Broccoli: combining phylogenetic and network analyses for orthology assignment

Authors

Romain Derelle [1]*, Hervé Philippe [2,3] and John K. Colbourne [1]

[1]. School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.
[2]. Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321, Moulis 09200, France.
[3]. Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC, Canada H3C 3J7.

* Corresponding author: r.derelle@bham.ac.uk

ABSTRACT

Orthology assignment is a key step of comparative genomic studies, for which many bioinformatic tools have been developed. However, all gene clustering pipelines are based on the analysis of protein distances, which are subject to many artefacts. In this paper we introduce Broccoli, a user-friendly pipeline designed to infer, with high precision, orthologous groups and pairs of proteins using a phylogeny-based approach. Briefly, Broccoli performs ultra-fast phylogenetic analyses on most proteins and builds a network of orthologous relationships. Orthologous groups are then identified from the network using a parameter-free machine learning algorithm. Broccoli is also able to detect chimeric proteins resulting from gene-fusion events and to assign these proteins to the corresponding orthologous groups. Tested on two benchmark datasets, Broccoli outperforms current orthology pipelines. In addition, Broccoli is scalable, with runtimes similar to those of recent distance-based pipelines. Given its high level of performance and efficiency, this new pipeline represents a suitable choice for comparative genomic studies.
Broccoli is freely available at https://github.com/rderelle/Broccoli.

Keywords

orthology, orthologous groups, label propagation algorithm, LPA, gene fusions

Introduction

Orthologous genes are genes originating from a speciation event, as opposed to paralogous genes originating from a gene duplication event (Koonin 2005). The identification of either orthologous pairs or orthologous groups of genes (i.e. independent sets of orthologs found at a given taxonomic level) is the primary step of most comparative genomic studies, since it provides genetic equivalences between species. For instance, the extrapolation of functional genetic discoveries made from experimental model species to distantly related species, including to humans in medicine and in environmental toxicology, requires a precise mapping of orthologs across species.

Assigning gene orthology across distantly related species typically consists of identifying ancient speciation and gene duplication events from the comparisons of present gene or protein sequences. This task is highly challenging for many reasons. The combination of successive speciation and gene duplication events, with the latter often being associated with gene losses and gene conversions (Kondrashov 2012, Pich and Kondrashov 2014, Harpak, Lan et al. 2017), tends to blur the distinction between orthologs and paralogs. In addition, incomplete lineage sorting (Maddison 1997), and the transfers of genetic material between species (i.e. lateral gene transfers) (Soucy, Huang et al. 2015) and between genes (i.e. gene-fusions) (Zmasek and Godzik 2012), all create complex reticulate gene histories. Finally, the heterogenous evolutionary rate of proteins, with known variations across species and over time (Dorus, Vallender et al. 2004, Kawahara and Imanishi 2007), and gene prediction errors (e.g. missing, truncated or fused genes) are also important sources of background noise in orthology inferences.

Current *de novo* clustering algorithms are all based on the analysis of pairwise protein distances. Two main approaches have been proposed: distances can be analysed (i) using the best bi-directional hits (BBH) approach or one of its derivative to infer orthologous pairs (Huynen and Bork 1998, Schreiber and Sonnhammer 2013, Sonnhammer and Ostlund 2015, Cosentino and Iwasaki 2019), or (ii) using the Markov Cluster algorithm (MCL) to infer orthologous groups from the network of similarities (Dongen 2000, Li, Stoeckert et al. 2003, Emms and Kelly 2015), orthologous groups that can further be analysed using phylogenetic analyses and a species tree reconciliation approach to infer orthologous pairs (DM Emms 2018). The BBH approach is highly precise but is inclined to miss orthologous pairs due to its highly constrained nature (Dalquen and Dessimoz 2013). By contrast, the MCL approach is generally inclusive but unavoidably merges orthologous groups with high sequence similarity, thus lacks precision. Finally, it is important to note that similarity distances are always an underestimate of the true evolutionary distances due to the saturation of sequences, making it difficult for these distance-based approaches to resolve ancient gene histories.

As an alternative, the use of phylogenetic analyses as a first step has been proposed (Gabaldon 2008). The basic principle of this approach is to build a phylogenetic tree for each protein and its similarity hits, and to infer orthologous relationships based on the taxonomic distribution of hits in the trees. The delineation between orthologs and paralogs is made here from the analysis of phylogenetic relationships rather than protein distances (Huerta-Cepas, Dopazo et al. 2007, Vilella, Severin et al. 2009, Huerta-Cepas, Capella-Gutierrez et al. 2014). The promise of this ‘phylogeny-based’ approach at improving orthology inferences has three important caveats: (i) the many hundreds of thousands of phylogenetic analyses required by this approach must be computationally efficient, (ii) new methods for the delineation of orthologous groups must be proposed and (iii) a phylogeny-based pipeline must be made freely available to the research community.

