SonicParanoid: extremely fast, accurate, and easy orthology inference

Salvatore Cosentino and Wataru Iwasaki

Graduate School of Science, The University of Tokyo

Abstract

Accurate inference of orthologous genes encoded on multiple genomes plays a key role in many fields of present genomics. For example, elucidation of species-specific genomic signatures at the sequence, structural, and/or functional levels (comparative genomics), and the reconstruction of phylogenetic trees using genomic information (phylogenomics), depend on reliable orthology inference.

Dozens of tools are available for orthologous inference, which can be grouped into two classes: graph-based (e.g., InParanoid[1] and its extension MultiParanoid[2], for the inference of multi-species ortholog groups, OrthoMCL, and Proteinortho[3]) and tree-based methods (e.g., PANTHER and Ensembl Compara). Graph-based methods accept sequence-similarity graphs of proteins as input, use best reciprocal hit-based approaches to find orthologs between species, and adopt various strategies to identify paralogs within species. Tree-based methods require phylogenetic trees as input, and even thought generally more accurate than graph-based ones, they are much more computionally expensive. Despite the importance of orthology inference most prediction tools are difficult to deploy and use for users with limited computer skills, and the lengthy computation they require are a big limitation, especially if we consider that present studies involve dozens of genomes.

We present SonicParanoid, a fast and easy to use orthology prediction tool which is 750X faster than InParanoid (on which it is based on), and has an accuracy comparable to that of the well-established orthology inference tools.

Implementation and testing

Features and implementation

Features

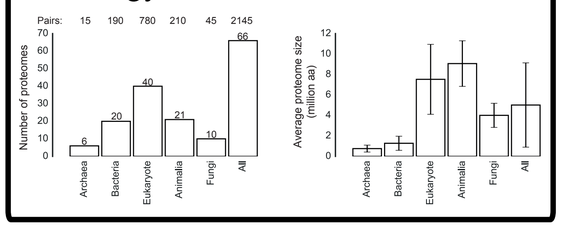
- **Fast**: <10 hours to analyze the 66 benchmark proteomes in *fast* mode (8 CPUs)
- **Easy to employ**: virtually no installation required
- **Hardware requirements**: a 64-bit multi-core CPU and 8 Gygabytes of memory
- **Update mode**: possibility to mantain a database of orthologous genes that can be updated by adding or removing species
- **Persistent:** execution can be halted and restarded without losing results
- **Multi-species clustering on subsets of species:** perfomed in few seconds

Implementation

SonicParanoid was implemented in the Python programming language; it uses MMseqs2[5] to perform sequence alignments and a C++ version of MultiParanoid.

Dataset

SonicParanoid was tested on a proteome dataset from the Quest for Orthologs (QfO) consortium[4], commonly used for benchmarking orthology inference methods.



The fastest method available

The execution time of SonicParanoid, InParanoid and Proteinortho on the QoF dataset, and its subsets were compared. SonicParanoid was >160X faster (almost 750X in the *fast* mode) than InParanoid in inferring orthologs for the 2145 proteome combinations for the 66 species in the test dataset. When compared to Proteinortho, which is to our knowledge the fastest prediction tool available, SonicParanoid was up to 6.8X faster (28.4X in the *fast* mode) and predicted a higher amount ortholog pairs.

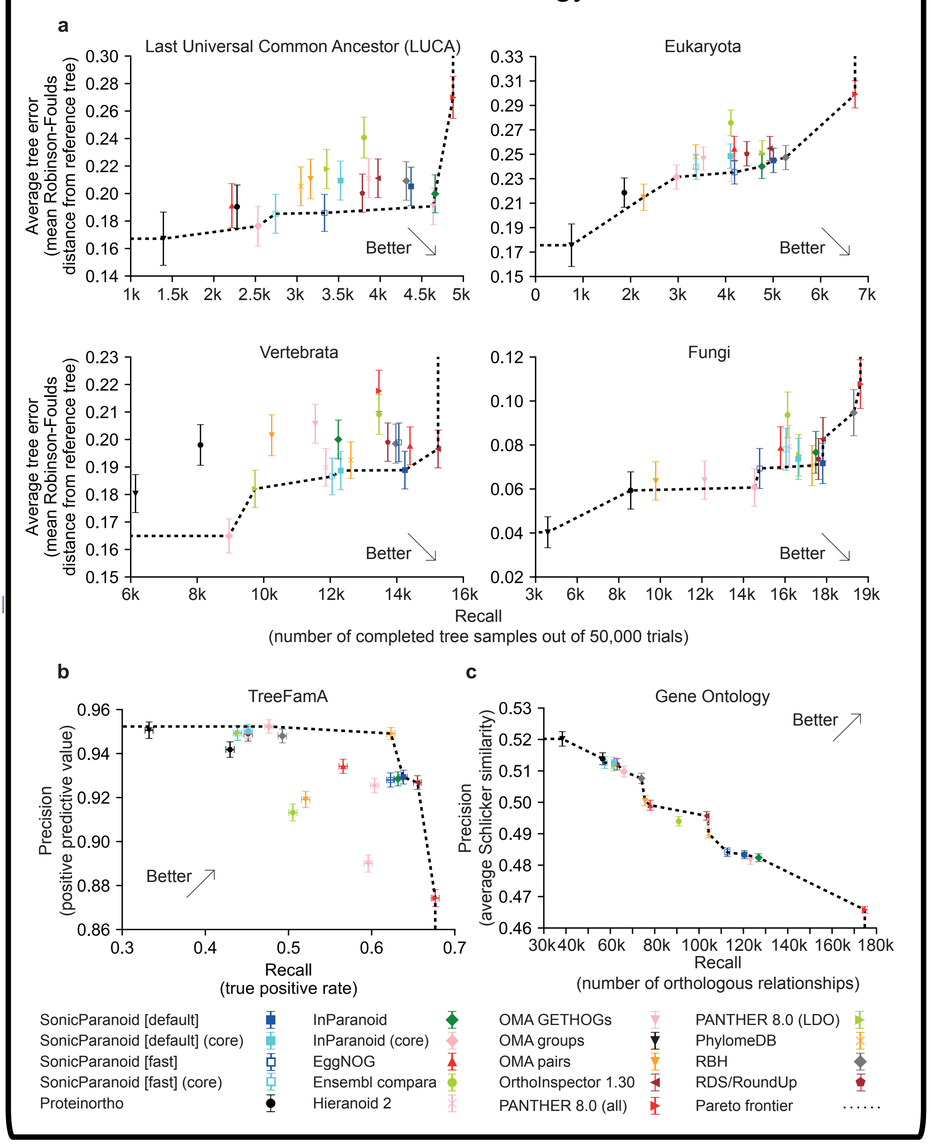
Test data			Execution time in hours		Speedup VS InParanoid		Speedup VS Proteinortho	
Dataset	Number of proteomes	Proteome pairs	Fast mode	Default mode	Fast mode	Default mode	Fast mode	Default mode
All	66	2145	9.72h	45.28h	747.8X	160.5X	21.9X	4.7X
Eukaryote	40	780	6.90h	28.94h	691.3X	164.8X	28.4X	6.8X
Archaea	6	15	0.02h	0.04h	83.8X	46.8X	1.8X	1.0X
Bacteria	20	190	0.28h	0.81h	139.8X	48.4X	3.1X	1.1X
Animalia	21	210	3.12h	15.27h	521.9X	106.7X	23.5X	4.8X
Fungi	10	45	0.15h	1.26h	345.1X	39.8X	18.6X	2.1X

