
...

E=0.0013 32.3 bits rmark3 6S 10135 10261 +
E=0.0026 27.6 bits rmark6 tRNA 789278 789466 +
E= 0.0061 28.3 bits rmark2 Cobalamin 32032 31787 -
E=0.0231 25.4 bits rmark 6 FALSE 673200 673340 +
E=0.0670 25.3 bits rmark6 tRNA 789278 789116 -
...

E=103.3 16.4 bits rmark 4 FALSE 783222 782803 -

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



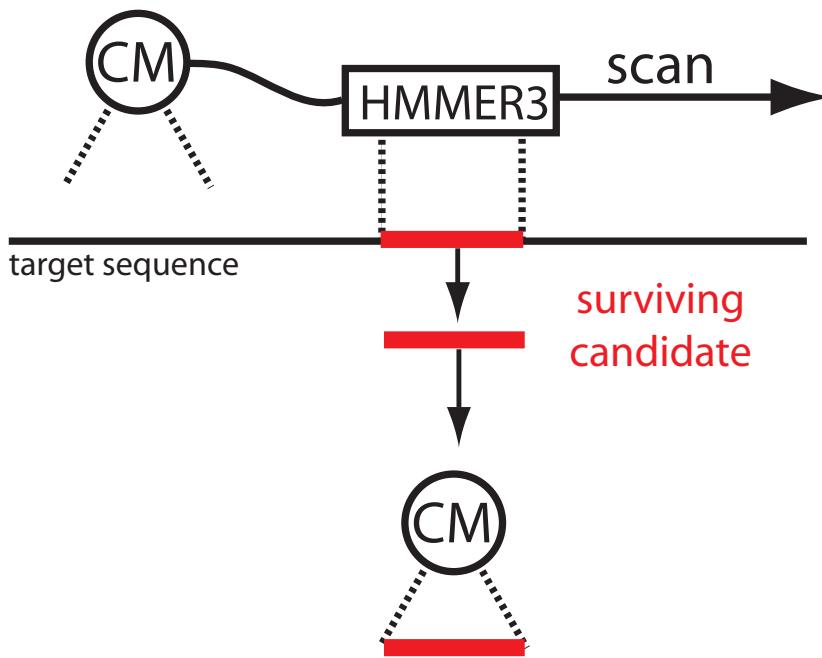
Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

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

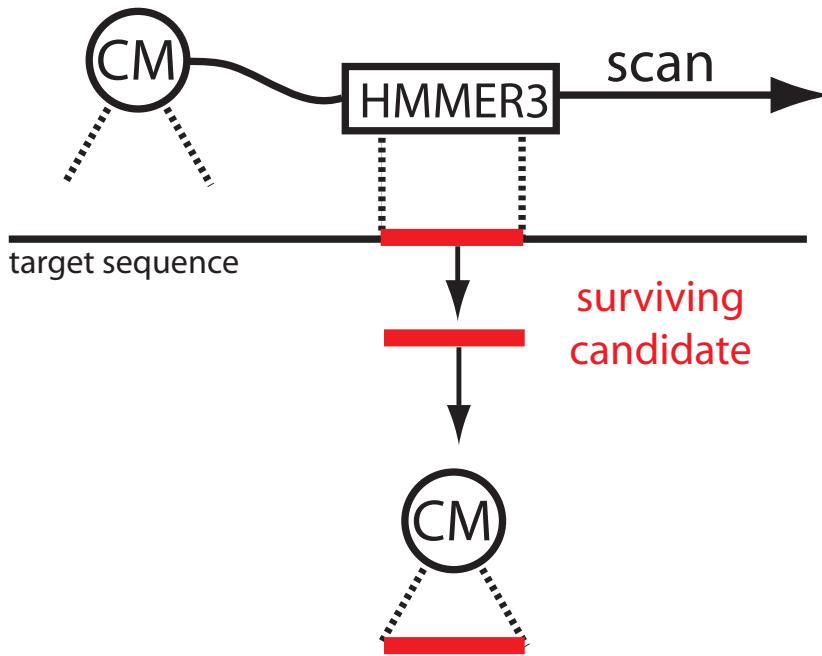
Filter target database using profile HMMs*

HMM filter first pass



Filter target database using profile HMMs*

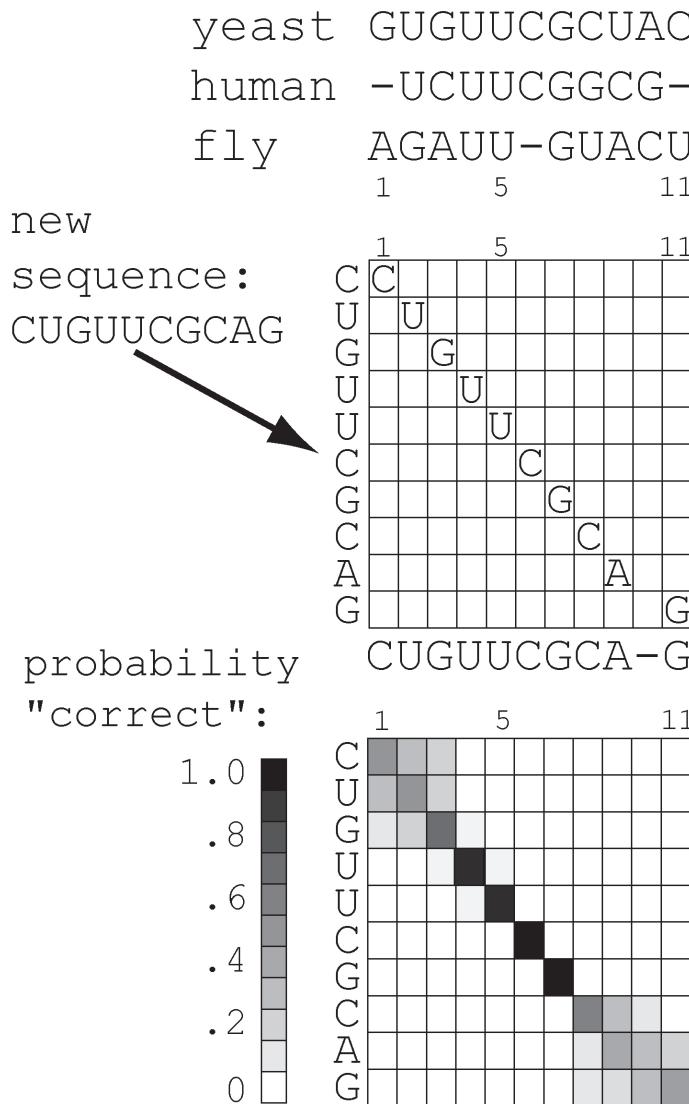
HMM filter first pass



- 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.

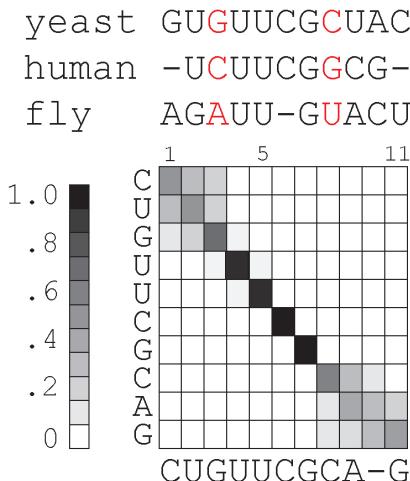
Accelerating CM alignment step 1: HMM posterior decoding to get confidence estimates



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

HMMs -

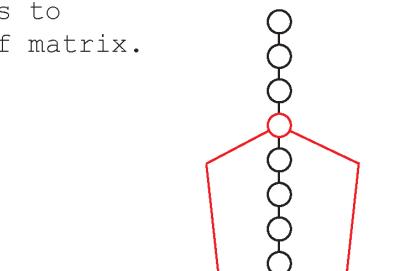
Each column of seed alignment corresponds to a column of matrix.



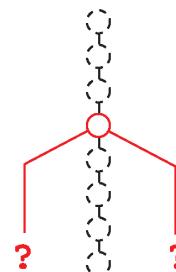
CMs -

Each column of seed alignment corresponds to a state.

yeast	human	fly
U C	U C	U
U G	U G	U G
G•C	C•G	A•U
U•A U	U•G C	G•C A
G•C		A•U



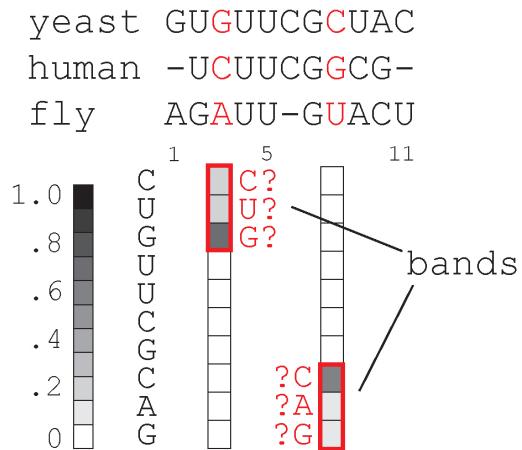
struct <<----->->
 yeast GUGUUCG**C**UAC
 human -UCUUCGG**G**CG-
 fly AG**A**UU-G**U**ACU



CUGUUCGCAG
 45 possibilities

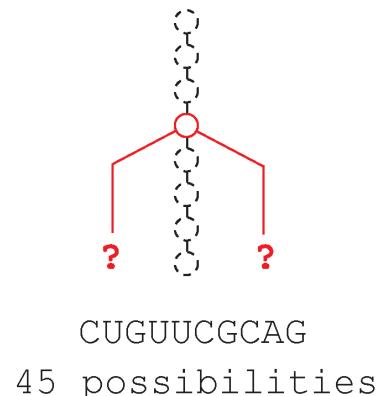
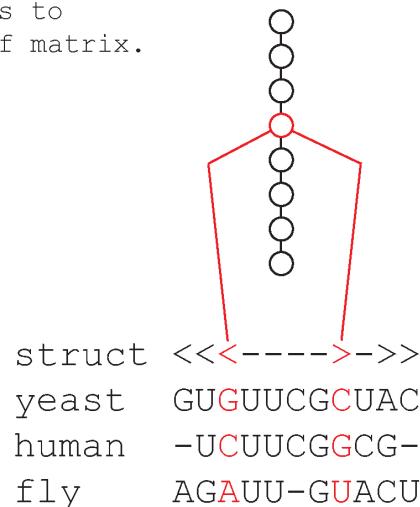
Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*

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

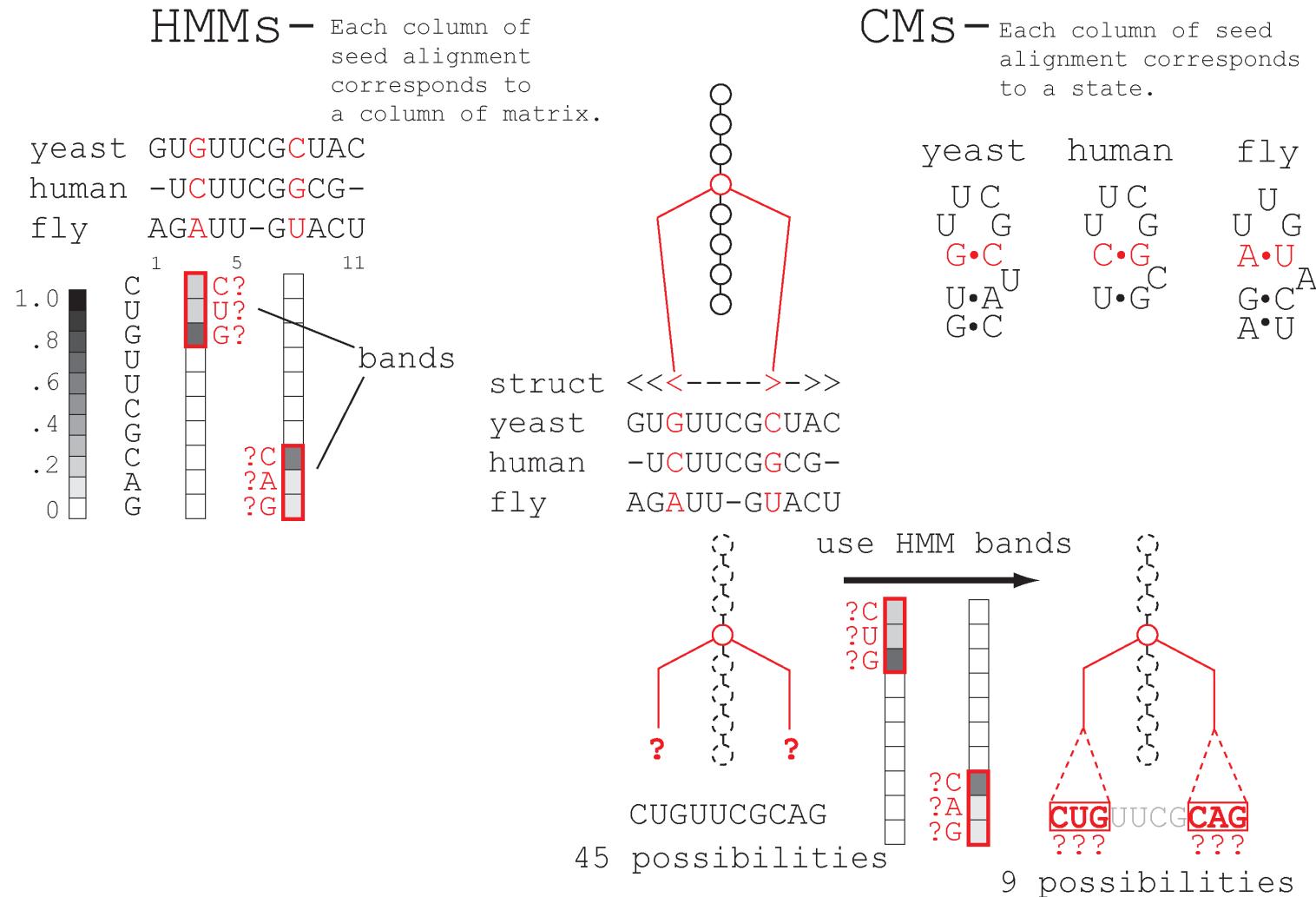


CMs - Each column of seed alignment corresponds to a state.

