2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

# Disease-Gene Association Using a Genetic Algorithm

Koosha Tahmasebipour
Department of Computer Science
Brock University
St. Catharines, ON, Canada

Sheridan Houghten, Member, *IEEE*
Department of Computer Science
Brock University
St. Catharines, ON, Canada

***Abstract*—Understanding the relationship between genetic diseases and the genes associated with them is an important problem regarding human health. The vast amount of data created from a large number of high-throughput experiments performed in the last few years has resulted in an unprecedented growth in computational methods to tackle the disease gene association problem. Nowadays, it is clear that many of genetic diseases are not consequence of defects in a single gene. Instead, the disease phenotype is a reflection of various genetic components interacting in a complex network. In fact, most of genetic diseases occur as a result of various genes working in sync with each other in a single or several biological module(s). Using a genetic algorithm, we have devised a computational approach for disease-gene association. As a proof of concept, we apply this method to the problem of identifying genes involved in breast cancer.**

***Keywords*—Disease-Gene Association, Community Identification, Genetic Algorithms, Complex Networks**

## I. Introduction

Associating genes with phenotypes has been one of the most demanding research subjects in bioinformatics. One of the major parts of gene-phenotype association research is to associate genes with a specific group of phenotypes, i.e. the genetic diseases. The specific research area dealing with this problem is called the *disease-gene association problem*, or *identification of disease genes*, or *disease gene prediction*, i.e. identifying the genes which are in any way involved in the existence of a given genetic disease or disorder. Associating genes with genetic diseases and disorders is crucial to identify the genetic basis of human diseases. Procuring more information and knowledge about the genes behind a given disorder can lead to more effective diagnosis, improved prognosis, and development of improved therapeutic strategies to treat the genetic diseases and disorders.

Bioinformatics tries to tackle the disease-gene association problem via developing *computational disease-gene prediction methods*. These methods or tools apply the discoveries and findings from years of research in related areas by examining useful information from the literature to associate genes with diseases [18].

Disease-gene prediction methods choose potential disease genes from a set of candidate genes usually determined by different experimental and other computational methods including Genomewide Association Studies (GWAS) and Linkage Analysis. In Genomewide Association Studies many and various SNPs (Single Nucleotide Polymorphism) are statistically investigated for association with genetic disease in hundreds or thousands of individuals [8]. Linkage Analysis typically associates certain chromosomal loci (linkage interval) with a particular disease phenotype [13].

Different types of evidence in the literature are applied by disease gene prediction methods to identify the genes associated with diseases. *Text-mining of biomedical literature* discovers gene-disease association by using natural language processing techniques over huge quantities of related written knowledge. These uses data resources such as PubMed [1] and OMIM [14] which contain a multitude of biomedical abstracts and studies (e.g. [2], [5], [22], [30]). *Functional annotation* identifies unknown genes involved in disease by examining the same pathways or functionalities as other known disease genes, as these are likely to be connected to the diseases under study (e.g. Gene Ontology [3], KEGG[19], and Human Phenotype Ontology [34]). *Phenotype relationships* uses the similarities among various diseases and phenotypes to predict potential disease genes (e.g. Mimminer[38], [9]). *Protein-protein interaction networks* (PPI) represent physical interactions between proteins; this is one of the strongest and most frequently used types of evidence for disease gene prediction [27] (see also [13], [44], [29], [17], [43], [20], [39], [7], [23]). *Gene expression information* is one of the least biased pieces of evidence provided by high-throughput experiments [31]. *Gene regulatory networks* indicate how genes' expression levels are controlled by regulators, i.e. other genes, proteins and molecules in the cell[32]. *Next-generation sequencing data* and inherent gene or protein properties (e.g. length, phylogenetic analysis of genes, conversation degree etc.) have also been used as clues to associate genes with disease [31]. For further information on the various evidence, tools, and computational approaches so far developed and used to associate disease with genes, see [13], [31], [18], [37], [6] and [26].

