A Centrality Based Multi-Objective Approach to Disease Gene Association

Tyler K. Collins and Sheridan Houghten

Computer Science Department, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, Ontario L2S 3A1, Canada

Abstract

Disease Gene Association finds genes that are involved in the presentation of a given genetic disease. We present a hybrid approach which implements a multi-objective genetic algorithm, where input consists of centrality measures based on various relational biological evidence types merged into a complex network. Multiple objective settings and parameters are studied including the development of a new exchange methodology, safe dealer-based crossover. Successful results with respect to breast cancer and Parkinson's disease compared to previous techniques and popular known databases are shown. In addition, the newly developed methodology is also successfully applied to Alzheimer's disease, further demonstrating its flexibility.

Across all three case studies the strongest results were produced by the shortest path-based measures stress and betweenness, either in a single objective parameter setting or when used in conjunction in a multi-objective environment. The new crossover technique achieved the best results when applied to Alzheimer's disease.

1. Introduction

The disease gene association problem (DGAP) concerns the understanding of the link between genes and diseases. Disease gene association uses search strategies to find and rank various genes based on their involvement in the presentation of a given disease. Information gathered is often used in two ways. Firstly, it is used for *gene prioritization*, the ranking of known problematic genes, representing how likely they are to be involved in presentation of the disease. Secondly, it is used to identify a reduced subset of genes which are considered to be highly involved, to be examined for further study; some of these may be genes that have yet to be investigated in relation to the given disease.

The *modularity principle* states that there does not exist a one-to-one relationship between genes and genetic diseases [33]. For example the gene BRCA1 is frequently present in cases of breast cancer. However, with this gene highly interacting on the protein level with other genes, its presence is not by itself enough to guarantee a presentation of breast cancer [2]. As such, the DGAP

and its respective methodologies need to be able to return more than a single individual gene as "guilty" of being involved in the disease. Thus if it is found that the interaction between a group of genes and BRCA1 is common, they are considered to be "guilty by association", i.e. likely to also contribute to the presentation of the disease. Genes involved in the same disease tend to closely interact [17]. Search strategies can take advantage of this and examine frequent gene interactions.

In this paper we use genetic algorithms to study genes involved in breast cancer, Parkinson's disease and Alzheimer's disease. Our methodology works with multiple types of evidence which together form a complex network. A multi-objective approach is used that takes into account several centrality measures of this network.

2. Background

This section briefly reviews the mathematical, biological and computational concepts required in this work.

2.1. Graph Theory: Formal Definitions

A graph G is defined as $G = (V, E)$, where V is a set of *vertices* (also known as *nodes*) and E is a set of *edges*, pairs of vertices implying a relationship between the vertices. A graph with weights for each edge is called *weighted*, while a graph without weights is *unweighted*; in this case, all edges are considered to have the same weight. Graphs are *directed* if the edges each have an associated direction, otherwise they are *undirected*. Topological properties of graphs include the following.

Degree of node v: The number of nodes u for which (u, v) is an edge.

Neighbour of v: Any other node directly connected to v by an edge.

Traversal or path: An ordered list of nodes that implies visiting each node in the list via edge connections in the order specified by the list.

Shortest path between nodes u and v: The path of lowest cost, of all paths between u and v.

Diameter of a graph: The length of its maximal shortest path.

Clique: A subset of nodes in a graph which, with respect to the group, form a complete graph.

2.1.1. Complex Networks

Graphs are often also called *networks*. Belonging to certain families of networks implies adherence to several additional definitions and constraints [6]. For example, road networks in cities, maps of the Internet, personal contact networks used in epidemic modeling, and other biologically inspired graphs all fall into the category of complex networks [7][18].

A *complex network* is most commonly defined as a network in which there are non-trivial topological features. However, this definition is not entirely formal. It does not exclude the appearance of some of the topological features introduced earlier, but it does imply that they are perhaps difficult both visually and computationally to discover. As such, it can be difficult to extract information and meaning from complex networks. Difficulty aside, complex networks still tend to have inherent properties, the most notable being that they tend to have small highly interconnected areas known as *communities* [38].

