LOOP : Linkage Optimized Ortholog Predictions

Noah Lee, Chaitanya Srinivasan

Carnegie Mellon University, Department of Computational Biology, School of Computer Science, Pittsburgh, PA

Background

Orthologous genes provide quantitative approaches to assess relatedness and inform research into gene groups and evolutionary history. However, technological and logistical constraints prevent the full existing suite of prediction tools from benefiting every research group. Certain methods rely on assumptions which are not applicable to every genome assembly, such as confidence in the global alignment to induce synteny or experimental data for protein-protein interaction networks. We seek to provide a method which can be applied broadly to more sparsely populated investigations at a low computational cost. .

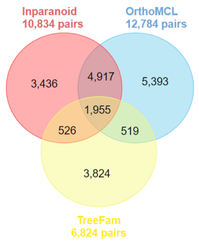


Fig. 1. Overlap of results from prediction tools is often varied

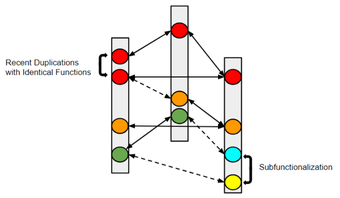
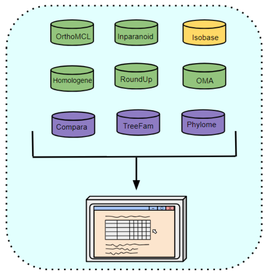
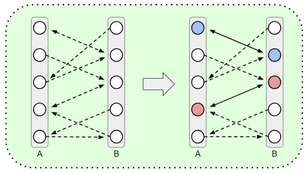


Fig. 2. Orthologs predictions are confounded by evolutionary events such as duplication and subfunctionalization

Tools





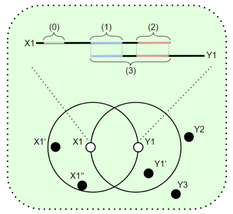
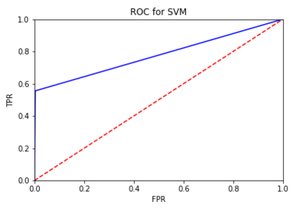


Fig. 3. Clockwise from blue: the general structure for DIOPT (DRSC Integrative Ortholog Prediction Tool), the method of RBH (Reciprocal Best Hit), and the algorithm behind InParanoid

Initial Investigations



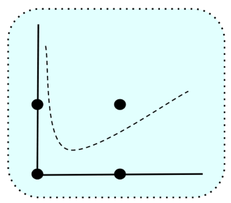
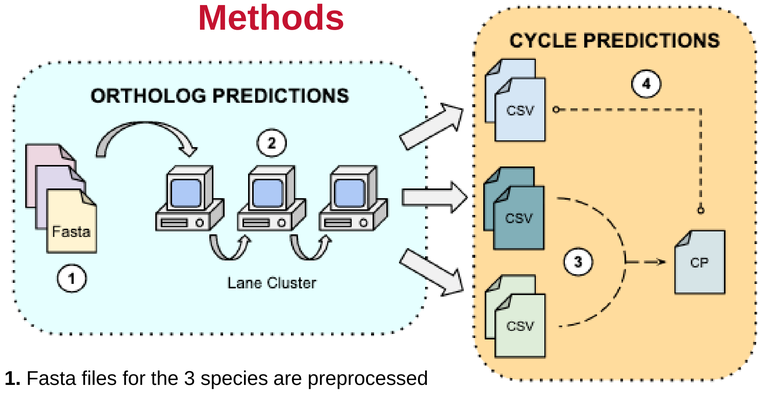


Fig. 4. A purely statistical approach to ortholog predictions in which a sparse feature space of available predictions is tackled with SVM

Methods

Fig. 5. Traditional ortholog prediction tools are used to generate two pairwise lists for *S. purpuratus* to *D. melanogaster*, and from *D. mel* to *H. sapiens*. These two lists are connected to form the "cycle prediction" method to generate *S. purp* to *H. sap* predictions, which are then verified by RBH and InParanoid



1. Fasta files for the 3 species are preprocessed
2. Data from DIOPT, RBH, and InParanoid is generated (17.5 hours)
3. Cycle predictions link SPU<>DMEL to DMEL<>HSAP predictions (0.5 hours)
4. Predictions are verified using RBH and InParanoid (11 hours)

