Modeling Structural RNA Families with Infernal

Eric Nawrocki

Sean Eddy's Lab

Howard Hughes Medical Institute
Janelia Research Campus

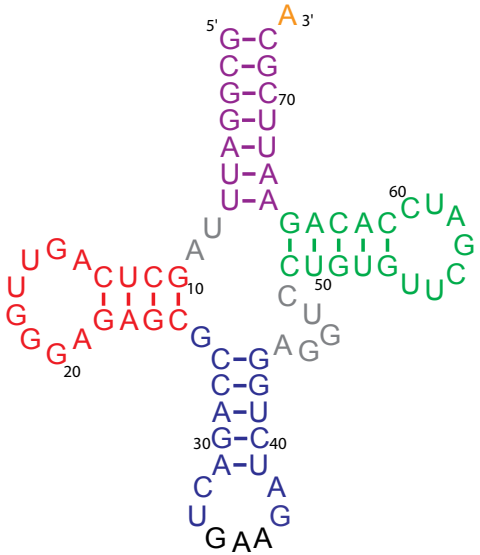
Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



3-dimensional structure

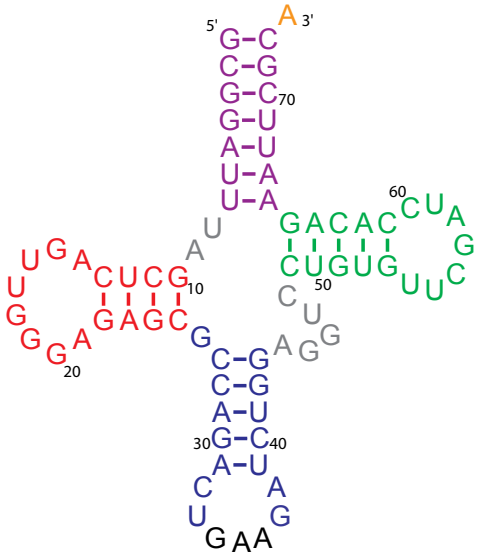
Most proteins and RNAs adopt a conserved 3-dimensional structure that is responsible for their function in the cell

Three representations of a transfer RNA:

Primary sequence

GCGGAUUUAGCUCAGUUGGG
AGAGCGCCAGACUGAAGAUC
UGGAGGUCCUGUGUUCGAUC
CACAGAAUUCGCA

Secondary structure



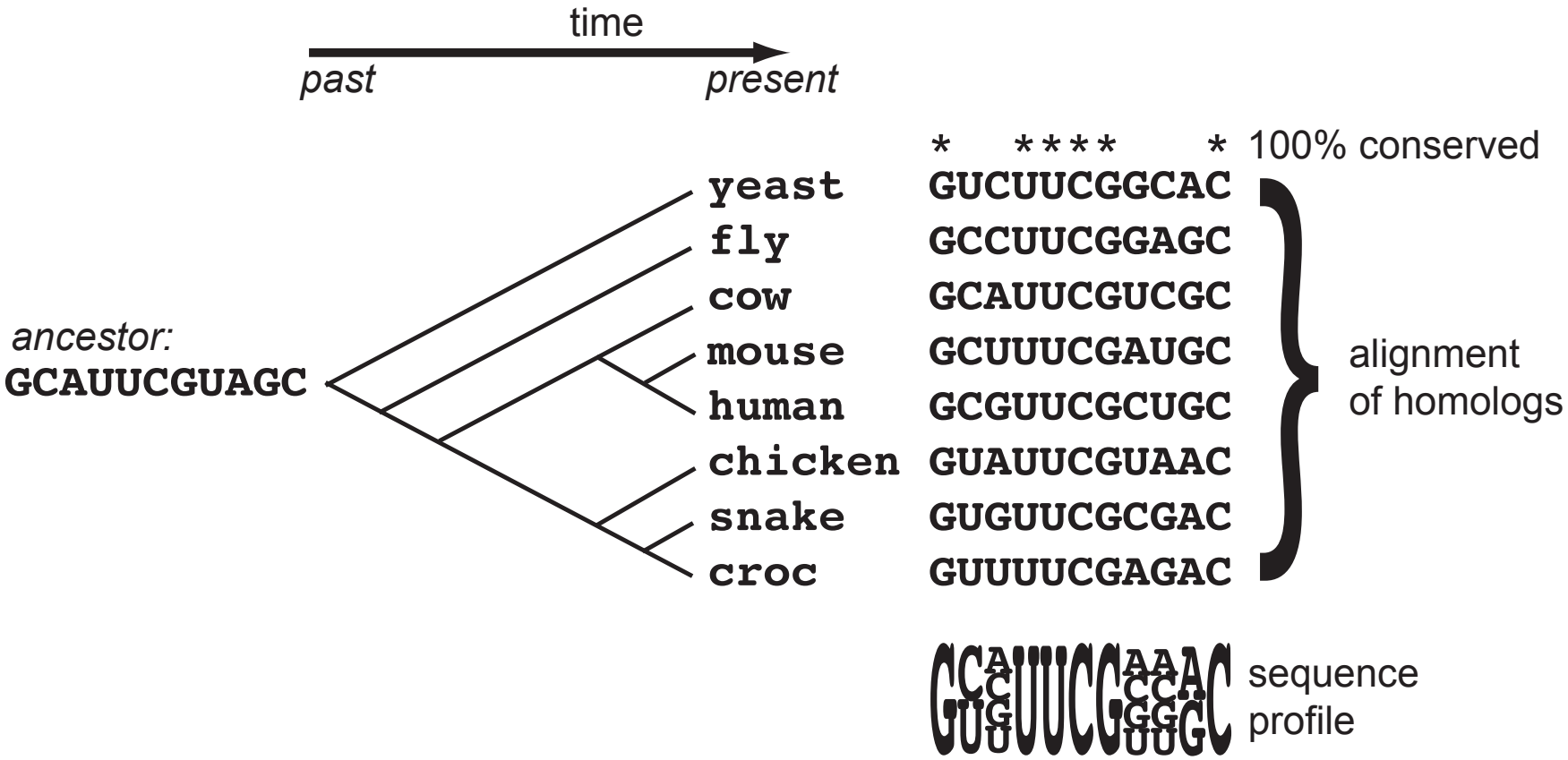
3-dimensional structure

BLAST: given a single sequence, search genomes for similar sequences.

Homologous proteins and RNAs conserve different sequence and structural features to different degrees.

