HMMERHEAD

HMMer Hashing Enabled Acceleration Device

*99% of the matches in **6.6%** of the time*

Elon Portugaly and Matan Ninio, School of Computer Science and Engineering, Hebrew University
{elonp,ninio}@cs.huji.ac.il
http://www.cs.huji.ac.il/labs/compbio/hmmerhead

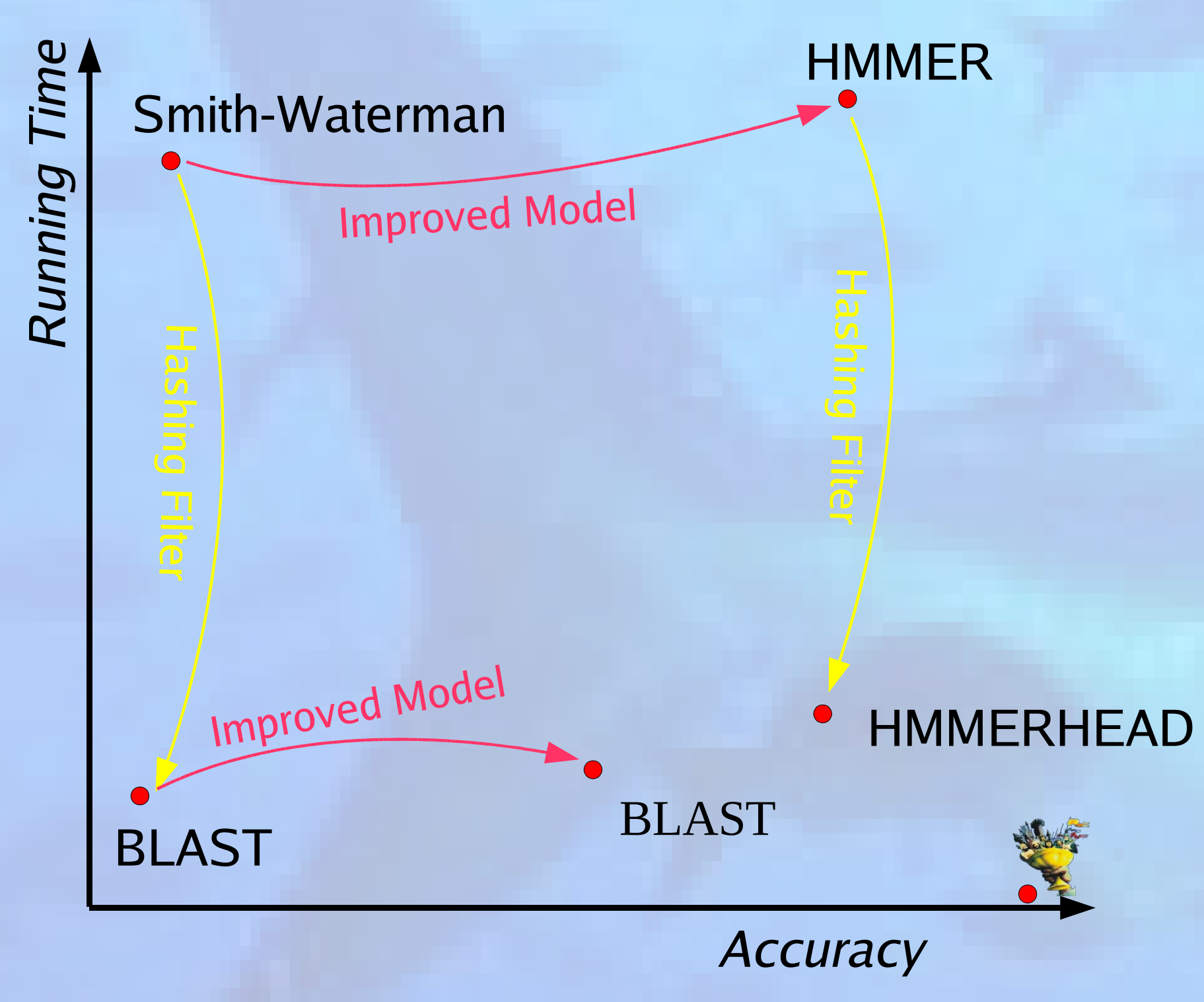
Abstract

Profile HMMs are amongst the strongest remote homologue sequence search tools available today. However HMM scans are computationally expensive – scanning SWISS-PROT with all Pfam HMMs would take over 3 months on a 2.8GH Pentium 4.

