

RNA homology search using covariance models

Eric Nawrocki

National Center for Biotechnology Information
National Institutes of Health

Howard Hughes Medical Institute
Janelia Research Campus



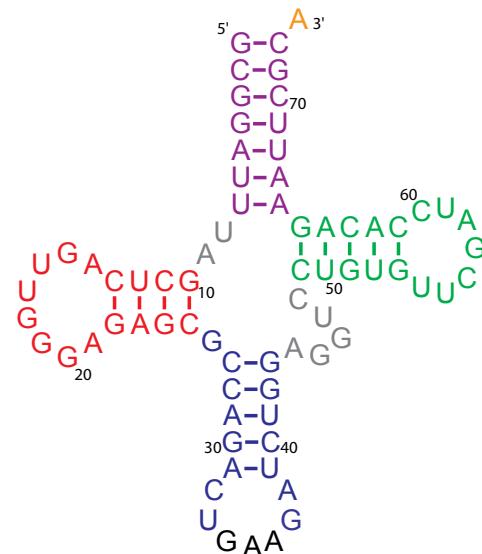
Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence

GC₁GGAUUUAAGCUCAGUUGGG
AGAGC₂GCCAGACU₃GAAGAUC
UGGAGGUCC₄UGUGUUCGAUC
CACAGAAUUCGCA₅

Secondary structure



3-dimensional structure



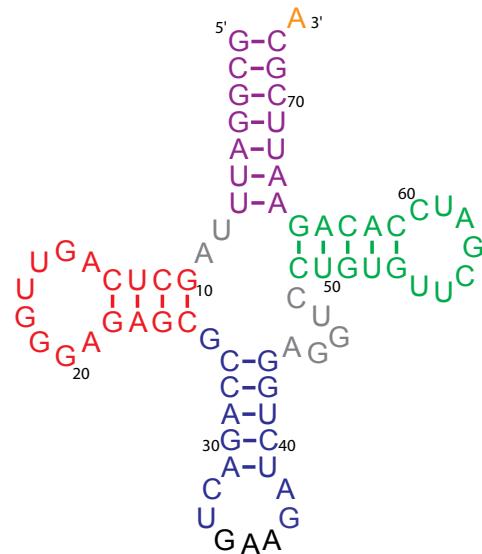
Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence

GC₁GGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUC₂CUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



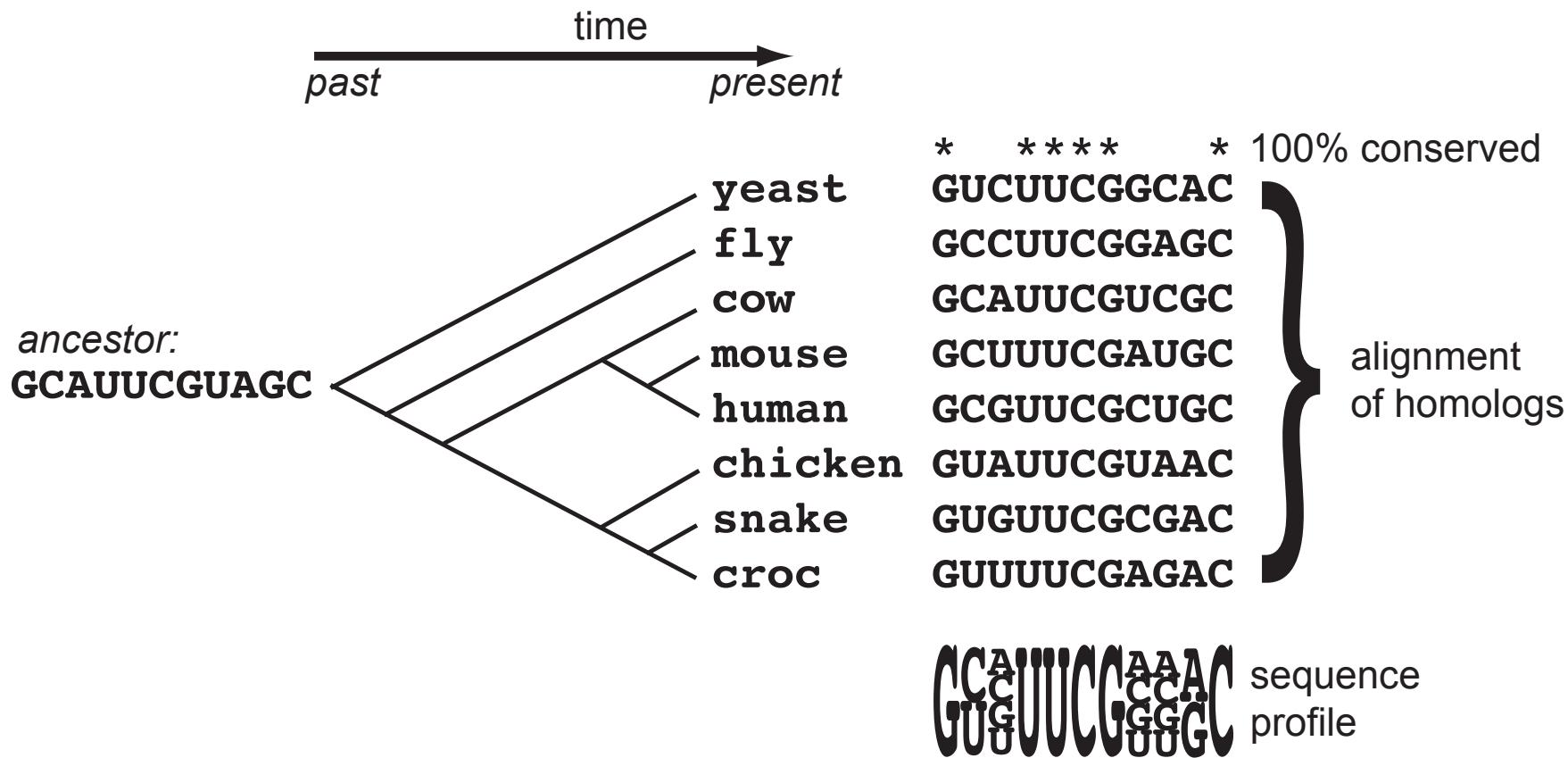
3-dimensional structure



- BLAST: given a single sequence, search genomes for similar sequences.
- BLAST cannot take advantage of:
 - sequence conservation, which varies across the gene
 - secondary structure

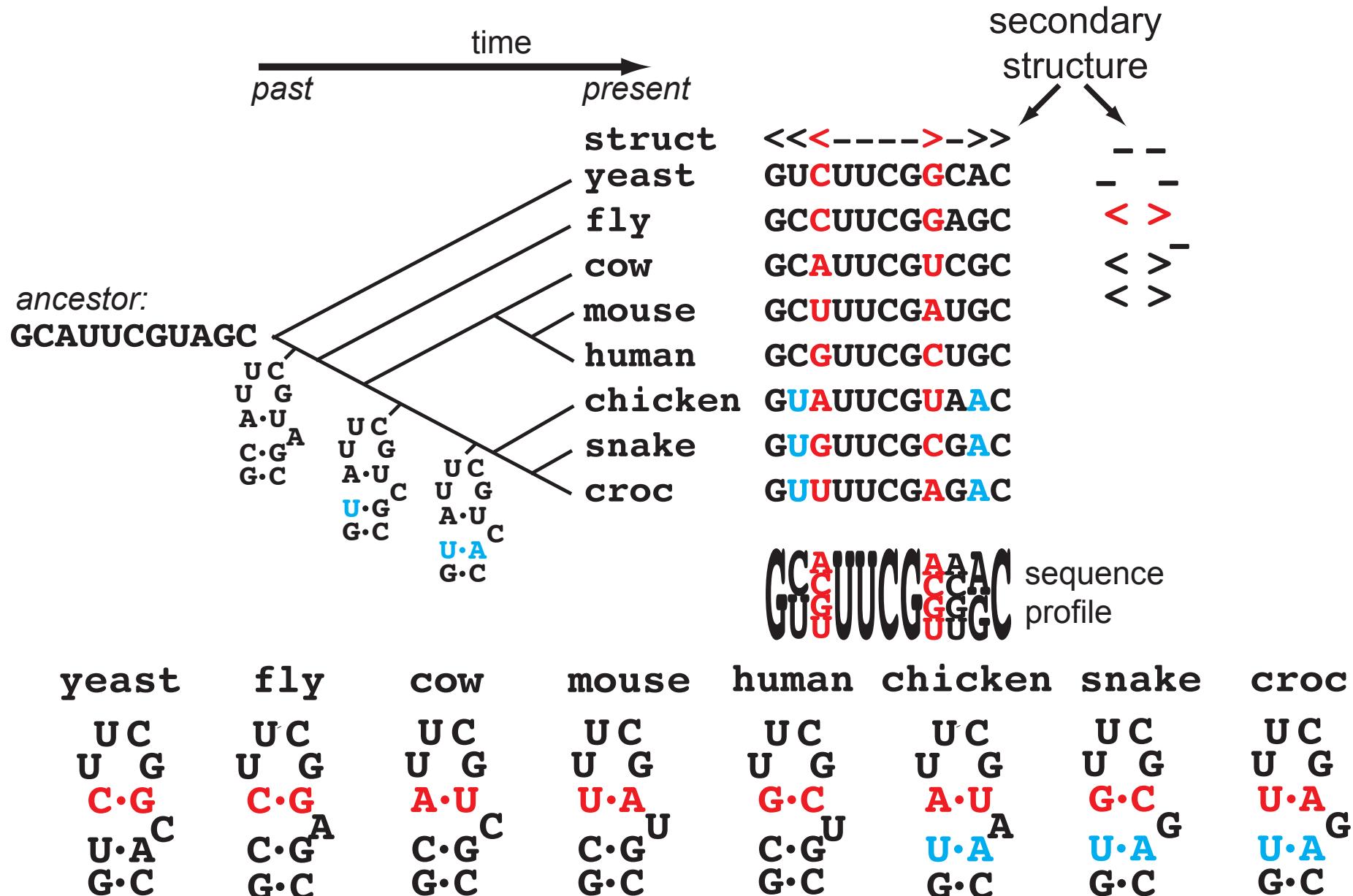
Sequence conservation provides information for homology searches

Conservation levels vary across alignment columns.

