31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Disease Gene-fishing in Molecular Interaction Networks: a Case Study in Colorectal Cancer

Hui Huang, Jiao Li, and Jake Y. Chen, *Senior Member*, IEEE

***Abstract*—In the post-genome era, disease-relevant gene finding and prioritization have focused on genome-wide association studies and molecular interaction networks, due to their power in characterizing the functions of genes/proteins in genomics and network biology contexts. In this paper, we describe a simple yet generic computational framework based on protein interaction networks to perform and evaluate disease gene-hunting, using colorectal cancer as a case study. We applied statistical measurements including specificity, sensitivity and Positive Predictive Value (PPV) to evaluate the performance of disease gene ranking methods, which we broke down into seed gene selection, protein interaction data quality and coverage, and network-based gene-ranking strategies. We discovered that best results may be obtained by using curated gene sets as seeds, applying protein interaction data set with high data coverage and decent quality, and adopting variants of local degree methods.**

I. Introduction

DISEASE gene finding is a central topic in biomedical research. If the causal genes are found for a disease, health care solutions may be developed to prevent disease occurrence, diagnose disease early, and make tailored treatment plans, e.g., in [1, 2]. For nearly a century, there have been two approaches to discover genes related to a specific disease experimentally: biochemical analysis approach and genetic analysis approach [3]. The first approach attempts to first separate and purify proteins characteristic of disease conditions in model organisms or tissues, and then study the disease-related proteins' biochemical or biophysical altered properties that can be mapped to gene mutations. The second approach normally relies on first studying genetic markers identified in families of diseased populations, and then applying positional cloning techniques and linkage analysis to identify microsatellite markers, chromosomal aberrations, or DNA polymorphisms. However, experimental characterization of proteins or genes involved in diseases is a slow meticulous process. Today, even with advances of genomics technology, one third of all the genes and most of the disease related genes remain functionally uncharacterized [4]. A promising new experimental technique is genome-wide association studies (GWAS), which may help identify candidate single-nucleotide polymorphism (SNP) genetic markers associated with disease risks.

While most computational approaches to disease gene finding rest on statistical association studies or computational sequence analysis, there are surging interests in taking advantage of molecular interaction networks. The concept is to put candidate genes and proteins in specific disease biology contexts defined by molecular interaction networks or biomolecular pathways, with which a researcher can infer functions of uncharacterized genes or proteins. Such disease biomolecular network context may be particularly useful for the study of polygenic diseases such as cancer, in which conventional reductionist approaches are ineffective [1]. In this new approach, disease networks are developed to rank disease relevance of genes/proteins based on properties such as node *degree*s (count of direct PPI connections to a node), *closeness* (path distance of a given node to all other nodes), or *betweenness* (count of geodesic paths that pass through a node). For example, Morrison *et al* used gene expression network and gene ontology information to rank genes similar to Google's PageRank method [5]. *Chen et al* were the first to propose a method that applied disease-specific protein-protein interaction (PPI) networks and modified local node degree measures to prioritize Alzheimer's disease genes [6].

While many network-based disease-gene ranking methods have been developed recently, there has not been a consensus how to evaluate their performances. In this paper, we describe a simple yet generic computational framework to perform and evaluate network-based disease gene-hunting methods. Using colorectal cancer gene finding as a case study, we report how various seed gene selection, PPI data quality, and ranking strategy could affect final gene-finding results. We also defined how specificity, sensitivity, and positive predictive values (PPV) could be used for performance evaluation criteria. We choose colorectal cancer because it is the third leading cause of cancer death in the US and our current knowledge of colorectal cancer genes is limited, making our results to carry special significance. Next, we will describe our methods and report our findings.

Manuscript received on April 23, 2009.

Hui Huang is with the School of Informatics, Indiana University, Indianapolis, IN 46202, USA (e-mail: huanghui@iupui.edu).

Jiao Li is with Dept. of Computer Science and Technology, Tsinghua University Beijing 100084, China. She is now with the Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202 (e-mail: jiao-li04@mails.tsinghua.edu.cn).

Hui Huang and Jiao Li contribute equally to this paper.

Jake. Y. Chen is with the School of Informatics, Indiana University, Indianapolis, IN 46202, USA and the Indiana Center for Systems Biology and Personalized Medicine, Indianapolis, IN 46202 (e-mail: jakechen@iupui.edu).