2.2. Centrality Measures

As it is difficult to extract meaning from complex networks, researchers studying them can use *centrality measures* that provide information on the level of importance of a node in a graph. These allow for meaning to be extracted from extremely large networks and afford the ability to rank nodes. Measures vary not only in computation, but also in their inspiration. For example, one measure may reward a node with a higher score for being isolated, while another may favour nodes with higher degree. Often, more complex measures use hybrid approaches.

The following definitions of centrality measures and the properties used to compute them are based on [40] and [23]. All of the following are considered in this study, with some being found to be more useful than others when applied to the problem of disease gene association. Note that this is a small sample of measures and the notion of centrality continues to be improved in the field of complex network analysis [5][44].

2.2.1. Degree

As defined previously, degree is the number of nodes in direct connection to a given node v. A high degree can imply that a node is highly connected in a network, thus inferring that it is central and important.

2.2.2. Eigenvector

Eigenvector is a recursive measure of the reachability of a node in a graph. A node v with a high eigenvector has many neighbours, which also have many neighbours, thus contributing to the overall score for v. If v has a low eigenvector then its local neighbourhood is sparsely connected with few overall nodes. Therefore a node with a high eigenvector relative to the rest of the graph can be thought of as a central node. The eigenvector for node v is computed recursively using the equation $Eig(v) = \frac{1}{\lambda} \sum_{w \epsilon N(v)} Eig(w)$, where λ is an eigenvalue of the graph represented as an adjacency matrix, and $N(v)$ is the set of neighbours of v. Initially, all nodes are given a default eigenvector value of one.

2.2.3. Closeness

The closeness of node v is computed by finding the shortest path from v to all other nodes in the graph. By taking the summation of these distances and computing the reciprocal, a measure for connectivity relative to the whole

network is defined. A node with high closeness is central to a graph as it takes few steps to travel from it to all other nodes in the graph.

2.2.4. Eccentricity

The eccentricity for a given node v is computed by finding shortest paths from v to all other nodes in the graph. The reciprocal of the distance of the longest of these paths is the eccentricity value. Eccentricity can be a difficult measure to work with. For example, a very "central" node which is far away from one other node has very low eccentricity. To combat such problems eccentricity should only be used when the topology of the graph is known to be fairly regular and without trails of nodes branching off from the main body of the graph.

2.2.5. Radiality

Radiality is a centrality measure computed respective to the diameter Δ_G of a graph: $C_r(v) = \frac{\sum_{w \epsilon N(v)} (\Delta_G + 1 - dist(v,w))}{n-1}$. Subtracting shortest paths from the longest shortest path, this centrality measure will return a high value if the node is involved in shortest paths.

2.2.6. Centroid Value

The centroid value of a node, like eigenvector, is highly influenced by neighbours more than one edge away. This increases the complexity drastically. Centroid value is $C_c(v) = min\{f(v,w) : w \epsilon N(v)\}$, where $f(v,w) = \gamma_v(w) - \gamma_w(v)$ and $\gamma_v(w)$ is the number of nodes in the graph that are closer in distance to v than w. Similar to eccentricity, centroid must be compared to the overall average to obtain a meaningful interpretation. A node with a high centroid value relative to the rest of the network can be interpreted as being a central node.

2.2.7. Stress

The stress of a given node v is computed by finding all shortest paths in the network and counting how many of these paths contain v. Stress differs from the previous measures as it does not necessarily reward a higher value based on connectivity relative to the graph, but instead based on how "travelled" the edges are. In a complex network based on communication, stress can reward nodes that connect dense areas. Comparison to average stress value of the network as a benchmark can provide a more accurate comparison between nodes.

2.2.8. Betweenness

Betweenness functions much like stress, but scales the number of shortest paths that pass through node v. The motivation is that stress does not account for the situation of a node being redundant. Betweenness attempts to combat this by dividing the total number of shortest paths that use v by all other shortest paths of equal length. Thus if betweenness returns a high value, then not only is the node providing an important connection, but it is also a critical non-redundant connection.

