GENOTYPE COHERENCE DETECTION AND CLASSIFICATION

Bhargav Challagulla[1], Tusar k Mishra[2], G. Gagan Teja[3], D. Lakshman Kumar[4]

{challagullabharghav[1], tusar.k.mishra[2],gaganpranayteja1998[3],lakshmankumardhanukonda999[4]}@gmail.com

Department of Computer Science and Engineering, ANITS, Visakhapatnam

ABSTRACT

A fundamental problem in human health is the inference of disease-causing genes, with important applications to diagnosis and treatment. Previous work in this direction relied on the knowledge of multiple loci associated with the disease, or causal genes for similar diseases, which limited its applicability. Here we present a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under a condition of interest. The latter is used to derive a set of disease-related genes which is assumed to be in close proximity in the network to the causal genes. Our method applies a set-cover-like heuristic to identify a small set of genes that best "cover" the disease-related genes. We perform comprehensive simulations to validate our method and test its robustness to noise. In addition, we validate our method on real gene expression data and on gene specific knockouts. Finally, we apply it to suggest possible genes that are involved in myasthenia gravis.

I. INTRODUCTION

The diseased gene identifying based on a network of interconnected proteins. Identifying disease genes from the human genome is an important and fundamental problem in biomedical research. Despite many publications of machine learning methods applied to discover new disease genes, it still remains a challenge because of the pleiotropy of genes, the limited number of confirmed disease genes among the whole genome and the genetic heterogeneity of diseases. The past two decades have witnessed an explosion in the identification, largely by positional cloning, of genes associated with Mendelian diseases. The roughly 1,200 genes that have been characterized have clarified our understanding of the molecular basis of human genetic disease. The principles derived from these successes should be applied now to strategies aimed at finding the considerably more elusive genes that underlie complex disease phenotypes. Typically, with these features available, a method for prioritizing disease genes computes a score quantifying the association between a gene and a disease, and then uses the computed scores to rank the candidates and select plausible susceptibility genes. However, various factors, such as the pleiotropy of genes, the interactions among genes, the genetic heterogeneity of diseases, and the ambiguous boundary between different diseases, as well as the incompleteness and false-positive data sources, might prevent the direct inference of single gene-disease association. We show that the correlation between phenotype similarities and gene closeness, defined by the concordance score, is a strong and robust predictor of disease genes. With the use of this score, we propose a new method, CIPHER to prioritize candidate genes and to explore gene cooperative behavior in human disease.

If we know something about the relationships between the genes, we can assess whether some genes (which may reside in different loci) functionally interact with each other, indicating a joint basis for the disease ethology. There are various repositories of information on pathway relationships. To consolidate this information, we developed a functional human gene network that integrates information on genes and the functional relationships between genes.

II. RELATED WORK

A method for prediction of disease-relevant human genes from the phenotypic appearance of a query disease is presented. Diseases of known genetic origin are clustered according to their phenotypic similarity. Each cluster entry consists of a disease and its underlying disease gene.

Potential disease genes from the human genome are scored by their functional similarity to known disease genes in these clusters, which are phenotypically similar to the query disease. [8]

One of the major promises is that these advances will lead to personalized medicine, in which preventive and therapeutic interventions for complex diseases are tailored to individuals based on their genetic profiles. Personalized medicine already exists for monogenetic disorders such as Huntington disease, phenylketonuria (PKU) and hereditary forms of cancer, in which genetic testing is the basis for informing individuals about their future health status and for deciding upon specific, often radical interventions such as lifetime dietary restrictions and preventive surgery. Yet, the ethology of complex diseases is essentially different from that of monogenic diseases, and hence translating the new emerging genomic knowledge into public health and medical care is one of the major challenges for the next decades. [5] A network of disorders and disease genes linked by known disorder-gene associations offers a platform to explore in a single graph-theoretic framework all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. Genes associated with similar disorders show both higher likelihood of physical interactions between their products and higher expression profiling similarity for their transcripts, supporting the existence of distinct disease-specific functional modules. We find those essential human genes are likely to encode hub proteins and are expressed widely in most tissues. This suggests that disease genes also would play a central role in human interaction [10] Here we present a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under a condition of interest. The latter is used to derive a set of disease-related genes which is assumed to be in close proximity in the network to the causal genes. Our method applies a set-cover-like heuristic to identify a small set of genes that best “cover” the disease-related genes. We perform comprehensive simulations to validate our method and test its robustness to noise. In addition, we validate our method on real gene expression data and on gene-specific knockouts [7] Here the gene-gene relations are extracted by taking a hybrid approach which is a combination of syntactic analysis and co-occurrence-based approaches. Specifically, we perform the syntactic parsing on the text and then, Keywords - gene ranking; text-mining; relation extraction; disease-related genes; microarray data analysis. within each clause of the parsed sentence, the co-occurred gene names are considered to be mutually related. Both the gene network derived from the gene-gene relations obtained in the above way and the gene expression scores is given as the inputs to the Gene Rank algorithm. [6]

