Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment

Romain Derelle, Herve Philippe, John Colbourne

Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment

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

[1]School of Biosciences, University of Birmingham, Birmingham, United Kingdom
[2]Station d'Ecologie Théorique et Expérimentale, UMR CNRS 5321, Moulis, France
[3]Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, QC, Canada

***Corresponding author:** E-mail: r.derelle@bham.ac.uk.

Associate editor: Daniel Falush

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 artifacts. In this article, 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 ultrafast 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 data sets, 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.

Key words: **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 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 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 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 analyzed 1) using the best bidirectional hits (BBH) approach or one of its derivative to infer orthologous pairs as implemented in Hieranoid or OMA (Huynen and Bork 1998; Roth et al. 2008; Schreiber and Sonnhammer 2013; Sonnhammer and Ostlund 2015; Cosentino and Iwasaki 2019), or 2) using the Markov Cluster algorithm (MCL) to infer orthologous groups from the network of similarities (Dongen 2000; Li et al. 2003; Emms and Kelly 2015), orthologous groups that can further be analyzed using phylogenetic analyses and a species tree reconciliation approach to infer orthologous pairs (Emms and Kelly 2019). 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

Article

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 et al. 2007, 2014; Vilella et al. 2008). The promise of this "phylogeny-based" approach at improving orthology inferences has three important caveats: 1) the many hundreds of thousands of phylogenetic analyses required by this approach must be computationally efficient, 2) new methods for the delineation of orthologous groups must be proposed, and 3) 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 ultrafast 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 (LPA). 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 data set for orthologous groups, and the Quest of Orthologs 2018 benchmark data set for orthologous pairs (Altenhoff et al. 2016; Glover et al. 2019). In these tests, we compared Broccoli with recent distance-based pipelines combined with fast similarity search algorithms (e.g., DIAMOND, MMseq2; Buchfink et al. 2015; Steinegger and Soding 2017) since BlastP, which is two orders of magnitude slower, would not be usable for large data sets.

Materials and Methods

Broccoli is a pipeline written in Python 3 that requires the ete3 library (Huerta-Cepas 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 Preclustering

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, whereas others are set-aside and will be reinjected into the orthologous groups and orthologous pairs at the corresponding steps. This step aims at reducing the number of proteins to be analyzed 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 analyzed (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 et al. 2007) for each species by comparing its proteins against other proteomes and performing phylogenetic analyses in possible cases of gene duplications (i.e., cases of multiple hits for at least one species). 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 analyzed using FastTree2 (Price 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: 1) only sequence fragments matching the query sequence are used for phylogenetic analyses, whereas MSA, which operate on full-length sequences, usually include unaligned blocks that create phylogenetic noise, and 2) 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 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 is compared (an example is provided in fig. 1B). The two sets of leaves are considered part of the same orthologous group if 1) there is no common species between the two sets (i.e., no "overlap" species) or 2) 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