One of the most important pieces of information in the literature which can aid computational disease gene prediction methods to associate diseases with genes is the already established set of *known disease genes*, i.e. those genes that have already been observed to be involved in a given disease. Knowledge about known disease genes can be crucial for disease gene prediction methods to identify new novel disease genes, mainly because genes causing the same or similar diseases tend to lie close to one another in the networks in which they interact. For example, genes related

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.38

to the same disease are known to have protein products that physically interact. Also, previous studies have shown that mutations in multiple proteins that form a protein complex may lead to the same disease phenotype [39]. This is called the principle of *guilt by association*. This is also called the *modular nature of genetic disease*, i.e. that various genes involved in the same phenotypes work together in a single biological module [28].

Due to the complexity of the interactions among the genes and the remarkable amount of discovered interactions among genes and their products (i.e. proteins), disease-gene association is not a matter of establishing simple connections between known disease genes and candidate genes using a set of straight-forward evidence [4]. It is crucial to note that complex phenotypes and diseases are far beyond the sum of independent effects of the individual genes. In many of the cases,"the impact of a specific genetic abnormality is not restricted to the activity of the gene product that carries it, but can spread along the links of the network, and alter the activity of gene products that carry no defects" [4]. Furthermore, "the phenotypic impact of a defect is not necessarily determined only by the known function of the mutated gene, but also by the functions of components with which the gene and its products interact and of their interactions partners, i.e., by its network context" [4]. From these key assumptions, it is concluded that genetic disorders could be reflection of various components (genes, proteins, etc.) interacting in a *complex network*.

Classic forms of disease gene prediction approaches, in the absence of considering the notion of complex networks, were not able to process the myriads of interactions among the genes and proteins. It has been shown that it is necessary to process and observe the interactions among the complex system's components in order to find facts and draw conclusions about these systems. As stated in [25], "Many of the ideas are disguised behind the interactions among the components, and not inside the component". This corresponds to what Aristotle said in 335-323 BCE, "The whole is more than the sum of its parts" [12].

## II. Community Identification and Disease Gene Association

Complex networks have a tendency to form different communities, usually each pursuing different functionality working in sync with each other [35]. Some examples include different regions of the brain, web pages about the same or similar topics on the World Wide Web, or proteins interacting with one another to perform a specific function. The network inside the communities is relatively dense compared to the parts of the network residing outside the communities. Descriptively, communities are sub-graphs with a higher level of internal interconnection than other sub-graphs of the network [21]. Based on how great a difference there is between the number of internal and external connections of a given community, the community can be considered a *weak* community or a *strong* one.

Different phenotypes including diseases are fairly likely to be assigned to a group of nodes (i.e. genes, RNA molecules, proteins and other distinct cellular components) working closely with one another as a community of the gigantic complex network of the *human interactome*, a network which is believed to contain some 25000 protein-encoding genes and a still higher number of unknown and undefined proteins and their interactions [4].

In 2002, the field of *complex network community identification* was first studied by Newman and Girvan [11]. The community identification problem contains within it two subproblems: (1) given a network, identify all communities working in sync with each other in the network, and (2) given a node or a group of nodes, identify the community or communities to which these nodes belong. Both subproblems have been shown to be NP-complete. The latter problem is little studied in comparison to the former one [40].

From the modern viewpoint of genetics (referred to as the *molecular revolution* [25]), disease-gene association can be redefined as: given a number of genes known to be involved in a disease, computationally examine the current available human interactome in the vicinity of the known disease genes to detect the community or communities in which these genes closely interact with one another. Discovering such communities sheds light on genes closely working with known disease genes and steers us towards predicting new disease genes. Therefore, computational disease genes prediction matches the second subproblem of community identification mentioned above.