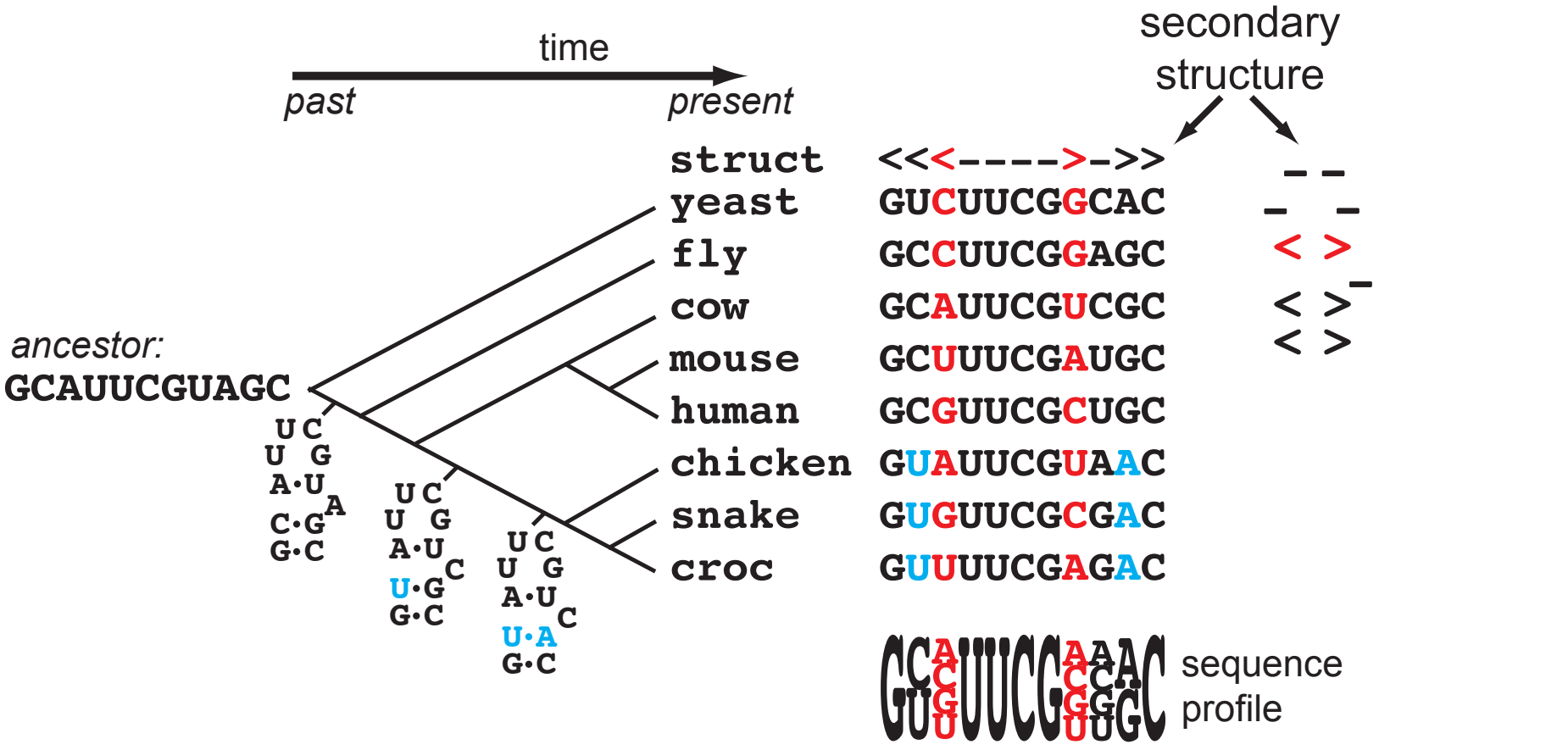
Sequence conservation provides information for homology searches

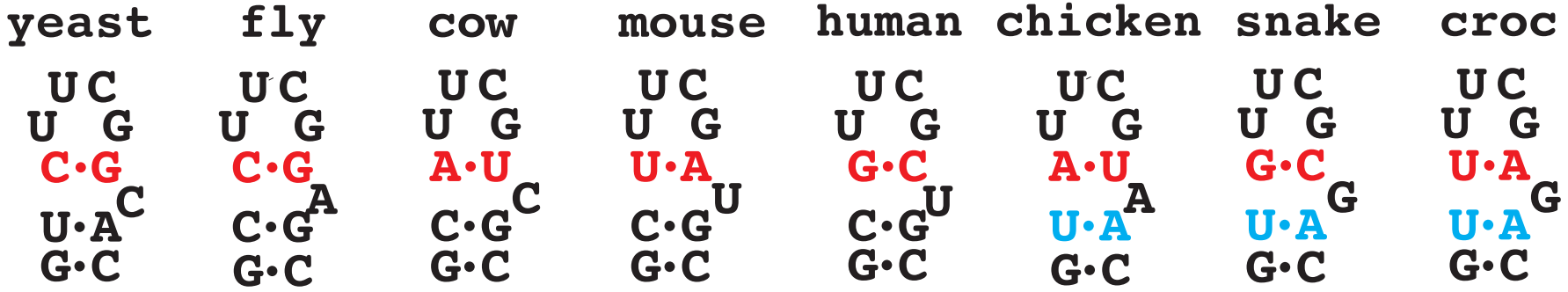
Conservation levels vary across alignment columns.



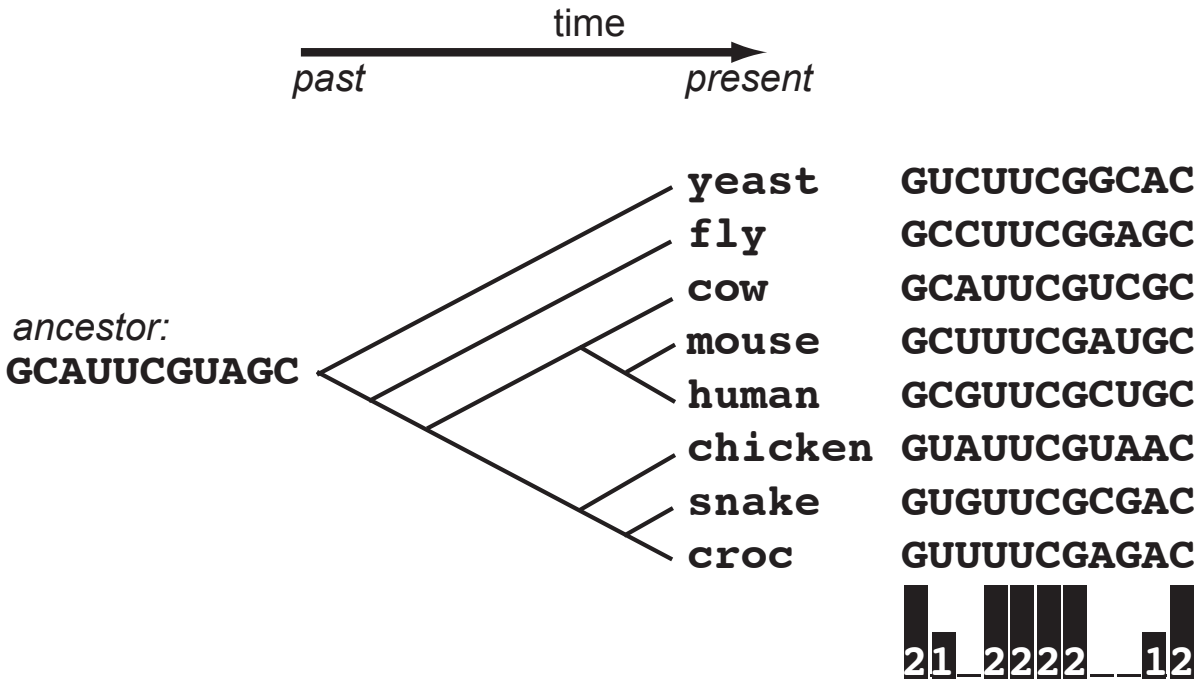
Structure conservation provides additional information

