HMMER AND APPLICATIONS

MVE360 – Bioinformatics 2013

Fredrik Boulund
fredrik.boulund@chalmers.se
Bioscience / Mathematical Statistics

This is me

- MSc Biotechnology / Mathematical statistics
- PhD student Bioscience / Mathematical statistics
  - Research on large scale data analysis (metagenomics)

- Researcher by day, musician by night
  - Play guitar in rock band **DÖDAREN**
  - (check us out on Spotify or
    come see us play sometime ;)

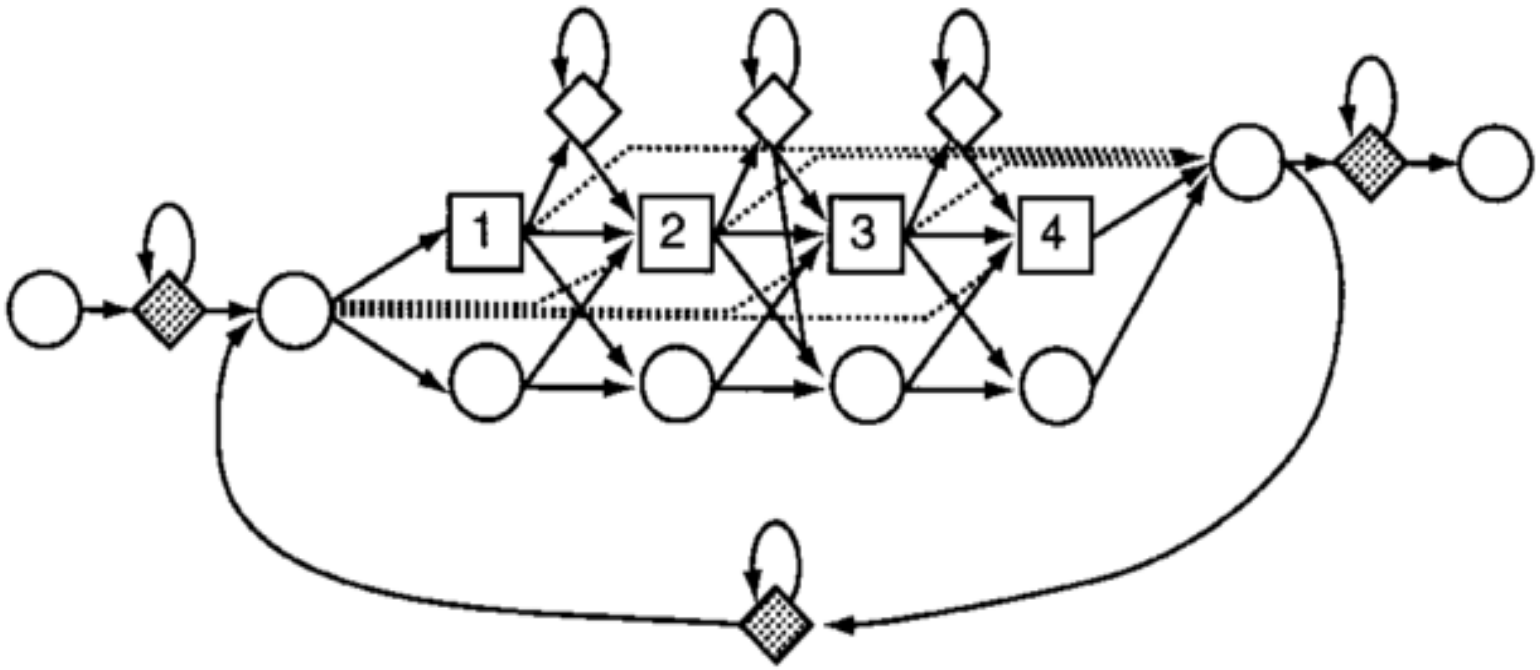
This talk

- HMMER
  - What it is, what are profile HMMs etc.
  - A brief history of [HMMER in] time

- Example of an application of HMMER in metagenomics:
  - Finding antibiotic resistance genes in the environment

