

MultiDomainBenchmark: A Multi-domain Query and Subject Database Suite: Supplementary Material

Hyrum D. Carroll¹, John L. Spouge² and Mileidy Gonzalez²

¹TSYS School of Computer Science, Columbus State University, Columbus, GA, 31907, USA,

²National Center for Biotechnology Information, Bethesda, MD 20894, USA

1 MULTI-DOMAINS IN UNIPROT-SWISSPROT

The UniProt-SwissProt database (downloaded February 27, 2018) has 556,567 sequences. 137,997 of those sequences have a DOMAIN and/or TOPO_DOM Feature Table annotation. 60,311 (or 43.7%) of those sequences have more than one DOMAIN and/or TOPO_DOM annotation, illustrating the prevalence of multi-domain sequences. Of the sequences with DOMAIN and/or TOPO_DOM annotations, the average number of domains is 2.4. For the multi-domain sequences, the average is 4.2. Figure 1 details the distribution of the number of domains per entry.

2 DISTRIBUTION OF TAP- K SCORES

Figure 2 illustrates the the distribution of TAP- k scores for the non-iterative results and Figure 3 iterative results from the Case Study (see the main paper for background details). In each box-and-whisker plot, the whiskers represent the best and worst TAP- k score. Given the broad range of queries, each plot has scores of 0.0. In fact, the lowest number of 0.0 scores is for $k = 1$ for non-iterative HMMER (with 32 scores of 0.0). The most number of 0.0 scores is for $k = 1$ for iterative HMMER (with 92 scores of 0.0). All of the plots include at least one score of 1.0, with the exception of $k = 20$ for iterative HMMER (which has a highest score of 0.99). In each plot, the blue box indicates the first and third quartiles and the thick black line the second quartile (or median).

