Covariance models and Infernal 1.1

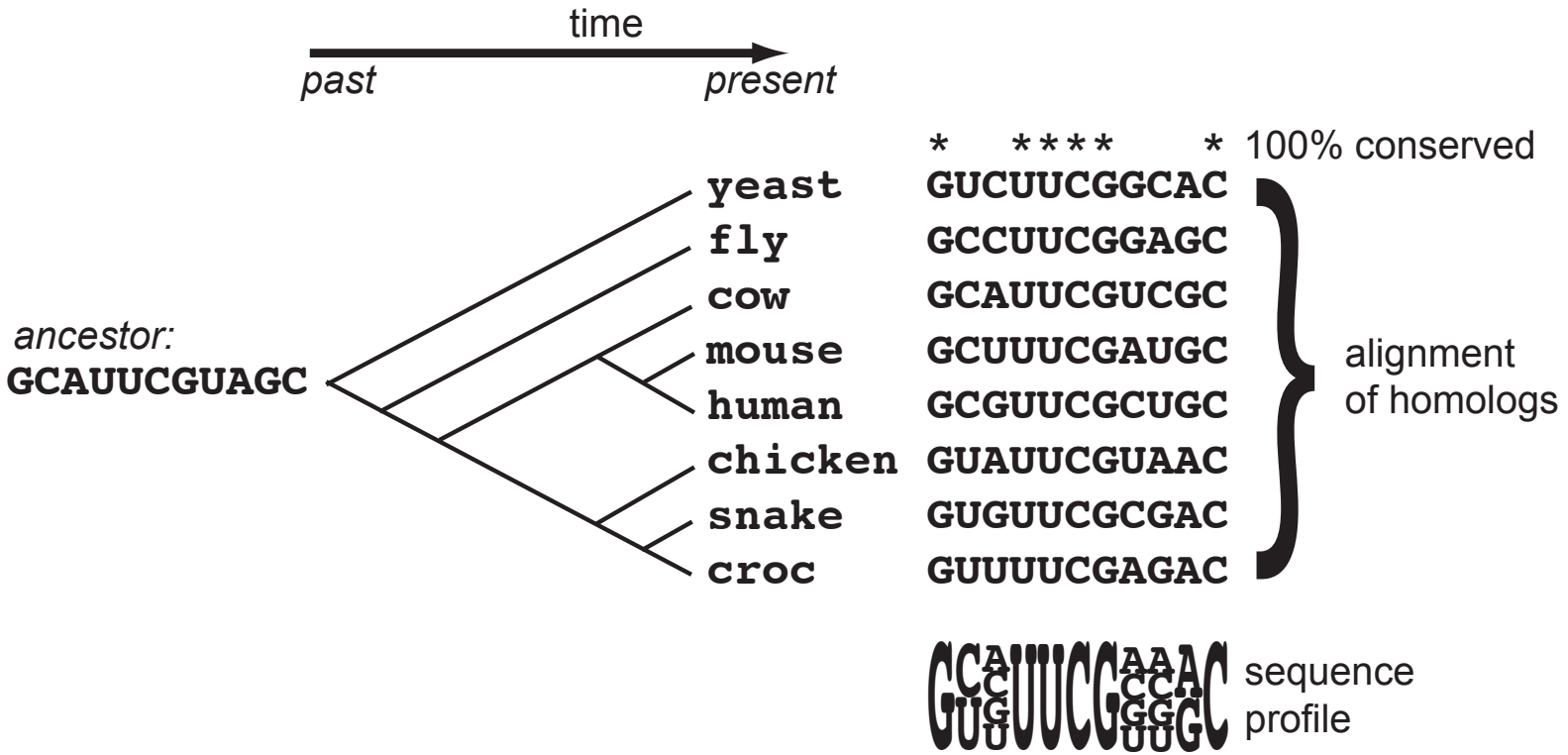
Eric Nawrocki

Sean Eddy's Lab

Howard Hughes Medical Institute
Janelia Farm Research Campus

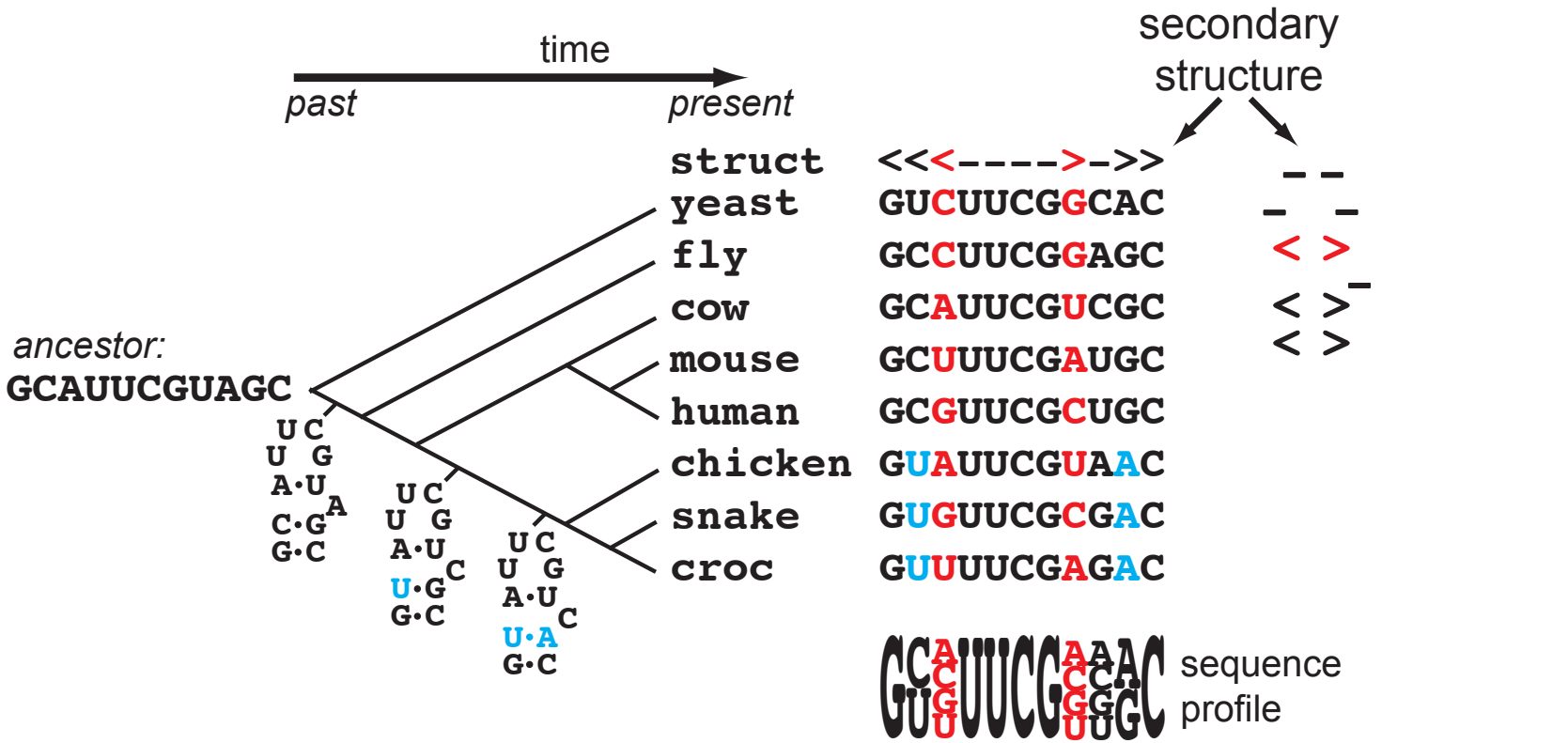
Sequence conservation provides information for homology searches

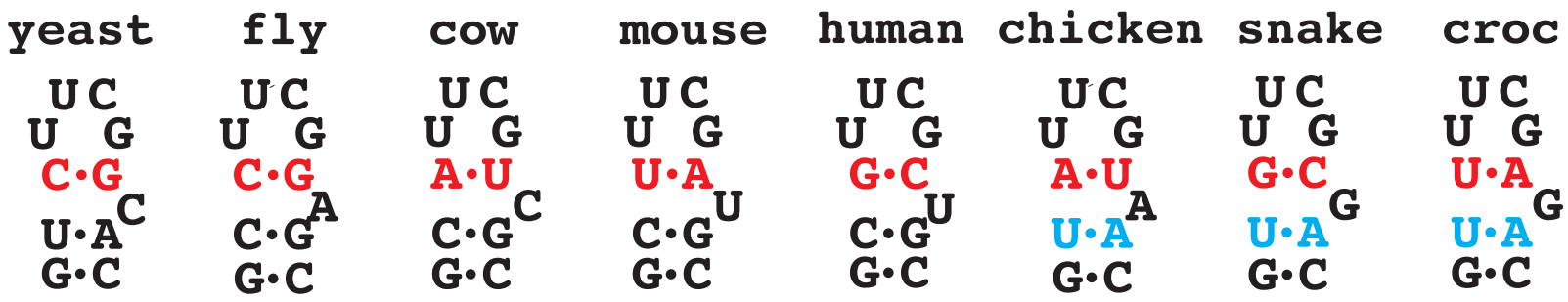
Conservation levels vary across alignment columns.