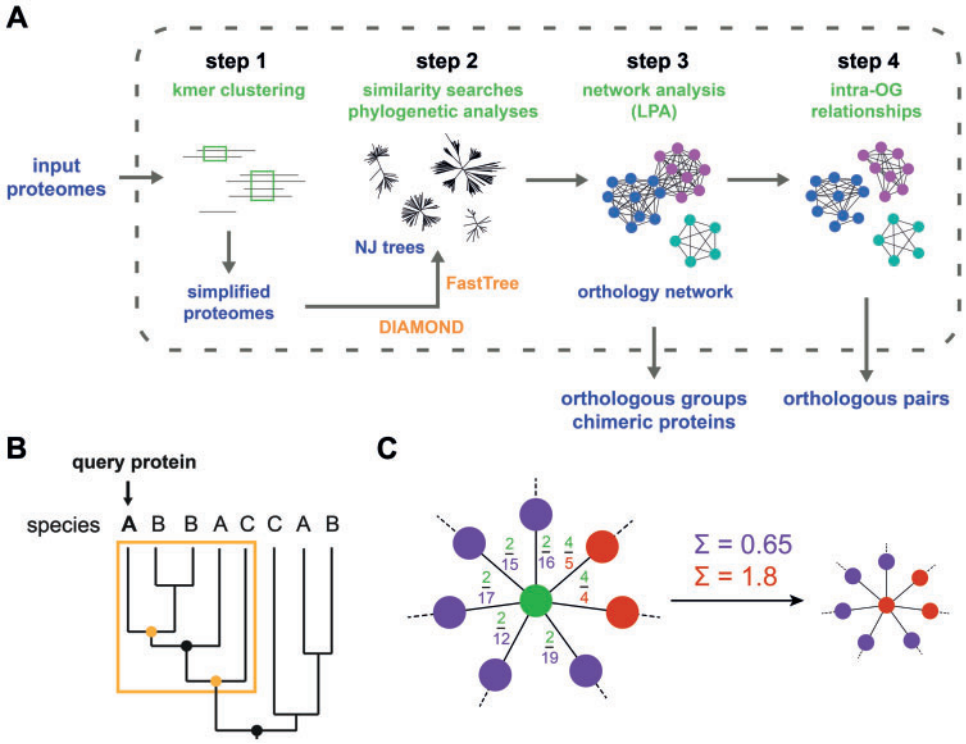

FIG. 1. Key aspects of Broccoli. (*A*) Overview of the pipeline. Data, external programs, and processes are colored in blue, orange, and green, respectively. (*B*) Example of the species overlap approach on a gene tree obtained from three species A, B, and C. The nodes fulfilling the species overlap 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 colors (green, purple, and red). The node with the green label is about to exchange its label with one of its neighbors. The fractions present on each edge represent the weights AB, where A is the green node and B its neighbor. The green node takes the red label since the sum of the "red weights" is higher than the sum of the "purple weights."

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 1) orthologs at least twice (since at the very least A has been compared with B and B compared with A) and 2) more often as orthologs than as paralogs. The edge weight w(AB) from A to B is defined as:

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

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 LPA (Raghavan et al. 2007) is applied to the orthology network to identify node communities (i.e., orthologous groups). The LPA used here, described in supplementary material 1, Supplementary Material online, 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 (supplementary material 1, Supplementary Material online) 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 LPA 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 supplementary material 1, Supplementary Material online. 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

Although 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: 1) proteins not belonging to the orthologous group are first removed from the "set-aside orthologous pairs" and from the rooted trees, and 2) 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:

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

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).

Performance Tests

The paraBench data set was built from an in-house collection of phylogenetic markers. The data set and the performance metrics are fully described in supplementary material 2, Supplementary Material online. As opposed to the benchmark of orthologous pairs, the performance metrics were calculated considering all possible pairs within each orthologous group. The data set, 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 (QfO) benchmark 2018 data set was downloaded from the EBI ftp server, then analyzed by Broccoli using the "not_same_sp" option and by varying the r threshold value using the "-ratio_ortho" option. The resulting sets of orthologous pairs were submitted to the OpenEBench website to run the QfO benchmark. Benchmarks of Broccoli under default parameters are available online at https://orthology.benchmarkservice.org, and other benchmark outputs are available to download in the Zenodo research data archive https://zenodo.org/record/3710751.

The versions of the pipelines and programs, and command lines, used in this study are indicated in supplementary material 3, Supplementary Material online. All benchmark outputs and QfO raw results are available to download in the Zenodo research data archive at https://zenodo.org/record/3710751.

Running Time Analyses

The running time analyses based on fungal data sets were performed using 4 CPUs of an Intel Xeon Gold 6248 processor and 40 GB of RAM memory. The fungal data sets correspond to those used in Emms and Kelly (2019), with one modification performed on the 64 species data set (see readme file in the Zenodo archive). All data sets are available to download in the Zenodo research data archive at https://zenodo.org/record/3710751.

The QfO 2018 runtimes were measured using 8 CPUs of the same processor and 60 GB of RAM memory.

