

# Network Properties of a Gene Network Built from *Drosophila melanogaster* Data

James Costello  
School of Informatics  
Drosophila Genome  
Resource Center  
jccostel@indiana.edu

Mehmet Dalkilic  
School of Informatics  
dalkilic@indiana.edu

Justen Andrews  
Dept. of Biology  
Drosophila Genome  
Resource Center  
jandrew@bio.indiana.edu

Center for Genomics and Bioinformatics  
Indiana University  
Bloomington, IN 47405

## 1. INTRODUCTION AND MOTIVATION

Recently, much attention has been given to the analysis of protein interaction networks built from high-throughput (HT) experimental data, such as yeast-two-hybrid assays. Specifically, because of long-standing interest and study, comprehensive experimental data, and simplicity of having a single cell type, *Saccharomyces cerevisiae* has been the focus of most computational research [6, 11]. Research topics in protein interaction networks run the gamut of biology from evolution [5] to functional annotation [9]. These studies have unveiled a great deal of information, however the scope of potential questions that can be asked from the yeast protein interaction network is limited. We believe a much richer representation of gene function can be sought through the exploration of different data sources in more complex organisms, where the focus of this work is in the model organism *Drosophila melanogaster*.

In addition to protein interaction data, the availability of HT genomic data (*e.g.* transcriptional profiling or RNAi screens) has made it possible to better interpret currently available and always increasing amounts of genetic data (*e.g.* genetic screens or *in situ* hybridizations). Each one of these data sources on their own offers insight into genes and their functional relationships; however, each dataset has a rather narrow focus of biological inquiry. The ultimate elucidation of gene function within an organism (considering time, developmental stage, localization, etc.) will involve knowledge derived from multiple (apparently) *orthogonal* data sources. Although these data sources individually aim to answer different questions, they can be complementary in many cases, where data from one source can be leveraged to inform or confirm the existence of a relationship in another.

This logic is corroborated by previous studies in yeast and worm which have shown that integration of different biological data sources results in biologically meaningful results. Specifically, Gunsalus, *et al.* [4], integrated transcriptome, proteome, and phenome data in *C. elegans* resulting in “wet-lab” testable predictions which subsequently lead to discovery of new gene function. As another illustration, Kemmeren, *et al.* [8], combined multiple HT data sources in *S. cerevisiae* to show that data integration provides a reliable

means to computationally predict unknown gene function.

Inspired by this trend, recent work by our group [2] has shown the utility of integrating disparate sources of data into a gene network of functional relationships in *Drosophila melanogaster*. This network has been shown to be biologically consistent with genes of known functions (Sec. 3).

*Drosophila* is one of the most widely and deeply studied model organisms, but little has been published on the creation of such an integrated network. This type of network is a valuable platform of research because, *i)* *Drosophila* is a metazoan, so compared to yeast, the complexity of multiple cell types and a greater number of genes will have to be considered—consequently, understanding of our own genome can be significantly improved (*e.g.* 70% of human disease genes have homologs in fly [10]); *ii)* the genetic and genomic data available for *Drosophila* is different than other organisms; and *iii)* there has been little done on network analysis of functional gene relationship networks built from multiple data sources.

In this paper, we aim to show some of the construction and basic properties of an integrated *Drosophila* network and will also highlight a few interesting questions that have arisen from our research.

## 2. NETWORK CONSTRUCTION

An integrated network was constructed from microarray co-expression, protein-protein interaction, and genetic interaction data specific to *Drosophila*. Also, data from transcription factor binding sites and allele phenotypic annotation was considered, but because of the nature of this data, was used to a much lesser extent. (Please refer to [2] for references to all datasets used)

The network construction was done by first creating a filtered, non-redundant set of genes from each dataset. Microarray data was normalized and consolidated. Next for a pair of genes  $g_{i,j}$ , a vector was created,  $g_{i,j} = \langle m_1, m_2, \dots, m_n \rangle$ , where  $m$  is a value (correlation or binary) connecting  $i$  to  $j$  from a particular data source, and  $n$  is the number of data sources. In order to create an edge (representing a functional relationship between two genes) within the network, the vector of information associated with each gene pair is subjected to a set of “biological” rules over the vectors. These rules require thresholds to be set for each individual dataset,

while employing both standard and Bayesian techniques.

The resulting network consists of  $\sim 12\text{K}$  vertices (genes) and  $\sim 203\text{K}$  edges (functional relationships). Further analysis of the network was done on the largest connected component, which consisted of 11,768 vertices and 202,594 edges. These 11,768 genes represents  $\approx 86\%$  of the *Drosophila* genes as reported by the version 4.3 annotation<sup>1</sup>.

### 3. RESULTS

Several types of analysis were carried out to answer the following: *i*) Is the integrated network consistent with known biological data; *ii*) Does the integrated network offer more information than can be obtained from an individual data source; and *iii*) What are the inherent properties of the integrated network.

The first two questions are more extensively addressed in Costello, et al. [2], but are briefly highlighted. The relative contributions (measured as the percentage of edges created by each dataset) of the datasets to the integrated network are as follows: the total of all three major microarray datasets is 0.853, protein-protein interaction is 0.127, and genetic interactions are 0.025. The Relative contributions of each dataset to the network is most likely a function of the number of genes represented in the different datasets, where microarray platforms can cover an entire transcriptome as compared to a genetic complementation test of two alleles.

