Modeling structural RNA families with Infernal

Eric Nawrocki

Alejandro Schäffer's group

National Center for Biotechnology Information
National Institutes of Health

Howard Hughes Medical Institute
Janelia Research Campus

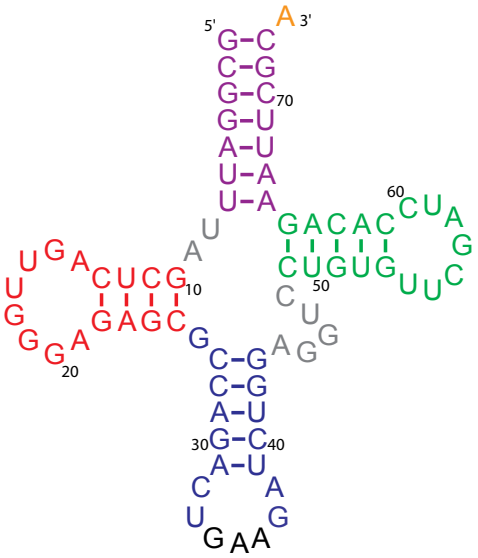
Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



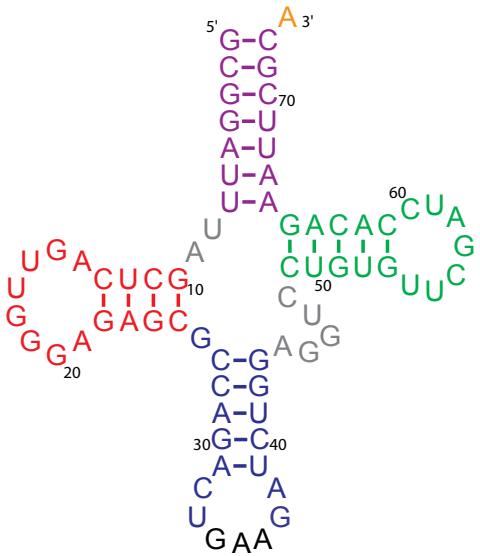
3-dimensional structure

Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence	Secondary structure	3-dimensional structure

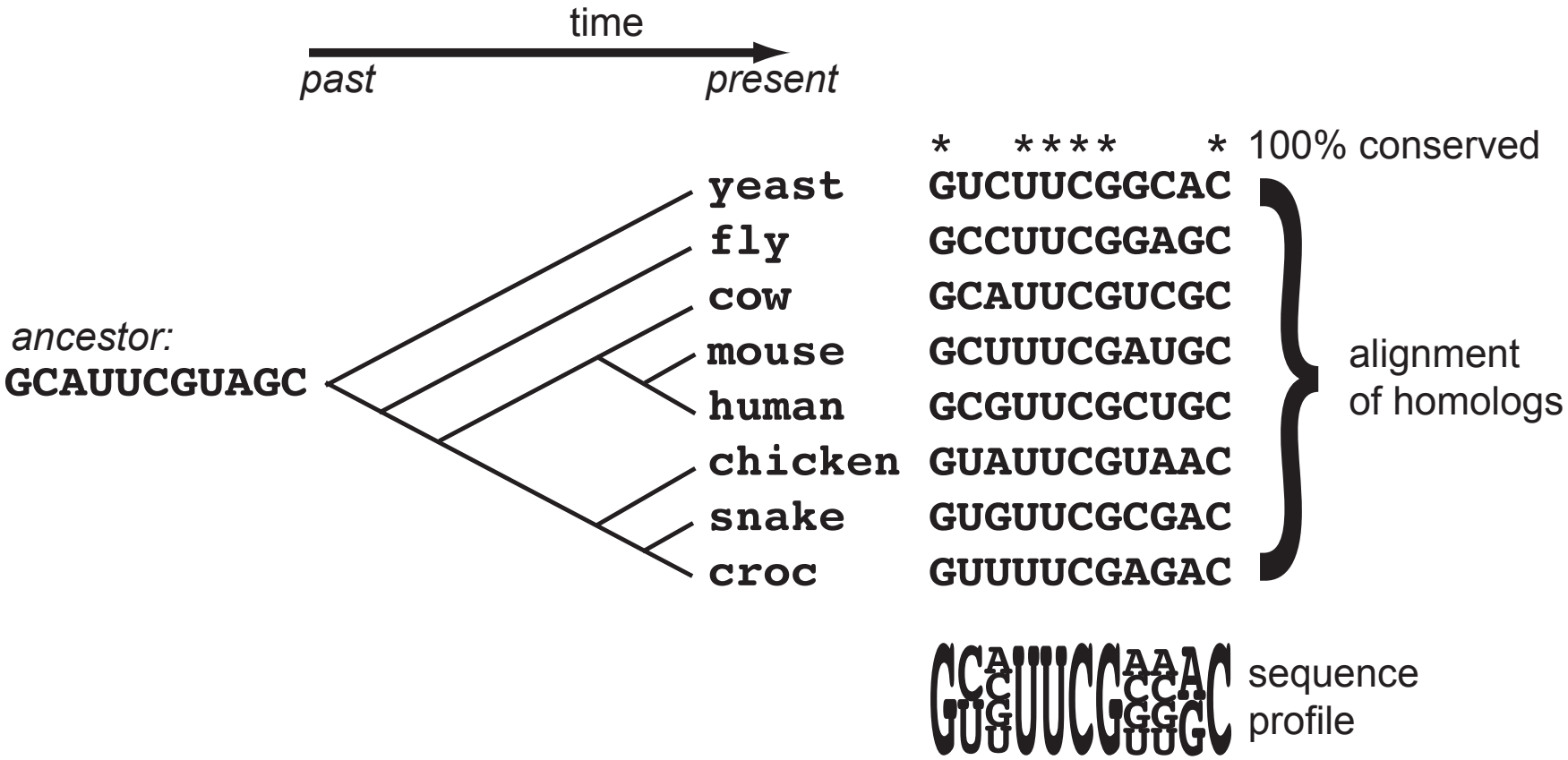
GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA



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

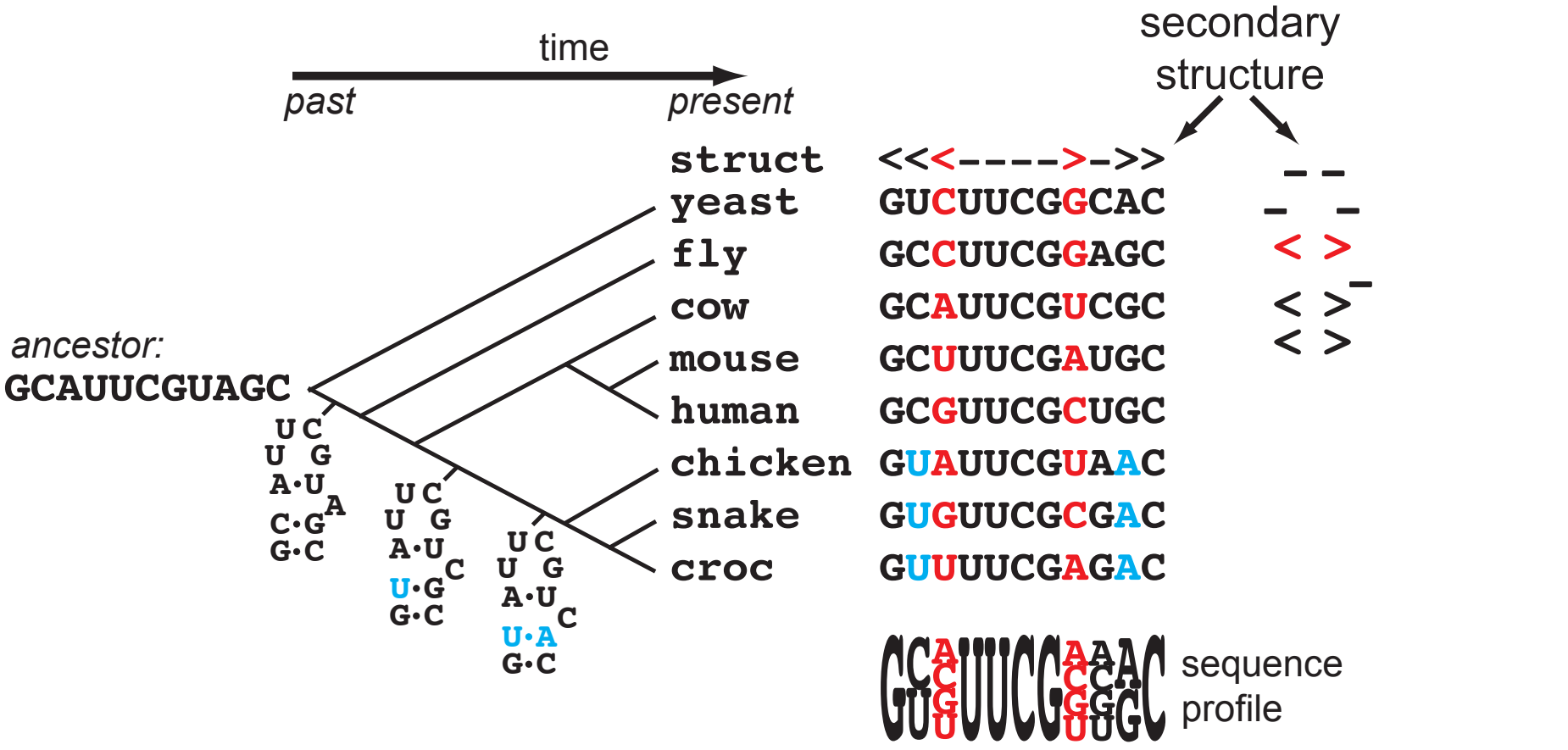
Sequence conservation provides information for homology searches

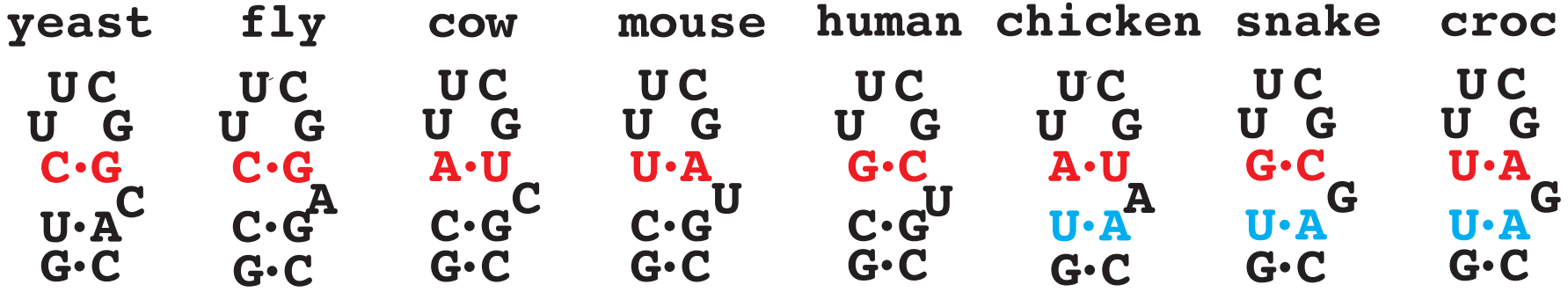
Conservation levels vary across alignment columns.