## III. Methodology

In this study, we propose a computational disease-gene association method using a genetic algorithm. Genetic algorithms have been proven to be effective in finding near-optimal solutions for many NP problems [25]. The main idea of GA is to mimic evolution and the Darwinian natural selection process to find a satisfactory solution for a given computational problem. The first serious effort to apply GAs to practical problems was work done by John Holland and his students at the University of Michigan in the 1970s [25]. Generally, a genetic algorithm tries to evolve a population of *candidate solutions* for the problem at hand. These candidate solutions are also called *chromosomes* or *individuals*. Each candidate solution is assigned a fitness value measured by a fitness function. The fitness value measures how good the solution is for the given problem. GA individuals undergo cross-over which is applied on two or more individuals (chromosomes) for offspring reproduction, and mutation which mimics possible mutations when new individuals are reproduced. A general recipe for a genetic algorithm can be described as having the following steps: 1) Generate an initial population of candidate solutions. These solutions are usually created randomly. 2) Measure the fitness of each individual in the current population. 3) If there is a solution in the population with the desired fitness value, or if the number of generations has reached the predefined maximum number of generations, then the program stops 4) Randomly select a number of the individuals and select a number of them to be the parents reproducing individuals for the next generation. 5) Generate the population for the next generation. 6) Go to 2.

We chose to use a genetic algorithm because we wished to produce a population of various communities. Note that our ranking for a given gene actually depends on all of the communities in the population over generations. See Sections III-E and VIII for further information on this point.

Prior to clarifying the technical details of our GA-based disease gene prediction method, we first review its general properties.

Using a genetic algorithm, our computational method tries to evolve a community containing the set of potential disease genes likely to be involved in a given genetic disease. Having a set of known disease genes already believed to be involved in a disease, we first obtain a protein-protein interaction network containing all the selected known disease genes, as described in Section IV-A. All the other genes inside the procured PPI network are then considered as candidate disease genes as they lie in the vicinity of the known disease genes in the human interactome. Our method attempts to find a community of potential disease genes strongly working with one another and with the genes in the known disease genes set.

In our GA-based approach, each individual within a population is a candidate community of genes that may be involved in a given disease. The initial population consists of individuals that are simply randomly-created communities. At each generation, the GA tries to evolve individuals (communities) that have a relatively high degree of collaboration between the known disease genes and the other genes within the individual (community), and that also that have a high degree of collaboration among all the genes inside the individual (community). The more collaborative the genes are in the evolved individual (community) with the known disease genes and with one another, the more likely it is that the individual (community) contains genes potentially involved in the disease under study. Intuitively, in our method, the GA tries to find subnetworks of the interaction network that are as modular and collaborative as possible, where the subnetworks also contain the known disease genes.

### A. Fitness Evaluation

As previously stated, our GA tries to increase the degree of collaboration (i.e. the *modularity*) of the nodes (i.e. genes) inside the evolving communities. This is mainly because a community is recognized as a subgraph that is denser inside in comparison to the areas of the network around the subgraph. Therefore, as we can see in the community identification problem literature, measuring the modularity of a given community revolves around comparing the density of the area residing inside the community to the areas around the community. To see a general review of the various metrics used to measure the quality of a given community, see [40].

Hence, reasonably enough, the fitness value should assess the modularity degree of the candidate communities (individuals in a given generation) to increase the likelihood of reproduction for the more modular individuals across the generations.

For the experiments we have performed so far, we have used the simple modularity function $Q$ proposed by Luo et al. [24], also known as subgraph modularity. This modularity function simply divides the number of the edges inside the community $C$ by the number of all the edges involved with the community:

$$Q(C) = \sum_{i \in C} \frac{K_i^{in}(C)}{K_i(C) = K_i^{in}(C) + K_i^{out}(C)} \tag{1}$$

where

$$K_i^{in}(C) = \sum_{j \in C} E(i,j) \tag{2}$$

is the number of edges connecting $i$ to other nodes in $C$ and

$$K_i^{out}(C) = \sum_{j \notin C} E(i,j) \tag{3}$$

is the number of edges connecting $i$ to the nodes out of $C$.

### B. Individuals (Chromosomes)

As mentioned in the previous sections, individuals (candidate solutions) are induced subnetworks that are potential disease gene communities. Individuals are represented using binary arrays of length $N$, where $N$ is the size of the protein-protein interaction network (i.e. the number of all the genes in the network). If the $i$th element of the array is 1, then the $i$th gene of the network is considered to be in the candidate community; otherwise it is outside of the candidate community.