yeast	human	fly
U C	U C	U
U G	U G	U G
G•C	C•G	A•U
U•A U	U•G C	G•C A
G•C		A•U

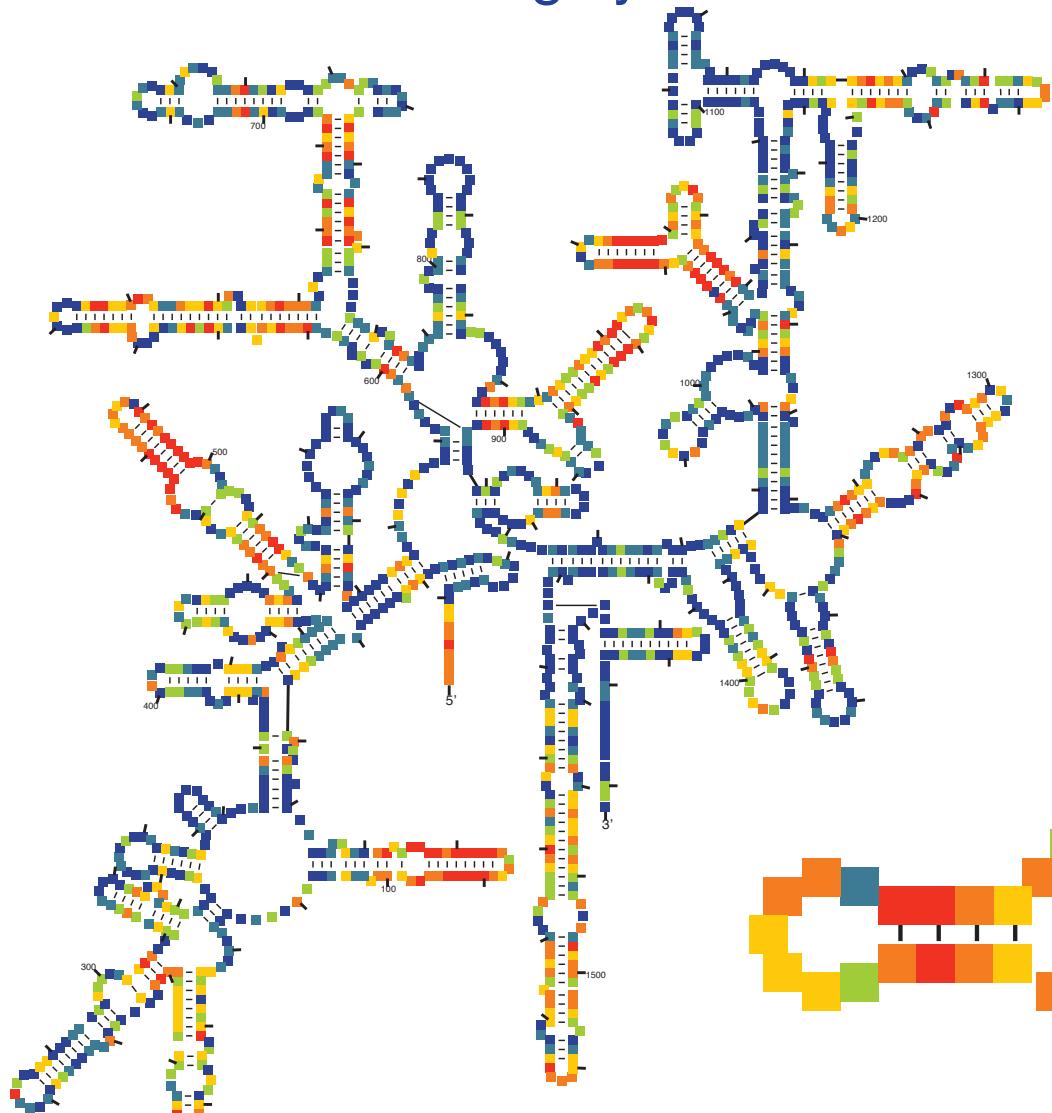


Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*

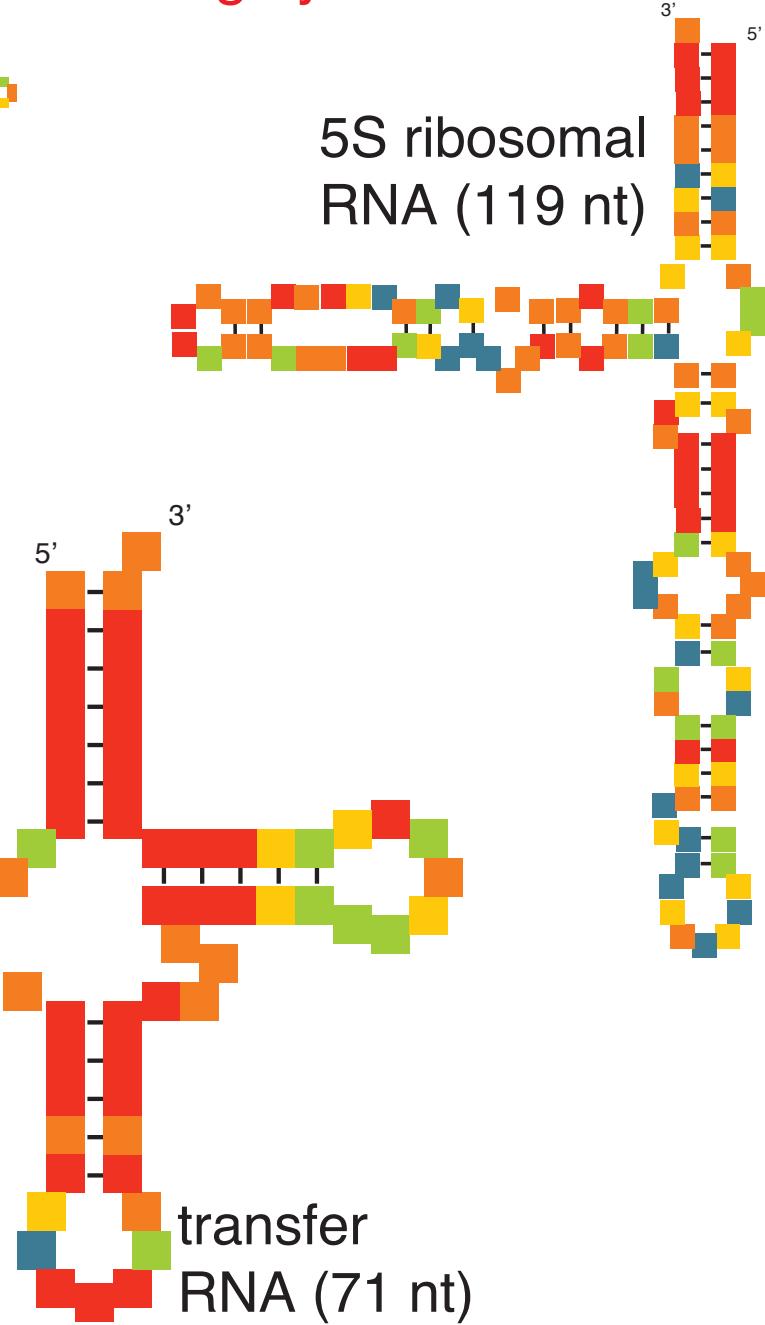


Sequence conservation per position

blue:highly conserved red: highly variable



small subunit
ribosomal RNA
(SSU rRNA, 1582 nt)

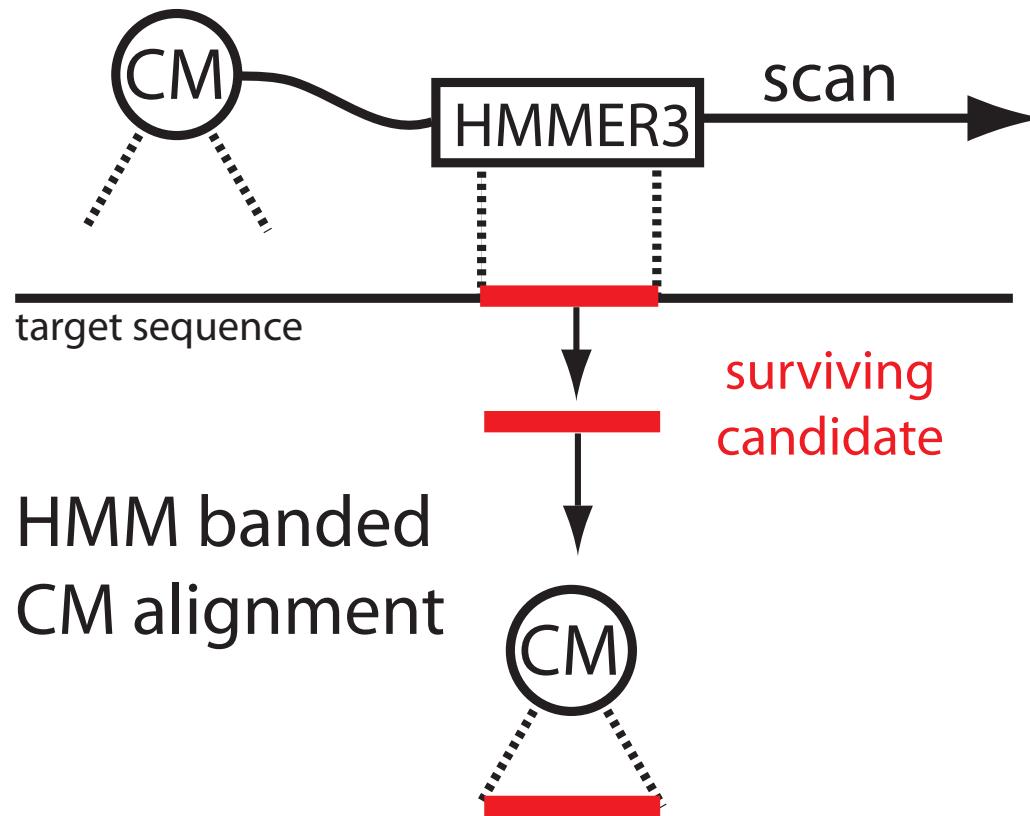


5S ribosomal
RNA (119 nt)