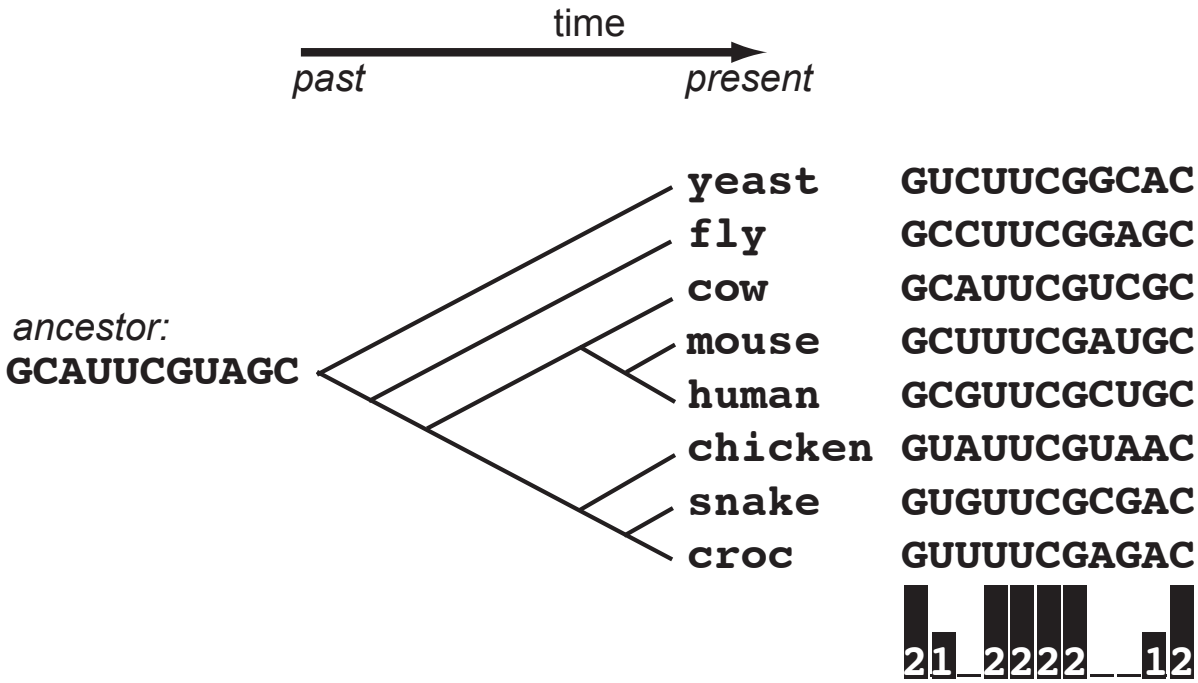
Structure conservation provides additional information

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





Amount of information in a profile can be measured in bits



Each column is independent

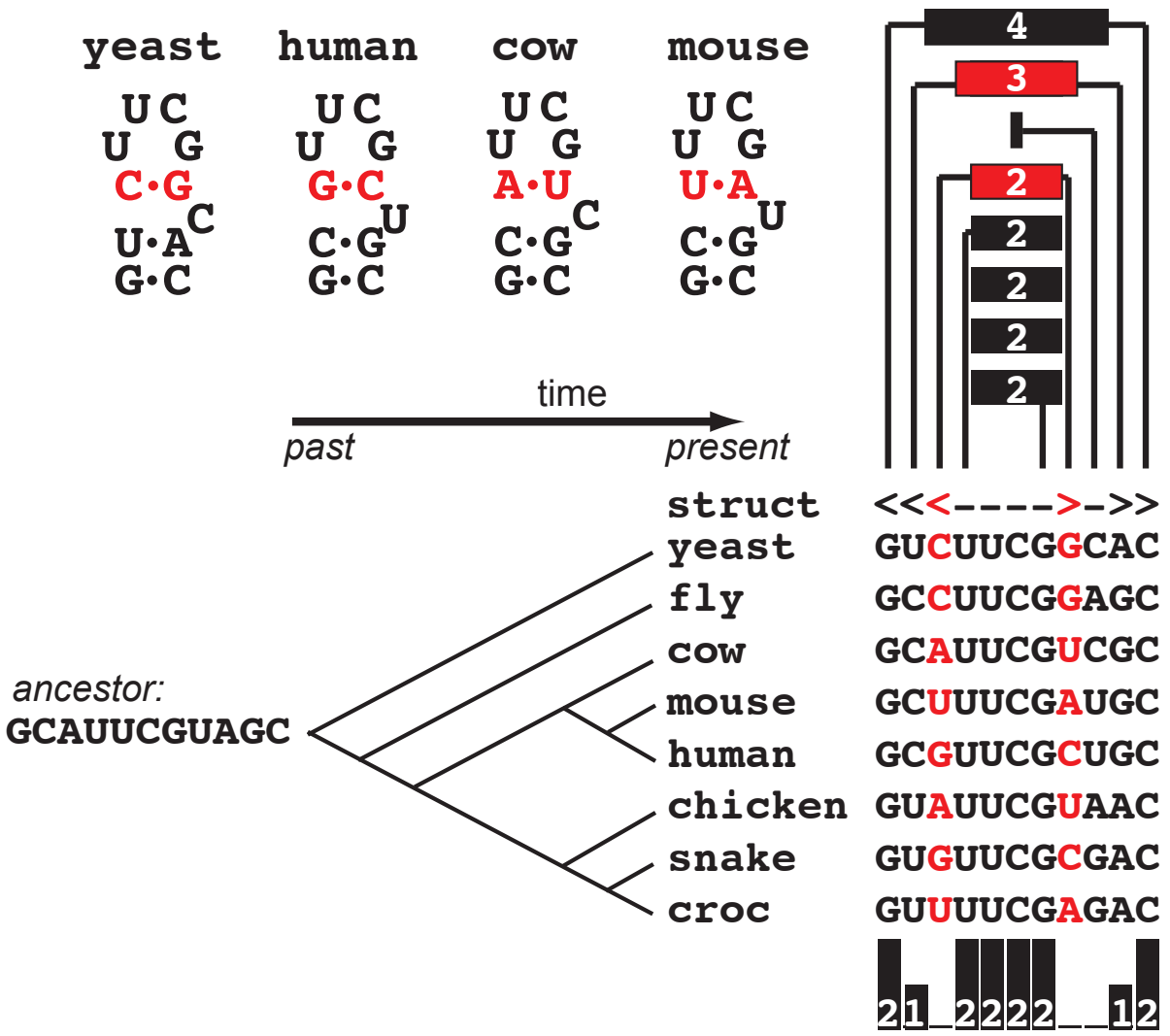
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ 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

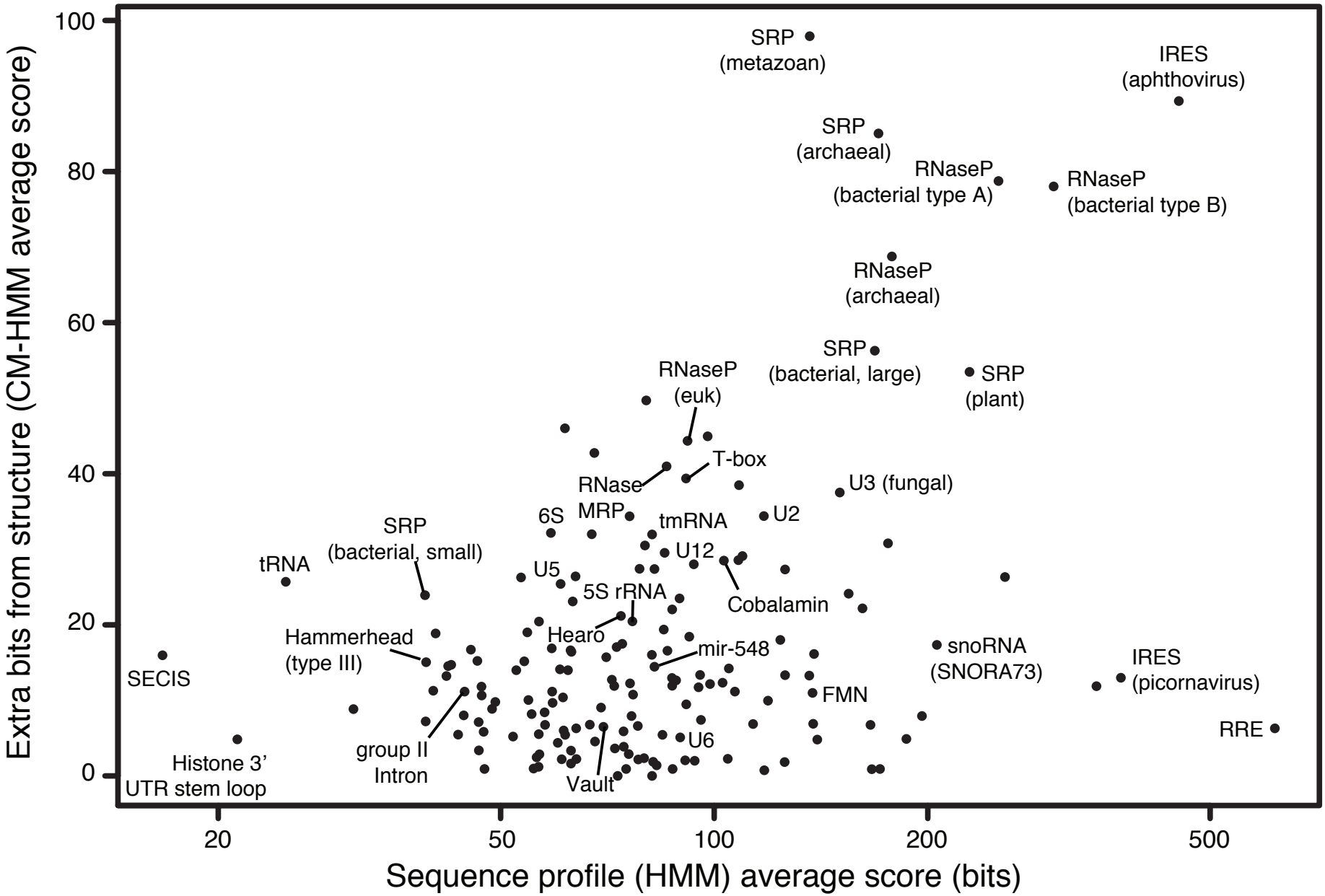
$\log_2 (16/4) = 2$ bits

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 (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