What is HMMER

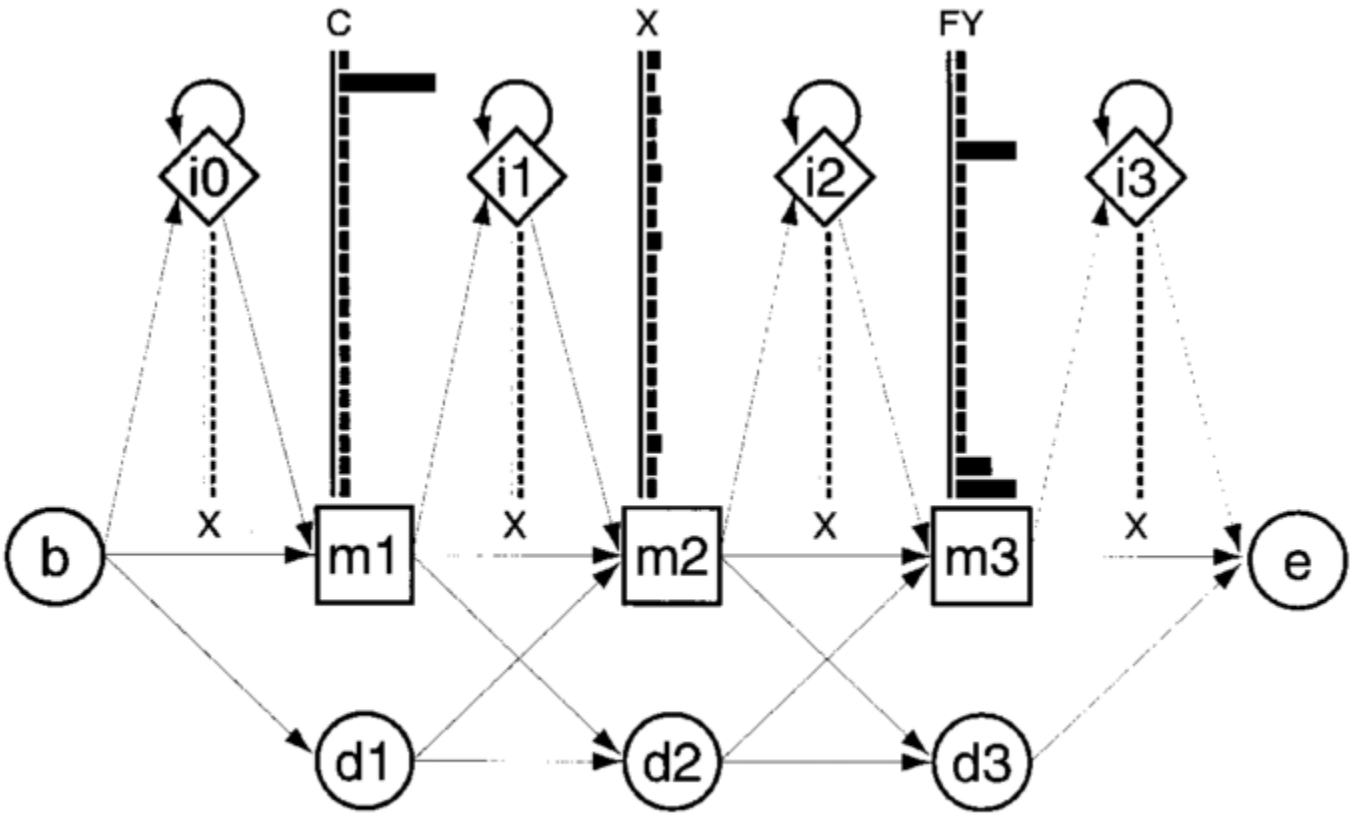
- Sequence alignment software based on a statistical framework using profile hidden Markov models (HMM)



Profile HMMs

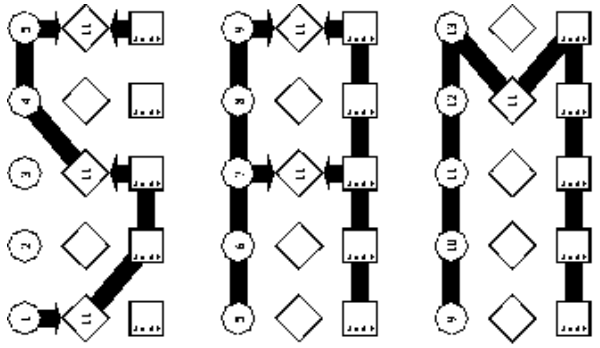
- Probabilistic models of multiple sequence alignments

1	2	3
C	A	F
C	G	W
C	D	Y
C	V	F
C	K	Y



Alternatives

- SAM *(1994)*
  (Sequence Alignment and Modeling system)
  - Richard Hughey
  - Kevin Karplus
  - Anders Krogh



- PSI-BLAST *(1997)*
  (Position-Specific Iterative BLAST)
  - Stephen F. Altschul et al.

(PSI-)BLAST vs HMMER

BLAST

- Single query sequence
- String matching with advanced heuristics for speed
- Mainly good for finding closely related sequences

  (PSI-BLAST)
- Uses position-specific scoring matrices to detect more remote homologs

HMMER

- Based on profile HMMs
- Higher accuracy
- Able to detect even more remote homologs than PSI-BLAST

A brief history of HMMER

- Based on the principles in:
  - Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) *Hidden Markov models in computational biology: Applications to protein modeling*. J. Mol. Biol., 235, 1501–1531.
  - Durbin, Richard; Sean R. Eddy, Anders Krogh, Graeme Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.
- Historically very slow
  - 100-1000 times slower than BLAST
- Instrumental in the construction of:
  - Pfam
  - PROSITE
  - InterPro

HMMER3

- Complete rewrite of HMMER2, focus on improving speed:
  - Eddy, S.R., 2011. Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10), p.e1002195.