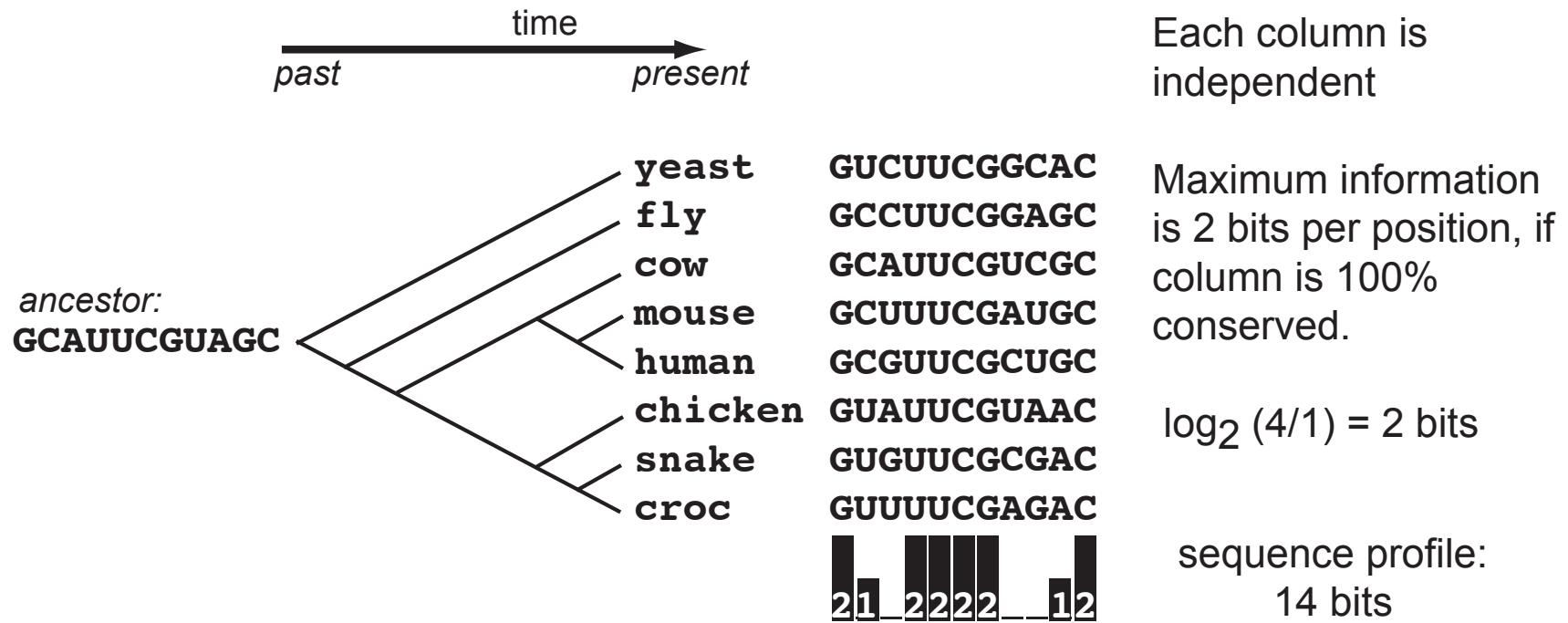


Structure conservation provides additional information

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

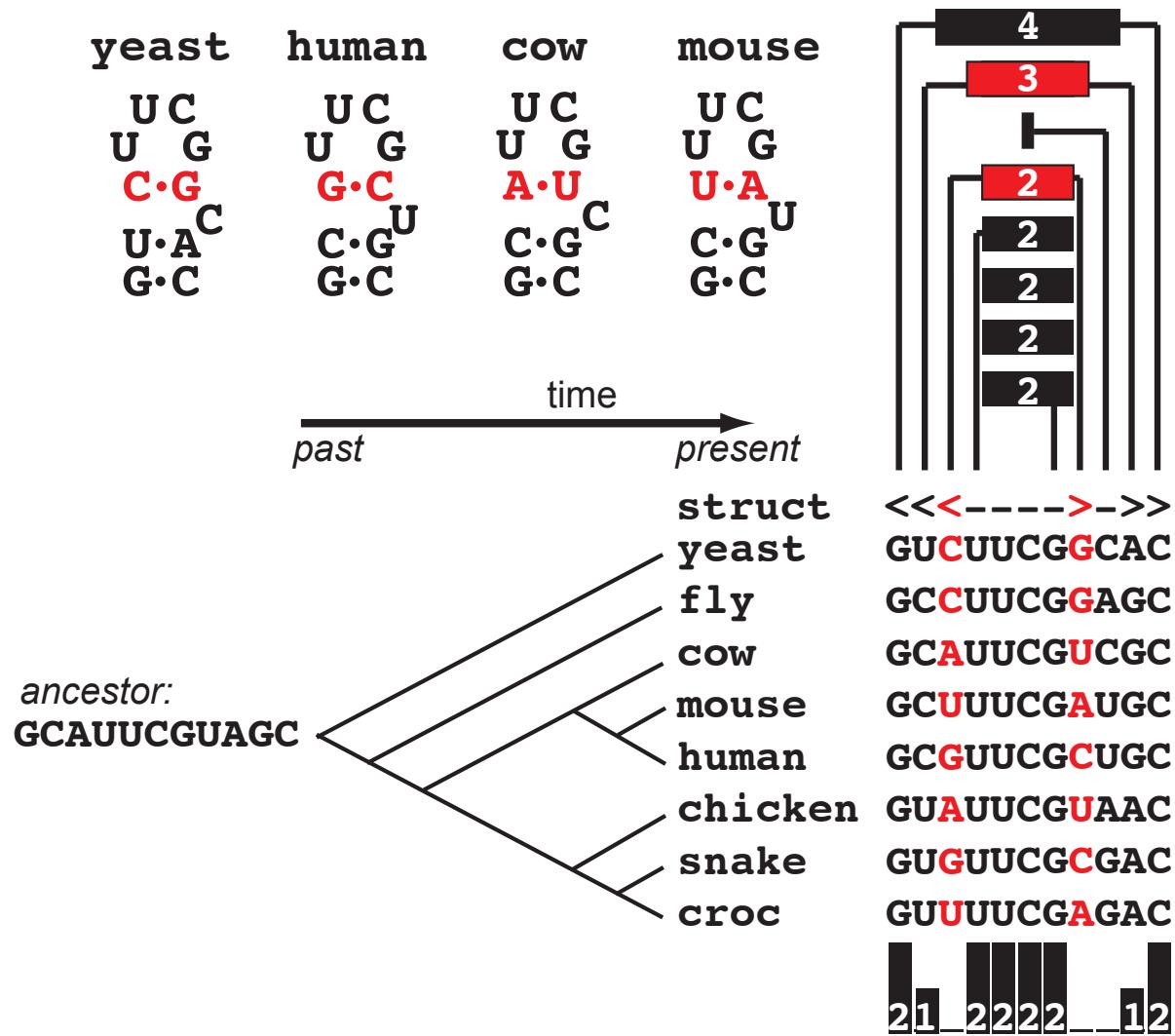


Amount of information in a profile can be measured in bits



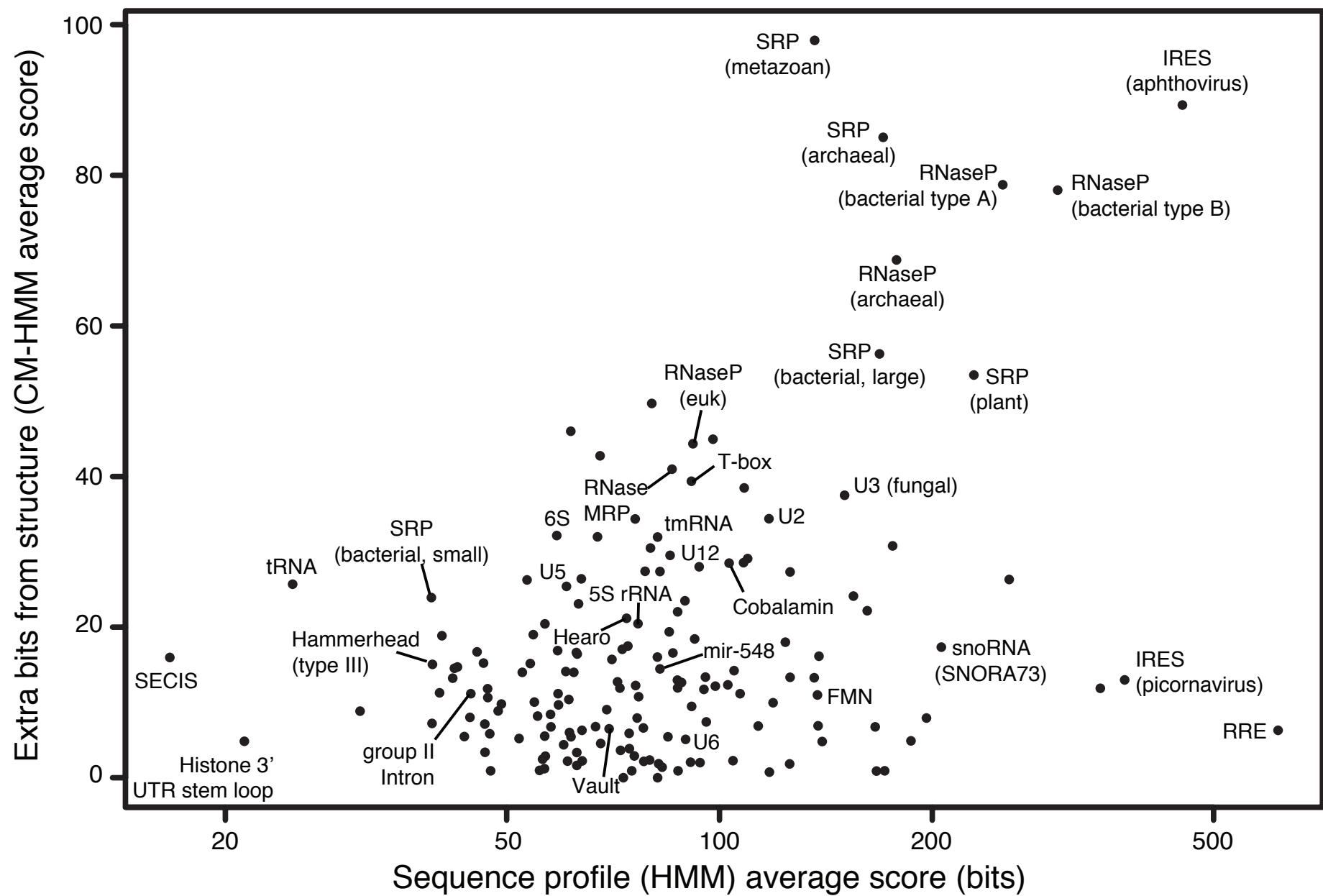
expect a match by chance: 1 in 2^{14} nt $= \sim 16$ Kb

Structure contributes additional information from covariation



expect a match by chance: 1 in 2^{17} nt \approx 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 (14831 and 1132 entries)	Rfam (2450 families)
performance for RNAs	faster but less accurate	slower but more accurate

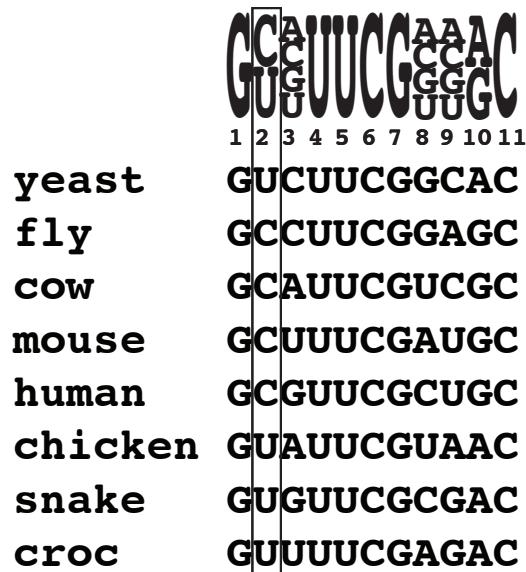


<http://hmmer.janelia.org>
Eddy, SR. PLoS Comp. Biol.,
7:e1002195, 2011.
Eddy, SR. PLoS Comp. Biol.,
4:e1000069, 2008.
Eddy, SR. Bioinformatics,
14:755-763, 1998.



<http://infernal.janelia.org>
Nawrocki EP, Eddy SR
Bioinformatics,
29:2933-2935, 2013.
Eddy SR, Durbin R.
Nucleic Acids Research,
22:2079-2088, 1994.

Profile HMMs: sequence family models built from alignments



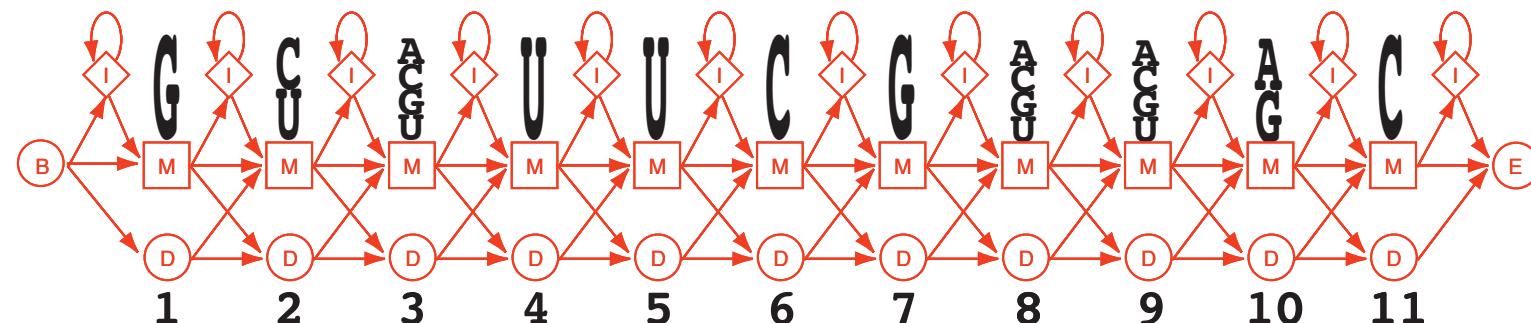
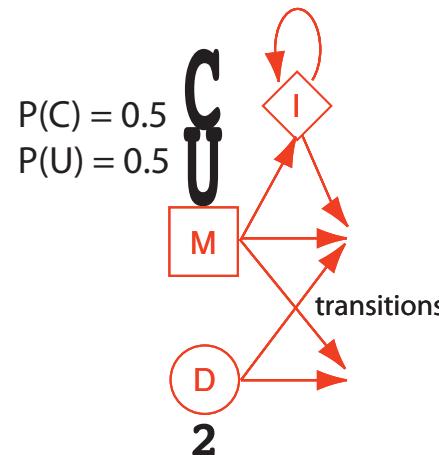
One HMM node per alignment column

3 states per node:

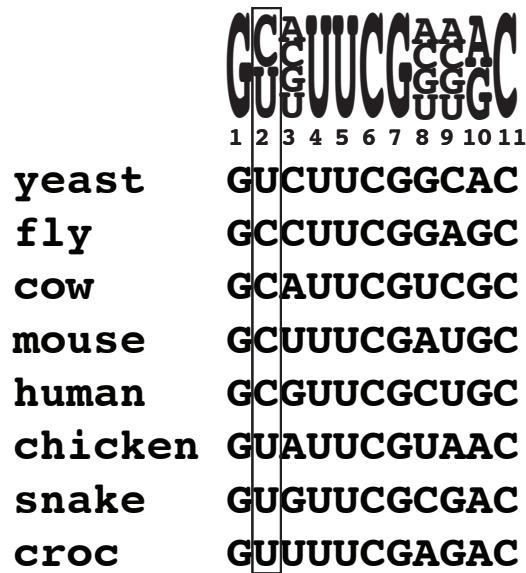
- (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

HMMs generate homologous sequences.

