The integrative use of anatomy ontology and protein-protein interaction networks to study evolutionary phenotypic transitions

Pasan C. Fernando, Erliang Zeng, Paula M. Mabee
Biology Department
University of South Dakota
Vermillion, USA
pasan.fernando@coyotes.usd.edu, erliang.zeng@usd.edu, paula.mabee@usd.edu

***Abstract*— Studying evolutionary phenotypic transitions, such as the fin to limb transition, is popular in evolutionary biology. The recent advances in next-generation technologies have accumulated large volumes of genomics and proteomics data, which can be used to analyze the genetic basis for evolutionary phenotypic transitions. Protein-protein interaction (PPI) networks can be used to predict candidate genes and identify gene modules related to evolutionary phenotypes; however, they suffer from low gene prediction accuracy. Therefore, an integrative framework was developed using PPI networks and anatomy ontology, which significantly improved the accuracy of network-based candidate gene predictions in zebrafish and mouse. This integrative framework will also be used to identify gene modules associated with the fin to limb transition and to study the changes in these modules which lead to the phenotypic change.**

Keywords- Anatomy ontology; network analysis; protein-protein interactions; data integration; gene prediction.

I. INTRODUCTION

The process of evolution is accompanied by numerous important phenotypic transitions, such as the fin to limb transition in vertebrates, which contributed to the wealth of phenotypic diversity observed among different species today. Understanding the relationship between genes and their phenotypes is important in explaining the changes in those phenotypes. Traditionally, wet lab methods were used to discover genes to phenotype relations. Despite the higher accuracy of their predictions, wet lab candidate gene prediction methods are high in resource and time consumption, which lead to the popularity of faster computational candidate gene predictions methods [1] that use the genomic and proteomic data accumulated in public databases.

The use of PPI networks for candidate gene prediction has become popular due to the availability of large PPI datasets for model organisms. Network analysis algorithms can be used to analyze PPI networks and detect gene modules corresponding to phenotypes in question [1]. Other gene prediction methods only discover direct gene to phenotype relationships, but network analysis further identifies gene interactions that are important in regulating the phenotype. Understanding the modular structure of gene interactions is extremely important in studying their role in the development of phenotypes because it is the gene interactions that determine the outcome rather than the individual genes.

The biggest challenge of using PPI networks is the low candidate gene prediction accuracy due to the low quality of the networks [1]. The PPI networks are known to contain a higher amount of false positive interactions, and some networks are still incomplete [2]. Before using PPI networks to study evolutionary phenotypic transitions, their quality must be improved to obtain better results. Because we are focusing on anatomical phenotypes, such as the pectoral fin development and the forelimb development, we propose an integrative framework that uses anatomy ontology to incorporate known information about gene-phenotype relationships in literature with the PPI networks. This integration is expected to improve the PPI network quality and predict candidate genes with a higher accuracy. To test this hypothesis, we use known anatomical phenotype annotations from mouse and zebrafish. After the evaluation, the integrated networks will be used to detect gene modules associated with the fin to limb transition in mouse and zebrafish, and the modules will be compared to observe the genetic changes corresponding to the phenotypic transition.

II. METHODS

The first step of the integrative framework is constructing gene networks that are entirely based on the known gene to anatomical phenotype annotations. The anatomical profiles for mouse and zebrafish were downloaded from the Monarch initiative data repository (https://monarchinitiative.org/), which retrieves data from model organism databases. Monarch initiative data is manually pre-processed to remove unwanted annotations and the genes are annotated to Uberon anatomy ontology terms [3]. Uberon (http://uberon.github.io/) is a cross-species anatomy ontology that integrates species-specific anatomy ontologies, such as Mouse Anatomy Ontology (MA) and Zebrafish Anatomy Ontology (ZFA), which makes it suitable for evolutionary analyses involving multiple species [4].

Semantic similarity scores between anatomy ontology terms were calculated to obtain pairwise gene similarity values for all the genes in mouse and zebrafish. Semantic similarity is a quantitative value that represents similarity between two ontology terms based on their location in the ontological structure and their gene annotations [5]. Four different semantic similarity methods (Lin, Resnik, Schlicker, and Wang) were used to generate pairwise gene similarity matrices, which in turn were used to generate gene networks that are entirely based on the anatomy ontology annotations of the genes (anatomy-based gene networks). These networks were filtered using a gene similarity score cutoff to remove interactions with low scores. In these networks, the genes with higher similarity scores are the ones that are annotated to similar anatomy ontology terms.