II. METHODS

In Figure 1, we show an overview of the computational framework used in this study. It consists of two components: (1) *Disease Gene Identification,* in which we expand seed genes to disease-specific protein sub-network and subsequently generate a ranked list of disease-relevant genes; (2) *Disease Gene Assessment*, in which we quantitatively assess disease genes using statistical measurements including sensitivity, specificity and PPV. The relationships between the two components is the following: First, disease gene identification will be performed using a fixed set of gene-seeding, PPI sub-network construction, and disease gene ranking strategies; then, we evaluate how sensitivity, specificity, and PPV are affected by varying choices of seed genes, PPI networks, and ranking strategy.

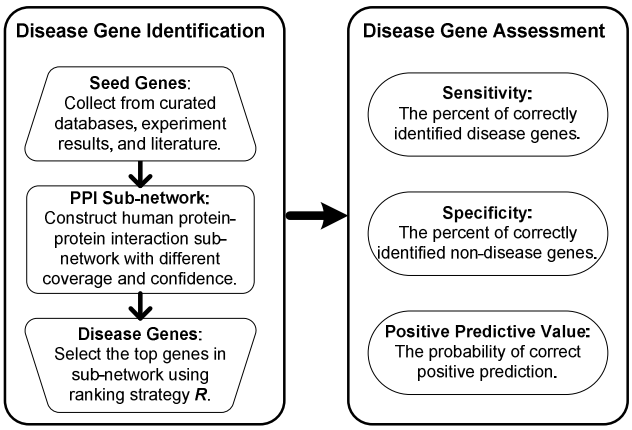

Figure 1. Computational Framework for Disease Gene Identification and Assessment.

A. Seed Gene Selection

We consider three sets of colorectal cancer-related genes collected from different resources as *seeds*, which are: (1) the **CORE1** set, derived from human curated databases by querying the OMIM [7] and KEGG [8] database for "colorectal cancer" and manually curating the set of genes/proteins; (2) the **CORE2** set, derived from high-throughput microarray data in the ONCOMINE [9] database by keeping only differentially expressed genes with p-value<0.05 performed for colorectal cancer samples against controls; (3) the **CORE3** set, derived from the Comparative Toxicogenmics Database (CTD [10]) by searching for colorectal cancer genes associated with >2 chemicals in the database.

B. Protein Interaction Sub-network Construction

We expand seeds, using PPIs recorded in the Human Annotated and Predicted Protein Interactions (HAPPI) database [11] to construct colorectal cancer-specific PPI sub-network. A unique feature of the HAPPI database is that the quality of PPIs comes with estimated confidence scores (a real value between 0 and 1) and star grades (an integer between 1 and 5). The higher the confidence score or the star grade number, the more likely the PPI is attributable to physical PPI events. In this study, we use PPI star grade to control disease-specific sub-network quality and coverage. We refer to the disease-specific PPI sub-network constructed from HAPPI quality star grade *n* and above as PPI-*n*. For example, PP1-3 includes all PPIs from HAPPI with quality star grade of 3, 4, and 5.

C. Disease Gene Ranking Strategy

We treat the disease gene ranking problem as a problem to calculate a weight for each protein in the disease-specific PPI sub-network. There are three ranking strategies being considered in this study: (1) *Global degree* strategy, in which we use the protein's node degree in the global PPI-n network as the weight; (2) *Local degree* strategy, in which we use the protein's node degree in the local (colorectal-specific) PPI-n network as the weight; and (3) *Edge-weighted Promiscuous Hub subtraction* (EPHS) strategy developed in [6], which is a variant of *local degree* strategy adapted by penalize the impact of low-quality promiscuous protein hubs on ranks defined by the following formula:

$$r_p = k * \ln(\sum_{q \in NET} conf(p,q)) - \ln(\sum_{q \in NET} N(p,q)) \quad (1)$$

Here, p and q are indices for proteins in the constructed network *NET*. k is an empirical constant. *conf(p, q)* refers to confidence score in HAPPI Database. *N(p, q)* holds the value of 1 if the protein p interacts with q. The r_p score is the weight calculated to rank each protein in the network.

In addition, we use *TOP_M* to refer to the *M* highest ranked disease-relevant proteins/genes given by a specific disease gene ranking strategy.

D. Disease Gene Assessment

To evaluate the disease-related gene list, the sets of Gold Standard Positive (GSP) and Gold Standard Negative (GSN) are constructed as illustrated in Figure 2.

