Modeling Structural RNA Families with Infernal

Eric Nawrocki

Sean Eddy's Lab

hhmi | janelia Research Campus

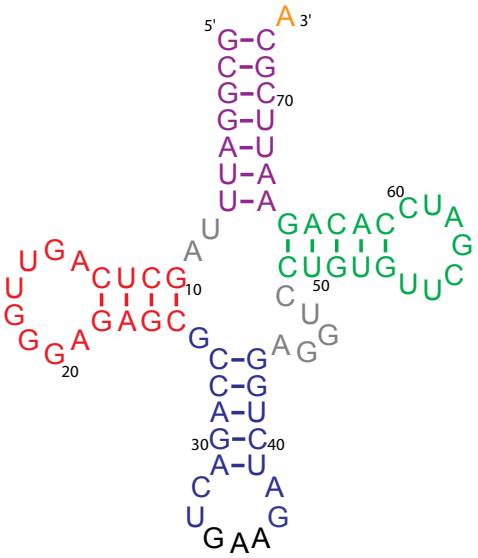
Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



3-dimensional structure

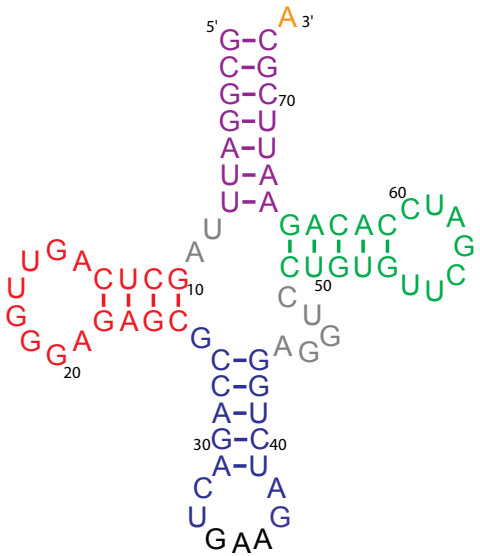
Many functional RNAs adopt a conserved 3-dimensional structure

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure

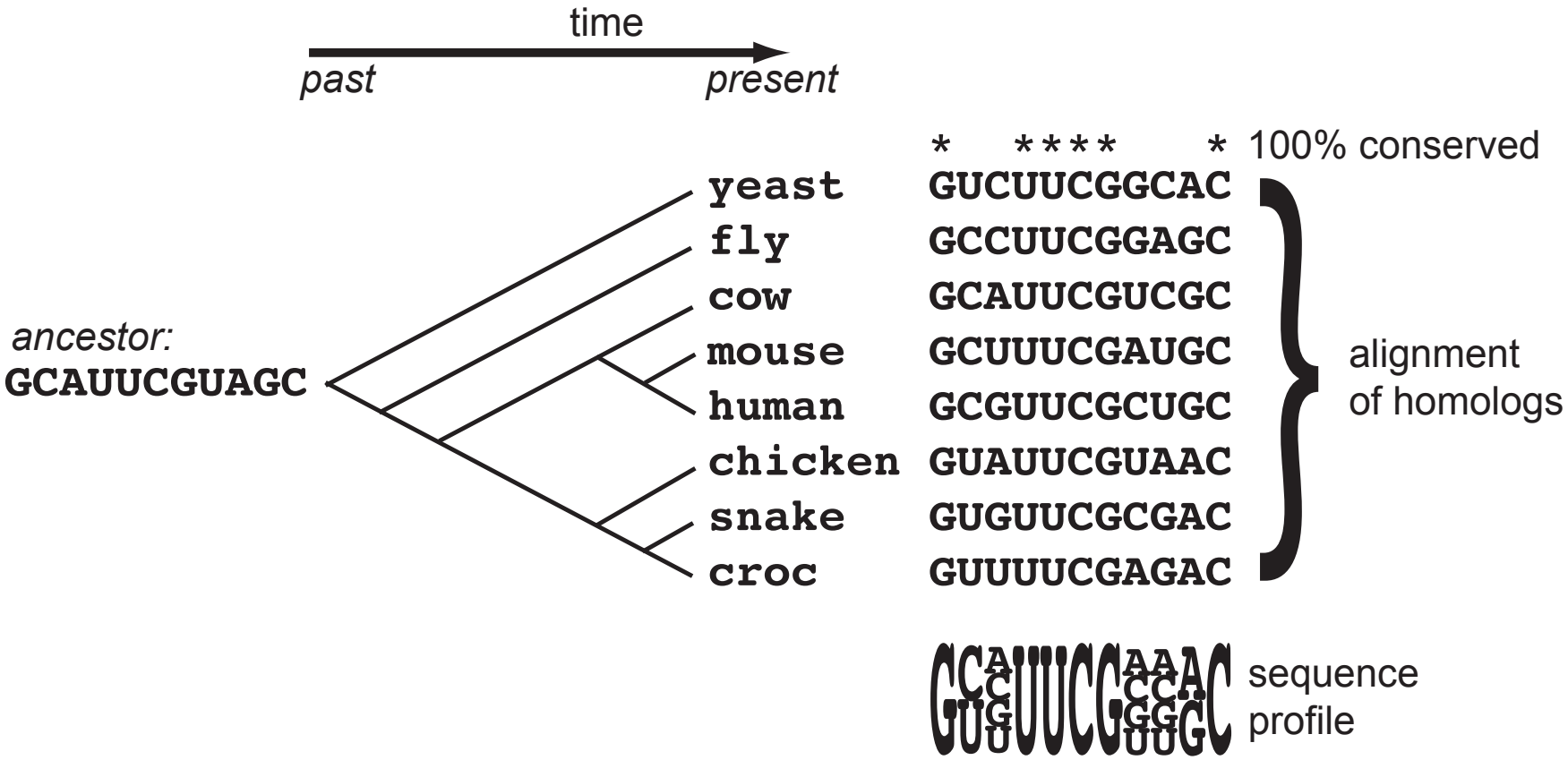


3-dimensional structure

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

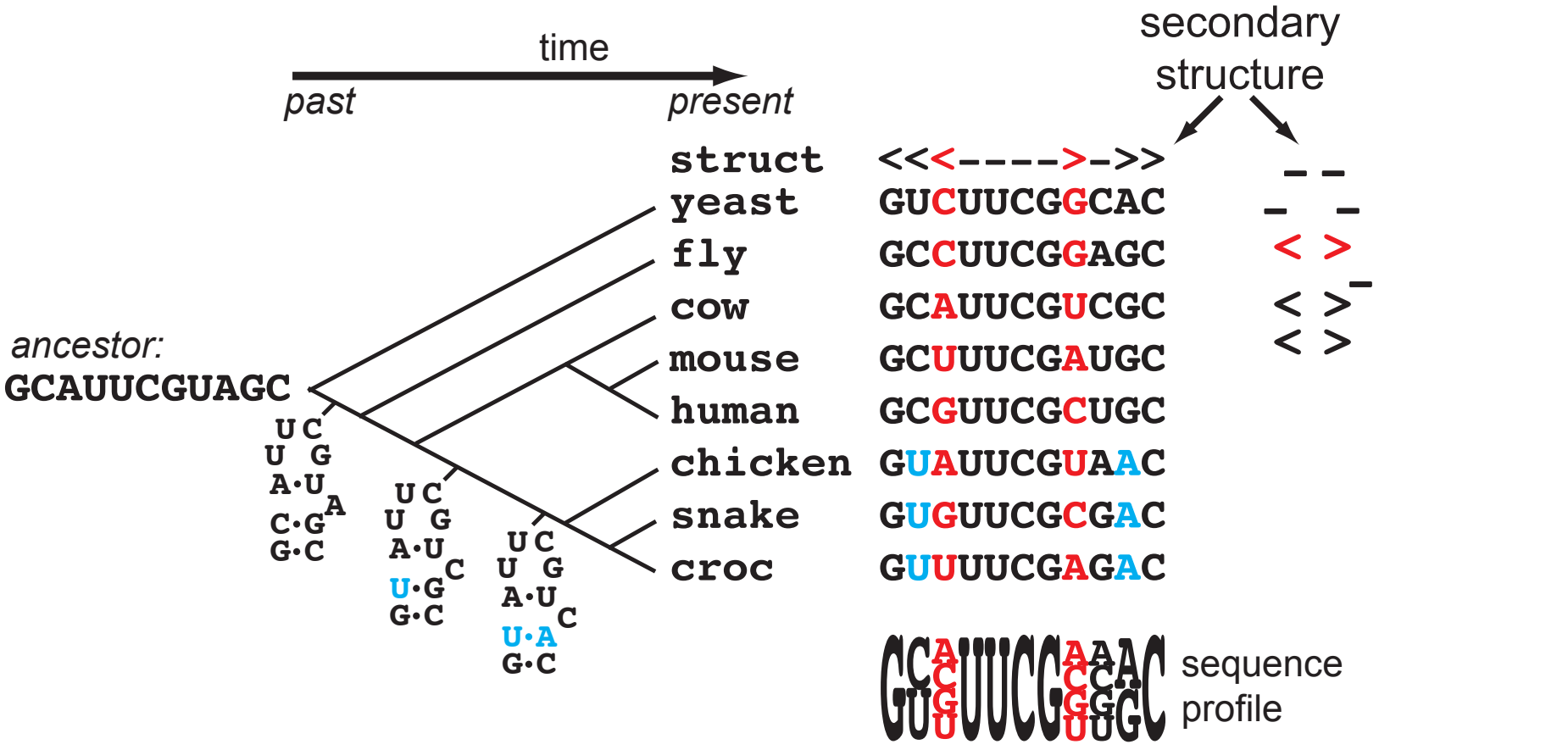
Sequence conservation provides information for homology searches