Here we introduce Broccoli, an open-source pipeline for *de novo* orthology assignment using a phylogeny-based approach. Briefly, Broccoli performs ultra-fast phylogenetic analyses and extracts successively two sets of orthologous relationships from the trees. The first set is used to build an orthology network (as opposed to networks of similarity distances), from which orthologous groups are identified using a label propagation algorithm. Then a more precise second set is defined to identify pairs of orthologous genes within each orthologous group.

The performance of Broccoli was assessed by using a custom benchmark dataset for orthologous groups, and the Quest of Orthologs 2011 benchmark dataset for orthologous pairs (Altenhoff, Boeckmann et al. 2016, Glover, Dessimoz et al. 2019). In these tests, we compared Broccoli to recent distance-based pipelines combined with fast similarity search algorithms (e.g. DIAMOND, MMseq2 (Buchfink, Xie et al. 2015, Steinegger and Soding 2017)) since blastp, which is two orders of magnitude slower, would not be usable for large datasets.

Materials and methods

Broccoli is a pipeline written in Python 3 that requires the ete3 library (Huerta-Cepas, Dopazo et al. 2010). It is composed of four steps as summarized in Figure 1A and described below. The rationale of Broccoli is that, since single gene trees are expected to be too inaccurate to directly infer orthology relationships, as many trees as the number of sequences will be inferred (Steps 1 and 2) and orthology will be inferred from the consensus of information extracted from these multiple trees using a network analysis (Step 3).

Step 1: kmer pre-clustering

The objective is to simplify proteomes without loss of information and therefore to decrease the computational time of steps 2 and 3. Broccoli first converts the protein names into unique identifiers. The proteome of each species is then independently clustered using kmers of amino-acids. For each cluster of sequences, the longest one is retained for further analysis while others are set-aside and are re-injected into the orthologous groups and orthologous pairs at the corresponding steps. This step aims at reducing the number of proteins to be analysed by removing allelic variants and ‘recent’ duplicates. By default, the kmer size is set to 100 amino-acids. This high value prevents the grouping of paralogs between closely related species. But the kmer size can be reduced when distantly related species are analysed (e.g. species belonging to different eukaryotic supergroups).

Step 2: similarity searches and phylogenetic analyses

Broccoli then builds a phylome (i.e. the set of gene trees (Huerta-Cepas, Dopazo et al. 2007)) for each species by comparing its proteins against other proteomes and performing phylogenetic analyses in possible cases of gene duplications. For each simplified proteome, similarity searches against all proteomes are individually performed using DIAMOND under the 'most-sensitive' option and the *N* best hits per species are reported (*N* is set to 6 by default). Then, for each query protein, all its hits are considered orthologs to each other if no species have multiple hits (referred thereafter as 'set-aside orthologous pairs'). Otherwise, the DIAMOND pairwise alignments between the query and each of its target sequences are combined together to build a trimmed alignment by allowing a fraction *g* of missing data per position (*g* is set to 0.7 by default). The trimmed alignment is then analysed using FastTree2 (Price, Dehal et al. 2010) to produce a BioNJ tree that is rooted using the midpoint method.

To our knowledge, it is the first time DIAMOND (or blastp) alignments are used to perform phylogenetic analyses instead of classical multiple sequence alignments (MSA). The main advantage of this approach is a considerable decrease of the computational time since alignments are already computed during the similarity searches. But the use of these pairwise alignments also have two additional advantages: (i) only sequence fragments matching the query sequence are used for phylogenetic analyses while MSA, which operate on full length sequences, usually include unaligned blocks that create phylogenetic noise, and (ii) short sequences are often misaligned in MSA but not in pairwise alignments.

Step 3: identification of orthologous groups

In this third step, Broccoli builds an orthology network from which orthologous groups are isolated using a machine learning algorithm. Broccoli first delineates orthologous groups in each rooted tree using a relaxed 'species overlap' approach as defined in (Huerta-Cepas, Dopazo et al. 2007). Briefly, the trees are traversed from the query protein to the root and, at each node, the taxonomic composition of the two sets of leaves emerging from that node are compared (an example is provided in Figure 1B). The two sets of leaves are considered part of the same orthologous group if (i) there is no common species between the two sets (i.e. no 'overlap' species) or (ii) there is only one common species and at least two unique species in both sets (i.e. species not present in the other set). Broccoli identifies the deepest node of the tree fulfilling this 'species overlap' criteria, and builds orthologous pairs between all leaves belonging to that node and also paralogous pairs between these leaves and all remaining leaves of the tree.