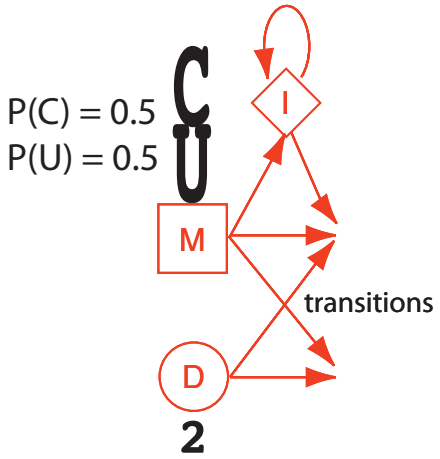
yeast	GUCUUCGGCAC
fly	GCCUUCGGAGC
cow	GCAUUCGUCGC
mouse	GCUUUCGAUGC
human	GCGUUCGCUGC
chicken	GUAUUCGUAAC
snake	GUGUUCGCGAC
croc	GUUUUCGAGAC

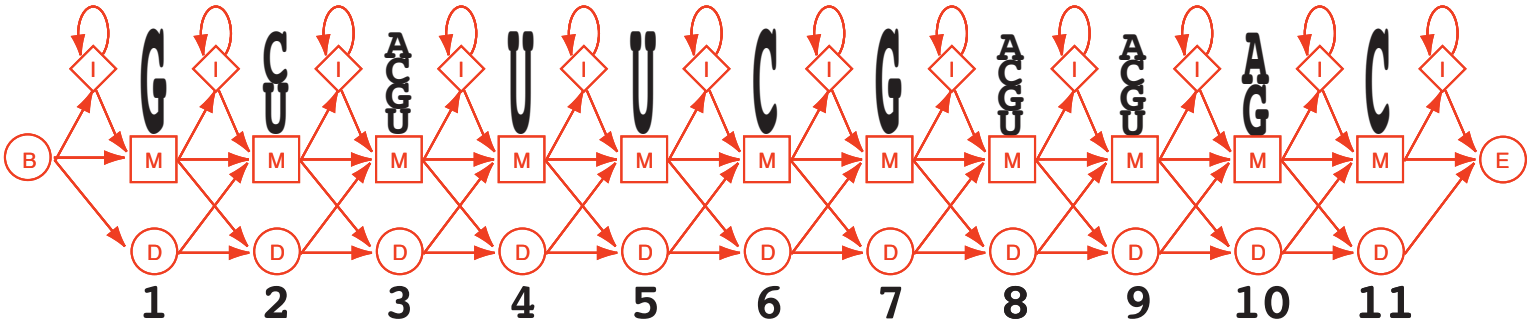
One HMM node per alignment column

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

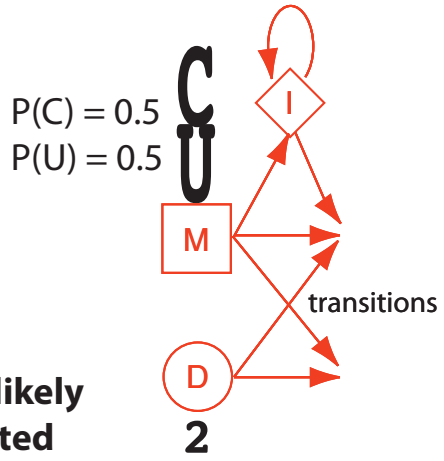
One HMM node per alignment column

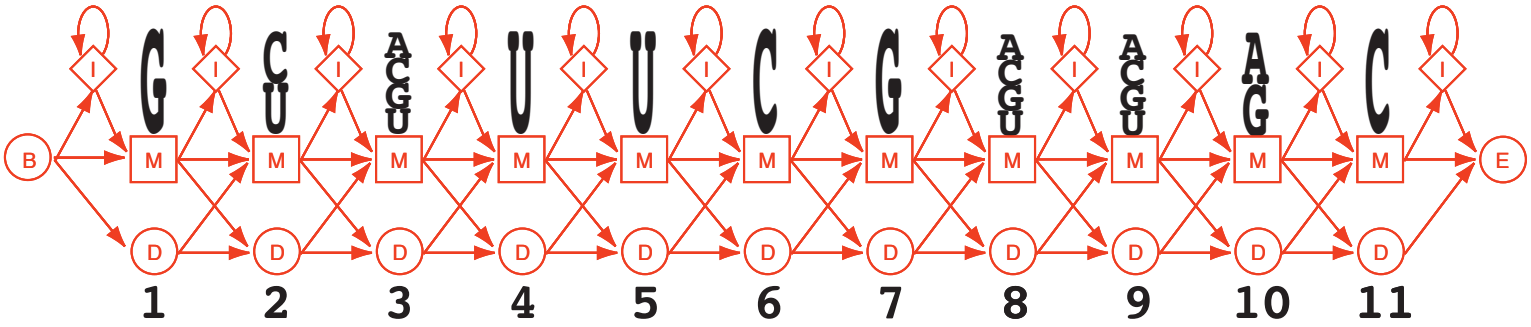
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

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

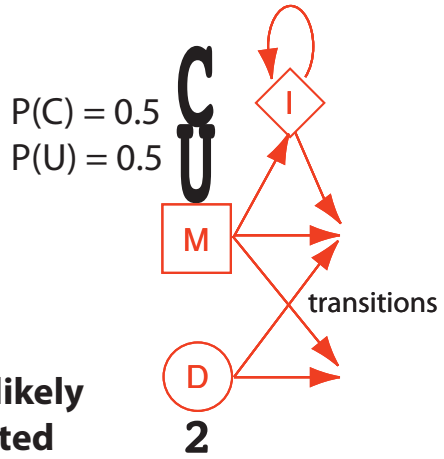
One HMM node per alignment column

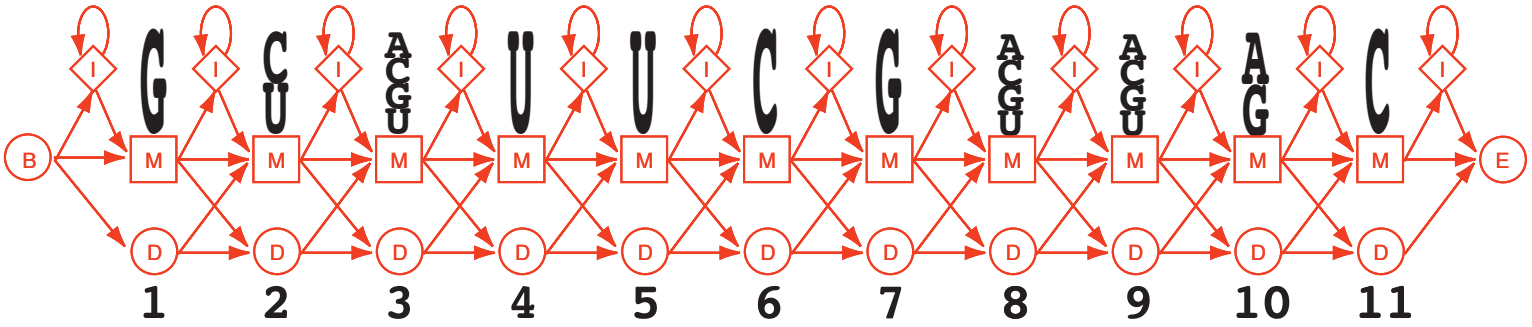
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

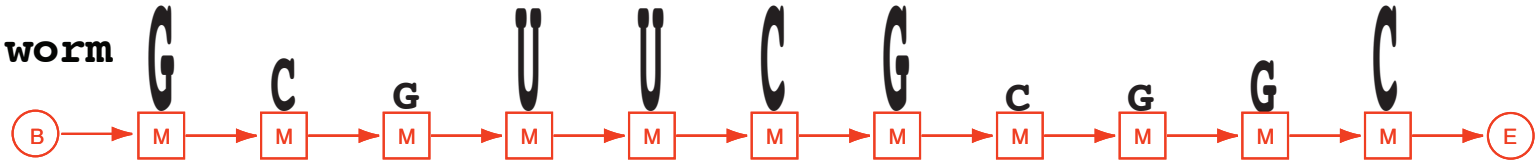
HMMs generate homologous sequences.

