Improved HMMERHEAD for Better Sensitivity

Elon Portugaly[1] **Steve Johnson**[2] **Matan Ninio**[1] **Sean R. Eddy**[3]

Keywords: HMM, HMMER, acceleration, sequence, search

1 Introduction

HMMs have been proven useful in protein sequence analysis. However, a full search of a sequence database using an HMM is computationally expensive. HMMERHEAD is a software package, incorporated into the HMMER [4] HMM suite, that efficiently filters out most of the unrelated sequences while retaining nearly all related ones. A full scan of the NR database [7] takes 767 seconds using HMMERHEAD, while an unfiltered search takes 5 hours.

In RECOMB 2004[6] we presented the first version of the HMMERHEAD acceleration package. There, we have reported a fifteen fold acceleration while retaining over 99% of the matches of Pfam [1] HMMs scanning the Swiss-Prot database [2]. While still the most widespread use of protein sequence HMMs, Pfam HMMs proved too easy a benchmark when compared to HMMs constructed from structural families. In this work we present improvements of the HMMERHEAD algorithm that enable it to maintain high sensitivity even when applied to more demanding HMMs modeling protein structural families.

The current HMMERHEAD filtering process is composed of an initialization step followed by a four tier filtering scan of the target sequences. Those sequences that pass the filtering are scored by full dynamic programming, as in unfiltered HMM scan.

HMMERHEAD is provided both as a stand-alone *C++* template library which can be employed by different model-to-sequence scan software, and incorporated into HMMER, at `http://www.cs.huji.ac.il/~elonp/hmmerhead`.

2 HMMERHEAD filtration

HMMERHEAD filtration is achieved by seeking a high-scoring partial alignment between the target sequence and the HMM model. The procedure that seeks this partial alignment is composed of four filtering stages with increasing computational cost. At the end of each stage, candidates that are unlikely to reach a good alignment are filtered out. At the end of the filtration process, the remaining sequences are passed to standard HMMER scoring.

The first filter seeks short k-mers in the target sequence that get a high score from match-states alone. Next, each such hit is extended without gaps in both directions, and low-scoring hits are discarded. In the third filter, pairs of hits whose positions on both query and target are close to each other, and whose diagonals are close are passed. Finally, each pair of hits is extended to a limited alignment by first connecting the two hits in a gapped alignment, and then extending the resulting hit in an ungapped manner in both directions. Sequences with at least one such hit that scored above a threshold are passed on to HMMER.

[1] `{elonp,ninio}@cs.huji.ac.il` The Selim and Rachel Benin School of Computer Science and Engineering, Hebrew University, Jerusalem 91904, Israel.

[2] `stevej@pathology.wustl.edu` Department of Pathology, Washington University School of Medicine, St. Louis, MO 63108

[3] `eddys@janelia.hhmi.org` HHMI Janelia Farm Research Campus, 19700 Helix Drive, Ashburn VA 20147

3 Results

We followed the benchmark of remote homology detection described in [5]. This benchmark uses HMMs based on Astral SCOP [3] sequences as queries. The searched database is composed of all 2,521 Astral SCOP sequences plus over two million shuffled sequences as decoys. Hits to sequences of the same SCOP superfamily as the query HMM are considered true positives, while hits to decoy sequences are considered false positives. Search outputs of all queries are combined and sorted by e-value.

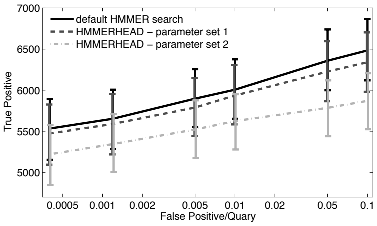

Figure 1: **HMMERHEAD Performance.** Results for three parameter settings of HMMER 2.5 local Forward search. Error bars represent 95% confidence intervals

This benchmark is significantly harder than benchmarks defined using Pfam HMMs, mainly because of the need to identify more distant homologous at the superfamily, rather than family level. To accommodate for the requirements of this benchmark, we employ the much more sensitive but slower Forward algorithm rather than the more commonly used Viterbi algorithm . We show results for three searches: an unfiltered (no HMMERHEAD) search, and two HMMERHEAD filtered searches with two different parameter settings.

Figure 1 shows performance of HMMER 2.5, local forward searches, for unfiltered search and for two HMMERHEAD parameter settings. The unfiltered forward search takes an average of 5 hours per HMM query. The first parameter setting takes 767 seconds per query (24 fold acceleration), while the difference between its sensitivity and that of unfiltered search is statistically insignificant. Some loss of sensitivity is suffered when using the second parameter setting, but it reaches a 27 fold acceleration (676 seconds per query).

We conclude that by achieving 24 fold acceleration, HMMERHEAD filtration allows searches of large databases using the Forward algorithm, with insignificant loss of sensitivity when compared with unfiltered Forward searches and significant increase of sensitivity when compared with Viterbi searches.

References

[1] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res.*, 30(1):276–80, 2002.

[2] B. Boeckmann, A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, 31(1):365–70, 2003.

[3] J. M. Chandonia, N. S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S. E. Brenner. ASTRAL compendium enhancements. *Nucleic Acids Res*, 30(1):260–3, 2002.

[4] S. R. Eddy. Hmmer: Profile hidden markov models for biological sequence analysis. 2001.

[5] Steven Johnson. *Remote Protein Homology Detection Using Hidden Markov Models.* PhD thesis, Department of Genetics, Washington University School of Medicine, 2006.

[6] E. Portugaly and M. Ninio. HMMERHEAD - Accelerating HMM Searches On Large Databases. *Currents in Computational Molecular Biology - Poster Abstracts from RECOMB*, pages 250–1, 2004.

[7] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 28(1):10–14, Jan 2000.