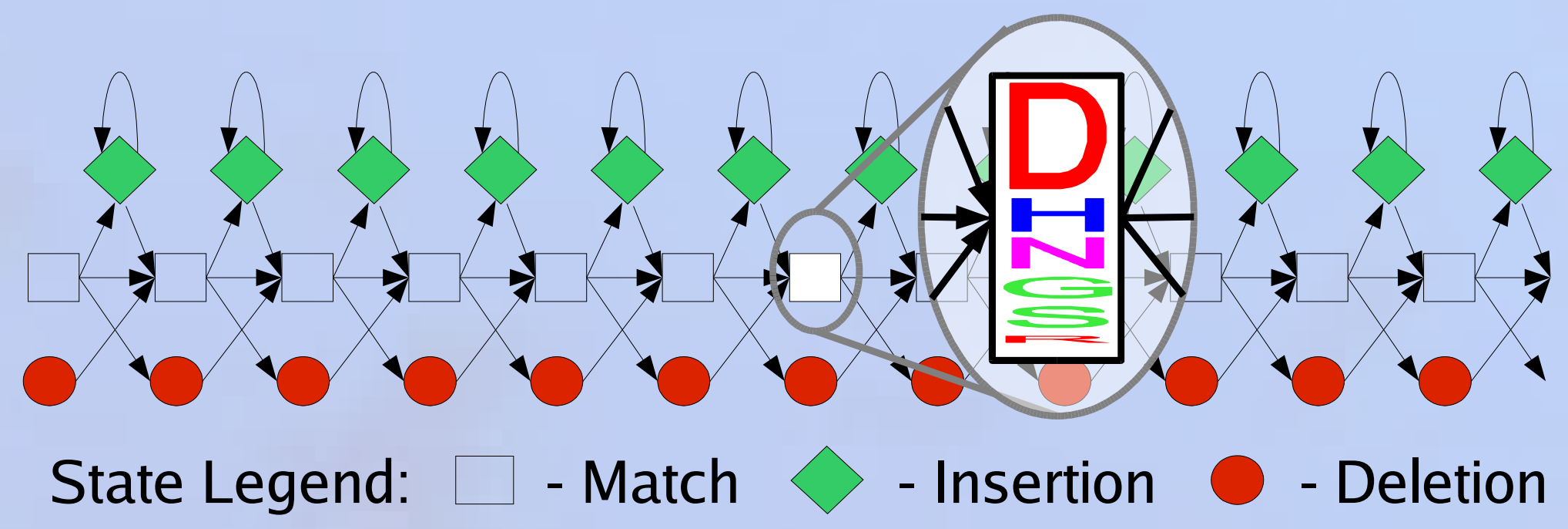
We present HMMERHEAD – a BLAST like two-hit-method filter for profile HMMs. HMMERHEAD filters sequence databases and presents a reduced set of candidate sequences to HMMER's hmmsearch.

HMMERHEAD achieves a 15-fold acceleration while retaining 99% of the results.

