OXFORD

Genome analysis

InParanoid-DIAMOND: faster orthology analysis with the InParanoid algorithm

Emma Persson (ID) **and Erik L. L. Sonnhammer** (ID) *

Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, 17121 Solna, Sweden

*To whom correspondence should be addressed.
Associate Editor: Tobias Marschall

Abstract

Summary: Predicting orthologs, genes in different species having shared ancestry, is an important task in bioinformatics. Orthology prediction tools are required to make accurate and fast predictions, in order to analyze large amounts of data within a feasible time frame. InParanoid is a well-known algorithm for orthology analysis, shown to perform well in benchmarks, but having the major limitation of long runtimes on large datasets. Here, we present an update to the InParanoid algorithm that can use the faster tool DIAMOND instead of BLAST for the homolog search step. We show that it reduces the runtime by 94%, while still obtaining similar performance in the Quest for Orthologs benchmark.
Availability and implementation: The source code is available at (https://bitbucket.org/sonnhammergroup/inparanoid).
Contact: erik.sonnhammer@scilifelab.se
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Orthologs are commonly defined as genes in different species having a common ancestry (Fitch, 1970). They tend to have similar functions (Altenhoff *et al.*, 2012), and are therefore highly useful in several areas of bioinformatics. Identification of orthologs is a challenging task, and many tools and algorithms of different types have been developed for ortholog prediction (Altenhoff *et al.*, 2020; Linard *et al.*, 2021).

InParanoid is a well-known algorithm for inferring orthologs, based on all-versus-all sequence comparison with BLAST (Altschul *et al.*, 1990), which has been shown to perform well in benchmarks with a good balance between specificity and recall (Altenhoff *et al.*, 2016). However, the speed of the InParanoid algorithm has become an increasing problem with the need for running it on ever larger collections of species with completely sequenced genomes, such as the Quest for orthologs (Altenhoff *et al.*, 2020) reference proteomes or the set of reference proteomes from UniProt (UniProt Consortium, 2021).

In recent years several efforts have been made to develop tools to perform sequence search faster than the BLAST algorithm (Altschul *et al.*, 1990), while retaining the high accuracy (Buchfink *et al.*, 2021; Steinegger and Söding, 2017). By switching out the underlying sequence search tool used in the InParanoid algorithm from BLAST to DIAMOND in the new InParanoid-DIAMOND algorithm, it can perform much faster orthology analysis with similar results as the original InParanoid tool.

2 Implementation

InParanoid-DIAMOND is a command line application written in perl, utilizing the same codebase as InParanoid-BLAST, but using DIAMOND (v2.0.8) (Buchfink *et al.*, 2021) for homolog searching. For detailed implementation and methods, see Supplementary Methods.

3 Results

To optimize the settings for InParanoid-DIAMOND, we compared the resulting orthologs (Supplementary Fig. S1) and runtimes of InParanoid-BLAST and InParanoid-DIAMOND run on the QFO proteomes for different sensitivity levels (Supplementary Table S1). The results show that InParanoid-DIAMOND with the very-sensitive option results in the best tradeoff between runtime and similarity to the InParanoid-BLAST results, with a reduced runtime of 94% compared to InParanoid-BLAST, and an average Jaccard index of 0.74. As seen in Supplementary Figure S2, InParanoid-BLAST detected more ortholog pairs than InParanoid-DIAMOND for most species. On average, 17% of the InParanoid-BLAST ortholog pairs were not detected, while on average 9% of the ortholog pairs were not detected by InParanoid-BLAST. The less sensitive DIAMOND options provide a greater speedup to InParanoid-DIAMOND, but at the cost of a lower similarity to InParanoid-BLAST. The more sensitive option 'ultra-sensitive' results in a slight increase in jaccard index, but at the cost of tripling the runtime in CPU hours compared to the very-sensitive option. Running with composition-based statistics similar to BLAST resulted in a significantly increased runtime, yet the resulting ortholog pairs were less similar to InParanoid-BLAST.