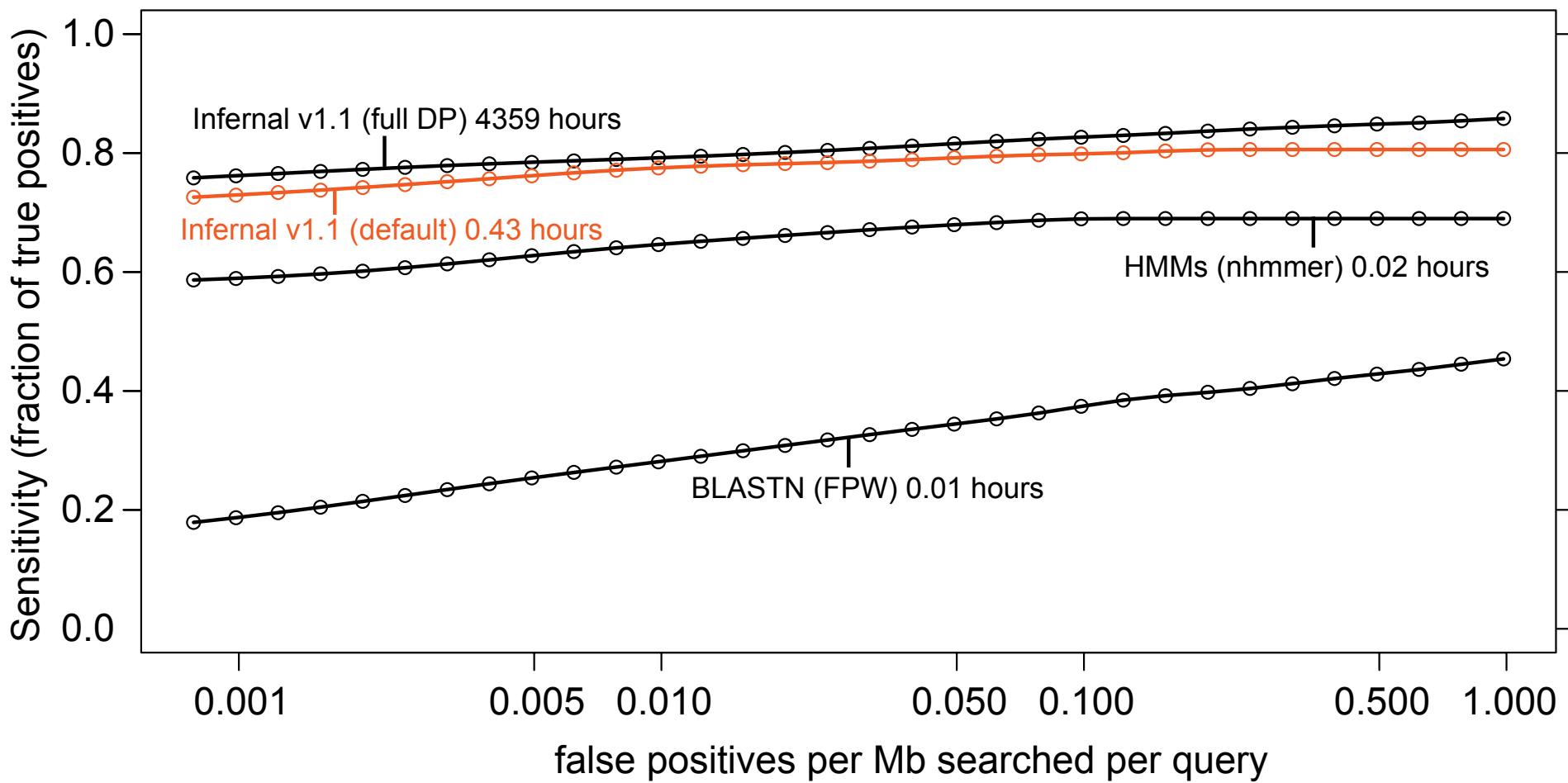
transfer
RNA (71 nt)

Use HMMs as filters and to constrain CM alignment

HMM filter first pass



HMM-based acceleration makes Infernal 10,000 times faster



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

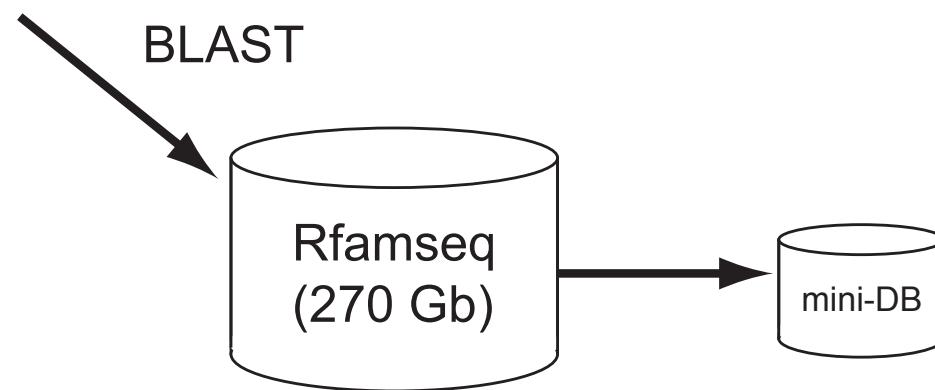
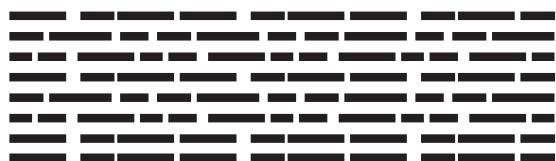
Outline of talk

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

Rfam originally used BLAST filters

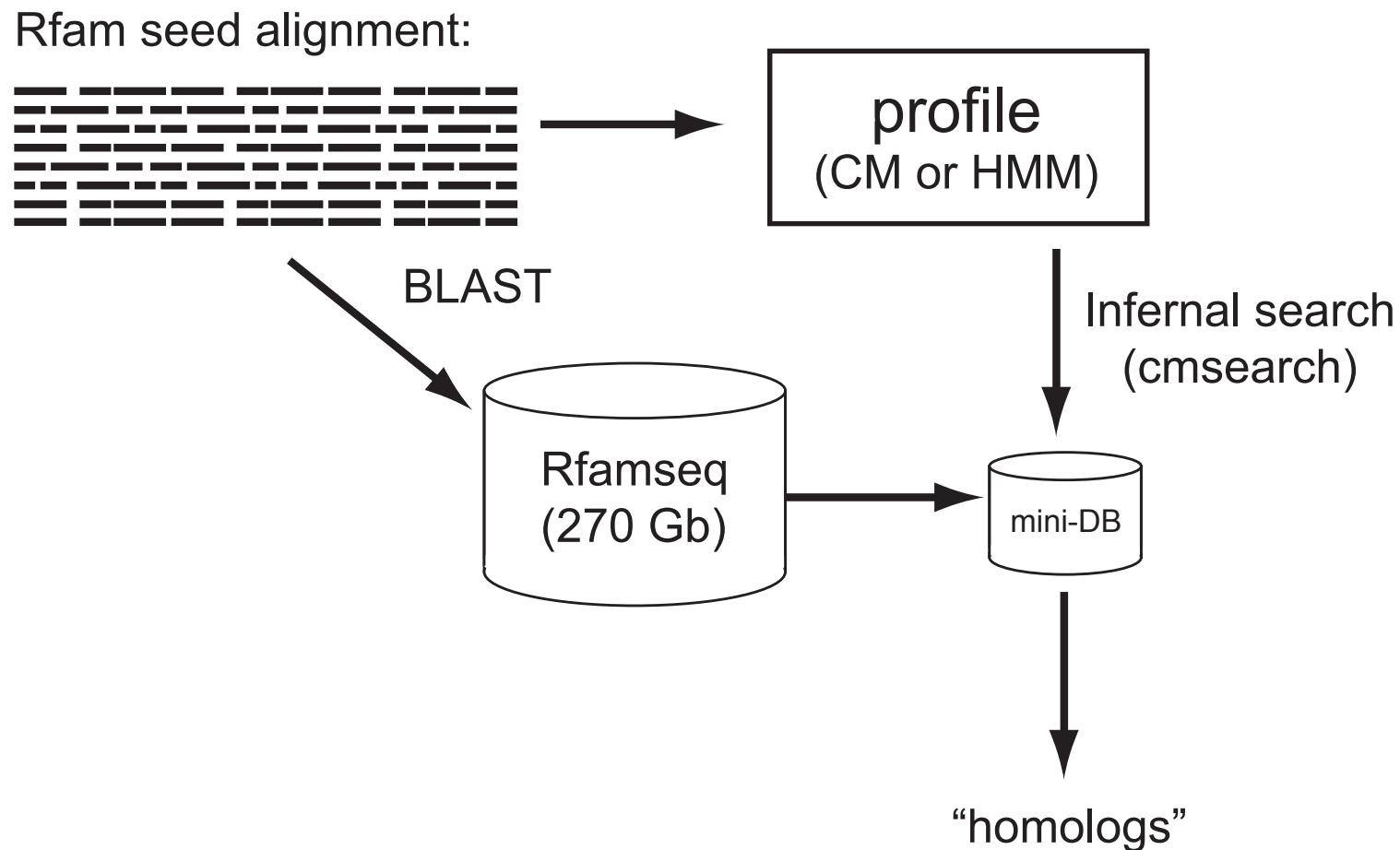
- Rfam includes > 3000 RNA families, each represented by an alignment, CM and set of predicted homologs in a large database (Rfamseq).

Rfam seed alignment:



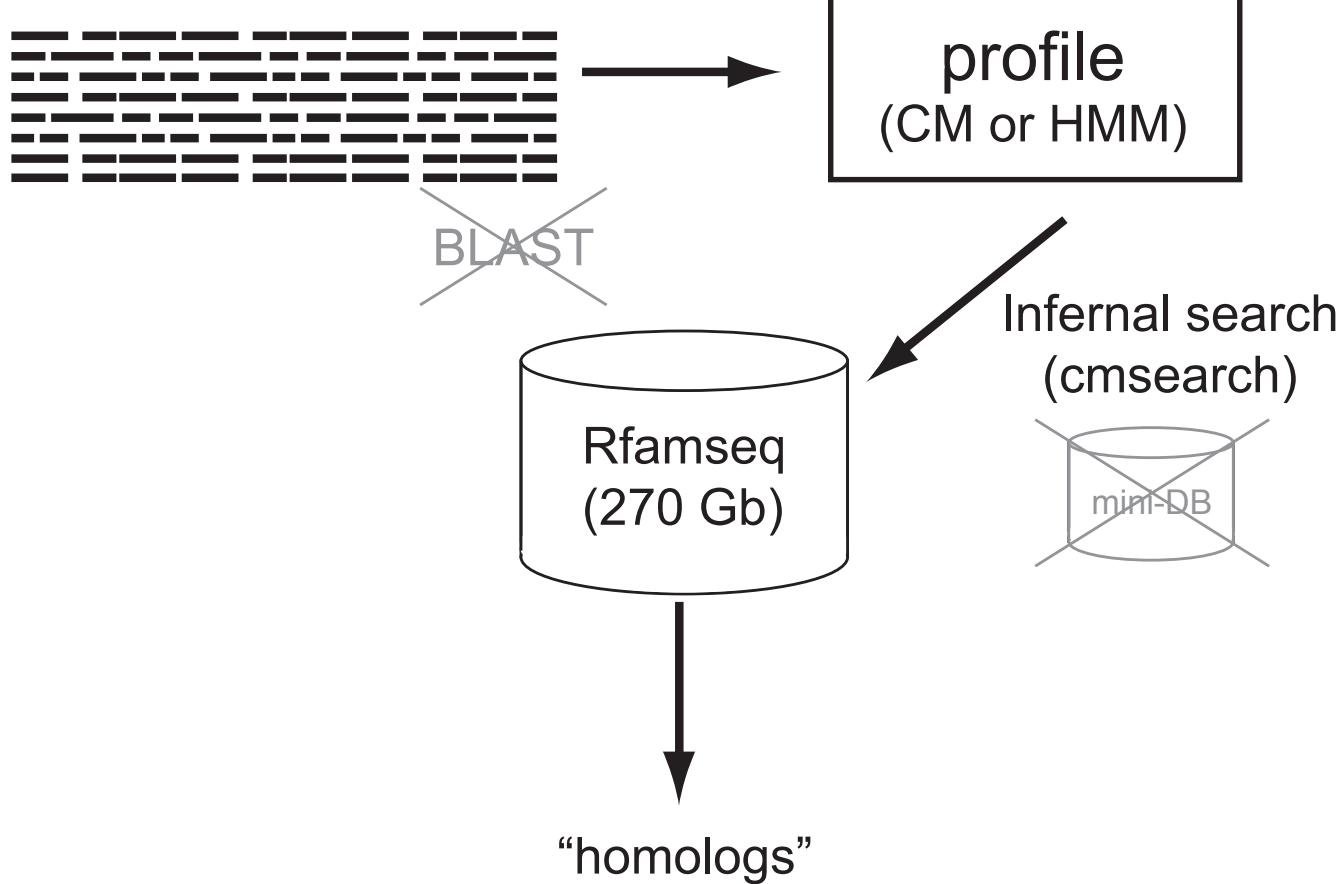
Rfam originally used BLAST filters

- Rfam includes > 3000 RNA families, each represented by an alignment, CM and set of predicted homologs in a large database (Rfamseq).



