02-510 Computational Genomics
Project Report
5/12/2019

Group Members: Noah Lee (noahl), Chaitanya Srinivasan (csriniv1)

Learned Ortholog Optimized Predictions (LOOP)

Introduction

Identifying orthologs, which are common genes in different species that have undergone speciation events, delineates the genealogy of genes to investigate evolutionary mechanisms and creates groups of genes with the same or similar biological functions. Reconstructing evolutionary relationships between genes in different species can predict gene function without using expensive experimental procedures. Orthologous genes also provide quantitative approaches to access relatedness and inform research into gene groups and evolutionary history. While orthologs are clearly defined for general examples, the edge cases and practicality begin to blur these lines. Furthermore, technological and logistical constraints restrict many powerful prediction tools from benefiting every research group. Certain methods rely on assumptions which are not applicable to all genome assemblies, such as confidence in the global alignment to induce synteny or experimental data for protein-protein interaction networks. Current approaches follow a model similar to the DRSC Integrative Ortholog Prediction Tool (DIOPT) which performs ortholog mapping with high confidence using a suite of 17 prediction tools. As discussed previously, a large quantity of data which includes manual curation is utilized to power the suite of tools. However, this scale of data quantity and quality is not available to every research group. We seek to provide a method which can be applied broadly to more sparsely populated investigations and relies on straightforward and transparent assumptions.

Initial Investigations

Machine learning has not been applied in any of the 17 methods compiled within DIOPT, despite promise from many scientific fields outside of comparative genomics. A descriptive feature space is a strong statistical assumption we initially attempted to use by assigning predicted orthologous gene pairs scores based on a subset of the 17 methods along with classes labeling “true” orthologs based on DIOPT confidence scores. This provided a training set we sought to bring to a less populated ortholog prediction space in the Echinobase Ortholog Project (EBOP).

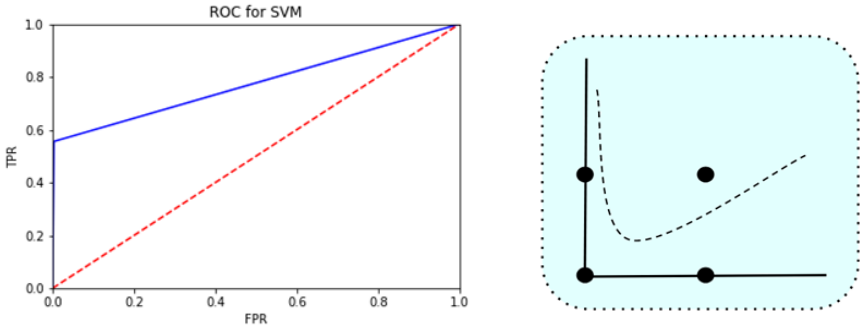

Fig. 1 The graph on the left represents the receiver operating characteristic (ROC) curve produced when testing EBOP data. The blue line represents the true-positive vs. false-positive rate for this model as opposed to the random model (red dotted line). The graphic to the right shows the limitations of a low dimension feature space, wherein the sophistication of the model has little impact on the classification.

While quickly operable, using Support Vector Machines (SVM) appeared to highlight correlations between prediction tools more so than any significant underlying factors influencing traditional ortholog predictions. The predictions made by SVM are likely operating on too small of a feature space given the subset of methods utilized in the test data. Adding additional existing prediction methods to the model would defeat the purpose of pursuing a low-computational cost method applicable to a wide range of data. After yielding no apparent insights by adjusting thresholds or type of learning model, the strong statistical assumptions were dropped and instead a strong biological assumption was pursued: transitivity.

Methods

Transitivity in orthologs refers to the continuity of conserved genes across species. This strong biological assumption is the basis for our method LOOP which creates "cycle predictions" as visualized below.

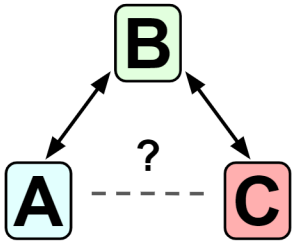

Fig. 2 Gene "A" is predicted as orthologous to gene "B". Gene "B" is predicted as orthologous to gene "C". Transitivity would imply that gene "A" is therefore orthologous to gene "C".

To test this approach, data was taken from both DIOPT and EBOP to procure ortholog predictions based on established prediction tools. From EBOP, predictions from *S. purpuratus* to *D.*

melanogaster using Reciprocal Best Hit (RBH) and InParanoid were generated for this project. From DIOPT, predictions from *D. melanogaster* to *H. sapiens* were graciously provided by Dr. Claire Yanhui Hu.