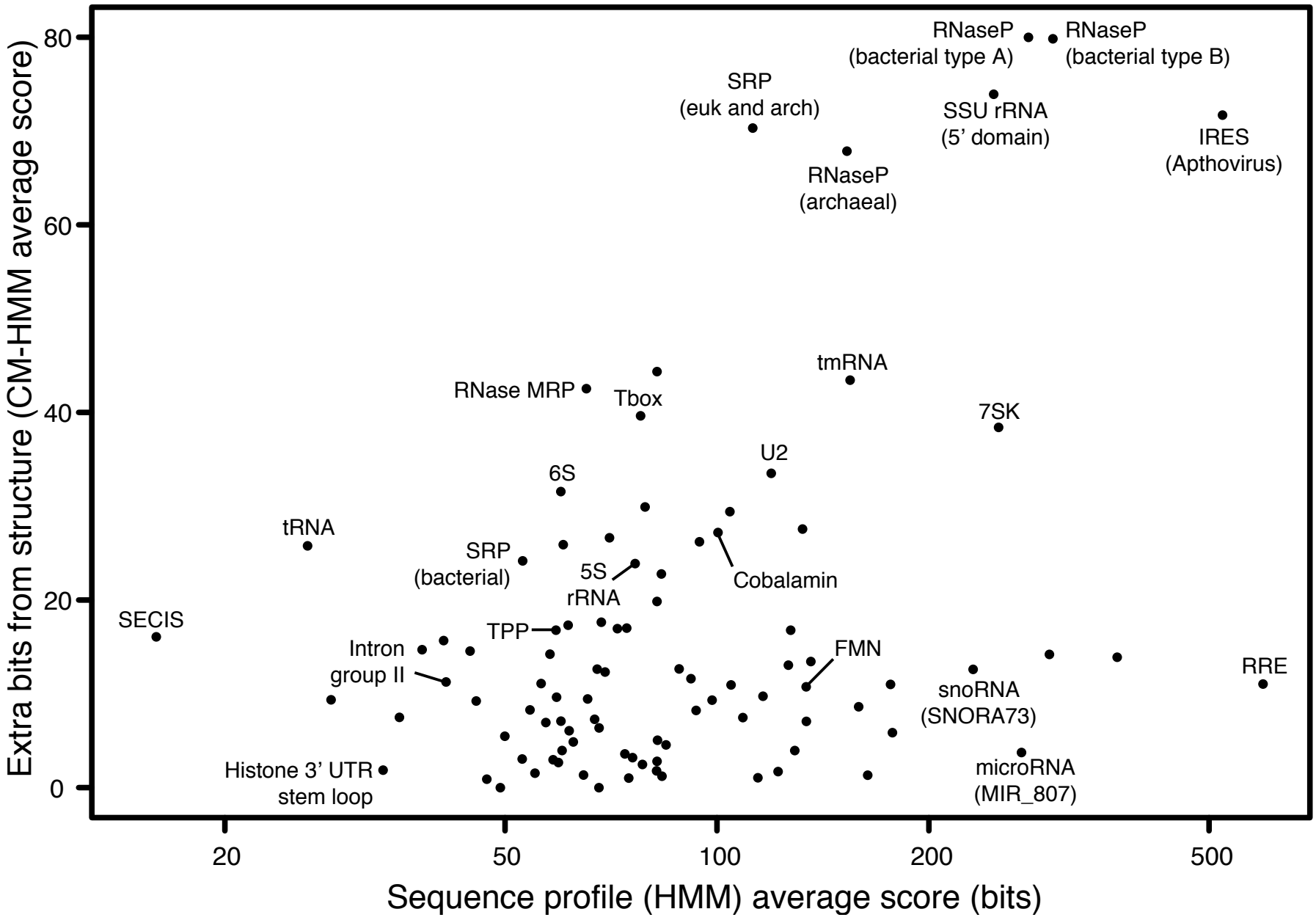
Structure conservation provides additional information

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





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 (prev. COVE)
main use	proteins	RNAs
database	Pfam (12273 families)	Rfam (1973 families)
performance for RNAs	faster but less accurate	slower but more accurate

http://hmmer.janelia.org

Eddy, SR. PLoS Comp. Biol., 4:e1000069, 2008.

Eddy. SR. Bioinformatics, 14:755-763, 1998.

http://infernal.janelia.org

Nawrocki EP, Kolbe DL, Eddy SR Bioinformatics, 25 (10):1335-1337, 2009.

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

Profile HMMs: sequence family models built from alignments

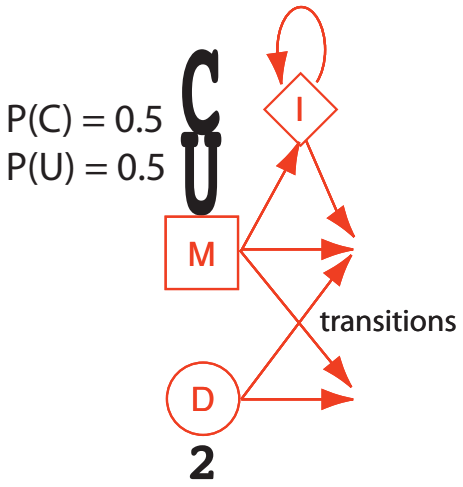
yeast	GUCUUCGGCAC
fly	GCCUUCGGAGC
cow	GCAUUCGUCGC
mouse	GCUUUCGAUGC
human	GCGUUCGCUGC
chicken	GUAUUCGUAAC
snake	GUGUUCGCGAC
croc	GUUUUCGAGAC

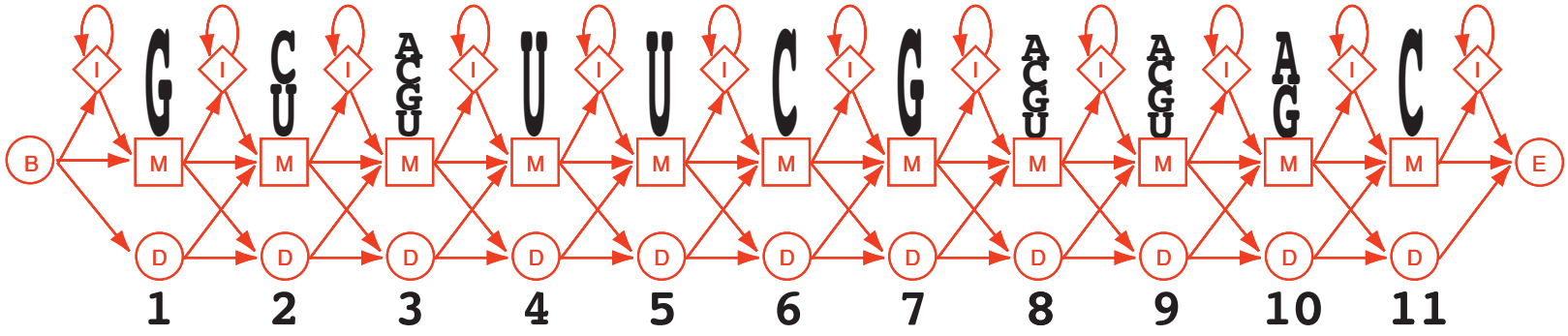
One HMM node per alignment column

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
```

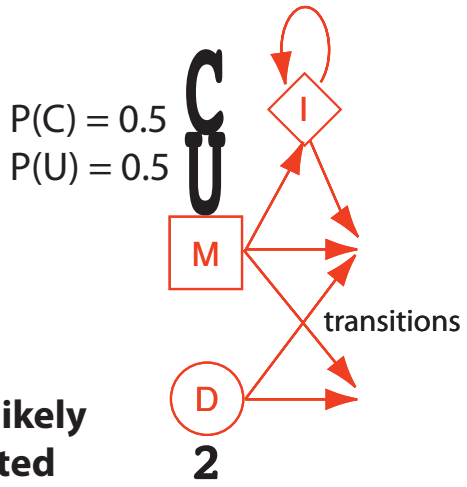
One HMM node per alignment column

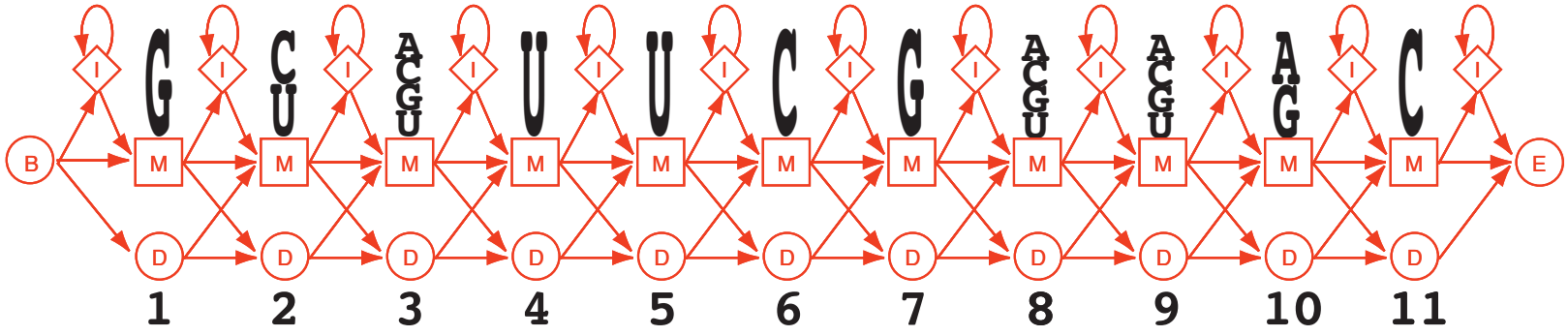
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

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
```

