Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Using *Drosophila melanogaster* Data to Discover Disease-Related Protein Interactions in Humans

James C. Costello*†‡¶, Jade E. Buchanan-Carter*†, Mehmet M. Dalkilic*†, and Justen Andrews†‡§
*School of Informatics
†Center for Genomics and Bioinformatics
‡Drosophila Genome Resource Center
§Department of Biology
Indiana University
Bloomington, IN 47405
¶Contact Author: jccostel@indiana.edu

***Abstract*— The discovery and understanding of functional relationships amongst genes and their gene products are fundamental to our understanding of human disease. Data regarding protein-protein interactions in humans are not complete and computational techniques along with data provided from other organisms can be used to better inform relationships in humans. We demonstrate the utility of using genome scale data from *Drosophila melanogaster* to predict protein-protein relationships for human proteins specific to disease. To find the most likely candidates for protein-protein interaction, the predicted relationships are tested against human protein interaction and known disease data, then ranked through a Support Vector Machine. An illustrative example related to hBrm and hSNF5 shows the validity of our approach.**

I. Introduction

Most of our knowledge about human disease genes is based on associations between phenotypes and genetic variation. The current state of knowledge includes approximately 7,000 human genes associated with 1,500 diseases as found in OMIM (Online Mendelian Inheritance in Man)[1]. However, the vast majority of human genes are poorly characterized for function, if at all[2]. The fact that the products of most genes function at the cellular level in interconnected pathways of interacting molecules, provides a means of prioritizing research efforts on likely disease-related candidate genes. Elucidating the interacting partners and the genetic pathways in which human disease genes act, enhances our understanding of these respective diseases and also identifies additional potential drug targets[3].

The recent blossoming of genome-wide studies of molecular gene function in humans provides a ready means to identify putative functional relationships between disease genes and previously uncharacterized or poorly characterized genes[4], [5]. The general approach is to identify functional relationship-based assays of human genes, such as protein-protein interactions, microarray co-expression, gene knockdown phenotypes, or integration of such data sets.

A complimentary approach takes advantage of the fact that many genes and gene pathways are widely conserved across genera including well characterized model organisms such as *S. cerevisiae, C. elegans, D. melanogaster and M. musculus*. For instance, approximately 75% of known human disease genes have homologs in *D. melanogaster*[6]. Additionally, many of the human disease relevant signaling pathways are conserved across metazoans[7], [8], [9]. It follows that gene-gene functional relationships established in these experimentally tractable and well-studied organisms can be used to infer putative functional relationships between pairs of orthologous human genes[6], [10]. As a recent illustration, Cooper *et al.* [11] used yeast to study the molecular effects of α-synuclein misfolding and discovered a yeast gene, *Rab1*, whose ortholog in mouse – a mammalian model for Parkinson's disease [12], [13] – protected against neuron loss when expressed.

Drosophila melanogaster is arguably one of the most well-studied model organisms and with this distinction comes a great deal of genetic and genomic data. Consequently, it has been used to model a number of diseases [6], [14], including cancer [10] and neuromuscular diseases [15], [16], [13]. In this study we aim to leverage *Drosophila* genetic and genomic, experimental data to perform a systematic genome-scale analysis to better inform potential functional relationships of human genes involved in disease.

The term *interolog* was first introduced by Walhout, *et al.* [17] to describe interacting gene pairs that are conserved across species, where both their sequence and interacting relationship is conserved. This term has propagated through the literature [18], [19], but has mostly been used in the context of conserved protein-protein interactions. In fact, work by Bandyopadhyay, *et al.* [20], used protein interaction network alignments between fly and yeast to assign gene function annotations. Although it has been shown that protein-protein interactions are more conserved within species than across species [21], as described above, there are still many examples of conserved pathways across species that can be used to study biological processes.

Our approach considers the potential complications that may arise from looking solely within protein-protein interactions and draws from the multitude of data and data sources available in *Drosophila.* Conceptually, we are looking to find interologs between human and fly, but from the unique perspective of human disease genes, where the functional re-