Rfam 12.0*, first release without BLAST filtering

Rfam seed alignment:



Rfam 12.0*, 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

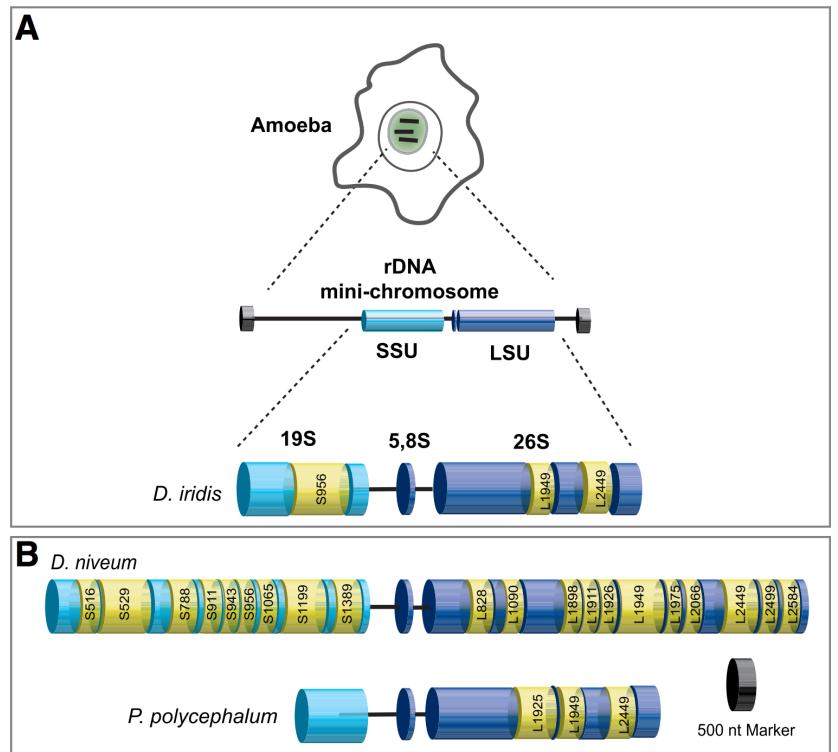
Infernal 1.1 finds 11,000 new group I intron candidates

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

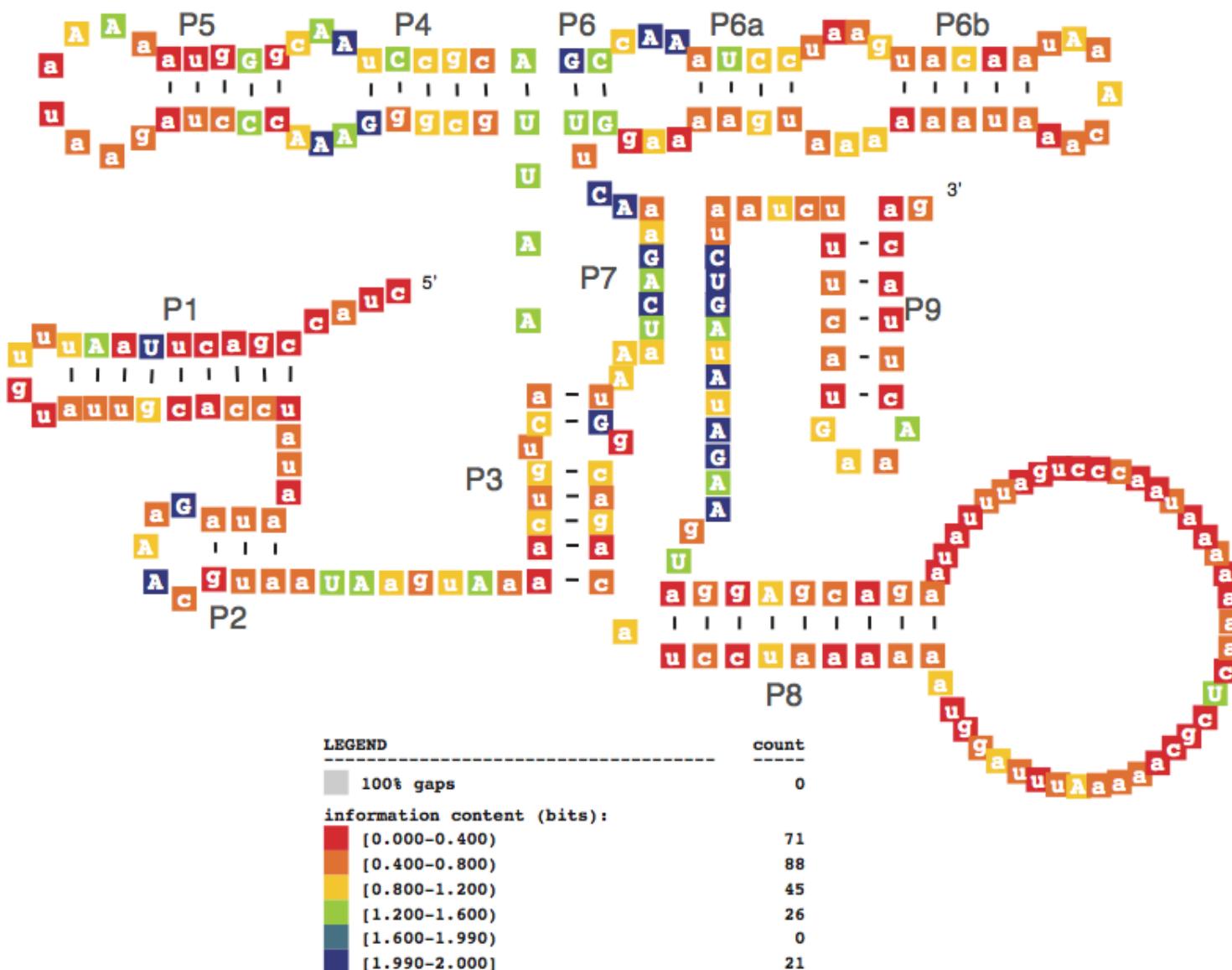
Accession	Family ID	Length (nt)	#of seed seqs	Time new (h)	Time old (h)	Time (old/new)	New total hits	Old total hits	New unique hits	Old unique hits
Top five families										
RF00028	Intron-gpI	251	12	125.0	357.2	2.8	71 433	60 264	11 175	1
RF00026	U6	104	188	31.2	181.1	5.8	66 517	62 174	4367	14
RF00003	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

Group I catalytic introns

- self splicing ribozymes found in lower eukaryotes, higher plants, bacteria and bacteriophages
- often have ORFs (homing endonucleases) inserted in loop regions
- genes they are found in:
 - bacteria and mitochondria and chloroplast of lower euks: rRNA, mRNA, and tRNAs
 - higher plants mitochondria and chloroplast: a few tRNA and mRNA genes
 - nuclear lower eukaryotic genomes: only rRNA

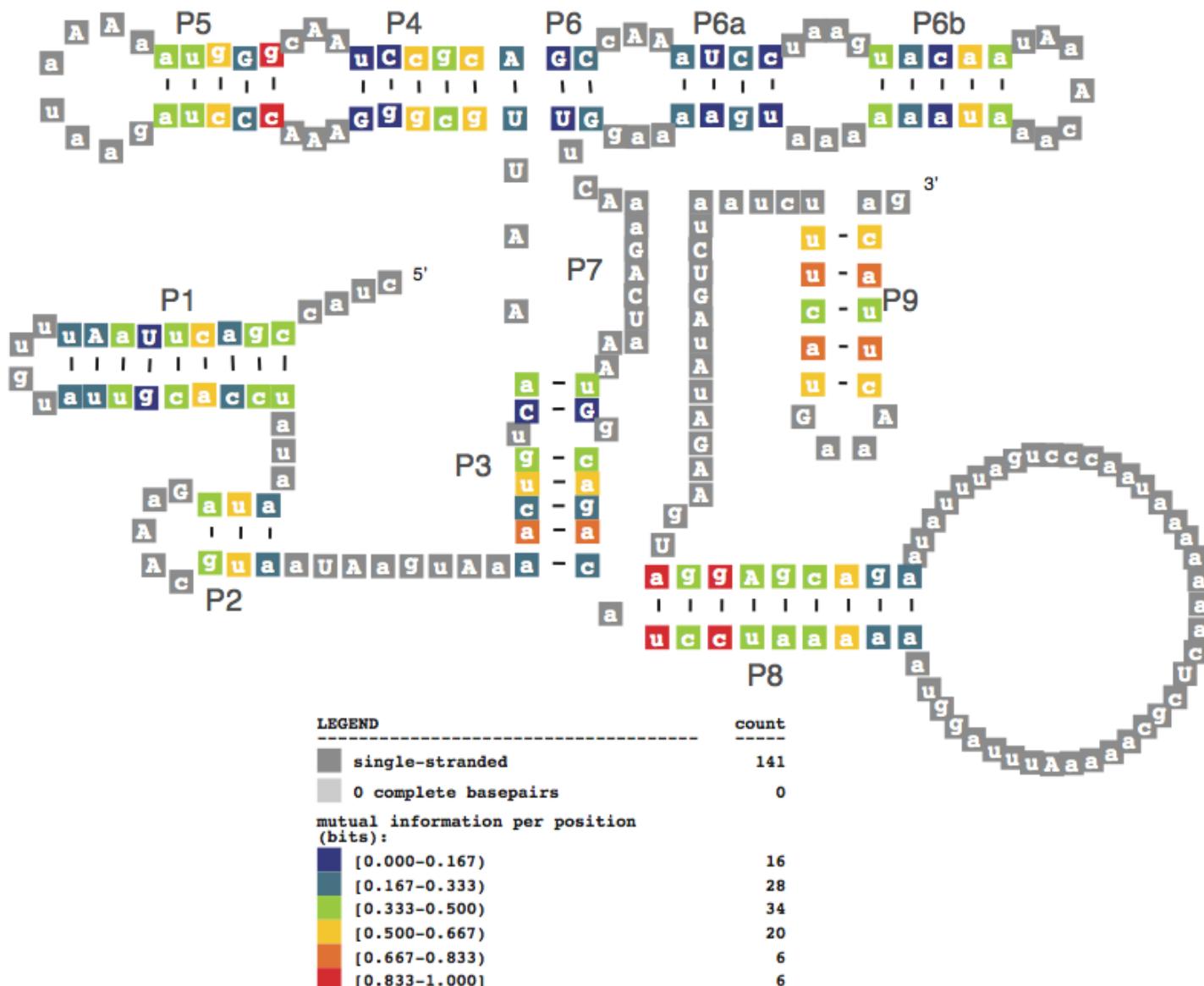


model	#pos	#bps	#seqs	description
Group I Intron	251	55	12	information content per position



Consensus nucleotides (nt) are displayed, defined as the most frequent nt at each position.
Capitalized nts occur in ≥ 0.75 fraction of sequences that do not have a gap at the position.

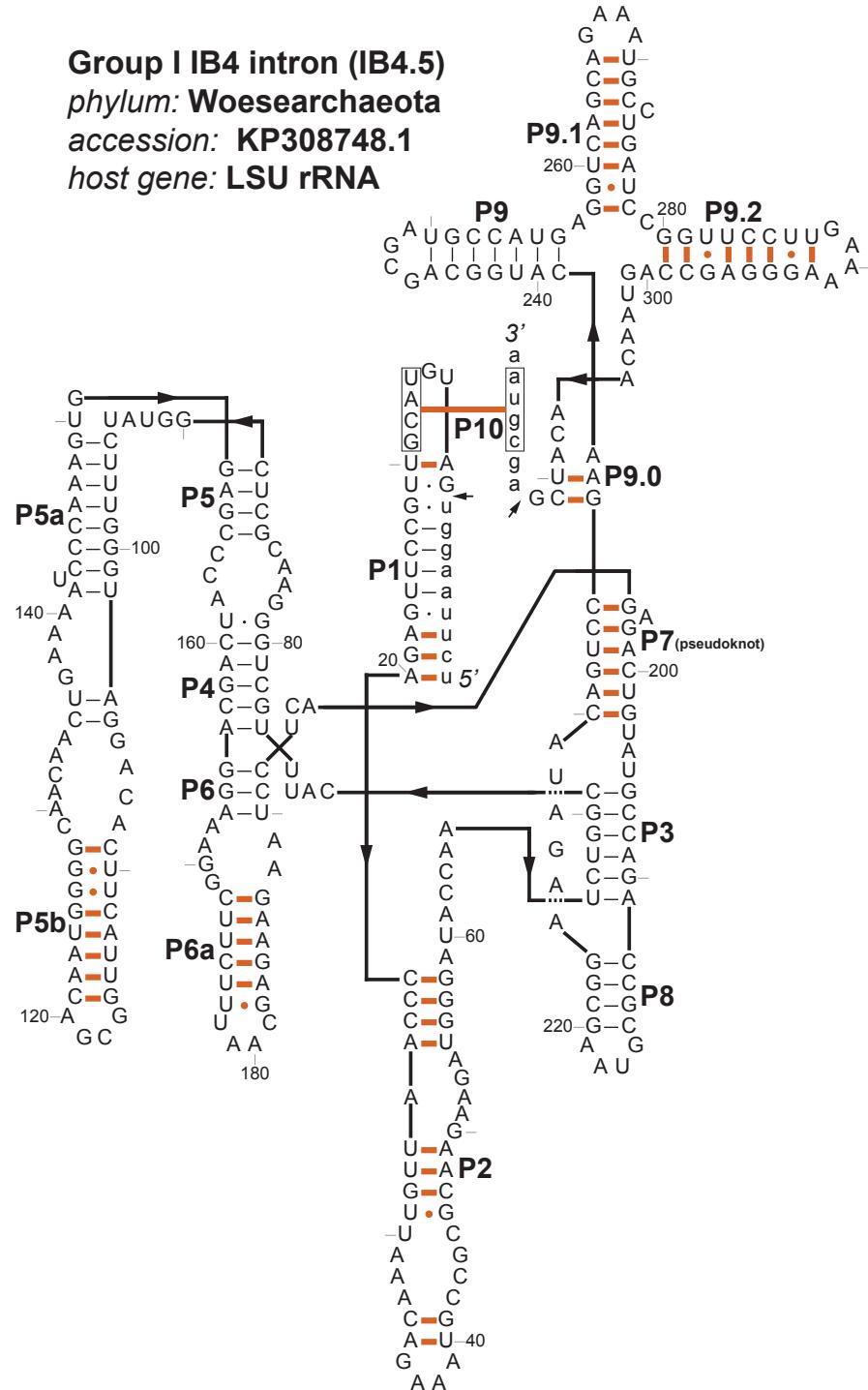
model	#pos	#bps	#seqs	description		
Group	I	Intron	251	55	12	mutual information per basepaired position