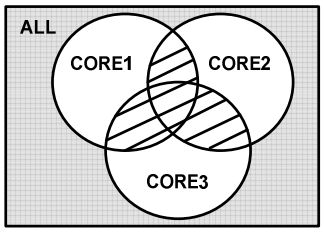

Figure 2. Gold standard construction for disease gene assessment. *As* shown in the striped area, *GSP= (CORE1∩CORE2)∪(CORE1∩CORE3)∪(CORE2∩CORE3).* As shown in the gray area, *GSN=ALL-(CORE1∪ CORE2∪ CORE3).* Note that ***ALL*** refers to all HAPPI human proteins.

The following measurements are calculated to evaluate the performance of each disease gene identification method: (1) **Sensitivity**, calculated as the percent of correctly identified disease genes $|TOP\_M \cap GSP| / |GSP|$; (2) **Specificity**, calculated as the percent of correctly identified non disease genes $|GSN-(TOP\_M-GSP)| / |GSN|$; (3) **Positive Predictive Value (PPV)**, calculated as the probability of correct positive predictions $|TOP\_M \cap GSP| / |TOP\_M|$.

III. RESULTS

We developed three colorectal cancer seeds: CORE1, consisting of 148 proteins; CORE2 containing 42 proteins, a subset of 7410 proteins with p-value <0.05 consisting of 81 samples from Oncomine data; and CORE3, consisting of 721 proteins. With three choices of seeds gene selections (CORE1, CORE2, and CORE3), four PPI qualities (PPI-3, PPI-4, PPI-5, PPI-1), three ranking strategies (EPHS, Local Degree, Global Degree), we tested different combinations to conduct the disease gene findings and assessment for colorectal cancer.

A. *Effect of Various Seed Gene Selection Methods*

In Figure 3, we show how seed selections affect the ranking results. In this experiment, we used PPI-3 as PPI network data source and the EPHS disease protein ranking method. The *ranking index* on the x-axis refers to a number, *TOP_M*, used to indicate the number of all rank-ordered proteins in a given expanded protein set consisting of both seed proteins and PPI-expanded disease sub-network. PPV for the initial top-10 or top-20 proteins for both core-1 and core-2 seeded strategies were at 0.7-0.8 range, suggesting high predictive power of top-ranked proteins for disease-relevance. As ranking index increases, PPV decrease for all core seeded strategies. However, the performance for core-1 is superior to both core-2 and core-3. This is perhaps due to the highly curated nature of core-1 seeds as compared with possible noises introduced by Omics data for core-2 and text mining data for core-3. Core-3 shows an overall poorer PPV performance, particular within top-20 compared with core-1 and core-2. Beyond ranking index of 250, all core seeded strategies converged to low PPV within 0.15. Therefore, the relatively high predictive powers of all disease gene rankings seem to be restricted to the top 50.

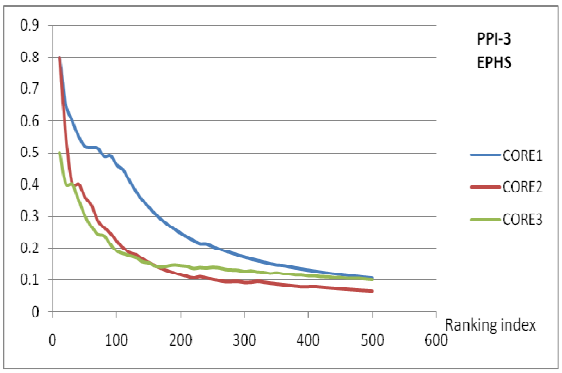

Figure 3. PPV performance using different seed choices.

B. *Effect of Various PPI Data Quality and Coverage*

In Figure 4, we show how PPI data used for network expansion affect the ranking results. In this experiment, we compared results using PPI-1, PPI-3, PPI-4, and PPI-5, using core1 as seeds and the EPHS ranking method. All PPI-n except for PPI-5 showed a similar trend of decreasing PPV. Again, the relatively high predictive powers (PPV>0.5) seem to be achieved at the top 50, except for PPI-5, then continue to decrease to very low levels (PPV<0.15) beyond a ranking index>400. It's counter-intuitive that PPI-5's performance, being the poorest, has a rising phase from ranking index between 10 and 50 before decreasingly significantly. This may be primarily due to the poor coverage of true colorectal cancer proteins in current physical PPI data sets representative of PPI-5 until enough proteins are covered in the top 40 or 50 set. Therefore, data coverage seems quite important in gene ranking performance overall. Also, at least in the top 10 case, the fact that PPI-3 has the best PPV of 0.8 over PPI-1 that has much higher data coverage suggest that PPI data quality is also important to discover disease genes in the most highly ranked protein set. Therefore, balanced data coverage and quality are essential for disease gene finding from such networks.