Node for column 2:



Profile HMMs: sequence family models built from alignments



One HMM node per alignment column

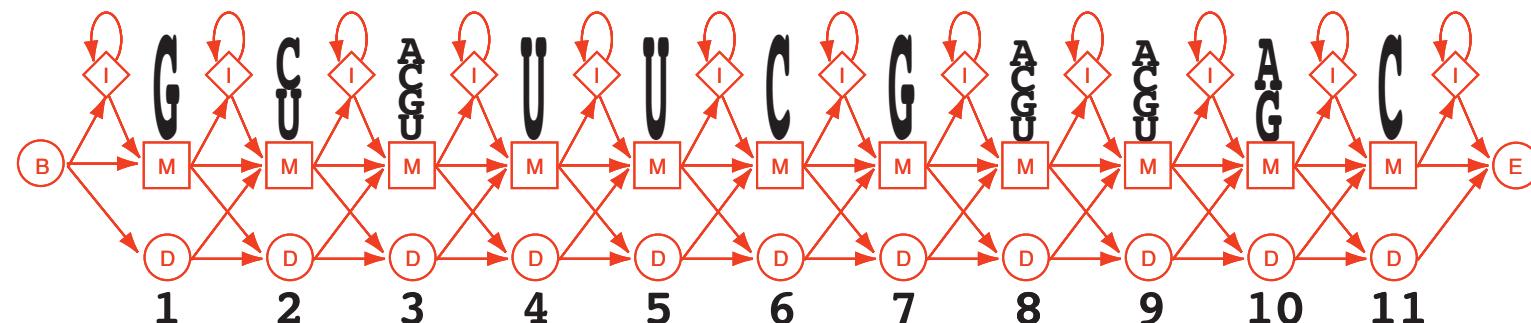
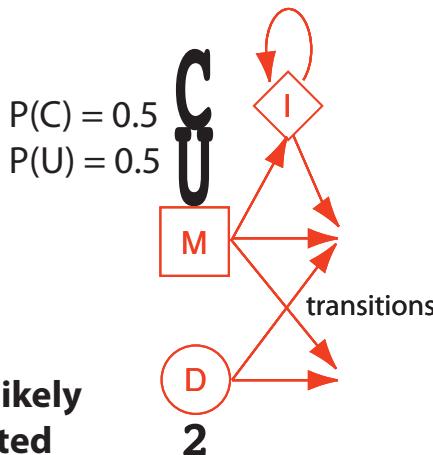
3 states per node:

- (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

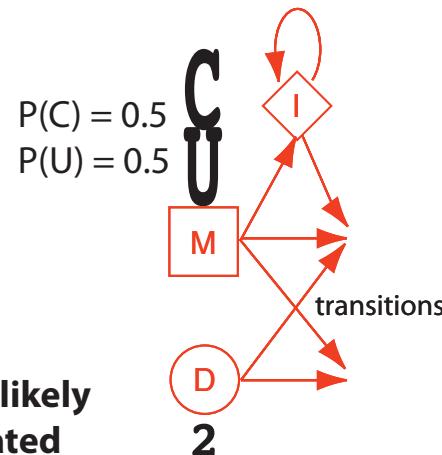
	yeast	GCA GUUUCGGAC 1 2 3 4 5 6 7 8 9 10 11
	fly	GCCUUUCGGAGC
	cow	GCAUUCGUCGC
	mouse	GCUUUCGAUGC
	human	GCGUUCGCUGC
	chicken	GUAUUCGUAAC
	snake	GUGUUCGCGAC
	croc	GUUUUCGAGAC
	worm	GCGUUCGCGGC

One HMM node per alignment column

3 states per node:

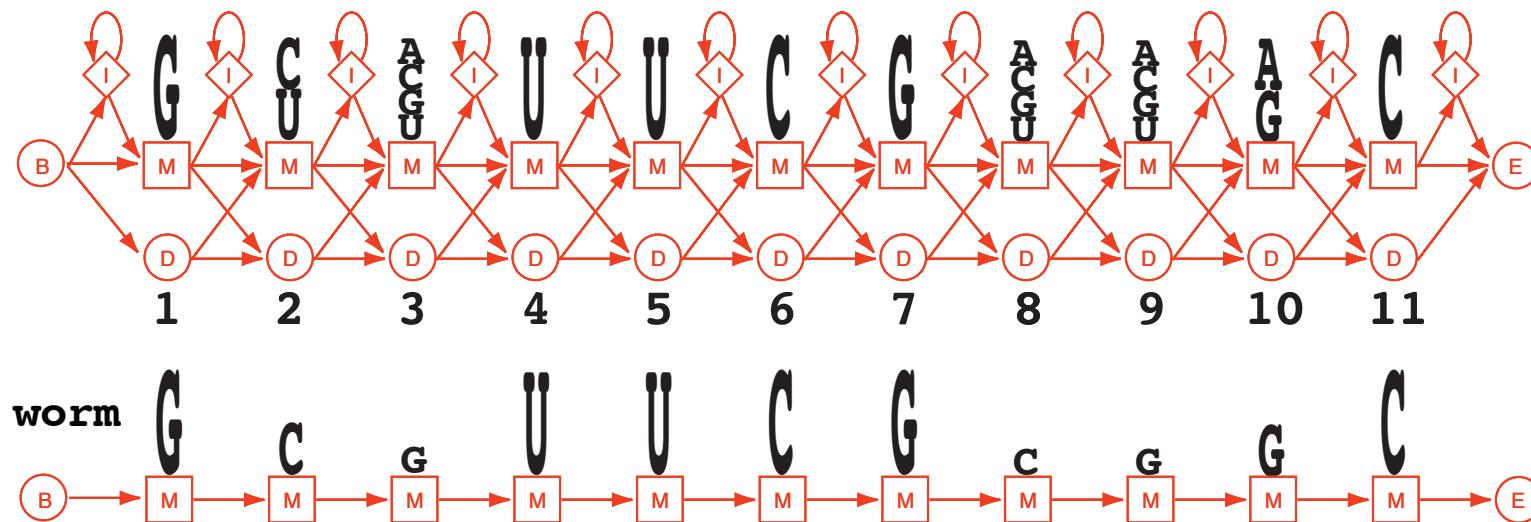
- (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

Node for column 2:

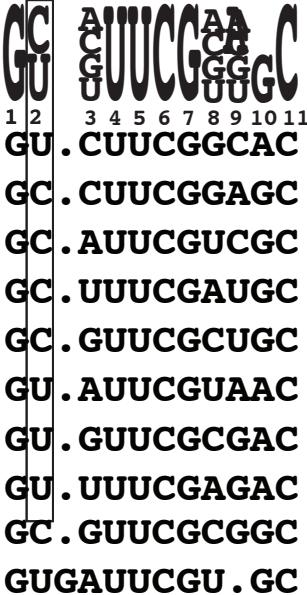


HMMs generate homologous sequences.

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



Profile HMMs: sequence family models built from alignments

	
yeast	GU. C UUCGGC A C
fly	GC. C UUCGGAGC
cow	GC. C AUUCGUCGC
mouse	GC. C UUUCGAUGC
human	GC. C GUUCGCUGC
chicken	GU. C AUUCGUAAC
snake	GU. C GUUCGCGAC
croc	GU. C UUUCGAGAC
worm	GC. C GUUCGCGGC
corn	GUGAUUCGU. C G

One HMM node per alignment column

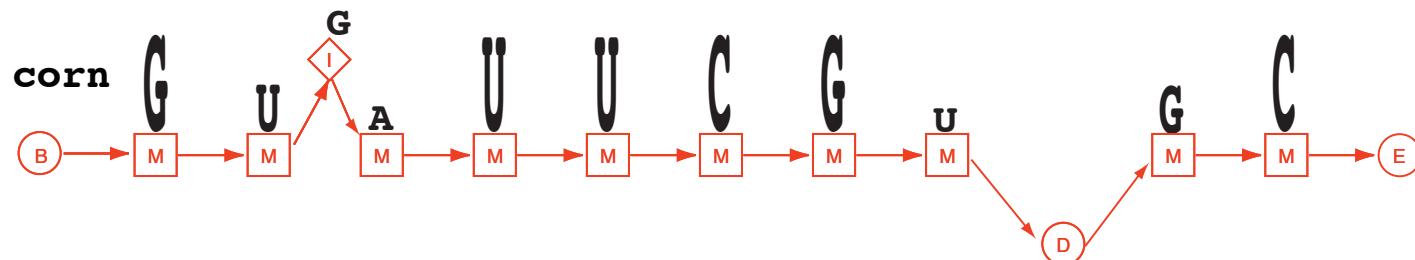
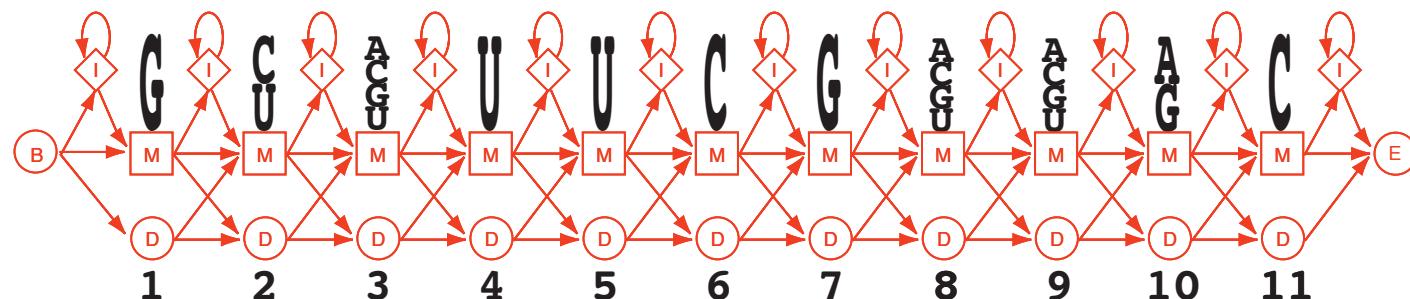
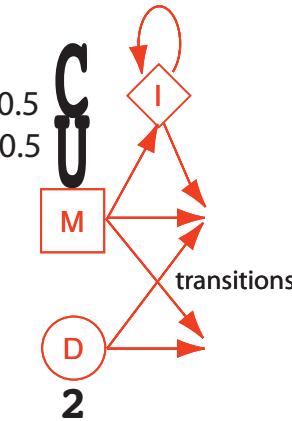
3 states per node:

- (M) Match: emits residues
- (I) Insert: inserts extra residues
- (D) Delete: deletes residues

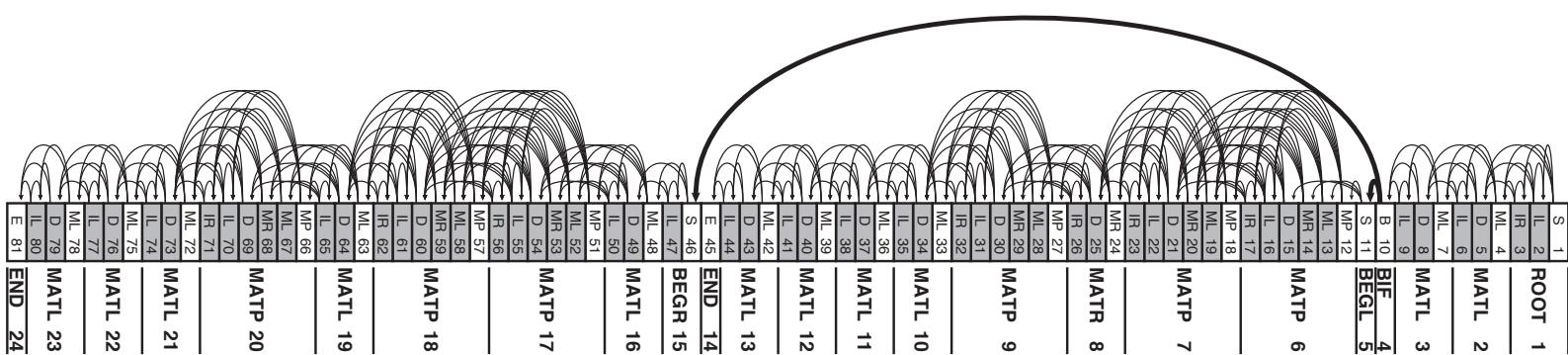
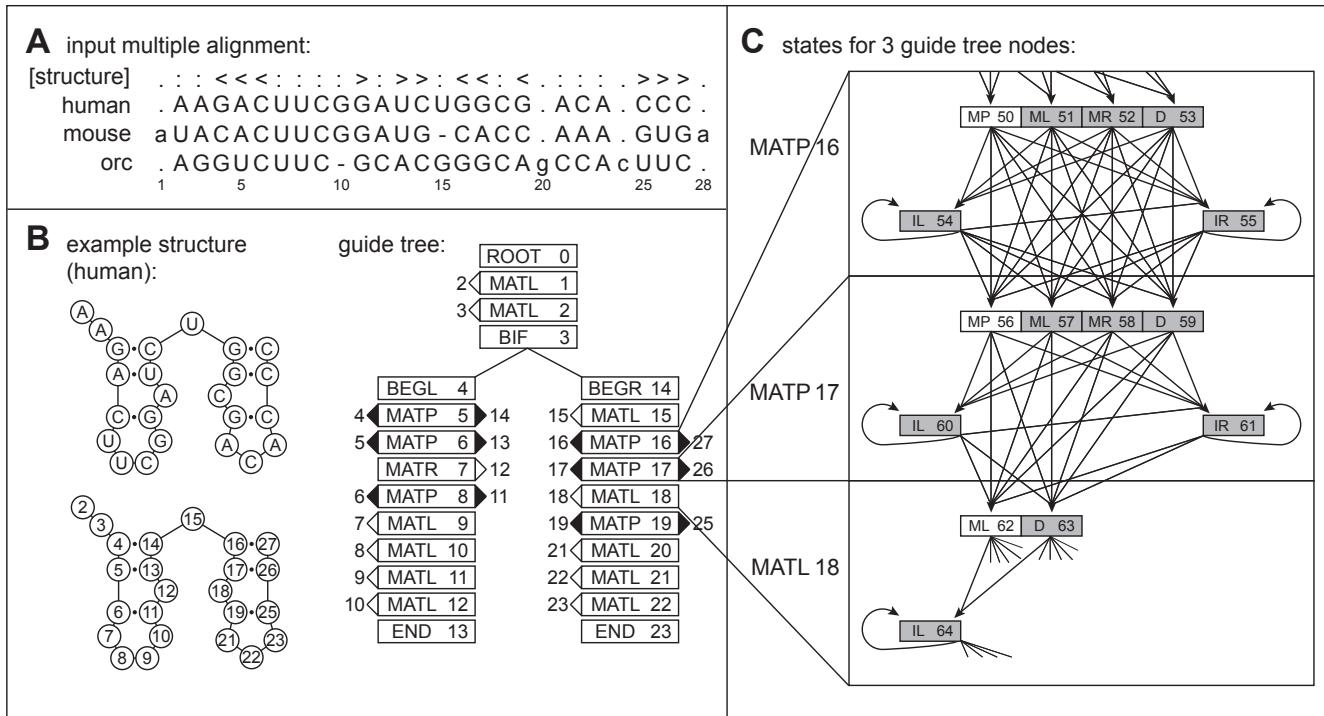
HMMs generate homologous sequences.