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





Amount of information in a profile can be measured in bits



Each column is independent

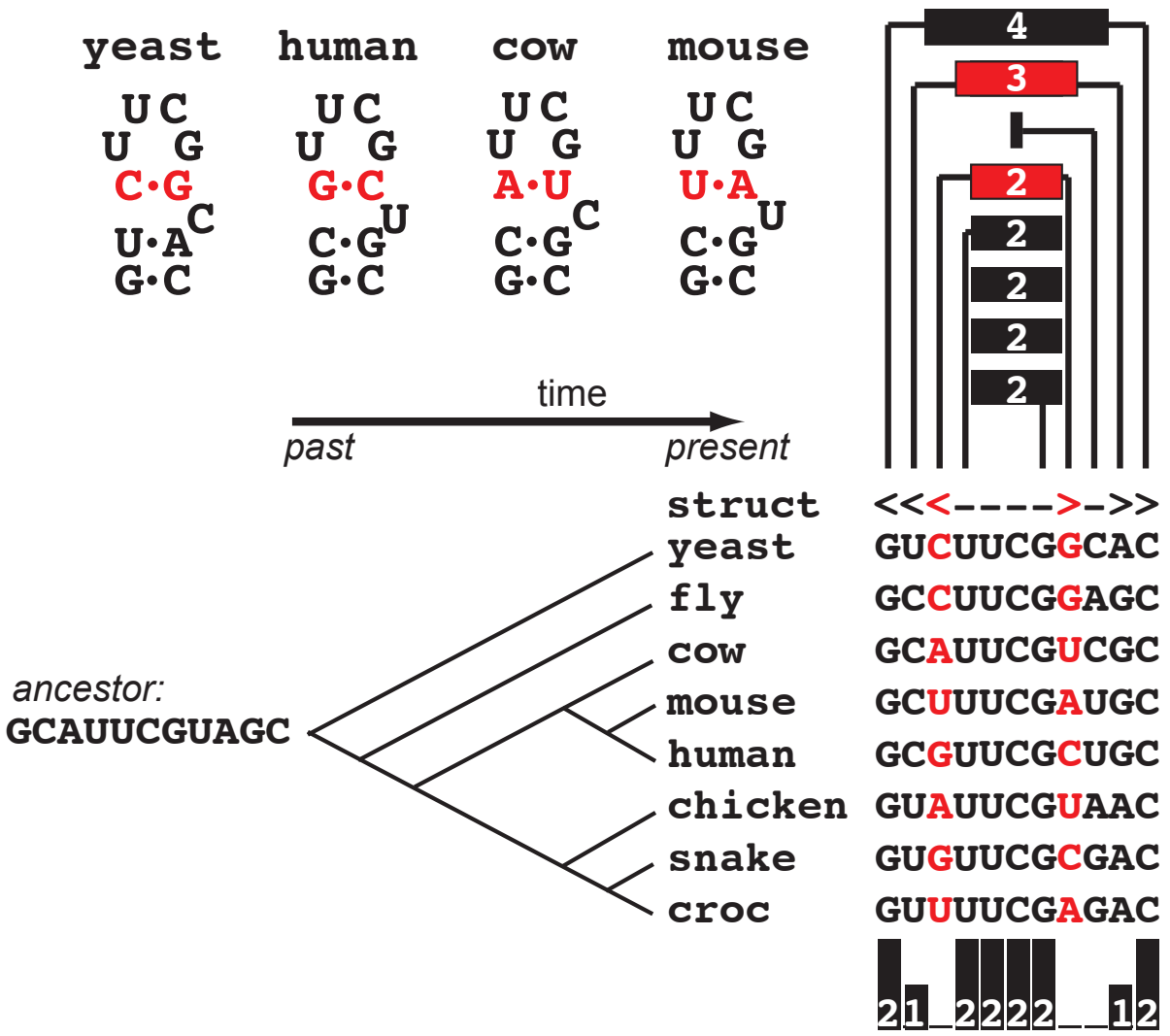
Maximum information is 2 bits per position, if column is 100% conserved.