Homologue Search

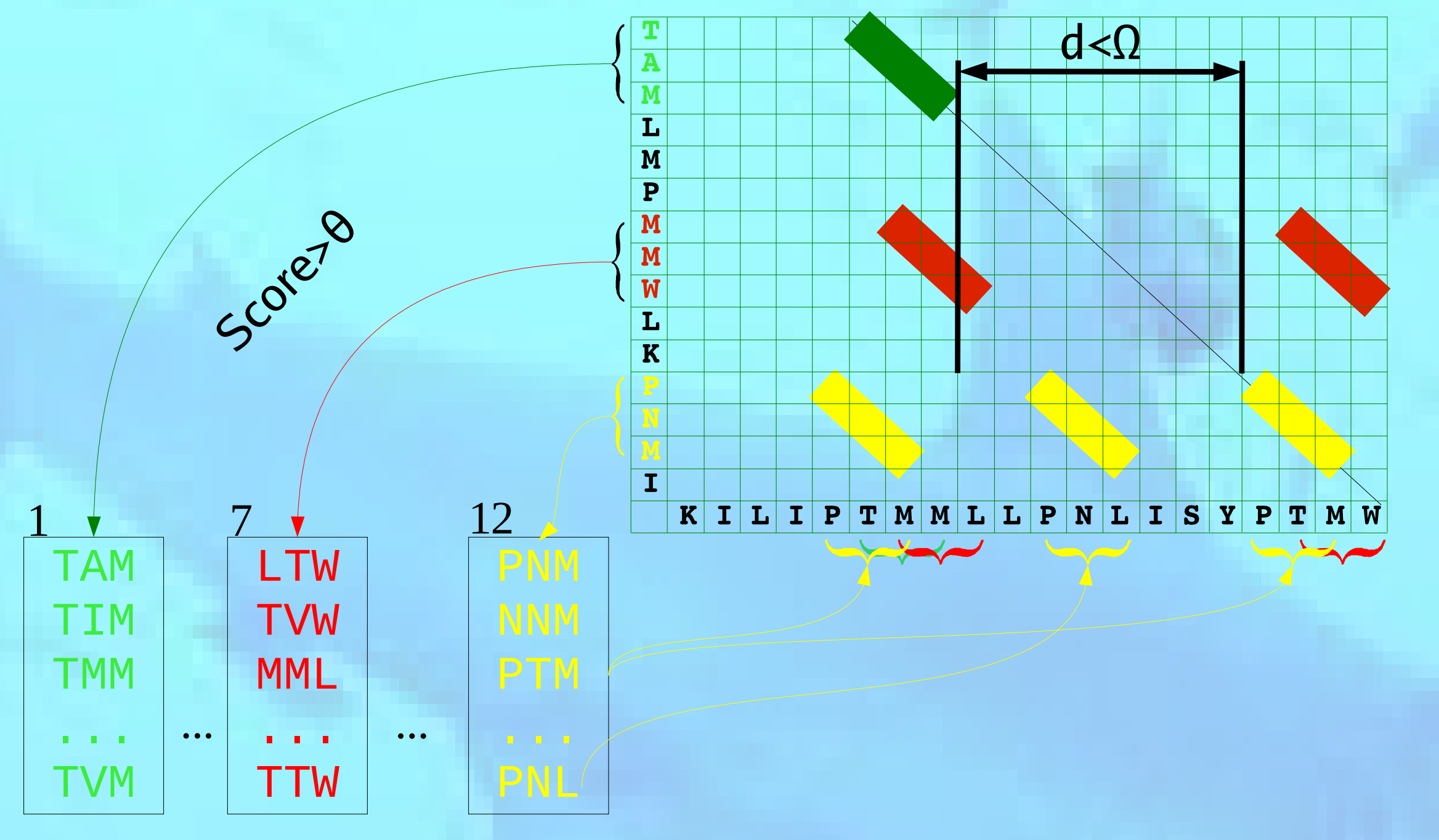


Profile HMMs

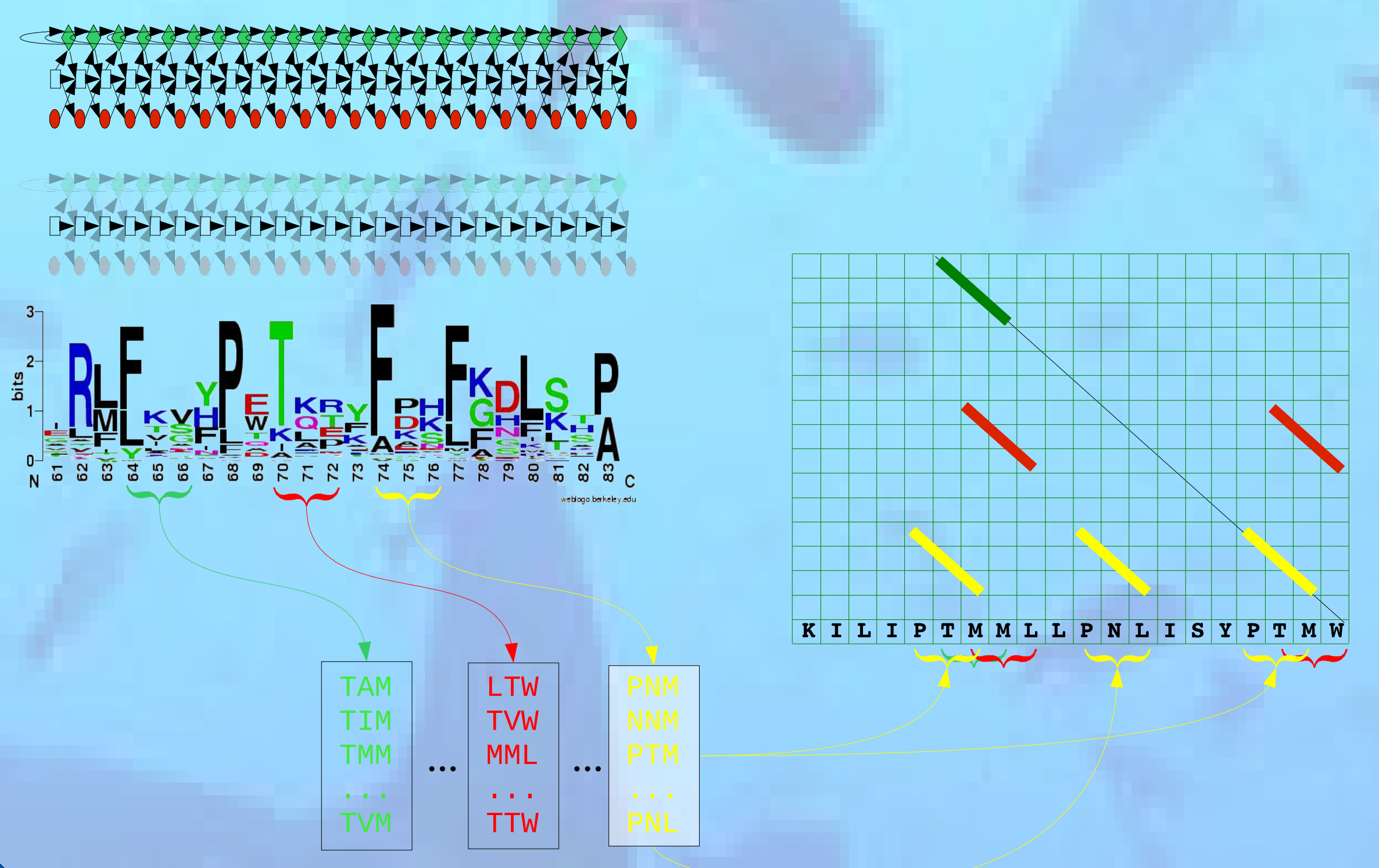


- Each triplet of match/insertion/deletion states corresponds to a position in the sequence family
- HMMs model different environments for different positions (levels of conservation, conservation along specific chemo-physical property, etc.)
- They can model different probabilities for insertion/deletion in different positions

The Two-Hit Method


T A M L M P M M W L K P N M I
K I L I P T M M L L P N L I S Y P T M W
d<Ω
Score>θ
1
TAM
TIM
TMM
...
TVM
7
LTW
TVW
MML
...
TTW
12
PNM
NNM
PTM
...
PNL