The orthologous and paralogous pairs extracted from all trees are then combined with the 'set-aside orthologous pairs' from Step 2, to build an undirected network of orthologous relationships. An edge between two proteins A and B is formed if they have been identified (i) orthologs at least twice (since at the very least A has been compared to B and B compared to A) and (ii) more often as orthologs than as paralogs. The edge weight w(AB) from A to B is defined as:

$$w(\mathrm{AB}) = \frac{ortho(AB)}{\max\_ortho(B)}$$

where 'ortho(AB)' corresponds to the number of times A and B have been identified as orthologs, and 'max_ortho(B)' corresponds to the maximum number of times B has been found to be an ortholog with any other protein. Therefore, the weight w(AB), ranging from infinitesimal to 1, represents the relevance of the orthologous relationships between A and B with respect to the reference node B. The weights are asymmetric since w(AB) might be different from w(BA).

Given the fast phylogenetic analyses, tree rooting and orthologous group delineations performed by Broccoli, the orthology network is expected to be noisy. But one can expect that truly orthologous proteins will be much more often connected and with higher weights among themselves than with paralogous proteins. A label propagation algorithm (LPA; (Raghavan, Albert et al. 2007)) is applied to the orthology network to identify node communities (i.e. orthologous groups). The LPA used here, described in Supplemental Material 1, is asynchronous and weighted (using the asymmetric edge weights described above), resulting in a highly precise community delineation. An example is given in Figure 1C, in which the 'green' node is assigned by Broccoli to the 'red' community due to the high relevance of its orthologous relationships with this community (an unweighted LPA would assign this node to the 'purple' community with which it has more connexions). This algorithm is also fast, with convergence of the labels being reached after only a few generations (Supplemental Material 1) and, in the absence of any random choice, fully deterministic.

Finally, two types of error corrections are applied to the detected communities (i.e. orthologous groups). First, Broccoli attempts to remove spurious hits, which are defined as proteins having less than *n* proteins of the orthologous group in their own similarity hits (*n* is set to 2 by default; connected components of three or less proteins are not subject to LBA and corrections). Proteins considered as spurious hits are then removed from their orthologous group, and therefore from the classification. Second, since proteins are initially assigned to a unique orthologous group, Broccoli aims at detecting gene-fusions and corrects the classification accordingly. Proteins resulting from gene-fusion events are detected among nodes connected to several communities using the method described in Supplemental Material 1. Proteins that are identified as chimeric proteins are then added to all orthologous groups involved in their corresponding fusion event.

Step 4: identification of orthologous pairs

While orthologous relationships were extracted at Step 3 to delineate orthologous groups, Broccoli builds a new set of orthologous relationships that considers gene duplication events within each orthologous group. The method here is the same as described in Step 3 but with two differences: (i) proteins not belonging to the orthologous group are first removed from the 'set-aside orthologous pairs' and from the rooted trees, and (ii) orthologous and paralogous pairs are built at each tested node from the rooted trees – not only at the deepest node fulfilling the species overlap criteria. Finally, for each pair of proteins A and B belonging to this orthologous group, a ratio R(AB) is calculated as

$$R(\mathrm{AB}) = \frac{ortho(AB)}{ortho(AB) + \ para(AB)}$$

where 'ortho(AB)' and 'para(AB)' represents the number of times A and B have been found as orthologs and as paralogs respectively. The two proteins will thus be reported as orthologs if their ratio R is superior to a threshold *r* (*r* is set to 0.5 by default).

Perfomance tests

The paraBench dataset was built from an in-house collection of phylogenetic markers. The dataset and the performance metrics are fully described in Supplemental Material 2. As opposed to the benchmark of orthologous pairs, the performance metrics were calculated considering all possible pairs within each orthologous group. The dataset, reference clustering, python script to compute the performance metrics and the results obtained in this study are all available at https://github.com/rderelle/paraBench.

In addition, the Quest for Orthologs benchmark '2011 dataset' was downloaded from the EBI ftp server, then analysed by Broccoli using the 'not_same_sp' option. The resulting orthologous pairs were submitted back to the QfO website. Results obtained from other clustering pipelines and databases were downloaded from the QfO website. Finally, the performance metrics were calculated as described in (DM Emms 2018). The orthologous pairs obtained with Broccoli under default parameters were submitted to the Quest for Orthologs website as 'Public Results'.

The versions of the pipelines and programs used in this study are indicated in Supplementary Material 3.

Running time analyses

The running time analyses were performed using 8 CPUs of an Intel Xeon Gold 6248 processor and 40 GB of RAM memory. 64 fungal proteomes, each representing a unique taxonomic order, were downloaded from NCBI ftp server and combined to build eight datasets containing 8 to 64 proteomes. The list of fungal species is available in Supplemental Material 3.