$\log_2 (4/1) = 2$ 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

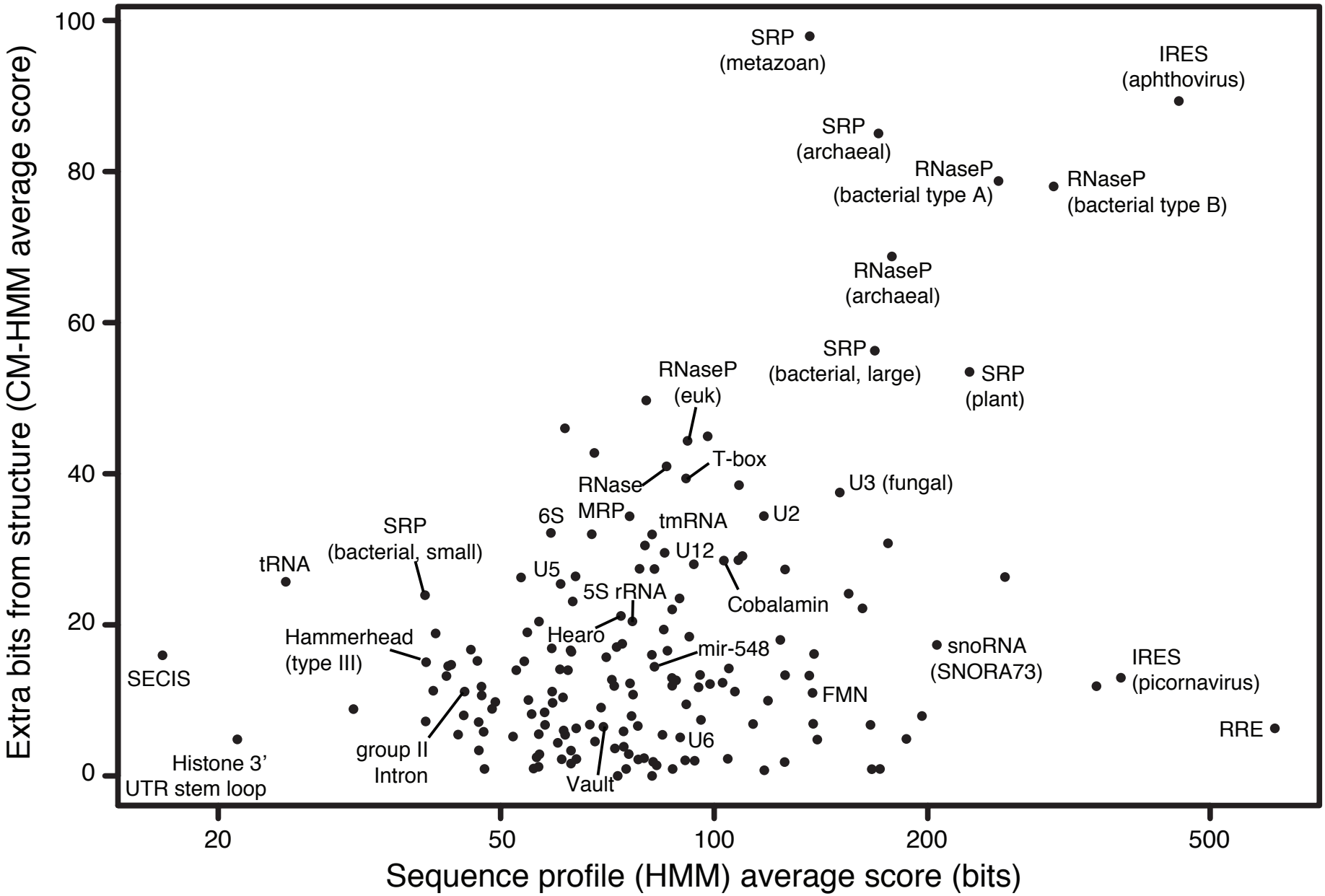
$\log_2 (16/4) = 2$ bits

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

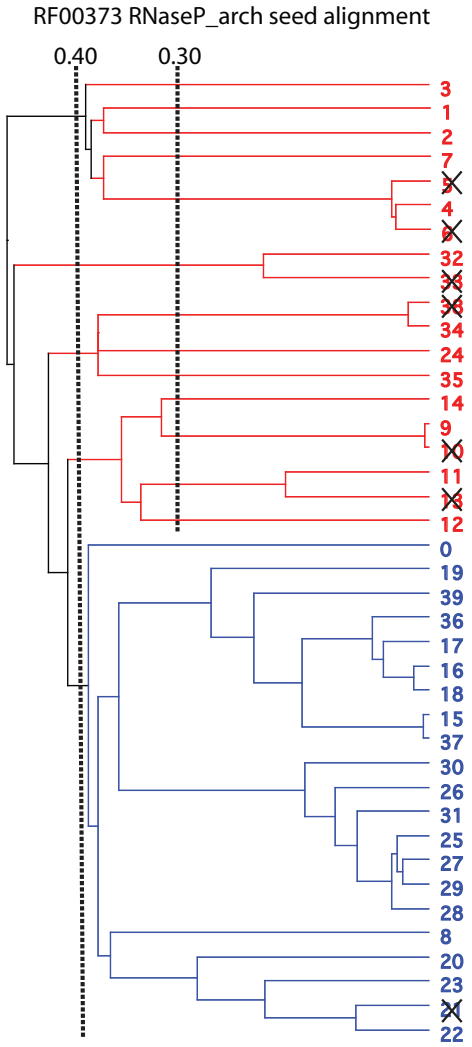
Nawrocki EP, Kolbe DL, Eddy SR Bioinformatics, 25 (10):1335-1337, 2009.

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

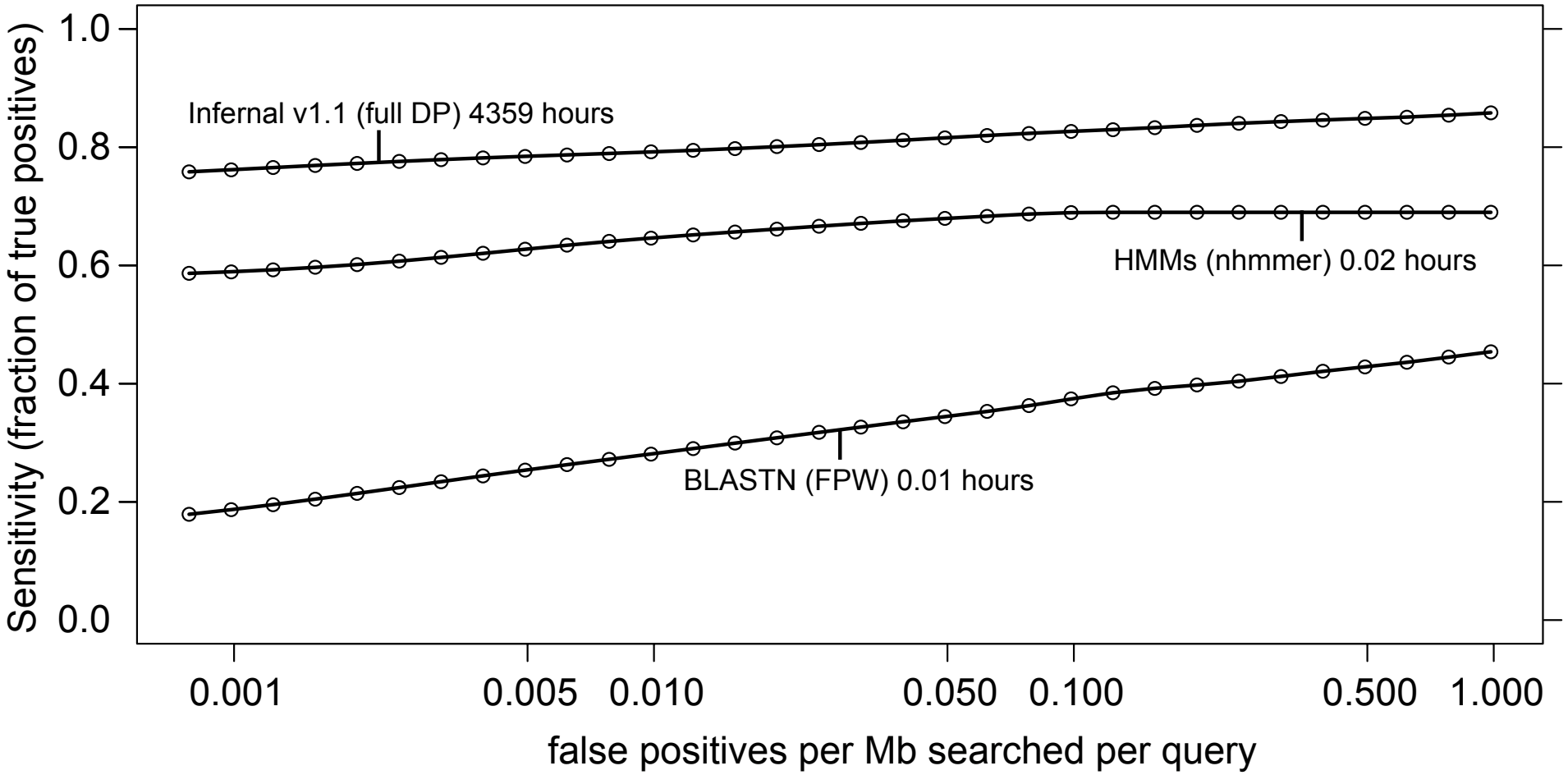
Is the added complexity worth it?
RMARK: a challenging internal RNA homology search benchmark for use during Infernal development