2.2.9. Bridging

This centrality measure also rewards nodes that occur in shortest paths rather than direct connectivity. However, it differs vastly in computation. It combines betweenness and eigenvector to create a new value that rewards critical connections and places a lower emphasis on degree. This relies on the notion that nodes with high degree typically exist in cliques, and bridging rewards nodes that "bridge" gaps between highly connected areas. Minimizing bridging finds the opposite, i.e. nodes part of highly connected areas that are non-bridging.

2.2.10. Limitations of Centrality Measures

Though centrality measures provide an effective way to analyze complex networks, they are not without flaws. Not every centrality measure or combination of measures will work for all graph types. For example, bridging may be useful when searching for important nodes in a telecommunication network, but less useful for a biologically generated network. These two types of networks often differ in both size and connectivity [18].

Additionally, issues when ranking with centrality measures can occur when there are many nodes with a low score. This is due to centrality measures only being functions which allow for a simple numeric comparison between nodes. These functions are often not normalized or scaled. The meaning is easy to discover when working with something simple such as degree, but when looking at (for example) centroid value it is not so clear.

Finally, most if not all centrality measures are related. Hence, although these measures are effective, we need to be aware of their relationships. Most centrality rankings are based on neighbours and as such are highly influenced by a node's degree. All the measures discussed in this paper, except for stress and betweenness, take into account neighbours in some way.

2.3. Community Detection

Often, real world data tends to have several highly connected nodes (*hubs*) while the remaining system is sparsely interconnected in comparison [45]. The process of finding and detecting these hubs is known as *community detection*. This is an ongoing field of research which seeks to make the process both more efficient and accurate [45]. This can be a challenging problem as networks vary in size, topology, and connectivity. Specifically, the problem of community detection can be shown to be NP-complete [45]. A proof sketch of this can be found in Fortunato's work in [15].

2.4. DGAP Input Data

The inputs to the DGAP are various gene interaction measures, which are briefly described in this section. Even a small subset of interactions contains many relations between numerous genes. Representing all of these relations at once in a graph forms a complex network [2].

2.4.1. Protein-Protein Interaction

One of the most widely used and most effective evidence types for this problem is that of protein-protein interaction (PPI) networks [32][9], which represent physical protein interactions (as edges) between genes (as nodes).

2.4.2. Co-expression

Genes under this evidence type are said to be related if they behave similarly across different environments, i.e. cell type. The strength of co-expression is that it is not biased to the highly studied genes, and as such can expose new and interesting relationships between genes [35].

2.4.3. Phenotype

Phenotypical evidence is based on pairs of genes leading to the presentation of the same phenotypes. Often this evidence type is used to strengthen already known relations. As a result, this evidence type is biased to highly studied genes and their effects [35].

2.4.4. Functional Annotations

Functional annotation relations refer to genes which are involved in creation of particular phenotypes and serve the same function during the process. This can be thought of as a low-level hybrid of both phenotypical and co-expression based relations.

2.4.5. Text Mining

This is based on the principles of meta analysis and seeks to relate genes that are often mentioned in conjunction with each other. Typically, these approaches mine databases of publications such as OMIM [19] via natural language methodologies. Text mining is among the first large scale approaches taken for the DGAP. More recent approaches, however, often use text mining in conjunction with other previously discussed evidence types.

2.5. Analysis and Benchmarking

As the success of a DGAP methodology is not directly related to the best score produced, various post processing and benchmarking techniques are used.

2.5.1. Leave-One-Out (LOO) Validation

This is the most common approach to benchmarking for such studies [24] and focuses on whether a methodology is able to recover known genes during execution. Formally, given N known disease related genes, fix $N\text{-}1$ of these to be in all candidate solutions. A LOO validation test is defined as successful if the left out gene is recovered by the method upon completion. This process is repeated for all N genes [24].

2.5.2. Fold Enrichment (FE)

This is an extension to LOO and seeks to define a pseudo-sensitivity measure. A methodology has an m/n average fold enrichment, where if a LOO was successful the technique correctly ranks known disease genes in the top $m\%$ for $n\%$ of the known genes [49]. A fold enrichment is counted as a success if the ratio is less than a defined threshold. This threshold often differs between studies and methodologies.