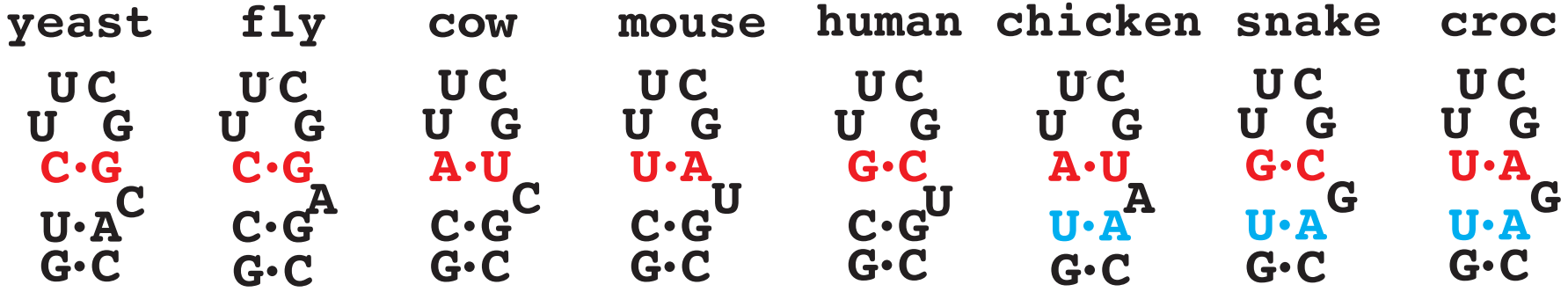
Conservation levels vary across alignment columns.



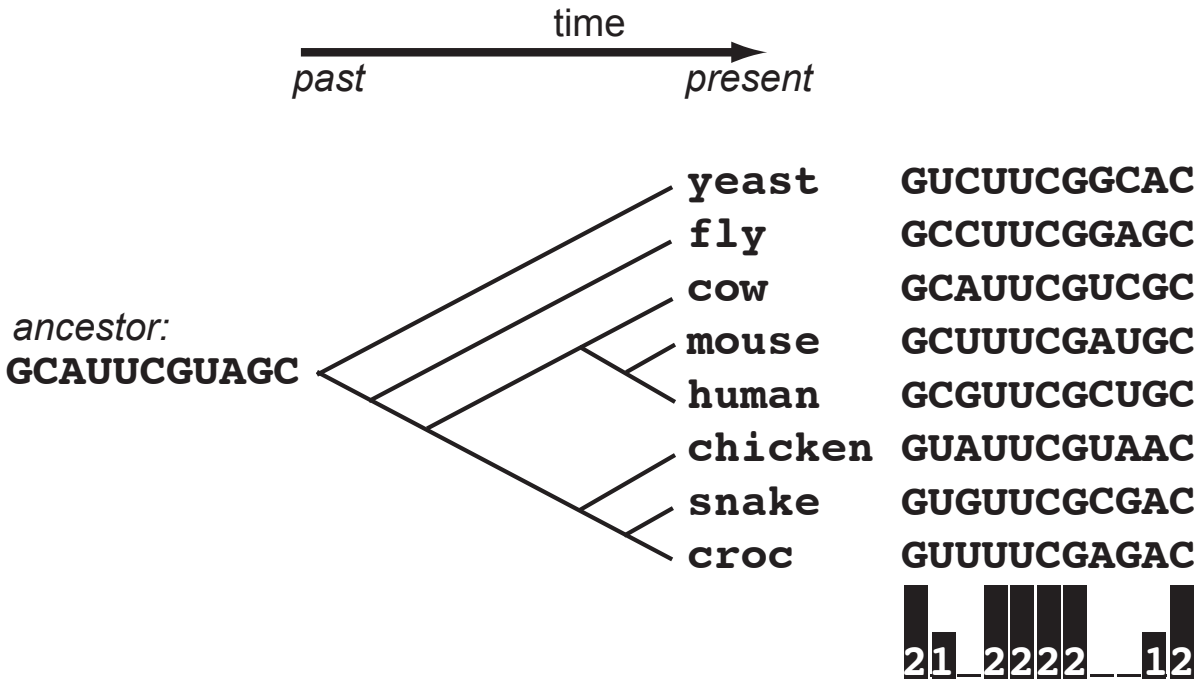
Structure conservation provides additional information

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





Amount of information in a profile can be measured in bits



Each column is independent

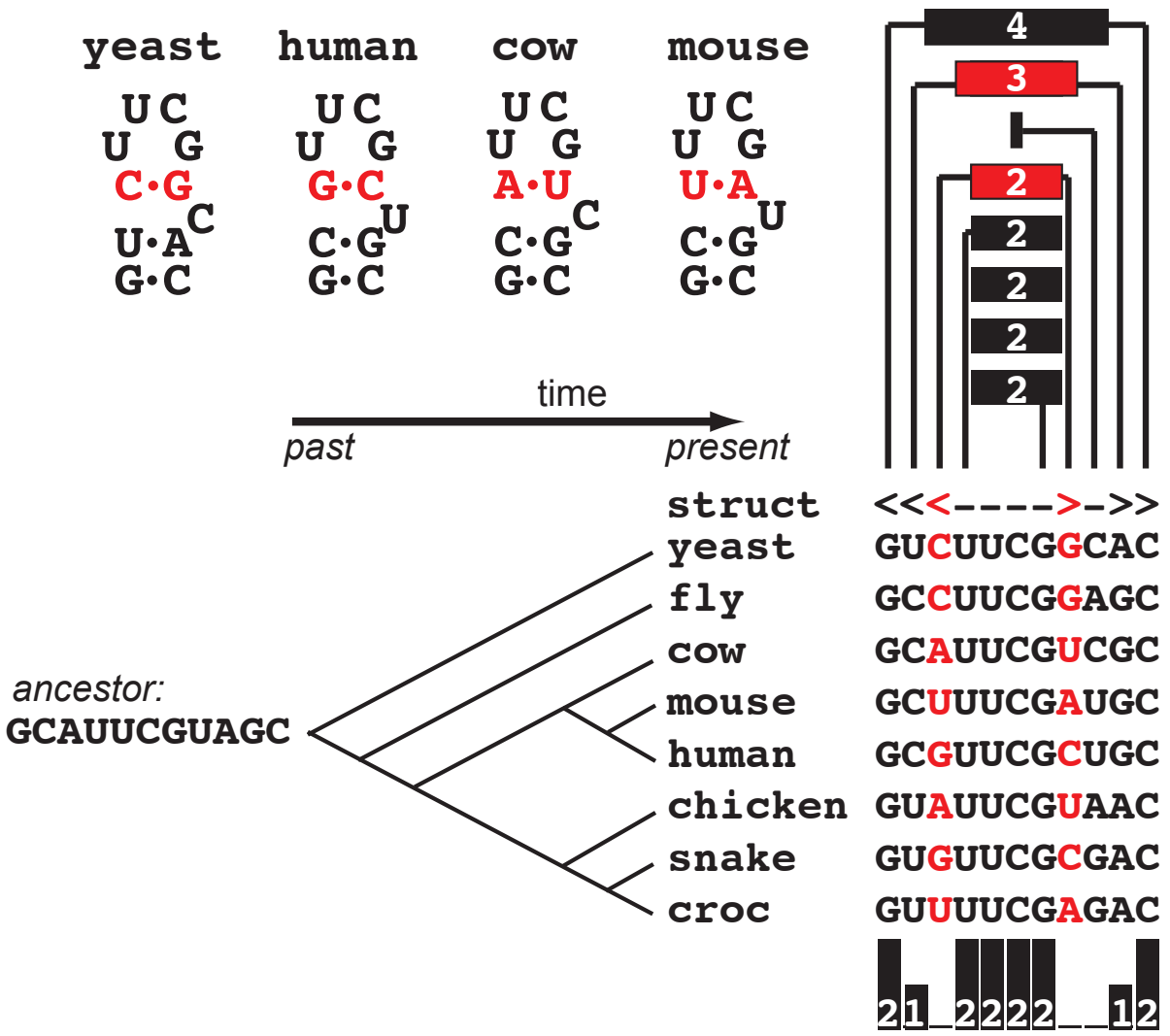
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ 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