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

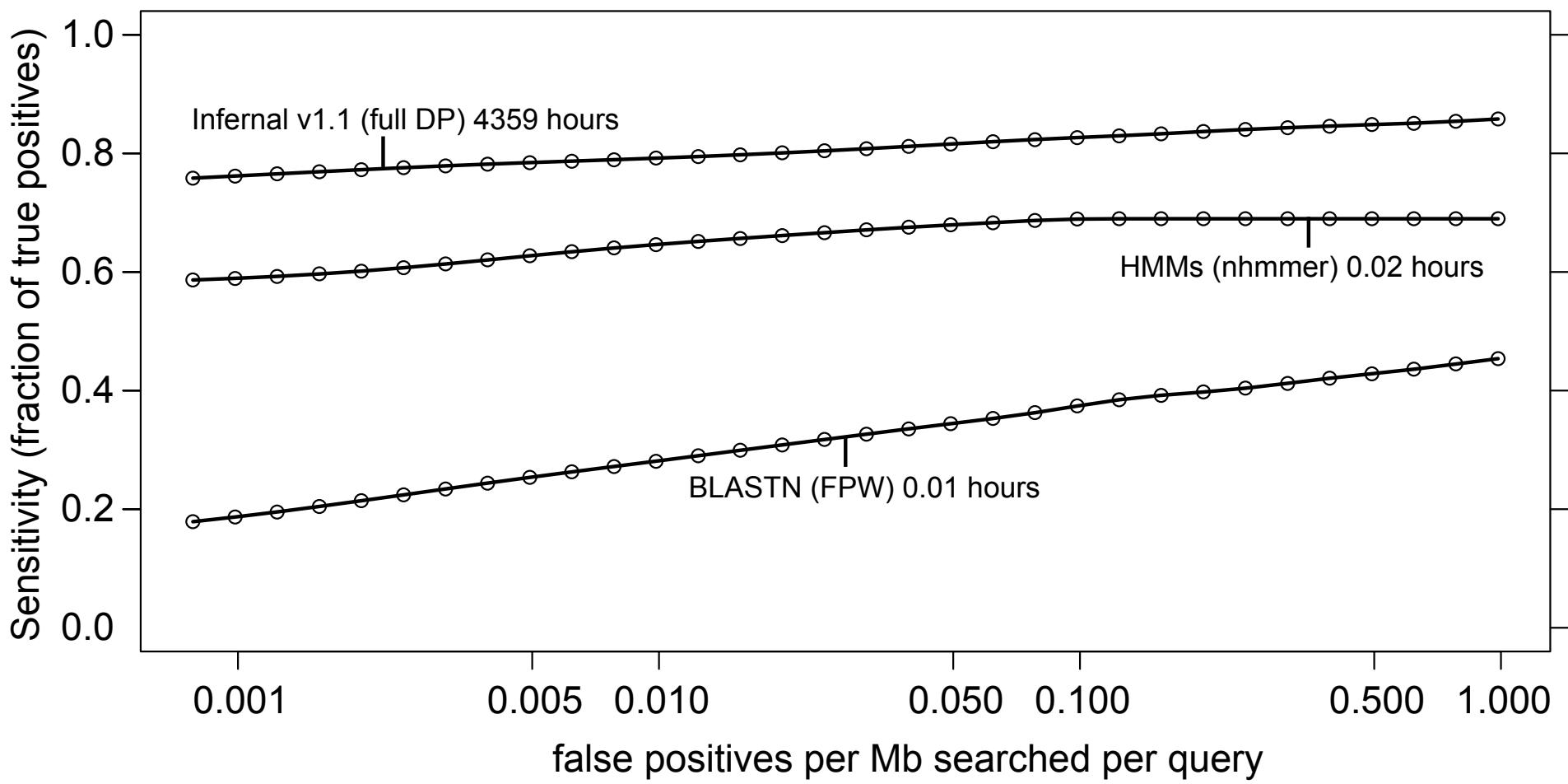
Node for column 2:



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



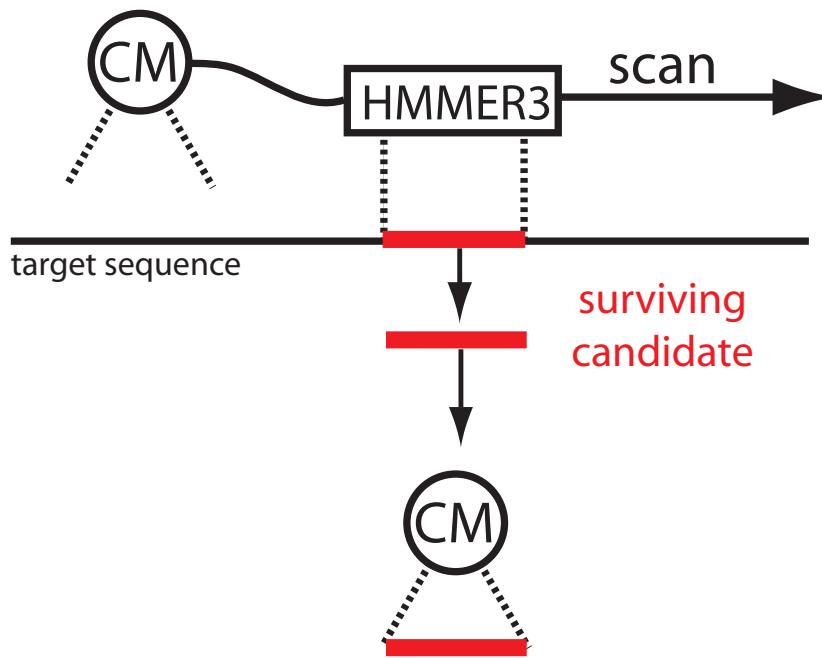
Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)



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

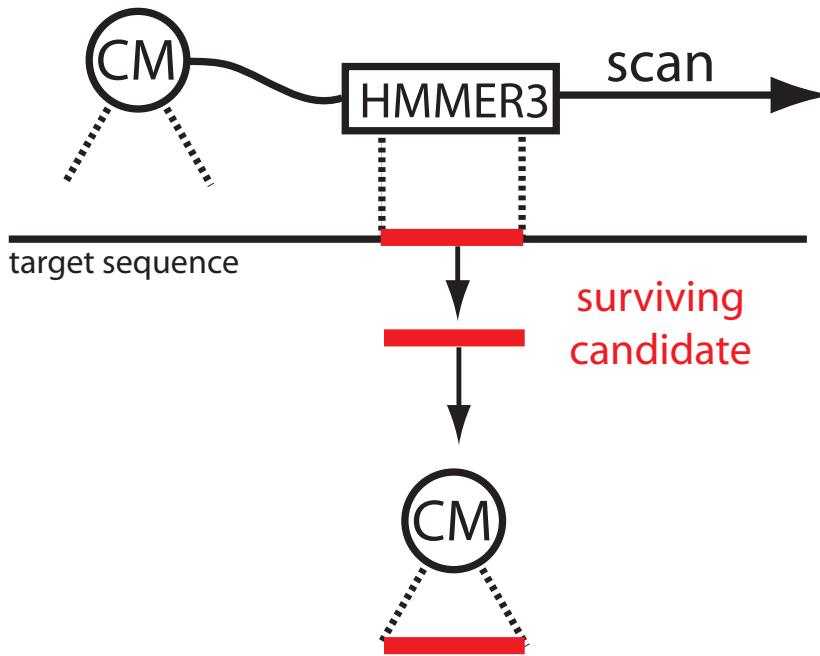
Filter target database using profile HMMs*

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.

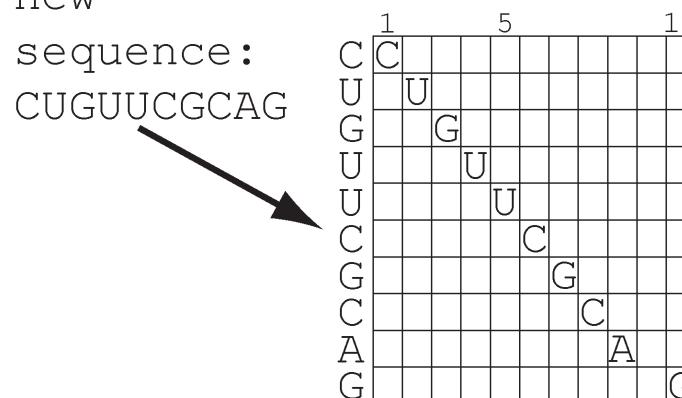
*Weinberg, Ruzzo, RECOMB, 243-251, 2004; Weinberg, Ruzzo, Bioinformatics, 22(1) 35-39 2006.

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

yeast	GUGUUCGCUAC
human	-UCUUCGGCG-
fly	AGAUU-GUACU

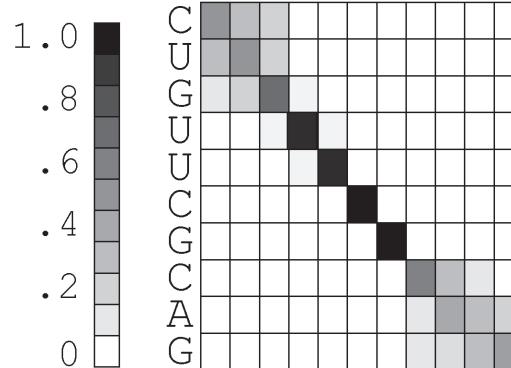
new

sequence:
CUGUUCGCGAG



probability

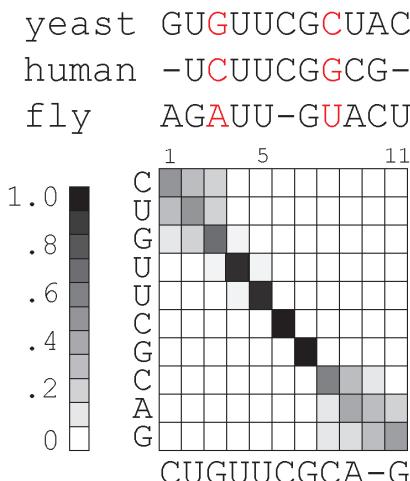
"correct":



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

HMMs -

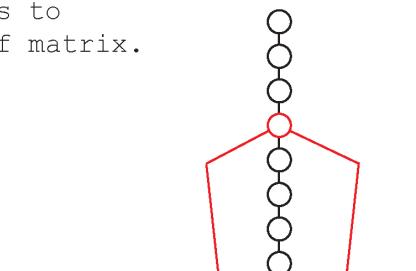
Each column of seed alignment corresponds to a column of matrix.



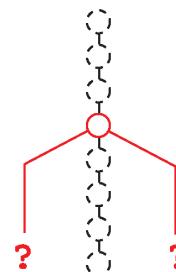
CMs -

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 AGAUU-G**U**ACU

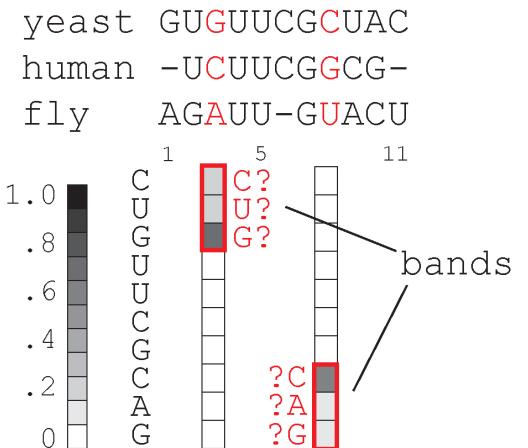


CUGUUCGCAG
 45 possibilities

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

HMMs -

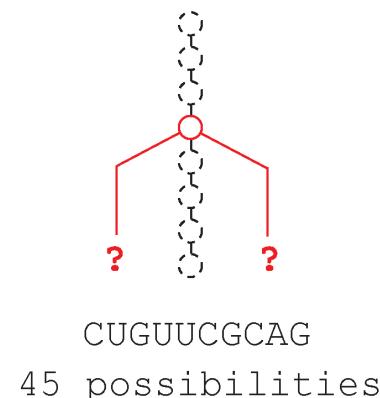
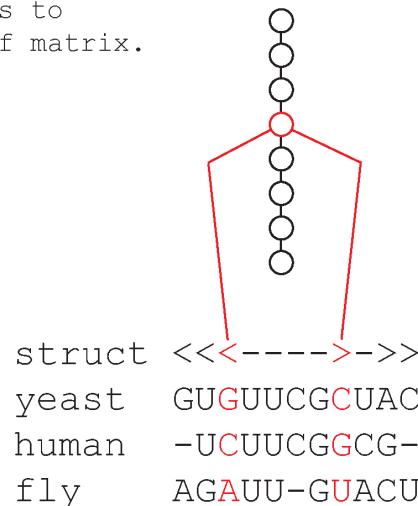
Each column of seed alignment corresponds to a column of matrix.