Results

Benchmark of Orthologous Groups

The quality assessment of orthologous group predictions was performed using a custom-built benchmark data set (named "paraBench") comprising 17 eukaryotic species and 52 orthologous groups (see supplementary material 2, Supplementary Material online). In this benchmark, we compared Broccoli with two recent distance-based pipelines: OrthoFinder2 (Emms and Kelly 2019), 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 BBH approach. Broccoli produced the highest recall score value, closely followed by OrthoFinder2, thanks to its distance corrections (fig. 2 and supplementary material 3, Supplementary Material online). Finally, Sonicparanoid, which oversplit orthologous groups due to the stringency of the BBH approach, scored the lowest. On the precision side, Broccoli also performed better than the two other pipelines with a score of 0.973. OrthoFinder2 scored the lowest precision value indicating a tendency to overmerge 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 fig. 2).

Benchmark of Orthologous Pairs

We tested the orthologous pairs predicted by Broccoli, with its r threshold (Step 4) ranging from 0.3 to 0.7 in 0.1 increment (including the default value of 0.5), by using the large-scale Quest for Orthologs 2018 benchmark data set (referred thereafter as QfO 2018 data set) (Altenhoff et al. 2016; Forslund et al. 2018). This benchmark includes three groups of tests: 1) pairs of orthologs are compared with manually curated sets of gene phylogenies (using the TreeFam-A and SwissTree databases), 2) the function of orthologs are compared with each other using the Gene ontology and Enzyme classification, assuming that orthologous genes should have the same function in different species, and 3) the phylogenetic relationships of orthologous proteins are compared with reference species trees (the species Tree Discordance Benchmark, STDB, and its generalized version GSTDB), assuming that the gene trees should mirror the species trees. All benchmark values and

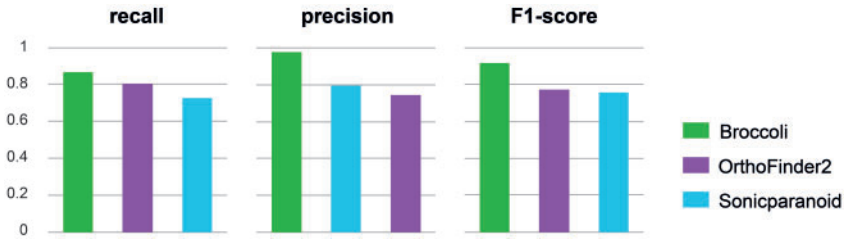

FIG. 2. Benchmark of orthologous groups (paraBench data set). Pipelines were ranked by their performance metric from left (highest value) to right (lowest value).