1-4244-0710-9/07/$20.00 ©2007 IEEE

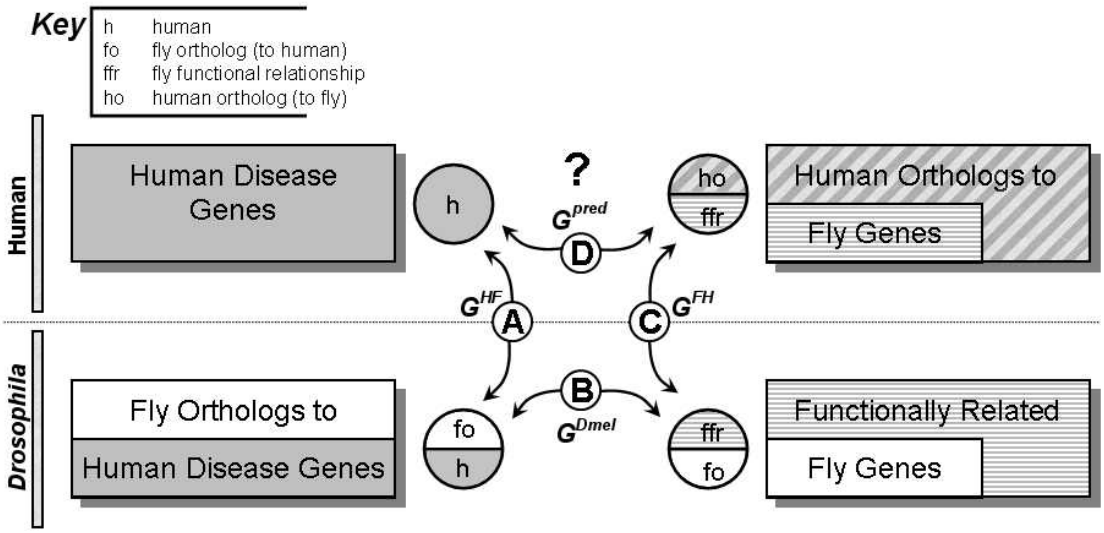

Fig. 1. Conceptual diagram of the methodology used to predict human protein relationships. Edge A represents the set of relationships between human proteins associated with disease and orthologous fly proteins, G^{HF}. Edge B represents the set of functional relationships between fly genes, G^{Dmel}. Edge C represents the set of fly genes that have human orthologs, G^{FH}. G^{pred} are the predicted relationships.

lationships are derived from physical and genetic interactions, along with gene co-expression, shared transcription factor binding sites, and annotated localization in *Drosophila*.

In Figure 1, we illustrate the flow of our methodology. Initially, we gather fly orthologs to human disease genes (A). Next, we find fly genes that are functionally related to these fly orthologs (B). We then look for human orthologs to the functionally related fly genes (C). Lastly, we draw potential connections between human genes represented by the question mark next to (D). Traveling this path allows us to gather a set of features which support the gene pair relationships, such as sequence similarity or experimental measures. Some of the disease relationships we predict in human can be verified against known data, while the vast majority cannot. Ultimately we break this problem down to a classification task, where we employ the proficiency of a Support Vector Machine to determine which predicted human gene pairs are most likely to interact and be involved in the same disease or disease pathway.

To test whether our prediction method returns biologically relevant results, we show one example of a high scoring, predicted relationship that is concordant with published data, but this relationship has yet to be reported in OMIM.

The procedures and computational methods we employed are covered in Section II. The results of our approach along with data validation and an illustrative example are covered in Section III, and a short discussion closes the paper in Section IV.

II. Methods

The methods used to predict human protein relationships with respect to fly data is conceptually illustrated in Figure 1. To best understand the computational techniques and procedures, we use this figure as a guide and refer back to it in the following subsections.

A. Collection of Human to Fly Candidate Orthologs

The methods to determine relationships between human proteins associated with disease and *Drosophila* proteins are described in this section representing edge A in Figure 1.

To start, a set of proteins involved in human disease and their homologous sequences in *Drosophila* are downloaded from the v2.1 Homophila database[22]. Homophila provides a mapping from human proteins associated with disease to fly proteins, which are ranked on BLAST[23] e-values. Disease relationships within Homophila are defined by NCBI's OMIM database. For our analysis, we created a candidate set of human-fly homologs, G^H, based on an e-value cutoff of 10^{-20}.

To increase the likelihood of comparing functionally similar proteins, the candidate set taken from Homophila were filtered using the Inparanoid[24] algorithm, which utilizes reciprocally best-matching BLAST scores to determine orthologs and additionally in and out-paralogs[25]. Protein pairs determined to be orthologs were only considered, thus ignoring the in and out-paralog relationships. The justification for such a filter is demonstrated by Hulsen *et al.*[26], who found by combining sensitivity and selectivity of ortholog identification, Inparanoid ranked highest among other methods in terms of identifying functionally equivalent genes. The set of human-fly protein pairs filtered through Inparanoid is designated, G^{HF}, where $G^{HF} \subset G^H$.

B. Finding Functional Relationships through Fly Data

The methods to determine functional relationships between fly genes are described in this section representing edge B in Figure 1.

The popularity of *Drosophila* as a model organism has resulted in the generation of large amounts of genome-scale experimental data; gene expression profiles, protein-protein interaction, and genetic interaction to name a few. Each one of these techniques captures a narrow and specific perspective of the entire system and the integration of these orthogonal datasets can provide a more informed and comprehensive perspective of relationships between genes, as compared to any individual dataset[27], [28]. This perspective is consistent with work from *C. elegans*[29], [30] and *S. cerevisiae*[31], [32].