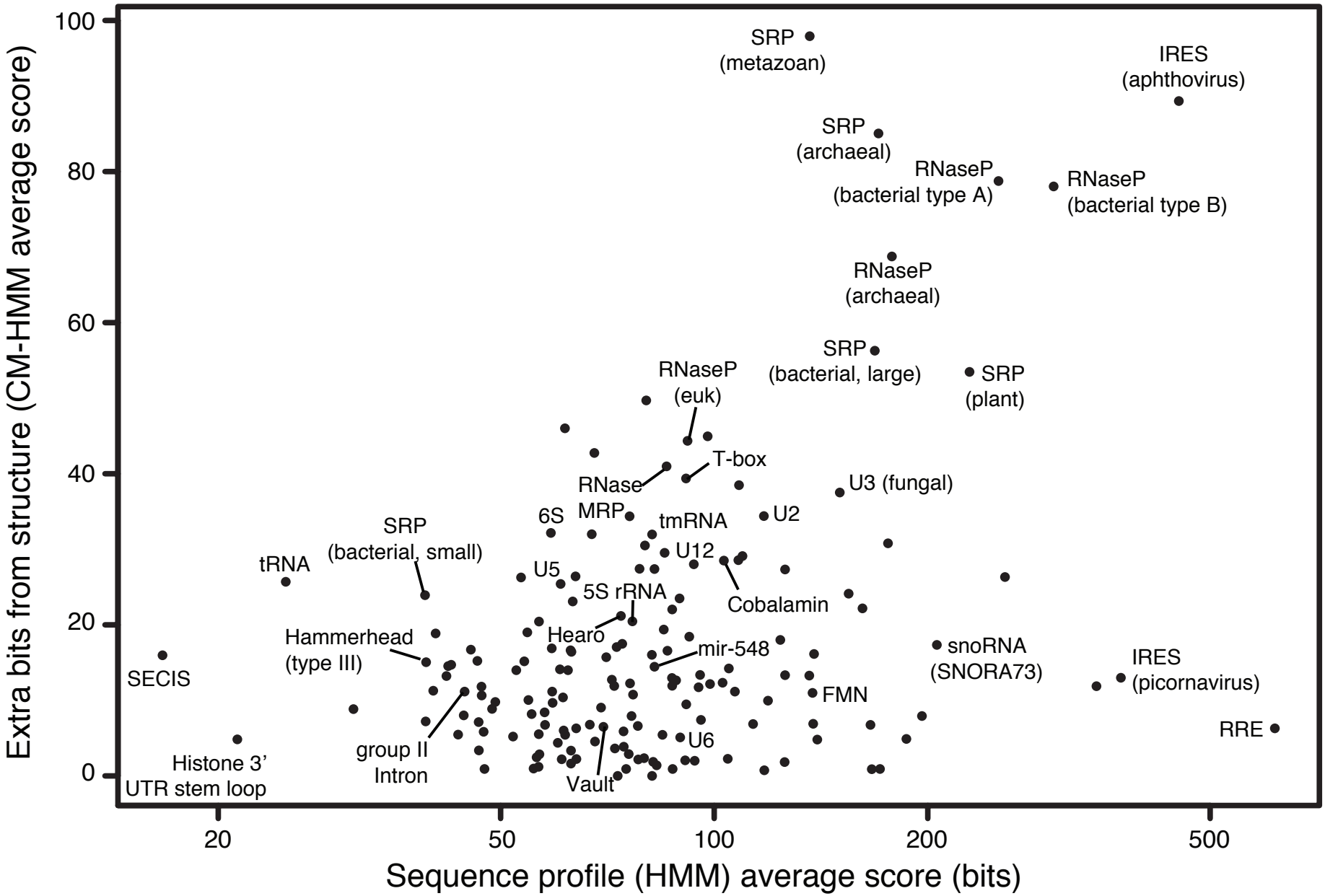
$\log_2 (16/4) = 2$ bits

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

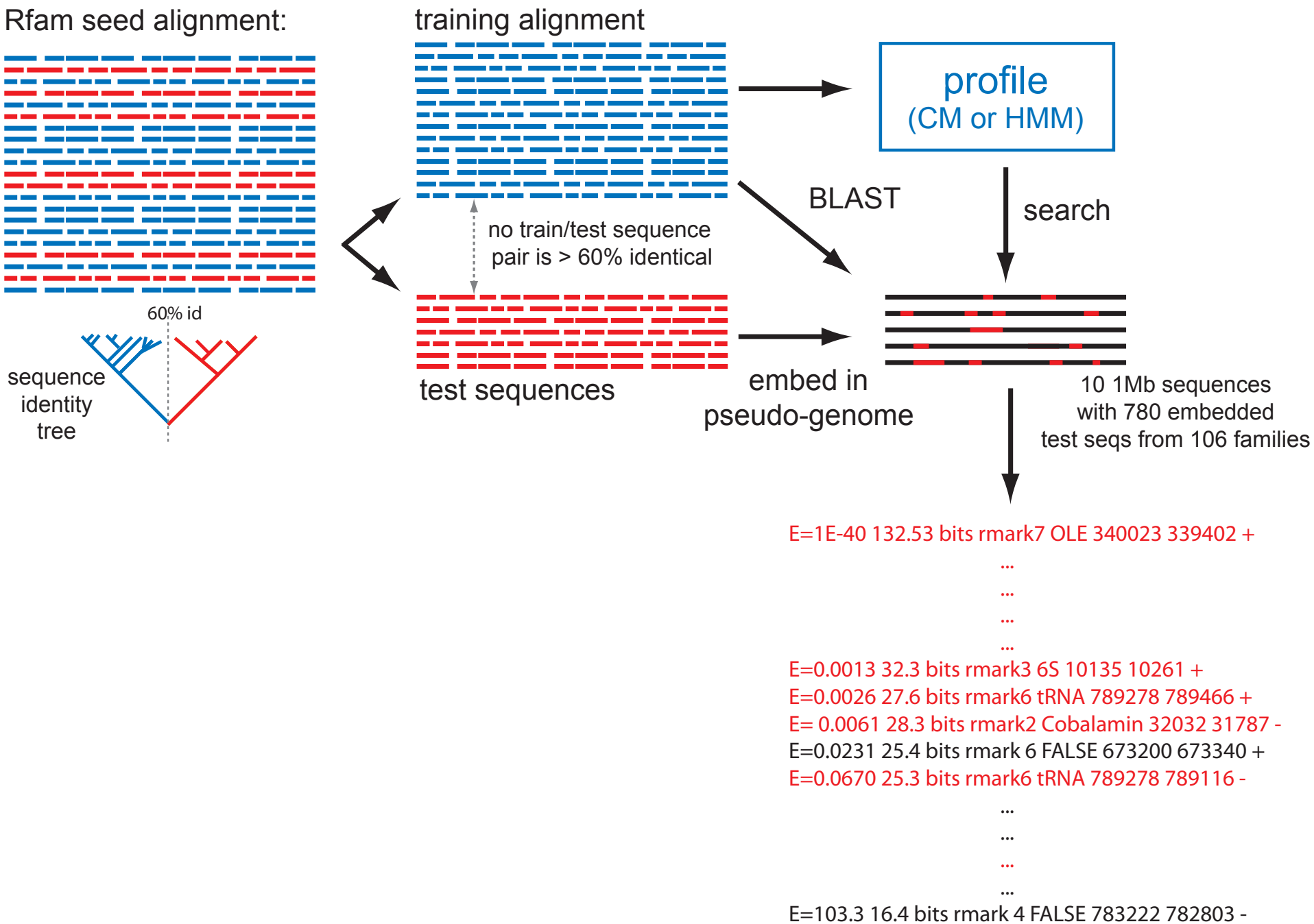
Nawrocki EP, Eddy SR Bioinformatics, 29:2933-2935, 2013.