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
```

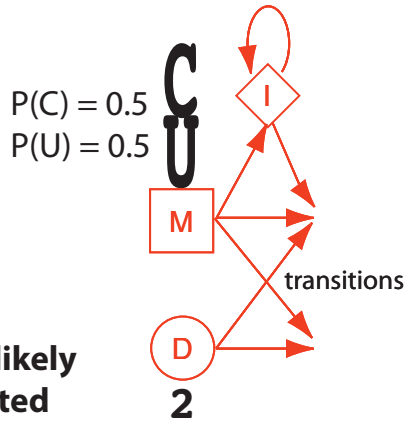
One HMM node per alignment column

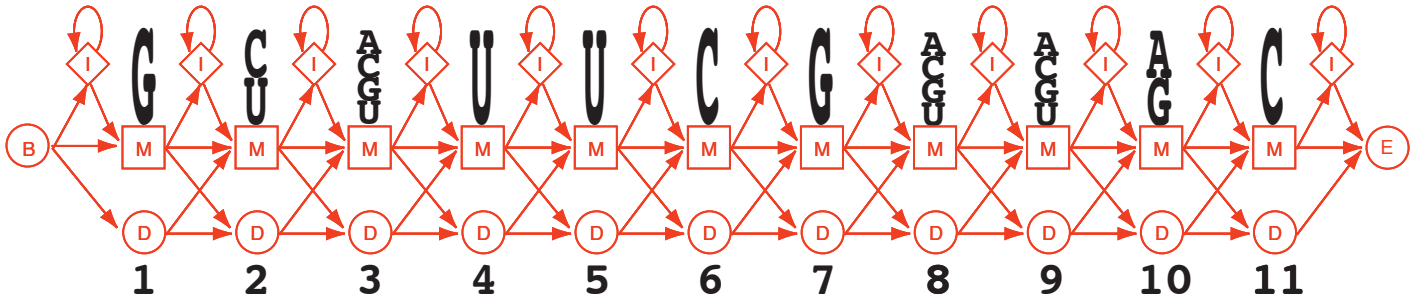
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

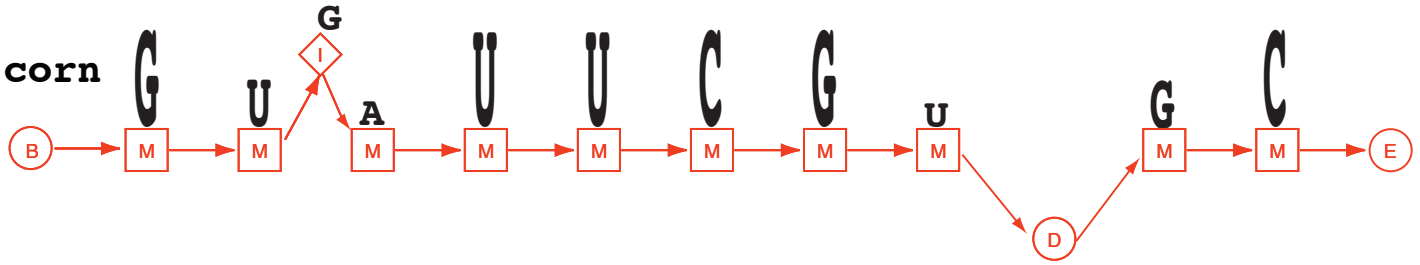
HMMs generate homologous sequences.

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

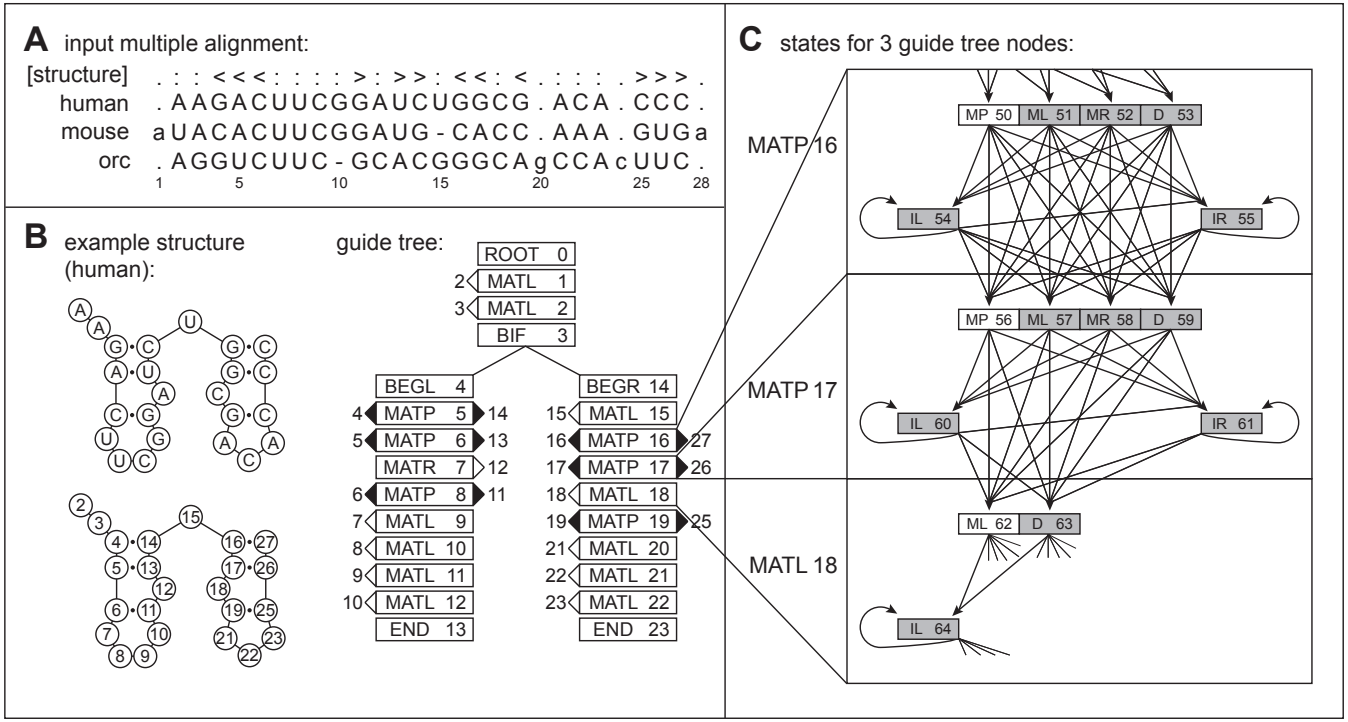
Node for column 2:

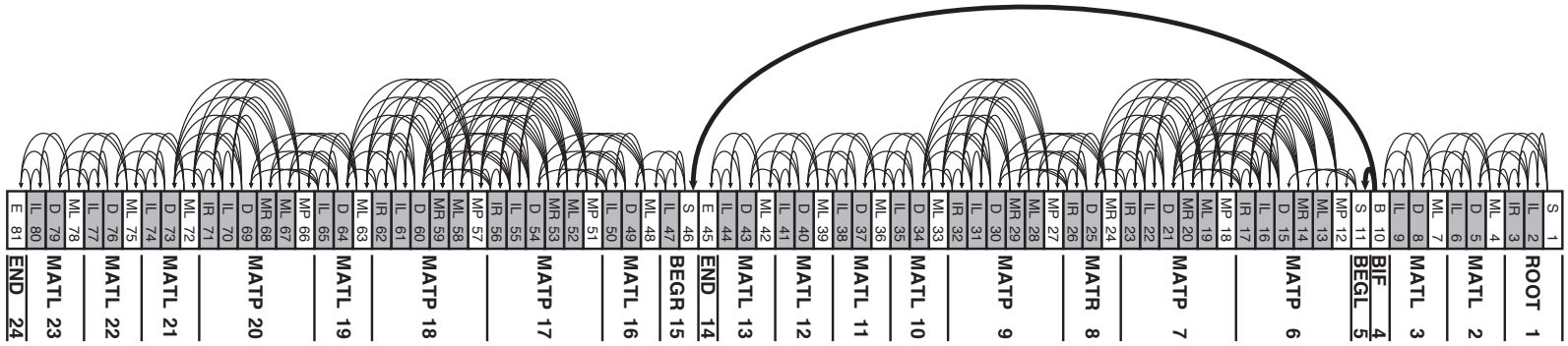




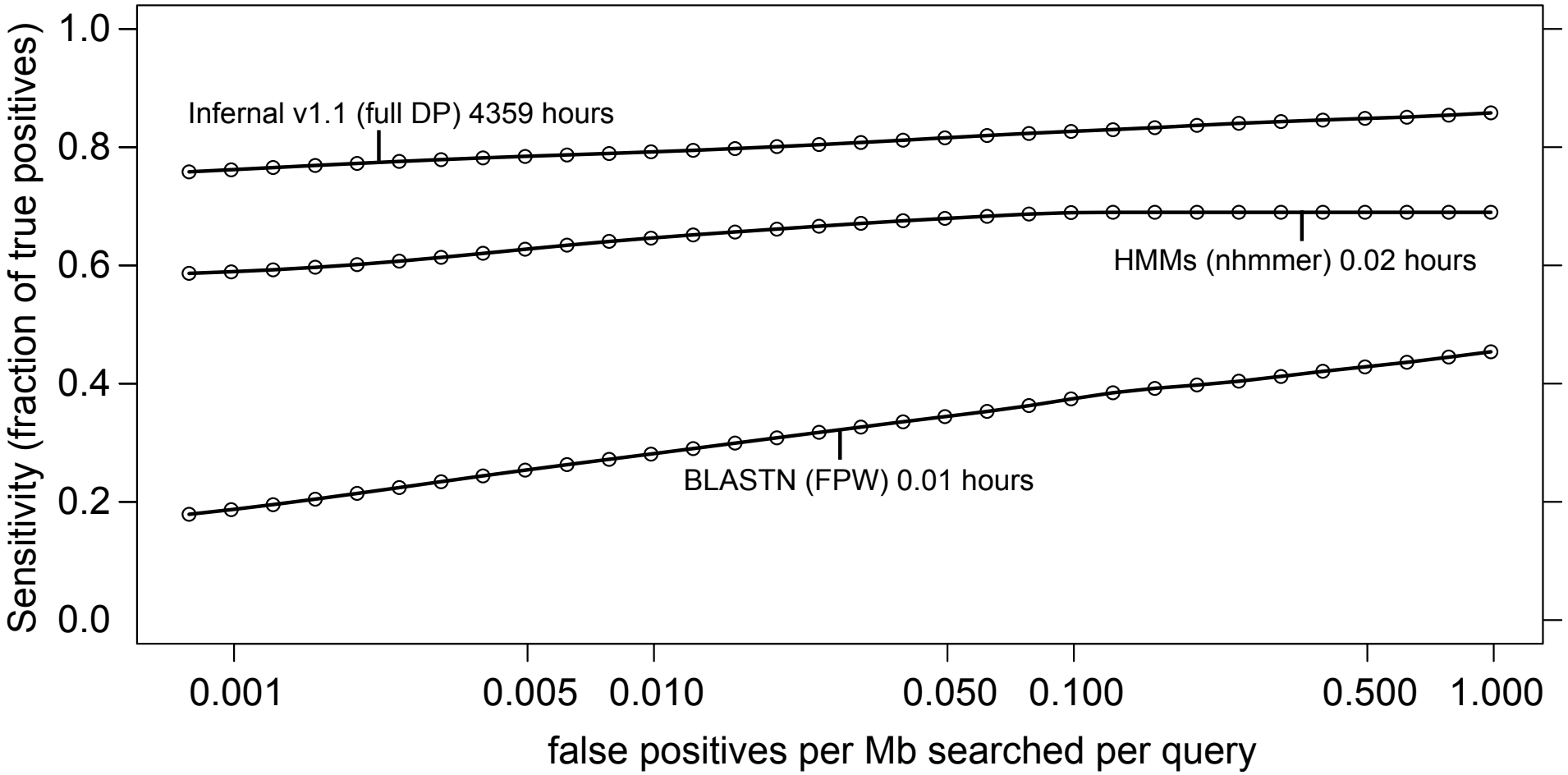


Covariance models (CMs) are built from structure-annotated alignments





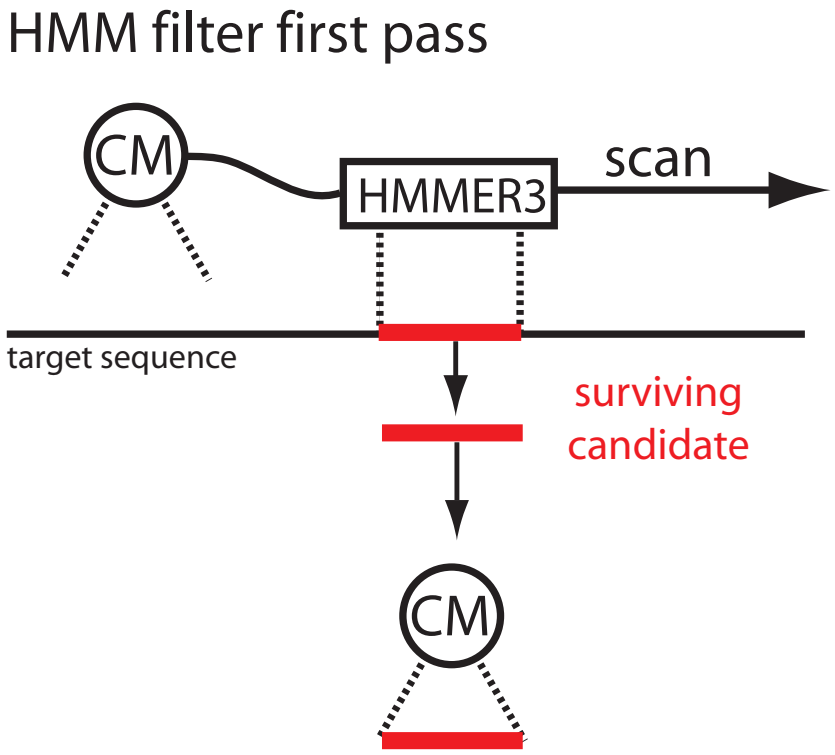
Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)



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

