shmlast: An improved implementation of Conditional Reciprocal Best Hits with LAST and Python

Camille Scott[1]

[1]University of California, Davis

21 November 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00142
Software Repository: https://github.com/camillescott/shmlast
Software Archive: http://dx.doi.org/10.5281/zenodo.260437

Summary

Conditional Reciprocal Best Hits (CRBH) was originally described by Aubry et al.(Aubry et al. 2014) and implemented in the crb-blast package. CRBH is a method for finding orthologs between two sets of sequences which builds on the traditional Reciprocal Best Hits (RBH) method; it improves RBH by finding an expect-value cutoff per alignment length, and then selecting non-reciprocal alignments which meet the minimum threshold.

Unfortunately, the original implementation uses the relatively slow NCBI BLAST+ (Altschul et al. 1990), and is implemented in Ruby, which requires users to leave the Python-dominated bioinformatics ecosystem. shmlast makes CRBH available to users in Python, while also greatly improving performance by using the LAST aligner (Kiełbasa et al. 2011) for initial homology searches. Other improvements include outputting the list of cutoffs generated by its model along with a plot of the decision boundary to aid in quality control, as well as using gnu-parallel (Tange 2011) to parallelize execution across multiple cores or nodes in a cluster environment.

Methods

RBH is a relatively old method for determining orthologs between two sequence databases. Orthology is distinguished from sequence similarity by descent; while two sequences with high similarity are likely to have structural of functional homology, they are orthologous if they share a common ancestor sequence. It is difficult to know for sure whether two sequences are orthologs, but orthologous groups of sequences are critical for a variety of analyses surrounding structure, function, and evolution, and particularly, for annotation. As such, many methods have been developed for separating high-similarity alignments from their orthologous counterparts. RBH is the regal elder of these methods, and although it is simplistic compared to newer clustering and graph-based methods, it remains in wide use due to its low false-positive rate and ease of implementation. It is performed as follows: given two sets of sequences A and B, sequences $a_i \in A$ and $b_j \in B$ are Reciprocal Best Hits if b_j has the highest scoring sequence alignment in B for a_i and a_i has the highest scoring sequence alignment in A for b_j. a_i and b_j then have a high probability of being orthologs.

While this method works well for finding orthologs between two sets of proteins from different species, it is less effective for annotating newly assembled transcriptomes from existing protein databases. Transcriptomes

are confounded by alternative splicing, causing several transcripts to share subsequences, which may prevent RBH detection between a translated transcript and its protein, even when an orthology relationship exists. Aubry et al., and this implementation, circumvent that problem by first using the reciprocals to establish a score cutoff for each alignment length, and then keeping *any* alignment which passes that cutoff. This prevents alignments with high-likelihood of being orthologs based on sequence identity from being discarded due to the high specificity of RBH.

Performance

shmlast benefits immensely from the use of LAST over BLAST. It scales well by using gnu-parallel, and can be distributed across clusters for particularly large runs.

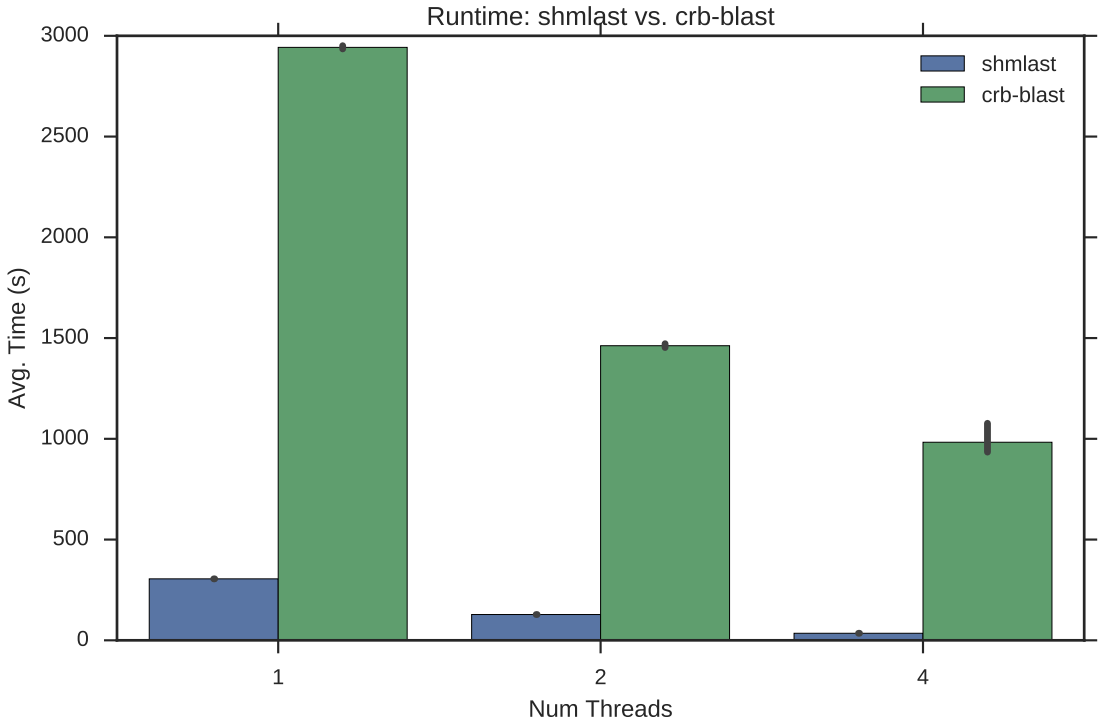

Figure 1: Performance comparison with *Schizosaccharomyces pombe* (Wood et al. 2012) as the query transcriptome and *Nematostella vectensis* (Apweiler, Bairoch, and Wu 2004) as the target proteome.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Apweiler, Rolf, Amos Bairoch, and Cathy H Wu. 2004. "Protein Sequence Databases." *Current Opinion in Chemical Biology* 8 (1): 76–80. doi:10.1016/j.cbpa.2003.12.004.