2.5.3. Receiver-Operating Characteristic (ROC)

This uses sensitivity, based on True Positives (TP) and False Negatives (FN). For a gene to be considered "found" and a TP its associated ranking must not exceed a given threshold, otherwise it is labeled as a FN. Sensitivity is defined as $\frac{TP}{TP+FN}$. This type of analysis is often studied in classifiers.

3. Previous Work

In several previous studies, researchers have taken the route of enriching input datasets to improve results, or to automate the entire process via construction of a pipeline. An example of this is [28], in which the proposed system improved the accuracy of results by including additional evidence types (e.g. co-expression, phenotype) on top of protein-protein interactions. The main technique used was that of a random walk algorithm.

The comparative study in [39] tested the effectiveness of a clustering based single-objective genetic algorithm technique across multiple distinct PPI disease data sets. While the authors found the genetic algorithm technique highly effective, it struggled compared to other approaches on diseases such as cerebral vascular disease. This could be due to the genetic algorithm's method of exploration of the fitness landscape having a more difficult time finding viable solutions upon which to build. The authors discussed that this relative difficulty may arise due to differences in the amount of existing study on the already known disease genes: for example, breast cancer is more highly studied than cerebral vascular diseases.

To combat this, HGPEC [25], a recently developed application for Cytoscape (see Section 4.2.1), sought to refine the decision making process for choosing known genes on a disease by disease basis. The plugin is based on random walk based measures. The authors of the application report novel disease gene relations discovered as a part of their case study of breast cancer.

An additional approach was introduced in Wu *et al.* [48] which sought to refine PPI based evidence via aggregating large amounts of standard PPI data together and building a weighted network, which was then trimmed into smaller sub graphs by selecting areas of dense weighting for continued analysis. As part of the study, the authors tested their methodology on several cancer types, including breast cancer. Significant improvements over similar families of methodologies were reported not only in accuracy, but also stability.

A genetic algorithm based on the principles of community detection was used in [43] and [21]. Candidate solutions in the GA took the form of communities of

potentially highly interacting genes, with some *known disease genes* identified ahead of time and automatically assigned to all communities. As part of the study breast cancer [43], and later Parkinson's disease [21], were investigated. Input data included only PPI based evidence types. Favourable results were achieved in comparison to popular disease gene databases.

As a continuation of the previous study, a GP approach was presented in [20]. The study added two new aspects: the inclusion of multiple evidence types in the data generation stage of the DGAP, and the use of centrality measures. By adding multiple evidence types the authors' hope was to strengthen important relationships between genes that are not strictly communicating on the protein level. This necessitated the introduction of centrality measures being computed on given DGAP input networks due to data size increases. The choice of which centrality measures to use was also briefly explored. The authors concluded that the measures of stress and betweenness when used in conjunction produced the strongest results. These results include an improvement over the previous GA-based work on Parkinson's disease. However, the technique was found to be slightly less effective when applied on the breast cancer dataset. In both case studies, the results improved on popular known DGAP databases.

As stated previously, there potentially exists merit in incorporating advanced community detection techniques into evolutionary computation based approaches. The locus-based adjacency representation (LAR), presented in [34], is a chromosome implementation focused on the ability of the candidate solution to possess multiple communities, and to vary in size. This ability not only comes from encoding and decoding of the chromosome, but also from modified genetic operators (i.e. crossover). LAR has been shown to be an effective methodology on benchmarking problems [36]. An extension to the LAR representation proposed in [27] introduced a local search based genetic algorithm (GALS). This approach focused on adapting the typical mutation operator in a LAR based scheme to be more focused on making local optimizations. Experimental results showed the technique to be both efficient and highly effective. A multi-objective optimization approach to the community detection problem was used in [41], in which the objectives were the in-degree and out-degree of nodes in the network.