### C. Linkage with known disease genes (Guilt by Association)

Like most of the computational disease gene prediction frameworks, this method also revolves around the Guilt by Association principle. This principle is crucial here to make sure that the final selected disease community contains the genes interacting with a high-degree of modularity (collaboration) to the known disease genes as well as to each other.

To provide this important feature for our GA-based method, all the individuals in all the populations (i.e. all the induced evolving subgraphs) always hold the known disease genes inside them. In fact, we force all evolving subnetworks to include all known disease genes. In other words, all the known disease genes are added to candidate communities while their fitness values are measured. This approach steers the whole process to finding communities in which genes are in high collaboration with the known disease genes.

### D. Maximum Community Size

The maximum number of the nodes (genes) that can be inside a candidate community (individual representing a sub-network) is determined in advance of starting the GA. As a matter of fact, we deliberately force all candidate communities to contain no more than a specific fraction of the network's genes. This is mostly for the following reasons: (1) to avoid obtaining unwanted futile solutions such as a big community containing most of the network's genes (which cannot be considered a community any more), (2) to provide more a competitive environment across the communities' evolution, and above all (3) to supervise the convergence trajectory of the sub-networks towards the disease genes community we wish to obtain, i.e. the way in which the individuals converge towards the community of disease genes over generations in order to define a scoring system to prioritize candidate genes (see Section III-E).

In other words, candidate communities of the predetermined maximum size could be interpreted as frames that move over the various areas of the network during the generations of the GA. Being evolved by the GA, these frames have a tendency to move towards the areas of the network that show higher collaborations and modularity with the known disease genes.

To avoid acquiring biased output, we perform the same GA experiment several times to make sure that the population, i.e. all induced sub-networks, converges towards the same area of the network as the disease genes community.

### E. Prioritization of Candidate Genes

Each experiment performed by our GA contains the following as input data: a set of known disease genes and a PPI network containing all known disease genes and a set of candidate genes. All the genes cumulatively gather scores while the candidate communities (individuals) are evolved over generations. The scoring method is based on the idea that genes which are more frequently selected in candidate communities, while they are being evolved over generations, should accumulate higher scores. Therefore, genes which survive for a greater number of generations, and which are in a greater number of communities, acquire higher scores. The cumulative score of a given gene is proportional to the fitness value of the communities for which it is selected. In other words, in being selected for better communities (i.e. individuals with higher fitness value), genes get relatively higher scores. Over generations, candidate communities (individuals of every generation) gradually converge into sets of genes that are in higher collaboration with the known disease genes and with one another. Therefore as the candidate communities are evolved by the GA, they are more likely to be in the vicinity of the community we are looking for, i.e. the predicted disease genes set.

The scoring function for a given candidate gene is calculated based upon Algorithm 1, where $Q(c)$ is the modularity of the community $C$ as defined in Section III-A.

**Algorithm 1** Calculate score of gene $i$

```
for each generation do
    for each candidate community $C$ in population do
        if gene $i$ is in $C$ then Score($i$) = Score($i$) + Q($C$)
        end if
    end for
end for
```

## IV. Benchmark Tests

### A. Leave-One-Out Cross-Validation

To assess the performance of our GA-base method, we used the *leave-one-out cross-validation* procedure. This procedure is the most frequently used procedure to assess many of the state-of-the-art disease-gene associations approaches, including [43], [39], [20] and many others.

According to the leave-one-out cross-validation procedure, for every time performing the experiment, one of the known disease genes is removed from the known disease genes set, to observe whether or not the method can again recover this gene as a high associative gene. In other words [23], given a set of known disease genes $D$ and a set of candidate genes $C$ a gene $d \in D$ is left out and all genes in $C \cup \{d\}$ are ranked using the computational approach under study. This process is repeated for every gene in the known disease genes set.