- RMARK construction - for each of the 1446 Rfam 10 seed alignments:
  - 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
    - at least five sequences are in the training set
  - filter test set so no two test seqs $> 70\%$ identical
  - 106 families qualify, with 780 test sequences
  - test seqs are embedded in a 10 Mb pseudo-genome of "realistic" base composition

Example:



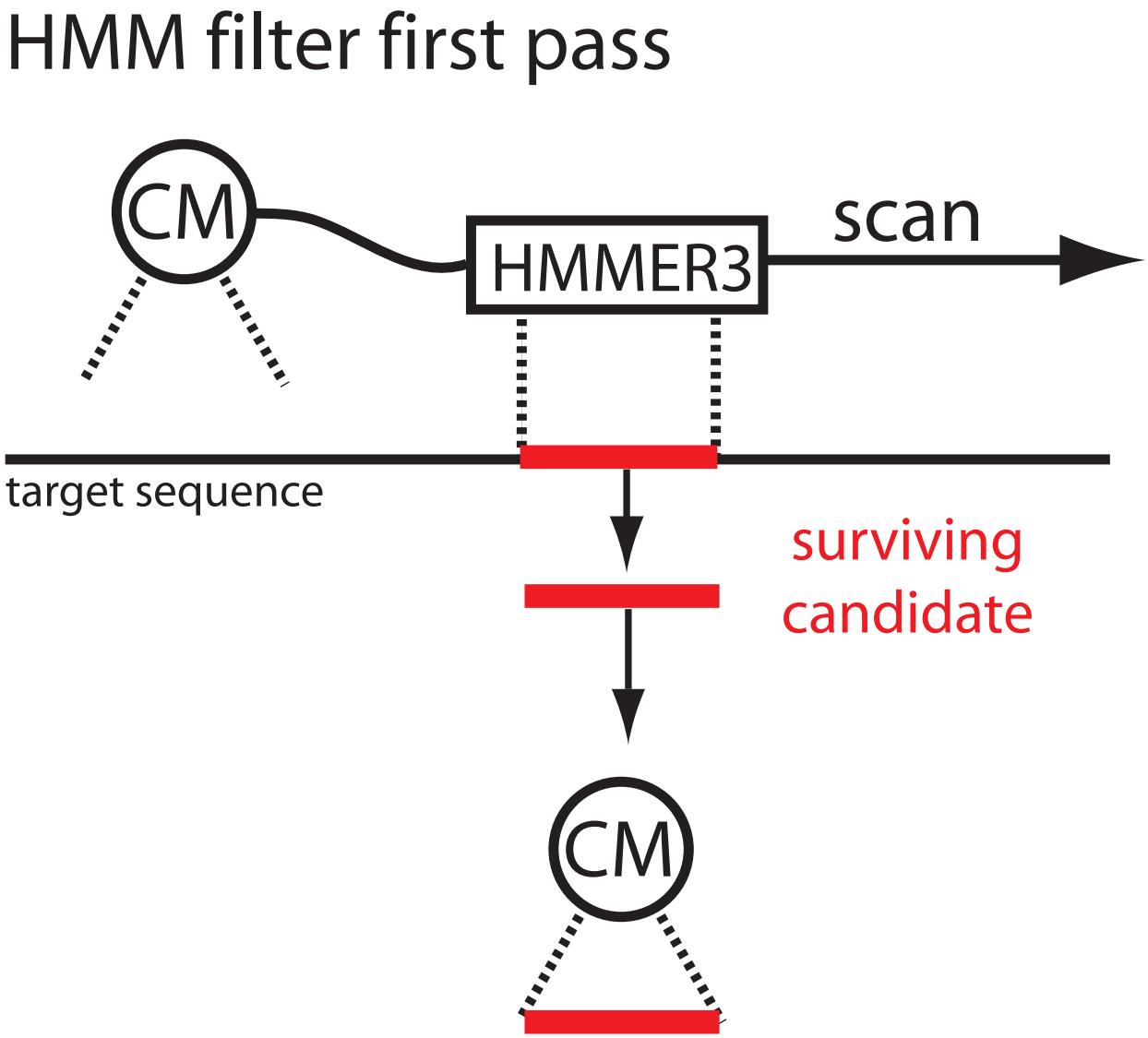
Infernal outperforms primary-sequence based methods on our benchmark (and others*, not shown)



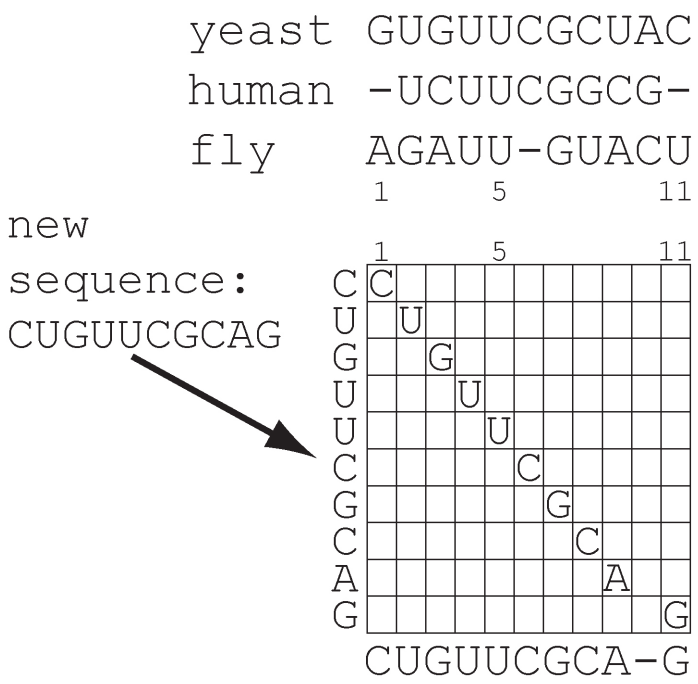
Nawrocki EP, Eddy SR. Bioinformatics, 29:2487-2489, 2013.

