*Genome Analysis*

# PorthoMCL: Parallel orthology prediction using MCL for the realm of massive genome availability

Ehsan S. Tabari[1,] and Zhengchang Su[1*]

[1]Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**ABSTRACT**

**Motivation:** Finding orthologous genes among multiple sequenced genomes is a primary step in comparative genomic studies. With the availability of exponentially increasing number of sequenced genomes, comparative genomics becomes more powerful than ever for genomic analysis. However, the very large number of genomes needing to be analyzed makes conventional orthology prediction methods incapable for the tasks. Thus an ultrafast tool is urgently needed.

**Results:** Here, we present PorthoMCL, an improved version of OrthoMCL with parallelization, for finding orthologous genes among a very large number of genomes. We have demonstrated its capability by identifying orthologs in 2,758 prokaryotic genomes, the results are available for downloading at: http://bioinfo.uncc.edu/ehsan.tabari/porthomcl/.

**Availability:** PorthoMCL (source code, executables, sample datasets and documentation) is available under the MIT license in the github repository: github.com/etabari/PorthoMCL.

**Contact:** zcsu@uncc.edu.

## 1 INTRODUCTION

The rapid advance in sequencing technologies has made sequencing a prokaryotic genome at an unprecedented fast speed and low cost. As a result, thousands of prokaryotic genomes have been fully sequenced, and this number can soon reach tens of thousands. The availability of large number of completed genomes renders comparative genomics ever a powerful approach for gene annotations and addressing many important theoretical and application problems. However, the rate at which genomes are sequenced outpaces that at which CPU speed increases. This poses a great challenge in comparative analyses of the very large number of genomes, soliciting new faster algorithms or adapting existing tools in parallel environments.

Orthologs are genes in different species that are derived from a single gene in their last common ancestor by speciation events. Orthology indicates the conservation in both sequence and function between genes in different genomes. Identification of orthologous genes among a group of genomes is crucial to almost any comparative genomic analysis (Alexeyenko *et al.*, 2006). In contrast, paralogs are genes that are resulted from gene duplication within a species, thus may have different functions though their sequences can be conserved. Depending on the duplication occurring before or after speciation, they are called outparalogs or inparalogs, respectively (Sonnhammer *et al.*, 2002). A major challenge in orthologs predictions is to differentiate true orthologs of a gene from the orthologs of its paralogs.

OrthoMCL is one of the most widely used algorithms for predicting orthologous genes across multiple genomes. Similar to many other orthology prediction algorithms, it is based on reciprocal best hits in all-against-all BLAST searches of complete proteomes of the genomes. OrthoMCL represents the similarity among the sequences using a weighted graph, where the nodes are the genes, and two nodes/genes are connected by an edge, if there is a pair of reciprocal best hits between them with a similarity greater than a cutoff. If there is a pair of reciprocal best hits between genes $x_A$ and $x_B$ in genomes $A$ and $B$, respectively, the weight of the edge that connects them is a normalized score ($\overline{w}$) based on the E-values of the reciprocal hits, and is defined as,

$$w(x_A, y_B) = -\frac{\log_{10} \textbf{Evalue}(x_A \rightarrow y_B) + \log_{10} \textbf{Evalue}(y_B \rightarrow x_A)}{2} \qquad (1)$$

$$\overline{w}(x_A, y_B) = \frac{w(x_A, y_B)}{\underset{\forall \alpha,\beta}{\text{average}}(w(\alpha_A, \beta_B))} \qquad (2)$$

Similarly, within-species reciprocal hits that have a better normalized score than between-species hits are identified as paralogs (Li *et al.*, 2003). Ortholog and paralog groups are then identified by finding the heavily connected subgraphs using the Markov Clustering algorithm (Van Dongen, 2000). However, OrthoMCL relies on a relational database system to store the BLAST results and issues SQL commands to find reciprocal best hits and score them, making it costly and inefficient when the number of genomes becomes large. To overcome this problem and to further speed up the method, we developed a parallel orthology prediction tool using MCL, PorthoMCL. In addition to the parallelization, our more efficient sparse file structure makes PorthoMCL ultrafast and highly scalable. Furthermore, PorthoMCL is platform independent, thus can be run on a wide range of high performance computing clusters and cloud computing platforms.

## 2 PORTHOMCL

[*]To whom correspondence should be addressed.