For this study, we collected genome-scale data from *D. melanogaster* in the form of gene co-expression[33], [34], [35], protein-protein interaction from FlyGrid[1], DIP[2], MINT[3], and BIND[4], genetic interaction from Flybase[5], phenotypic annotation from Flybase[36], and shared transcription factor binding sites[37]. These data were then cleaned, normalized, and stored in a MySQL database. Since each of the data sources convey different information regarding functionally related genes, each data source has its own measure of a gene-gene relationship. For protein-protein interaction data, a pair of genes is given a value of 1 if they are found to interact. Gene pairs that are reported to genetically interact are given a value of 1 and a value of 0 if they are experimentally determined not to genetically interact. The gene pair relationship measure for microarray data is given as the Pearson correlation between the expression profiles of two genes, across the experiment. For transcription factor binding site data, a correlation measure is assigned to the number of known binding sites shared in the upstream regions of any two genes. Lastly, the measure of phenotypic annotation was measured using Lord's [38] measure of semantic similarity, since the annotated terms belong to an anatomical ontology [39]. Semantic similarity is an *information content* based measure which leverages the fact that terms further from the root of the ontology will be used less frequently and thus carry more information.

A connection between two fly genes is drawn to reflect the functional relationships between the two genes derived from the aforementioned experimental data. Specifically, connections were created if there is a reported genetic interaction, protein-protein interaction, significant co-expression correlation, or significant similarity in phenotypic annotation. Often times a gene pair has supporting evidence from more than one data source, which increased our confidence in that relationship and exemplifies the complementary nature of the different data sources.

[1]http://biodata.mshri.on.ca/fly_grid/servlet/SearchPage
[2]http://dip.doe-mbi.ucla.edu/dip/Main.cgi
[3]http://mint.bio.uniroma2.it/mint/
[4]http://www.bind.ca/Action
[5]http://www.flybase.net

The fly orthologs from G^{HF} are matched to the list of determined relationships, and the query gene's partner was taken to construct a set of fly-fly functionally related gene pairs, G^{Dmel}. *Drosophila* genes from G^{HF} that did not match any integrated fly relationships are subsequently excluded.

C. Determining Human Orthologs from Fly Interactions

This section describes the methods used to determine fly and human orthologs representing edge C in Figure 1

The set G^{Dmel} contains fly-fly gene relationships, where fly orthologs of disease related human genes are matched with functionally related fly genes. In order to find new human relationships, the functionally related fly genes are mapped back to human orthologous proteins. To accomplish this task, we again utilized the Inparanoid [24] algorithm to create the set of fly-human orthologs, G^{FH}. Since we do not expect all fly genes derived from the *Drosophila* data to have orthologous sequences to human, the fly genes that do not map to human are ignored.

D. Predicting New Human Protein Relationships

To complete the cycle in Figure 1, the relationships from G^{HF} to G^{Dmel} to G^{FH}, are traced, thus providing a **Human** $\rightarrow$ **Fly** $\rightarrow$ **Fly** $\rightarrow$ **Human** sequence of genes/proteins. By drawing a connection between the two human genes, the prediction of gene relationships is made and subsequently tested. The final set of human-human predicted relationships is designated G^{pred}. At this point we have found relationships consistent with the interolog concept, where a functional related fly genes are used to predict potential relationships in humans.

The predicted human protein relationships require that the original human protein be associated with at least one disease, since that protein was extracted from Homophila, but the partner human protein may not be associated with a disease. Thus, disease associations are mapped onto the predicted relationships and verified against OMIM. As a second method of verification, relationships from G^{pred} are tested against the repository of human protein interactions, OPHID (The Online Predicted Human Interaction Database)[40], to determine how many predicted relationship are concordant with known and high-confidence human protein interaction data.

III. RESULTS

Applying our method, we produce the following numbers. From Homophila, we extracted 1,340 human-fly protein homologs associated with a disease. Of the 1,340, there are 774 unique human-fly protein pairs, G^H = 774. The reduced number is attributable to a single human protein being associated with more than one disease. To find putative orthologs, G^H is then filtered through Inparanoid, which reduces the human-fly protein homologs to 332 distinct human-fly protein orthologs, G^{HF} = 332. The genes in G^{HF} are associated with 556 distinct diseases. The *Drosophila* proteins in G^{HF} are then mapped to the integrated fly data where 274 of a potential 332 are present, meaning that there are 58 genes from G_{HF} that