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

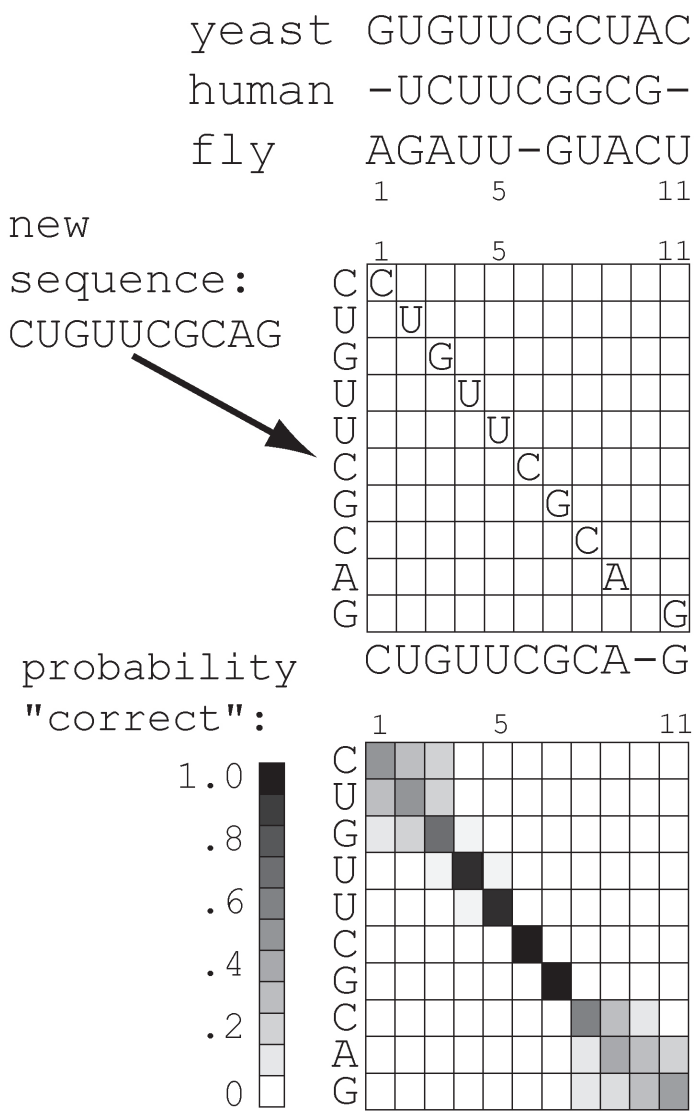
Exploit sequence-based conservation using HMM filters