- Substantially improves sensitivity and speed over HMMER2 (x100-x1000)

HMMER3 speed

- Heuristic filter: ”Multiple segment Viterbi” (MSV)
  - Computes optimal sum of multiple ungapped local alignment segments using striped vector-parallel (SIMD) Smith-Waterman alignment
- Also accelerates the two standard profile HMM algorithms (Forward/Backward)

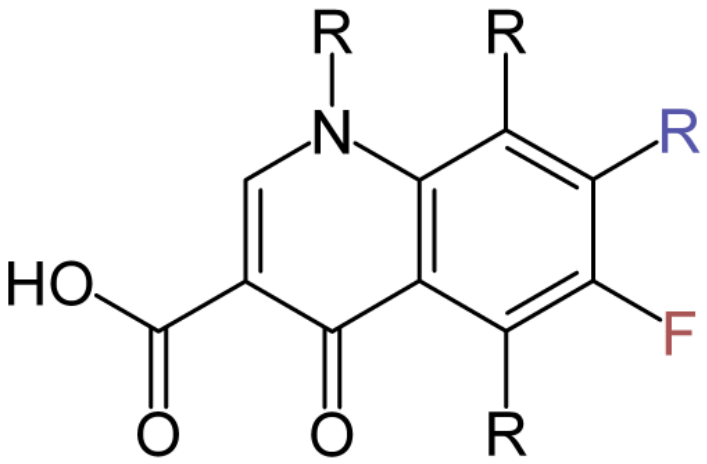
HMMER3 ❤ metagenomics

- Environmental impact of antibiotic production
- Antibiotic resistance
- Using profile HMMs to search for novel AR gene variants

Boulund et al. *A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences. BMC Genomics* 2012, **13**:695

Qnr

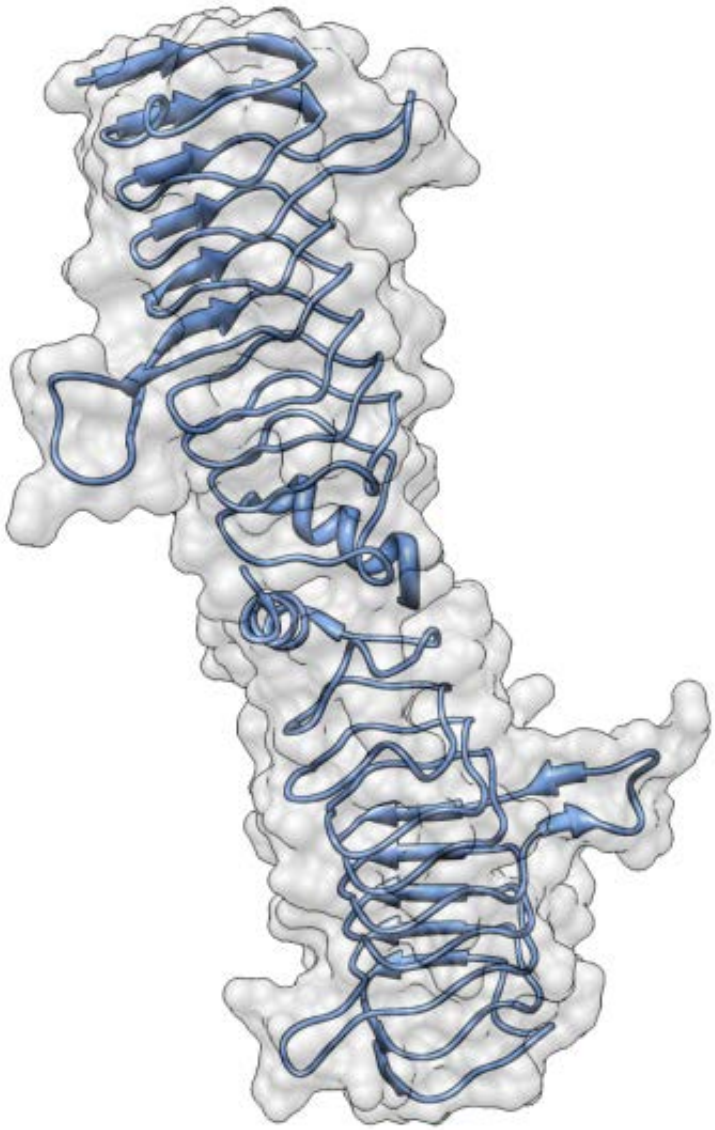
- Pentapeptide repeat proteins (PRP)
- 210-219 aa in length
- Inhibits Type II topoisomerases (Gyrase)
- Five known classes
- Provide bacteria with resistance to fluoroquinolones