### B. Fold enrichment analysis

Fold enrichment analysis is an important concept for using the leave-one-out cross-validation procedure. Here, we offer the simple and clear definition provided by X Wu et al [43]: if a method successfully ranks known disease genes in the top $m\%$ of all candidate genes for $n\%$ of the known disease genes, then there is on average a $n/m$-fold enrichment. Therefore, a threshold must be chosen to separate the two prediction classes. If a left-out known disease gene in leave-one-out cross-validation does not exceed the specified threshold in the candidate genes ranking, then it is considered as a successfully predicted gene in the validation.

### C. Receiver-Operating Characteristic (ROC) analysis

Another concept involved with the leave-one-out cross-validation performance measure is the *receiver-operating characteristic* (ROC) analysis. ROC is defined as a plot to measure the performance of a binary classifier according to a defined discriminating threshold. Here, for the disease-gene association/prioritization problem, ROC plots the true-positive(TP) rate versus the false-positive(FP) rate, subject to the threshold separating the prediction classes (see IV-B). We consider again the sets $D$ and $C$ and disease gene $d$ introduced in IV-A. Suppose that $\theta$ is the threshold which must not be exceeded by the rank of known disease gene $d$ in order for it to be considered as a successfully predicted known disease gene. TP(true positive) is the number of test cases where the ranking of $d$ is less than $\theta$, and FN(false negative) is the number of test cases where the ranking of $d$ exceeds $\theta$. The *sensitivity* of the computational disease-gene prioritization method for a given experiment is then defined by equation 4:

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

## V. Databases and Information Resources

### A. Protein-Protein Interaction Data

We obtained human PPI interaction network data using GeneMANIA [42], reported to be among the best performing methods in a public large scale computational biology challenge [33]. The homo sapiens (human) interaction database of GeneMania records different types of interactions (co-expression, protein interaction, pathway, co-localization etc.) based on 395 different interaction databases developed so far. It contains the interactions among 21435 genes from the human genome.

### B. *Disease-Gene Association Data*

In order to select disease genes and make the association between particular diseases and particular genes evidenced, we used Genotator [41], developed at Harvard University. Genotator is a realtime aggregation tool that provides both comprehensive coverage and reliable gene-to-disease rankings for many of the diseases. It integrates data from 11 well-known frequently used clinical genetics resources including Pubmed [1], OMIM [15], GeneCards [36], HuGE Navigator [45], GenAtlas [10], WikiGenes [16] and other phenotype-gene association data sources to rank genes in order of disease relevance.

## VI. Experimental Results

As a proof of concept, we applied our methodology to the problem of finding genes involved in breast cancer. We selected breast cancer due to the diversity of related studies and because there is a significant amount of research literature.

### A. *Input Data*

We used the same input data as Xuebing et al [43] used to demonstrate the ability of CIPHER, one of the most well-known and frequently used disease-gene association framework. Using OMIM[15], they selected 16 disease genes known to be involved in breast cancer. They measured the ability of CIPHER to predict disease genes based on these 16 genes. Table I outlines these genes.

TABLE I. KNOWN DISEASE GENES FOR BREAST CANCER

| Gene | NCBI ID |
|---|---|
| BRCA1 | 672 |
| BRCA2 | 675 |
| TP53 | 7157 |
| AR | 367 |
| ATM | 472 |
| CHEK2 | 11200 |
| STK11 | 6794 |
| RAD51 | 5888 |
| PTEN | 5728 |
| BARD1 | 580 |
| RB1CC1 | 9821 |
| NCOA3 | 8202 |
| PIK3CA | 5290 |
| PPM1D | 8493 |
| CASP8 | 841 |
| TGF1 | 281527 |

Using GeneMania (see V-A), we obtained an un-weighted protein-protein interaction network containing the 16 disease genes shown above. We defined the candidate genes to be the first 2000 genes scored by GeneMania in the order of interactions with these 16 disease genes.

### B. *GA Parameters*

We ran our GA-based method on this PPI network 20 separate times. Table II outlines our GA parameters for all 20 executions of the GA. These parameters were empirically determined. It should be noted that further study should be performed on determining the most appropriate values for these parameters for other experiments (e.g. for other diseases).

