False Discovery Rates for Iterative Database Searching

Alex C. Williams, Anthony G. Davis, Hyrum D. Carroll
Middle Tennessee State University

Introduction

In Bioinformatics, the reliability of a similarity score between a genetic query sequence and a database sequence is represented by an Expectation value, or E-value. Modern homology search algorithms such as BLAST or HMMER, scrutinize the retrieval list with a uniform E-value threshold in order to determine truly significant results. In iterative searching situations such as PSI-BLAST in which the results from one iteration are used as input for the next, this can be especially problematic because the likelihood of a truly irrelevant result being reported as relevant (false positive) increases with the number of performed hits.

While many different aspects of homology search algorithms have been rigorously studied, the inclusion threshold has not received the same attention. In order to improve search sensitivity, we propose the use of the false discovery rate (FDR) as a method for controlling the proportion of false positives. We introduce PSI-BLAST$_{FDR}$, an extended version of the iterative version of BLAST, PSI-BLAST, that uses a FDR method for the threshold criterion. PSI-BLAST$_{FDR}$ achieves 4.90% better retrieval performance than PSI-BLAST on a large test database and a 20.90% better retrieval score for queries belonging to small superfamilies. Furthermore, PSI-BLAST$_{FDR}$ retrieved only 4.3 irrelevant sequences per query compared to 28.7 for PSI-BLAST.

Objectives

The purpose of this study was to effectively reduce the number of irrelevant sequences yielded during the iterative search and retrieval process used in modern genetic database retrieval algorithms. Including too many irrelevant sequences has been shown to lead to search corruption (Gonzalez *et al*, 2010).

Methods

Development of PSI-BLAST$_{FDR}$

PSI-BLAST$_{FDR}$ extends the PSI-BLAST algorithm by replacing the uniform E-value threshold criterion with a false discovery rate controlled threshold. It is based on version 2.2.27 of NCBI's PSI-BLAST.

Controlling the False Discovery Rate

We evaluated four different FDR methods with five values of α. An exemplary FDR control method is the Benjamini-Hochberg method. This method calculates the threshold value for each sequence retrieved and considers the first k ranked sequences as significant that satisfy the following criterion: $P_k \leq k\alpha/m$, where P_k is the P-value of the k^{th} sequence and m is the size of the database searched. Because PSI-BLAST utilizes E-values, and given that E-value = P-value * m, we implemented the Benjamini-Hochberg method as: $E_k \leq k\alpha$ with E_k being the E-value of the k^{th} sequence.

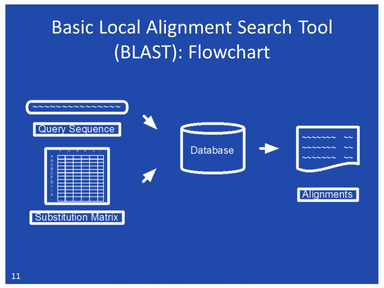

Fig. 1 The BLAST algorithm.

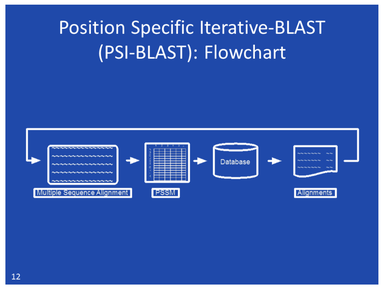

Fig. 2 The PSI-BLAST algorithm.

Retrieval Efficacy and Test Datasets

In this study, we utilize the TAP method (Carroll *et al*, 2010) as the evaluation criterion for retrieval efficacy. The TAP method calculates the median Average Precision-Recall with a moderate adjustment for irrelevant sequences just before the threshold. TAP values range from 0.0 for a retrieval with no relevant sequences to 1.0 for a search that retrieves all of the relevant sequences and only relevant sequences. For testing, we leveraged the query sequences in the ASTRAL40 database (Chandonia *et al*, 2004), a subset of the Structural Classification of Proteins (SCOP) 1.75A database (Murzin *et al*, 1995) with sequences having less than 40% sequence identity to each other.

Results

On a subset of 103 random queries from the training database, PSI-BLAST$_{FDR}$ with the Benjamini-Hochberg method received the best TAP value. On the full training database, we only evaluated α values for the Benjamini-Hochberg FDR method. Of these parameters, PSI-BLAST$_{FDR}$ with α=0.05 received the best TAP of 0.423 while PSI-BLAST received 0.403. Consequently, we adopted this method and α level as the defaults for PSI-BLAST$_{FDR}$. On the test datasets, PSI-BLAST$_{FDR}$ received a TAP value of 0.407 and PSI-BLAST a value of 0.388. In terms of irrelevant records, PSI-BLAST$_{FDR}$ retrieves an average of only 4.3 irrelevant records per query whereas PSI-BLAST retrieves 665.8% more with 28.7 per query.

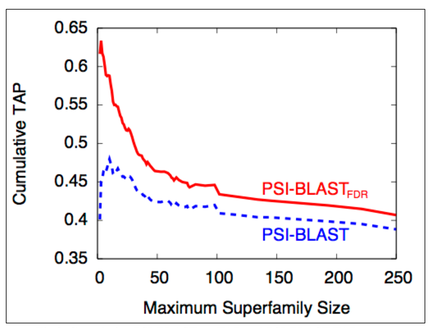

Fig 3. Cumulative PSI-BLAST$_{FDR}$ TAP and PSI-BLAST TAP versus aggregate superfamily size for the test datasets in ASTRAL40.

Conclusions

Using accepted evaluation procedures, PSI-BLAST$_{FDR}$ had a TAP value 4.90% higher than PSI-BLAST on the ASTRAL40 test datasets. This difference is significant given the extremely wide use that PSI-BLAST enjoys. Furthermore, PSI-BLAST$_{FDR}$ is particularly appropriate for queries with small superfamily sizes as evidenced by it obtaining a TAP value 20.90% higher than PSI-BLAST. For queries in larger superfamilies, if the goal is to assign function to a query, then adequately identifying the superfamily is sufficient. For example, retrieving 50% of a large superfamily clearly indicates which superfamily the query belongs. This objective is not currently captured in retrieval evaluation metrics and may make evaluation values misleading for large superfamilies.

While we only addressed FDR's applicability to the BLAST family in this study, additional database search algorithms that use uniform thresholds, such as Jackhmmer (hmmer.janelia.org), could also benefit from the implementation of a FDR controlled threshold. Furthermore, employing more advanced false discovery rate methods, such as the Q-value method (Storey *et al*, 2002) could also yield improvements. Implementation of the Q-value, because it requires the entire distribution of statistical scores, is inherently challenging for a heuristic algorithm like PSI-BLAST.

Acknowledgments

The authors appreciate John L. Spouge for suggesting the idea of this project.

References

Altschul, S., Gertz, E., Agarwala, R., Schaffer, A., and Yu, Y. (2009). PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Research*, 37(3), 815–824.

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. *Nucleic Acids Research*, 25(17), 3389–3402.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Carroll, H. D., Kann, M. G., Sheetlin, S. L., and Spouge, J. L. (2010). Threshold Average Precision (TAP-k): A Measure of Retrieval Efficacy Designed for Bioinformatics. *Bioinformatics*, 26(14), 1708–1713.

Chandonia, J., Hon, G., Walker, N., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Research*, 32(Database Issue), D189–D192.

Gonzalez, M. and Pearson, W. (2010). Homologous over-extension: a challenge for iterative similarity searches. *Nucleic acids research*, 38(7), 2177–2189.