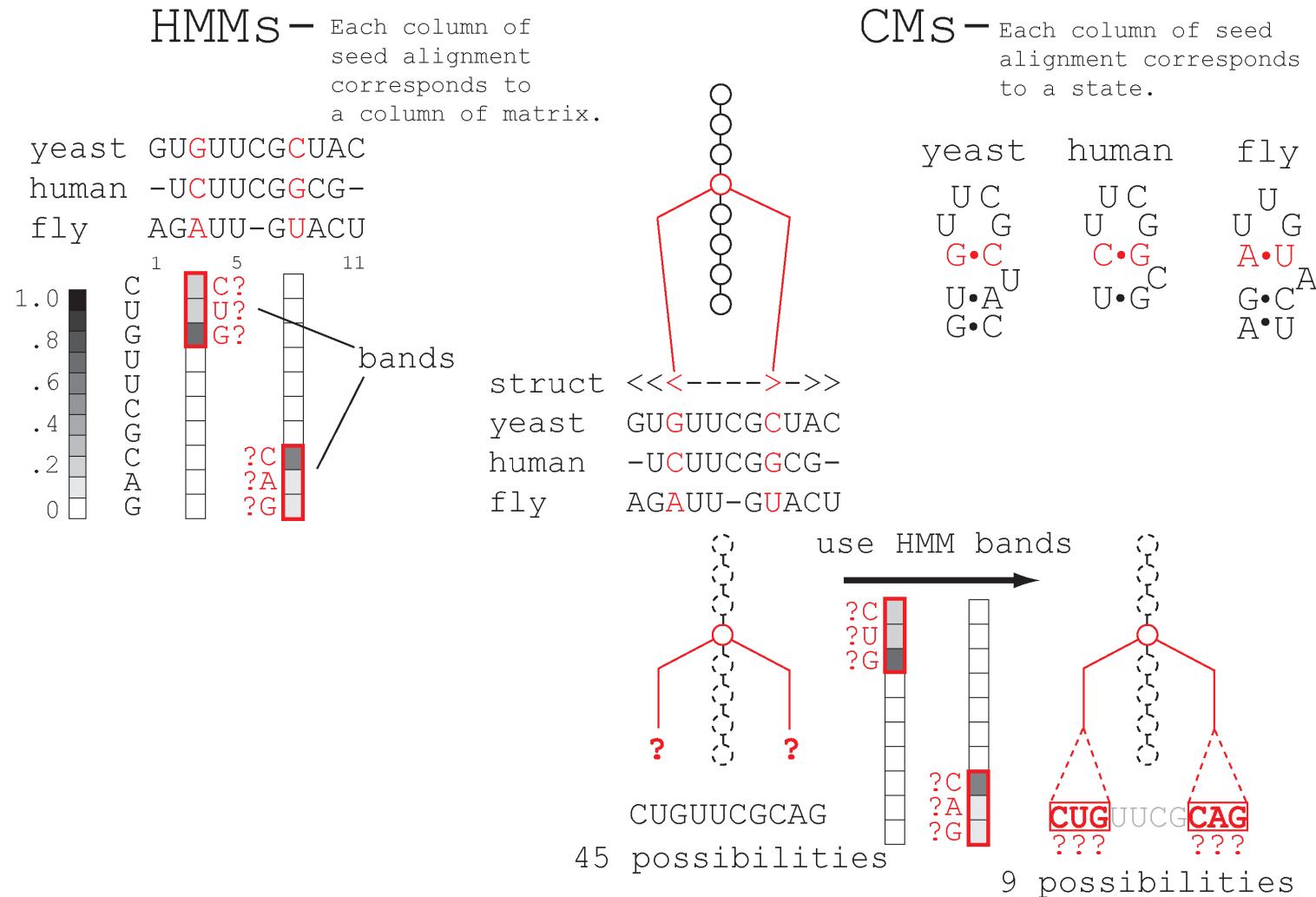
CMs -

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

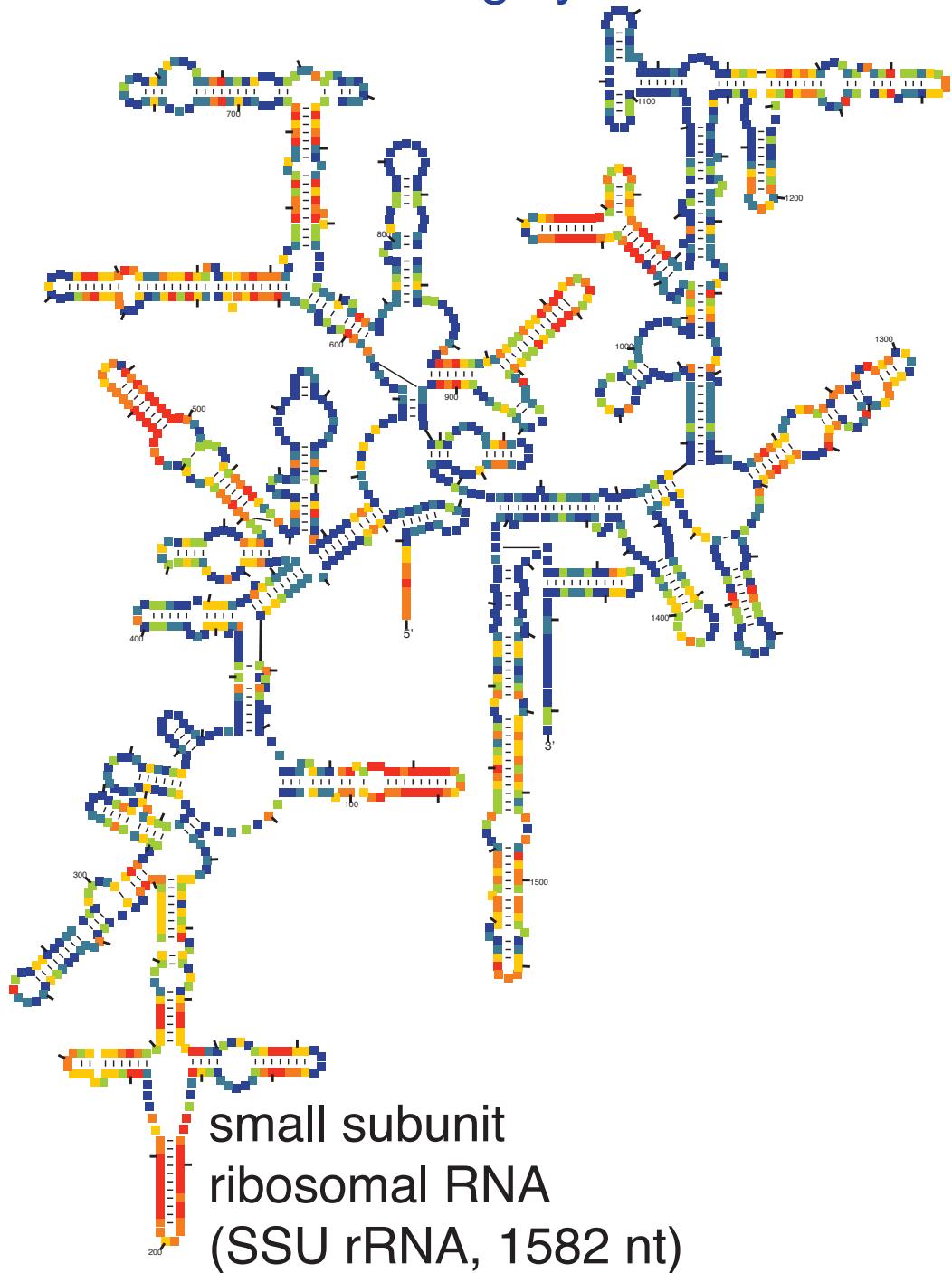


Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*



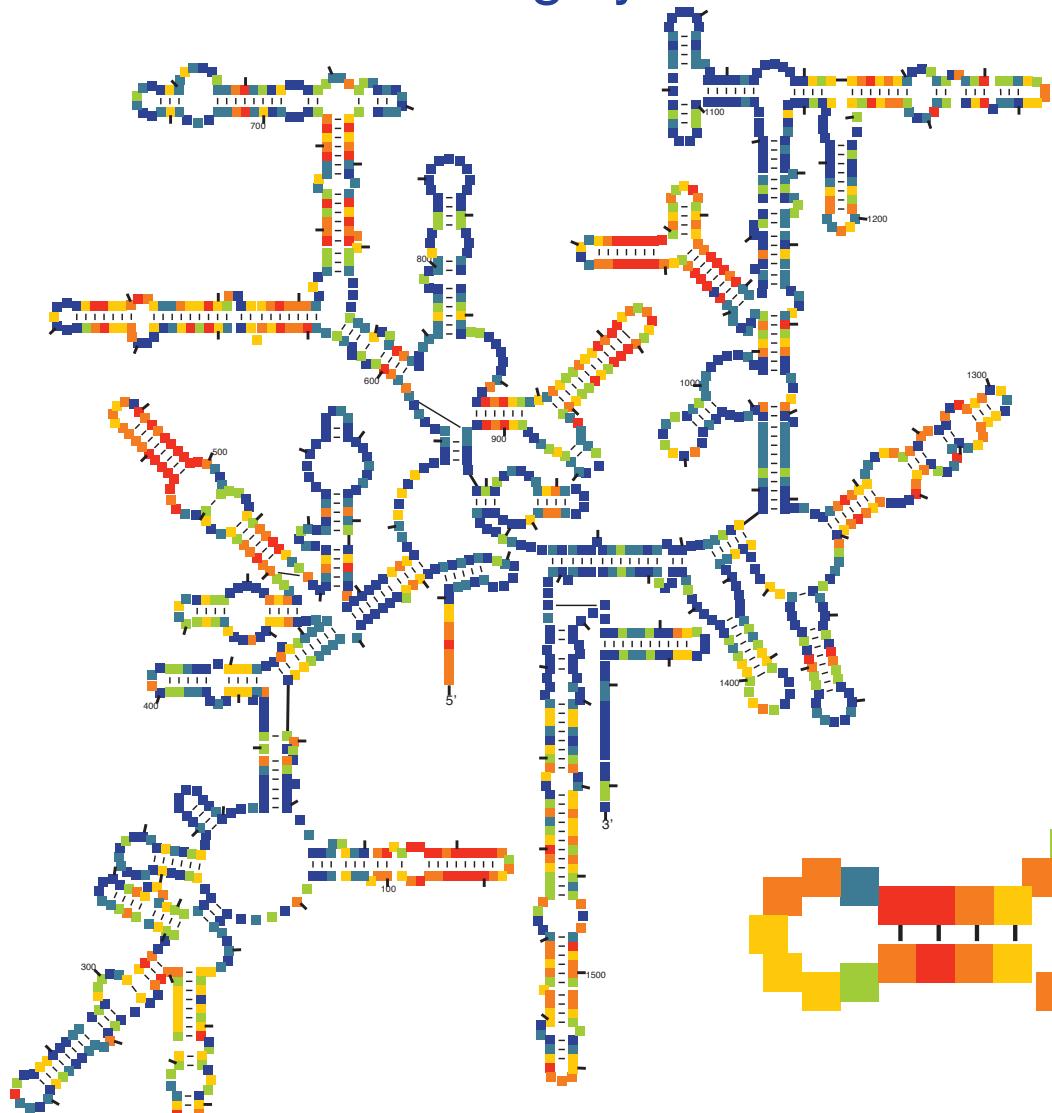
Sequence conservation per position

blue:highly conserved red: highly variable

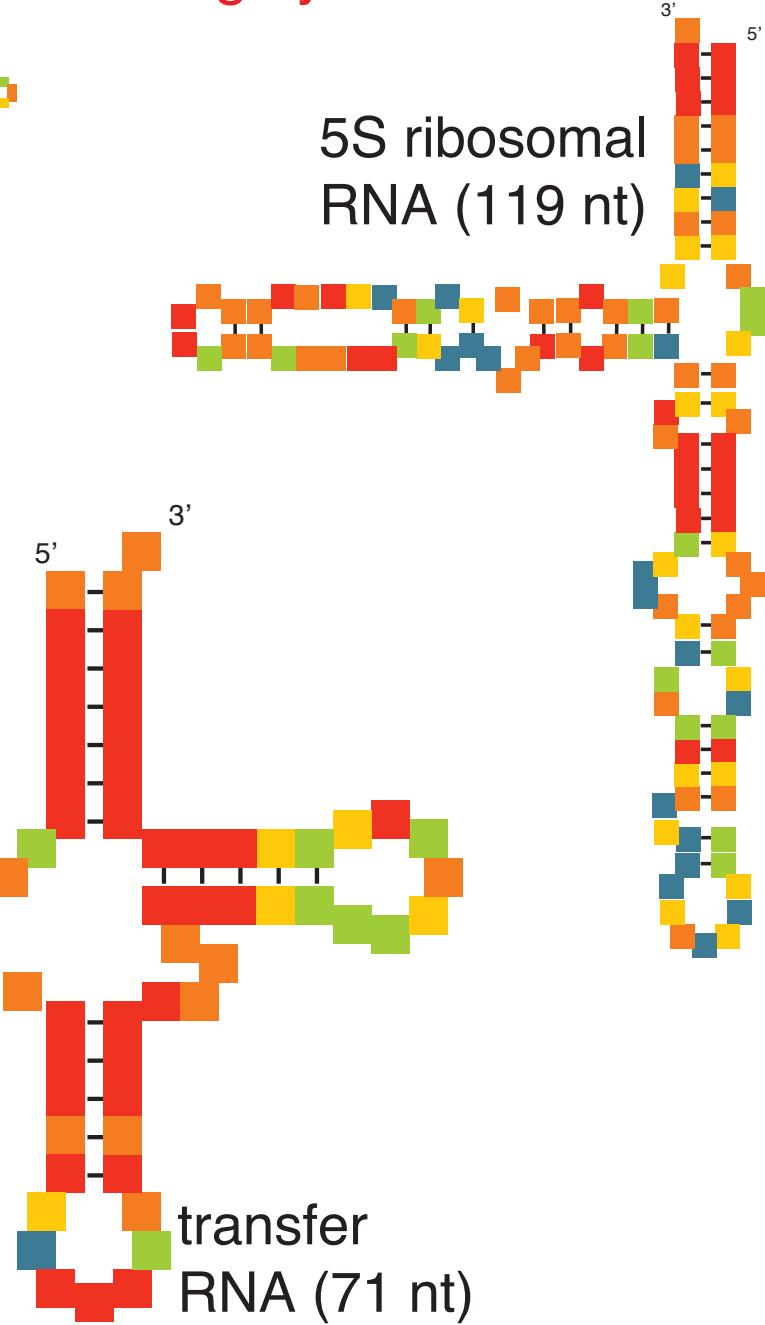


Sequence conservation per position

blue:highly conserved red: highly variable



small subunit
ribosomal RNA
(SSU rRNA, 1582 nt)