TABLE II. GA Parameter Settings

| Parameter | Value |
|---|---|
| number of generations | 300 |
| population size | 5000 |
| cross-over rate | %95 |
| mutation rate | %5 |
| tournament size | 3 |
| fitness evaluation | $Q(C)$ |
| maximum community size | 50 |

Figure 1 depicts the progression of the averages of best fitness values, and the averages of average fitness values for the 20 runs of the GA over the generations.

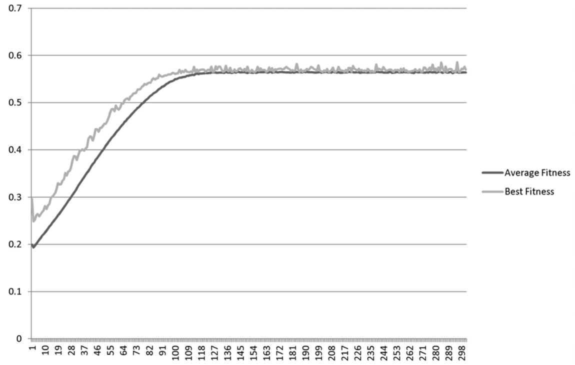

Fig. 1. averages of best/average fitness values over generations

## VII. Analysis

### A. *Leave-One-Out Cross-Validation*

Running our GA-based method on the input data described in VI-A and based on the parameters outlined in Table II, and using leave-one-out cross-validation method (explained in IV-A), our method successfully predicted 12 out of 16 genes as being involved in breast cancer. The threshold ($\theta$ - see Section IV) was set to the size of the community of the disease-genes the GA was evolving over generations (here, 50). Table III shows the 12 predicted genes and median of their 20 different rankings as a result of running the experiment 20 times.

### B. *Comparison with CIPHER*

CIPHER [43] used the same known genes, successfully ranking known disease genes in the top 0.917% of all candidate genes in 49.1% of test cases, giving them a fold enrichment of 53.50 on average. If we consider the top 1% of all candidate genes, Our GA-based approach gives fold enrichment of 50 on average, fold enrichment of 56.25 for the ranking offered by the best solution procured by GA, and fold enrichment of 62.50 if we consider the best rankings of the known disease genes among all 20 rankings given by GA.

TABLE III. 12 SUCCESSFULLY PREDICTED GENES OUT OF THE 16 KNOWN DISEASE GENES LISTED IN TABLE I

| Gene | Median of 20 rankings | Best Ranking of all |
|---|---|---|
| BRCA1 | 1 | 1 |
| TP53 | 1 | 1 |
| BRCA2 | 5 | 1 |
| ATM | 6 | 3 |
| AR | 8 | 2 |
| PPM1D | 16.50 | 8 |
| BARD1 | 19.50 | 5 |
| RAD51 | 20 | 13 |
| RB1CC1 | 22 | 10 |
| PTEN | 34.50 | 22 |
| STK11 | 42.50 | 17 |
| NCOA3 | 46 | 27 |

Based on the acquired results, ROC analysis (see IV-C) gives us 12 TPs and 4 FNs, thus the sensitivity of this experiment is 75%. The sensitivity of the experiment for CIPHER was 62.50% as a result of 10 TPs and 6 FNs.

### C. Predicting new Disease Genes

We briefly studied other candidate genes ranked among the top 1% of the candidate genes on average using the Genotator (see V-B). Our GA successfully predicted some other genes involved in breast cancer which were not in the known disease-genes set of the performed experiment. Among the other potential breast cancer genes ranked among the top 1% are ESR1, NBN, CHECK1, H2AFX, EP300, SMAD3, and other genes for which all have evidence of being involved in breast cancer by different data sources according to Genotator.