Results

Benchmark of orthologous groups

The quality assessment of orthologous group predictions was performed using a custom-built benchmark dataset (named 'paraBench') comprising 17 eukaryotic species and 52 orthologous groups (see Supplemental Material 2). In this benchmark, we compared Broccoli to two recent distance-based pipelines: OrthoFinder2 (DM Emms 2018), which uses the MCL algorithm after distance corrections to mitigate the impact of evolutionary rate differences between species, and Sonicparanoid (Cosentino and Iwasaki 2019), which employs a BRH approach. Broccoli produced the highest recall score value, closely followed by OrthoFinder2, thanks to its distance corrections (Figure 2; Supplemental Material 3). Finally, Sonicparanoid, which over-split orthologous groups due to the stringency of the BRH approach, scored the lowest. On the precision side, Broccoli also performed better than the two other pipelines, with a nearly perfect score of 0.975. OrthoFinder2 scored the lowest precision value indicating a tendency to over-merge closely related orthologous groups. Running Sonicparanoid using the 'most-sensitive' option as recommended for distantly related species yielded a slightly different protein clustering, yet achieving the same performance metrics. Overall, Broccoli scored the highest on this performance benchmark (F1-score in Figure 2).

Benchmark of orthologous pairs

We tested the orthologous pairs predicted by Broccoli (Step 4) by using the large-scale Quest for Orthologs 2011 benchmark dataset (referred thereafter as QfO 2011 dataset). The QfO 2011 results were combined into two groups corresponding to (i) analyses based on reference gene phylogenies (TreeFam-A and SwissTree databases) and (ii) analyses based on reference species trees (the Species Tree Discordance Benchmark, STDB, and its generalised version GSTDB).

Broccoli showed both high precision and recall when orthologous pairs were compared to reference gene phylogenies (Figure 3; Supplemental Material 3). Indeed, Broccoli's results were better than those of the database PhylomeDB (Huerta-Cepas, Capella-Gutierrez et al. 2014), which also identifies orthologous pairs using a phylogeny-based approach but using complicated phylogenetic analyses, and its results were similar to those of the database MetaphOrs (Pryszcz, Huerta-Cepas et al. 2011), which creates consensus orthologous pairs from several public databases. As expected, by varying the ratio of orthology (Step 4) the sensitivity and specificity of Broccoli can be adjusted: an increase of this ratio improved the precision while a decrease improved the recall. On the other hand, and as seen in the paraBench results, OrthoFinder2 scored high at recall but low on precision while BBH-like methods (RSD, RBH-BBH, Sonicparanoid and Hieranoid2) produced higher precision but lower recall scores. These results highlighted again the limitations of distance-based methods, with the mutually exclusive choice between precision and recall. OrthoFinder2 with MSA produced the highest F1-score, although this option of OrthoFinder2 is associated with long running times (see efficiency tests below), followed closely by Broccoli (Supplemental Material 3).

When orthologous pairs were compared to reference species phylogenies (i.e. SBD and SGBD benchmarks), Broccoli scored higher recall values but lower precision values than OrthoFinder2 (Supplemental Material 3). This result was unsurprising considering that OrthoFinder2 infers orthologous pairs using a species tree reconciliation approach, thereby maximizing these precision values, while Broccoli uses a simple species overlap approach that does not rely on species relationships. Yet more importantly, a gene tree might very well be different from the species tree for reasons that are explained in the introduction, hence questioning the relevance of this benchmark.

Gene-fusions

In the absence of a specific gene-fusion benchmark, it is difficult to assess the quality of the predictions made by Broccoli, in particular the proportion of missed gene-fusions. Nevertheless, a total of 1402 proteins were predicted to be the result of gene-fusion events from the QfO 2011 dataset (representing 0.2% of all proteins). The number of fused proteins per species was highly heterogeneous, ranging from 0 for *Korarchaeum cryptofilum* to 163 for *Branchiostoma floridae*. Broccoli was able to identify chimeric proteins that resulted from gene-fusions from up to five orthologous groups and genes-fusions events shared by up to sixteen proteins. The latter case corresponding to the known fusion of two tRNA synthetases shared by most metazoan species (Ray, Sullivan et al. 2011). However, Broccoli failed to recover the well-known fusion of the dihydrofolate reductase and thymidylate synthase proteins present in 'unikonts' and absent in most 'bikonts' (Cavalier-Smith 2003). This known gene-fusion that is missed by Broccoli can be explained by the low number of 'unikonts' compared to 'bikonts' in the QfO 2011 dataset (7 and 33 respectively), with fused and non-fused proteins interpreted by Broccoli as full-length and partial proteins respectively.

Running time analyses

Considering the large number of phylogenetic analyses performed by Broccoli (e.g. 501,570 phylogenies for the QfO 2011 dataset), it is expected to be several orders of magnitude slower than distance-based pipelines. Yet compared to two recent distance-based algorithms by using fungal proteome datasets (from 8 to 64 species), Broccoli showed high efficiency and scalability as it was found on average 13% faster than OrthoFinder2 (excluding the two smallest datasets for which the differences between the two pipelines were higher; Figure 4). On the other hand, OrthoFinder2 with the MSA option was by far the slowest pipeline, and Sonicparanoid, which only performs half of the similarity searches (Cosentino and Iwasaki 2019), was the fastest with a near linear running time.