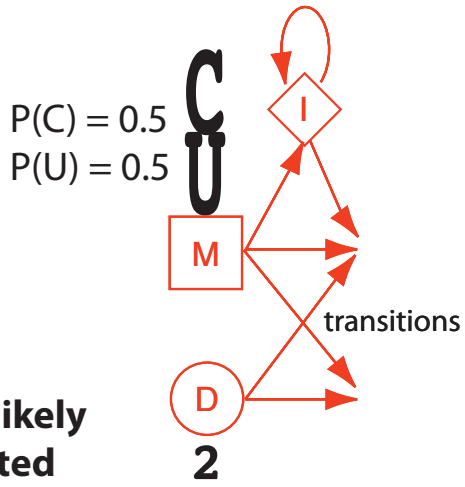
One HMM node per alignment column

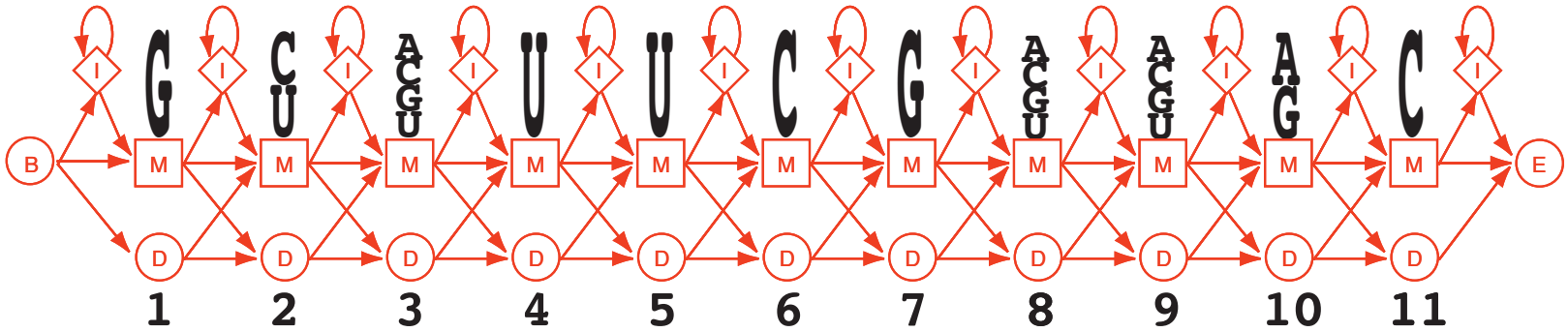
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

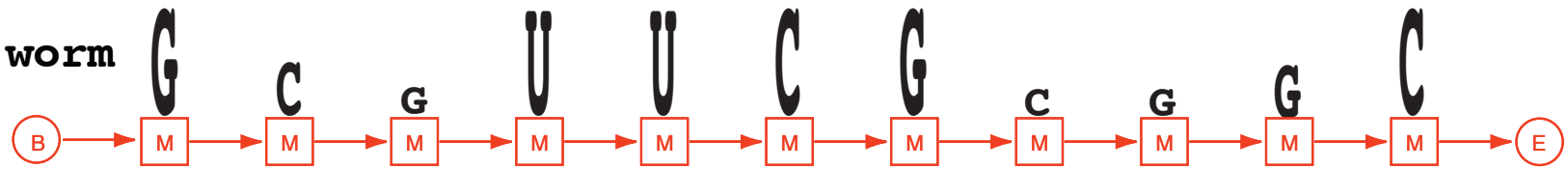
HMMs generate homologous sequences.