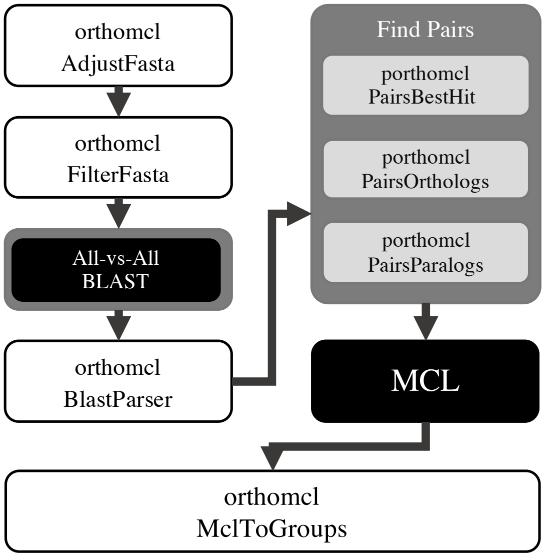

**Fig. 1.** Workflow of PorthoMCL. Original OrthoMCL steps are shown in white, and PorthoMCL steps are in grey shades. Black boxes are the external applications that PorthoMCL requires.

### 2.1 Workflow

The workflow of PorthoMCL is similar to that of OrthoMCL (Figure 1). However, instead of depending on an external database server, PorthoMCL uses a sparse file structure for more efficient data storage and retrieval. In addition, we parallelized all the computationally intensive steps of OrthoMCL. First, PorthoMCL performs all-against-all BLAST searches in parallel by performing individual-against-all BLAST searches for every genome independently. Second, it identifies the best between-species BLAST hits for each two genomes by scanning the individual-against-all BLAST results in parallel. The hit $x_A \rightarrow y_B$ is identified as the best hit if its E-value is the best E-value for all the searches of $x_A$ in genome $B$ and meets an E-value/match-percentage threshold. This step results in a single best hit file per each genome, and a self-hit file for paralogy-finding. Third, it finds reciprocal best hits between every two genomes and calculates the normalized score using Formula 2. This is the most computationally intensive step in the algorithm, so we used a sparse file for storage and parallel processing. Specifically, for each parallel process, PorthoMCL loads at most two best hit files at the same time to reduce the memory footprint. Also, every best hit file is only loaded once to lower the I/O costs. Finally, PorthoMCL finds within-species reciprocal best hits and normalizes the score with the average score of all the paralog pairs that have an ortholog.

These steps are embarrassingly parallel problems and are readily designed to be executed in parallel on a variety of high performance computing (HPC) environments. They are clearly scalable and can exploit the capacity available to the HPC. However, these steps are not totally independent as each step requires the output of the preceding step. The output of these steps are eventually collated to construct a sequence similarity graph that is then cut by the MCL program to predict orthologous and paralogous gene groups.

### 2.2 High performance computing support

PorthoMCL is designed to predict orthologs in ever an increasingly large number of sequenced genomes in a HPC environments such as computing clusters or cloud computing platforms. We have included a TORQUE script with the package to facilitate its use in such environments. However, PorthoMCL also runs on a desktop or a server without the need for a database server which is advantageous over OrthoMCL.

## 3 RESULTS

To illustrate the power of PorthoMCL, we have applied it to all the 2,758 sequenced bacterial genomes in GenBank (downloaded: April 2015) using their annotated protein sequences. These genomes contain a total of 8,661,583 protein sequences with a median length of 270 amino acids. They serve as both the query and the database for all-against-all BLAST searches. We split the query into smaller files each containing about10,000 sequences, we ran BLAST searches (e-value cutoff: 1e-5; database size: 1e8) in parallel using PorthoMCL. The combined output of the BLAST contained 2,957,375,578 hits. The total runtime of the BLAST searches were 11 days on a cluster with 60 computing nodes (each nodes has 12 cores and 36GBs of RAM), which would need 549 days if run on a single node. PorthoMCL identified 763,506,331 ortholog gene pairs and identified 230,815 ortholog groups. PorthoMCL finished this step in only 7 days (same computing cluster, total runtime 1,634 days), while OrthoMCL could not finish this after 35 days of running on a database server with 40 cores and 1TBs of RAM.

The orthologous pairs (file size: 6.2GB) and orthologous groups (file size: 50MB) as well as paralogous pairs are available for download at http://bioinfo.uncc.edu/ehsan.tabari/porthomcl/.

In addition to PorthoMCL, we have also provided a sample dataset for convenience. The options and arguments required at each step are discussed in detail in the documentation that accompanies with PorthoMCL.

## ACKNOWLEDGEMENTS

We thank Jonathan Halter for his support and valuable contributions to this project.

*Funding*: This work was funded by the National Science Foundation (EF0849615 and CCF1048261) and NIH (R01GM106013).

*Conflict of Interest*: none declared.

## REFERENCES

Alexeyenko, A., Lindberg, J., Pérez-Bercoff, A., & Sonnhammer, E. L. L. (2006). Overview and comparison of ortholog databases. *Drug Discovery Today. Technologies*, 3(2), 137–43. doi:10.1016/j.ddtec.2006.06.002

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–89. doi:10.1101/gr.1224503

Sonnhammer, E. L. ., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, 18(12), 619–620. doi:10.1016/S0168-9525(02)02793-2

Van Dongen, S. (2000). Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, Netherlands.