Hardware used: The tests were performed on a RedHat Linux 4.8.5 server, using 8 Intel Xeon CPUs at 2.6 GHz, and 8 GBytes of memory.

Results

Accuracy benchmark

The predicted ortholog pairs were uploaded to a benchmark web-service[6] in order to assess the accuracy of SonicParanoid and compare it to that of other 14 orthology inference tools. The benchmark consists of tests which evaluate the predictions based on species tree discordance (a), agreement with reference trees (b), and functional conservation (c). Overall SonicParanoid showed a good balance between precision and recall, regardless of the mode in which it was executed. Moreover, in most of the tests it is pushing the pareto frontier proving its competitiveness in relation to the other 14 tested orthology inference tools.



Multi-species ortholog groups prediction

Input orthologous relationships	Groups (total)	Groups with tree conflicts	Groups with tree conflicts (%)	Groups with 66 species
SonicParanoid (all) [default]	16540	4297	26.0	42
SonicParanoid (core) [default]	24452	5913	24.2	51
SonicParanoid (all) [fast]	18004	4670	25.9	37
SonicParanoid (core) [fast]	26388	6249	23.7	50
InParanoid (all)	18468	4807	26.0	64
InParanoid (core)	24976	6040	24.2	57
Proteinortho	59086	NA	NA	0

* Core pairs are predicted ortholog pairs with the maximum confidence score (1.0). Such confidence scores are not present in predictions from Proteinortho.

SonicParanoid predicted an amout of ortholog relationships and multi-species ortholog groups similar to that of InParanoid for the complete QfO dataset. Multi-species ortholog prediction in SonicParanoid is perfomed using a C++ implementation of MultiParanoid. Proteinortho estimated a much larger number of ortholog groups, which were apparently very fragmented–the average number of species contained in each group was only approximately 6 species, and there was no group containing proteins from all 66 proteomes.

Conclusions

In this work we presented SoniParanoid, an orthology inference tool which improves on InParanoid, one of the most used ortholog prediction methods. SonicParanoid was >160X and about 750X faster than InParanoid, when executed in the *default* and *fast* mode, respectively. It processed a set of 10 fungal proteomes in <10', and took <7 hours to process a set of 40 eukaryotic proteomes, using only 8 CPUs in the *fast* mode. Aside from the reduced execution time, SonicParanoid showed an accuracy comparable to that of the well-established orthology inference tool. Lastly, SonicParanoid is easy to install and use, works on multiple platforms, and allows users to maintain a collection of orthologous genes with ease.

Considering the pace at which new genomes are made available every month, with genomic studies involving a growing number of species, SonicParanoid could be the tool to achieve fast, accurate, and easy orthology inference.

References

1) Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res. 43, D234–9 (2015)

2) Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22, e9–15 (2006)

3) Lechner et al. Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis. BMC Bioinformatics, 12(1):124, (2011)

4) Gabaldón, T. et al. Joining forces in the quest for orthologs. Genome Biol. 10, 403 (2009)

5) Steinegger et al. Sensitive protein sequence searching for the analysis of massive data sets. bioRxiv, https://doi.org/10.1101/079681, (2016)

6) Altenhoff et al. Standardized benchmarking in the quest for orthologs. Nat. Methods 13, 425–430 (2016)

Website

iwasaki.bs.s.u-tokyo.ac.jp/sonicparanoid

Contacts

salvocos@bs.s.u-tokyo.ac.jp