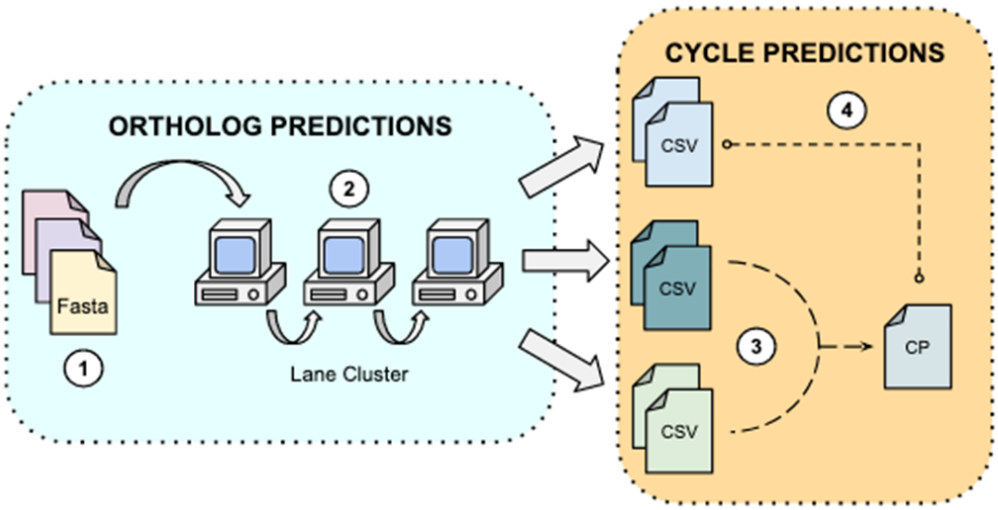

Fig. 3 (1) Fasta files for the 3 species are preprocessed (2) Data from DIOPT, RBH, and InParanoid is generated (17.5 hours) (3) Cycle predictions link SPU<>DMEL to DMEL<>HSAP predictions (0.5 hours) (4) Predictions are verified using RBH and InParanoid (11 hours)

LOOP then generated new ortholog predictions between *S. purpuratus* and *H. sapiens* based on the cycles connected within the other two sets of data. This was achieved by standardizing gene IDs to the first appearance within prediction files which required a large gene synonym file. Below is an example of connecting three genes which are identified by different names across different predictions.

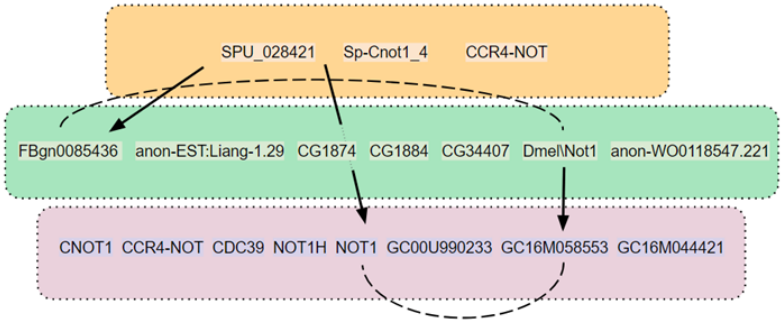

Fig. 4 Each color represents a single gene, with the gene synonyms contained inside. Bold arrows represent ortholog predictions, with dotted arrows representing the matches between synonyms with predictions which should be consolidated. Ideally, all genes would be identified by universal IDs but this ignores the often iterative nature of genome annotation and curation.

The predictions from LOOP were then assessed by using RBH and InParanoid to generate ortholog predictions between *S. purpuratus* and *H. sapiens*. Verification was based on overlap between the LOOP predictions and the RBH and InParanoid predictions.

Results

Many ortholog studies indicate ranges from 5,000 - 15,000 orthologous genes between species, depending on evolutionary distance and species involved. LOOP generated 8,437 ortholog predictions between *S. purp* and *H. sap* with a 43% overlap with predictions generated by the two traditional ortholog prediction methods, RBH and InParanoid.

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. 5 The left table represents the number of LOOP predictions which overlap with at least one other prediction method and are considered "verified". The table on the right breaks down the overlap into predictions made by each tool between *S. purpuratus* and *H. sapiens*.