The high speed of Broccoli was achieved thanks to the parallelization of most tasks, the ultra-fast phylogenetic analyses that is implemented (an average of 11.3 phylogenetic analyses per second and CPU on the QfO 2011 dataset, from DIAMOND outputs to rooted trees) and an efficient network analysis. Due to the high computational time required by the similarity searches, Step 2 was the most time-consuming step of the Broccoli pipeline, representing 85% to 95% of the total running times in the analyses of fungal datasets (data not shown). Finally, should orthologous groups be the desired output, OrthoFinder2 would be found slightly faster than Broccoli (423mn for OrthoFinder2 compared to 438mn for Broccoli respectively using the 64 fungal dataset).

Discussion

In this study, we introduced and tested a new phylogeny-based pipeline for orthology assignment. Since high-throughput phylogenetic analyses are challenging and time-consuming, the main idea behind Broccoli's design is to perform ultra-fast phylogenetic analyses (i.e. pairwise alignments, simple trimming, NJ trees, midpoint rooting), and to rely on a performant community detection algorithm for the identification of relevant orthologous relationships. Broccoli has achieved this objective, as it obtained both high recall and precision values on two benchmark datasets. Noticeably, none of these benchmarks consider the chimeric nature of genes. Therefore, the performances of Broccoli with respect to other pipelines are presumably underestimated.

With a small subset of proteins being assigned to several orthologous groups, the clustering generated by Broccoli lies between classical gene classifications and protein domain subdivisions (e.g. Pfam database (El-Gebali, Mistry et al. 2019)). We believe that Broccoli, which combines high performances and efficiency, will be suitable for most comparative genomic studies. Given its high level of precision, we also recommend this pipeline for the identification of phylogenetic markers in phylogenomic studies. Finally, Broccoli should also greatly facilitate evolutionary studies of chimeric genes created by gene-fusions, since it is the only available pipeline able to detect these proteins alongside orthologous groups (see also (Pathmanathan, Lopez et al. 2018)).

Given the large variety of analyses performed by this pipeline (kmer clustering, phylogenetic analyses and network analysis), there are combinations of parameters that have not been tested, and parts that have not been fully optimised (e.g. trimming of the alignments, species overlap criteria). We are continuing to improve Broccoli by investigating parameters that should provide greater performances. In addition, its relatively high efficiency leaves much room for the implementation of more complex analyses. In its current form, Broccoli categorizes proteins (i.e. ortholog, chimeric) but does not infer evolutionary events (i.e. gene duplications, gene-fusions), which would require a reference species tree. We plan to implement an automatic species tree reconstruction using the supermatrix method (de Queiroz and Gatesy 2007), that will enable Broccoli to predict these evolutionary events as well.

Acknowledgments

The authors thank Franz Lang and Luisa Orsini for their valuable comments on an earlier version of the manuscript. The authors also wish to thank Toni Gabaldón for his seminal work on the phylogeny-based approach. Broccoli was developed as part of the DeepEuk collaborative project. Romain Derelle was supported by a NERC Highlight Topic grant (NE/N006216/1). John Colbourne was supported by UK NERC award Cracking the Code of Adaptive Evolution (NE/N016777/1).

Figure Legends

Figure 1: Key aspects of Broccoli

A: Overview of the pipeline. Data, external programs and processes are coloured in blue, orange and green respectively. **B**: Example of the species overlap approach on a gene-tree obtained from 3 species A, B and C. The nodes fulfilling the species overlap criteria are indicated by orange dots, and the resulting orthologous group is delineated by the orange rectangle. **C**: Example of the label propagation, with labels represented by colours (green, purple and red). The node with the green label is about to exchange its label with one of its neighbours. The fractions present on each edge represent the weights AB where A is the green node and B its neighbour. The green node takes the red label since the sum of the ‘red weights’ is higher than the sum of the ‘purple weights’.

Figure 2: Benchmark of orthologous groups (paraBENCH dataset).

Pipelines were ranked by their performance metric from left (highest value) to right (lowest value).

Figure 3: Benchmark of orthologous pairs (SwissTree + TreeFam-A; QfO 2011 dataset).

Broccoli was tested with its *r* parameter (Step 4) ranging from 0.3 to 0.7 in 0.1 increment. Pipelines performing open clustering have been classified here as ‘database’ since all species of the QfO 2011 dataset are present in their reference database (e.g. Orthoinspector (Nevers, Kress et al. 2019)). Note that the scales of the two axes are different.

Figure 4: Efficiency tests.
Runtimes were obtained using eight CPUs and a dataset composed of 64 fungal species. Pipelines were ran using default parameters unless specified otherwise.

References

Altenhoff, A. M., B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca, K. Forslund, J. Huerta-Cepas, B. Linard, C. Pereira, L. P. Pryszcz, F. Schreiber, A. S. da Silva, D. Szklarczyk, C. M. Train, P. Bork, O. Lecompte, C. von Mering, I. Xenarios, K. Sjolander, L. J. Jensen, M. J. Martin, M. Muffato, c. Quest for Orthologs, T. Gabaldon, S. E. Lewis, P. D. Thomas, E. Sonnhammer and C. Dessimoz (2016). "Standardized benchmarking in the quest for orthologs." Nat Methods **13**(5): 425-430.