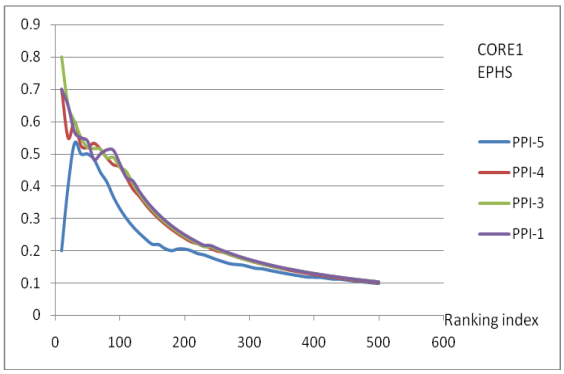

Figure 4. PPV performance using different PPI-n networks.

C. *Effect of Various Disease Gene Ranking Methods*

In Figure 5, we show how the choices of different ranking methods affect the ranking results. The results are performed by fixing seed protein to core1 and using PPI-5 for the expansion network. EPHS and Local Degree methods performed equally, while global degree performed extremely poor—although by sharing similar performance trend of the top-performing methods. The trend for all methods shows two phases: a PPV rising phase from top 10 to top 60-80; and a PPV decreasing phase from top 80 onwards. The separation of two phases is likely due to balanced PPI data coverage and quality as explained earlier.

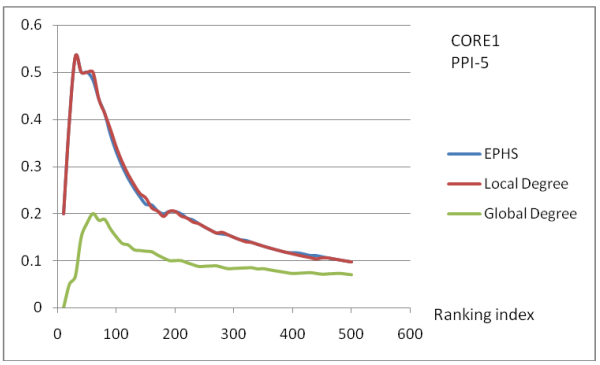

Figure 5. PPV performance using different disease gene ranking methods.

D. *Sensitivity and Specificity Comparisons of Top Disease Gene Ranking Methods.*

We further compared the sensitivity and specificity performances for the best two disease gene ranking methods, EPHS and Local Degree.

Figure 6 shows a comparison of their specificity (on the y-axis) performance distributed over different ranking index ranges (on the x-axis). The specificity performances of both methods are quite good overall, even at top 100 range (specificity>0.9). The EPHS ranking method is slightly better (more specific) than Local Degree ranking method. This is primarily because local degree method cannot distinguish nodes with the same number of node degrees, particularly when the node degree drops to small numbers such as 2 or 3 in the high ranking index region.

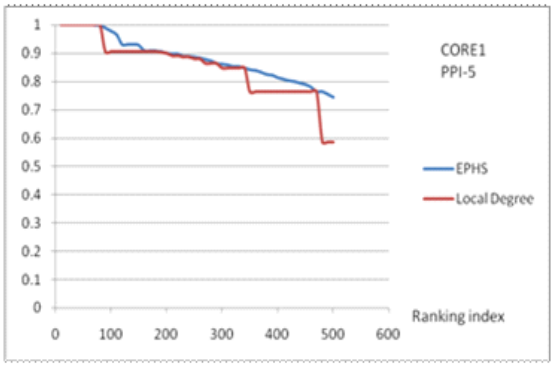

Figure 6. A comparison of specificity performance between the EPHS and Local Degree ranking methods.