In addition to the quantity of predictions scaling favorably when compared to other methods, the overlap shows a few favorable trends. InParanoid generates a greater quantity of predictions compared to RBH, which is a stricter method based purely on stringent sequence alignment. While the lowest overlap for LOOP predictions are on IP-only hits at 10.7%, the overlap increases with the more reliable RBH, and a further increase is shown when only gene pairs predicted by both methods were assessed up to 33.33%. Additionally, a total overlap of 43% by LOOP matches an assessment by Fang et. al which shows current prediction tools like InParanoid generating an 68% overlap, OrthoMCL with a 57% overlap, and TreeFam with a 43% overlap when comparing *D. melanogaster* to *H. sapiens*.

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. 6 An ortholog cycle as predicted by LOOP shows the expected similarities in manual curations.

Discussion

The initial aim of creating LOOP was to provide a method with simple assumptions, high utility, and effectiveness as a practical screening tool. While the accuracy, speed, and data pulled from existing predictions for LOOP work to accomplish those goals, the slightly smaller quantity of ortholog predictions relates to a flaw in LOOP's design- a reduction in dimension at each successive pass through the pipeline for each species. If extended beyond three species, we would expect increasingly diminishing numbers of ortholog predictions.

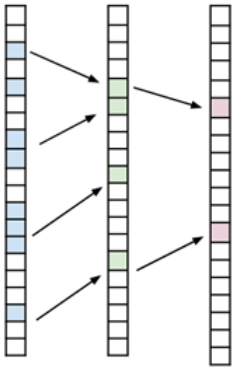

Fig. 7 As predictions move to each successive species, the number of predictions which are shared across all species is reduced, resulting in a reduced dimension space

Additionally, the strong assumption of transitivity belies the nature of orthologs predictions, which are often confounded by evolutionary events such as duplication and subfunctionalization. Duplication is an event by which a gene, or segment of a gene, has an identical copy, and subfunctionalization is the event by which the duplicated genes specialize in different functional roles that were carried out by the ancestral gene. These events complicate the categorization of in-paralogs, out-paralogs, and functional orthologs. LOOP will not be able to categorize these differences between species which have undergone major speciation events reflected in their genomes.

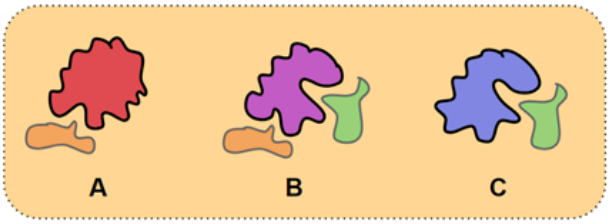

Fig. 8 Through speciation events, gene "A" may have a biological function shared with "B", and "B" may have an additional function shared with "C" which is not found in "A", thus breaking the transitive property.

However, the end results for LOOP remain promising. Other methods, like InParanoid in its 8th release, have undergone many iterations. As a first version, LOOP is a strong contender based on comparative accuracy and simplicity.

Future Steps

Two potential updates for LOOP could address current issues. A stricter threshold for connections could be induced by removing InParanoid predictions which have lower scores, and across additional databases similar thresholds can be implemented at basic levels such as only checking ortholog pairs with a certain amount of agreement between already available methods. Additionally, filtering methods can be applied similar to InParanoid v.8 which are designed to reduce false positives, shown in the figure below.

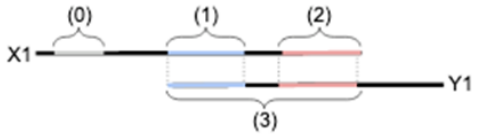

Fig. 9 Potential filtering methods such as (0) masking (1) & (2) overlap distances within genes (3) overlap coverage per gene

Better verification for ortholog predictions could be applied as well. The current analysis method only involved two prediction tools, RBH and InParanoid, but extending to more prediction methods which utilize more sophisticated assumptions could create a better image for how LOOP works across different orthological relationships. Species with closer evolutionary distances would likely work better with LOOP as well, given that less speciation events have occured between more closely related species. These refinements and adjustments point to additional benefits possible with the low-cost LOOP method.

Acknowledgements

Many thanks to Dr. Gregory Cary for guidance on orthologs, Dr. Claire Yanhui Hu for providing data from DIOPT, and the 02-510 staff: Prof. Ziv Bar-Joseph, Prof. Jian Ma, and TAs Cathy Su and Ruochi Zhang for great course material and feedback.

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).