The PPI networks for mouse and zebrafish were downloaded from the STRING database (https://string-db.org/). Then, the PPI networks were integrated with the anatomy-based gene networks using pairwise gene similarity scores of the two networks in a probabilistic model. In the integrated network, only the gene pairs that receive high similarity scores from both the input networks have high gene similarity scores. To assess the candidate gene prediction performance of the integrated networks and the PPI networks, Uberon anatomy ontology terms that have at least 10 or more gene annotations were used from the zebrafish and mouse anatomical profiles downloaded from the Monarch initiative data repository. Hishigaki prediction method [6] was used as the network-based candidate gene prediction algorithm and leave-one-out-cross-validation was used as the evaluation technique. Receiver operating characteristic (ROC) and precision-recall curves were generated for the comparison of different network types. Although the goal was to compare the integrated versus PPI networks, the anatomy-based gene networks were also included in the comparison.

III. Preliminary Results and Discussion

The ROC and precision-recall curve comparisons for mouse and zebrafish indicate that the integrated networks significantly outperform the original PPI networks when predicting candidate genes (Only the zebrafish ROC curve comparisons of the four semantic similarity calculation methods are shown in Fig. 1). This result is consistent among the four semantic similarity calculation methods used. The higher candidate gene prediction accuracy of the integrated networks means that their network quality was increased during the integration. Although anatomy-based gene networks (shown in blue in Fig. 1) have the highest performance among most of the semantic similarity calculation methods, they are not suitable for candidate gene prediction or identifying network modules because they only contain genes that have at least one anatomy ontology term annotation. This number is low compared to the integrated and PPI networks. For instance, the zebrafish anatomy-based gene network constructed using the Schlicker method contains 5,386 genes, whereas the corresponding integrated network contains 12,755 genes. The integrated networks contain a large number of unknown genes coming from PPI networks, which can be potential candidates for anatomical phenotypes. Therefore, integrated networks are more useful for downstream network analysis.

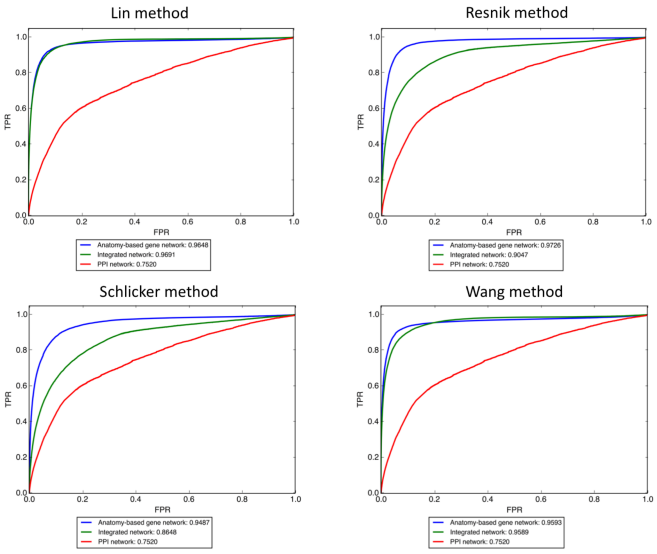

Fig. 1. Comparison of ROC curves for PPI (red), integrated (green), and anatomy-based gene (blue) networks for the four semantic similarity methods. The integrated and anatomy-based gene networks clearly outperform the PPI networks when predicting candidate genes.

The integrated network with the highest performance for mouse and zebrafish will be used for detecting gene modules associated with the fin to limb transition. Because the quality of the integrated networks is higher than the PPI networks, the gene modules will be more accurate. The gene modules for pectoral fin and pelvic fin in zebrafish will be compared with gene modules for forelimb and hindlimb in mouse, respectively, to identify modular changes genes during the fin to limb transition. This work showcases how anatomy ontology can be used to improve the quality of candidate gene predictions and to perform efficient network analyses to study evolutionary transitions.

References

[1] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," Molecular systems biology, vol. 3, 2007, p. 88.

[2] C. von Mering *et al.*, "Comparative assessment of large-scale data sets of protein-protein interactions," Nature, vol. 417, 2002, pp. 399-403.

[3] C. J. Mungall *et al.*, "The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," Nucleic Acids Res, vol. 45, 2017, pp. D712-D722.

[4] M. A. Haendel *et al.*, "Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon," Journal of biomedical semantics, vol. 5, 2014, p. 21.

[5] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, "Semantic Similarity in Biomedical Ontologies," PLoS Comput Biol, vol. 5, 2009, p. e1000443.

[6] H. Hishigaki, K. Nakai, T. Ono, A. Tanigami, and T. Takagi, "Assessment of prediction accuracy of protein function from protein–protein interaction data," Yeast, vol. 18, 2001, pp. 523-531.