Threshold Average Precision (TAP-k): A Retrieval Efficacy Measure for Bioinformatics

Hyrum D. Carroll[1], Maricel G. Kann[2], Sergey L. Sheetlin[1], and John L. Spouge[1]

[1]National Center for Biotechnology Information, NLM, NIH, Bethesda, MD 20894, USA; [2]University of Maryland, Baltimore County, Baltimore, MD 21250, USA

Introduction

Database retrieval is central to bioinformatics and therefore assessment of its performance is crucial for many studies. Two of the most widely used measures are the ROC$_n$ and the pooled ROC$_n$ scores. A ROC$_n$ score is the area under a receiver operating characteristic (ROC) curve truncated at n irrelevant ("false positive") records [1]. A pooled ROC$_n$ score is calculated in the same manner, after all of the results are "pooled" together [2]. While these scores have been invaluable to assessing relative algorithmic performance, they have two main drawbacks:

1. Unreliability because they do not account for typical usage of database retrieval,
2. Susceptibility to being skewed by a single query.

As a solution to these deficiencies, we present the retrieval measures Threshold Average Precision (TAP) and TAP-k (for k median errors per query). TAP's and TAP-k's are based on the average precision measure from information retrieval with an additional term to account for trailing irrelevant records [3].

Retrieval with HMMER [4] and PSI-BLAST [5] provide compelling examples in favor of the TAP methods. First, the distribution of E-values corresponding with the 50-th irrelevant record does not reflect actual usage in bioinformatics. Second, an example of pooled ROC$_n$ scores being skewed by a single query. TAP's and TAP-k's do not suffer from these drawbacks and provide an intuitive analyses of retrieval results.

Methods

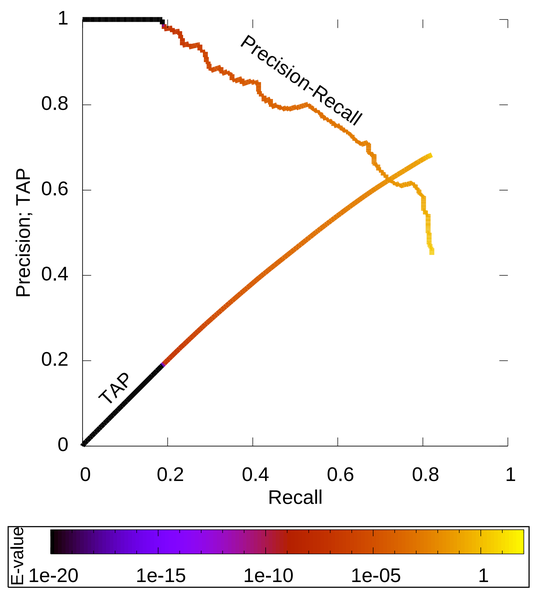

Figure 1. The Precision-Recall and TAP curves, with threshold E-values E_0 indicated with a color gradient.

The TAP and TAP-k measures leverage key concepts from information retrieval. The TAP is the average precision up to a threshold E_0 (usually an E-value) of a retrieval q plus an additional term of the precision at the threshold, $p(E_0)$, to account for trailing irrelevant records:

$$\bar{p}\,(E_0; q) = \frac{1}{T_q + 1}\left[\sum_{m=1}^{j(E_0)} p(m) + p(E_0)\right] \qquad (1)$$

Methods, cont'd

where T_q is the total number of relevant ("true positive") records for q and $j(E_0)$ is the rank of the last relevant record with a statistical score of E_0 or lower. Figure 1 shows an example of a Precision-Recall plot for PSI-BLAST retrieval results of a single query. Additionally, it illustrates Equation 1 with the TAP shown for all thresholds up to E_0.

While the TAP is the score for a single retrieval, the TAP-k is the query-average of multiple TAP's, each with the common threshold E_k for k median errors per query. The TAP-k is then $\bar{p}\,(E_k(A))$ for an algorithm A. A different E_k should be calculated for each algorithm tested to mitigate differences in their respective statistical calculations.

Results

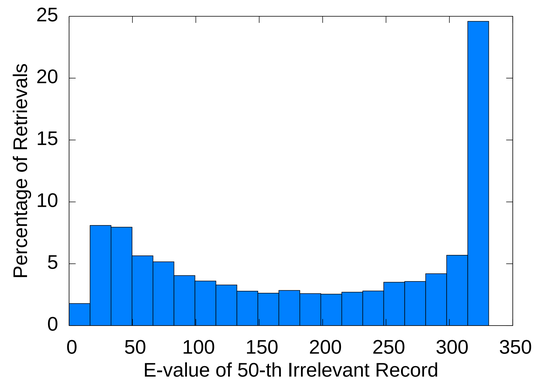

Figure 2. Histogram of the HMMER E-values of the 50-th irrelevant record for each query in the database of 331 queries.