Buchfink, B., C. Xie and D. H. Huson (2015). "Fast and sensitive protein alignment using DIAMOND." Nat Methods **12**(1): 59-60.

Cavalier-Smith, T. (2003). "Protist phylogeny and the high-level classification of Protozoa." European Journal of Protistology **39**(4): 338-348.

Cosentino, S. and W. Iwasaki (2019). "SonicParanoid: fast, accurate and easy orthology inference." Bioinformatics **35**(1): 149-151.

Dalquen, D. A. and C. Dessimoz (2013). "Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals." Genome Biol Evol **5**(10): 1800-1806.

de Queiroz, A. and J. Gatesy (2007). "The supermatrix approach to systematics." Trends Ecol Evol **22**(1): 34-41.

DM Emms, S. K. (2018). "OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences." BioRxiv.

Dongen, S. V. (2000). "Graph clustering by flow simulation." Ph.D thesis, University of Utrecht, The Netherlands.

Dorus, S., E. J. Vallender, P. D. Evans, J. R. Anderson, S. L. Gilbert, M. Mahowald, G. J. Wyckoff, C. M. Malcom and B. T. Lahn (2004). "Accelerated evolution of nervous system genes in the origin of Homo sapiens." Cell **119**(7): 1027-1040.

El-Gebali, S., J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto and R. D. Finn (2019). "The Pfam protein families database in 2019." Nucleic Acids Res **47**(D1): D427-D432.

Emms, D. M. and S. Kelly (2015). "OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy." Genome Biol **16**: 157.

Gabaldon, T. (2008). "Large-scale assignment of orthology: back to phylogenetics?" Genome Biol **9**(10): 235.

Glover, N., C. Dessimoz, I. Ebersberger, S. K. Forslund, T. Gabaldon, J. Huerta-Cepas, M. J. Martin, M. Muffato, M. Patricio, C. Pereira, A. S. da Silva, Y. Wang, E. Sonnhammer and P. D. Thomas (2019). "Advances and Applications in the Quest for Orthologs." Mol Biol Evol **36**(10): 2157-2164.

Harpak, A., X. Lan, Z. Gao and J. K. Pritchard (2017). "Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates." Proc Natl Acad Sci U S A **114**(48): 12779-12784.

Huerta-Cepas, J., S. Capella-Gutierrez, L. P. Pryszcz, M. Marcet-Houben and T. Gabaldon (2014). "PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome." Nucleic Acids Res **42**(Database issue): D897-902.

Huerta-Cepas, J., H. Dopazo, J. Dopazo and T. Gabaldon (2007). "The human phylome." Genome Biol **8**(6): R109.

Huerta-Cepas, J., J. Dopazo and T. Gabaldon (2010). "ETE: a python Environment for Tree Exploration." BMC Bioinformatics **11**: 24.

Huynen, M. A. and P. Bork (1998). "Measuring genome evolution." Proc Natl Acad Sci U S A **95**(11): 5849-5856.

Kawahara, Y. and T. Imanishi (2007). "A genome-wide survey of changes in protein evolutionary rates across four closely related species of Saccharomyces sensu stricto group." BMC Evol Biol **7**: 9.

Kondrashov, F. A. (2012). "Gene duplication as a mechanism of genomic adaptation to a changing environment." Proc Biol Sci **279**(1749): 5048-5057.

Koonin, E. V. (2005). "Orthologs, paralogs, and evolutionary genomics." Annu Rev Genet **39**: 309-338.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos (2003). "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome Res **13**(9): 2178-2189.

Maddison, W. P. (1997). "Gene Trees in Species Trees." Systematic Biology **46**(3): 523-536.

Nevers, Y., A. Kress, A. Defosset, R. Ripp, B. Linard, J. D. Thompson, O. Poch and O. Lecompte (2019). "OrthoInspector 3.0: open portal for comparative genomics." Nucleic Acids Res **47**(D1): D411-D418.

Pathmanathan, J. S., P. Lopez, F. J. Lapointe and E. Bapteste (2018). "CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection." Mol Biol Evol **35**(1): 252-255.

Pich, I. R. O. and F. A. Kondrashov (2014). "Long-term asymmetrical acceleration of protein evolution after gene duplication." Genome Biol Evol **6**(8): 1949-1955.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). "FastTree 2--approximately maximum-likelihood trees for large alignments." PLoS One **5**(3): e9490.

Pryszcz, L. P., J. Huerta-Cepas and T. Gabaldon (2011). "MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score." Nucleic Acids Res **39**(5): e32.