General fluoroquinolone

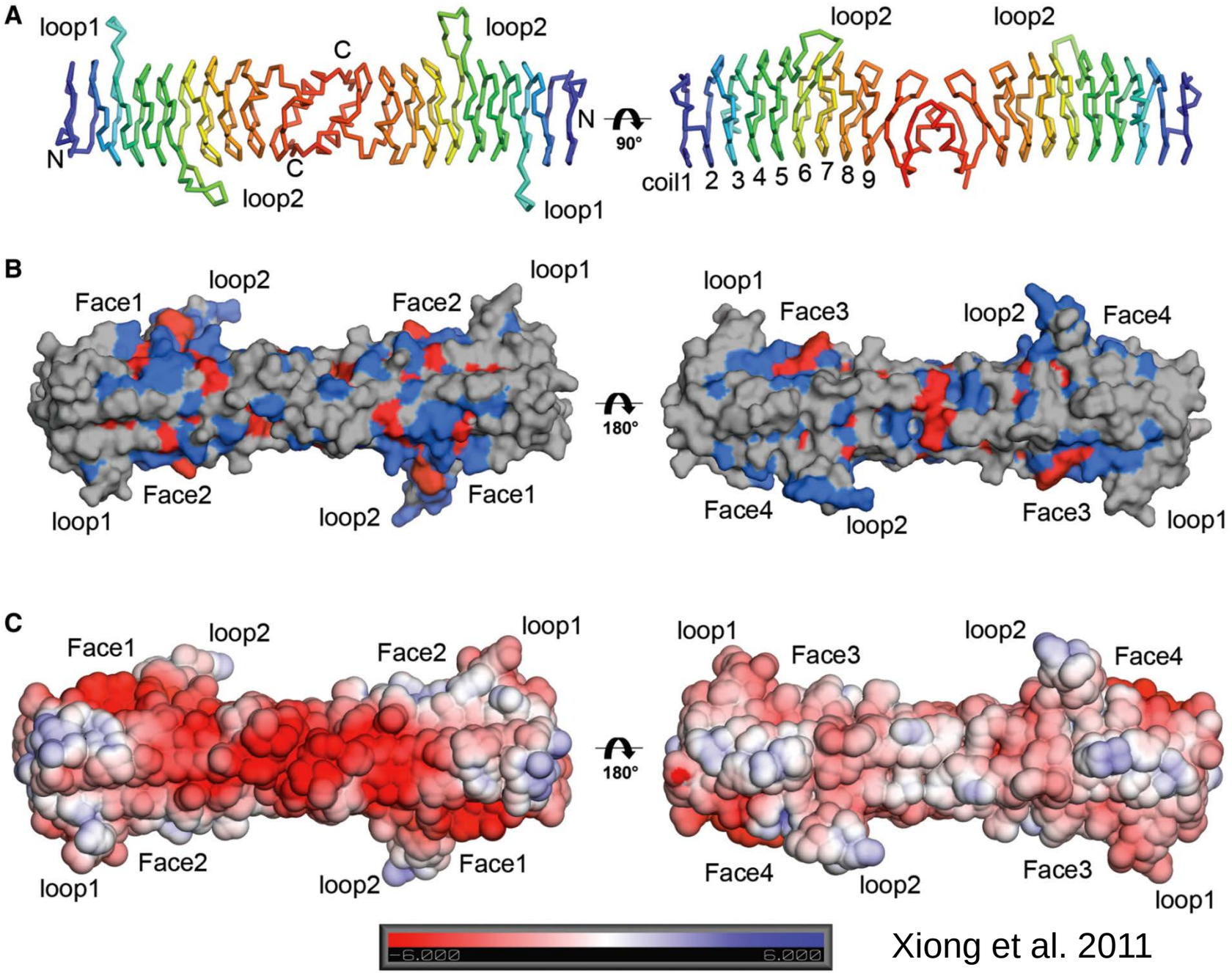
Qnr proteins

- Pentapeptide repeat protein (PRP)
- β-barrel structure
- Size ≤220 amino acids
- Plasmid mediated (PMQR)