are not accounted for in the experimental fly data. These 274 genes cover 466 distinct diseases. Of these 274 fly genes, 7,893 fly-fly functional relationships are extracted, G^{Dmel} = 7,893. From these fly relationships, there are 4,056 new partner genes. In other words, the 274 fly orthologs of human genes combine with 4,056 unique fly genes to create 7,893 unique fly-fly functional relationships. Thus, on average, each of the 274 fly genes has $\approx$ 27 partners determined from the fly data. From G^{Dmel}, the 4,056 unique fly genes are then mapped to human, where 1,890 are determined to be orthologs, G^{FH} = 1,890. Consequently, 2,166 fly genes and their corresponding relationships are removed from G^{Dmel}. The full procedure results in a final set of 3,941 predicted human-human relationships, or G^{pred} = 3,941. This set of 3,941 relationships is composed of 2,035 distinct human proteins. Lastly, to explore the disease association among the proteins in G^{pred}, each of the 2,035 proteins are annotated with known disease associations. The proteins extracted from Homophila have already been associated with disease, however, the partner protein in a given pair must be assigned an annotated disease. The partner protein is text-mined against the description field in OMIM, where associated diseases are matched to that protein. In total, 725 diseases are covered by the 2,035 proteins. From G^{pred}, 1,610 of the 3,941 relationships ($\approx 41\%$) have an associated disease with both proteins in the protein pair. The remaining 2,331 of the 3,941 relationships ($\approx 59\%$) have disease associations with the protein extracted from Homophila, but not its predicted partner protein.

The following three subsections describe the approaches taken to determine the statistical significance of the final predicted set of human proteins, rank and validate the interactions, and test a new prediction through literature review.

A. Validation and Statistical Significance

The premise of this approach is that functional relationships between genes are conserved between flies and humans, thus the human gene pairs identified in G^{pred} are predicted to have putative functional relationships in humans. Obviously, this method will be subject to errors stemming from inaccuracies in determining correct orthologs and the evolution of new gene functions. To test the utility of this approach, we assess whether G^{pred} is significantly enriched for true positives. We define a true positive as a human gene pair in G^{pred} that is known to either (*i*) be associated with the same disease, or (*ii*) have a protein-protein interaction.

The first statistical measure aims to test the recovery of gene pairs associated with the same disease. We found 71 pairs in G^{pred} that shared a common disease, where a common disease is an exact text match between disease names. We tested the significance of this recovery with respect to a set of randomly generated protein pairs of equivalent size over 10,000 iterations. As shown in Figure 2 A, B, two distributions were created, where the first distribution sampled random pairs from the list of 2,035 unique proteins in G^{pred}, and the second distribution sampled random pairs from a list of proteins defined by the HGNC (HUGO Gene Nomenclature Committee)[6]. So, although 71 protein pairs may seem like a low number, it has a p-value of $>> 10^{-20}$ when compared to the 2,035 proteins from G^{pred} and a p-value of 10^{-6} when compared to all HGNC proteins.

Using a similar approach we tested the recovery of gene pairs previously known or predicted to have human protein-protein interactions. We used OPHID (The Online Predicted Human Interaction Database)[40], June 2006, as the source of true positives. This data set contains over 9,000 proteins comprising over 50,000 human protein interactions. The protein interactions within OPHID are compiled from human, as well as model organism data, and uses supporting evidence from co-expression, Gene Ontology (GO) terms, protein domains, and literature mining to strengthen its predictions. Since data from other model organisms was used in the construction of OPHID, including *Drosophila*, any interactions derived from fly were removed from the OPHID data. Each of the 3,941 relationships in G^{pred} were tested against OPHID and 47 relationships are found. As with the association of a shared disease, the statistical significance of these numbers were then tested against a set of randomly generated protein pairs tested against OPHID interactions (fly data excluded). As shown in Figure 2 C, D, two distributions were created, where the first distribution sampled random pairs from the list of 2,035 unique proteins in G^{pred}, and the second distribution sampled random pairs from a list of proteins defined by HGNC. Consistent with the disease association calculations, the 47 protein pairs from G^{pred} found in OPHID may seem like a low number, but it has a p-value of $>> 10^{-20}$ when compared to the 2,035 proteins from G^{pred} and a p-value of $>> 10^{-20}$ when compared to all HGNC proteins.

B. Ranking Predicted Gene Relationships

The relationships predicted through *Drosophila* data are not specific to protein interactions. In fact, outside of the experimental protein-protein interactions in fly, the remaining data suggest the presence of functional relationships which may or may not include a physical protein interaction. Knowing this and considering the protein interaction relationships that are concordant with OPHID, we explore the possibility of finding other protein pairs in G^{pred} that share features with the 47 protein pairs found in OPHID. This same logic can be applied to the disease relationships, where we do not expect the fly data to specifically predict disease, but we can look for gene pairs that share similar features to the 71 proteins that have a common annotated disease. To accomplish this task, we employ a Support Vector Machine (SVM) to separate and rank order the predicted relationships.