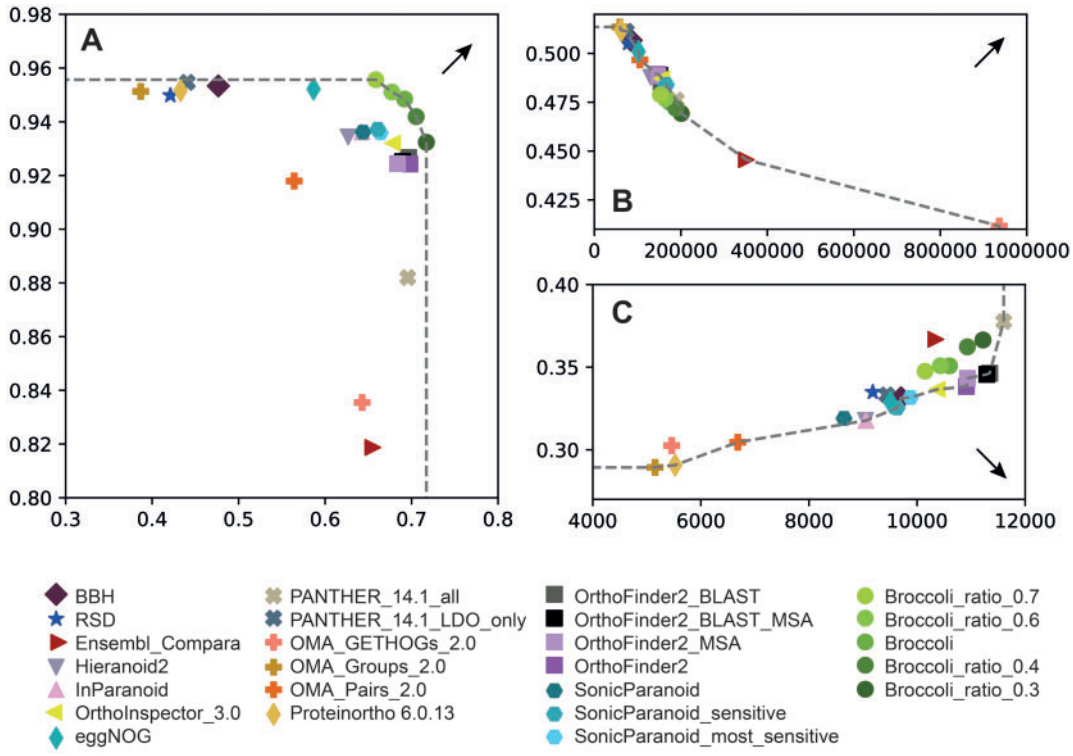

FIG. 3. Benchmark of orthologous pairs (Quest for Orthologs 2018 data set). The Pareto frontiers are represented by dotted lines and arrows indicate the direction toward better performances. In each of these three scatter plots, the *X* axis and *Y* axis correspond to a measure of recall and precision, respectively. (*A*) TreeFam-A benchmark (*X* axis: true positive rate; *Y* axis: positive predictive value). (*B*) Gene-ontology benchmark (*X* axis: number of orthologs; *Y* axis: average Schlicker similarity). (*C*) STD Fungi benchmark (*X* axis: number of completed tree samplings; *Y* axis: average Robinson–Foulds distance).

scatter plots are available in supplementary materials 3 and 4, Supplementary Material online, respectively.

Orthologous pairs produced by Broccoli showed a strong agreement with the reference gene-phylogenies databases. When compared with the TreeFam-A database, only results produced by Broccoli were located on the Pareto frontier (fig. 3A), which is defines by the set of methods that are not outperformed by any other method in both recall (*X* axis) and precision (*Y* axis) (Altenhoff et al. 2016). As expected, results corresponding to $r = 0.7$ (i.e., stringent threshold) scored the highest precision of all methods and results corresponding to $r = 0.3$ (i.e., relaxed threshold) were the most sensitive of all methods. In the case of the SwissTree database, a small reference data set compared with the TreeFam-A database, all results were located on or closed to the Pareto frontier (supplementary material 4, Supplementary Material online).

In the two functional conservation tests, Broccoli results were similar to those of other methods (excluding two outlier methods, Ensembl Compara and OMA GEtHoG), with one (Gene ontology; fig. 3B) or two (Enzyme classification; supplementary material 4, Supplementary Material online) Broccoli results on the Pareto frontier. As observed in the TreeFam-A benchmark, results corresponding to low *r* thresholds achieved higher precision (measured as the average Schlicker semantic similarity of functional annotations associated with orthologs), whereas results corresponding to high *r* thresholds showed higher sensitivity.

In contrast, the Broccoli performance in the species-trees discordance benchmarks are slightly worse than the ones of

other methods: in most of these benchmarks, Broccoli results were distant from the Pareto frontier, with high sensitivity but high Robinson–Foulds distances (i.e., low precision) compared with other methods (e.g., STD Fungi in fig. 3C). The only exceptions were the STD Bacteria benchmark in which all Broccoli results were found on the Pareto frontier, and the GSTD Eukaryota benchmark in which two Broccoli results were found on the Pareto frontier (supplementary material 4, Supplementary Material online).

Chimeric Proteins

In the absence of a specific gene-fusion benchmark, it is difficult to assess the quality of the predictions made by Broccoli. Nevertheless, a total of 1,675 proteins were predicted to be the result of gene-fusion events from the QfO 2018 data set (representing ~0.2% of all proteins; list available in supplementary material 3, Supplementary Material online and in the Zenodo research data archive). The number of chimeric proteins per species was highly heterogeneous, ranging from 0 (*Thermodesulfovibrio yellowstonii* and *Giardia intestinalis*) to 223 (*Zea mays*; supplementary material 3, Supplementary Material online). Four species of this data set showed particularly high numbers of chimeric proteins (namely *Z. mays*, *Phytophthora ramorum*, *Branchiostoma floridae*, and *Monosiga brevicolis*). We hypothesize that these high prevalences are the consequences of errors in the gene prediction of these genomes. Broccoli was able to identify chimeric proteins that resulted from the combination of up to six orthologous groups and genes-fusions events shared by up to 16 proteins. The latter case corresponds to the pentafunctional arom proteins present in multiple eukaryotic lineages (Richards et al. 2006). However, Broccoli failed to recover some well-known gene fusion events such as the fusion of the dihydrofolate reductase and thymidylate synthase proteins present in "unikonts" and absent in most "bikonts" (Cavalier-Smith 2003), and the fusion of two tRNA synthetases shared by most metazoan species (Ray et al. 2011).