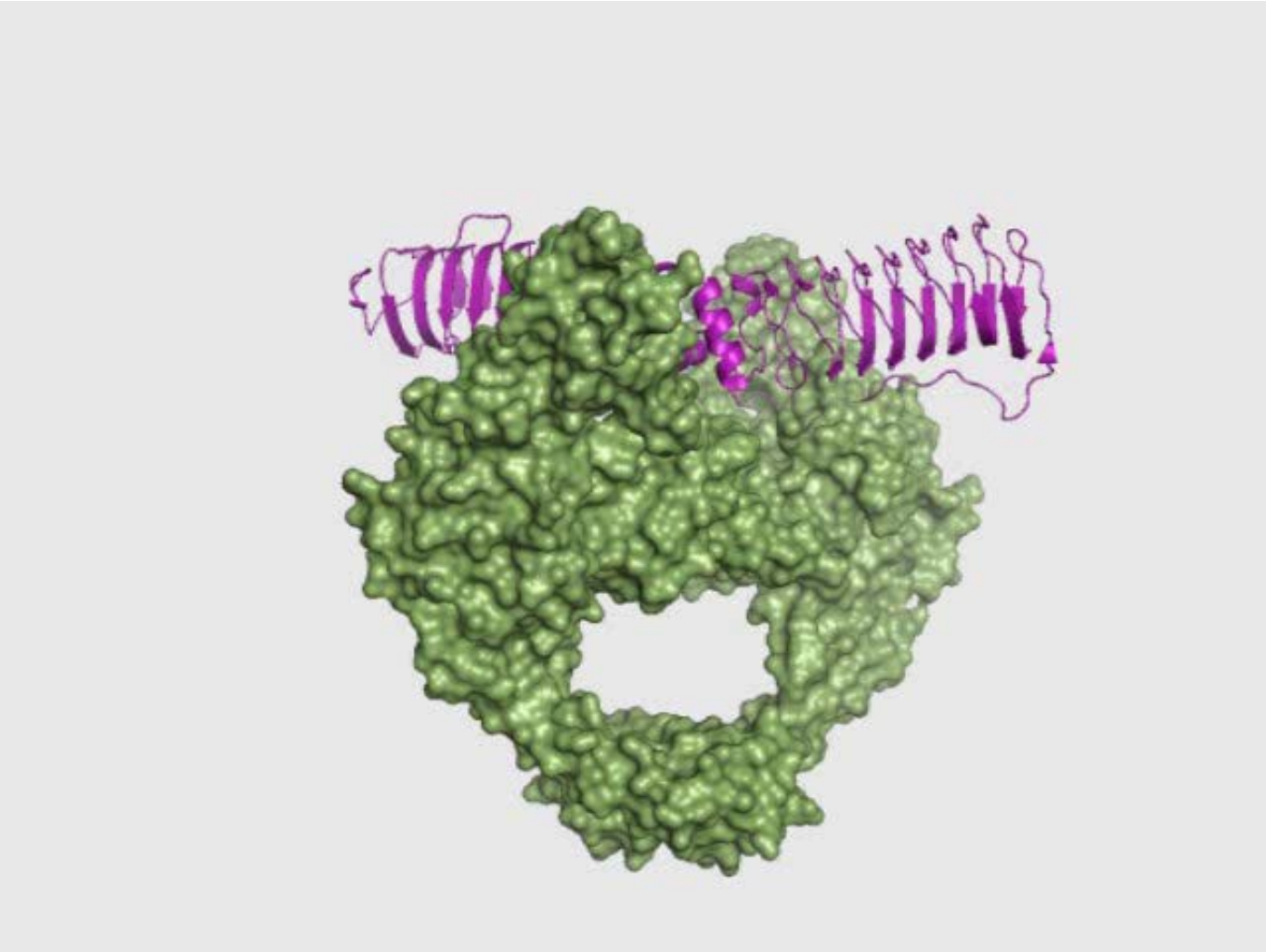
QnrB1 structure

Vetting, M.W., Hegde, S.S., Park, C.H., Jacoby, G.A., Hooper, D.C., Blanchard, J.S. Structure of Qnrb1, a Plasmid-Mediated Fluoroquinolone Resistance Factor. (2011) J.Biol.Chem. 286: 25265