The generation of features to test and train the SVM were derived from characteristics of the fly and human proteins resulting in a list of 24 features. These features include measures of similarity from Gene Ontology (GO)[41] terms between fly-human and fly-fly pairs (electronically inferred annotations were ignored), where the measure of similarity between two

[6] http://www.gene.ucl.ac.uk/nomenclature/

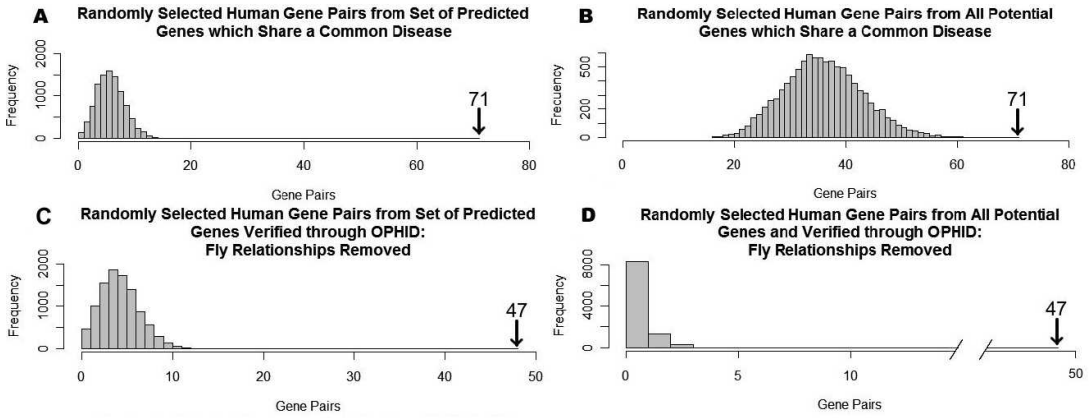

Fig. 2. Distributions showing the significance of the predicted human protein pairs, designated by the black arrows, as compared to randomly generated gene pairs. Plots (A) and (B) show the significance of predicted protein pairs as they relate to known disease. Plots (C) and (D) show the significance of predicted protein pairs as they relate to protein interactions found in OPHID. The randomly created pairs from plots (A) and (C) are sample from proteins found in G^{pred} only, while random pairs created in plots (B) and (D) are sampled from all genes in HGNC. The p-value of plot (B) is 10^{-6}, while the p-values for all other plots are $>> 10^{-20}$.

genes is Lord's measure of semantic similarity [38]. When considering the GO classes of Biological Process, Molecular Function, and Cellular Component across the human → fly, fly → fly, fly → human, and human → human relationships (A, B, C, and D in Figure 1); this results in 12 features. BLAST e-values for fly and human protein pairs add 2 more features. The functional data defining fly-fly relationships consisted of 3 microarray experiments, reported protein interactions, reported genetic interactions, phenotypic similarity, and common transcription factor binding sites, which qualifies as 7 features. Also related to the functional data, we wanted to take into account the number of experimentally determined partners of the fly genes in a functionally related pair, as some genes have many more partners than others. Thus, for each of the two genes in the interacting pair, we counted the number of neighbor genes for each gene and also considered the number of neighbor genes shared between the two paired genes, resulting in 3 more features. Considering all this data results in a total of 24 features.

For classification of protein interactions, the 47 protein pairs found in OPHID were treated as positives, and for classification of proteins involved with the same disease, the 71 protein pairs known to share a disease were treated as positives. The remaining relationships in G^{pred} in either set are treated as negatives.

For each classification task, a 10-fold cross-validation was done, where both positive and negative data were evenly split into 10 disjoint sets. One set was treated as the test set, while the remaining 9 were combined to form the training set. SVM^{light} v6.01[42] was used to train a linear kernel model on the training set, which was used to classify the test set. During training, the cost-factor parameter was set to balance the error associated with the size imbalance in the positive and negative data. This procedure was followed for each of the 10 disjoint sets. The data from each of the 10 classification results were combined to test the overall efficacy of classification, which is reflected in the ROC curves shown in Figure 3. When plotted, the AUC (area under the curve) for the protein interaction classification was calculated to be 75.7% and the disease classification was calculated to be 78.4%, which suggests there is a definite bias in the relationships predicted in G^{pred}. The false positive rate and false negative rate are reflected in the ROC plots (Figure 3).

Every protein pair from each of the classified 10-fold cross validation sets is given a logistic function score, which can then be converted into a probability. The probabilities of each protein pair are then used to rank order the predicted relationships. Since there is a probability associated with a prediction for both potential protein interactions and potential disease relationships, a joint probability can be created to reflect a potential physical interaction and a shared disease. This measure can then be used to rank order the predicted pairs. The results from the SVM classification have been made available in a searchable web interface.[7].