As predicted, genes annotated with a biological function from GO [3] are more tightly connected as compared to a random sampling of nodes from the network. This was tested by grouping genes according to GO terms, then measuring the number of direct connections and fully connected sets of three nodes. These results were compared to a random sampling in the network. Genes annotated with a common biological process GO term were significant ( $z$ -score of over 2) in 80% of the measured cases and fully connected graphs of three nodes were significant in 73% of the measured cases.

As a complementary verification of our ability to recover biologically significant relationships in the network, pathways from the KEGG [7] database were tested for coherency. These pathways offer a fundamentally different way to measure gene relationships in biological processes and were shown to be coherent in the integrated network above noise.

The integrated network is consistent with the power law distribution. If the network scales as  $P(k) \sim k^{-\gamma}$  [1], where  $k$  is the degree of a node, then our  $\gamma \approx -1.3$ .

The clustering coefficient,  $\langle C \rangle$ , for the entire graph was calculated to be  $\approx 0.188$ , where  $\langle C \rangle = \frac{1}{N} \sum_{i=1}^N C_i$  [1],  $N$  is the number of vertices in the network and  $C_i = \frac{2n_i}{k_i(k_i-1)}$ . Also, the mean path length,  $\langle l \rangle$ , was calculated to be  $\approx 5.3$ , where  $\langle l \rangle = \frac{2}{N(N-1)} \sum_{i < j} l_{i,j}$  [1],  $N$  is the number of vertices in the network and  $l_{i,j}$  is the shortest path between node  $i$  and node  $j$ .

### 4. DISCUSSION

Exploration of the integrated network raises a number of significant statistical and computational questions which re-

<sup>1</sup><http://www.flybase.net/annot/dmel-release4-notes.html>

late to how the properties of the network can inform relationships among genes. Although all the topics of interest cannot be fully discussed, a series of issues related to topological properties, their identification, and their biological relevance are discussed below.

Let  $G_C = \langle V, E, c \rangle$  be an undirected graph with edges annotated with a subset of colors  $c \in 2^C$  from a color set  $|C| \geq 2$ , where the color set represents support from multiple datasets. We pose a number of interesting problems: *i*) Given two vertices  $v_1, v_2$  find the shortest path such that any two adjacent edges along the path differ by at least one color (biological perspective—requires explicitly different data sources for functional support); *ii*) Given three vertices  $v_1, v_2, v_3$ , find the shortest path between  $v_1, v_2$  as in (*i*), including  $v_3$  (biological perspective— $v_3$  is a gene if unknown function); *iii*) Given a second smaller graph  $G'_C = \langle V' \subset V, E' \rangle$  find, if there exists, a “reasonably close” (biologically meaningful) isomorphism of  $G'_C$  in  $G_C$  (biological perspective—find whether a molecular pathway exists in the gene network); *iv*) Given a set of graphs over the same vertex set, but whose edge set can differ and are annotated with a single color  $\mathcal{G} = \{G_{red}^1, G_{red}^2, G_{blue}, C_{green}^3, \dots, G_{yellow}^k\}$  discover whether there is a meaningful “integration” of the graphs—do relationships in graphs seem to “support” each other (biological perspective—basic integration problem)

Ultimately, our research will turn to the bench where functional relationships predicted from the integrated network must be experimentally verified.

### 5. ADDITIONAL AUTHORS

Rupali Patwardhan (email: [rpatward@indiana.edu](mailto:rpatward@indiana.edu)),  
Sumit Middha (email: [smiddha@indiana.edu](mailto:smiddha@indiana.edu)),  
Brian Eads (email: [beads@cgb.indiana.edu](mailto:beads@cgb.indiana.edu)),  
John Colbourne (email: [jcolbour@cgb.indiana.edu](mailto:jcolbour@cgb.indiana.edu)),  
Junguk Hur (email: [juhur@indiana.edu](mailto:juhur@indiana.edu)),  
Keval Mehta (email: [kymehta@indiana.edu](mailto:kymehta@indiana.edu)).

### 6. REFERENCES

- [1] A.-L. Barabási, Z. N. Oltvai, and S. Wuchty. In *Complex Systems*. Springer Lecture Notes in Physics, New York, NY, 2003.
- [2] J. Costello, M. Dalkilic, R. Patwardhan, and et al. In *ISMB (submitted)*, 2006.
- [3] Gene Ontology Consortium. *Genome Research*, 11(8):1425–1433, 2001.
- [4] K. C. Gunsalus, H. Ge, A. Schetter, and et al. *Nature Letters*, 436:861–865, August 11 2005.
- [5] M. W. Hahn, G. C. Conant, and A. Wagner. *Journal of Molecular Evolution*, 58:203–211, 2004.
- [6] T. Ito, T. Chiba, R. Ozawa, and et al. *PNAS*, 98(8):4569–4574, 2001.
- [7] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. *Nucleic Acids Research*, 32(Database Issue):D277–D280, 2004.
- [8] P. Kemmeren, T. Kockelkorn, T. Bijma, and et al. *Bioinformatics*, 21(8):1644–1652, 2005.
- [9] S. Letovsky and S. Kasif. *Bioinformatics*, 19(Suppl.1):i197–i204, 2003.
- [10] L. Reiter, L. Potocki, S. Chien, and et al. *Genome Research*, 11(6):1114–1125, 2001.
- [11] P. Uetz, L. Giot, G. Cagney, and et al. *Nature*, 403(6770):623–627, 2000.