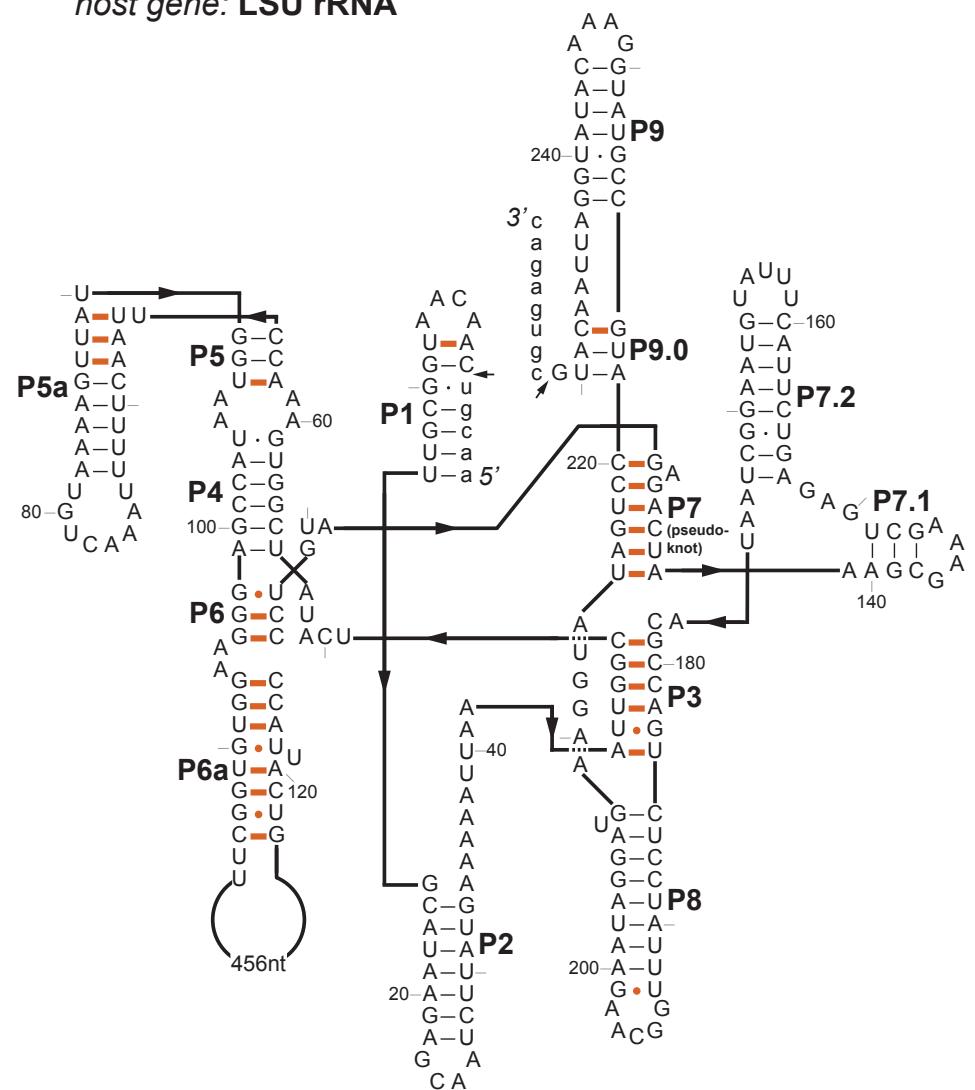
Consensus nucleotides (nt) are displayed, defined as the most frequent nt at each position.
Capitalized nts occur in ≥ 0.75 fraction of sequences that do not have a gap at the position.

Group I Introns?		
	previously known	Infernal v1.1 predictions
EUKARYOTA	insects	- +
	flatworms	- +
	vertebrates	- +
	jellyfish	+ +
	Choanoflagellata	- +
	fungi	+ +
	plants	+ +
	ciliates	+ +
ARCHAEA	Euryarchaeota	- -
	Crenarchaeota	- +
	Thaumarchaeaota	- +
BACTERIA	Proteobacteria	+ +
	Cyanobacteria	+ +
	Aquifex	- +
	Bacteriodetes	- +
	Firmicutes	+ +
	Actinobacteria	- +

Group I IB4 intron (IB4.5)
phylum: Woesearchaeota
accession: KP308748.1
host gene: LSU rRNA



Group I IA3 intron (IA3.1)
phylum: Woesearchaeota
accession: CP010426.1
host gene: LSU rRNA



*

Could archaeal group I introns have evolved into BHB introns?

Evolution of introns in the archaeal world

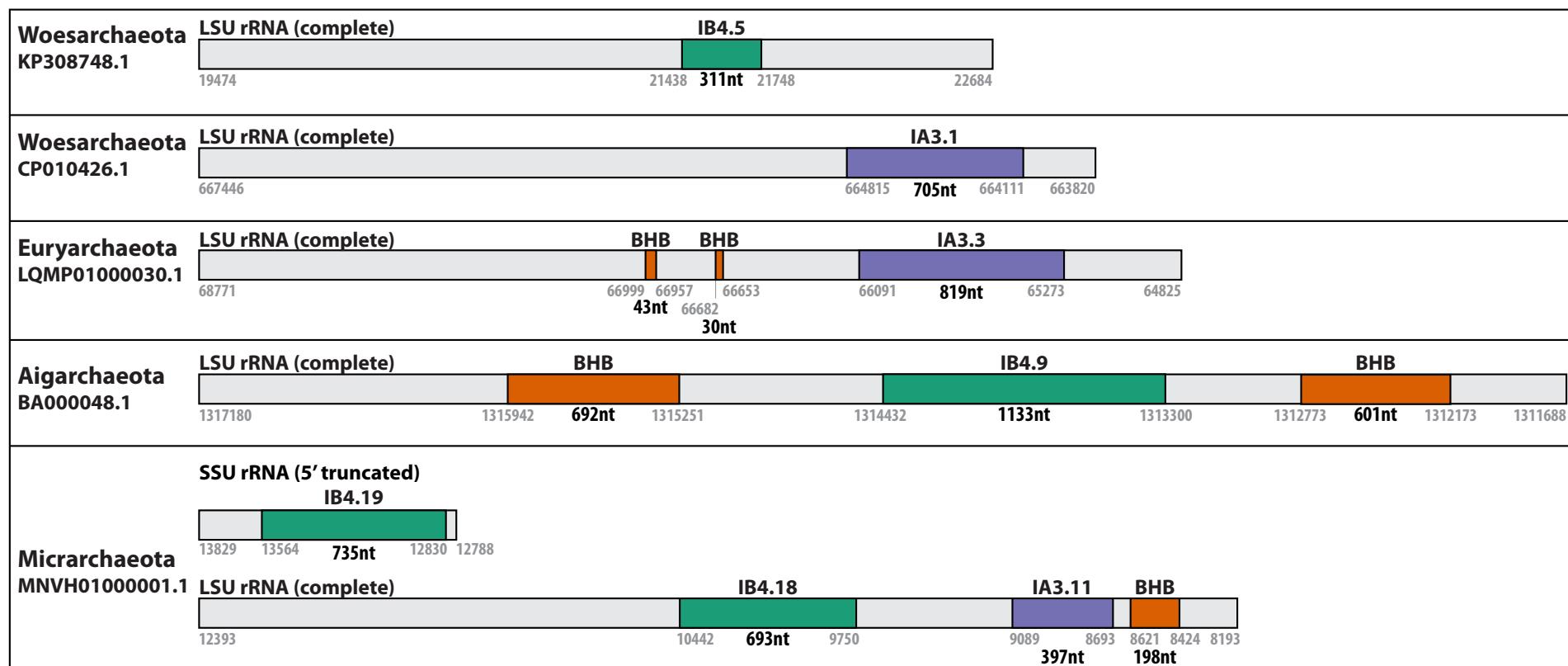
Giuseppe D. Tocchini-Valentini, Paolo Fruscoloni, and Glaucio P. Tocchini-Valentini¹

Istituto di Biologia Cellulare, Consiglio Nazionale delle Ricerche, Campus A, Buzzati-Traverso, Via Ramarini 32, Monterotondo Scalo, 00016 Rome, Italy

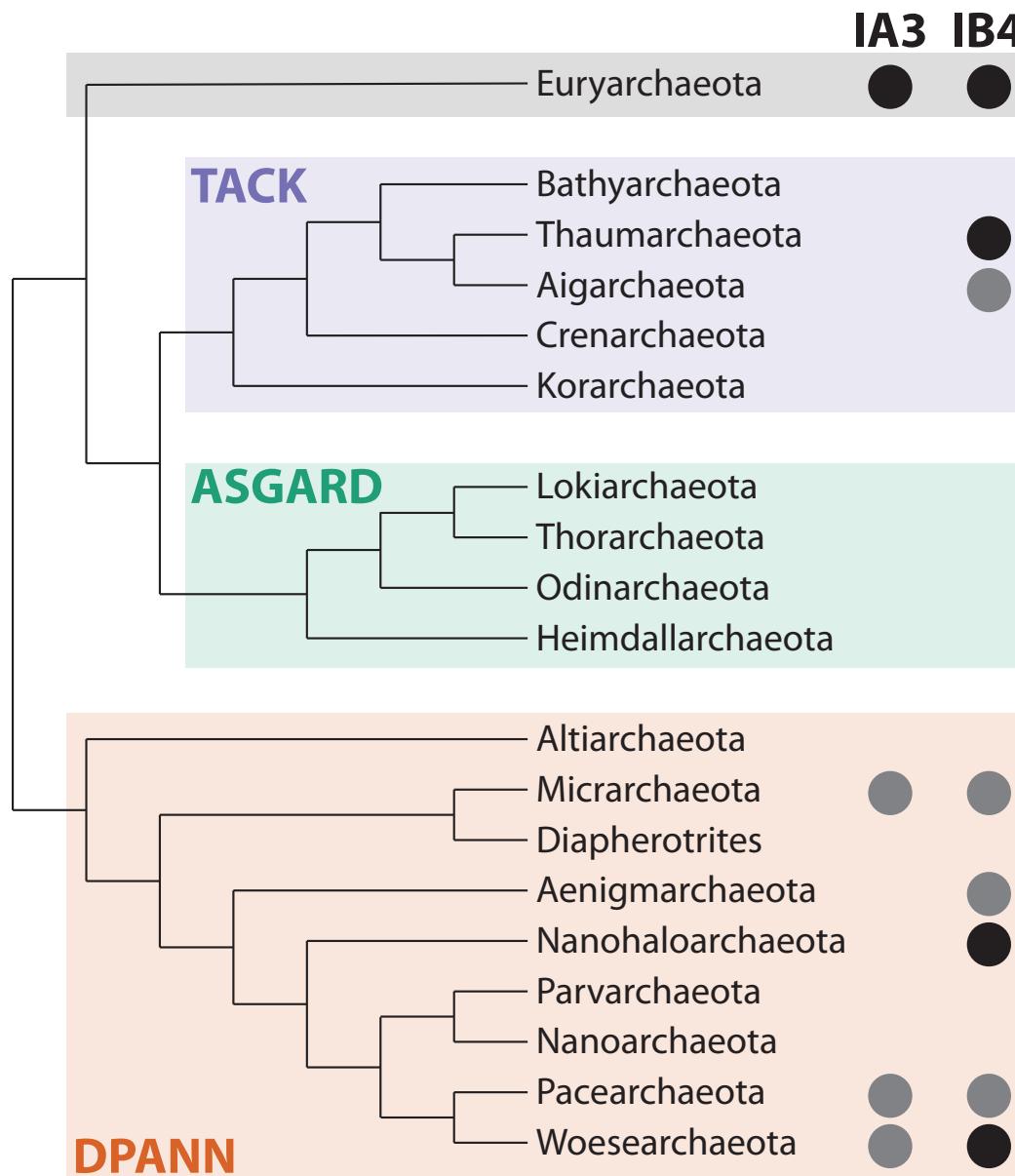
Contributed by Glaucio P. Tocchini-Valentini, January 24, 2011 (sent for review December 1, 2010)