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

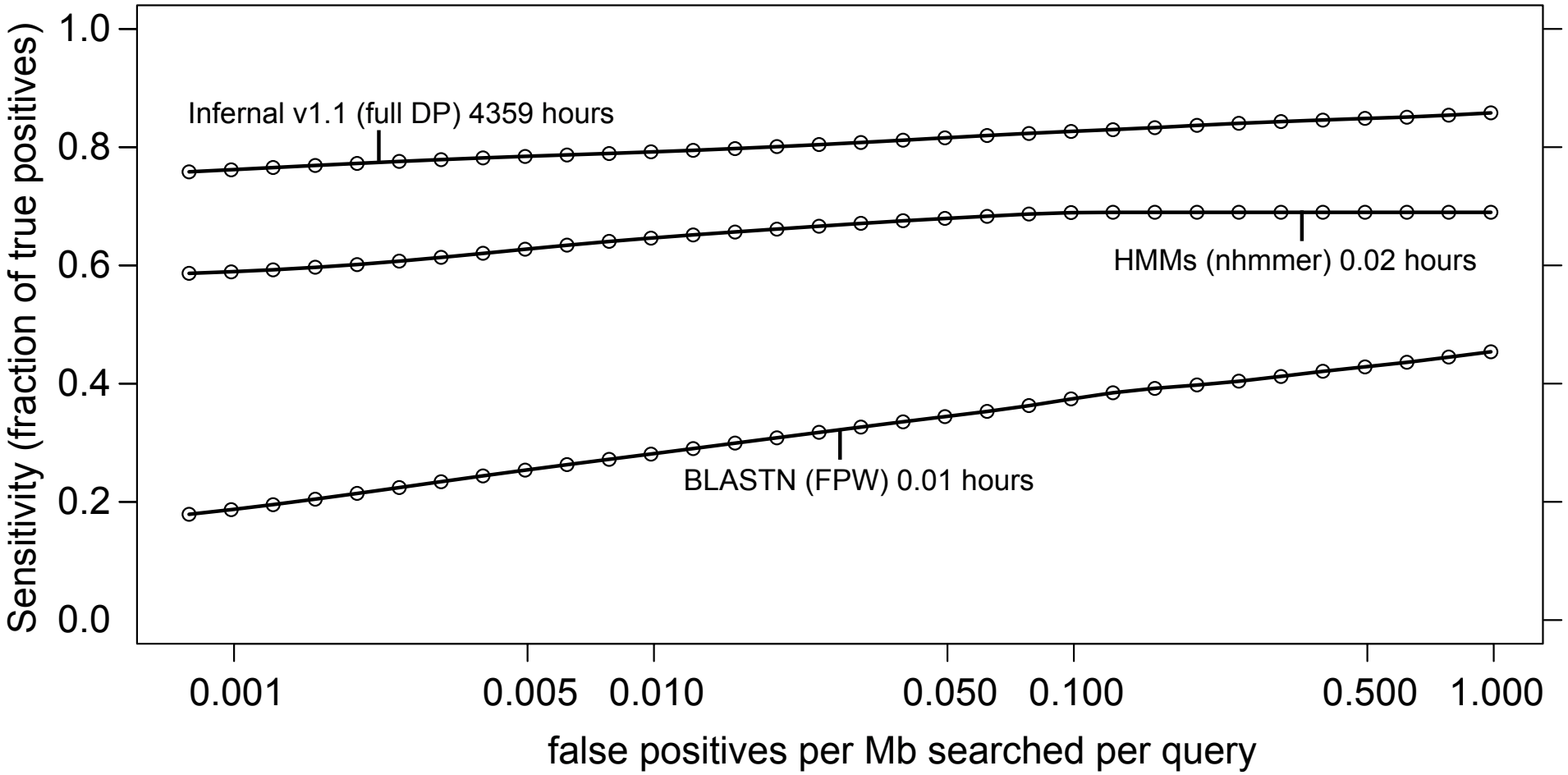
Is the added complexity worth it?

RMARK: a challenging internal RNA homology search benchmark

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



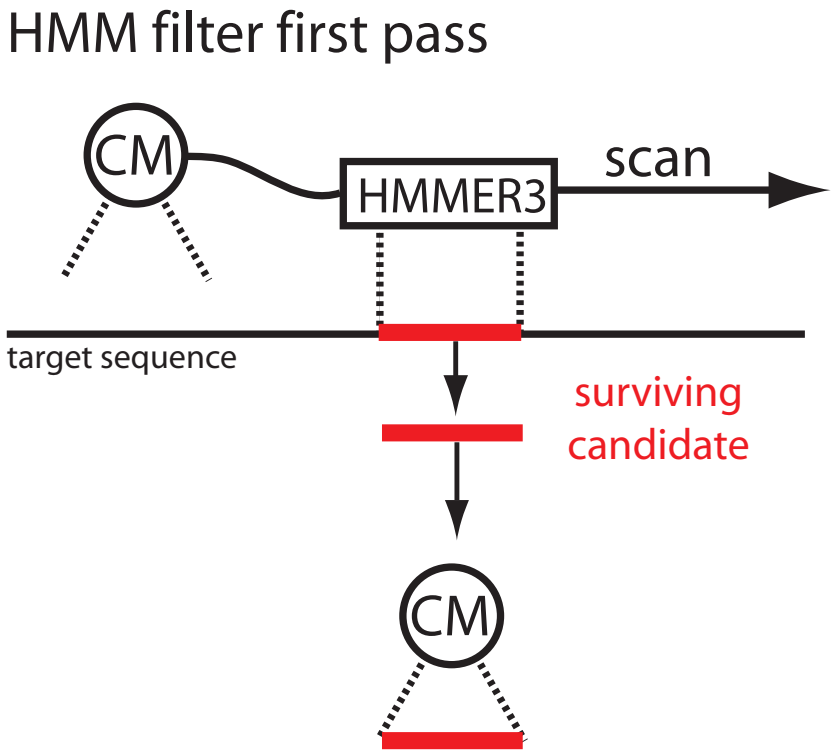
Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)



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

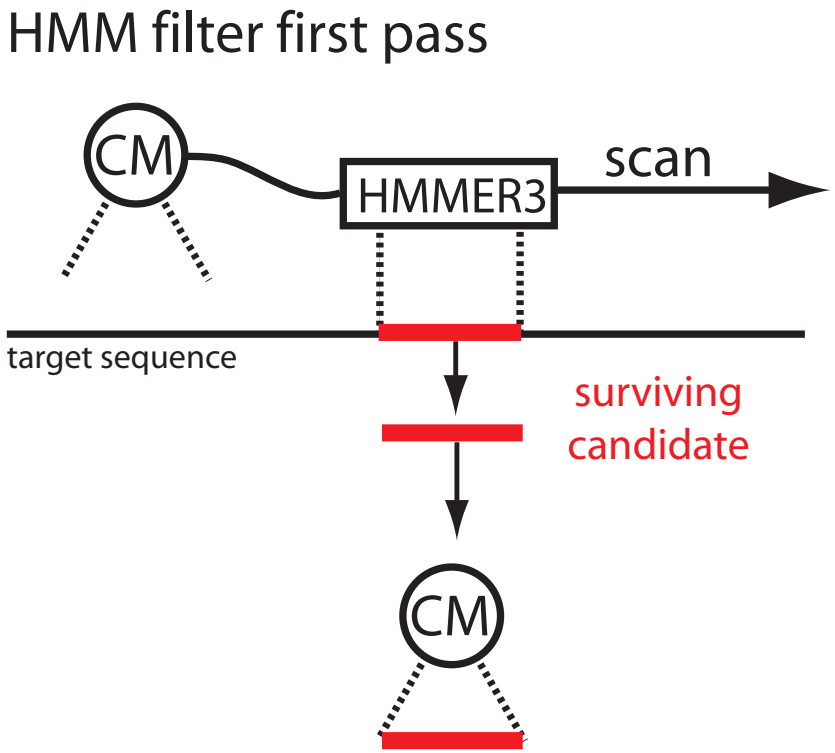
*Freyhult EK, Bollback JP, Gardner PP. Genome Res. 2007 17: 117-125.