III. PROPOSED WORK

In this section, we will see the proposed work and process flow to do Genotype Coherence Detection. The System Model has been depicted in Fig. 1.

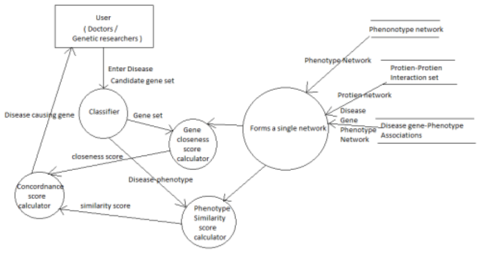

Fig.1: Gene Coherence Detection

Data sets:

The project requires datasets containing data on phenotypes, corresponding gene-phenotype and protein-protein interactions (PPI). The phenotypes and gene-phenotype data sets are obtained from OMIM database present in www.omim.org. The OMIM database is an authorized database consisting of 5206 known phenotypes and 7356 known genes and is being updated through medical research.

The PPI dataset is obtained from HPRD database from www.hprd.org. The HPRD database consists of 34,364 manually curated PPIs between 8919 human proteins.

Proposed Training Method

Using the above data sets a new merged dataset is obtained and is used to train the system. In the training, phenotype similarity score, gene closeness score and using the two-concordance score is calculated which is used to rank the system. The phenotype similarity score is calculated between the query phenotype and known phenotypes. It is determined by the cosine of their feature vector angle.

Using the score similar phenotypes and their causative genes are obtained. The distance between these genes and candidate genes are compared to calculate gene closeness score and thus concordance score to rank the candidate genes.

Phenotype Network, protein-protein interaction set, and Disease Gene-Phenotype Association are the data sets used to form the network. In this network, we develop two score gene closeness score calculator and phenotype similarity score calculator.

By combining these two-score calculators we will get the concordance to score calculator. This score is used by the doctor to find the disease gene. The doctor uses a classifier to get the gene closeness score calculator and phenotype similarity score calculator.

Data columns	Range Index
MIM Number	7536
MIM Entry Type	7536
Entrez Gene ID	7536
Approved Gene Symbol	6652
Prefix	7513
Preferred Title; symbol	7536
Alternative Title(s); symbol (s)	5706
Included Title(s); symbols	791
Sort	7536
Confidence	7536
Mapping Method	7536
Mouse Gene Symbol	2416
Phenotype	7536
Chromosome	7536
Genomic Position Start	7536
Genomic Position End	7536
Cyto Location	7536
Computed Cyto Location	6322
Gene Symbols	7536
Approved Symbol Entrez G ene ID	6652
Mouse Gene Symbol/ID	6045

Testing the system

To test the trained system's ability in uncovering known disease genes and predicting novel susceptibility candidates, we present a case study for breast cancer, which is the most commonly occurring cancer among women and accounts for 22% of all female cancers. Known susceptibility genes, including BRCA1 (Miki et al, 1994) and BRCA2 (Wooster et al, 1995), can only explain less than 5% of the total breast cancer incidence and less than 25% of the familial risk, suggesting that many susceptibility genes remain to be discovered.

Analyzing the system

In the analysis phase, we analyze the ranking of the candidate genes showing their contribution level with the known research data.

Accuracy

Finally, we put together all the data and divided them into 2 files namely test data, train data files based on their similarity. Since null values should not be considered we only considered the data.