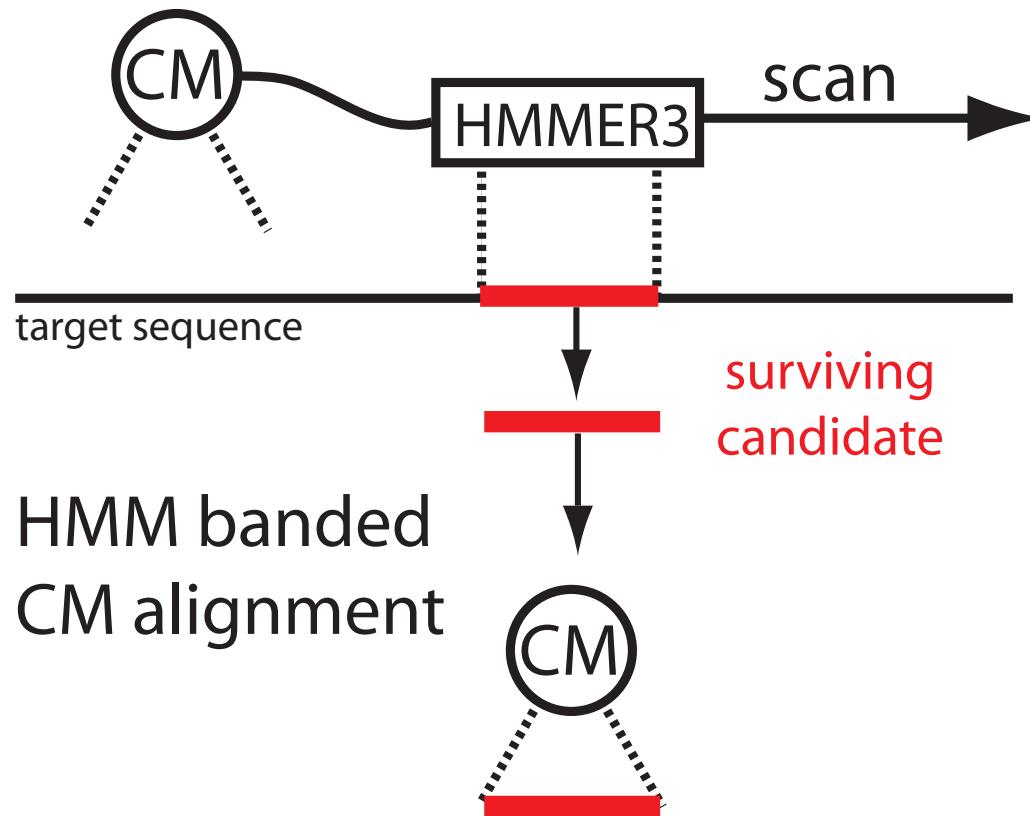


5S ribosomal
RNA (119 nt)

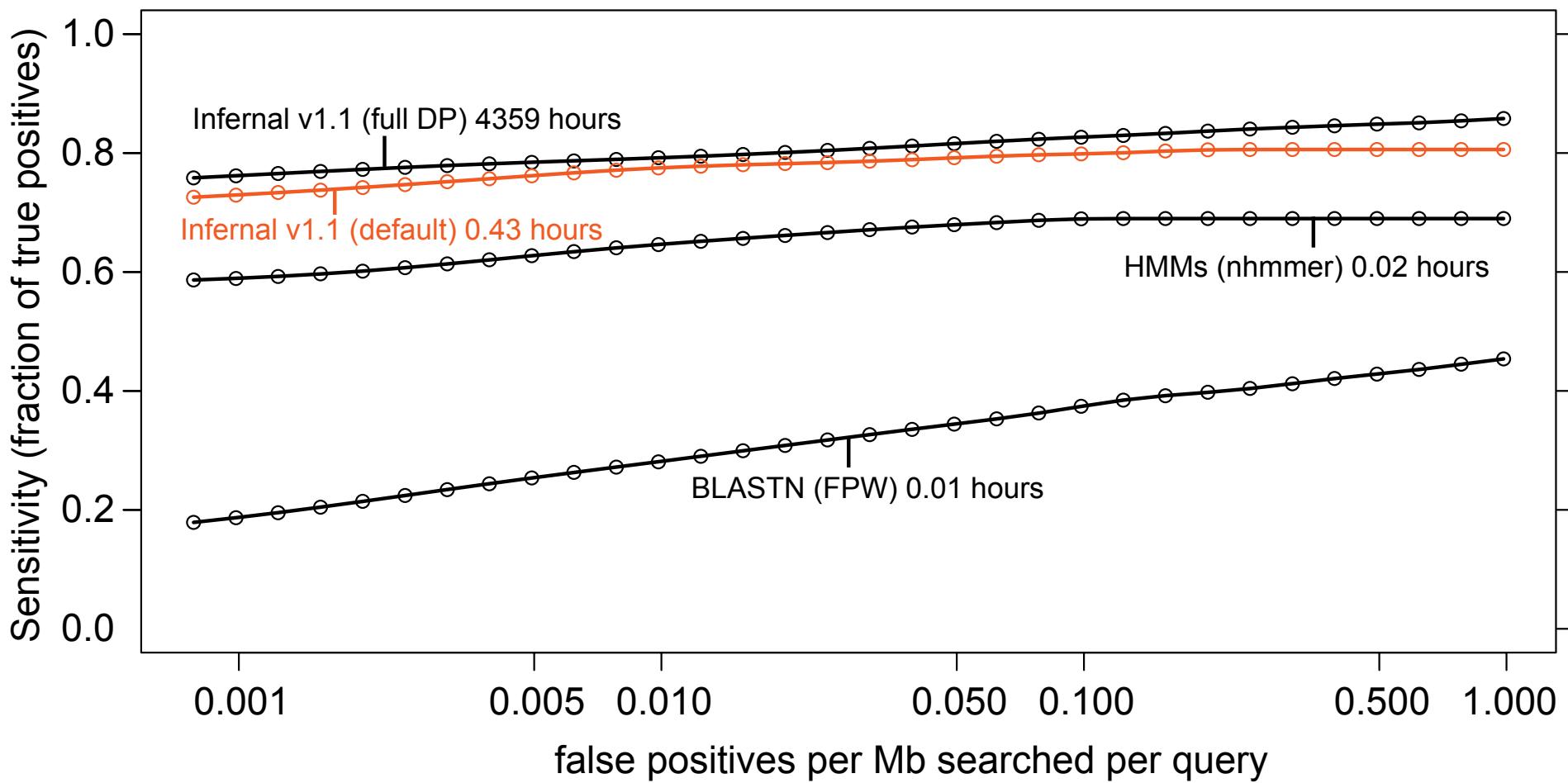
transfer
RNA (71 nt)

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.

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.

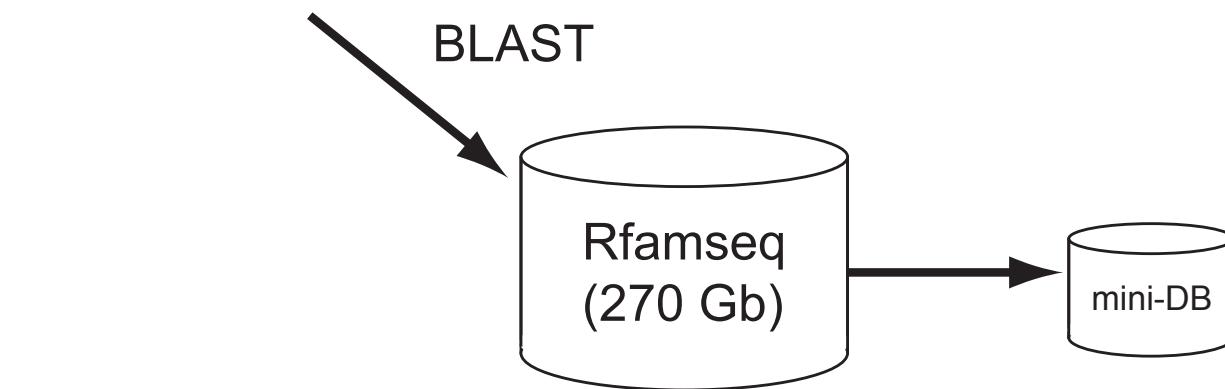
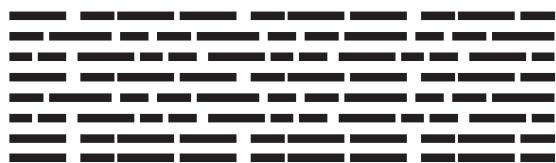
††E. P. Nawrocki. PhD Thesis: 2009, Washington University School of Medicine

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

Rfam used BLAST filters from 2003 to 2012

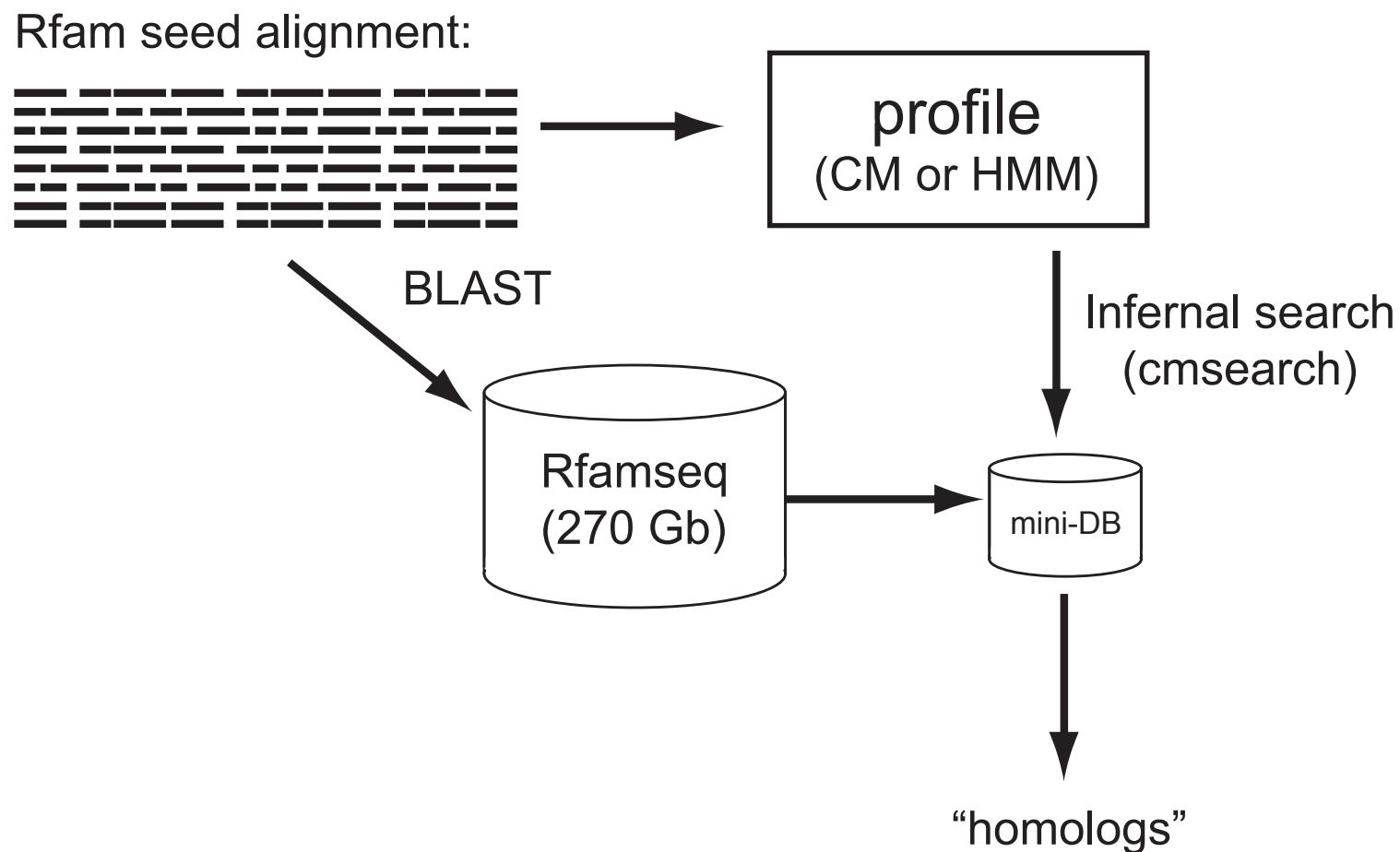
- Rfam includes > 2000 RNA families, each represented by an alignment, CM and set of predicted homologs in a large database (Rfamseq).