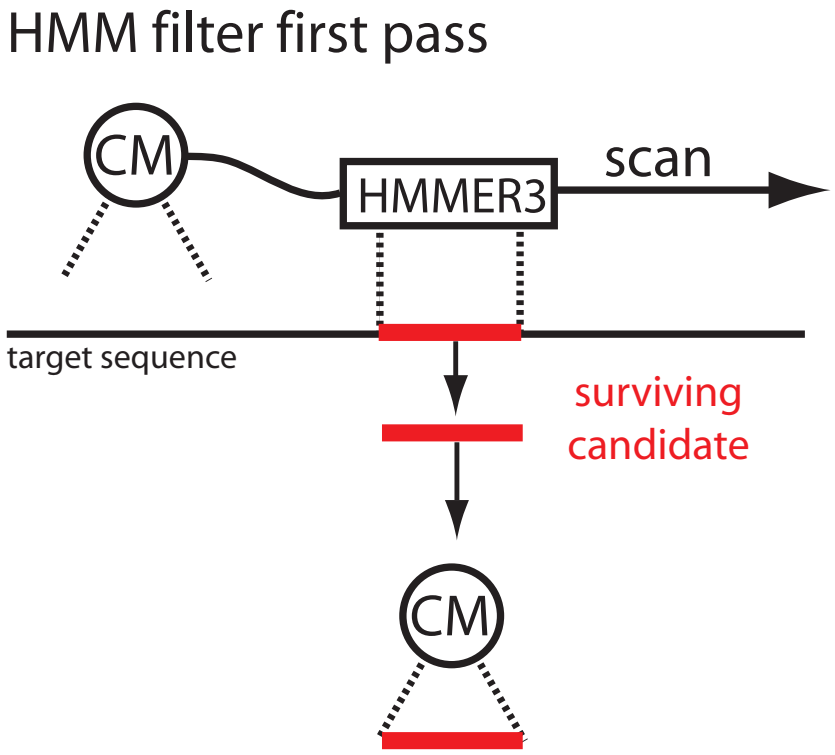
*Freyhult EK, Bollback JP, Gardner PP. Genome Res. 2007 17: 117-125.

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*

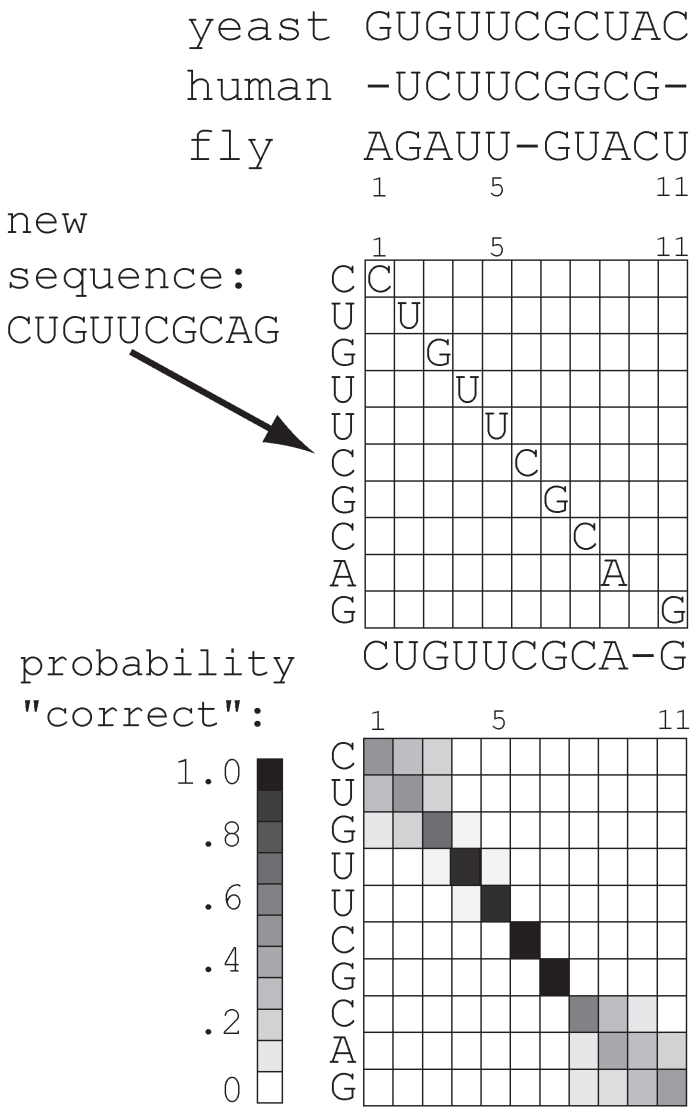


- Even if we filter out 99% of the database (for up to 100X acceleration), searches will still be too slow.

- CM step needs to be accelerated.

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

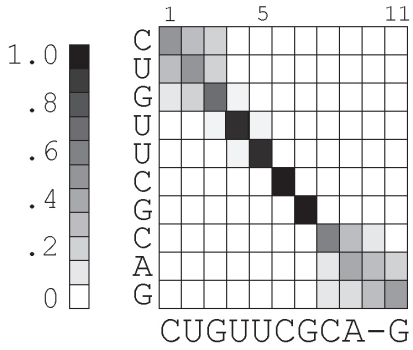
Accelerating CM alignment step 1: HMM posterior decoding to get confidence estimates

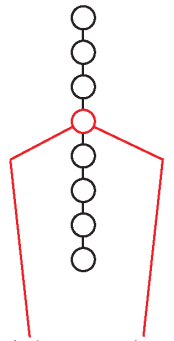


Accelerating CM alignment step 2: use HMM alignment confidence to constrain CM alignment*

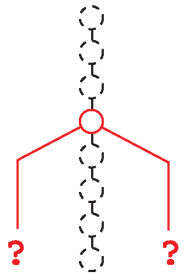