*

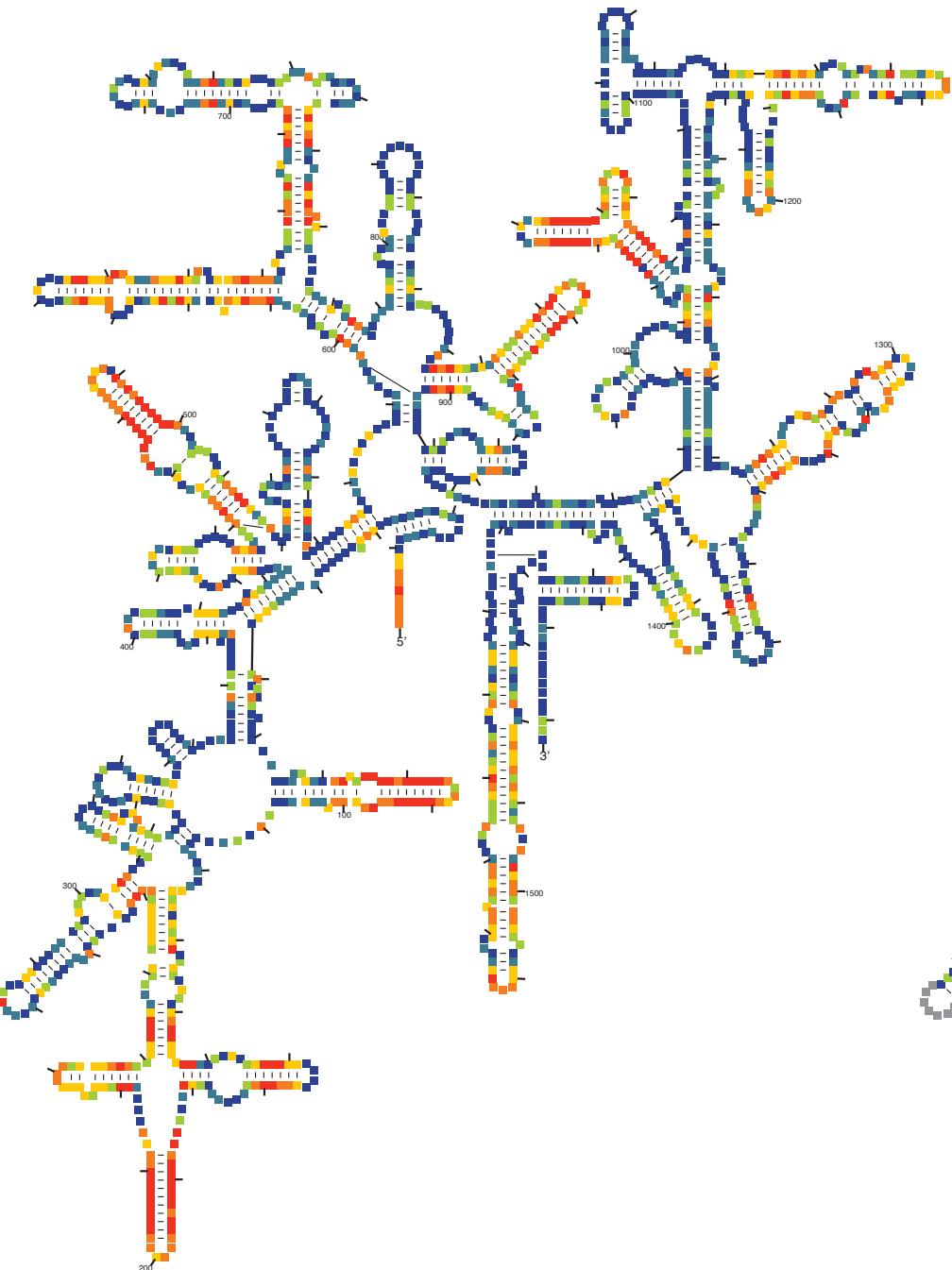
Archaeal group I introns can occur in same host gene as BHB introns



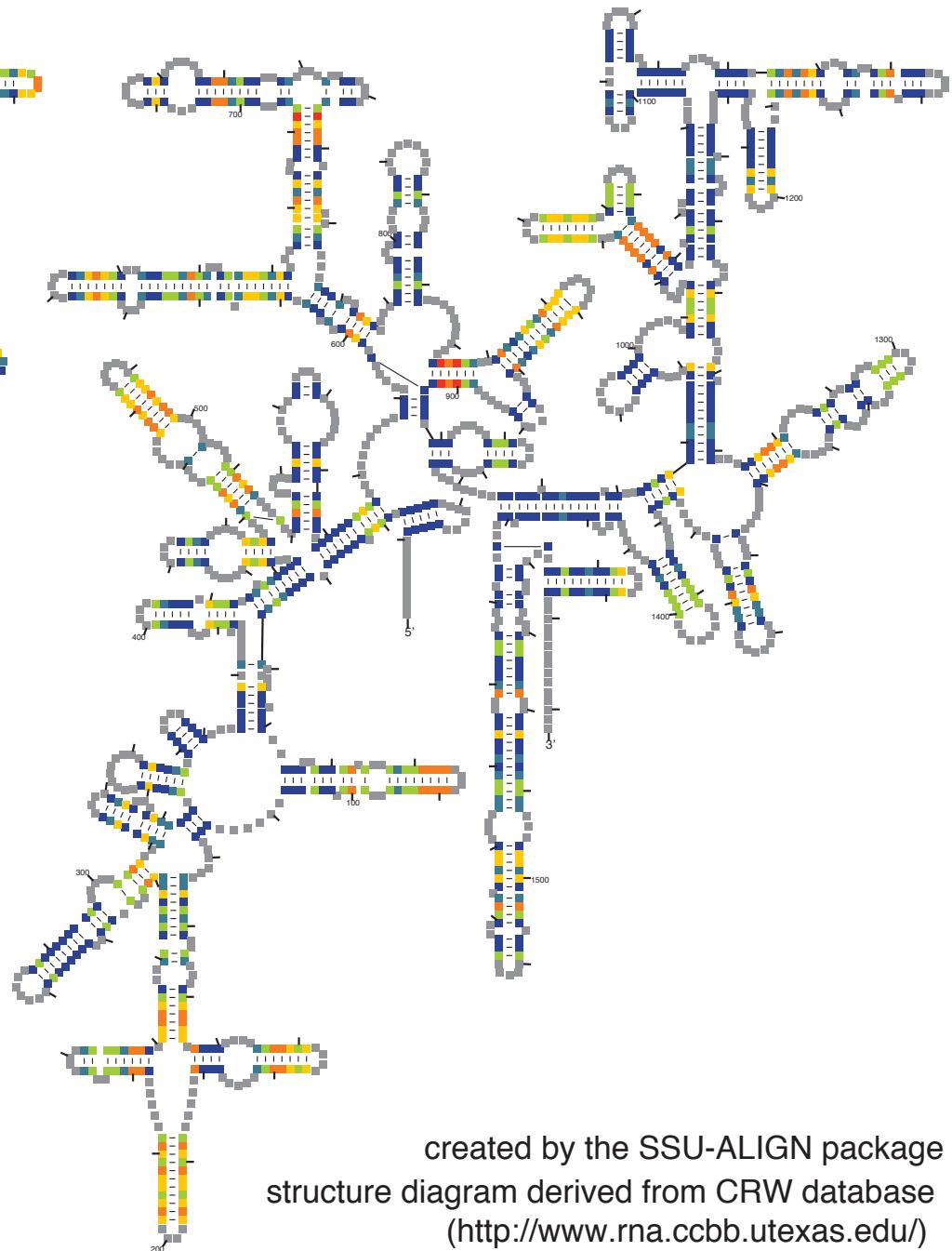
Group I introns are widespread in Archaea



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/>)