Rfam seed alignment:



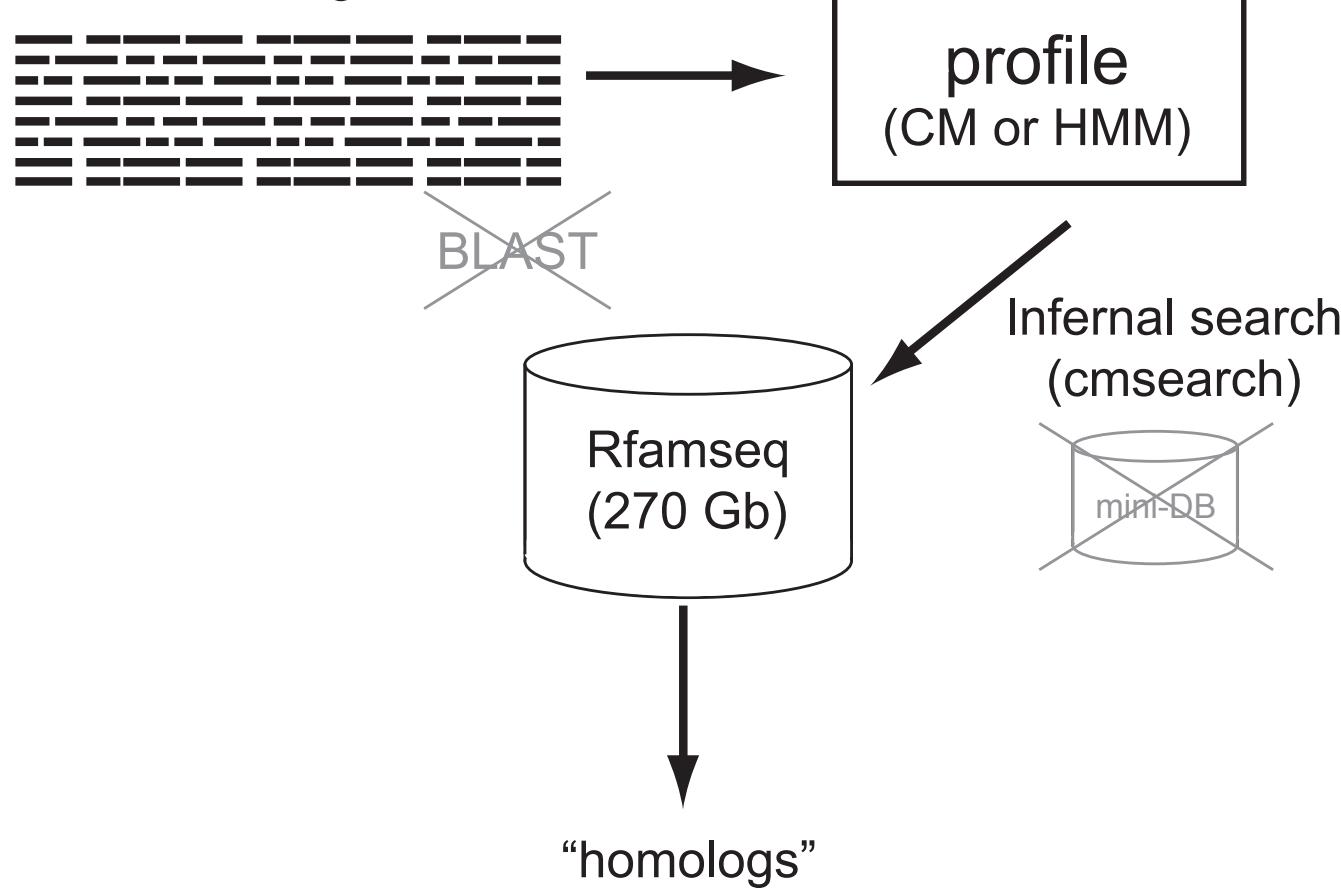
Rfam used BLAST filters from 2003 to 2012

- Rfam includes > 2000 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

Rfam seed alignment:



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.

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

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

The *cmsearch* program of Infernal 1.1 was used with Rfam 12.0 CM files and the following command-line options: `-noali -cut ga -rfam -nohmmonly -cpu 0`. Overlapping hits were removed such that no nucleotide was matched by more than one family by keeping the hit with the lower *E*-value in the case of overlaps (and higher bit score in the case of tying *E*-values). All searches were run as single execution threads on 3.0 GHz Intel Xeon processors. The *Homo sapiens*, *Sus scrofa*, *Drosophila melanogaster* and *Saccharomyces cerevisiae* genomes searched were obtained from Ensembl release 76 (<http://www.ensembl.org/>) (26) and the *Escherichia coli* (K12 substr MG1655), *Bacillus subtilis* (BSn5), *Methanocaldococcus jannaschii* (DSM 2661), *Aquifex aeolicus* (VF5) and *Borrelia burgdorferi* (CA-11 2A) genomes were obtained from release 23 of Ensembl Genomes (<http://ensemblgenomes.org/>) (27) for all of those the actual sequence file searched was downloaded via FTP and suffixed with .dna.toplevel.fa.gz. The HIV genome used is ENA accession AJ291720 and the four metagenomic samples were downloaded from the EBI Metagenomics Portal (<https://www.ebi.ac.uk/metagenomics/>) (29), and can be accessed by the sample accession listed in the table. ‘CPU time’ and ‘Mb/hour’ columns are rounded to two significant digits.

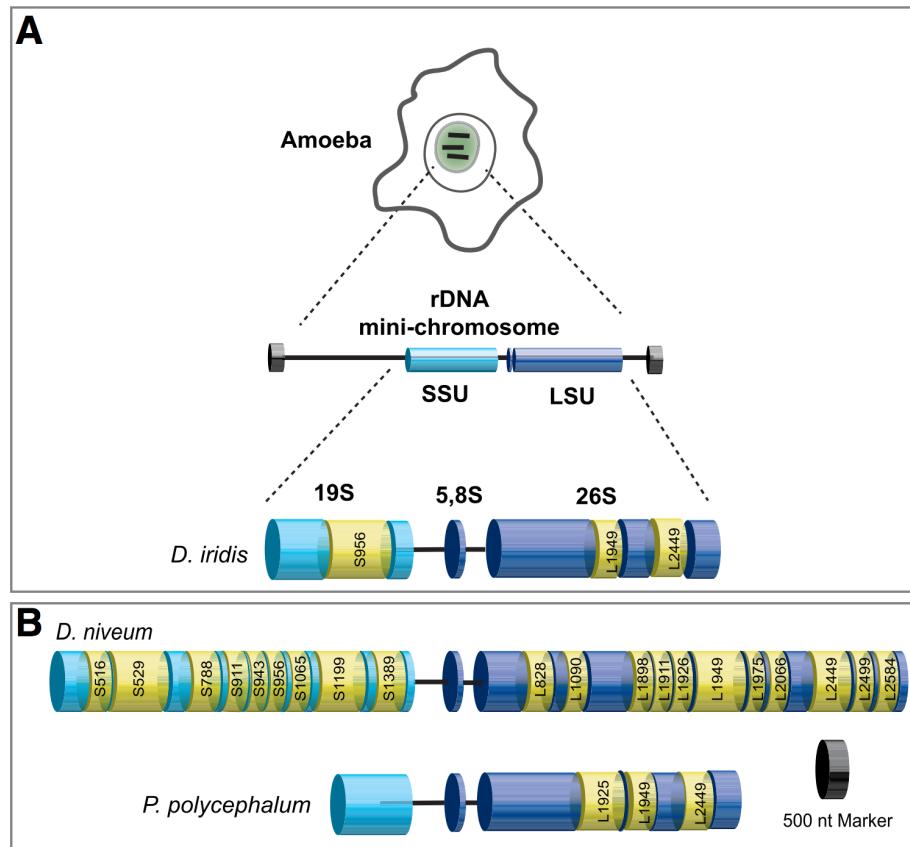
Infernal 1.1 finds 11,000 new group I intron candidates

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

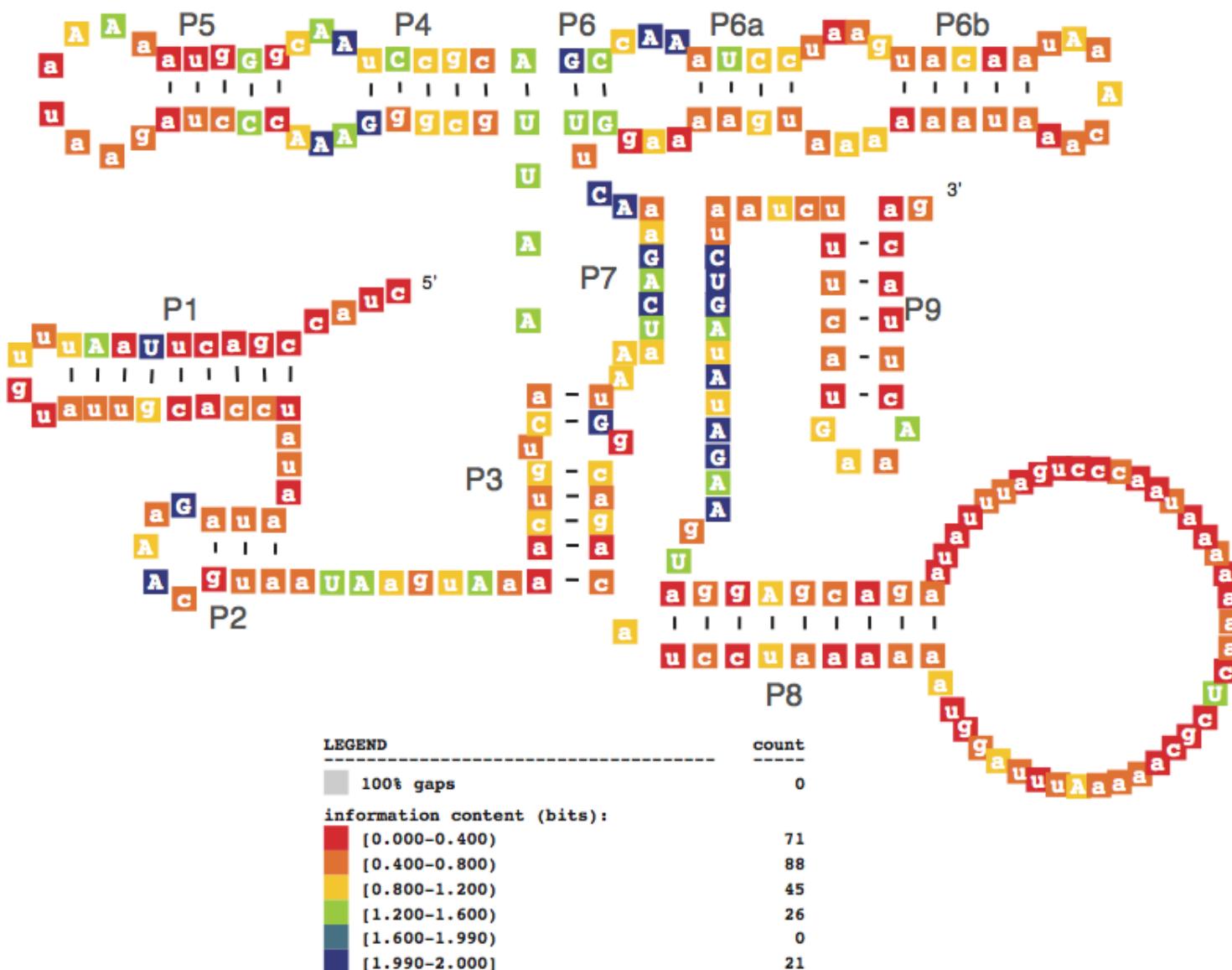
Accession	Family ID	Length (nt)	#of seed seqs	Time new (h)	Time old (h)	Time (old/new)	New total hits	Old total hits	New unique hits	Old unique hits
Top five families										
RF00028	Intron-gpI	251	12	125.0	357.2	2.8	71 433	60 264	11 175	1
RF00026	U6	104	188	31.2	181.1	5.8	66 517	62 174	4367	14
RF00003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