HMMs – Each column of seed alignment corresponds to a column of matrix.

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





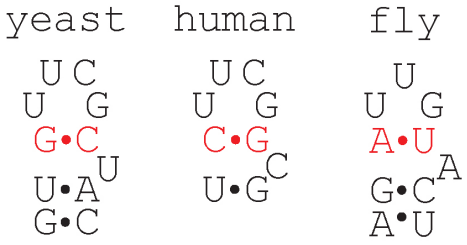
```
struct <<<<----->->>
yeast  GUGUUCGCUAC
human  -UCUUCGGCG-
fly    AGAUU-GUACU
```



CUGUUCGCAG

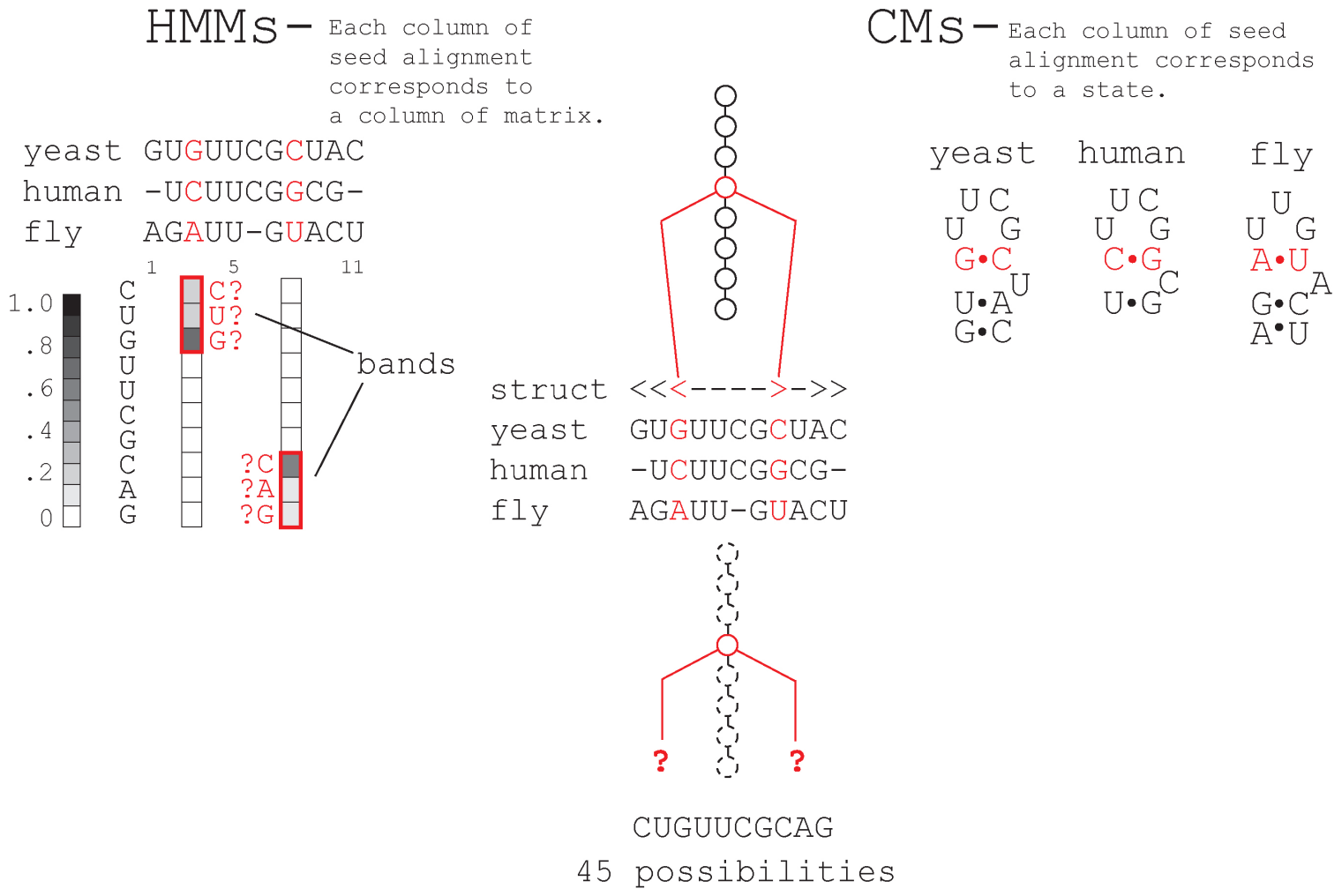
45 possibilities

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



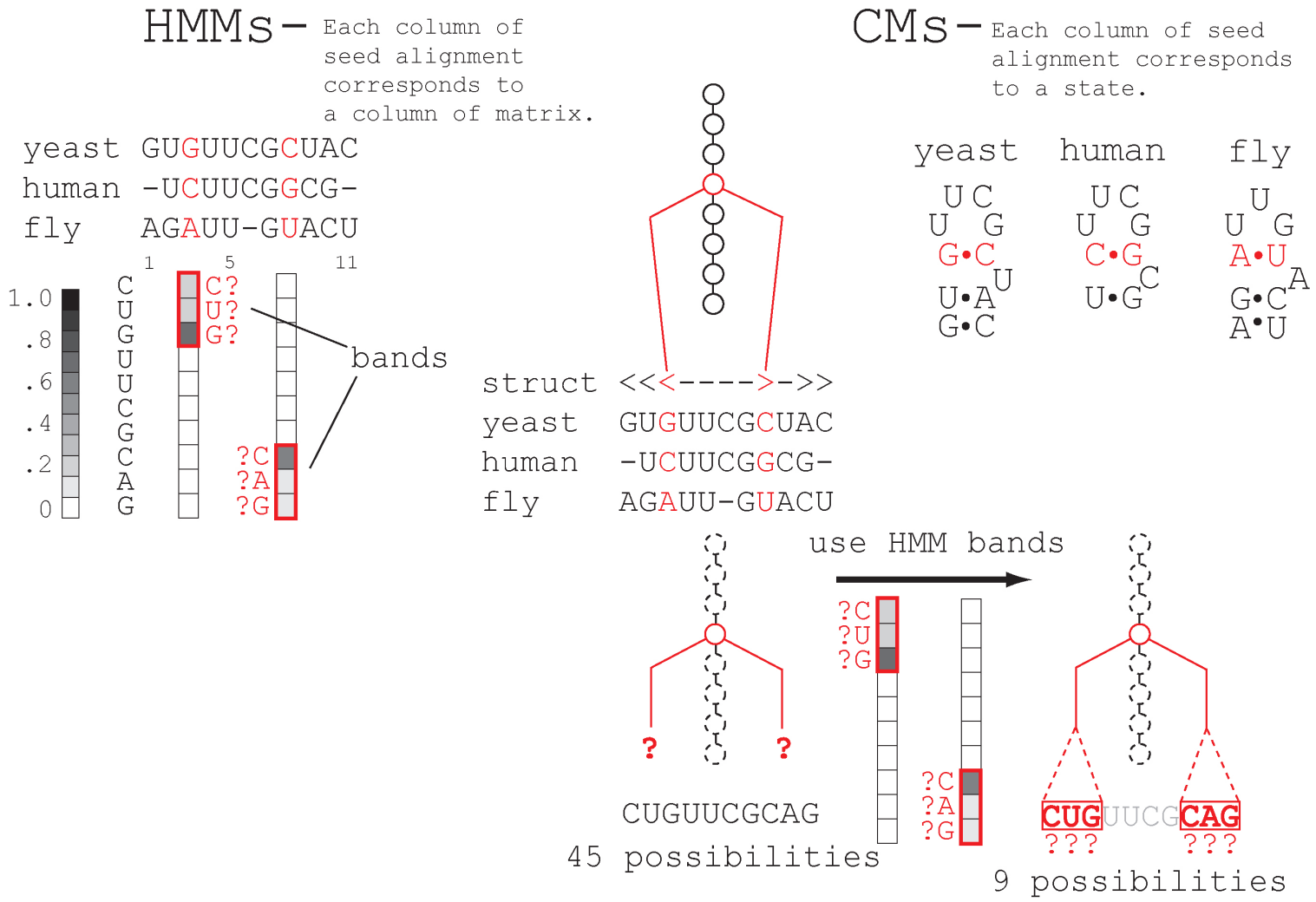
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

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



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

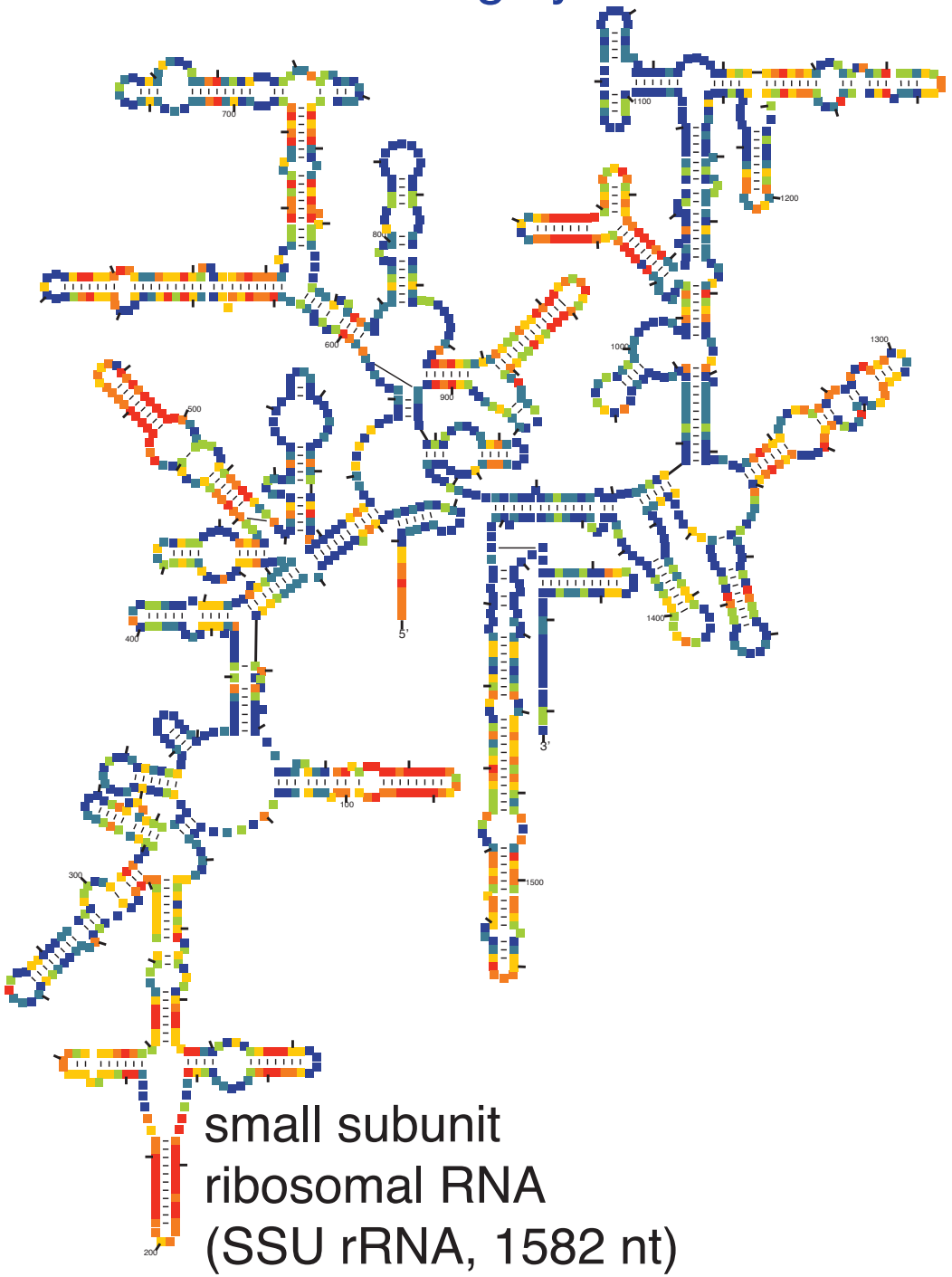
Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*



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

Sequence conservation per position