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
```

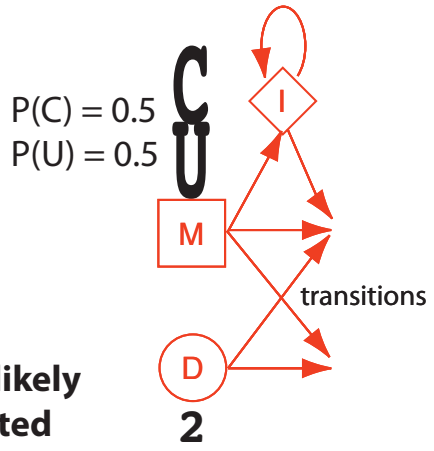
One HMM node per alignment column

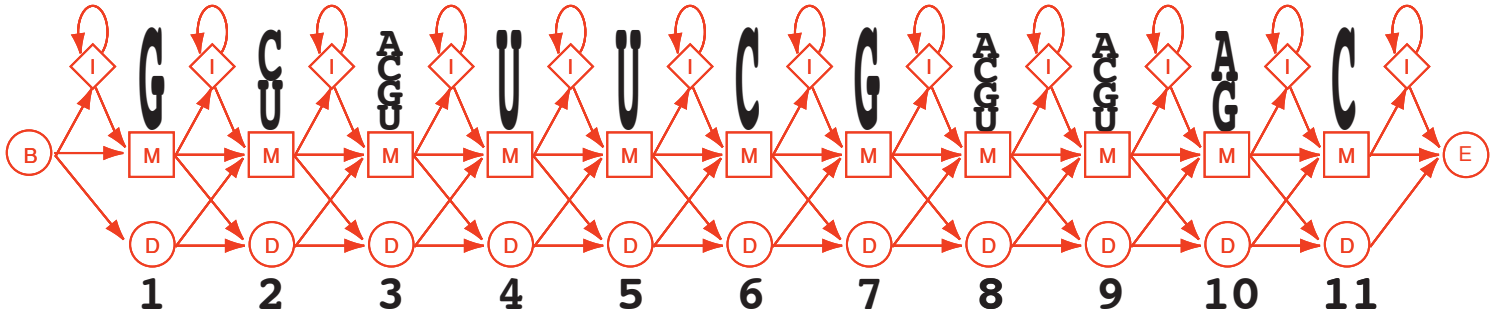
3 states per node:
(M) Match: emits residues
(I) Insert: inserts extra residues
(D) Delete: deletes residues

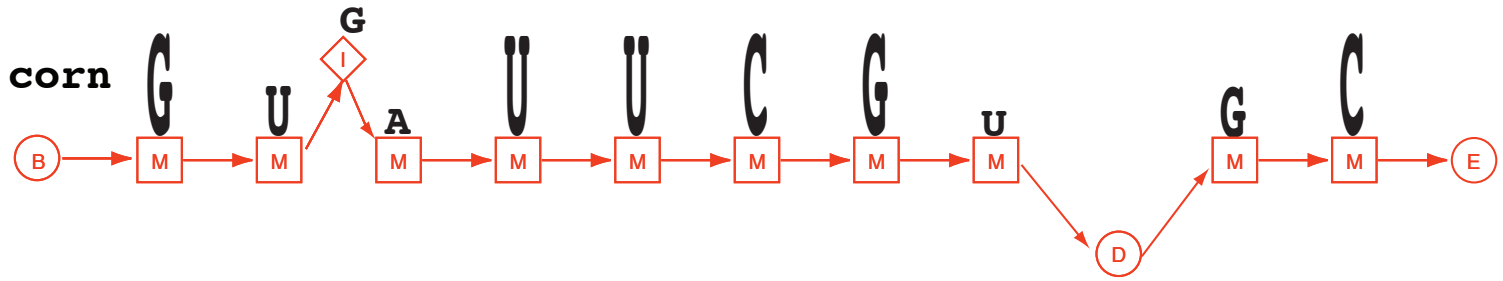
HMMs generate homologous sequences.

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

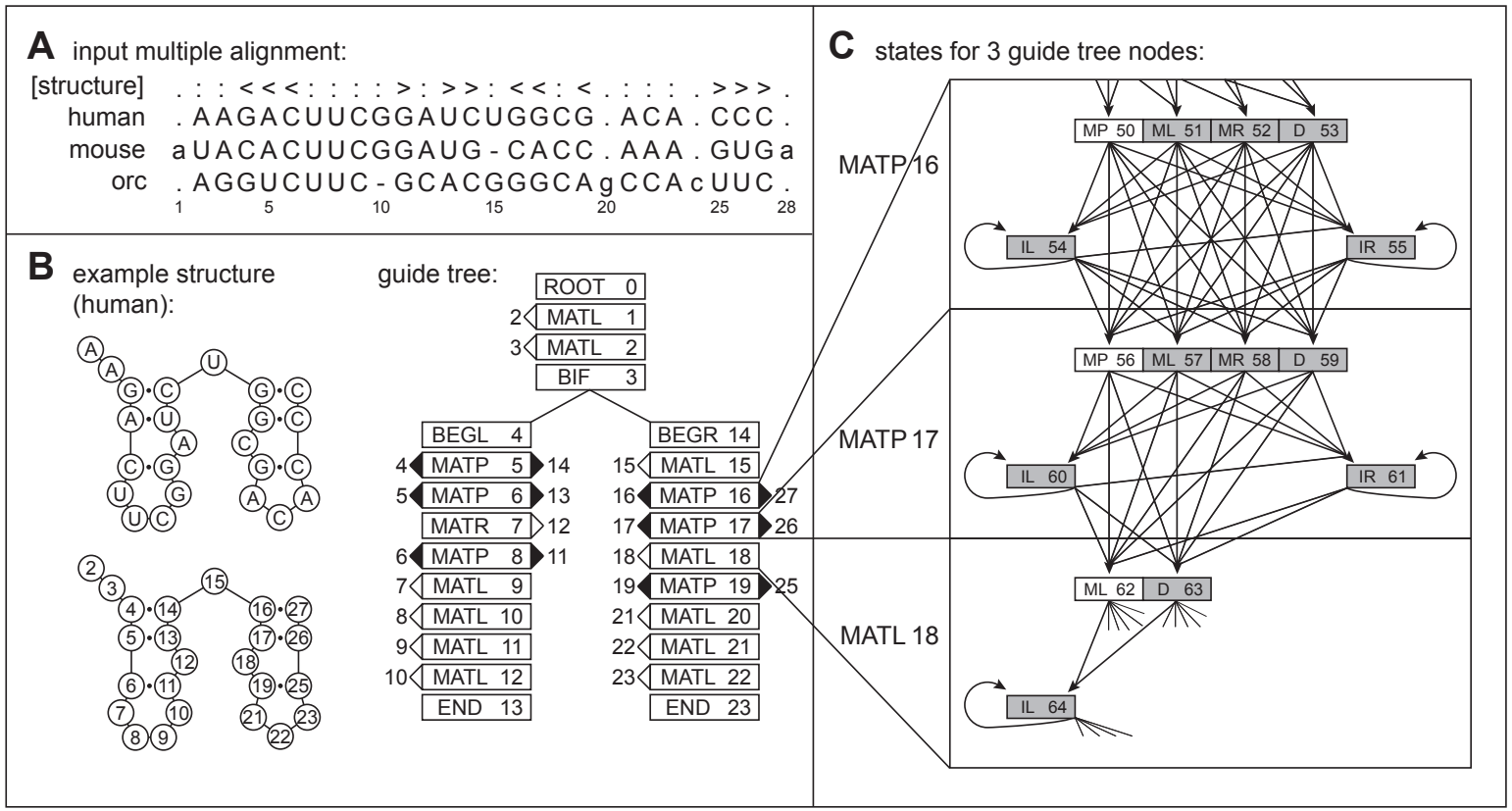
Node for column 2:

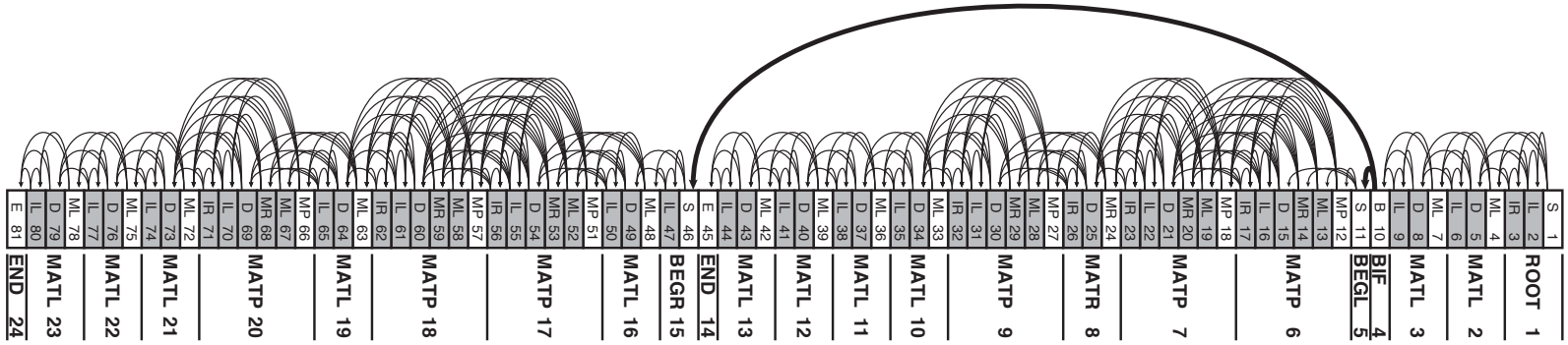






Covariance models (CMs) are built from structure-annotated alignments



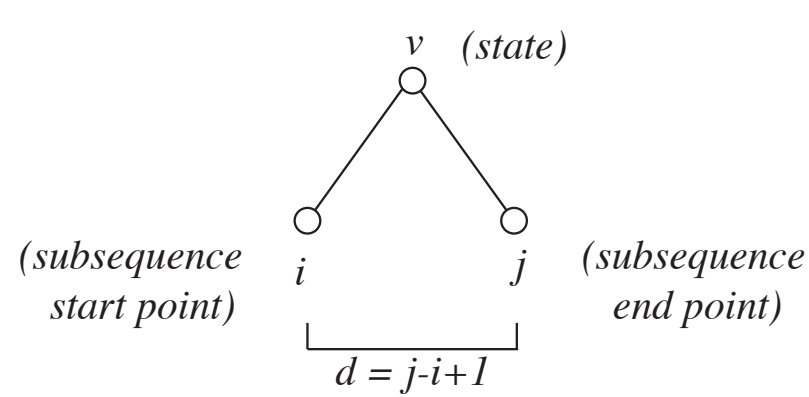


CM searches are slow (especially for large RNAs)

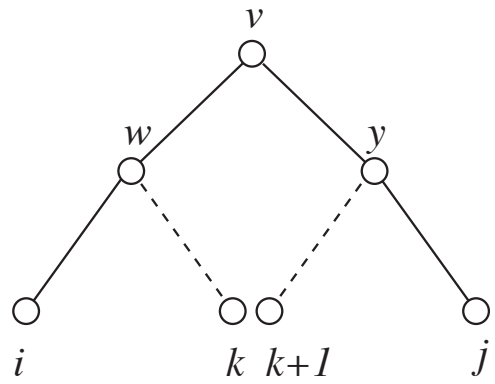
model	DP algorithms	direction	complexity
profile HMM	Viterbi, Forward	left to right	$O(N^2)$
CM	CYK, Inside	inside to outside	$O(N^3 logN)$

CYK and Inside fill in a 3-dimension matrix with (v, i, j) tuples:

Non-bifurcation states:



Bifurcation states:



Accelerating CM searches: filtering and banded DP

- Filtering: search database with faster method first, hits above some threshold survive the filter and are searched with the slow CM.
  - *tRNAscan-SE*[*]: filters database using tRNA-specific heuristics (1997)
  - Rfam uses BLAST filters (2002-present)