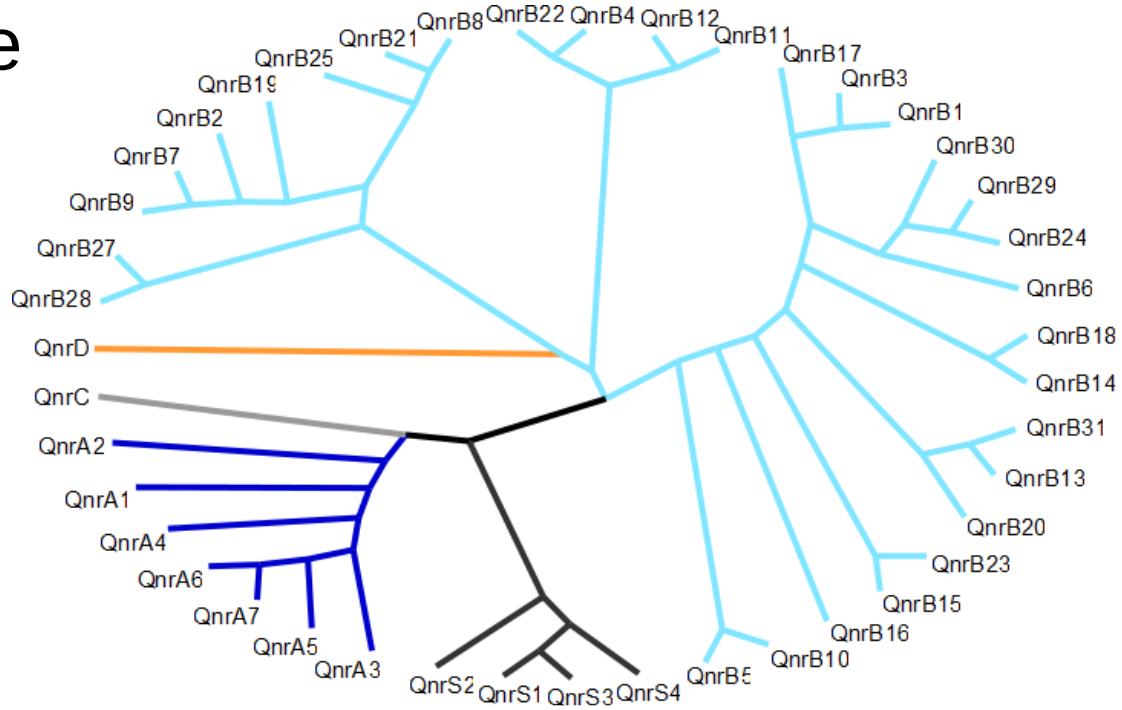
Xiong et al. 2011

QNR in GyrA



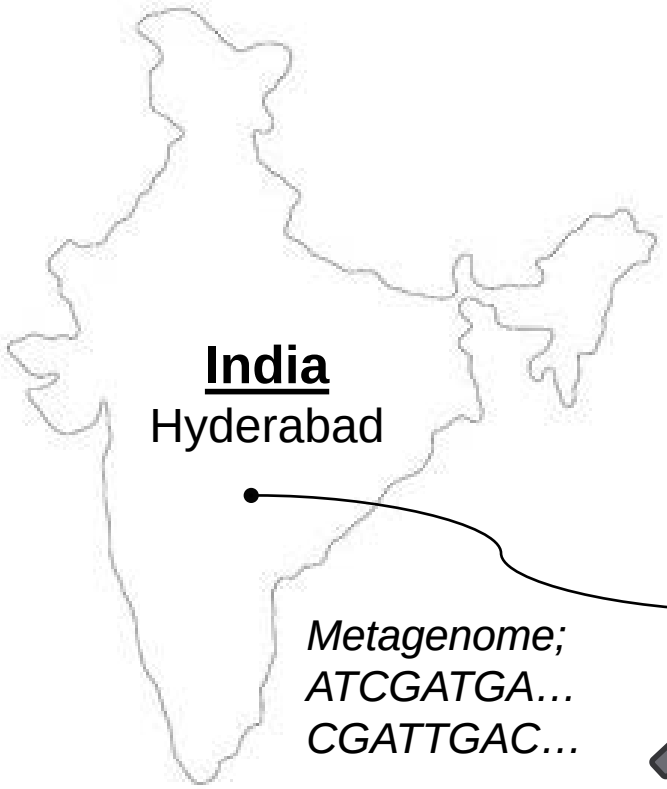
Current state of Qnr knowledge

- 5 classes of mobile Qnr, 61 different variants in total:
  - QnrA: 7
  - QnrB: 47
  - QnrC: 1
  - QnrD: 1
  - QnrS: 5
- Classification based on sequence similarity



Radial cladogram of identified Qnr variants February 2011

Concept



India
Hyderabad

Metagenome;
ATCGATGA…
CGATTGAC…

Search metagenome using model

Multiple alignment; qnr genes

```
QnrA MDIIDKVFQQEDFSRQDLSDS…
QnrB MA---LALVGEKIDRNRFTGE…
QnrC MNYSHKTYDQIDFSGQDLSSH…
QnrD ME---KHFINEKFSRDQFTGN…
QnrS METYNHTYRHHNFSHKDLSDL…
```

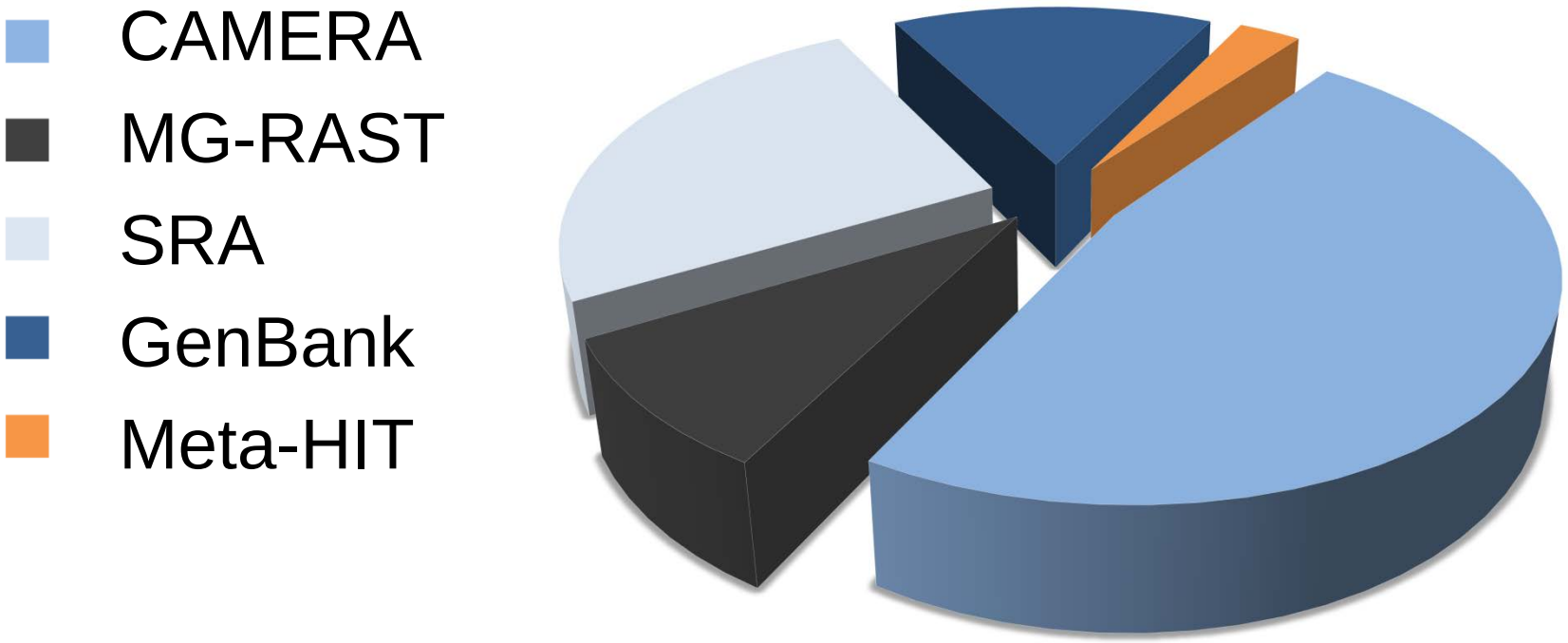
QNR MODEL (HMM)