Optimal alignment to an HMM

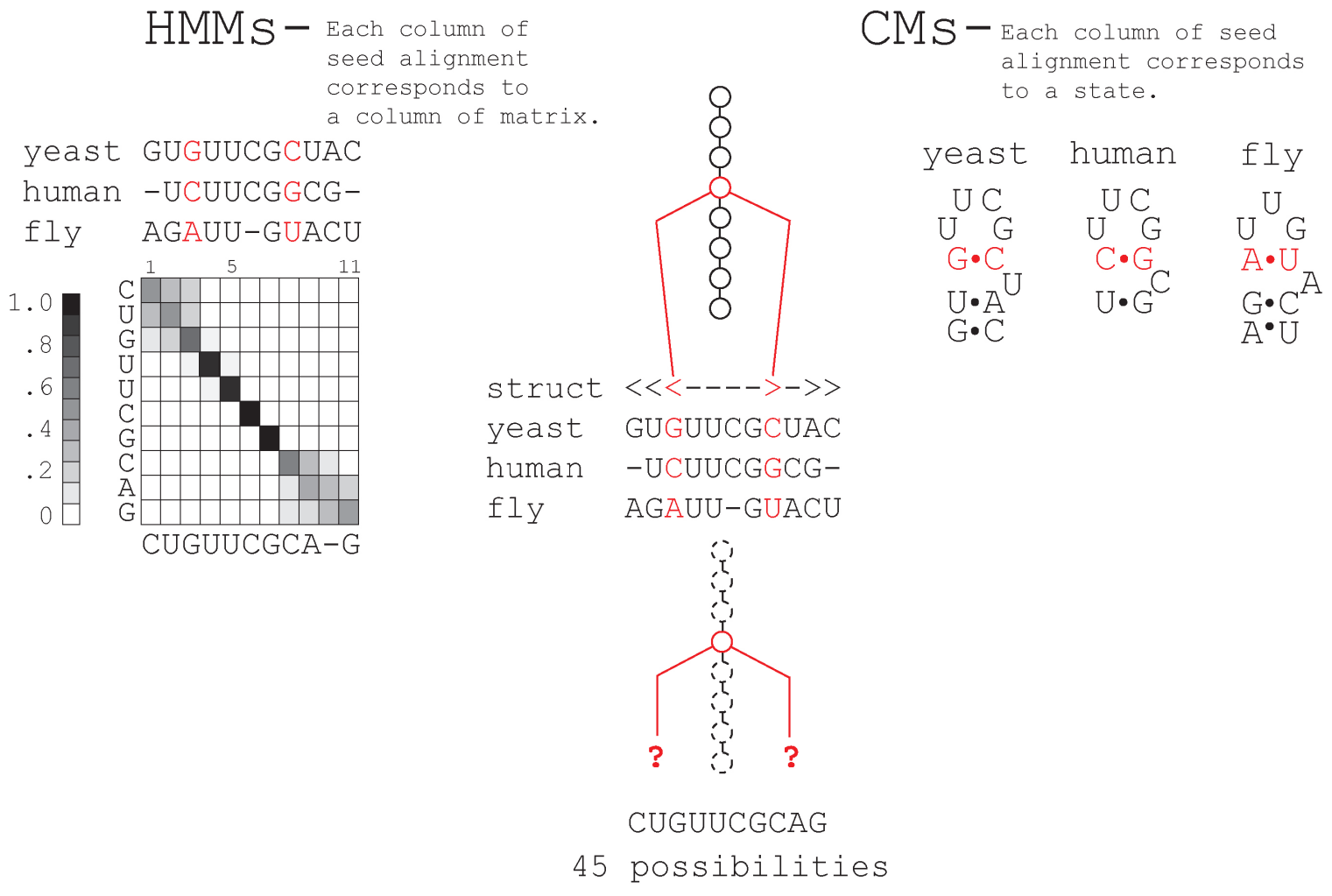


Posterior decoding of an HMM alignment gives confidence estimates



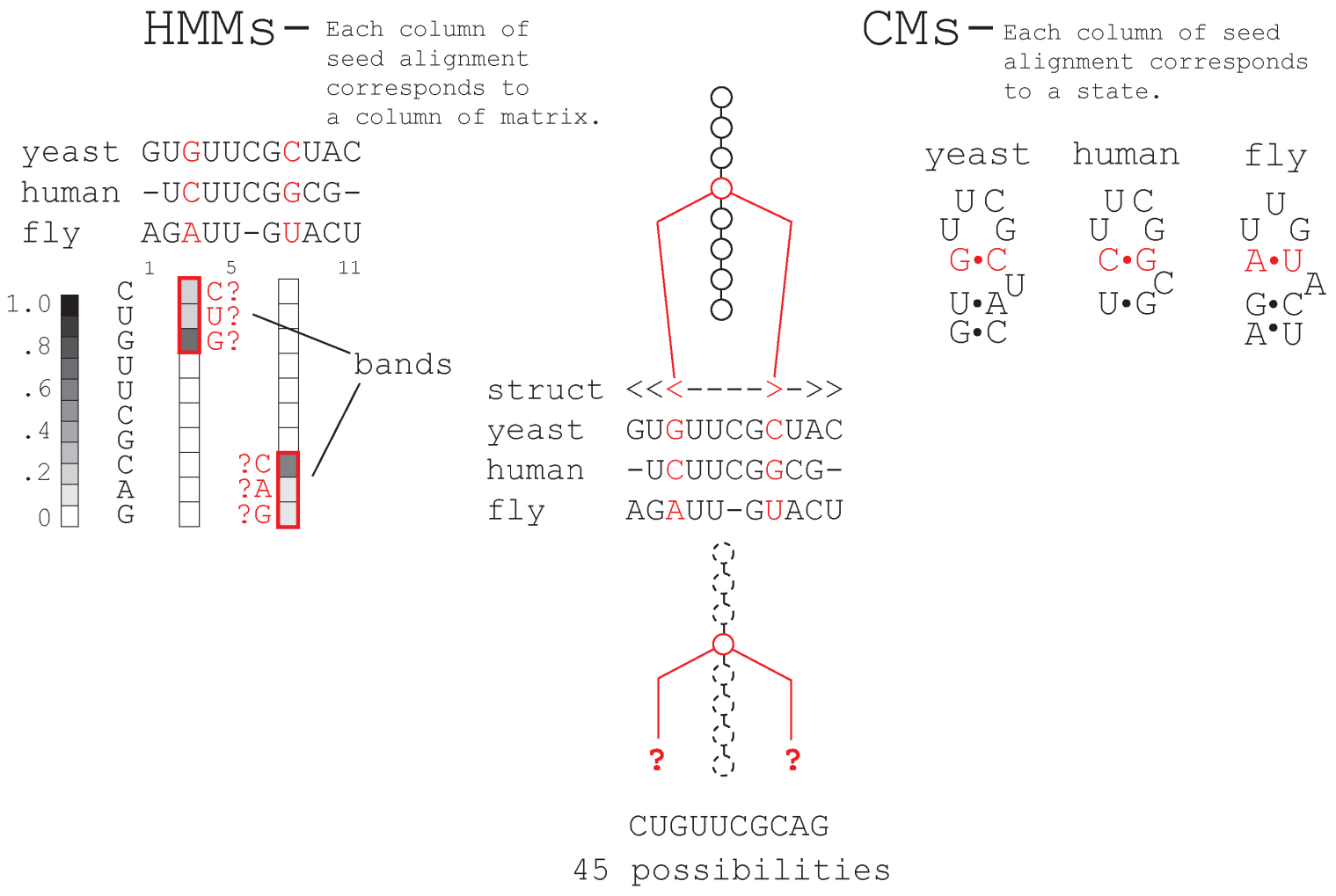
HMM bands accelerate CM alignment

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



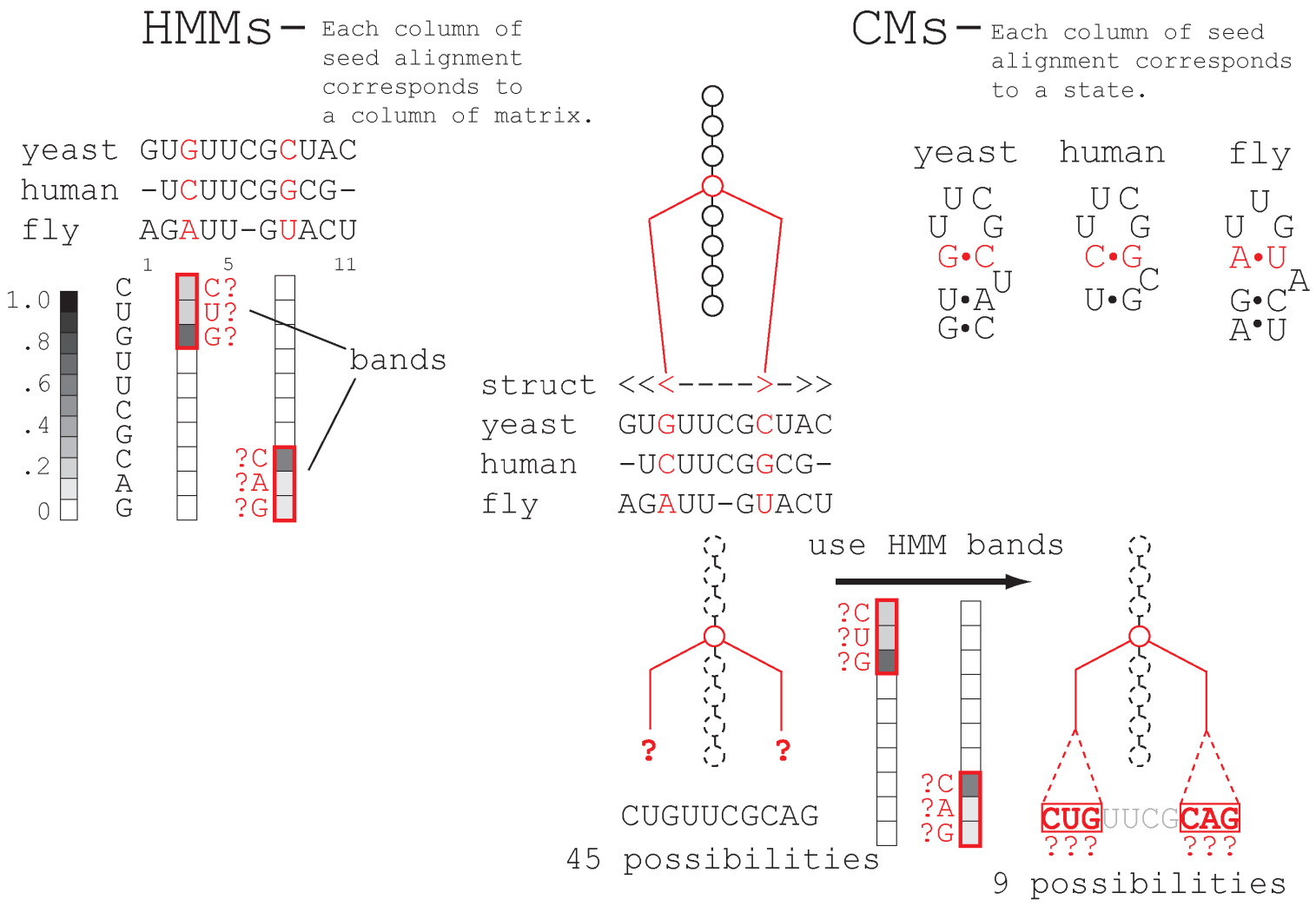
HMM bands accelerate CM alignment

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

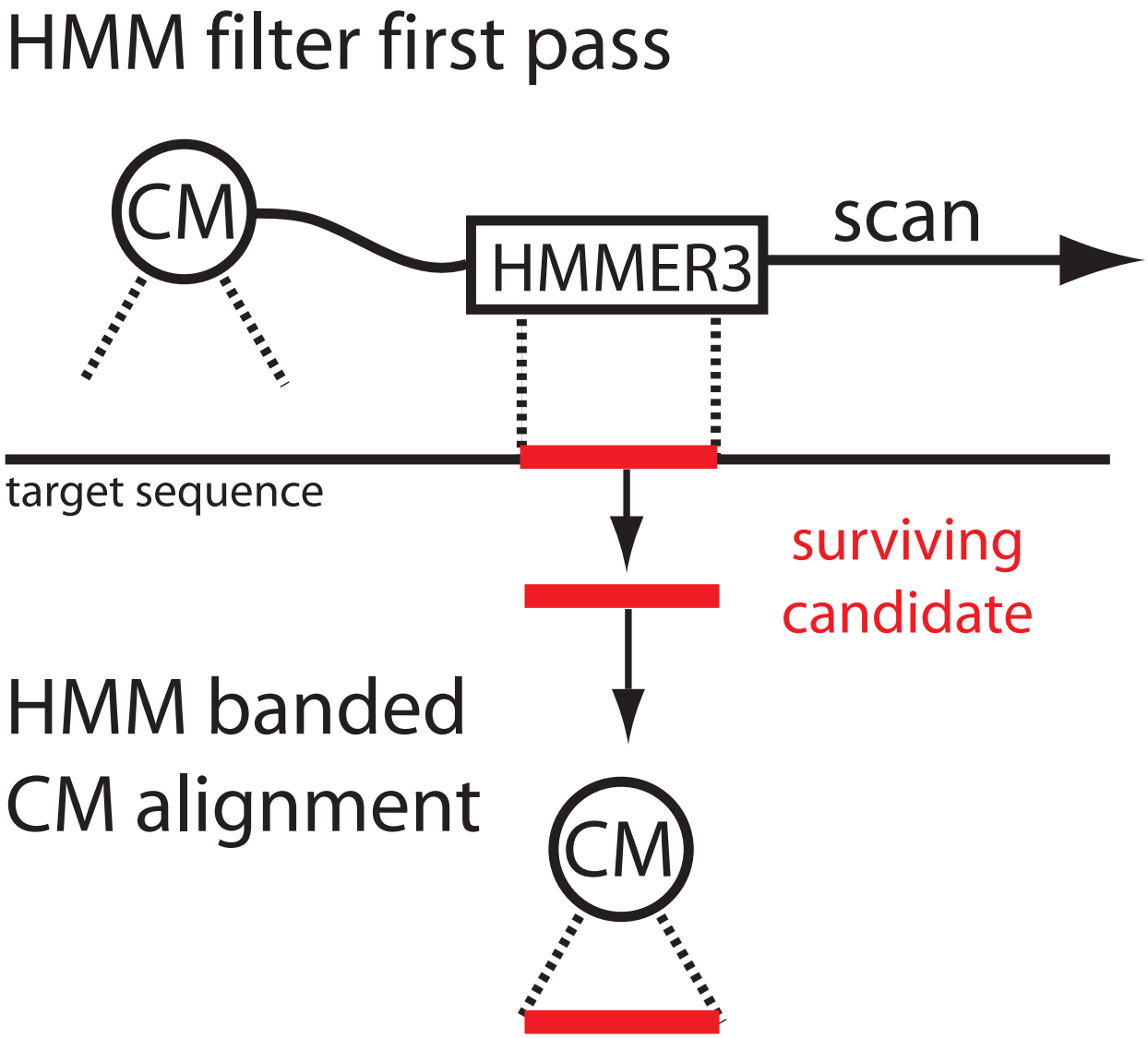


HMM bands accelerate CM alignment