Two-Hit Method in HMM



Implementatio

- HMMERHEAD is implemented as a filtering step before HMMER's hmmsearch

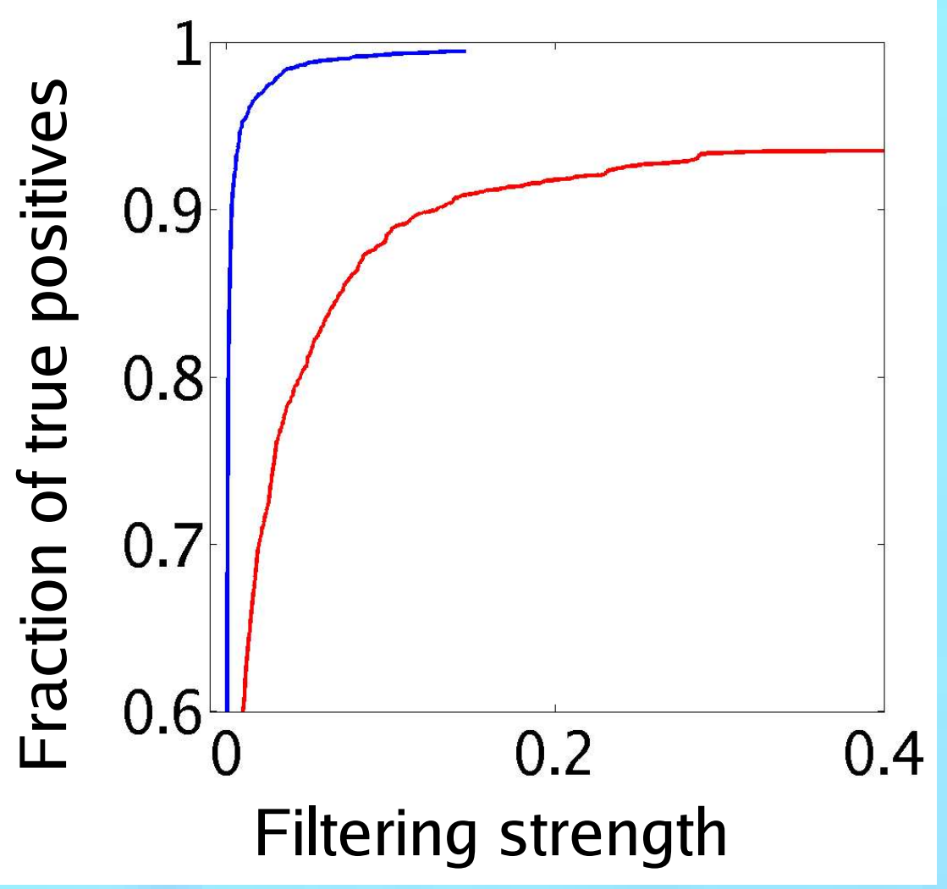
Experiments

- 476 randomly chosen Pfam HMMs
- All 133,312 SWISS-PROT (rel 41.21) sequences
- Database scanned by each HMM, matches collected as goals for filtered search
- Database scanned by each HMM after HMMERHEAD filtering. Goal:
  - Collect all matches of unfiltered scan
  - Filter as much as possible

    ⇕

  - Reduce running time as much as possible
- We used Ω=25, and different values for θ