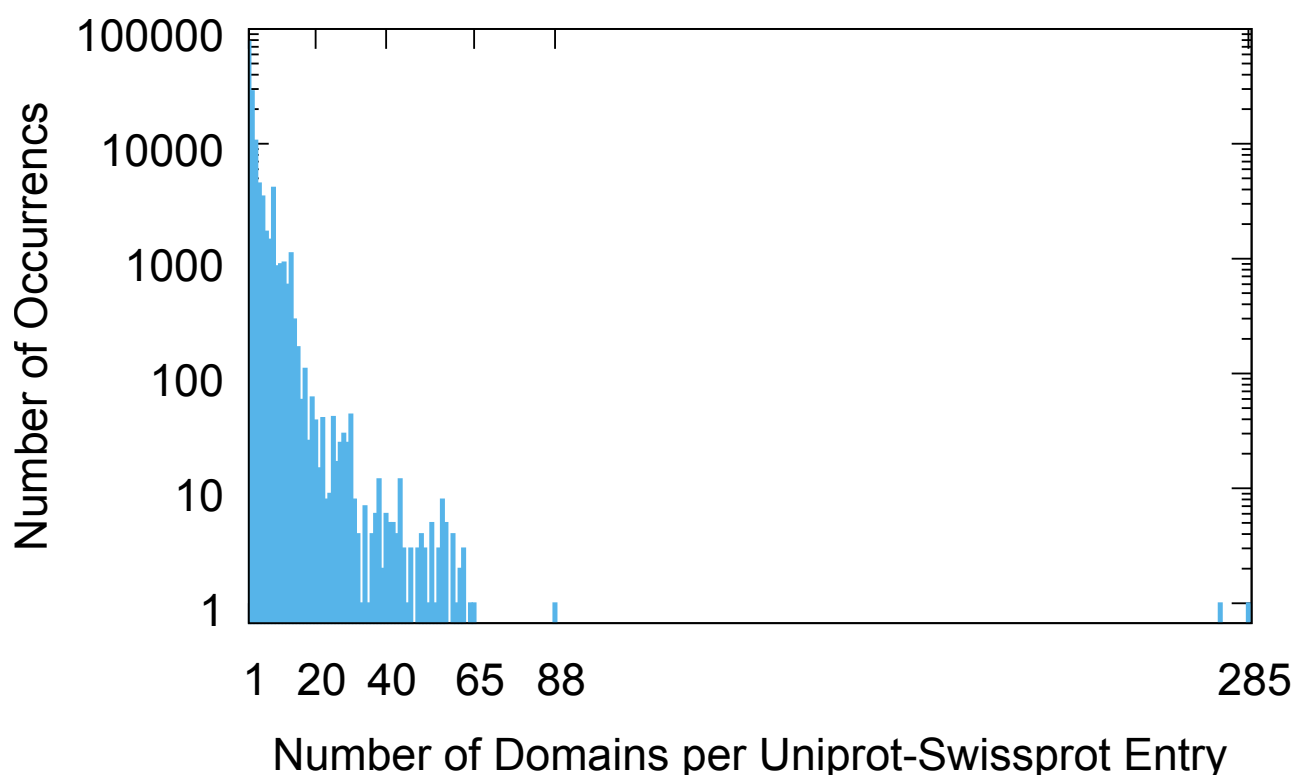


Fig. 1. Histogram of the number of domains per Uniprot-Swissprot entry. Note, the y-axis is logarithmic.

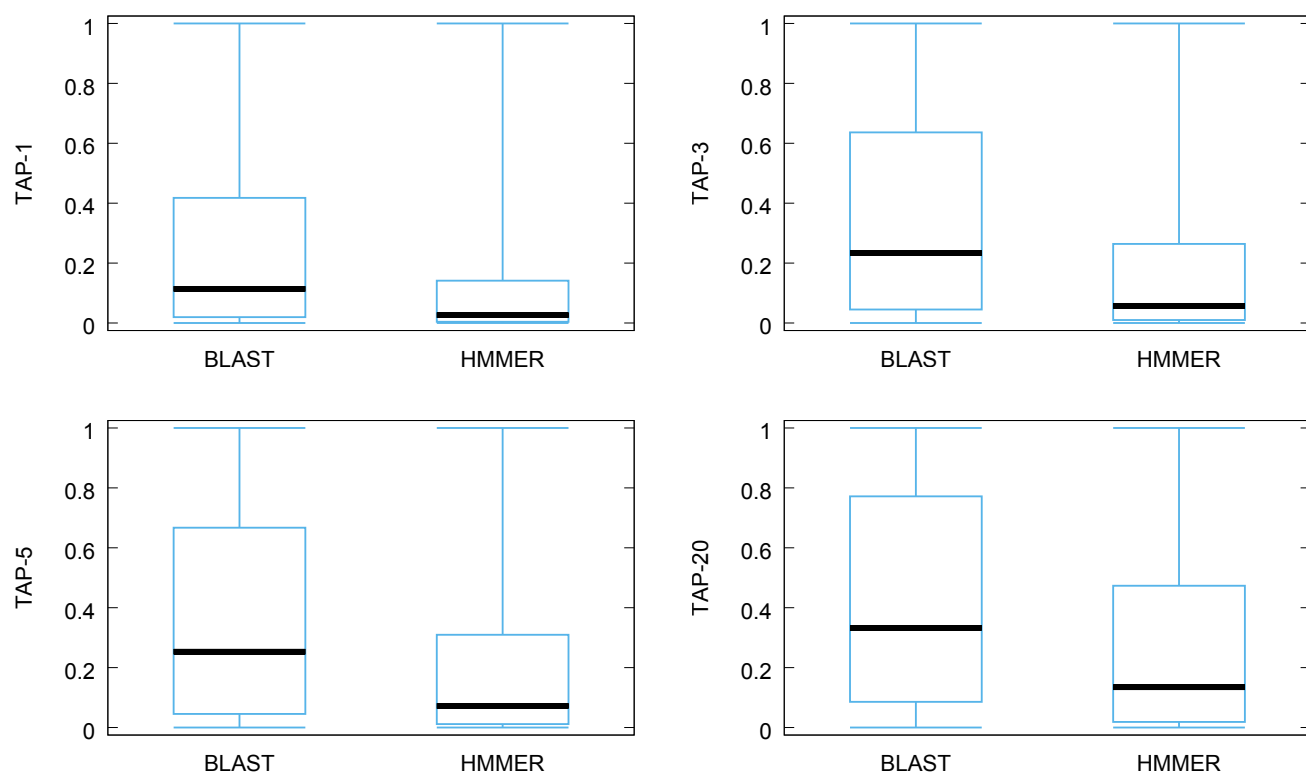


Fig. 2. Box-and-whisker plots of the non-iterative TAP- $\{1,3,5,20\}$ results of BLAST and HMMER for the MultiDomainBenchmark Test queries, searching against the MultiDomainBenchmark target database. The whiskers represent the maximum and minimum TAP scores. The blue box indicates the first and third quartiles and the thick black line the second quartile (or median).