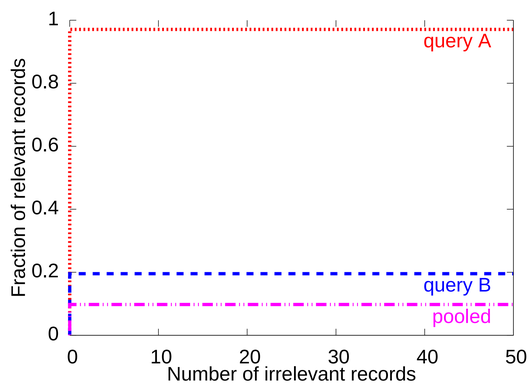

Figure 3. Individual ROC$_{50}$ curves, along with the corresponding pooled ROC$_{50}$ for the Homoserine Dehydrogenase Pfam family. Note that the pooled ROC$_{50}$ curve is lower than either of the queries.

ROC$_n$ and pooled ROC$_n$ analyses have two main drawbacks as retrieval measures. First, the threshold of n irrelevant records does not necessarily reflect actual usage in bioinformatics. Figure 2 quantifies this by reporting the occurrences of the 50-th irrelevant record for HMMER retrievals from DB_331_CDD [3]. Note, the most common value is the size of the database.

Another drawback is that pooled ROC curves can be skewed by the retrieval results from a single query. Figure 3 illustrates such pooled ROC$_{50}$ retrieval results for two queries from a single family. Note, that the pooled ROC$_{50}$ curve is skewed (i.e., below both of the ROC$_{50}$ curves of the queries). Figure 4 clearly illustrates that the aggregate TAP curve is query-averaged, having values between the TAP curves each of the queries.

Results, cont'd

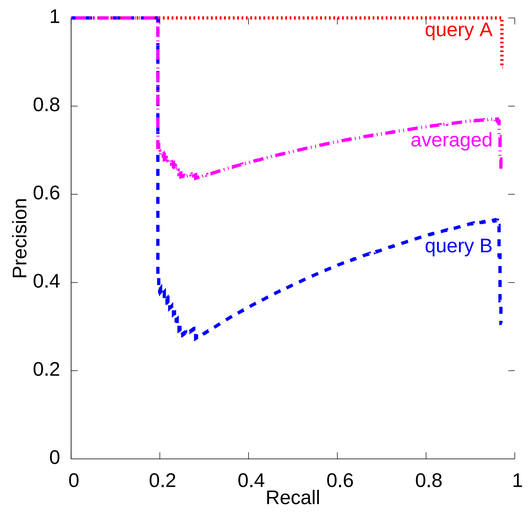

Figure 4. Precision-Recall curves (and average of the two) for the same queries as in Figure 3. The TAP for each curve is the normalized average precision, with the precision of last record repeated.

Conclusion

Retrieval is a common procedure in bioinformatics. Therefore, accurate evaluation methods have a large impact in several research areas. Unfortunately, the commonly used method for database retrieval 1) does not necessarily reflect actual usage in the field and 2) can be skewed by a single query. The Threshold Average Precision (TAP) and TAP-k are measures of retrieval results that reflect actual usage in bioinformatics. Furthermore, the TAP-k uses a query-average to aggregate scores to protect it against being skewed by a single query.

Availability

The TAP and TAP-k web server and downloadable Perl script are freely available at http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/tap/.

Funding

Funded by the Intramural Research Program of the National Library of Medicine (NIH) and partially by NIH 1K22CA143148, MGK (PI).

Contact Information

HDC: HyrumCarroll@gmail.com, JLS: spouge@ncbi.nlm.nih.gov

References

1. Gribskov, M., Robinson, N.L., (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching, Computers and Chemistry, **20**:1, 25–33.
2. Schäffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices, *Bioinformatics*, **15**:12, 1000–1011.
3. Carroll, H.D., Kann, M.G., Sheetlin, S.L., Spouge, J.L. (2010) Threshold Average Precision (TAP-k): A Measure of Retrieval Efficacy Designed for Bioinformatics, *Bioinformatics*, **26**:14, 1708–1713.
4. Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics*, **14**:9, 755–763.
5. Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., et al. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucleic Acids Research*, **29**:14, 2994–3005.