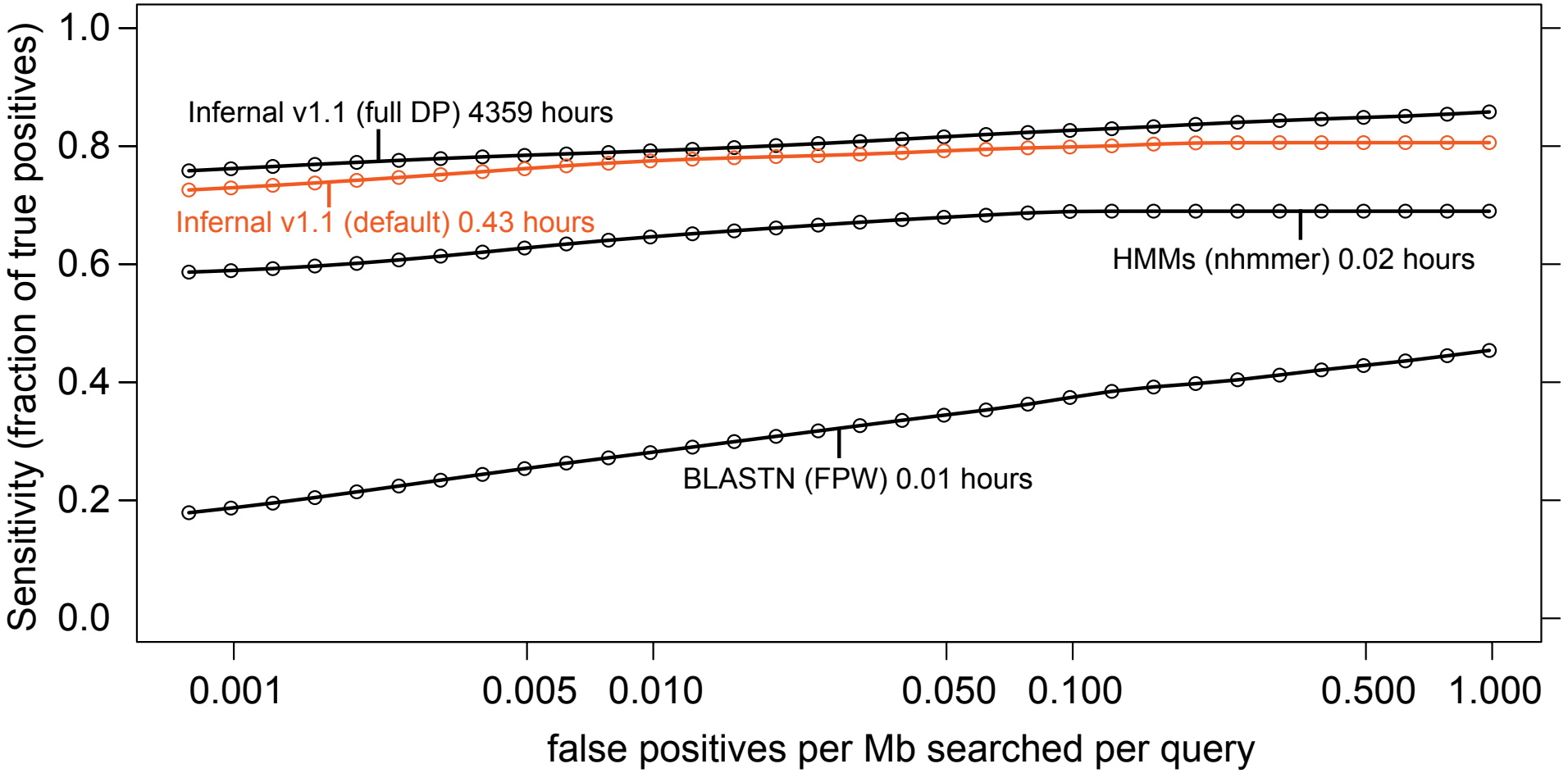
- **main idea:** eliminate potential alignments the HMM tells us are very improbable



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's use of BLAST to accelerate Infernal search of large 270Gb Rfamseq database

PICTURE HERE

Rfam 12.0: first release using Infernal without BLAST filters

SUMMARY LINE FROM TABLE 1 HERE

An example family

???

Rfam and Infernal-based genome/dataset annotation is simple

???