- Known qnr genes (identification)
- Novel qnr (discovery)

Metagenomic data

- Short ”reads”
- Several different technologies
- Mainly 454 pyrosequencing in this project
  - Read lengths approximately 200-600 bp.

Sequence data



More than 700 gigabytes!

Why HMMER3?

- Hidden Markov models very suitable for Qnr pentapeptide repeat structure

- Speed is improved from HMMER2:
  - Making it possible to apply to high-throughput sequencing data

HMMER3 output

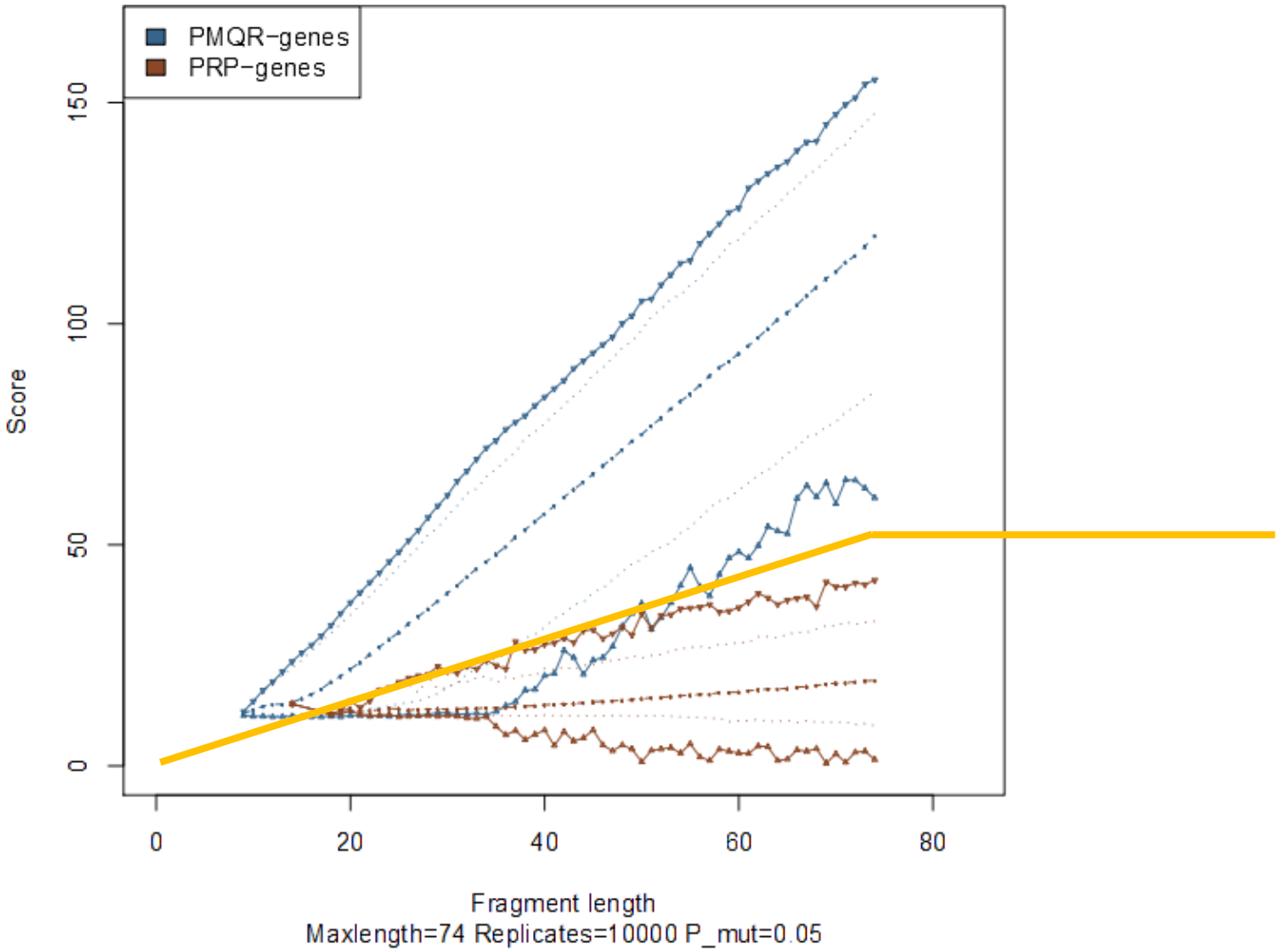
```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.0 (March 2010); http://hmmer.org/
# Copyright (C) 2010 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  hmm_model9_PMQR_aa_20110203.hmm
# target sequence database:        /db/genbank/nt.pfa
# output directed to file:         ./hmmsearchresults/nt.pfa.hmmsearched
# max ASCII text line length:      unlimited
# number of worker threads:        6
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       PMQR_aa  [M=217]
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence         Description
    ------- ------ -----    ------- ------ -----   ---- --  --------         -----------
   3.6e-122  418.1   6.7   4.3e-122  417.8   4.6    1.0  1  EU273755.1_1     Citrobacter freundii strain 05K657 Qnr…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EF517946.1_1     Enterobacter aerogenes plasmid pWCH-LM…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EF520349.1_1     Pantoea agglomerans plasmid pWCH-LM2 Q…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EF523819.1_1     Klebsiella pneumoniae strain HX0500638…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EF634464.1_1     Escherichia coli plasmid pGD005 QnrB (…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EU093091.1_1     Escherichia coli plasmid pGD006 QnrB (…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  EU443840.1_4     Klebsiella pneumoniae plasmid pGDK05 Q…
   4.1e-122  417.9   6.5     5e-122  417.6   4.5    1.0  1  GQ914054.1_1     Shigella sonnei strain 136 quinolone- …
```