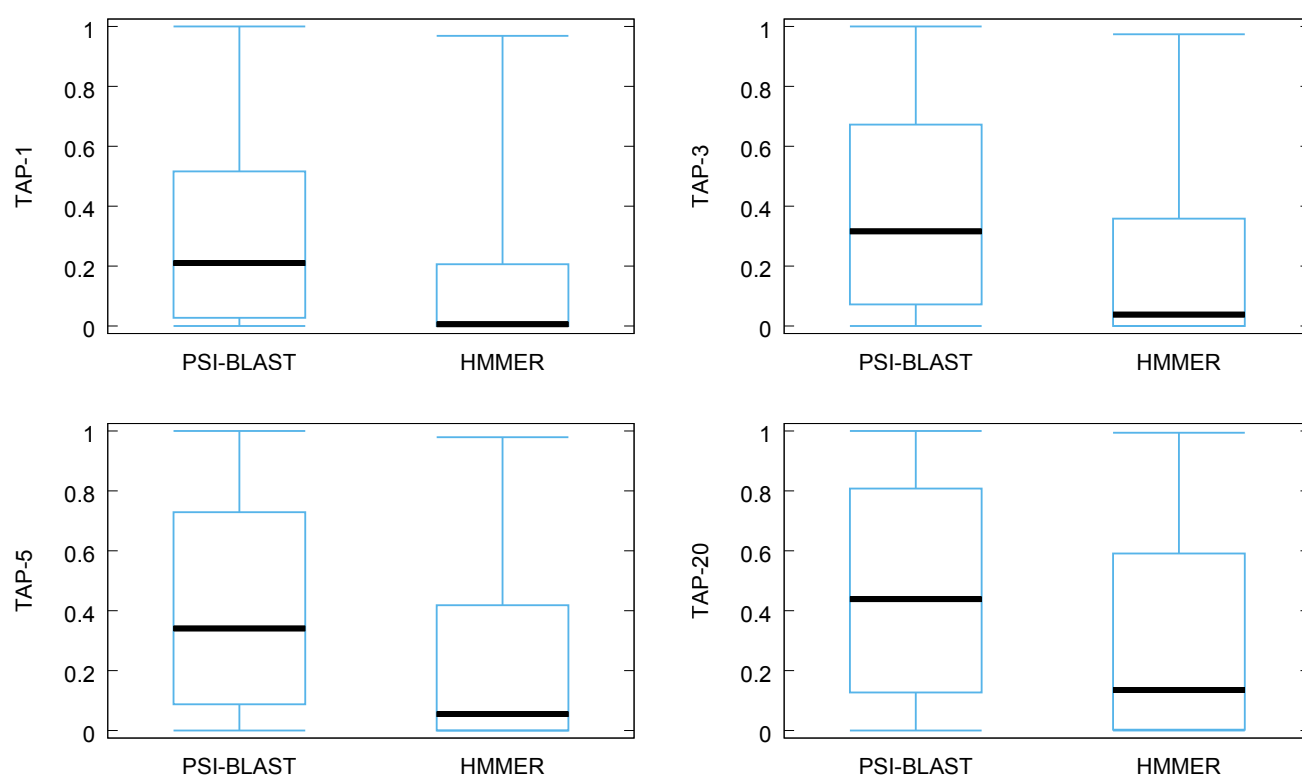


Fig. 3. Box-and-whisker plots of the iterative TAP- $\{1,3,5,20\}$ results of PSI-BLAST and HMMER for the MultiDomainBenchmark Test queries (using the profile generated from searching up to five iterations on a clustered version of the NR database). The whiskers represent the maximum and minimum TAP scores. The blue box indicates the first and third quartiles and the thick black line the second quartile (or median).