Acknowledgements

Harvard/Janelia

Sean Eddy
Tom Jones
Diana Kolbe
Travis Wheeler
Elena Rivas
Michael Farrar

NCBI

Alejandro Schäffer
David Landsman
Jim Ostell
David Lipman

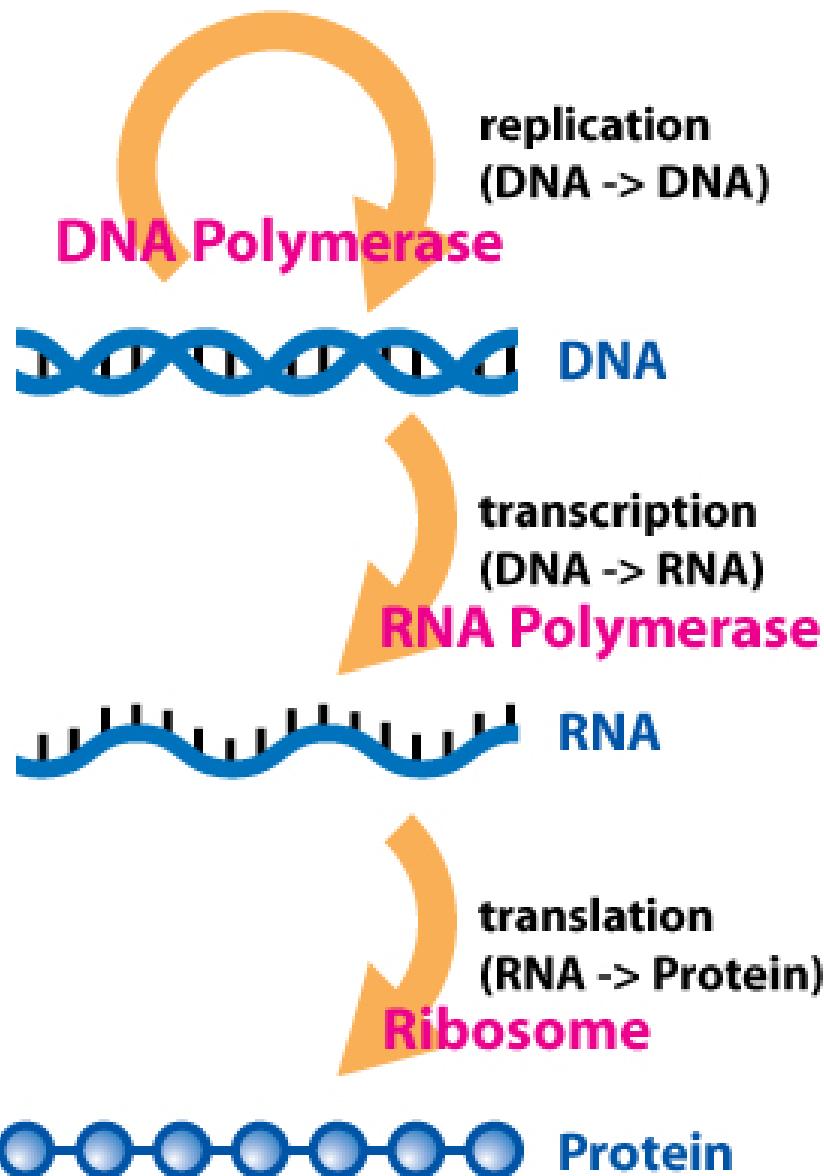
GISSD

Yu Zhou
Chen Lu
Qi-Jia Wu
Yu Wang
Zhi-Tao Sun
Jia-Cong Deng
Yi Zhang

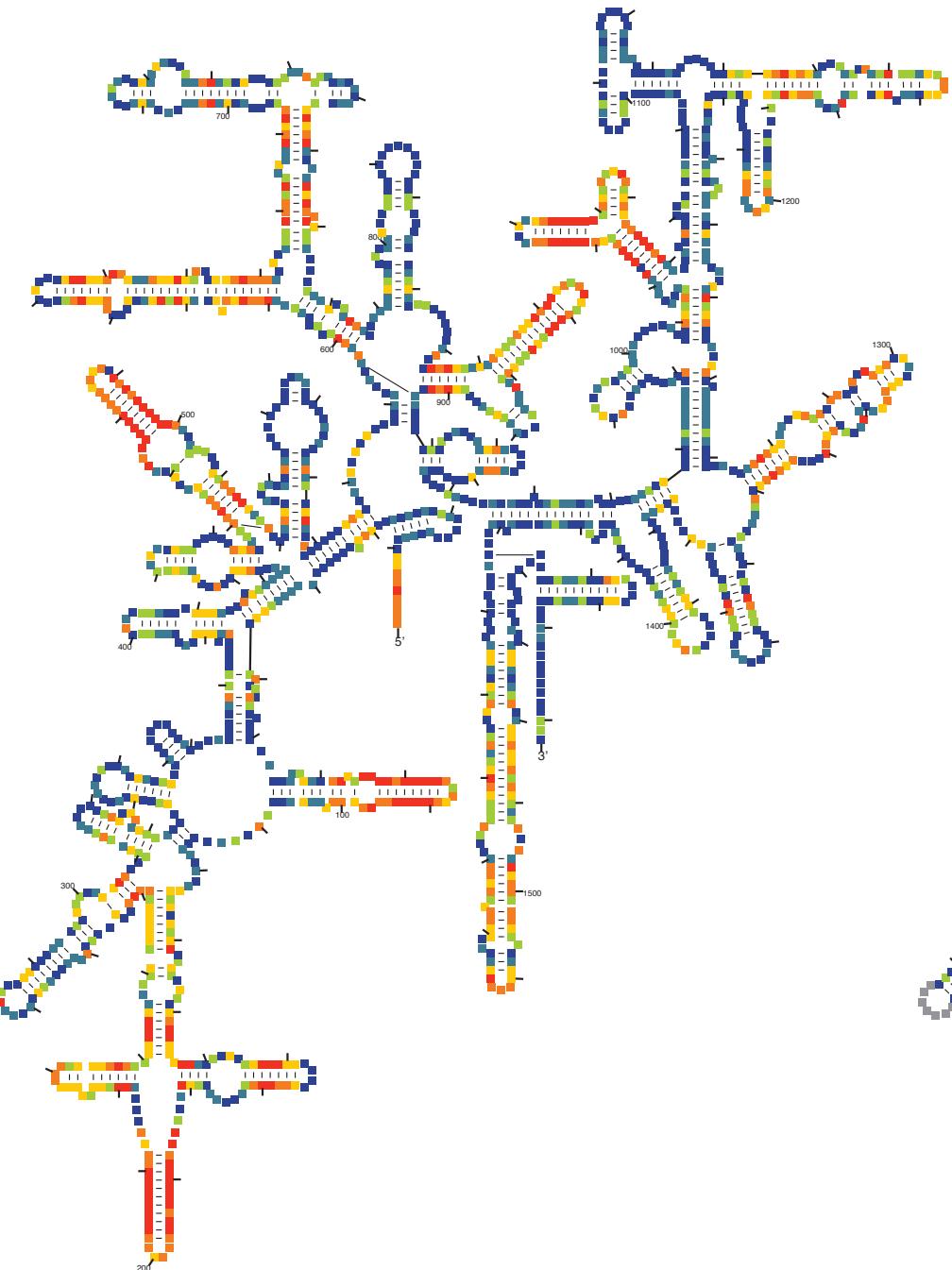
David Haussler's group

Yasu Sakakibara
(CM-like models in 1994)

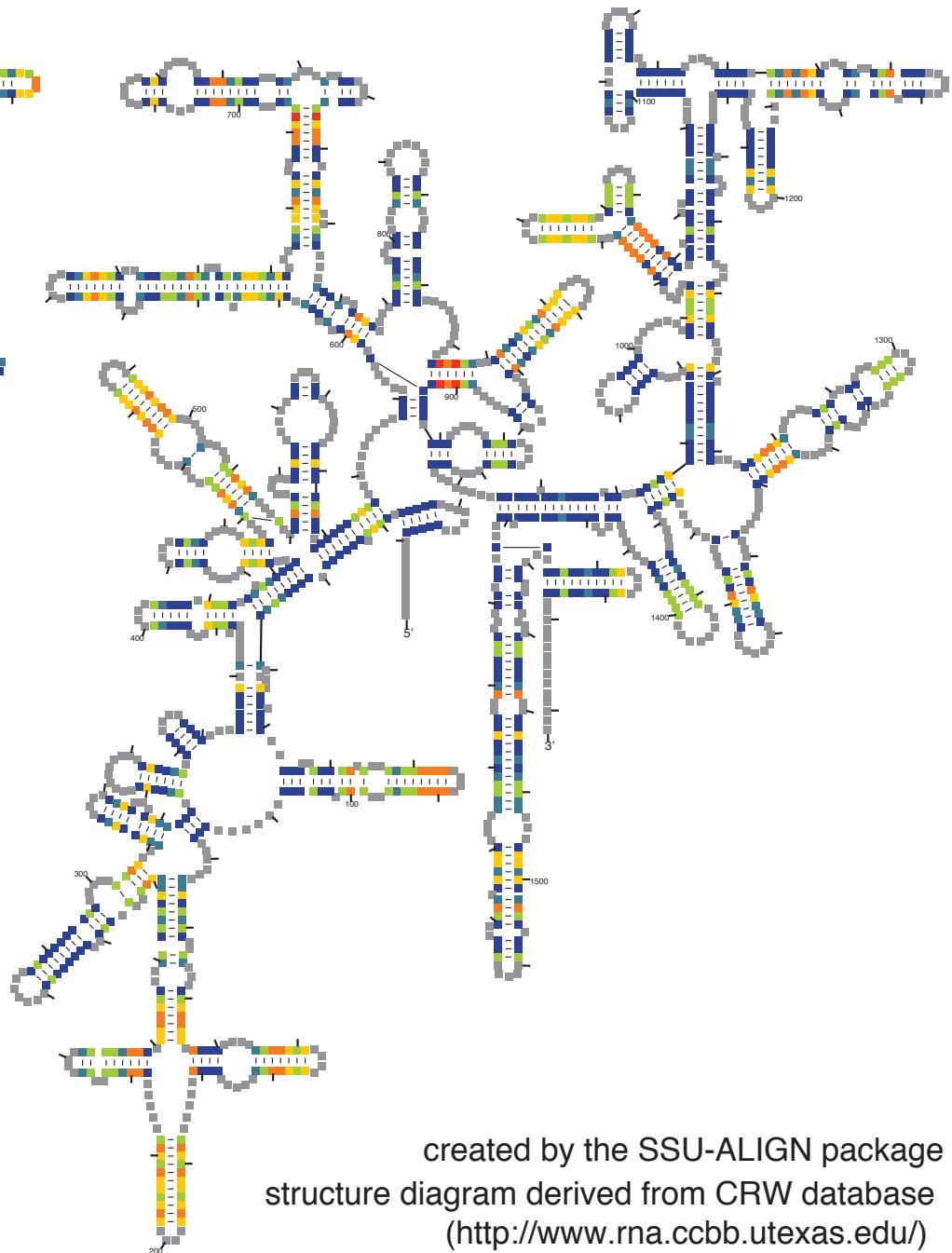
Central dogma of molecular biology



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

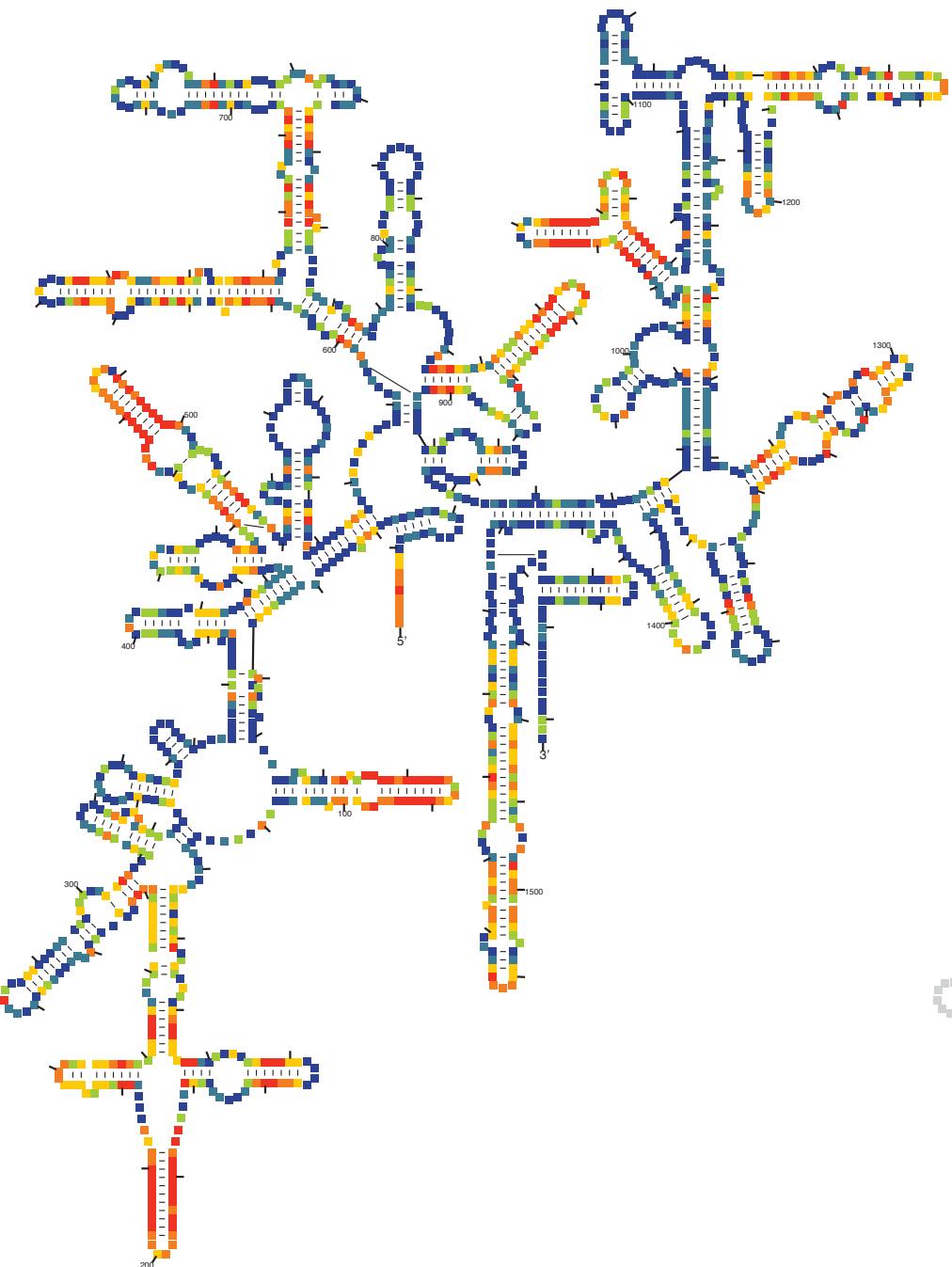


Secondary structure (mutual) information per position
blue: low information red: high information

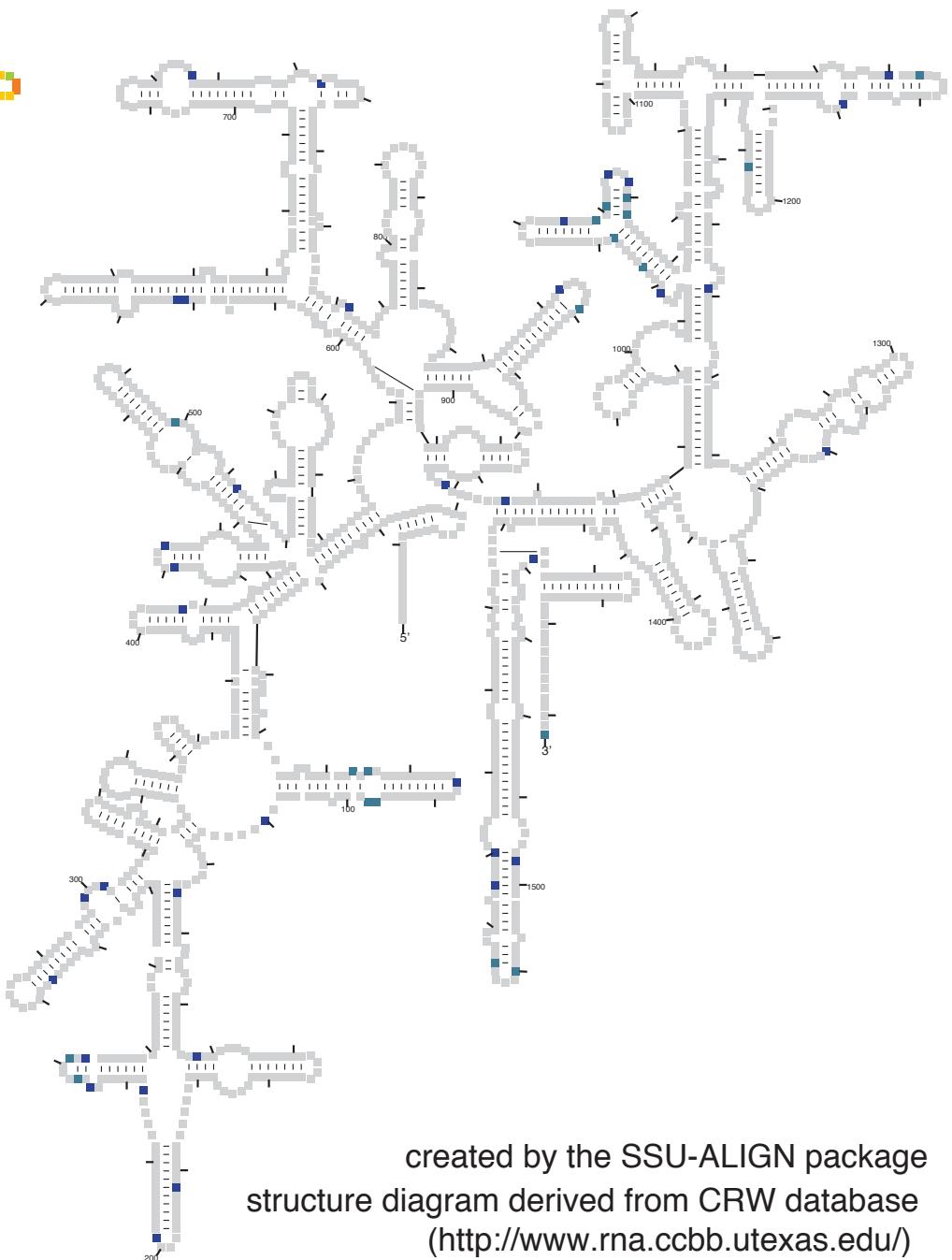


created by the SSU-ALIGN package
structure diagram derived from CRW database
(<http://www.rna.ccbb.utexas.edu/>)

