

Measuring the Average Accuracy Performance of Homology Search

XiaoLei Chen¹ Adam Cannane¹ Hugh E. Williams¹
xchen@cs.rmit.edu.au cannane@cs.rmit.edu.au hugh@cs.rmit.edu.au

¹ Bioinformatics Search Group, School of Computer Science and Information Technology, RMIT University, GPO Box 2476V, Australia, 3001

Keywords: homology search, sensitivity, selectivity, evaluation measures

1 Introduction

Homology searches of genomic databases are the essential step for new gene discovery. While novel search techniques are being developed to address this need, it is equally important to provide a unified experimental methodology for evaluating and comparing the different homology search methods.

Current evaluation measures such as the Receiver Operating Characteristic (ROC) [5], the Coverage Versus Errors per query (CVE) plot [3], and the average precision (AP) measure [4], provide biologists with sensitivity and selectivity tradeoffs for specific queries but fail to provide an average performance analysis for typical queries.

We propose a rapid and unbiased evaluation metric, recall-EPQ, to compare the average accuracy performance of different search methods.

2 Recall-EPQ

We propose a novel effectiveness measurement scheme to assess the average performance of homology search systems, called *recall-EPQ*. The recall-EPQ measure combines the merits of recall-precision and CVE measures, yet it avoids their shortcomings. As recall and EPQ are important indicators for sensitivity and selectivity, this measure better represents the average selectivity performance of a method at different levels of accuracy—recall. Importantly, the representational bias within the CVE plot is addressed by treating each superfamily equally in the recall-EPQ plot.

The recall-EPQ graph is based on 11 standard recall levels: 0%, 10%, 20%, ... 100%. For each query i , we measure the EPQ_i at each recall point r . The number of homologues, a_i , required at all recall points for a query is calculated by multiplying the superfamily size of the query S_i by the recall point. More formally, for n queries,

$$EPQ_i(r) = \frac{1}{n} \sum_{j=1}^x b_j \tag{1}$$

where $b_j = 1$ for non-homologues otherwise $b_j = 0$. x represents the rank of the homologous result that equals or exceeds a_i . In addition, we measure the mean EPQ over all queries n for each recall point r .

3 Results

We applied our novel effectiveness measure—recall-EPQ—to the results from five different search methods BLAST2 [1], BLASTNF [1] (BLAST2 with no filtering), FASTA3 [6] with ktup=1, FASTA3 [6] with ktup=2, and SSEARCH [6]. We also applied the ROC₅₀, CVE plot, and the AP measure to each of the results from the five search methods. The Structural Classification of Proteins (SCOP) database [2] is used as the basis of our test collection, PDB159-90.

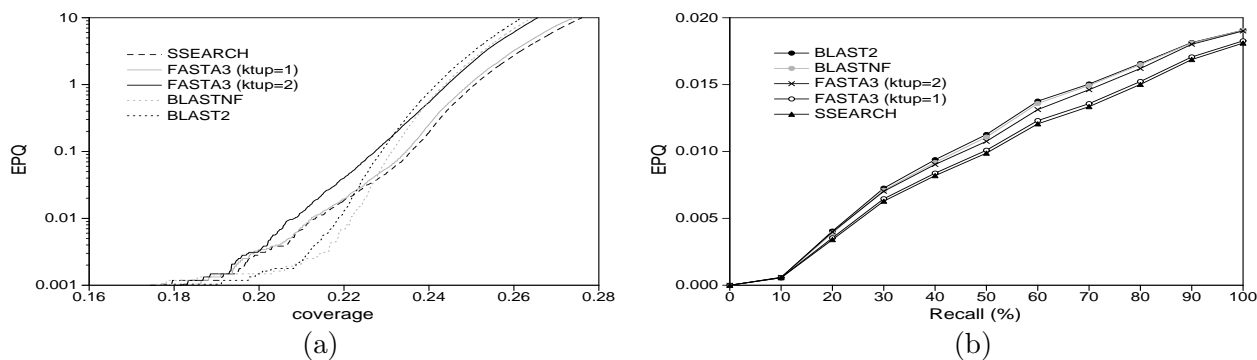


Figure 1: (a) CVE-plot and (b) recall-EPQ measures using the PDB159-90 collection

Table 1: Comparison of search techniques using Mean AP and ROC₅₀ for the PDB159-90 collection.

Measures	Matrices	SSEARCH	FASTA3 (ktup=1)	FASTA3 (ktup=2)	BLASTNF	BLAST2
Mean AP	BLOSUM62 (11,1)	0.4352	0.4257	0.4059	0.4068	0.3989
ROC ₅₀	BLOSUM62 (11,1)	0.4569	0.4471	0.4230	0.4233	0.4151

Figure 1 shows a side-by-side comparison of the CVE-plot and the recall-EPQ plot for the five search schemes on the test collection. The CVE-plot illustrates how the curves for separate schemes intersect one another at different levels of coverage. The recall-EPQ has smoother curves as it measures the average accuracy performance. Table 1 shows the single-value metrics (mean AP and ROC₅₀) for each of the search schemes. The differences between any two search methods in Table 1 are statistically significant, verified using the Wilcoxon signed ranked test.

The evaluation results from the various measures are consistent and give a powerful indication of sensitivity and selectivity of different search techniques. Our results suggest that any one of the measurement schemes can differentiate between the accuracy of the search schemes. However, we believe that our novel search measure, recall-EPQ, provides the best visual representation for biologists by presenting the sensitivity as an error-rate for different levels of selectivity (recall).

References

- [1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
- [2] A. Andreeva, D. Howorth, S.E. Brenner, T.J.P. Hubbard, C. Chothia, and A. G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, 32:D226–D229, 2004.
- [3] S.E. Brenner, C. Chothia, and T.J.P. Hubbard. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proceedings of the National Academy of Sciences USA*, 95:6073–6078, 1998.
- [4] Z. Chen. Assessing sequence comparison methods with the average precision criterion. *Bioinformatics*, 19(18):2456–2460, 2003.
- [5] M. Gribskov and N.L. Robinson. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Computers & Chemistry*, 20:25–33, 1996.
- [6] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences USA*, 85:2444–2448, 1988.