Filter target database using profile HMMs*



*Weinberg, Ruzzo, RECOMB, 243-251, 2004

Filter target database using profile HMMs*

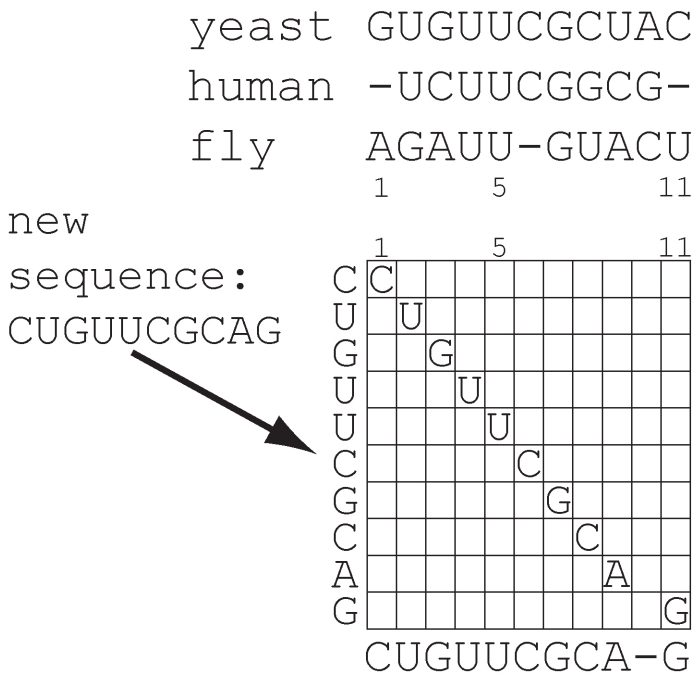


- Even if we filter out 99% of the database (for up to 100X acceleration), searches will still be too slow.

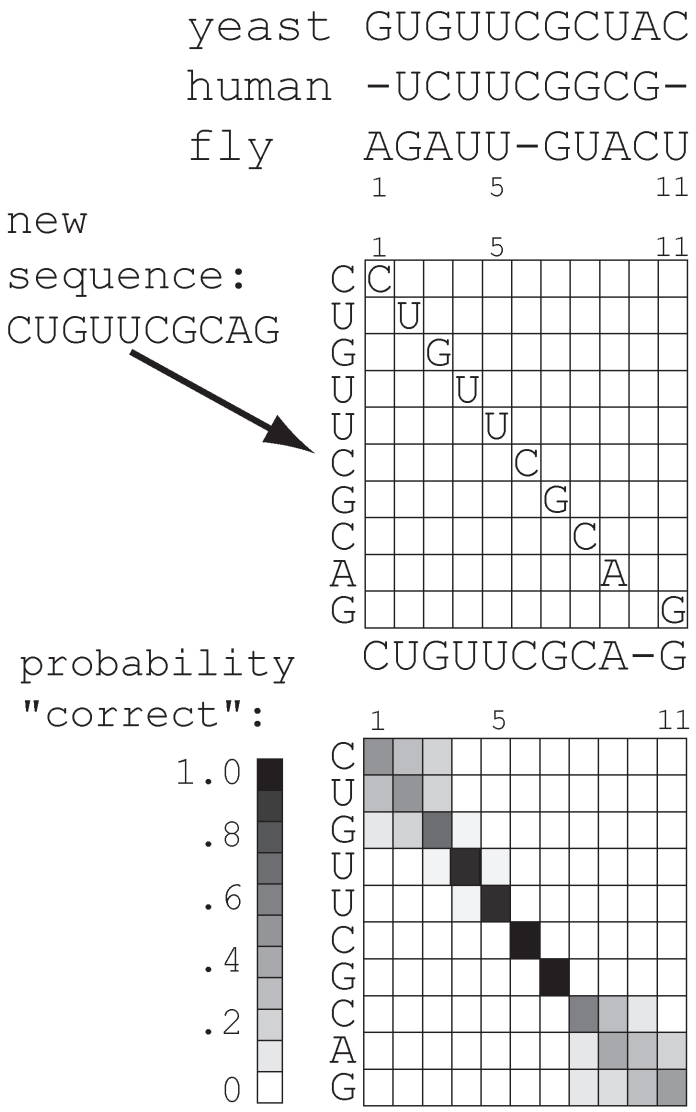
- CM step needs to be accelerated.

*Weinberg, Ruzzo, RECOMB, 243-251, 2004

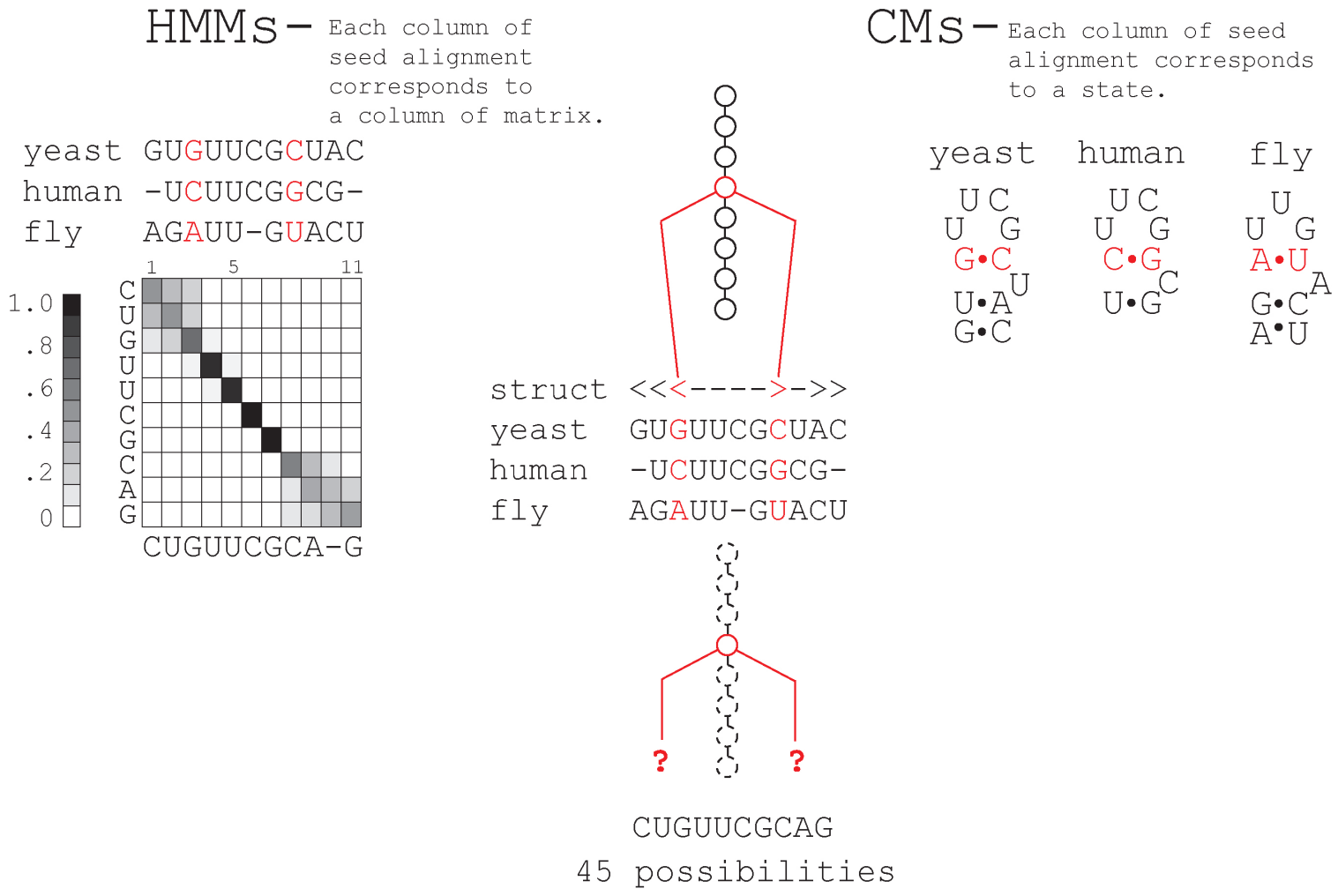
Accelerating CM alignment step 1: align sequence with HMM



Accelerating CM alignment step 2: HMM posterior decoding to get confidence estimates