## VIII. Discussion

As elaborated in previous sections, our method tries to evolve a set of candidate genes with as much collaboration with the whole set of known disease genes as possible. It starts from randomly created communities of genes and optimizes the communities over generations, based on the modularity of the community (equation 1). The idea behind the scoring system (algorithm 1) we used is that the more frequently a gene is chosen to be in the optimizing communities, i.e. the more a gene is selected for the more optimized communities of the higher generations, the more it is associated with the known disease genes, and therefore should have a higher score compared to others that are selected less often.

Genes associated with a specific disease may act in separate communities which work with one another, or separate communities which overlap. We intuitively believe that our GA-based computation is capable of finding disease genes working in different communities or in overlapped communities for the following reasons. First, as the evolving populations contain thousands of different communities, different communities which work with the known disease genes would have the chance to evolve and be in the population at the same time. Genes in all such communities will get high scores as they are often selected in a number of the population's communities over generations, thereby they can increase their scores. Secondly, potential disease genes which lie in more than one community working with disease genes are selected more frequently as they can have the chance to be in many of the communities of the populations.

As a proof of concept, we tested our approach on 16 breast cancer genes taken from the similar experiments performed by Xuebing et al [43] to test the performance of CIPHER, a well-known disease gene prioritization framework. The results are satisfactory as one of the first experiments using a newly developing approach, however, there should be specific and more accurate strategies to determine the known disease gene set with regarding to different diseases and choosing appropriate GA parameters.

Determining an appropriate maximum community size for use in our method is also a matter of discussion. Different maximum community sizes can obviously affect the number of times each gene would be allowed inside the evolving community, thereby affecting the scores acquired by each gene. Here we chose a maximum community of size of 50, however further investigation should be conducted on this parameter, considering supporting reasons for choosing a given value. The accuracy of the results may possibly be increased as a result.

Concisely speaking, appropriate strategies and different parameter values may vary significantly from one disease to another.

## IX. Conclusion

Although protein-protein interaction networks have been shown to be among the most powerful pieces of evidence for disease-gene association ([7], [27]), there are still major concerns about the amount, accuracy, and quality of the available data, and there are still a considerable amount of interactions that are not well-studied [31]. To obtain more accurate and reliable disease genes communities one may use a fusion of different network-based data resources along with PPI [18].

It should be noted nonetheless that this is a proof of concept and that much further study remains. Most importantly, this method should be applied to other diseases. Also, as specified in section VIII above, future work must include further study of appropriate parameters, especially with respect to appropriate parameters for different diseases.

## Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

## References

[1] Abigail Acland, Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A Benson, Colleen Bollin, Evan Bolton, Stephen H Bryant, Kathi Canese, Deanna M Church, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 42(D1):D7, 2014.

[2] Russ B Altman, Casey M Bergman, Judith Blake, Christian Blaschke, Aaron Cohen, Frank Gannon, Les Grivell, Udo Hahn, William Hersh, Lynette Hirschman, et al. Text mining for biology-the way forward: opinions from leading scientists. *Genome Biol*, 9(Suppl 2):S7, 2008.

[3] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[4] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[5] Christian Blaschke, Miguel A Andrade, Christos A Ouzounis, and Alfonso Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. In *Ismb*, volume 7, pages 60–67, 1999.

[6] Yana Bromberg. Disease gene prioritization. *PLoS computational biology*, 9(4):e1002902, 2013.

[7] Jing Chen, Bruce J Aronow, and Anil G Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, 10(1):73, 2009.

[8] W Gregory Feero, Alan E Guttmacher, and Teri A Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

[9] Jan Freudenberg and P Propping. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 18(suppl 2):S110–S115, 2002.

[10] Jean Frézal. Genatlas database, genes and development defects. *Comptes Rendus de l'Académie des Sciences-Series III-Sciences de la Vie*, 321(10):805–817, 1998.

[11] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[12] John Gribbin, Jeremy Cherfas, Douglas Palmer, and Adam Hart-Davis. *Science: the definitive visual guide*. Dorling Kindersley Ltd, 2009.

[13] Emre Guney and Baldo Oliva. Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PloS one*, 7(9):e43557, 2012.

[14] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.

[15] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.