  - Weinberg and Ruzzo[†] developed HMM filters for faster searches (2004, 2006).
  - Others have also worked on this (Sun and Buhler [‡], Zhang and Bafna[§])
- Query dependent banding (QDB)[¶]: restrict subsequence lengths that can align to each state of the model.
- Infernal 1.0: HMM filtering and QDB yield 100-fold acceleration (2009)

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

[†]Weinberg Z, Ruzzo WL. Bioinformatics 22(1):3539, 2006.

[‡]Sun Y, Buhler J, Comput. Systems Bioinf., p145-156, 2008.

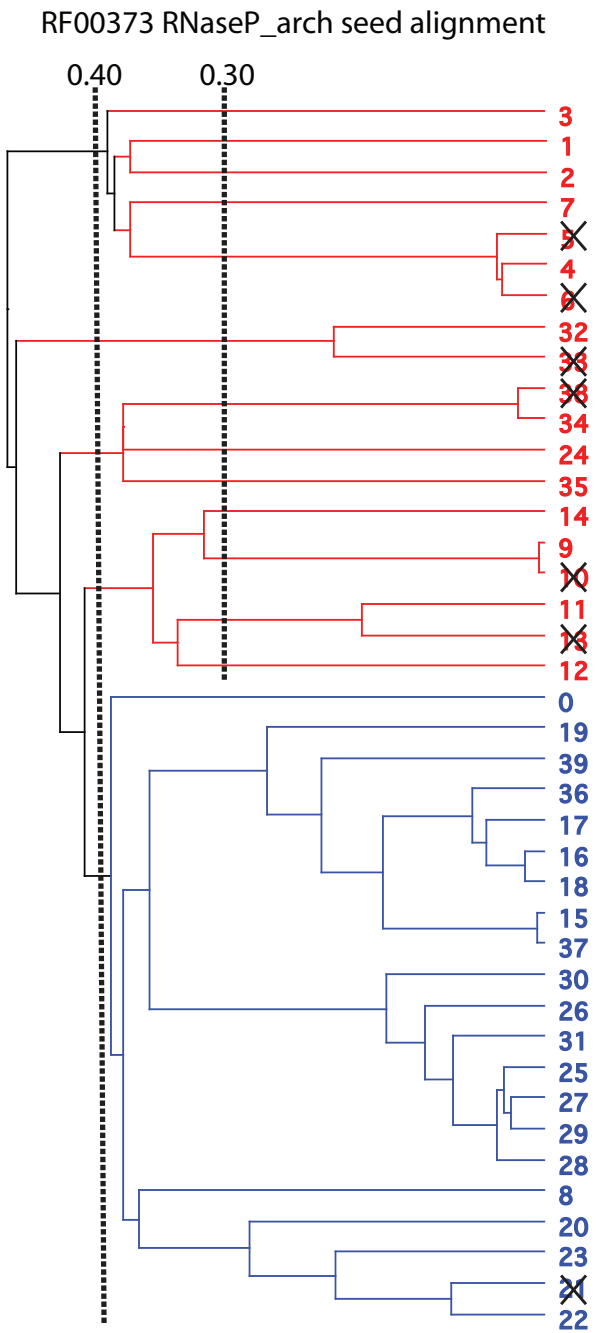
[§]Zhang S et al., Bioinformatics. 22(14):e557-e565, 2006.

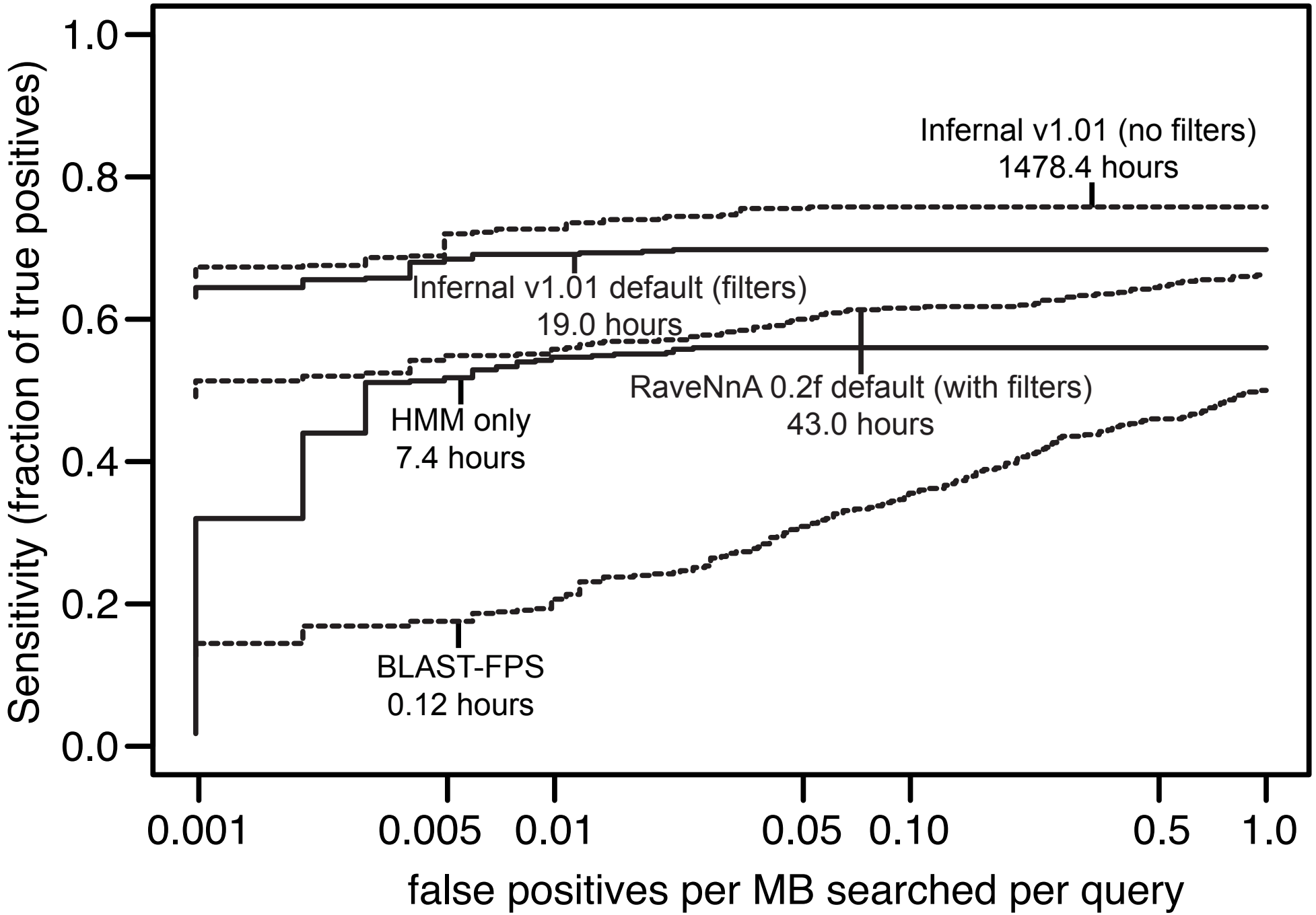
[¶]Nawrocki and Eddy, PLoS Comp Bio, 2007

RMARK: an internal RNA homology search benchmark

- RMARK construction - for each of the 503 Rfam 7.0 seed alignments:
  - single-linkage cluster sequences by sequence identity given the alignment
  - look for a training cluster and testing cluster such that:
    * no training/test sequence pair is $> 60\%$ identical
  - filter test set so no two test seqs $> 70\%$ identical
  - 51 families qualify, with 450 test sequences
  - test seqs are embedded in a 10 Mb (20 Mb counting both strands) pseudo-genome of "realistic" base composition (RMARK2)

Example:

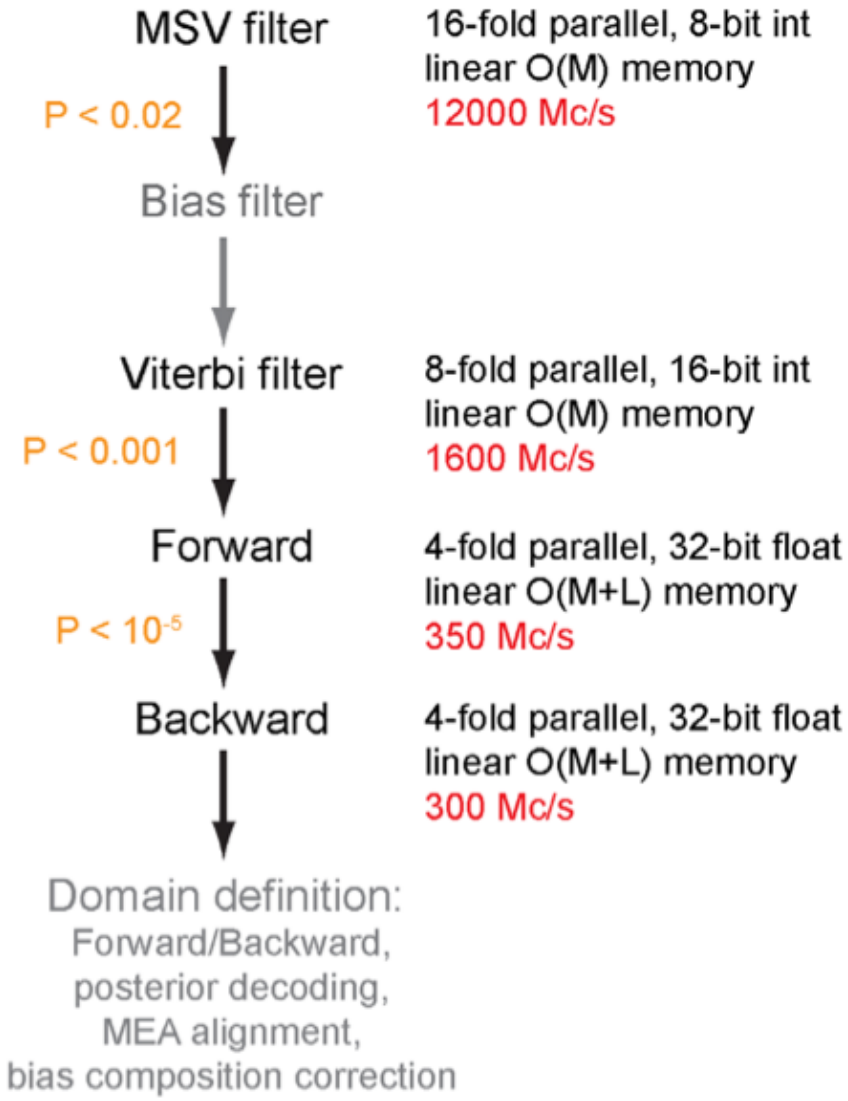