Finally, we carefully compared the Pfam domain contents of the seven *Caenorhabditis elegans* chimeric proteins identified by Broccoli in the QfO 2018 data set to those of their respective orthologous groups. These analyses showed that six of these chimeric proteins exhibit unique Pfam domain combinations, being the results of complete or partial fusions of proteins (supplementary material 5, Supplementary Material online).

Running Time Analyses

Considering the large number of phylogenetic analyses performed by Broccoli (e.g., 658,421 phylogenies for the QfO 2018 data set), it is expected to be several orders of magnitude slower than distance-based pipelines. We compared the runtimes of Broccoli to those of the two fastest distance-based algorithms Sonicparanoid and OrthoFinder2 (Emms and Kelly 2019) using data sets composed of 4 to 64 fungal proteomes. In these tests, the runtimes of Broccoli were found between those of Sonicparanoid and OrthoFinder2 (fig. 4). Regarding the two extremes of the speed spectrum, OrthoFinder2 with the MSA option was by far the slowest

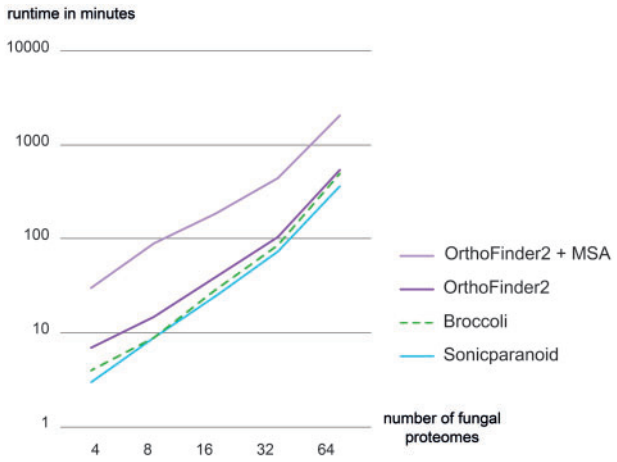

FIG. 4. Efficiency tests. Runtimes, rounded to the nearest minute, were obtained using four CPUs and a data set composed of 4–64 fungal species. Pipelines were run using default parameters unless specified otherwise.

pipeline, and Sonicparanoid, which only performs half of similarity searches (Cosentino and Iwasaki 2019), was found to be the fastest for every data set. The same speed rank was observed when analyzing the QfO 2018 data set, which contains 78 species, using 8 CPUs: Sonicparanoid (522 min), Broccoli (634 min), and OrthoFinder2 (850 min; we did not test the MSA option using this data set).

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 ultrafast 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 was found highly precise and sensitive on all tested benchmarks, with the noticeable exception of most QfO species tree benchmarks. Although these specific benchmarks could possibly point to some limitations of the inferences made by Broccoli, it should be noticed that disagreements between gene trees and species trees are extremely common (Marcet-Houben and Gabaldon 2009), with many well-known sources of these discrepancies (e.g., incomplete lineage sorting, lateral gene transfers, gene prediction errors; see Introduction). Indeed, the high frequency of these discrepancies is the main reason as to why Broccoli employs a species overlap and not a species tree reconciliation approach for orthology delineation. Therefore, we believe that the use of distances between gene and species trees as a surrogate of precision measurement is questionable.

Finally, Broccoli showed high efficiency, with runtimes similar to those of the fastest distance-based pipelines, thanks to the parallelization of most tasks, an initial kmer clustering to simply proteomes, ultrafast phylogenetic analyses, and an efficient network analysis.