Filtering & Coverage



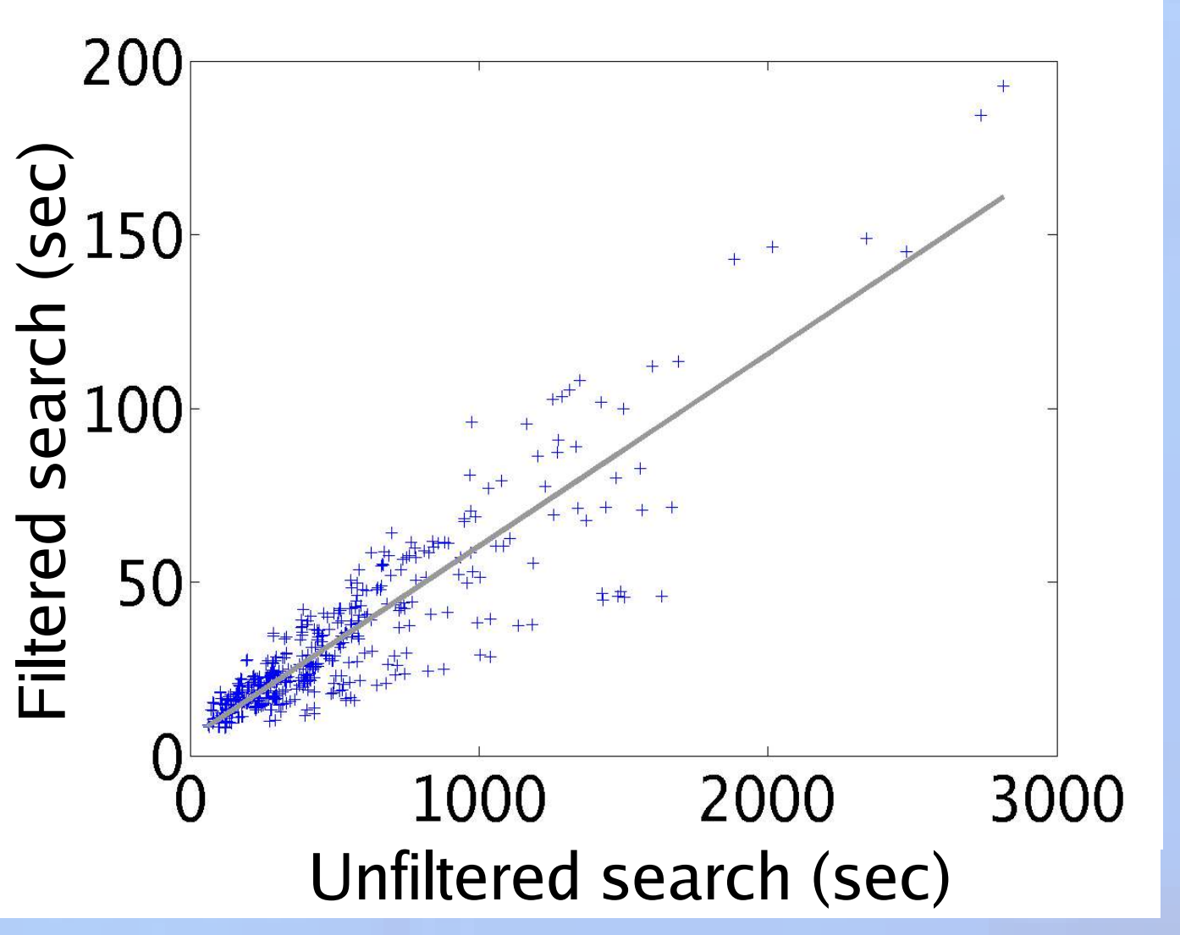
The number of true positives that pass the filter are shown in the graph for different filtering strengths.
Each match found by the unfiltered search is considered true positive.
Filtering strength is denoted by the fraction of sequences that pass the filter.
blue – filtering by Two-Hit method
red – filtering by single hits

Data on coverage and acceleration of HMMERHEAD at specific filtering thresholds are in the upper part.
Statistics of the performance for each hmm is shown in the lower part of the table below.

Filtering[1] (%)	10.0	5.0	**2.5**	1.0	0.5
Average recall[2]	99.4	99.3	**99.0**	90.6	88.3
Speedup factor[3]	5.5	8.4	**15.2**	34.4	52.5
Recall[4]					
Above 99%	93.6	92.0	**89.0**	82.8	77.2
99%-95%	4.3	5.4	**5.6**	4.0	4.0
95%-90%	1.6	2.1	**3.2**	4.0	5.4
Below 90%	0.5	0.5	**2.1**	9.1	13.4

[1] Percent of sequences that pass the filter
[2] Percent of total true positives that survive filtration
[3] Decrease factor in total CPU time (user+system)
[4] Percent of HMMs for which recall is within range

Improvement in Running Times



CPU time (user+system) for each HMM with and without HMMERHED filtering.

Filtering strength set to 2.5% sequences passing.

Discussion

HMMERHEAD was developed to assist the development of EVEREST - a fully automatic procedure that accepts a database of protein sequences and performs the dual role of splicing the sequences into domains, and clustering these domains into families. The speedup achieved by HMMERHEAD enabled us to run 20,000 HMMs automatically produced by EVEREST. The experiments we have performed are very preliminary, and yet show a significant increase in EVEREST's performance.

One of the strengths of profile HMMs is their ability to model different insertion and deletion probabilities at different positions. One can think of changes to the filtering algorithms that would take advantage of this ability, e.g. by adding flexibility to the diagonals in the Two-Hit method. However, the results we have achieved for the Pfam X SWISS-PROT experiments make it unclear that improving the algorithm is worth the effort.