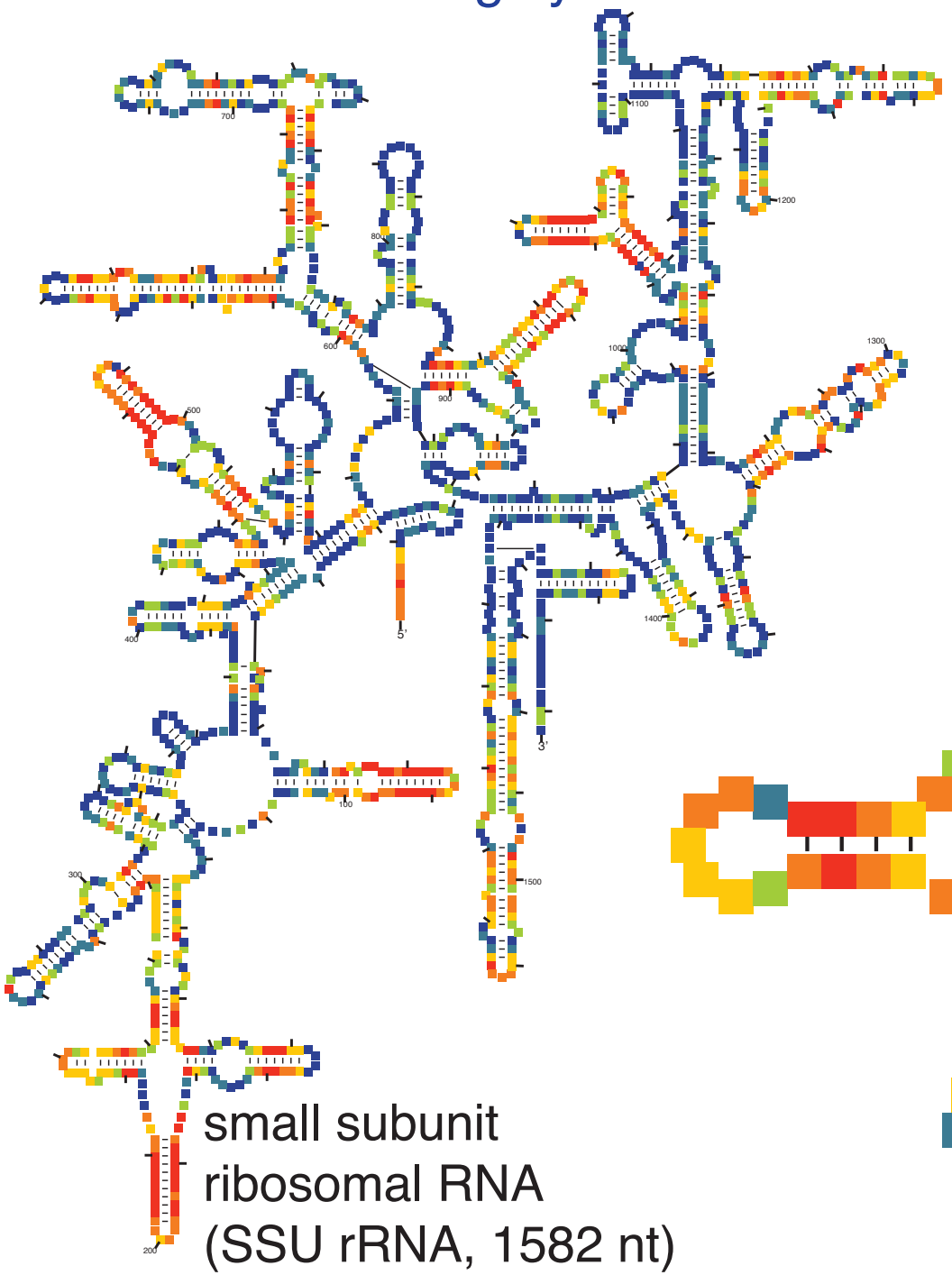
blue:highly conserved ..... red: highly variable



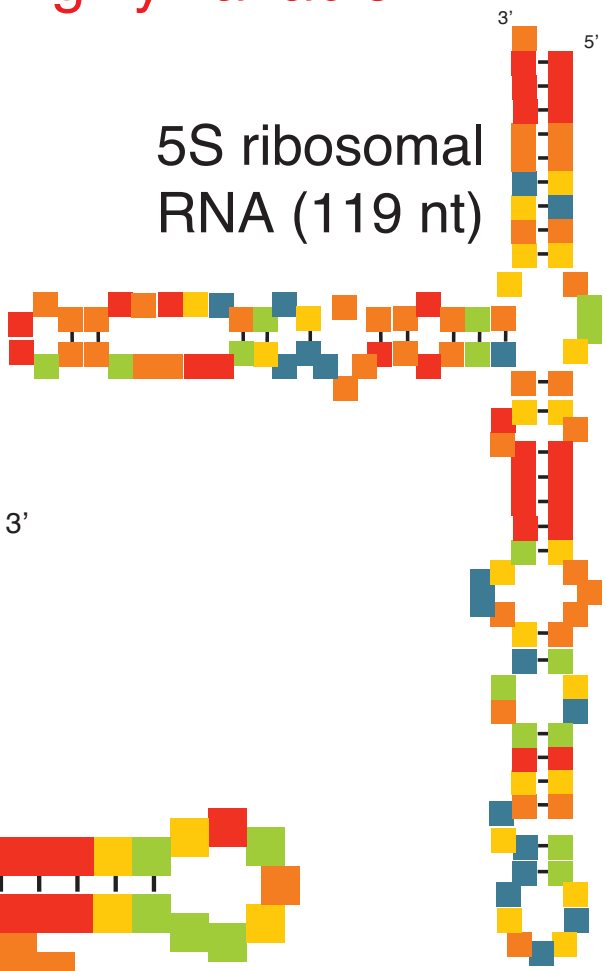
small subunit
ribosomal RNA
(SSU rRNA, 1582 nt)

Sequence conservation per position

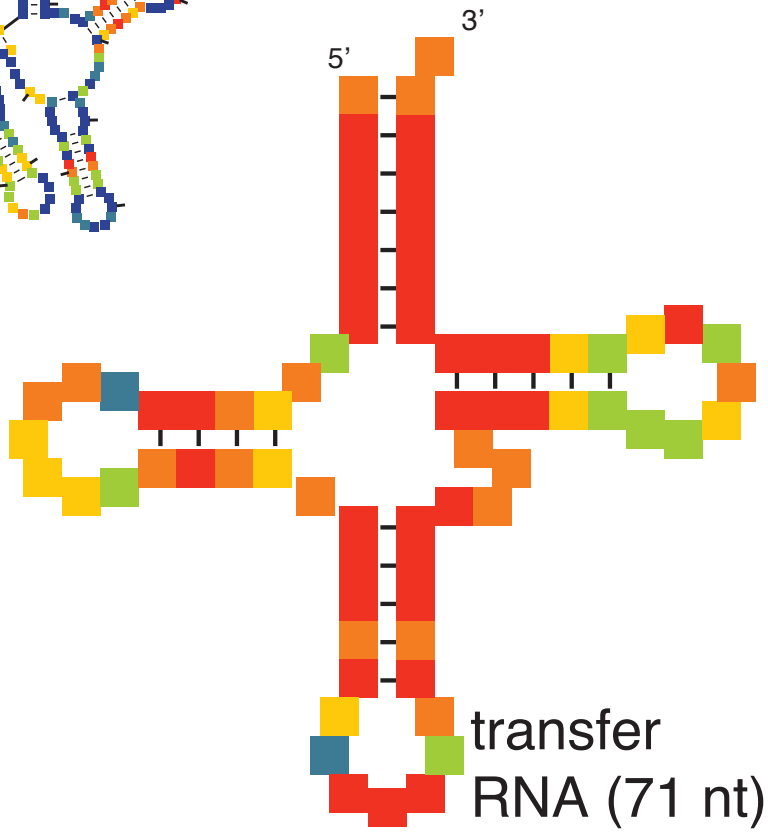
blue:highly conserved ..... red: highly variable



small subunit ribosomal RNA (SSU rRNA, 1582 nt)

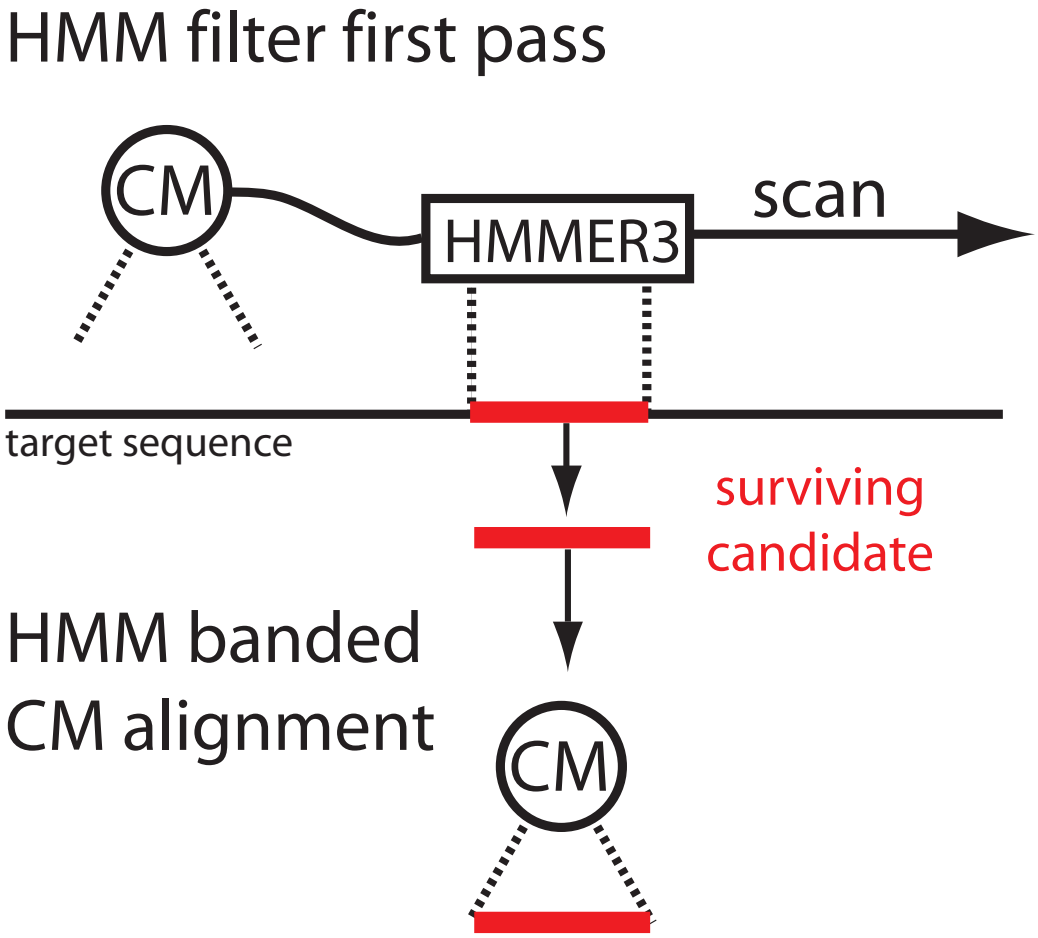


5S ribosomal RNA (119 nt)

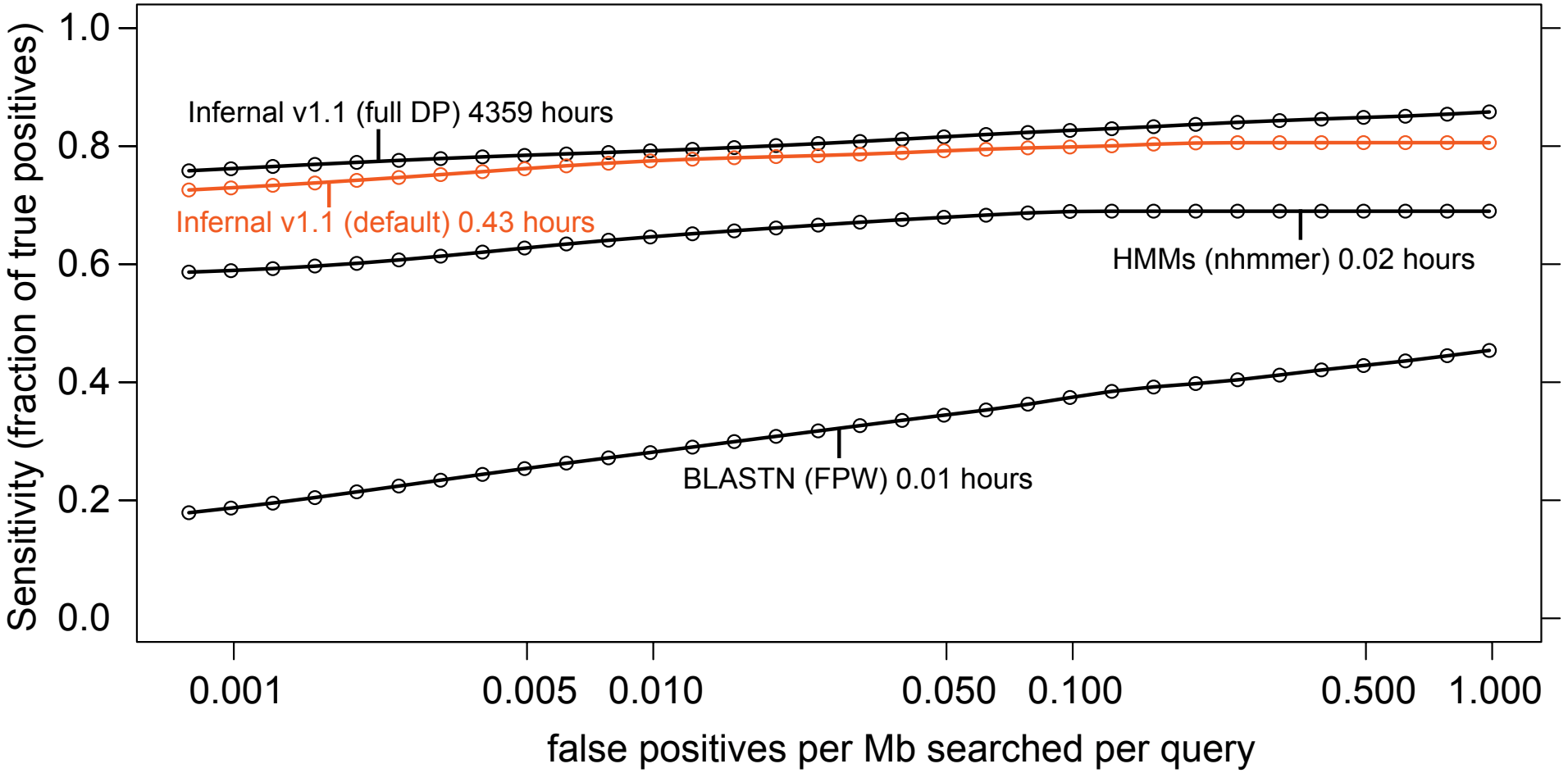


transfer RNA (71 nt)

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.

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.

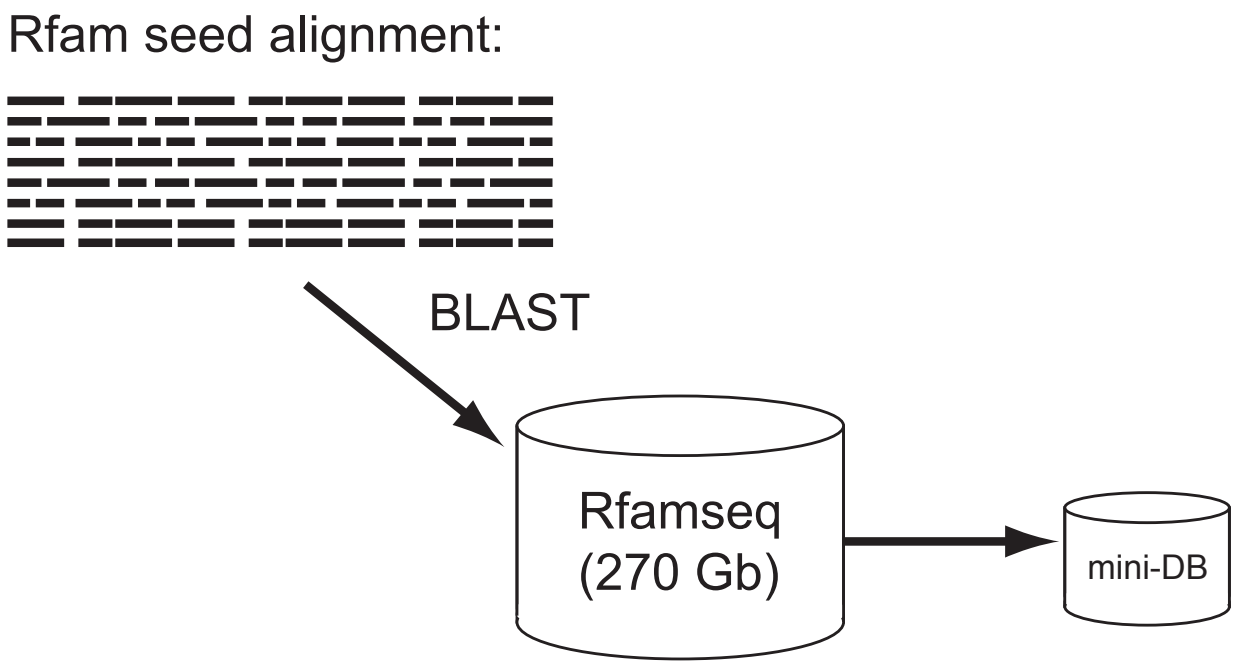
[**]T. M. Lowe, S. R. Eddy. NAR, 25:955-964, 1997.

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