Raghavan, U. N., R. Albert and S. Kumara (2007). "Near linear time algorithm to detect community structures in large-scale networks." Phys Rev E Stat Nonlin Soft Matter Phys **76**(3 Pt 2): 036106.

Ray, P. S., J. C. Sullivan, J. Jia, J. Francis, J. R. Finnerty and P. L. Fox (2011). "Evolution of function of a fused metazoan tRNA synthetase." Mol Biol Evol **28**(1): 437-447.

Schreiber, F. and E. L. L. Sonnhammer (2013). "Hieranoid: hierarchical orthology inference." J Mol Biol **425**(11): 2072-2081.

Sonnhammer, E. L. and G. Ostlund (2015). "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic." Nucleic Acids Res **43**(Database issue): D234-239.

Soucy, S. M., J. Huang and J. P. Gogarten (2015). "Horizontal gene transfer: building the web of life." Nat Rev Genet **16**(8): 472-482.

Steinegger, M. and J. Soding (2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nat Biotechnol **35**(11): 1026-1028.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin and E. Birney (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." Genome Res **19**(2): 327-335.
Zmasek, C. M. and A. Godzik (2012). "This Deja vu feeling--analysis of multidomain protein evolution in eukaryotic genomes." PLoS Comput Biol **8**(11): e1002701.