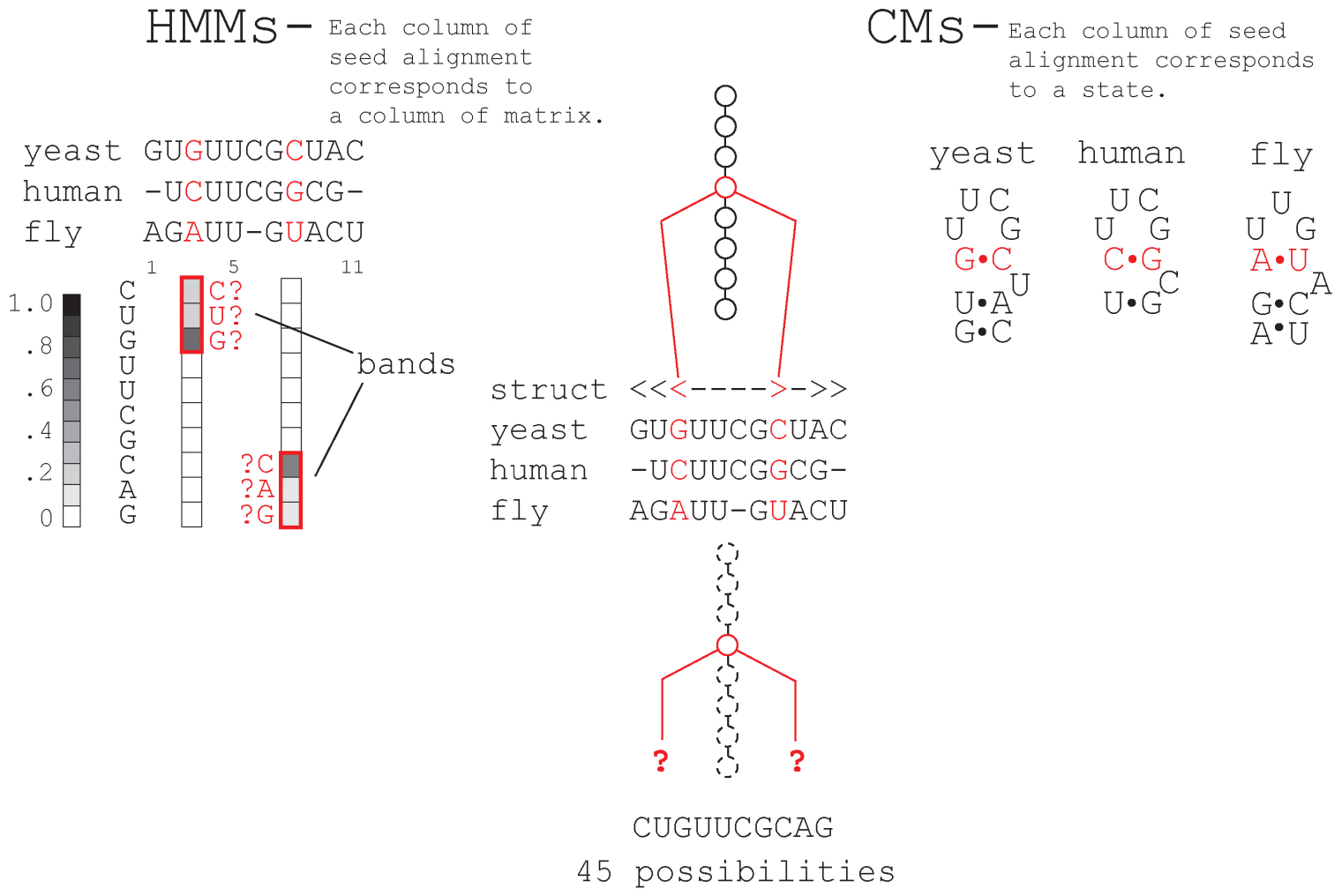


Accelerating CM alignment step 3: use HMM alignment confidence to constrain CM alignment*



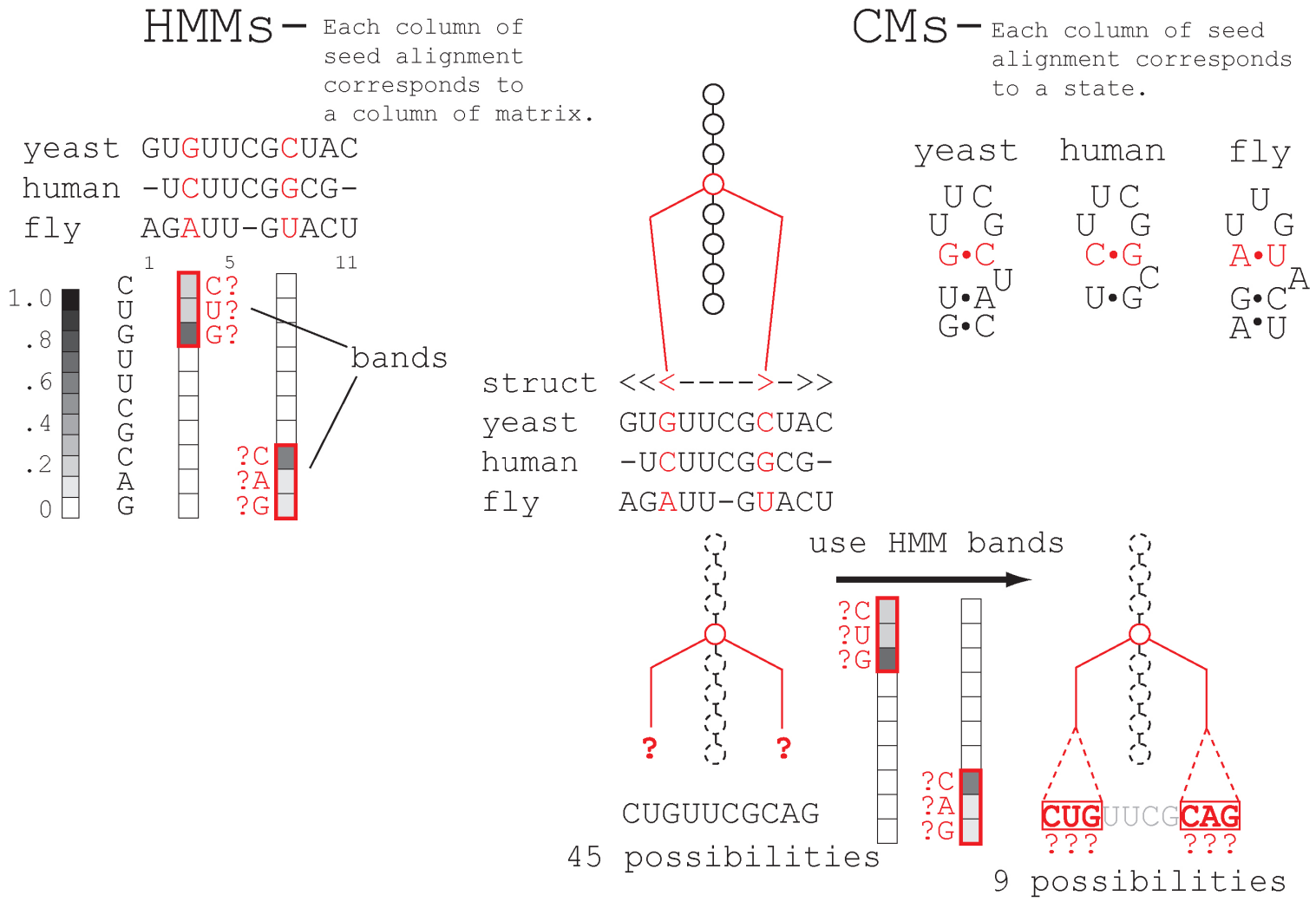
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