C. Verification of a Predicted Relationship

As a specific validation of our method, we explored the top ranked protein pairs from the combined protein interaction and disease relation scores with respect to relevant literature. Manual inspection of the literature revealed that many of the putative relationships indeed were supported; here we highlight one example whose protein interaction is not reported in OPHID and a connection can not be found in OMIM. We identified the human gene pair hBrm (also known as hSNF2α, Q9H836) and hSNF5 (P51531) via the following path: hSNF5 (human disease gene) → Snf5-related 1 (fly ortholog to human disease gene) → brahma (functionally related fly gene) → hBrm.

[7]http://www.monkey.informatics.indiana.edu/disease/

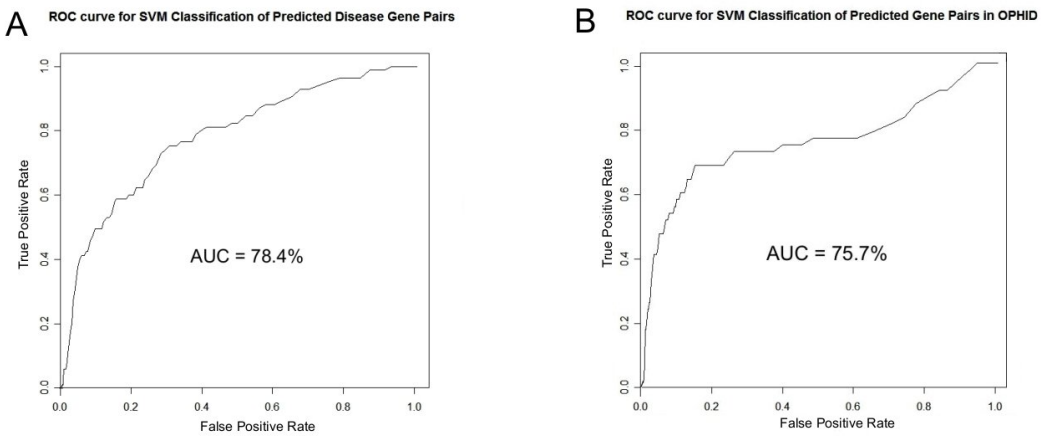

Fig. 3. ROC curves constructed from the 10-fold cross validation results of the SVM clssification. Plot A was generated from training a SVM on protein pairs which have a shared disease, and plot B was generated from training a SVM on protein pairs found in OPHID.

In both humans and flies, the proteins encoded by these genes are subunits of the large multimeric SWI/SNF chromatin remodeling complex. In humans, the literature reports that mutations in hSNF are associated with rhabdosarcomas and T cell lymphomas[43], [44]. OMIM, on the other hand, reports that hSnf5 is associated with Leukemia and Gene Expression, while hBrm is reported to be associated with Rhabdoid Predisposition Syndrome in OMIM. Proteins that are involved in large complexes may never physically interact with each other due to their location in the complex, so not finding a reported protein-protein interaction in OPHID is not surprising. However, their connection to each other is in their potential to be involved in the same disease as subunits in the same protein complex. Our prediction of this high scoring relationship complemented by the literature, and the fact that these two proteins are subunits of a protein complex suggest the pair should be investigated further.

IV. Discussion

The results of our method show the utility of using *Drosophila* genome-scale data to discover new relationships among human proteins. Specifically, the task was to explore *Drosophila* data to predict protein relationships in human specific to disease in a genome-scale, systematic manner. As shown through SVM classification of gene pairs found in OPHID and shared disease relationships among human proteins, the data suggest there are putative relationships that need to be either validated or simply investigated. In fact, the combination of ranks derived from the two classification tasks offers a simple measure of the most likely protein pairs that interact and may share a disease, which is shown through our findings between hBrm and hSNF5.

The approach we have taken to detect new protein relationships relies on integration of multiple kinds of *D. melanogaster* data. The nature of the fly data is not specific to any particular kind of functional relationships among genes. However, through the classification tasks performed by a SVM, the data *do* reveal separable associations that reflect specific molecular relationships, which is extremely encouraging for an applied systems approach with benefits to, for example, drug target prediction. Further, it would be of interest to explore the extent of what relationships are and are not predictable through our methodology beyond the ones tested in this paper.

Another consideration in predicting relationships through an integrative approach is whether the integration of data is supplying new information or whether all of the information can be garnered from one data source. Of the 47 predicted protein pairs that were found in OPHID, the functional connections made in fly are supported by 27 genetic interactions, 11 protein interactions, and another 15 were from microarray experiments. Of the 71 predicted protein pairs that have a shared disease from OMIM, the relationships in fly are supported by 32 genetic interactions, 17 protein interactions, and 28 from microarray data. Keep in mind that any particular predicted pair may be supported by multiple data sources. The distribution of different data sources amongst our predicted protein pairs that are concordant with known human data *do* validate that integration of the different data sources supplies new information. This method also adds support to the concept of interologs.