Figures

Figure1

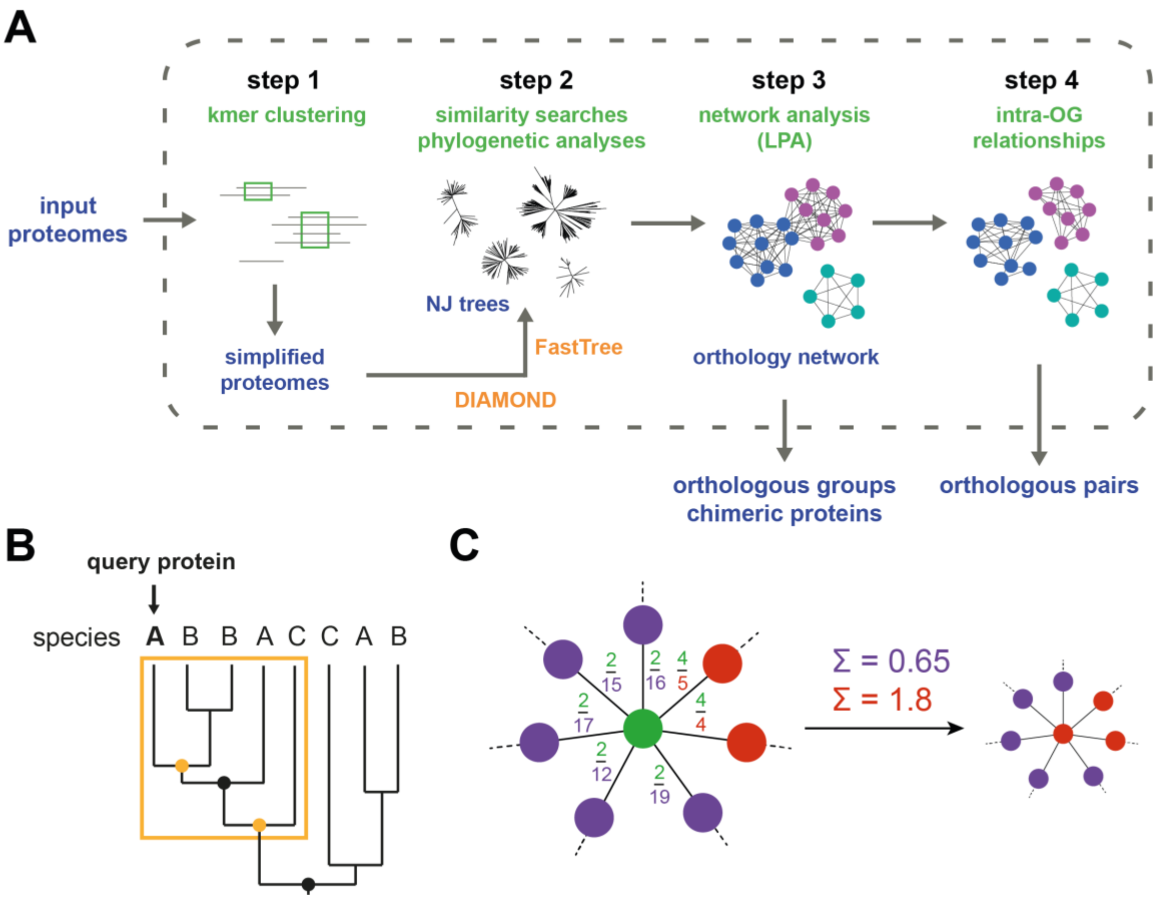


Figure 2

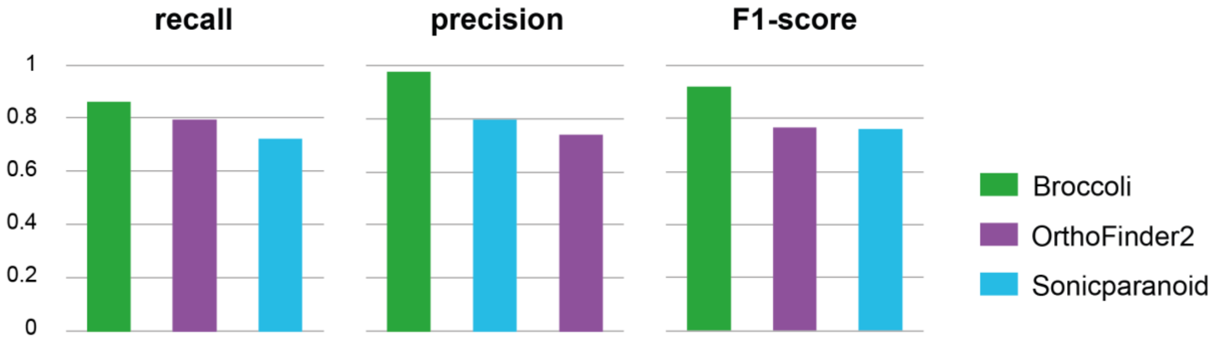


Figure 3

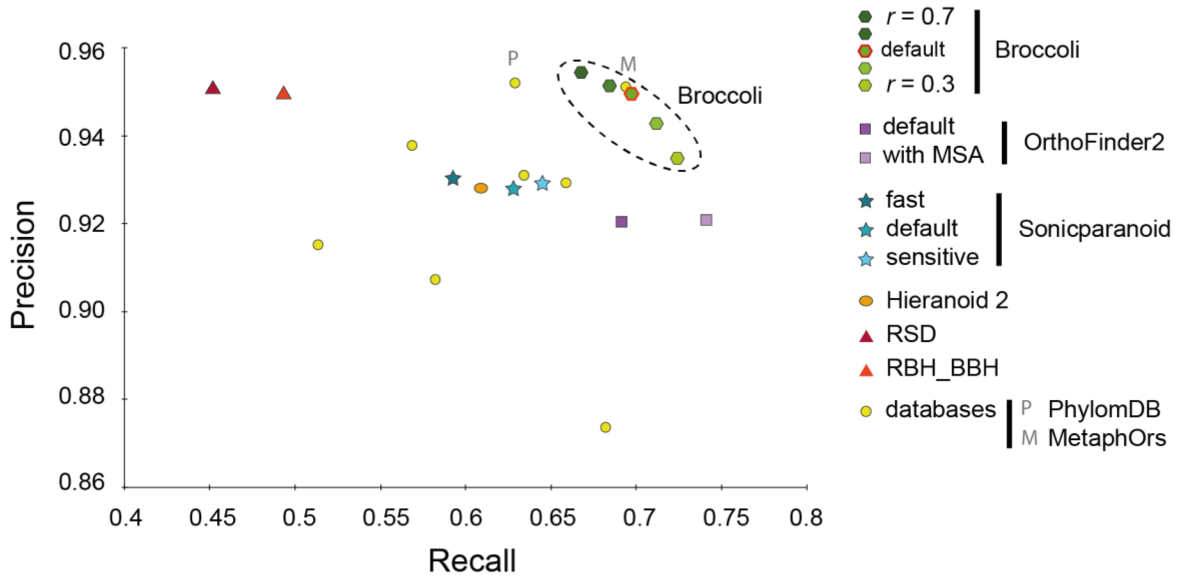
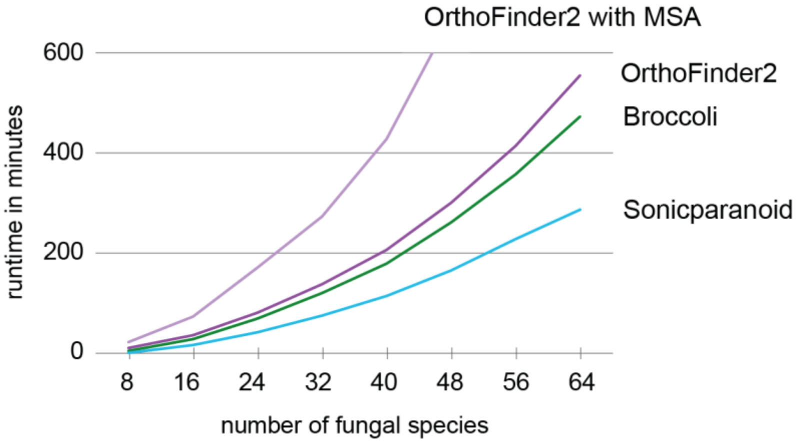


Figure 4


OrthoFinder2 with MSA
OrthoFinder2
Broccoli
Sonicparanoid
runtime in minutes
number of fungal species