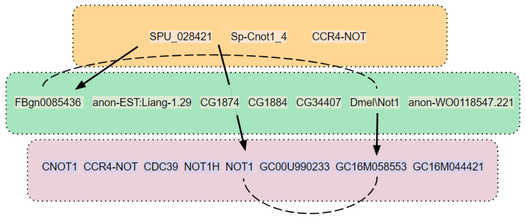
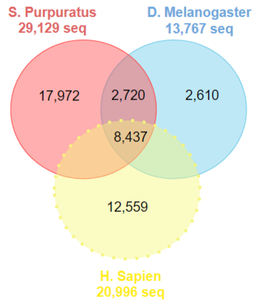


Fig. 6. Instances of gene IDs, gene names, and gene synonyms as appearing in fasta files and databases. Connecting the cycles from prediction tools requires identifying equivalent genes using large "synonym" files. Using standard gene IDs would eliminate this pain point, but belies the iterative nature of genome annotation and curation

Results



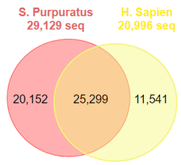


Fig. 8. Common predictions between SPU<>DMEL and from DMEL<>HSAP are used to generate the cycle predictions. The overlap is contrasted with the unique pairs of a full SPU<>HSAP analysis in the second Venn diagram

Amongst the 8,437 cycle predictions between *S. purp* and *H. sap*, there is a 43% overlap with predictions generated by two traditional ortholog prediction methods, InParanoid and RBH.

Verified	Count	Percent
Yes	3,631	43%
No	4,806	57%

Method	Overlap w/ Cycle	Total Predictions	Percent
IP	1,929	17,968	10.7%
RBH	592	4,004	14.7%
BOTH	1,110	3,327	33.33%

Fig. 7. Breakdown for the cycle predictions is shown on the left table, with further detail shown through identifying via which method the cycle predictions were verified on the right table

Species	S. purpuratus	D. melanogaster	H. sapiens
Gene Name	SPU_001817	PyK	PKLR
Function	Glycolytic process, magnesium ion binding, pyruvate kinase activity	Catalytic activity in glycolysis, pyruvate kinase	Glycolysis, pyruvate kinase

Fig. 9. A triplet ortholog group produced by cycle prediction shows expected similarity through manual annotation

Discussion

Current gene prediction tools leverage a balance of both statistical and biological assumptions, but by using a single strong biological assumption for a transitive property in orthologs, the cycle prediction method was able to generate a comparably sized list on par with first iterations of many tools.

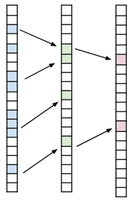


Fig. 10. A drop in coverage may be due to dimensionality reduction, which also impairs scalability. As predictions pass from one species to another, the number of considered genes is greatly reduced.

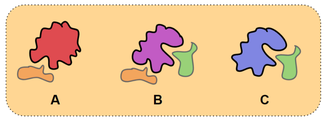


Fig. 11. The assumption of transitivity is complicated by the presence of in-paralogs, out-paralogs, and functional orthologs

Future Steps

A reverse cycle approach, in which *H. sap* predictions are traced toward *S. purp* predictions, could reduce the loss in dimensionality and improve scalability. Additionally, filtering methods could improve accuracy by removing less confident ortholog predictions and trending toward the increase in verification already seen in initial results.

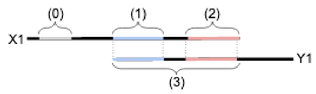


Fig. 12. A potential filtering method, similar to InParanoid, could filter results based on adjusted coverage scores

References

Erik L.L.Sonnhammer and Gabriel Östlund. "InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic". Nucleic Acids Res. 43:D234-D239 (2015)

Fang, Gang et al. "*Getting started in gene orthology and functional analysis*" PLoS computational biology vol. 6,3 e1000703. 26 Mar. 2010, doi:10.1371/journal.pcbi.1000703

Hu, Yanhui et al. "*FlyRNAi.org-the database of the Drosophila RNAi screening center and transgenic RNAi project: 2017 update*" Nucleic acids research vol. 45,D1 (2016): D672-D678.

Platt, John. *Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf, C. Burges, A. Smola, eds., MIT Press (1998).

Acknowledgements

Many thanks to Dr. Claire Yanhui Hu for DIOPT data, Dr. Gregory Cary for guidance, and the 02-510 TAs, Prof. Ziv Bar-Joseph, and Prof. Jian Ma