[16] Robert Hoffmann. A wiki for the life sciences where authorship matters. *Nature genetics*, 40(9):1047–1051, 2008.

[17] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome research*, 18(4):644–652, 2008.

[18] Jieun Jeong and Jake Y Chen. Techniques for prioritization of candidate disease genes. *Computational Intelligence and Pattern Analysis in Biology Informatics*, 13:309, 2010.

[19] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl 1):D480–D484, 2008.

[20] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958, 2008.

[21] Andrea Lancichinetti, Filippo Radicchi, and José J Ramasco. Statistical significance of communities in networks. *Physical Review E*, 81(4):046110, 2010.

[22] Jonas B Laurila, Nona Naderi, René Witte, Alexandre Riazanov, Alexandre Kouznetsov, and Christopher JO Baker. Algorithms and semantic infrastructure for mutation impact extraction and grounding. *BMC genomics*, 11(Suppl 4):S24, 2010.

[23] Duc-Hau Le and Yung-Keun Kwon. Gpec: a cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. *Computational biology and chemistry*, 37:17–23, 2012.

[24] Feng Luo, James Z Wang, and Eric Promislow. Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6(4):387–400, 2008.

[25] Melanie Mitchell. *Complexity: A guided tour*. Oxford University Press, 2009.

[26] Yves Moreau and Léon-Charles Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536, 2012.

[27] Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063, 2010.

[28] Martin Oti and Han G Brunner. The modular nature of genetic diseases. *Clinical genetics*, 71(1):1–11, 2007.

[29] Martin Oti, Berend Snel, Martijn A Huynen, and Han G Brunner. Predicting disease genes using protein–protein interactions. *Journal of medical genetics*, 43(8):691–698, 2006.

[30] Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285, 2008.

[31] Rosario M Piro and Ferdinando Di Cunto. Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS Journal*, 279(5):678–696, 2012.

[32] Mahbubur SM Rashid, Hasan Jamil, Raquel Hontecillas, Josep Bassaganya-Riera, et al. Gene regulatory network reveals oxidative stress as the underlying molecular mechanism of type 2 diabetes and hypertension. *BMC medical genomics*, 3(1):45, 2010.

[33] Matteo Re and Giorgio Valentini. Genes prioritization with respect to cancer gene modules using functional interaction network data. 2011.

[34] Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

[35] Francisco A Rodrigues, Guilherme Ferraz de Arruda, and Luciano da Fontoura Costa. A complex networks approach for data clustering. *arXiv preprint arXiv:1101.5141*, 2011.

[36] Marilyn Safran, Irina Dalah, Justin Alexander, Naomi Rosen, Tsippi Iny Stein, Michael Shmoish, Noam Nativ, Iris Bahir, Tirza Doniger, Hagit Krug, et al. Genecards version 3: the human gene integrator. *Database*, 2010:baq020, 2010.

[37] Nicki Tiffin, Miguel A Andrade-Navarro, and Carolina Perez-Iratxeta. Linking genes to diseases: its all in the data. *Genome Med*, 1(8):77, 2009.

[38] Marc A van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack AM Leunissen. A text-mining analysis of the human phenome. *European journal of human genetics*, 14(5):535–542, 2006.

[39] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

[40] Mahadevan Vasudevan and Narsingh Deo. Efficient community identification in complex networks. *Social Network Analysis and Mining*, 2(4):345–359, 2012.

[41] Dennis P Wall, Rimma Pivovarov, Mark Tong, Jae-Yoon Jung, Vincent A Fusaro, Todd F DeLuca, and Peter J Tonellato. Genotator: a disease-agnostic tool for genetic annotation of disease. *BMC medical genomics*, 3(1):50, 2010.

[42] David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl 2):W214–W220, 2010.

[43] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular systems biology*, 4(1), 2008.

[44] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[45] Wei Yu, Marta Gwinn, Melinda Clyne, Ajay Yesupriya, and Muin J Khoury. A navigator for human genome epidemiology. *Nature genetics*, 40(2):124–125, 2008.