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/>)

Acknowledgements

Alejandro Schäffer

David Landsman

David Lipman Jim Ostell

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	

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
<i>Homo sapiens</i>	3099.7	14 508	796	650	4.8
<i>Sus scrofa (pig)</i>	2808.5	6177	625	460	6.1
<i>Drosophila melanogaster</i>	168.7	4321	156	30	5.7
<i>Caenorhabditis elegans</i>	100.3	1022	175	20	5.2
<i>Saccharomyces cerevisiae</i>	12.2	376	96	1.7	7.3
<i>Escherichia coli</i>	4.6	256	112	0.46	10.2
<i>Bacillus subtilis</i>	4.1	211	52	0.57	7.2
<i>Methanocaldococcus jannaschii</i>	1.7	257	18	0.31	5.6
<i>Aquifex aeolicus</i>	1.6	52	7	0.22	7.3
<i>Borrelia burgdorferi</i>	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

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

Infernal version 1.1.2 (July 2016)

- First release after moving to NCBI (without Infernal as my primary project)
- Main difference is in the `cmscan` executable program
 - faster
 - annotates overlaps

cmsearch versus cmscan: a subtle but important difference

- cmsearch: hits reported by model
 - for each query CM
 - * for each target sequence
 - search for high-scoring hits
- cmscan: hits reported by sequence
 - for each query sequence
 - * for each target CM
 - search for high-scoring hits
- cmscan reads each CM for each sequence (all 2474 models)
- if sequences are short this makes reading CMs a bottleneck

cmsearch is faster than cmscan for datasets with many short sequences

Three sequence datasets:

1. *E. coli* genome: 4.6 Mb, 1 sequence
2. ERS167139 (human gut microbiome, 454): 166 Mb (avg 423nt) 393K sequences
3. ERS235581 (human gut microbiome, HiSeq): 53Mb (avg: 120nt), 440K sequences

	E.coli (4.6 mb, 1 seq)		ERS167139 (393K seqs, avg 423nt)		ERS235581 (440K seqs, avg 120nt)	
program	time	sec/seq	time	sec/seq	time	sec/seq
cmsearch v1.1.1	0.5h	1746.7	28.2h	0.26	9.8h	0.08
cmscan v1.1.1	0.5h	1678.6	37.3h	0.34	20.5h	0.17

cmscan v1.1.2 stores CM model parameters in memory instead of re-reading them for each sequence

Three sequence datasets:

1. *E. coli* genome: 4.6 Mb, 1 sequence
2. ERS167139 (human gut microbiome, 454): 166 Mb (avg 423nt) 393K sequences
3. ERS235581 (human gut microbiome, HiSeq): 53Mb (avg: 120nt), 440K sequences

	E.coli (4.6 mb, 1 seq)		ERS167139 (393K seqs, avg 423nt)		ERS235581 (440K seqs, avg 120nt)	
program	time	sec/seq	time	sec/seq	time	sec/seq
cmsearch v1.1.1	0.5h	1746.7	28.2h	0.26	9.8h	0.08
cmscan v1.1.1	0.5h	1678.6	37.3h	0.34	20.5h	0.17
cmsearch v1.1.2	0.5h	1808.2	25.3h	0.23	8.3h	0.07
cmscan v1.1.2	0.5h	1735.0	26.7h	0.24	9.5h	0.08

Example of overlapping hits in cmscan v1.1.1 output

Infernal v1.1.1:

#target name	accession	query name	accession	mdl	mdl	from	mdl to seq	from	seq to strand	trunc	pass	gc	bias	score	E-value	inc	description of target	
#																		
SSU_rRNA_bacteria	RF00177	G6NTHBW01DJ5GC	-	cm	963	1339	374	1	- 5'&3'	4	0.50	0.0	362.1	3.4e-114	!	-		
SSU_rRNA_archaea	RF01959	G6NTHBW01DJ5GC	-	cm	919	1294	374	1	- 5'&3'	4	0.50	0.0	235.5	1.9e-67	!	-		
SSU_rRNA_eukarya	RF01960	G6NTHBW01DJ5GC	-	cm	1218	1638	374	1	- 5'&3'	4	0.50	0.0	180.6	6.6e-52	!	-		
SSU_rRNA_microsporidia	RF02542	G6NTHBW01DJ5GC	-	cm	841	1142	374	1	- 5'&3'	4	0.50	0.0	120.5	3.5e-34	!	-		
LSU_rRNA_bacteria	RF02541	G6NTHBW01EMUQ5	-	cm	1	243	231	1	- 3'	3	0.49	0.0	189.0	2e-60	!	-		
LSU_rRNA_archaea	RF02540	G6NTHBW01EMUQ5	-	cm	1	224	232	1	- 3'	3	0.49	0.0	155.5	2.7e-46	!	-		
tRNA	RF00005	G6NTHBW01EMUQ5	-	cm	1	71	417	345	- no	1	0.56	0.0	62.6	3.6e-13	!	-		
5_8S_rRNA	RF00002	G6NTHBW01EMUQ5	-	cm	1	154	219	70	- no	1	0.55	0.0	52.3	7.3e-12	!	-		
LSU_rRNA_eukarya	RF02543	G6NTHBW01EMUQ5	-	cm	1	83	84	1	- 3'	3	0.42	0.0	49.3	2.2e-09	!	-		
CRISPR-DR25	RF01338	G6NTHBW01DAB1L	-	cm	1	25	330	307	- no	1	0.29	0.0	21.9	0.18	?	-		
CRISPR-DR45	RF01354	G6NTHBW01DAB1L	-	cm	1	24	330	310	- no	1	0.29	0.0	19.9	2	?	-		
CRISPR-DR17	RF01328	G6NTHBW01DAB1L	-	cm	1	25	330	309	- no	1	0.32	0.0	16.8	6.9	?	-		

cmscan v1.1.2 marks up overlaps

Infernal v1.1.1:

#target name	accession	query name	accession	mdl	mdl from	mdl to seq from	seq to strand	trunc pass	gc	bias	score	E-value	inc	description of target	
SSU_rRNA_bacteria	RF00177	G6NTHBW01DJ5GC	-	cm	963	1339	374	1	- 5'&3'	4 0.50	0.0	362.1	3.4e-114	!	-
SSU_rRNA_archaea	RF01959	G6NTHBW01DJ5GC	-	cm	919	1294	374	1	- 5'&3'	4 0.50	0.0	235.5	1.9e-67	!	-
SSU_rRNA_eukarya	RF01960	G6NTHBW01DJ5GC	-	cm	1218	1638	374	1	- 5'&3'	4 0.50	0.0	180.6	6.6e-52	!	-
SSU_rRNA_microsporidia	RF02542	G6NTHBW01DJ5GC	-	cm	841	1142	374	1	- 5'&3'	4 0.50	0.0	120.5	3.5e-34	!	-
LSU_rRNA_bacteria	RF02541	G6NTHBW01EMUQ5	-	cm	1	243	231	1	- 3'	3 0.49	0.0	189.0	2e-60	!	-
LSU_rRNA_archaea	RF02540	G6NTHBW01EMUQ5	-	cm	1	224	232	1	- 3'	3 0.49	0.0	155.5	2.7e-46	!	-
tRNA	RF00005	G6NTHBW01EMUQ5	-	cm	1	71	417	345	- no	1 0.56	0.0	62.6	3.6e-13	!	-
5_8S_rRNA	RF00002	G6NTHBW01EMUQ5	-	cm	1	154	219	70	- no	1 0.55	0.0	52.3	7.3e-12	!	-
LSU_rRNA_eukarya	RF02543	G6NTHBW01EMUQ5	-	cm	1	83	84	1	- 3'	3 0.42	0.0	49.3	2.2e-09	!	-
CRISPR-DR25	RF01338	G6NTHBW01DAB1L	-	cm	1	25	330	307	- no	1 0.29	0.0	21.9	0.18	?	-
CRISPR-DR45	RF01354	G6NTHBW01DAB1L	-	cm	1	24	330	310	- no	1 0.29	0.0	19.9	2	?	-
CRISPR-DR17	RF01328	G6NTHBW01DAB1L	-	cm	1	25	330	309	- no	1 0.32	0.0	16.8	6.9	?	-

Infernal v1.1.2:

#idx	target name	accession	query name	accession	clan name	mdl	mdl from	mdl to seq from	seq to strand	trunc pass	gc	bias	score	E-value	inc	olp	anyidx	afrct1	afrct2	
1	SSU_rRNA_bacteria	RF00177	G6NTHBW01DJ5GC	-	CL00111	cm	963	1339	374	1	- 5'&3'	4 0.50	0.0	362.1	3.4e-114	!	^	-	-	-
2	SSU_rRNA_archaea	RF01959	G6NTHBW01DJ5GC	-	CL00111	cm	919	1294	374	1	- 5'&3'	4 0.50	0.0	235.5	1.9e-67	!	=	1	1.000	1.000
3	SSU_rRNA_eukarya	RF01960	G6NTHBW01DJ5GC	-	CL00111	cm	1218	1638	374	1	- 5'&3'	4 0.50	0.0	180.6	6.6e-52	!	=	1	1.000	1.000
4	SSU_rRNA_microsporidia	RF02542	G6NTHBW01DJ5GC	-	CL00111	cm	841	1142	374	1	- 5'&3'	4 0.50	0.0	120.5	3.5e-34	!	=	1	1.000	1.000
1	LSU_rRNA_bacteria	RF02541	G6NTHBW01EMUQ5	-	CL00112	cm	1	243	231	1	- 3'	3 0.49	0.0	189.0	2e-60	!	^	-	-	-
2	LSU_rRNA_archaea	RF02540	G6NTHBW01EMUQ5	-	CL00112	cm	1	224	232	1	- 3'	3 0.49	0.0	155.5	2.7e-46	!	=	1	0.996	1.000
3	tRNA	RF00005	G6NTHBW01EMUQ5	-	CL00001	cm	1	71	417	345	- no	1 0.56	0.0	62.6	3.6e-13	!	*	-	-	-
4	5_8S_rRNA	RF00002	G6NTHBW01EMUQ5	-	-	cm	1	154	219	70	- no	1 0.55	0.0	52.3	7.3e-12	!	*	-	-	-
5	LSU_rRNA_eukarya	RF02543	G6NTHBW01EMUQ5	-	CL00112	cm	1	83	84	1	- 3'	3 0.42	0.0	49.3	2.2e-09	!	=	1	1.000	0.364
1	CRISPR-DR25	RF01338	G6NTHBW01DAB1L	-	CL00014	cm	1	25	330	307	- no	1 0.29	0.0	21.9	0.18	?	^	-	-	-
2	CRISPR-DR45	RF01354	G6NTHBW01DAB1L	-	-	cm	1	24	330	310	- no	1 0.29	0.0	19.9	2	?	*	-	-	-
3	CRISPR-DR17	RF01328	G6NTHBW01DAB1L	-	CL00014	cm	1	25	330	309	- no	1 0.32	0.0	16.8	6.9	?	=	1	1.000	0.917

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.