Figure 7 shows a comparison of their sensitivity (on the y-axis) performance distributed over different ranking index ranges (on the x-axis). The sensitivity performances of both methods are decent overall after ranking index range of top 80 (sensitivity>0.75). The local degree ranking method is slightly better (more sensitive) than EPHS ranking method. The reason that local degree method performed better than EPHS ranking method is that there are many tied genes in local degree method due to their sharing the same node degrees. However, since most rankings should be performed in the low ranking index region, this slight loss of sensitivity for EPHS method can be ignored.

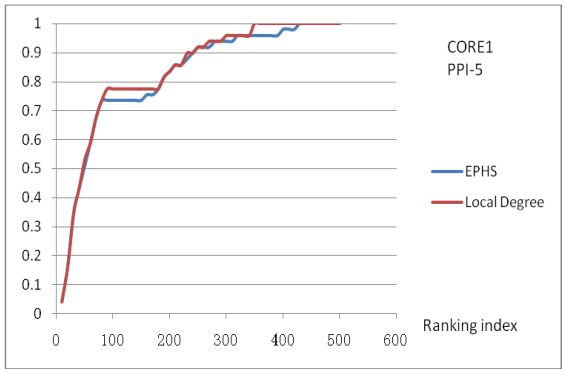

Figure 7. A comparison of sensitivity performance between the EPHS and Local Degree ranking methods.

IV. CONCLUSION

In this paper, we performed disease gene finding from protein-protein interaction networks specific to colorectal cancer. We examined the effects of different seeds, different PPI data quality, and different disease gene ranking methods on the final performance of the task. While all of these parameters may impact the final performance, our results show that (1) the initial quality of seeds should be based on prior curated knowledge as much possible, with Omics results being the next choice and text mining results being the last resort; (2) disease gene ranking should be performed using PPI data with reasonable quality but as high data coverage as possible; (3) the ranking algorithm that takes advantage of local network parameters should be chosen over those using global network parameters. There are several limitations to our current research approach. For example, the gold standard positive set of genes used for evaluation had to be built by considering seed gene sets used for research studies due to convenience of computation. The observations made for this framework should be carefully validated in other disease contexts before they are generalized.

REFERENCES

[1] A.C. Ahn, M. Tewari, C.S. Poon, and R.S. Phillips, "The clinical applications of a systems approach," *PLoS Med*, vol. 3, (no. 7), pp. e209, Jul 2006.

[2] A.N. Smith, J. Skaug, K.A. Choate, A. Nayir, A. Bakkaloglu, S. Ozen, S.A. Hulton, S.A. Sanjad, E.A. Al-Sabban, R.P. Lifton, S.W. Scherer, and F.E. Karet, "Mutations in ATP6N1B, encoding a new kidney vacuolar proton pump 116-kD subunit, cause recessive distal renal tubular acidosis with preserved hearing," *Nat Genet*, vol. 26, (no. 1), pp. 71-5, Sep 2000.

[3] C. Giallourakis, C. Henson, M. Reich, X. Xie, and V.K. Mootha, "Disease gene discovery through integrative genomics," *Annu Rev Genomics Hum Genet*, vol. 6, pp. 381-406, 2005.

[4] J. Chen, H. Xu, B.J. Aronow, and A.G. Jegga, "Improved human disease candidate gene prioritization using mouse phenotype," *BMC Bioinformatics*, vol. 8, pp. 392, 2007.

[5] J.L. Morrison, R. Breitling, D.J. Higham, and D.R. Gilbert, "GeneRank: using search engine technology for the analysis of microarray experiments," *BMC Bioinformatics*, vol. 6, pp. 233, 2005.

[6] J.Y. Chen, C. Shen, and A.Y. Sivachenko, "Mining Alzheimer disease relevant proteins from integrated protein interactome data," *Pac Symp Biocomput*, pp. 367-78, 2006.

[7] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, (no. Database issue), pp. D514-517, 2005.

[8] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucleic Acids Res*, vol. 36, (no. Database issue), pp. D480-4, Jan 2008.

[9] D.R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B.B. Briggs, T.R. Barrette, M.J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A.M. Chinnaiyan, "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia*, vol. 9, (no. 2), pp. 166-80, Feb 2007.

[10] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest, and J.L. Boyer, "The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies," *J Exp Zoolog A Comp Exp Biol*, vol. 305, (no. 9), pp. 689-92, 2006.

[11] J.Y. Chen, S. Mamidipalli, and T. Huang, "HAPPI: an Online Database of Comprehensive Human Annotated and Predicted Protein Interactions," *BMC Genomics (Accepted)*, 2009.