Sensitivity (fraction of true positives)
1.0
0.8
0.6
0.4
0.2
0.0
Infernal v1.01 (no filters)
1478.4 hours
Infernal v1.01 default (filters)
19.0 hours
RaveNnA 0.2f default (with filters)
43.0 hours
HMM only
7.4 hours
BLAST-FPS
0.12 hours
0.001
0.005
0.01
0.05
0.10
0.5
1.0
false positives per MB searched per query

100 to 1000-fold faster profile HMM searches using HMMER3

(Eddy, PLoS Comp Bio, 2011)

- Striped vector parallelization (Farrar, Bioinformatics, 2007)



Infernal 1.1's HMMER3-based cmsearch pipeline

	model	hmmer3	infernal 1.1
F1 (MSV/SSV)	HMM	P=0.02	P=0.35
*windows > 2*W are split up*			
F2 (Viterbi)	HMM	P=0.001	P=0.15
F3 (Forward)	HMM	P=0.00001	P=0.003
F4 (glocal Forward)	HMM		P=0.003
overlapping windows are merged			
F5 (glocal envelope defn)	HMM		P=0.003
F6 (HMM banded CYK)	CM		P=0.0001
F7 (HMM banded Inside)	CM		

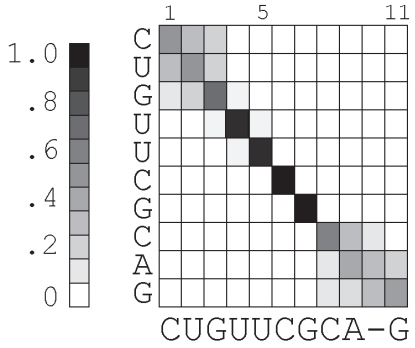
- Glocal HMM envelope definition trims off residues unlikely to be involved in the hit, enabling use of fast HMM banded CM algorithms.

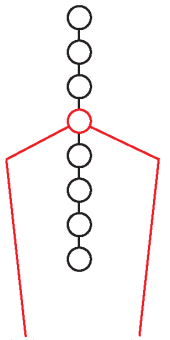
HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable

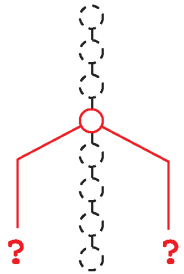
HMMs– Each column of seed alignment corresponds to a column of matrix.

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





