# Using fast homology search tools for protein sequence functional annotation: a comparison

Pronozin Artem, Genaev Michail, Afonnikov Dmitry
*IInstitute of Cytology and Genetics SB RAS, Novosibirsk, Russian Federation*

## ABSTRACT

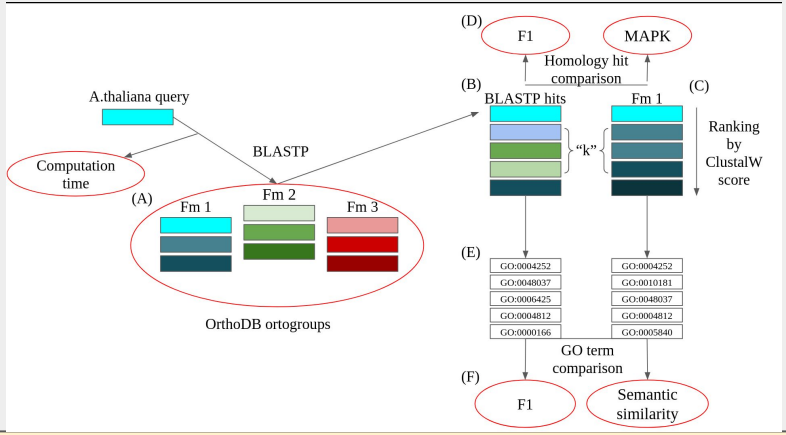
Annotation of the protein sequences by homology search and GO term transfer from highly homologous sequences is an important task for current genome and transcriptome sequencing projects. However, large size of sequence databases make homologous sequence search difficult in reasonable time. There exist tools that apply fast and ultrafast database search algorithms to find sequence homologs. These tools usually apply various heuristics for fast determining possible sequence matches. This result in different results of these programs with respect to returned set of homologous sequences and their rankings. These differences may lead to differences in the sets of GO terms and lead to errors in query sequence function annotation.

## INTRODUCTION

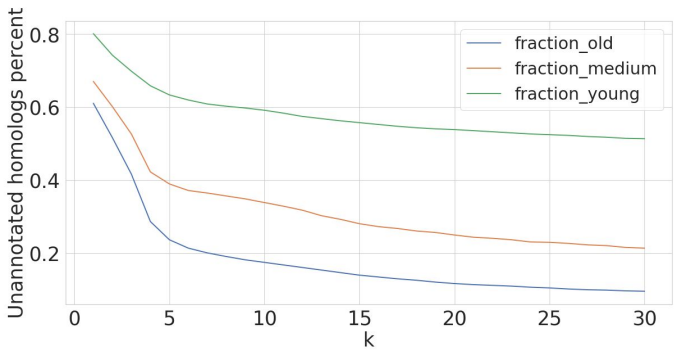
Due to ever-growing amount of new data being analyzed together with an exponential increase of database sizes makes protein similarity analysis using existing tools a dauntingly slow and inefficient task [1]. To solve the task of fast searching of the query homologs in the large sequence databases a number of tools developed. Results of these programs with respect to returned set of homologous sequences and their rankings. In this work we compare several sequence search tools for protein sequences (BLAST, BLAST fast, Diamond, Usearch ublast, Usearch local, Mmseq2) in their ability to identifying and score ranking highly homologous sequences. We tested these tools using all Arabidopsis thaliana proteins. As the search database we used OrthoDB.

## MATERIALS AND METHODS

- 27,636 sequences of the A. thaliana proteins from TAIR v10.
- We used orthologous genes from other organisms, as query homologs, which perform similar functions. The orthologous sequences identified from the OrthoDB database v10.
- We compared the following homology search programs: BLASTP and BLASTP-fast from BLAST v.2.9.0 package, Diamond, ‘ublast’ and ‘usearch_local’ algorithms from Usearch v11 , Mmseq2.
- ClustalW, was used to align homologous sequences from respective orthogroups with query and rank them by similarity score as standard.



