Evaluating Homology Search Algorithms

Timothy L. Bailey **Michael Gribskov**
tbailey@sdsc.edu gribskov@sdsc.edu
University of California, San Diego, NPACI/SDSC, MC 0537,
9500 Gilman Drive, La Jolla, California 92093-0505, USA

Keywords: accuracy comparison, homology algorithms, classification, *p*-values

1 Introduction

Both the users and developers of homology search algorithms have a natural interest in comparing the search accuracy of different algorithms. We show that traditional comparison methods [1, 4] may favor search algorithms that systematically underestimate the scores of either short or long sequences. The winning algorithm in such a comparison may not be ideal since it may tend to miss distant homologs that are short (or long). We show how to quantify the length bias of homology search algorithms that return *p*-values (as most modern homology search algorithms do). We also show how to present this information along with search accuracy results to more fully characterize the relative merits of two or more search algorithms.

2 Method and Results

We compare two search algorithms, called ML and MLH, that differ only in the way that they compute *p*-values from raw Smith-Waterman scores. Each of 321 test cases consists of a single query sequence selected at random from a SCOP version 37 [3] family and used to search a database of 2448 SCOP domain sequences. Sequence length is highly uniform within test families (Figure 1A).

We measure classification accuracy using the rate of false positives (RFP) [2], the number of false positives scoring better than the lowest-scoring true positive. (The results are similar using ROC or other metrics.) Algorithm ML has better RFP on far more test cases than MLH. The *sign test* [4] rejects the null hypothesis that the RFP is the same with Z-score -5.96 (highly statistically significant).

By plotting the difference in RFP for each test case (Figure 1B), we see that algorithm ML wins primarily when the average sequence length in the target family is greater than 50. Examining the results of one test case in detail (circled point in Figure 1B), we see in Figures 1C and 1D that ML is more biased against short sequences than is MLH. The points in Figures 1C and 1D are rank *p*-value ($\frac{1}{r+1}$) versus *p*-value, where r is the rank of the *p*-value among targets of the same length. The curves shown in the figures (*p*-value slope lines) are regression lines fit to the points for target sequences in a given length range. Completely accurate *p*-values would give *p*-value slope lines with slope (*p*-value slope) equal to 1.0. This suggests characterizing the overall *p*-value accuracy of a search algorithm by one minus the *p*-value slope (*p*-value slope error), averaged over the test cases. Figure 1E shows this for ten length ranges chosen to contain equal numbers of target sequences. Compared with MLH, ML systematically overestimates the *p*-values of short sequences.

The final visualization step is to overlay the *difference* between the two average *p*-value slope accuracy curves in Figure 1E onto the plot showing pairwise differences in search accuracy in Figure 1B. Numerous comparisons of different search algorithms consistently show strong correlation between the *p*-value slope error difference curve and the distribution of classification accuracy differences like that in Figure 1F. Overall, MLH may be the preferable algorithm of the two because it gives more accurate *p*-values (on average) and is less biased against short sequences, despite the results of the *sign test*.

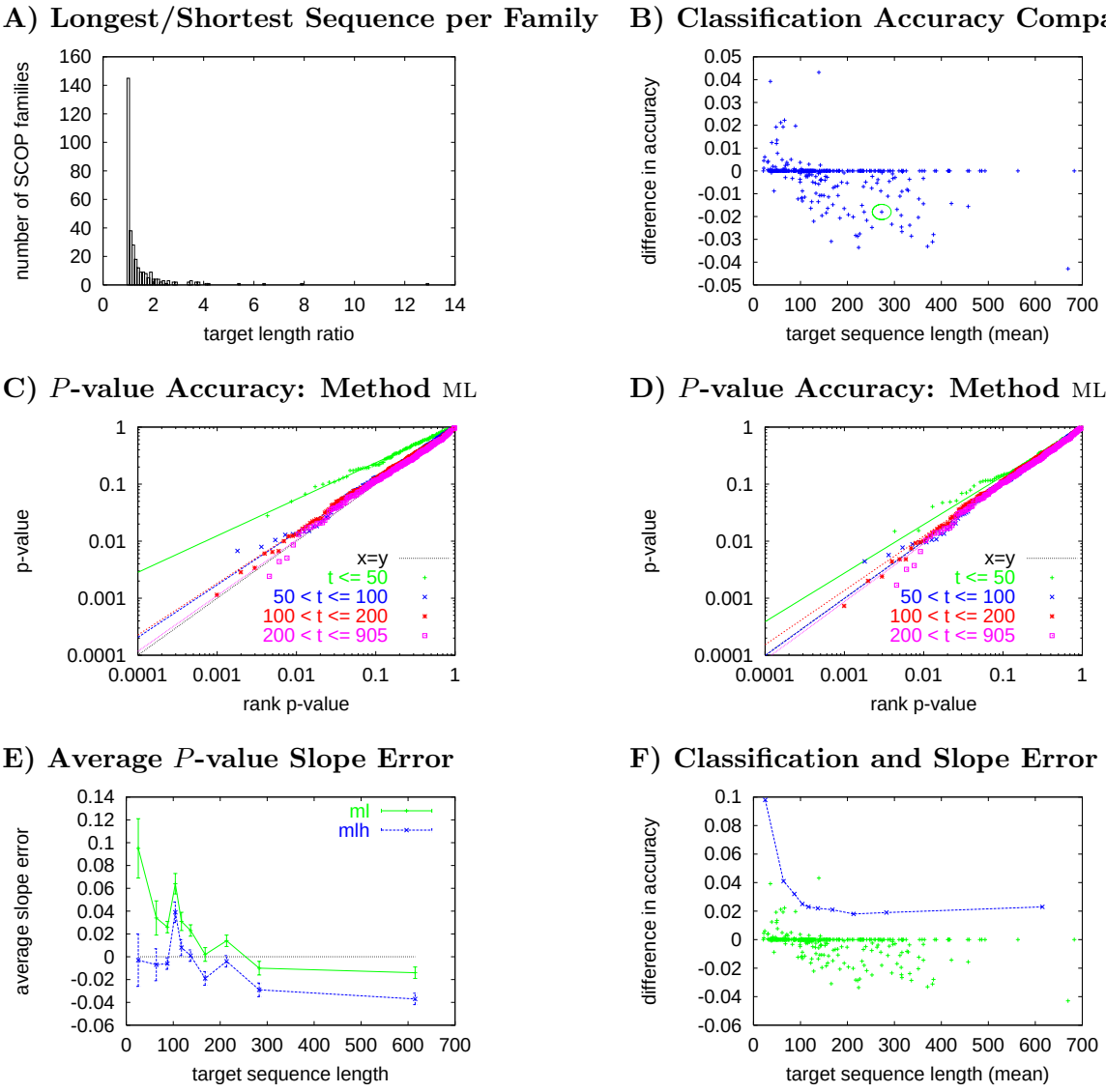

Figure 1:

References

[1] Henikoff, S. and Henikoff, J.G. Performance evaluation of amino acid substitution matrices, *Proteins: Struct. Funct. Genet.*, 17:49–61, 1993.

[2] Jaakkola, T., Diekhans, M. and Haussler, D. Using the fisher kernel method to detect remote protein homologies, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 149–158, Menlo Park, California, 1999.

[3] Murzin, A.G., Brenner, S.E., Hubbar, T., and Chothia, C., SCOP: A structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247:536–540, 1995.

[4] Pearson, W.R., Empirical statistical estimates for sequence similarity searches, *J. Mol. Biol.*, 276:71–84, 1998.