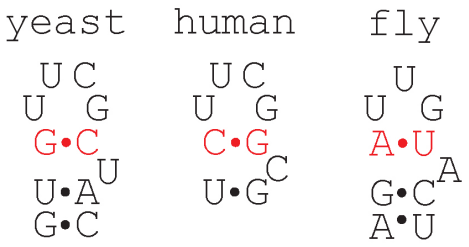
```
struct <<<<----->->>>
yeast  GUGUUCGCUAC
human  -UCUUCGGCG-
fly    AGAUU-GUACU
```



CUGUUCGCAG

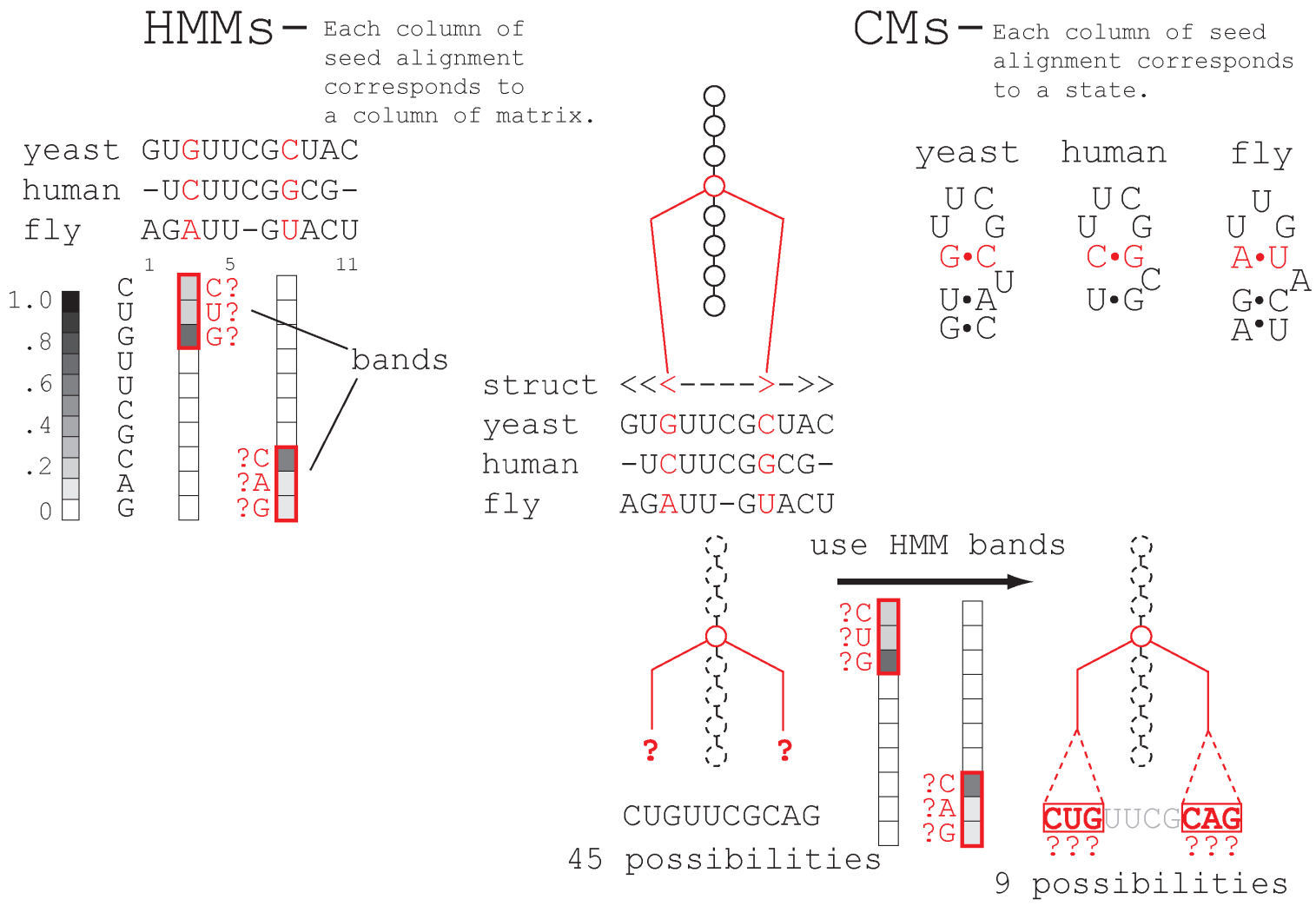
45 possibilities

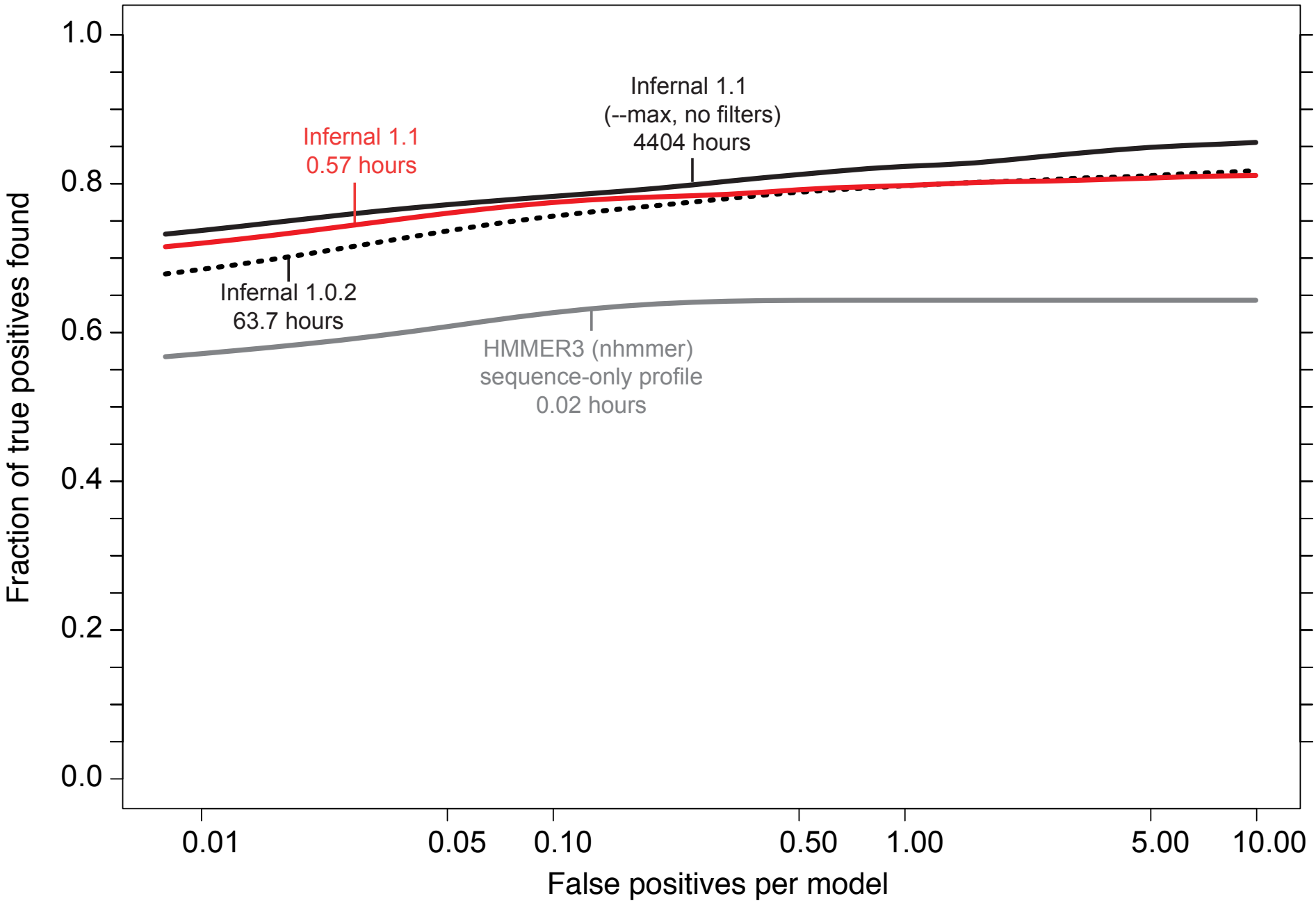
CMs– Each column of seed alignment corresponds to a state.

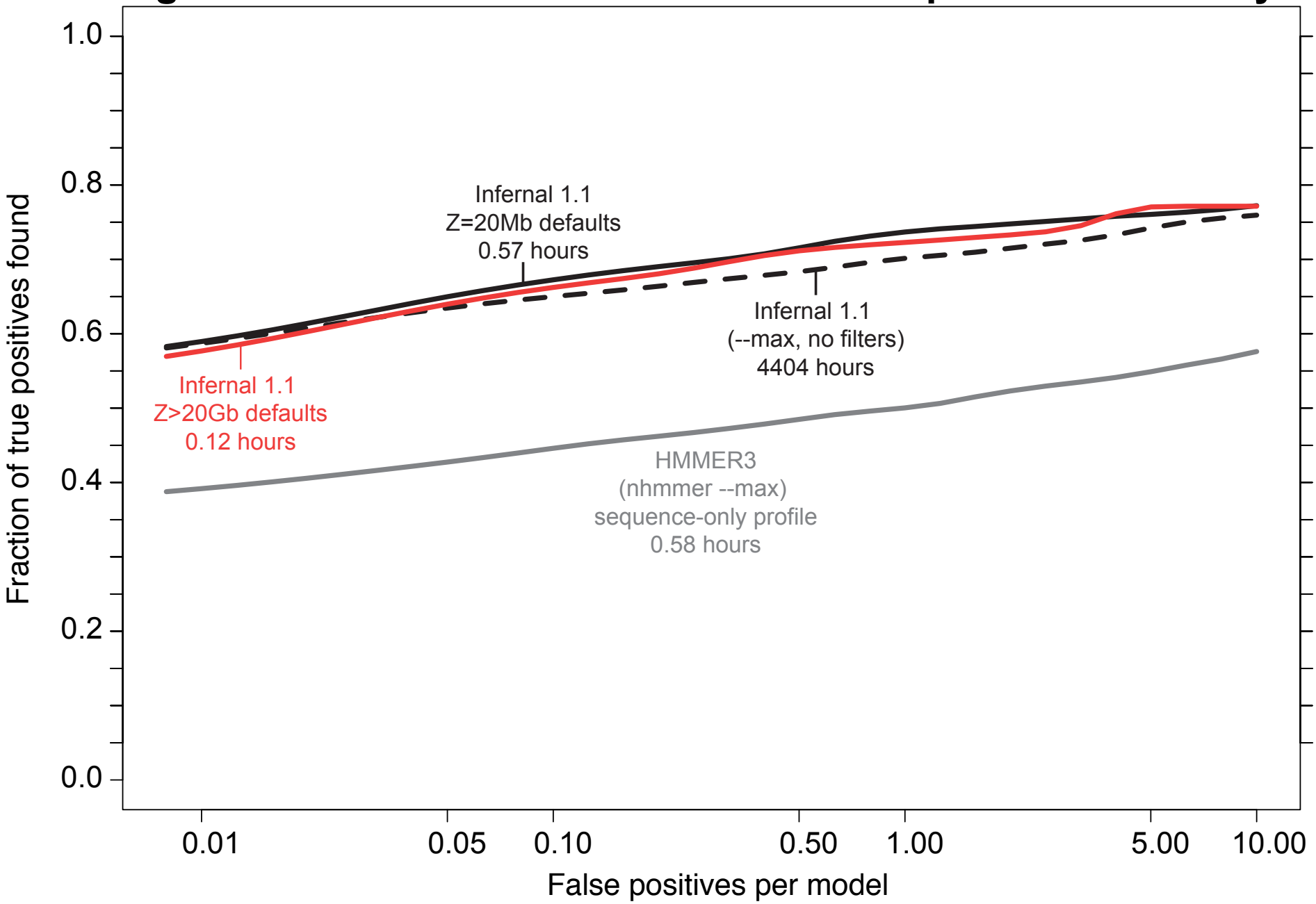


HMM bands accelerate CM alignment

- **main idea:** eliminate potential alignments the HMM tells us are very improbable



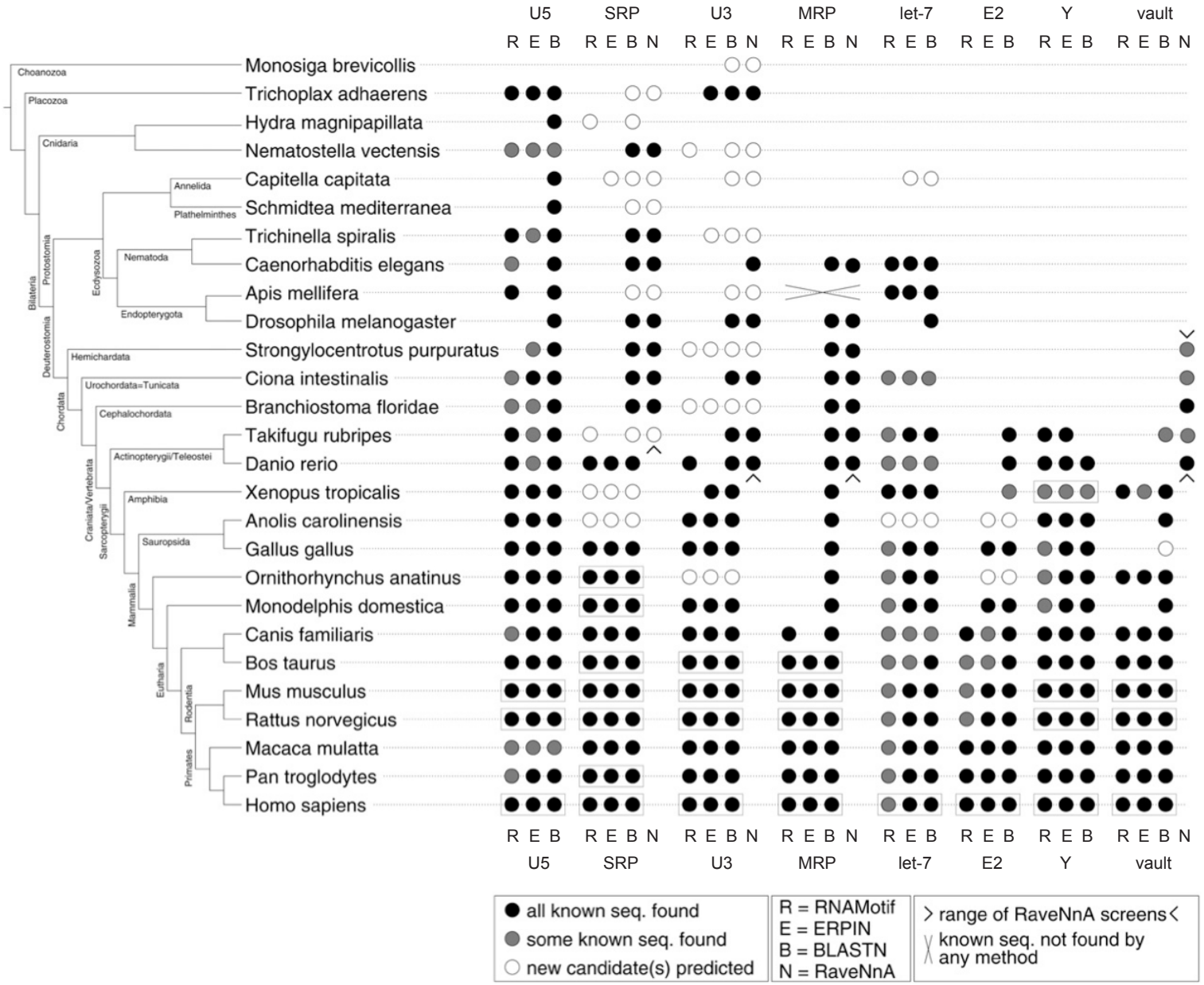

Infernal v1.1 is 100X faster and more sensitive than v1.0.2
Fraction of true positives found
1.0
0.8
0.6
0.4
0.2
0.0
Infernal 1.1
(--max, no filters)
4404 hours
Infernal 1.1
0.57 hours
Infernal 1.0.2
63.7 hours
HMMER3 (nhmmer)
sequence-only profile
0.02 hours
0.01
0.05
0.10
0.50
1.00
5.00
10.00
False positives per model


Tighter filter thresholds for big databases (Z>20Gb)
give 5X further acceleration with little impact on sensitivity
Fraction of true positives found
1.0
0.8
0.6
0.4
0.2
0.0
Infernal 1.1
Z=20Mb defaults
0.57 hours
Infernal 1.1
(--max, no filters)
4404 hours
Infernal 1.1
Z>20Gb defaults
0.12 hours
HMMER3
(nhmmer --max)
sequence-only profile
0.58 hours
0.01
0.05
0.10
0.50
1.00
5.00
10.00
False positives per model

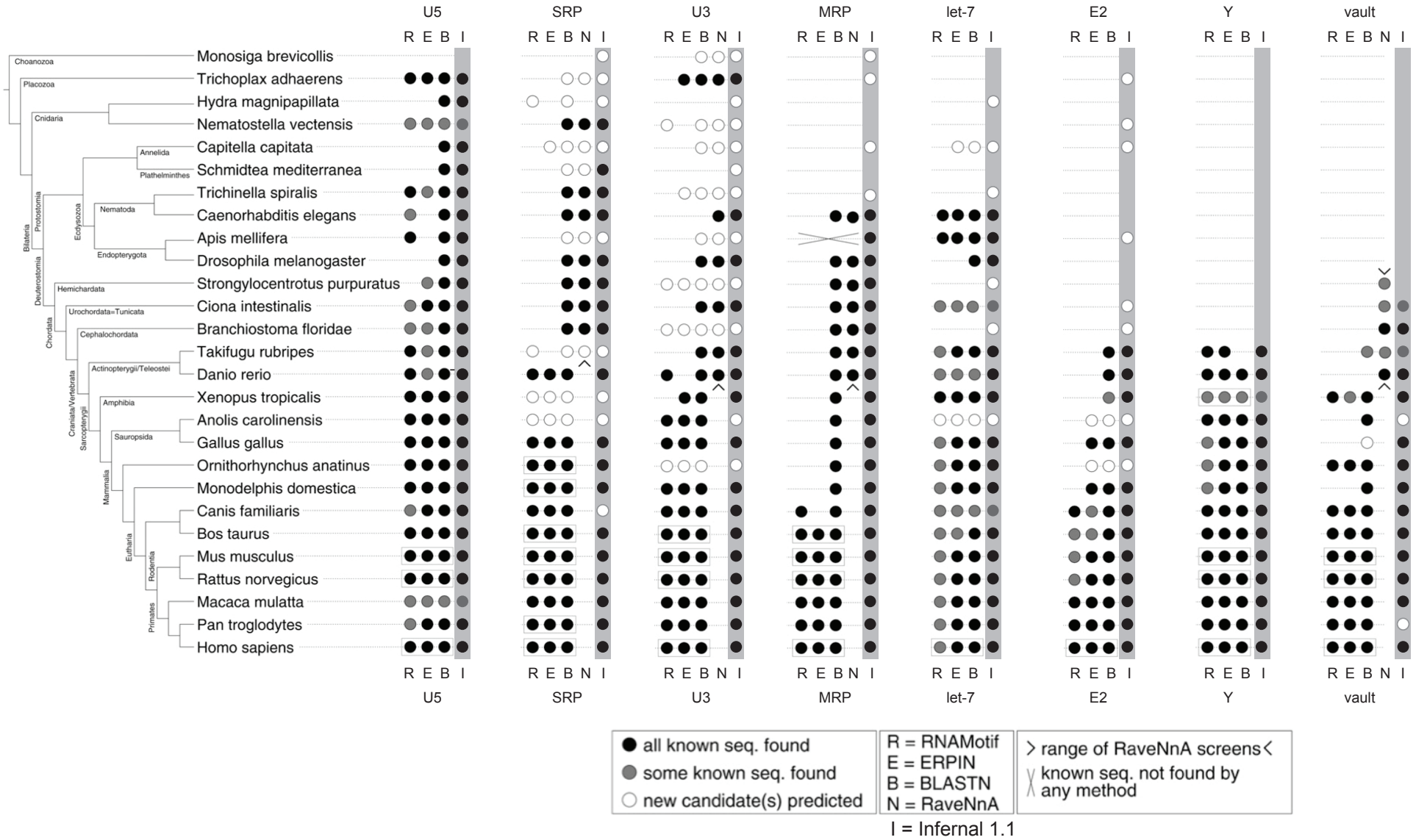