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 et al. 2019). This fast and precise identification of chimeric proteins alongside their corresponding orthologous groups represents a promising avenue that should facilitate evolutionary studies of gene-fusion events (see also Pathmanathan et al. 2018). However, future work is still required to widen the search of chimeric protein in the orthology network as the set of chimeric proteins currently identified by Broccoli appears to be incomplete.

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 optimized (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.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank Franz Lang and Luisa Orsini for their valuable comments on an earlier version of the article. The authors also wish to thank Benjamin Buchfink for implementing new output fields in Diamond, and Toni Gabaldón for his seminal work on the phylogeny-based approach. Broccoli was developed as part of the DeepEuk collaborative project. R.D. was supported by a NERC Highlight Topic Grant (NE/N006216/1). J.K.C. was supported by UK NERC award Cracking the Code of Adaptive Evolution (NE/N016777/1).

References

Altenhoff AM, Boeckmann B, Capella-Gutierrez S, Dalquen DA, DeLuca T, Forslund K, Huerta-Cepas J, Linard B, Pereira C, Pryszcz LP, et al. 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods*. 13(5):425–430.

Buchfink B, Xie C, Huson DH. 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. *Eur J Protistol*. 39(4):338–348.

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

Dalquen DA, Dessimoz C. 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, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol Evol*. 22(1):34–41.

Dongen SV. 2000. Graph clustering by flow simulation [Ph.D thesis]. The Netherlands: University of Utrecht.

Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, Wyckoff GJ, Malcom CM, Lahn BT. 2004. Accelerated evolution of nervous system genes in the origin of *Homo sapiens*. *Cell* 119(7):1027–1040.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res*. 47(D1):D427–D432.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*. 16(1):157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 20(1):238.

Forslund K, Pereira C, Capella-Gutierrez S, Sousa da Silva A, Altenhoff A, Huerta-Cepas J, Muffato M, Patricio M, Vandepoele K, Ebersberger I, et al. 2018. Gearing up to handle the mosaic nature of life in the quest for orthologs. *Bioinformatics* 34(2):323–329.

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

Glover N, Dessimoz C, Ebersberger I, Forslund SK, Gabaldon T, Huerta-Cepas J, Martin MJ, Muffato M, Patricio M, Pereira C, et al. 2019. Advances and applications in the quest for orthologs. *Mol Biol Evol*. 36(10):2157–2164.

Harpak A, Lan X, Gao Z, Pritchard JK. 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, Capella-Gutierrez S, Pryszcz LP, Marcet-Houben M, Gabaldon T. 2014. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res*. 42(D1):D897–D902.

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

Huerta-Cepas J, Dopazo J, Gabaldon T. 2010. ETE: a python environment for tree exploration. *BMC Bioinformatics* 11(1):24.

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

Kawahara Y, Imanishi T. 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(1):9.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc R Soc B*. 279(1749):5048–5057.

Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*. 39(1):309–338.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13(9):2178–2189.

Maddison WP. 1997. Gene trees in species trees. *Syst Biol*. 46(3):523–536.

Marcet-Houben M, Gabaldon T. 2009. The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One* 4(2):e4357.

Pathmanathan JS, Lopez P, Lapointe FJ, Bapteste E. 2018. CompositeSearch: a generalized network approach for composite gene families detection. *Mol Biol Evol*. 35(1):252–255.

Pich IRO, Kondrashov FA. 2014. Long-term asymmetrical acceleration of protein evolution after gene duplication. *Genome Biol Evol*. 6(8):1949–1955.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.

Raghavan UN, Albert R, Kumara S. 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 PS, Sullivan JC, Jia J, Francis J, Finnerty JR, Fox PL. 2011. Evolution of function of a fused metazoan tRNA synthetase. *Mol Biol Evol*. 28(1):437–447.

Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW. 2006. Evolutionary origins of the eukaryotic Shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryot Cell*. 5(9):1517–1531.

Roth AC, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9(1):518.

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

Sonnhammer EL, Ostlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res*. 43(Database issue):D234–D239.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 16(8):472–482.

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

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2008. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19(2):327–335.

Zmasek CM, Godzik A. 2012. This Deja vu feeling–analysis of multi-domain protein evolution in eukaryotic genomes. *PLoS Comput Biol*. 8(11):e1002701.