## ANNOTATION QUALITY RESULTS

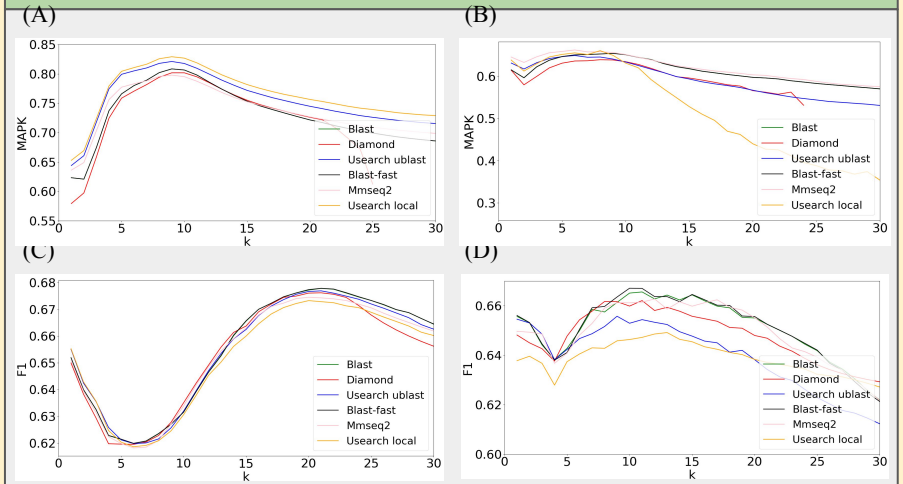



The proportion of query sequences for which among the k nearest hits received by the BLASTP program, there were not a single one annotated with GO terms. X axis represents k values. Y axis, the proportion of query sequences for which hits did not have the GO annotation.

## TIME TEST RESULTS

| | Indexing, min | Search, min | Index size, GB |
|---|---|---|---|
| Blast | 22 min 45 s | 91 h 6 min | 22 |
| BlastFast | 22 min 45 s | 9 h 81 min | 22 |
| Usearch local | 53 min 3 s | 10 min | 75 |
| UBLAST | 47 min 8 s | 5 h 23 min | 98 |
| Diamond | 3 min 45 s | 11 min | 17 |
| Mmseq2 | 2 min 5 s | 33 min | 19 |

## METRIC ANALYSIS RESULTS




X axis *k* value. Y axis MAPK and F1 for GO values, respectively. A, B, panels shows mean MAPK values for old and young group, respectively. C, D old and young group - F1 for GO, respectively. Different color lines respond for programm.

- Metric MAPK - for old group (A) minimal values of performance metrics checked on k =1 (0.65). The performance increase for hit number ranging up to *k*=10 (0.80-0.85) and gradually decreased. For young group (B), value stay on stable position (0.65) on k = 1-10, after this point gradually decrease.
- F1 for GO - old and young group show similar behavior: height vaule on k = 1 (0.64-0.66). On k = 3-10, observed minimum, on k = 20 maximum. Difference between old and young group that minimum for for young k = 5 smaller then for old k = 5 -10.

## SUMMARY

These work shows, dependencies of different metrics of orthologs identification accuracy and their annotations on the number of nearest homologs is not monotonous. MAPK show clear accuracy maximum on k = 10 and particularly not depends on accuracy metric and gene age. The presence of such a peak can be explained by the fact that for small k the probability of coincidence of sequence identifiers for orthologs and hits is small. With k increasing the number of terms are increasing as well, that means the probability of finding identifiers from the list of orthologs increases and reaches its maximum on k = 10. With a further increase in k, sequences with weaker homology add to hits list, as a result proportion of matching terms / identifiers against the background of their increasing total number, decrease.

Most significantly, the age of genes affects the accuracy of identifying orthologs. Compared to the old genes, the accuracy of the identification of orthologs for young genes is lower and decreases more sharply with increasing k. The greatest accuracy is achieved for small values of k.

Analysis of the operating time of the programs showed that Mmseq2 has the most optimal operating parameters (building an index and searching for orthologists).

## REFERENCES

[1] A. Conesa,S. Götz,J. García-Gómez,J.Terol,M. Talón,andM.Robles, “Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research,” Bioinformatics, 21(18), pp. 3674-3676, 2005.
[2] R. Vaser,, D. Pavlović, and M. Šikić, “SWORD—a highly efficient protein database search,” Bioinformatics, 32.17 , pp. i680-i684, 2016.
[3] Z. Mustafin,et al. “Phylostratigraphic Analysis Shows the Earliest Origination of the Abiotic Stress Associated Genes in A. thaliana,” Genes, 10.12 , pp. 963, 2019.
[4] E. Kriventseva, et al. “OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs,” Nucleic acids research, 47.D1, pp. D807-D811, 2019.
[5] S. Altschul,et al. “Basic local alignment search tool,” Journal of molecular biology,215.3, , pp. 403-410, 1990.
[6] B. Buchfink, C. Xie, and D. Huson, “Fast and sensitive protein alignment using DIAMOND,” Nature methods, 12.1,, 59, 2015.
[7] E. Usearch, “Lawrence Berkeley National Laboratory (LBNL),” Berkeley, CA (United States) (2010).
[8] M. Hauser, M. Steinegger, and J. Söding, “MMseqs software suite for fast and deep clustering and searching of large protein sequence sets,”Bioinformatics, 32.9,pp. 1323-1330, 2016.
[9] J. Thompson,, G. Desmondand T.Gibson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” Nucleic acids research, 22.22, pp. 4673-4680, 1994.
[10] N. Pentreath, “Machine learning with spark,” Packt Publishing Ltd, 2015.
[11] V. Rijsbergen CJ, “Information retrieval,” 2nd edn. Butterworths, London, 1979.
[12] C. Pesquita, “Semantic similarity in the gene ontology,” The gene ontology handbook. Humana Press, New York, NY, 2017. 161-173