HMMERHEAD - Accelerating HMM Searches On Large Databases

Elon Portugaly[1] **Matan Ninio**[1]

Keywords: HMM, acceleration, sequence, search, Pfam, SWISS-PROT, BLAST

1 Introduction

HMMs have been proven useful in protein sequence analysis [1]. However, a full search of a sequence database using an HMM is a computationally expensive process - running all the Pfam [3] HMMs on the SWISS-PROT database [4] takes almost three months of computer time. The two-hit method used by Altschul et al [2] allows BLAST to accelerate both sequence vs. sequence searches and profile vs. sequence searches. In this work we build a framework that uses a similar method for HMM searches.

We provide *HMMER Hashing Enabled Acceleration Device* (HMMERHEAD) - a software package that filters out sequences for hmmsearch. Our experiments show that we typically achieve a 15-fold acceleration of running time, while retaining 99% of the results.

2 The Two-Hit Method

The two-hit method was introduced in [2]. Following is a short description of the method.

Preprocessing: In a preprocessing step, a database of k-mers (i.e. words of size k over the alphabet used) is compiled from the sequence database. For each possible k-mer this database provides quick access to the list of all occurrences of the k-mer in the sequence database. This database is fixed for all queries, and need only be computed once for each sequence database.

Queries: When presented with a query, the algorithm finds all k-mers that locally match any part of the query with a local score above some given threshold. Next, the k-mer database is queried for the occurrences of each of the above k-mers in the target sequences. Each such occurrence resides within a *diagonal* path in the alignment graph of the sequence and the query. If two such occurrences share a diagonal, and are within a fixed distance from each other, the target sequence is reported as a candidate for dynamic programming search.

3 HMMER and HMMERHEAD

We have implemented the two-hit-method as a filter stage for HMMER [1]. HMMER is a software package that implements HMMs for families of protein and DNA sequences. Given a query HMM and a sequence database, HMMERHEAD filters the sequence database using the two-hit-method, and pipes the sequences that passed the filter to HMMER's hmmsearch program for the final search. HMMERHEAD accepts HMMER HMMs files as input. The HMMERHEAD package has been tested on Linux and is provided under the GNU license at `http://www.cs.huji.ac.il/labs/compbio/hmmerhead`.

[1]School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel.
E-Mail: {elonp,ninio}@cs.huji.ac.il

Filtering[1] (%)	90	95	97.5	99	99.5
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 are filtered out
[2]percent of total required matches that survive filtration
[3]decrease factor in total cpu time (user+system)
[4]Percent of HMMs for which recall is within range

Table 1. Recall and speedup by filtering

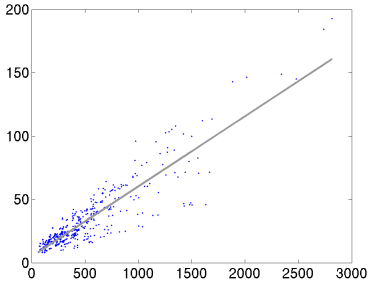

Figure 1. Running times

4 Results

HMMERHEAD performance is evaluated by two measures: *recall* - the percentage of true positives that survive the filtering, and *filtering* - the percent of sequences that pass the filter. We also measure machine running time directly. All runs where performed with k-mers of size 4 and a two-hit window of size 25.

We tested HMMERHEAD performance by searching for Pfam family members in the SWISS-PROT (rel. 41.21) database [4].

Each Pfam family is defined by two HMMs and two corresponding cutoff scores. We randomly chose about 5% of the Pfam families, and collected all 476 related HMMs. We searched the SWISS-PROT database with each of the HMMs, first with HMMER, and then with HMMERHEAD using several different levels of filtering. The results are shown in Table 1. Note that when filtering 97.5% of the sequences, we achieve a speedup factor of more then 15, and loose only 1% of the required matches. The bottom part of the table shows the performance over the different HMM profiles - here recall is computed separately for each family. Figure 1 shows the running times, in cpu seconds, of the different HMMs. X-axis – search without HMMERHEAD; Y-axis – HMMERHEAD filtration with 97.5%. A least squares linear fit sets $f = 5.028 + 0.0553 \times u$ for f filtered time and u unfiltered time.

References

[1] Eddy, S. R. 2001. HMMER: Profile hidden Markov models for biological sequence analysis (`http://hmmer.wustl.edu/`)

[2] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. in *Nucleic Acids Res* 25, 3389-3402.

[3] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S. R., Griffiths-Jones S, Howe K. L., Marshall M, and Sonnhammer E. L. 2002. The Pfam protein families database. In *Nucleic Acids Research* 30, 276-280

[4] Boeckmann B., Bairoch A., Apweiler R., Blatter M. -C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I., and Pilbout S., Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. In *Nucleic Acids Res.* 31, 365-370.