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

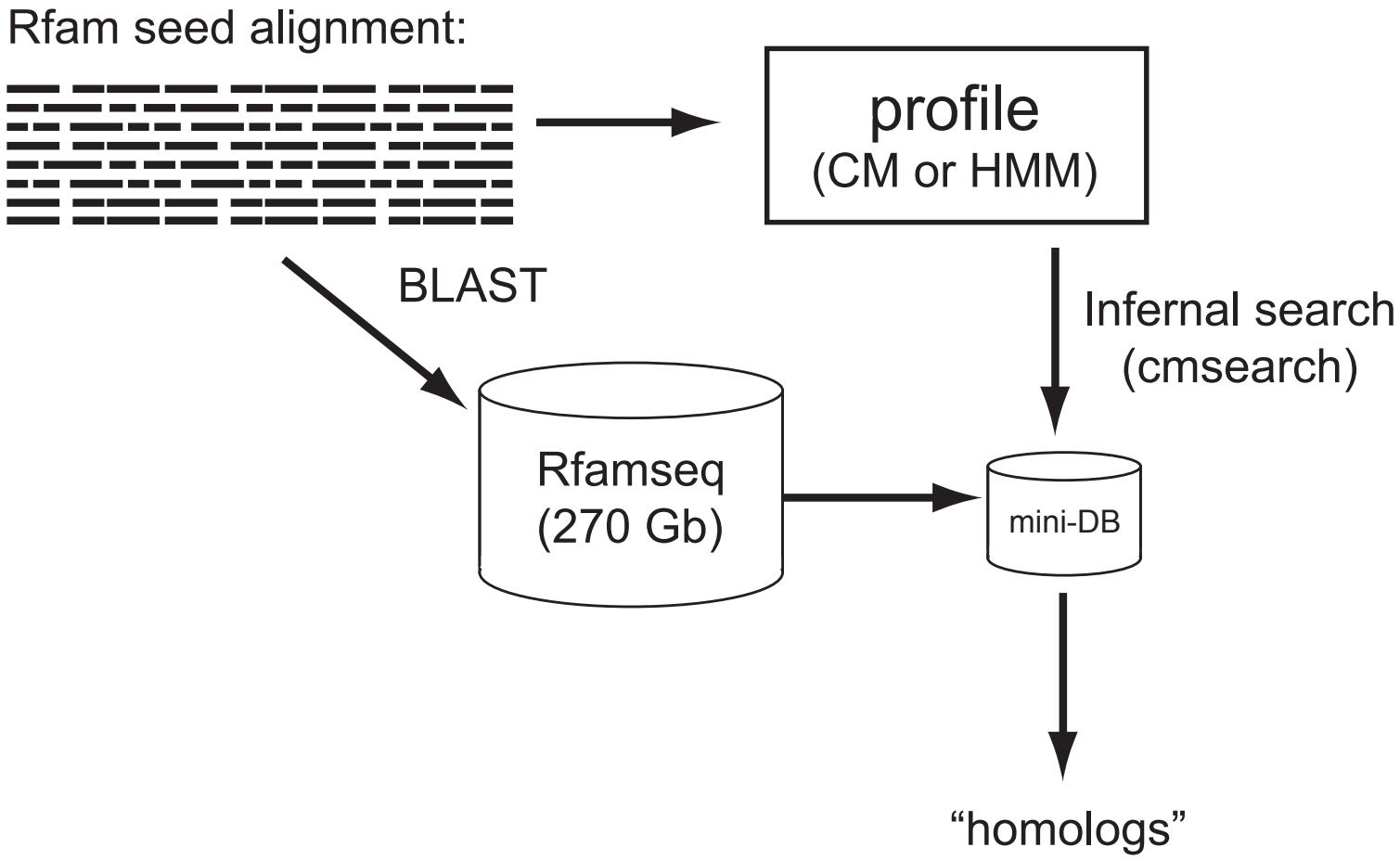
Rfam used BLAST filters from 2003 to 2012

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

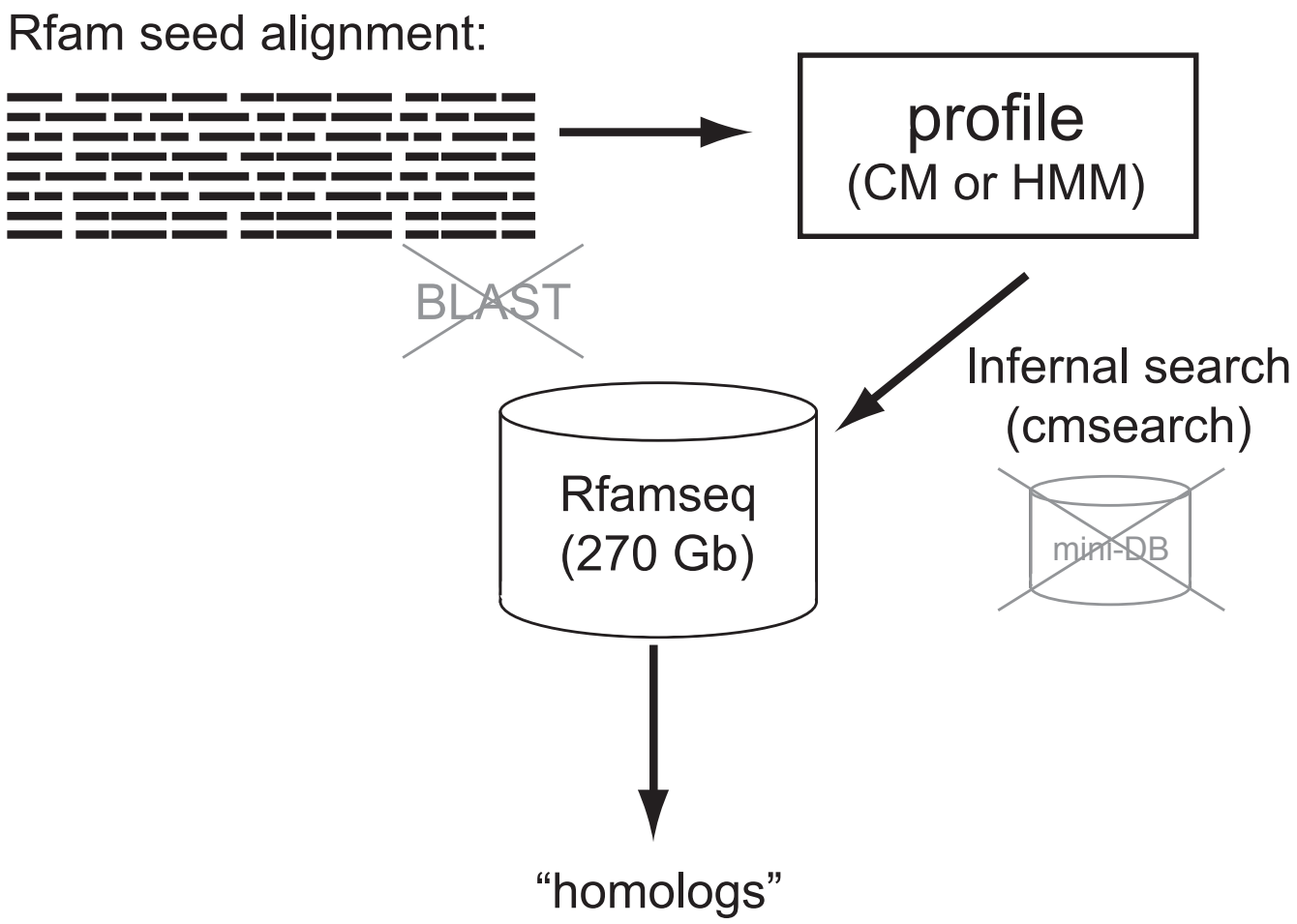


Rfam used BLAST filters from 2003 to 2012

- 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



*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 2450 models.
- Michael DiCuccio's group is working on adding Rfam/Infernal to the GPIPE pipeline.

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.

Acknowledgements

Alejandro Schäffer

David Landsman

David Lipman

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

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