In summation, the integrative approach of using model organism data to inform relationships in human is not new[7], [11], [45], [46]; however, our approach to focus of methods on human disease genes for prediction of new protein relationships in human through *Drosophila* experimental data is a novel application and resulted in testable predictions.

Acknowledgment

The authors would like to thank Rupali Patwardhan, Sumit Middha, Brian Eads, John Colbourne, Changqing Lin, Junguk Hur, Keval Mehta, Predrag Radivojac, and the Center for Genomics and Bioinformatics at IU for their comments, data support, and computer support.

REFERENCES

[1] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucl. Acids Res.*, vol. 30, no. 1, pp. 52–55, 2002.

[2] D. Kemmer, R. Podowski, D. Arenillas, J. Lim, E. Hodges, P. Roth, E. Sonnhammer, C. Hoog, and W. Wasserman, "Novelfam3000 - uncharacterized human protein domains conserved across model organisms," *BMC Genomics*, vol. 7, no. 1, p. 48, 2006.

[3] G. V. Borzillo and B. Lippa, "The hedgehog signaling pathway as a target for anticancer drug discovery," *Curr Top Med Chem*, vol. 5, no. 2, pp. 147–157, 2005.

[4] D. O. Ricke, S. Wang, R. Cai, and D. Cohen, "Genomic approaches to drug discovery," *Curr Opin Chem Biol*, 2006.

[5] C. Giallourakis and *et al.*, "Disease gene discovery through integrative genomics," *Annu Rev Genomics Hum Genet*, vol. 6, pp. 381–406, 2005.

[6] E. Bier, "*Drosophila*, the golden bug, emerges as a tool for human genetics," *Nat Rev Genet*, vol. 6, no. 1, pp. 9–23, 2005. 1471-0056 (Print) Journal Article Review.

[7] G. M. Rubin and *et al.*, "Comparative genomics of the eukaryotes," *Science*, vol. 287, pp. 2204–2215, Mar. 24 2000.

[8] D. H. Kim and F. M. Ausebel, "Evolutionary perspectives on innate immunity from the study of *caenorhabditis elegans*," *Curr Opin Immunol*, vol. 17, no. 1, pp. 4–10, 2005.

[9] F. Radtke, A. Wilson, and H. R. MacDonald, "Notch signaling in hematopoiesis: lessons from *drosophila*," *Bioessays*, vol. 27, no. 11, pp. 1117–1128, 2005.

[10] A. M. Brumby and H. E. Richardson, "Using *drosophila melanogaster* to map human cancer pathways," *Nat Rev Cancer*, vol. 5, no. 8, pp. 626–39, 2005.

[11] A. A. Cooper and *et al.*, "alpha-Synuclein Blocks ER-Golgi Traffic and Rab1 Rescues Neuron Loss in Parkinson's Models," *Science*, p. 1129462, 2006.

[12] E. Masliah, E. Rockenstein, I. Veinbergs, M. Mallory, M. Hashimoto, A. Takeda, Y. Sagara, A. Sisk, and L. Mucke, "Dopaminergic Loss and Inclusion Body Formation in -Synuclein Mice: Implications for Neurodegenerative Disorders," *Science*, vol. 287, no. 5456, pp. 1265–1269, 2000.

[13] T. Dawson, A. Mandir, and M. Lee, "Animal models of pd: Pieces of the same puzzle?," *Neuron*, vol. 35, pp. 219–222, July 2002.

[14] M. E. Fortini and *et al.*, "A survey of human disease gene counterparts in the *drosophila* genome," *J Cell Biol*, vol. 150, no. 2, pp. F23–30, 2000.

[15] M. E. Fortini and N. M. Bonini, "Modeling human neurodegenerative disease in *drosophila*: On a wing and a prayer," *TIG*, vol. 16, pp. 161–167, April 2000.

[16] N. M. Bonini and M. E. Fortini, "Human neurodegenerative disease modeling using *drosophila*," *Annual Review of Neuroscience*, vol. 26, pp. 627–656, 2003.

[17] A. J. Walhout, R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg, and M. Vidal, "Protein Interaction Mapping in C.elegans Using Proteins Involved in Vulval Development," *Science*, vol. 287, no. 5450, pp. 116–122, 2000.

[18] L. R. Matthews, P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent, and M. Vidal, "Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs"," *Genome Res.*, vol. 11, no. 12, pp. 2120–2126, 2001.

[19] H. Yu, N. M. Luscombe, H. X. Lu, and X. Zhu, "Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs," *Genome Research*, vol. 14, no. 6, pp. 1107–1118, 2004.

[20] S. Bandyopadhyay, R. Sharan, and T. Ideker, "Systematic identification of functional orthologs based on protein network comparison," *Genome Research*, vol. 16, pp. 428–435, 2006.