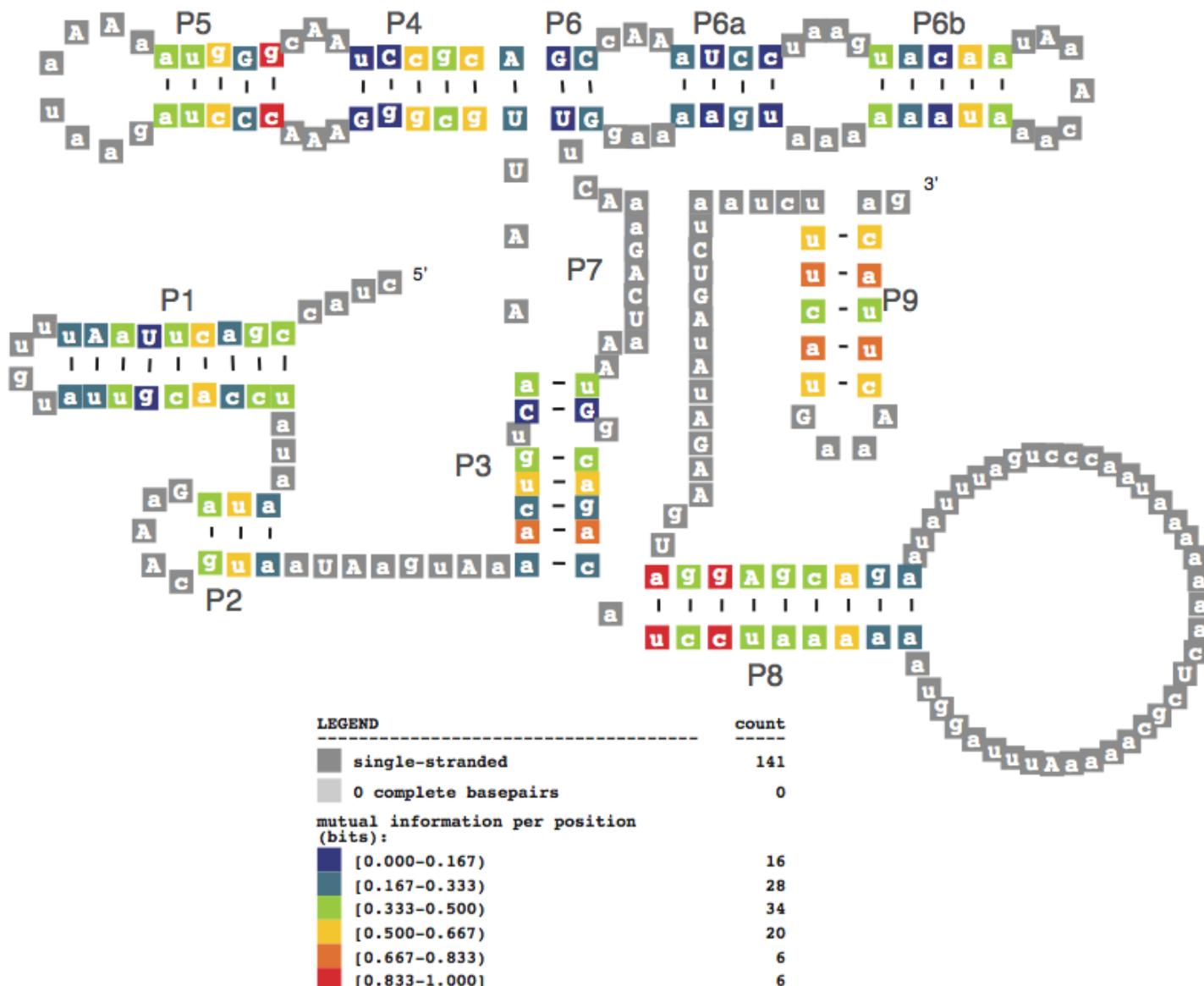


model	#pos	#bps	#seqs	description
Group I Intron	251	55	12	information content per position



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.

GISSD*: Group I Intron Sequence and Structure Database



GISSD Group I Intron Sequence and Structure Database Wuhan University, China

Home Search Sequence Structure Alignment Distribution gIRfam Submission Help

```
<[GISSD]> esl-alistat -1 GISSD-plus-RF00028.sto
```

#	idx	name	format	nseq	alen	nres	small	large	avlen	%id
#	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
1	IA1		Stockholm	76	1471	44356	216	1261	583.6	45
2	IA2		Stockholm	15	503	4152	228	313	276.8	38
3	IA3		Stockholm	56	602	15808	205	374	282.3	46
4	IB1		Stockholm	42	503	12515	213	407	298.0	72
5	IB2		Stockholm	18	502	4360	163	320	242.2	39
6	IB3		Stockholm	7	373	1944	235	345	277.7	52
7	IB4		Stockholm	89	662	25124	203	392	282.3	44
8	IC1		Stockholm	837	1408	364902	251	615	436.0	39
9	IC2		Stockholm	32	452	10247	279	348	320.2	66
10	IC3		Stockholm	328	1584	83888	196	675	255.8	67
11	ID		Stockholm	17	386	4123	197	328	242.5	53
12	IE1		Stockholm	38	714	13763	308	479	362.2	60
13	IE2		Stockholm	56	710	22396	265	480	399.9	54
14	IE3		Stockholm	110	706	44653	252	481	405.9	57
15	RF00028		Stockholm	12	891	4378	206	488	364.8	34

Searching Rfamseq with GISSD models

type	# RF00028 seed seqs	# hits total	# hits common	# hits unique
IA1	3	814	385	425
IA2	1	1722	823	899
IA3		958	401	557
IB1		3949	1033	2916
IB2		1861	467	1394
IB3		479	136	343
IB4	1	5717	2400	3317
IC1	3	8475	5385	3090
IC2		4870	3858	1012
IC3	4	72692	66033	6659
ID		572	0	572
IE1		1305	10	1295
IE2		1377	8	1369
IE3		1379	1	1378
total	12	106170*	80940*	16842
RF00028	-	71421	71421	-

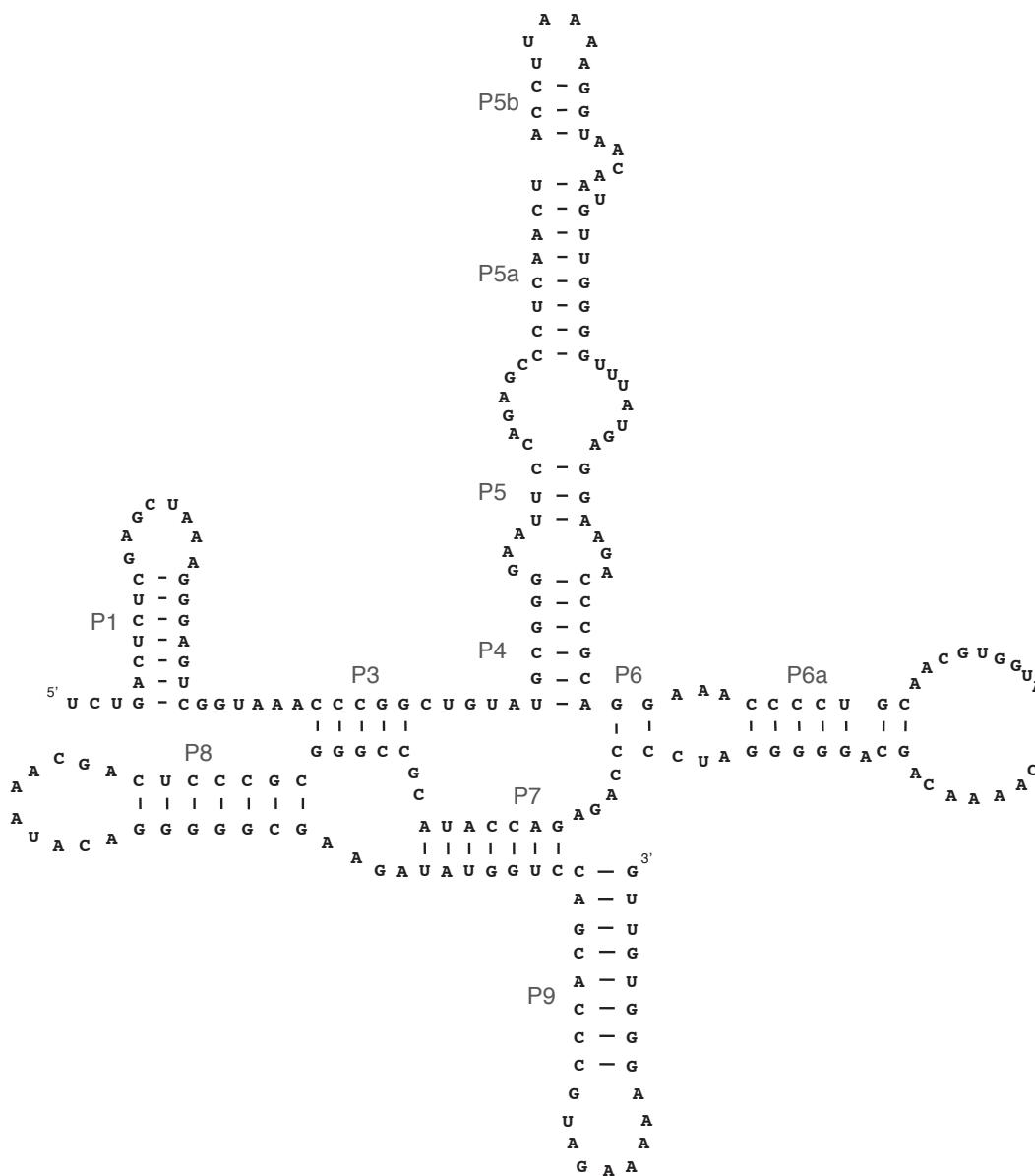
* contains overlaps

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	- +

Annotation of thaumarchaeota candidate

19000	18000	17000	16000	15000	14000	13000	
-----	-----	-----	-----	-----	-----	-----	
----- LSU ----- ---- HEG ----- ----- LSU ----- IB4 ----- HEG ----- IB4 -- LSU -- --- HEG --- -- LSU ----							
18821*	17588	16859	16078 15952*	15045* 14946*	14419	13819	13344
+5	-38		-10 0	0 0	0	0	

Thaumarchaeota group I structure



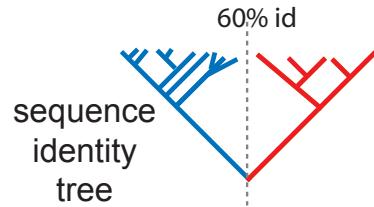
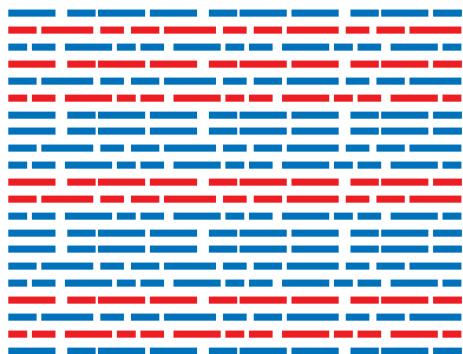
Acknowledgements

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

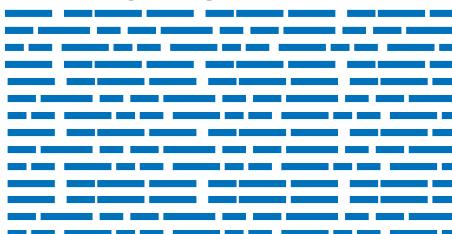
Is the added complexity worth it?

RMARK: a challenging internal RNA homology search benchmark

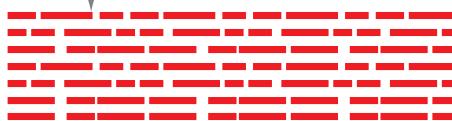
Rfam seed alignment:



training alignment

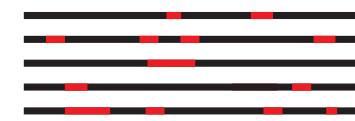


no train/test sequence pair is > 60% identical



test sequences

embed in
pseudo-genome

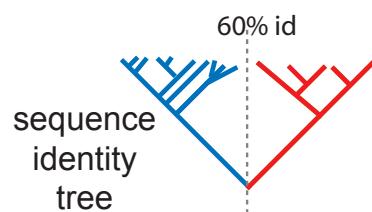
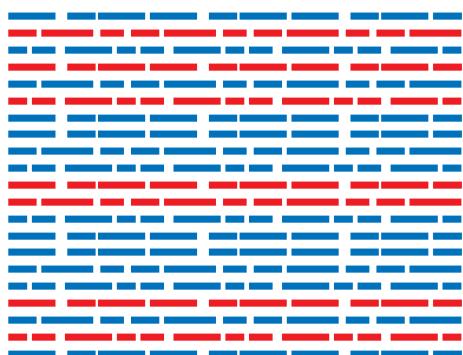


10 1Mb sequences
with 780 embedded
test seqs from 106 families

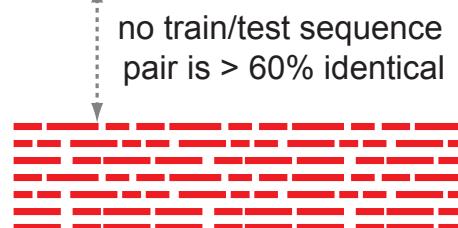
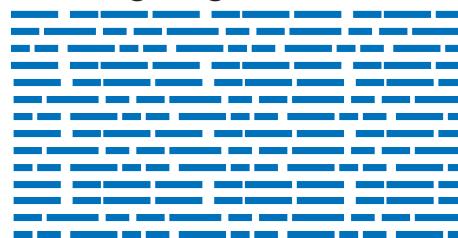
Is the added complexity worth it?

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 -
...