The tedious task of finding homologous noncoding RNA genes

PETER MENZEL,[1,2] JAN GORODKIN,[1] and PETER F. STADLER[2,3,4,5,6]

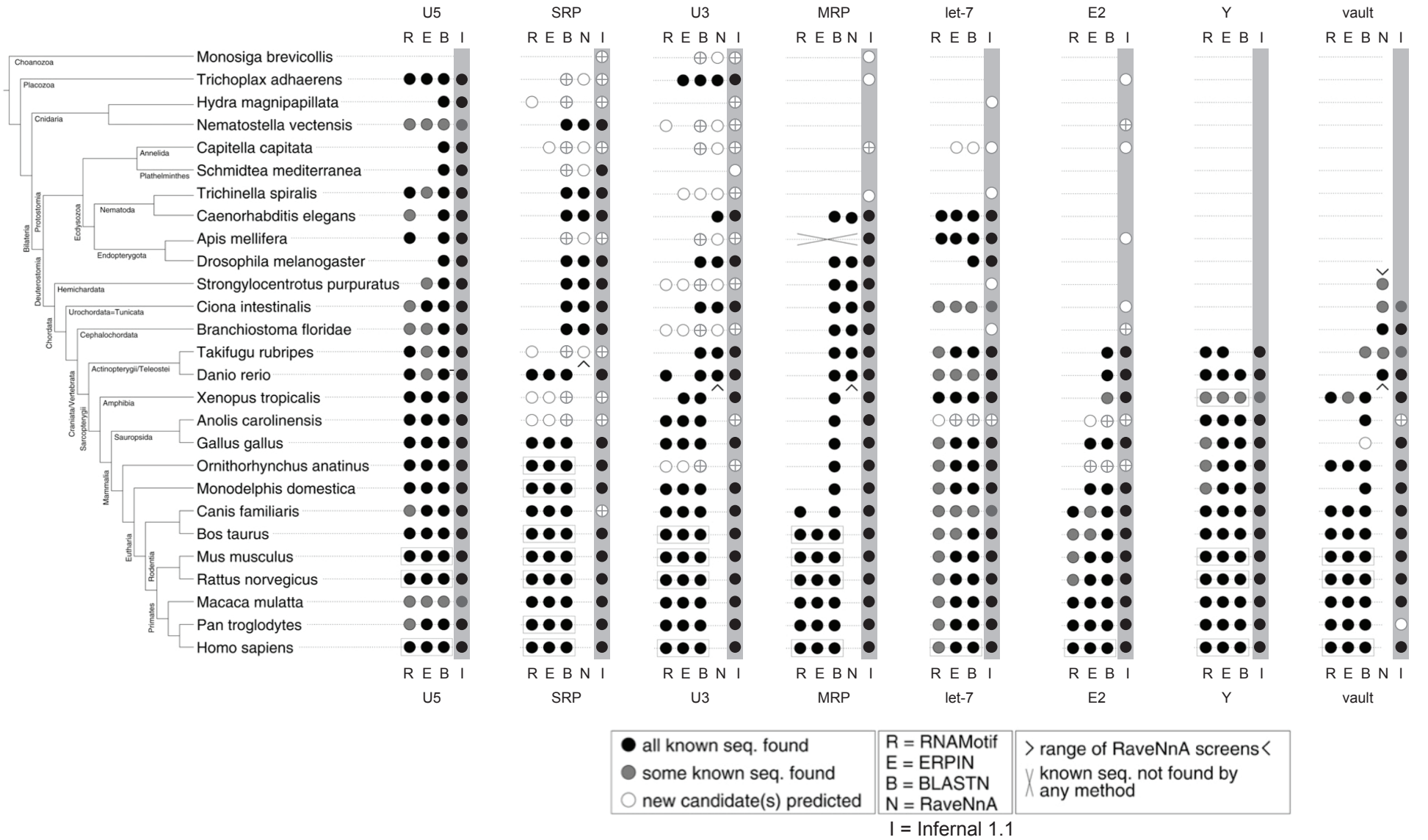
- Searched for 8 RNA families in 26 animal genomes plus a choanoflagellate with RNAMotif, ERPIN, BLASTN, and RaveNnA (for non-vertebrates).
  - families: U5, SRP, U3, RNase MRP, let-7, E2, Y, Vault.
- BLASTN does as well or better than other methods in most cases.
- RaveNnA finds some novel candidates missed by BLASTN in some cases, but is very slow.



Menzel et al., Bioinformatics, 2009.



Modified from Menzel et al., Bioinformatics, 2009.



Modified from Menzel et al., Bioinformatics, 2009.

Running times on all genomes (47 Gb) in hours

family	BLASTN	RNAMotif	ERPIN	Infernal
U5	0.12	0.44	3.68	8.50
SRP	1.82	120.73	2.14	290.94
U3	0.09	35.78	4.75	14.18
MRP	0.07	9.16	6.41	4.10
let-7	0.10	0.47	26.80	3.79
E2	0.08	0.58	3.16	4.81
Y	0.14	0.86	199.16	11.49
vault	0.08	140.81	87.67	4.32
total	2.52	308.83	333.77	342.12

Infernal 1.1

`infernal.janelia.org`

- faster homology searches
- no more BLAST filters for Rfam
- better handling of truncated sequences*
- `cmscan` program: search a sequence against a CM database (e.g. Rfam)

*Kolbe and Eddy, Bioinformatics, 2009

Acknowledgements

Sean Eddy
Michael Farrar
Travis Wheeler
Diana Kolbe