[21] S. Mika and B. Rost, "Protein-protein interactions more conserved with species than across species," *PLOS Computational Biology*, vol. 2, no. 7, p. e79, 2006.

[22] S. Chien and *et al.*, "Homophila: human disease gene cognates in*drosophila*," *Nucleic Acids Res*, vol. 30, no. 1, pp. 149–51, 2002. 1362-4962 (Electronic) Journal Article.

[23] S. F. Altschul and *et al*, "Basic sequence alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, Oct. 5 1990.

[24] K. P. O'Brien and *et al.*, "Inparanoid: a comprehensive database of eukaryotic orthologs," *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D476–80, 2005. 1362-4962 (Electronic) Journal Article.

[25] M. Remm and *et al.*, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *J Mol Biol*, vol. 314, no. 5, pp. 1041–52, 2001.

[26] T. Hulsen and *et al.*, "Benchmarking ortholog identification methods using functional genomics data," *Genome Biol*, vol. 7, no. 4, p. R31, 2006. 1465-6914 (Electronic) Journal Article.

[27] P. Kemmeren and F. C. P. Holstege, "Integrating functional genomics data," *Bioch Soc Trans*, vol. 31, no. 6, pp. 1484–1487, 2003.

[28] O. Troyanskaya, A. Owens, R. Altman, and D. Botstein, "A bayesian framework for combining heterogeneous data sources for gene function prediction," *PNAS*, vol. 100, no. 14, pp. 8348–8353, 2003.

[29] A. Walhout and *et al.*, "Integrating interactome, phenome, and transcriptome mapping data for the *c. elegans* germline," *Curr. Biol.*, vol. 12, pp. 1959–1964, November 19 2002.

[30] K. C. Gunsalus and *et al.*, "Predictive models of molecular machines involved in *caenorhabditi elegans* early embryogenesis," *Nature Letters*, vol. 436, pp. 861–865, August 11 2005.

[31] P. Kemmeren and *et al.*, "Predicting gene function through systematic analysis and quality assessment of high-throughput data," *Bioinformatics*, vol. 21, no. 8, pp. 1644–1652, 2005.

[32] S. L. Wong and *et al.*, "Combining biological networks to predict genetic interactions," *PNAS*, vol. 101, no. 44, pp. 15682–15687, 2004.

[33] M. Arbeitman and *et al.*, "Gene expression during the life cycle of *drosophila melanogaster*," *Science*, vol. 297, no. 5590, pp. 2270–2275, 2002.

[34] T. Li and K. White, "Tissue-specific gene expression and ecdysone-regulated genomic networks in*drosophila*," *Developmental Cell*, vol. 5, pp. 59–72, July 2003.

[35] M. Parisi and *et al.*, "Paucity of genes on the *drosophila* x chromosome showing male-biased expression," *Science*, vol. 299, no. 5607, pp. 697–700, 2003.

[36] R. Drysdale, "Phenotypic data in flybase," *Briefings in Bioinformatics*, vol. 2, pp. 68–80, 2001.

[37] C. M. Bergman, J. W. Carlson, and S. E. Celniker, "*Drosophila* dnase i footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *d. melanogaster*," *Bioinformatics*, vol. 21, pp. 1747–1749, 2005.

[38] P. W. Lord and *et al.*, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–83, 2003.

[39] G. Grumbling, V. Strelets, and The FlyBase Consortium, "Flybase: anatomical data, images, and queries," *Nucleic Acids Research*, vol. 34, no. Datbase Issue, pp. D484–D488, 2006.

[40] K. R. Brown and I. Jurisica, "Online Predicted Human Interaction Database," *Bioinformatics*, vol. 21, no. 9, pp. 2076–2082, 2005.

[41] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–9, 2000. 1061-4036 (Print) Journal Article.

[42] T. Joachims, *Making large-Scale SVM Learning Practical*, ch. 11. MIT-Press, 1999.

[43] C. W. Roberts and *et al.*, "Highly penetrant, rapid tumorigenesis throught conditional inversion of the tumor supressor gene snf5," *Cancer Cells*, vol. 2, pp. 415–425, 2002.

[44] N. Sevenet and *et al.*, "Spectrum of hsnf5/ini1 somatic mutations in human cancer and genotype-phenotype correlations," *Hum Mol Genet*, vol. 8, pp. 2359–2368, 1999.

[45] L. T. Rieter and E. Bier, "Using *drosophila melanogaster* to uncover human disease gene function and potential drug target proteins," *Expert Opin Ther Targets*, vol. 6, no. 3, pp. 387–399, 2002.

[46] E. Culetto and D. B. Sattelle, "A role for Caenorhabditis elegans in understanding the function and interactions of human disease genes," *Hum. Mol. Genet.*, vol. 9, no. 6, pp. 869–877, 2000.