Aubry, Sylvain, Steven Kelly, Britta M. C. Kümpers, Richard D. Smith-Unna, and Julian M. Hibberd. 2014. "Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans -Factors in Two

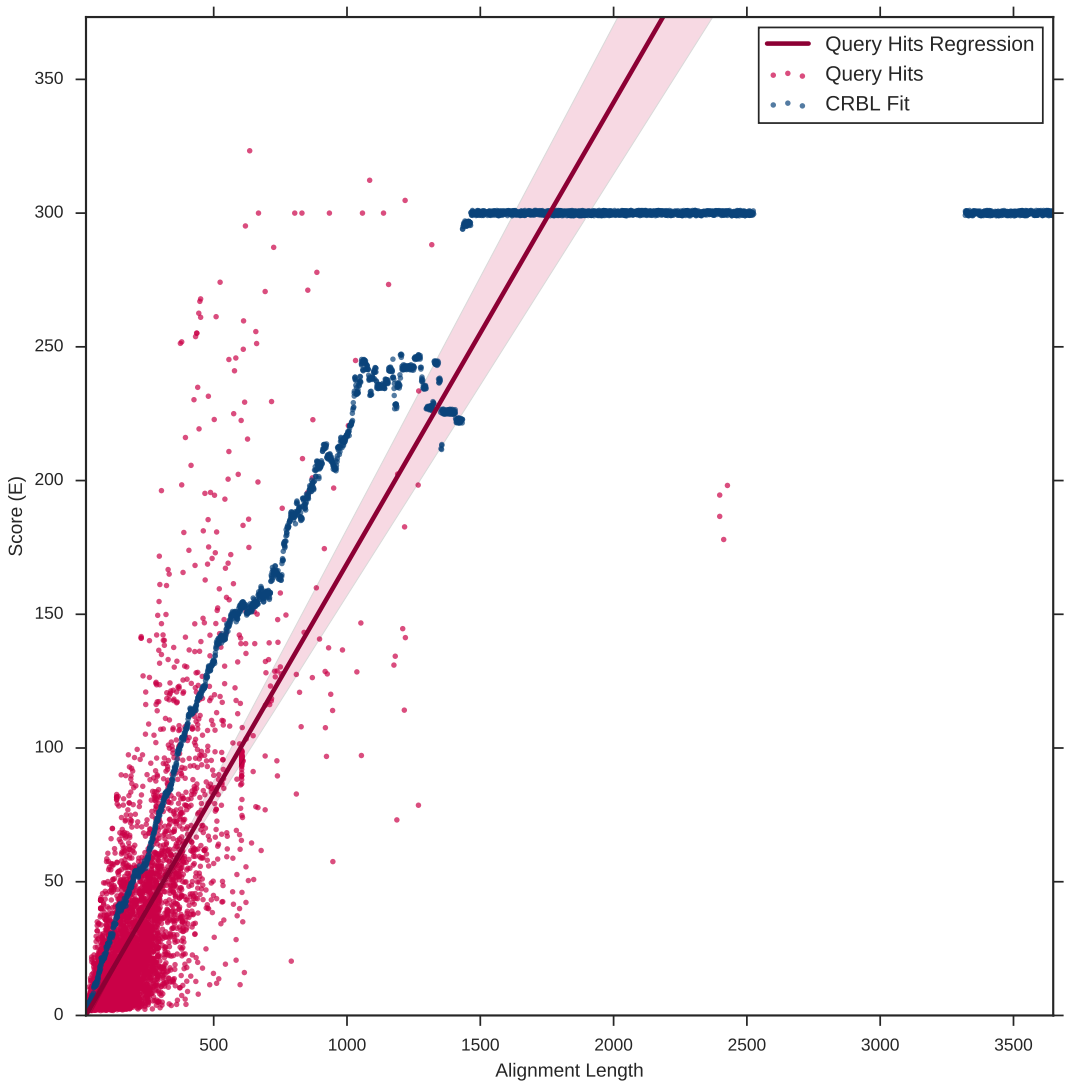

Figure 2: CRBH model generated from the performance comparison. Hits with scores above the blue dotted line will be kept.

Independent Origins of C 4 Photosynthesis." *PLOS Genet* 10 (6): e1004365. doi:10.1371/journal.pgen.1004365.

Kiełbasa, Szymon M., Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. 2011. "Adaptive Seeds Tame Genomic Sequence Comparison." *Genome Research* 21 (3): 487–93. doi:10.1101/gr.113985.110.

Tange, O. 2011. "GNU Parallel - The Command-Line Power Tool."*;Login: The USENIX Magazine* 36 (1): 42–47. doi:10.5281/zenodo.16303.

Wood, Valerie, Midori A. Harris, Mark D. McDowall, Kim Rutherford, Brendan W. Vaughan, Daniel M. Staines, Martin Aslett, et al. 2012. "PomBase: A Comprehensive Online Resource for Fission Yeast." *Nucleic Acids Research* 40 (D1): D695–D699. doi:10.1093/nar/gkr853.