With the help of the lstm, we obtain the Accuracy, concordance score ($A_{pp'}$), similarity score (Score $_{\text{combined}}$) and Recall using the formulas below

$$S_{pp'} = C_p + \sum_{g \in G(p)} \sum_{g' \in G(p')} \beta_{pg} e^{-L^2_{gg'}}$$

$$\text{score}_{\text{combined}} = \text{score}_{\text{original}} \times (1 - \alpha) + \text{score}_{\text{noise}} \times \alpha$$

IV. EXPERIMENTAL RESULTS & ANALYSIS

In Figure 2, a snapshot of the plots between genomic-start and genomic-end has been presented. Uniformity is better observed in the same. Similarly, other parameters also show fine-tuned progress of the iterations thereby. This indicates smooth learning of the system with finite convergence. A comparative analysis of the proposed scheme has been shown in Table I. The time of computation and rate of accuracy of the proposed scheme has been found to be comparatively better among all. Although the rate of accuracy is not so high, this can be considered as a good figure when we talk about health concerned risk.

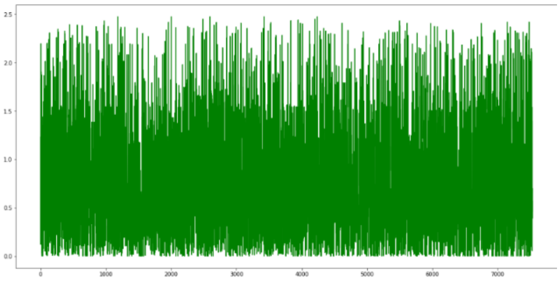

Fig.2: sample output

Table I

Tabulated Results

Accuracy (%)	concordance	Score combined
62.0	0.69	60.0

V. CONCLUSION and FUTURE WORK

The availability of an annotated dataset has reduced the difficulty of PROTIN-PROTIN interaction to some extent. As we proposed in our paper, Gene Detection can be extended for as long as it encounters new data which are not specified in datasets. But even then, there are possibilities for the existence of some drawbacks as one gene may not always give appropriate sentiment for the whole sentence. Though the accuracy obtained is74.74%, the above-mentioned flaw in our approach can be reduced by the use of improved heuristics, which can be an extension to the current approach.

Acknowledgments: The authors would like to thank DR. Tusar Kanti Mishra for his constant support and guidance in the process of working on this paper.

REFERENCES

[1] Briefings in Functional Genomics, Volume 10, Issue 5, 1 September 2011, pages 1-11

[2] B. Hu, P. Shuai, Z. Shan and C. Pang, "Define a function on gene order of DNA microarray data and use it to identify genes associated with Alzheimer's disease," 2011 IEEE International Conference on Granular Computing, Kaohsiung, 2011, pp. 836-839.

[3] From syndrome families to functional Genomics by Brunner, Han G. and van Driel, Marc A
doi =10.1038/nrg1383, number = {7}, pages = {545-551}

[4] author = Esposito, Daniel Christopher and Cursons, Joseph and Davis, Melissa Jane, title = Inferring edge function in protein-protein interaction networks, location-id = 321984, year = 2018, doi = 10.1101/321984, publisher = Cold Spring Harbor Laboratory

[5] Genome-based prediction of common diseases: advances and prospects A. Cecile J.W. Janssens, and Cornelia M. van Duijn Human Molecular Genetics, Volume 17, Issue R2, 15 October 2008, Pages R166–R173

[6] H. Lee, M. Shin and M. Hong, "A gene ranking method using text-mining for the identification of disease-related genes," *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Hong Kong, 2010, pp. 493-498

[7] author = Karni, Shaul and Soreq, Hermona and Sharan, Roded, title = A Network-Based Method for Predicting Disease-Causing Genes, journal = Journal of Computational Biology, volume = 16, number = 2 , pages = 181-189, year = 2009

[8] Freudenberg, J & Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics (Oxford, England). 18 Suppl 2. S110-5. 10.1093/bioinformatics/18.suppl_2S110.

[9] Classifying Gene Expression Data of CancerUsing Classifier Ensemble with Mutually
Exclusive Features by SUNG-BAE CHO, MEMBER, IEEE, AND JUNGWON RYU

[10] The human disease network by Kwang Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-Laszlo Baraba´ si Lopes, Julio & Figuerêdo Domingues, Bernardo. (2007). PharmGKB Network - Integrating Diseaseome, Pharmacome,andTargetome. 10.13140/RG.2.1.4634.1847.