Seeing that the composition-based statistics option in DIAMOND resulted in longer runtimes and lower similarity to the InParanoid-

Table 1. Ortholog prediction methods InParanoid-DIAMOND, InParanoid-BLAST, the three variants of SonicParanoid, Proteinortho, OrthoFinder default and with MSA (using DIAMOND for similarity search), their runtime in hours on the Quest for Orthologs reference proteomes on a 48-thread node, the runtime in CPU hours and the number of appearances on the Pareto frontier out of the 11 Orthology benchmark service tests

Ortholog prediction method	Runtime (hours)	Runtime (CPU hours)	Appearances on Pareto frontier
InParanoid-DIAMOND (very-sensitive)	10.49	135.8	6
InParanoid-BLAST	166.5	5940.1	5
SonicParanoid default	2.44	100.8	2
SonicParanoid sensitive	7.23	325.4	4
SonicParanoid mostSensitive	15.68	725.4	7
Proteinortho 6	1.69	74.2	4
OrthoFinder 2 default	2.75	97.8	7
OrthoFinder 2 default+MSA	22.07	749.2	5

BLAST algorithm, the original 2-pass approach (Ostlund *et al.*, 2010) appears unsuitable for DIAMOND. An example of this can be found in Supplementary Figure S3, where running *Homo sapiens* versus *Escherichia coli* with 1- and 2-pass strategies resulted in a lower overlap with InParanoid-BLAST for 2-pass, which also suffered from a 9-fold increase in runtime (see Supplementary Table S2). The lower overlap agrees with the fact that enabling composition-based statistics in DIAMOND results in fewer homologs detected (data not shown). The massive increase in runtime can be explained by DIAMOND not coping well with executing a large number of single query runs in the second pass, as well as longer runtimes in general when using the comp-based-stats option. In light of these results, we have set the default in InParanoid to run DIAMOND in one pass, with the 'very-sensitive' sensitivity option, and composition-based statistics disabled.

Assessing the differences between orthologs found by InParanoid-DIAMOND and InParanoid-BLAST, we could see that small differences in the homologs found by BLAST and DIAMOND can result in different constellations of ortholog groups, potentially resulting in very large differences in the number of ortholog pairs. An analysis and discussion on orthologs uniquely detected when using one of the tools can be found in Supplementary Results.

Since public results in the benchmark on the QFO reference proteomes 2020 (see Supplementary Table S4) were not available for all methods in the comparison when this analysis was done, the 2018 benchmark was used. Assessing ortholog prediction quality with the Orthology benchmark service showed that InParanoid-DIAMOND and InParanoid-BLAST have similar performances—InParanoid-BLAST appeared on the Pareto frontier in five of the eleven benchmark tests, while InParanoid-DIAMOND was on the Pareto frontier in six of the tests (Table 1). A breakdown on the appearances on the Pareto frontier for the different tests can be found in Supplementary Table S3. Overall, InParanoid-DIAMOND is placed close to InParanoid-BLAST, generally getting slightly lower recall but higher precision (Supplementary Fig. S4). Compared to the variations of SonicParanoid (Cosentino and Iwasaki, 2019), InParanoid-DIAMOND appears on the Pareto frontier more often than the variants default and sensitive, while the most sensitive option performed slightly better than InParanoid-DIAMOND in the Orthology benchmark, but had a longer runtime. Proteinortho (Lechner *et al.*, 2011) performed better than all other tools in terms of speed, but was on the Pareto frontier less often than a majority of the tools. OrthoFinder (Emms and Kelly, 2019) with the MSA option was both slower and less often on the Pareto frontier than InParanoid-DIAMOND. On the other hand, the default version was on the Pareto frontier in one more test than InParanoid-DIAMOND, and also faster. It is worth noting however that this tool generally tends to have considerably lower precision and higher recall than InParanoid-DIAMOND, potentially making it a suitable tool to use when coverage is higher valued than low error rate, while

InParanoid-DIAMOND retains a similar balance between recall and precision as InParanoid-BLAST.

In conclusion, we have shown that InParanoid-DIAMOND reduces the runtime by 94% while achieving similar results to InParanoid-BLAST in terms of detected ortholog pairs and benchmark performance, making it much better suited for large scale orthology detection.

Funding

Data availability

The data underlying this article can be found at http://www.ebi.ac.uk/reference_proteomes/.

References

Altenhoff,A.M. *et al.* (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.*, **8**, e1002514.

Altenhoff,A.M. *et al.*; Quest for Orthologs Consortium. (2016) Standardized benchmarking in the quest for orthologs. *Nat. Methods*, **13**, 425–430.

Altenhoff,A.M. *et al.* (2020) The Quest for Orthologs benchmark service and consensus calls in 2020. *Nucleic Acids Res.*, **48**, W538–W545.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Buchfink,B. *et al.* (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.

Cosentino,S. and Iwasaki,W. (2019) SonicParanoid: fast, accurate and easy orthology inference. *Bioinformatics*, **35**, 149–151.

Emms,D.M. and Kelly,S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Lechner,M. *et al.* (2011) Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, **12**, 124.

Linard,B. *et al.* (2021) Ten years of collaborative progress in the Quest for Orthologs. *Mol. Biol. Evol.*, **8**, 20.

Ostlund,G. *et al.* (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

UniProt Consortium. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.