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



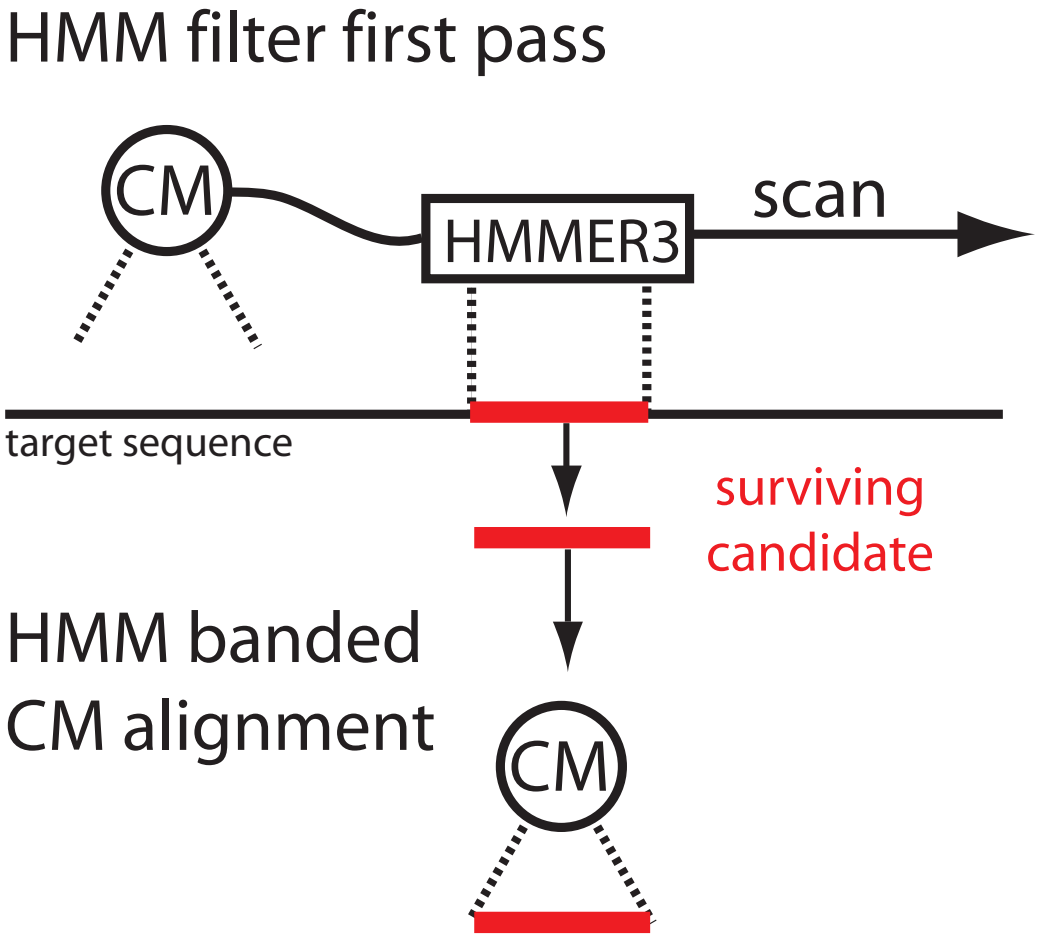
*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

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

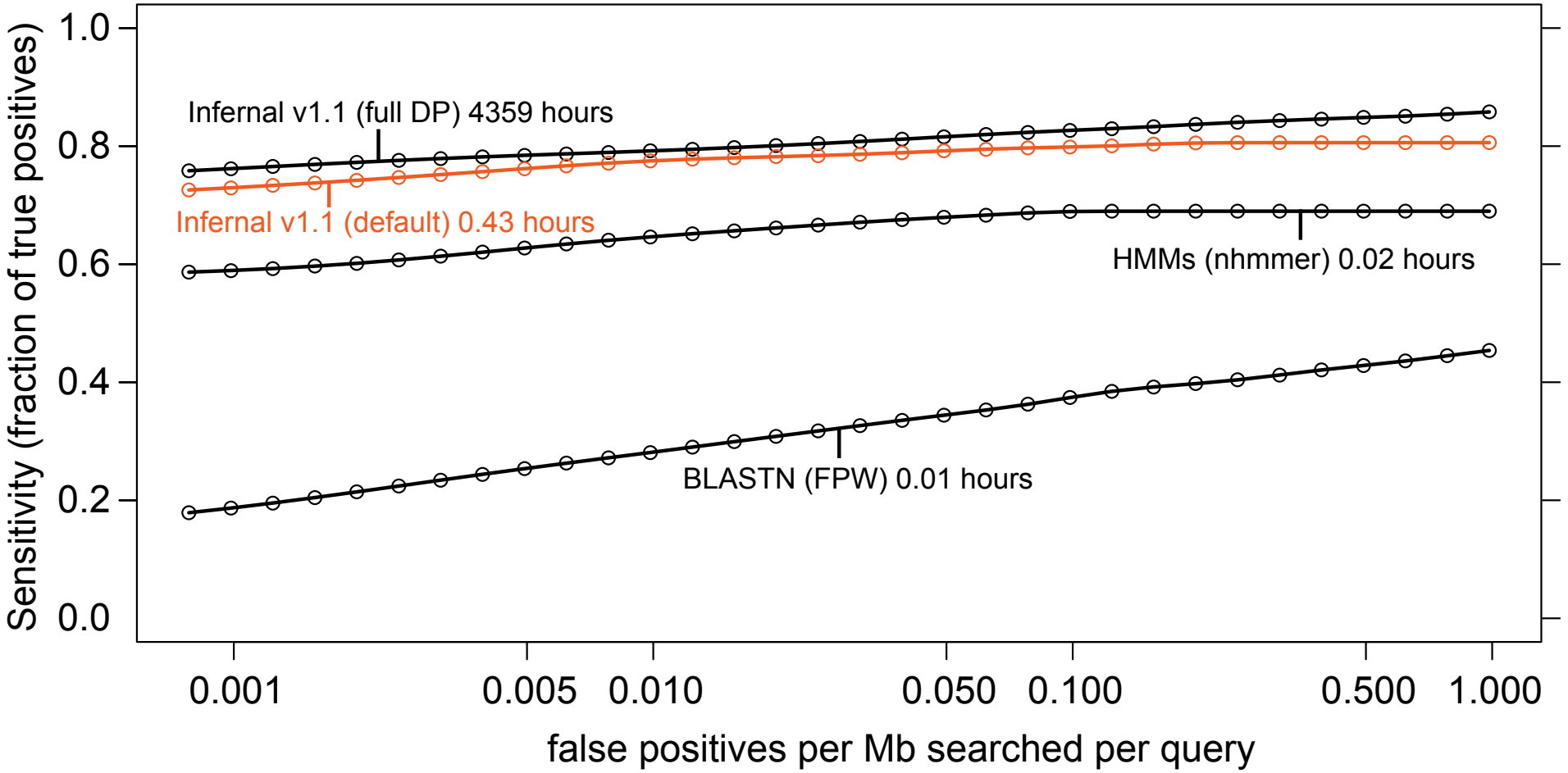


*M. P. Brown. Proc. Int. Conf. ISMB, 8:5766, 2000.