HMMER3 bit scores

- Log-odds score for sequence against model under null-hypothesis
  - A statistically based measure of how well the sequence aligns/fits to the model

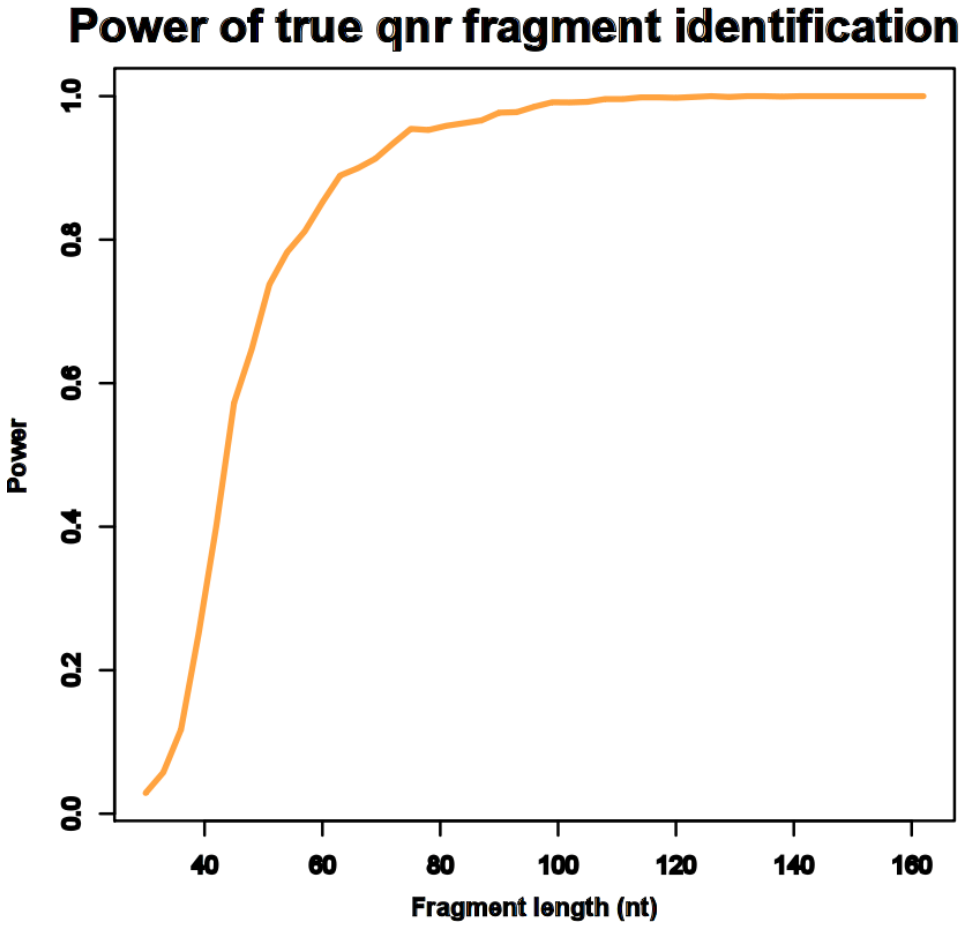
- Common to use E-value cutoffs, but difficult here since it depends on dataset size

Plot of bit score distribution



Classification of small fragments

- Function of fragment length and HMMER3 bit score
- Optimized then validated using cross-validation



Results

- Identified all previously known qnr genes in metagenomic data
- Reconstructed a complete QnrB35 gene from metagenomic data
- Found several putative novel variants of qnr
- Discovered sequences with lacking and/or incorrect annotation in GenBank

Conclusions

- HMMER3 opens up possibilities:
  - It **is** fast enough to apply to very large datasets
  - Using HMMER3 the pipeline is capable of detecting novel qnr fragments with 99% true positive rate when combined with our classifier
  - The high sensitivity and speed provided by HMMER3 allowed us to find a lot of fun stuff!