Use HMMs as filters and to constrain CM alignment

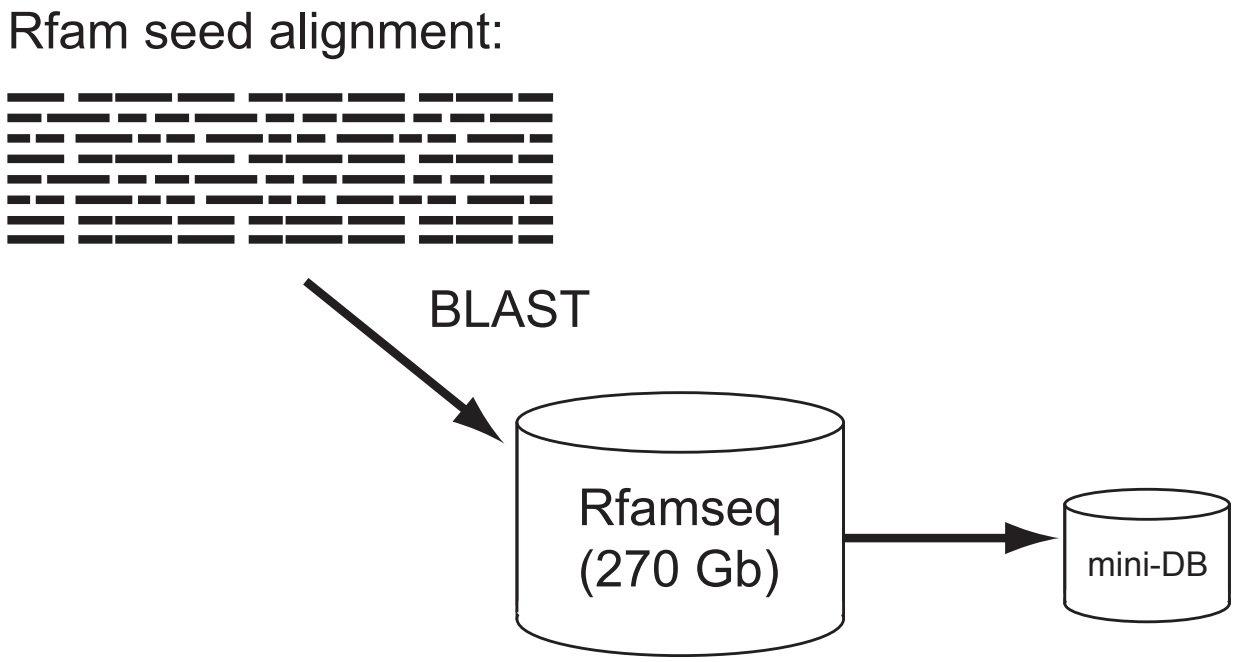


HMM-based acceleration makes Infernal 10,000 times faster

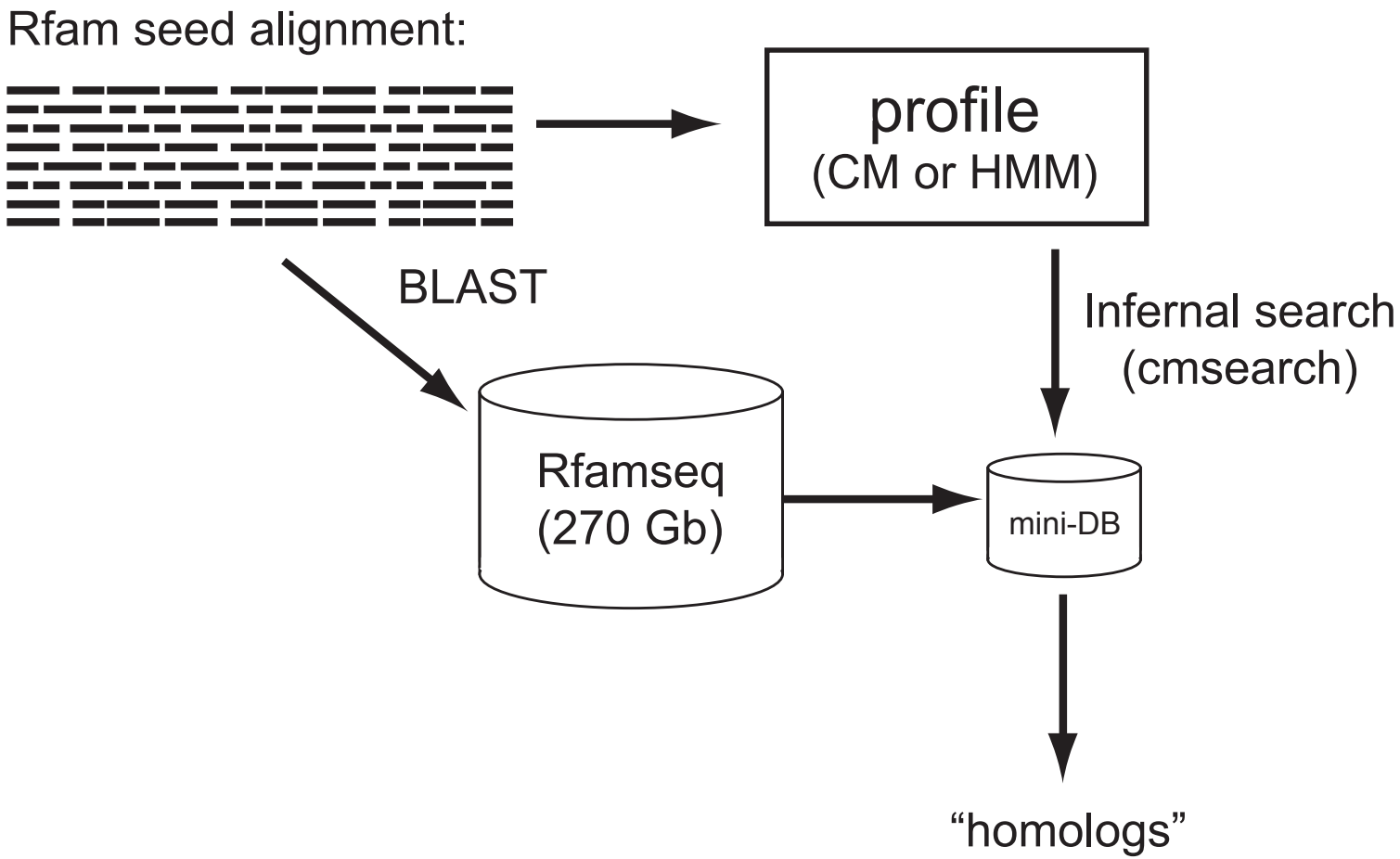


Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

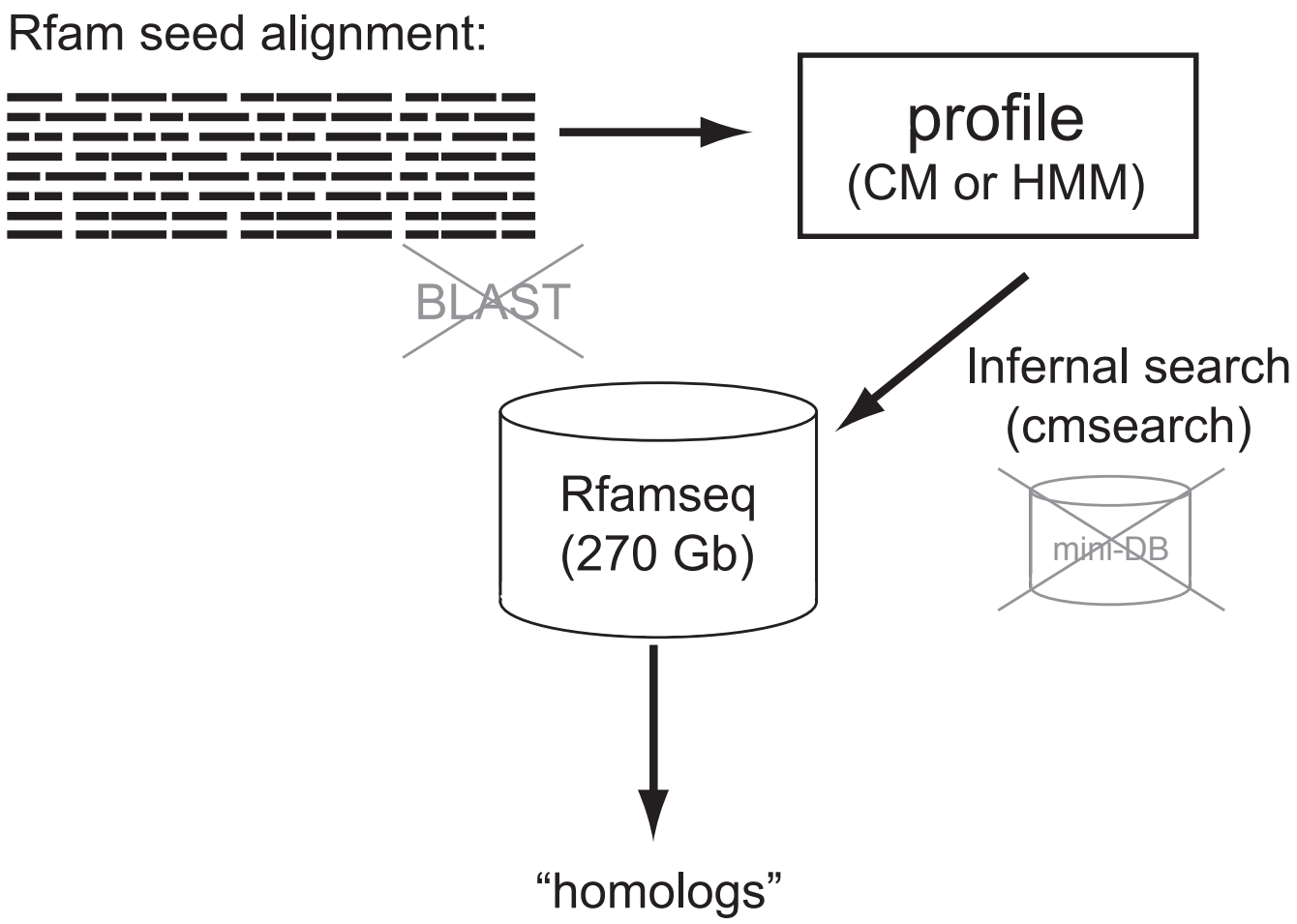
Rfam used BLAST filters from 2003 to 2012



Rfam used BLAST filters from 2003 to 2012



Rfam 12.0 (2014)* first release without BLAST filtering



*Nawrocki, Burge et. al, NAR 2014, in press

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 2014, in press

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