4. Methodology

Our proposed system is a hybrid that takes into account multiple evidence types and multi-objective techniques. This new technique is inspired by the initial work in [43]. Although also related to [20], this novel approach uses a genetic algorithm rather than genetic programming. The GA is multi-objective, where each objective is the sum of the centrality measures of each node across the community (see Section 4.1.6).

4.1. Genetic Algorithm Details

4.1.1. Individual Structure

The internal representation of a candidate solution is a bitstring in which each index represents a gene, with a value of one indicating that the gene is

inside the community. The number of indices with a value of one (true) in a chromosome is equivalent to the community size, which is an input parameter. Chromosomes also possess a number of "fixed genes", which are genes known to already be involved in the disease; as such, their indices are always set to true.

4.1.2. Self Correction

Note that the genetic operators must ensure that the known disease genes are always present in the chromosome. However, despite this, through genetic operators a chromosome may end up not having the required community size, and if so it must undergo self correction. This can lead to destructive behaviour (see further discussion in Section 4.1.5). Self correction occurs as follows:

- Over Community Size: If the chromosome possesses more genes in the candidate community than the required number, self correction randomly removes genes from the community until the size is correct. Note that the randomly removed genes cannot be the fixed genes.
- Under Community Size: If a candidate community is under the required size, random new genes are added until the required size is reached.

4.1.3. Selection

All selection is implemented via tournament selection.

4.1.4. Mutation

Mutation is achieved by randomly selecting one (non-fixed) true index and one false index in the bitstring. These two indices then have their states inverted thus preserving the community size and incurring no self corrections as a result of mutation.

4.1.5. Crossover

Two crossover methods are used. The first method is that of a standard one-point crossover. This was selected as the results of initial testing showed that it performed better that two-point crossover for the problem at hand; this is possibly due in part to the fact that the chromosome is sparse. Note that although this method encourages strong exploration of the fitness landscape, it can be quite destructive to the chromosome, forcing many self corrections.

To combat self correction, a second crossover technique named *Safe Dealer-Based Crossover* (SDB) was created. A small example is shown in Figures 1 and 2. SDB first selects two individuals and finds the intersection of their true indices inside of the bitstrings. The two children are then initialized with these intersections also set as true. Next, all of the true indices that the parent chromosomes do not have in common are stored in a list and are shuffled. Finally, for each item in the list, evenly distribute values to be set as true to the children. Upon completion, each child should have exactly the required community size with no self corrections necessary. One downside to this methodology is that it can potentially lower the amount of exploration of the fitness landscape. This

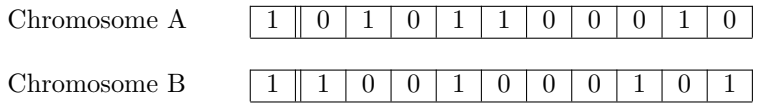

Figure 1: Before crossover. The network consists of 11 genes. Each chromosome represents a community of 5 genes and includes a single known disease gene shown as the leftmost gene.

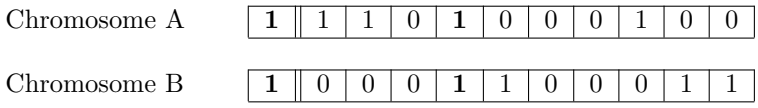

Figure 2: After SDB crossover. Both children retain the genes the parents held in common (shown in **bold**), which includes the known disease gene. The other genes held by the parents have been shuffled and evenly distributed to the children, ensuring that both children represent a community of the same size.

is due to new indices never being created, only "passed" between individuals. As such, for SDB to be successful it relies on a large population.

4.1.6. Fitness Methodology

Sum of Ranks (SoR) [4][11] differs from other popular multi-objective techniques such as NSGA-II [12] and SPEA2 [51] as it does not create or maintain a Pareto front, but rather returns the relative ranks of each individual to the population as a whole. Consequently, measuring the raw ranks of the individuals for any reason is no longer viable as ranks are only relative within a generation. However, SoR is not entirely without benefits. Consider the example of a small test population in Table 1, that contains the raw values of six individuals, along with SoR ranking. In this table, each objective is linearly ranked by raw fitness. Then for each individual, the sum of their ranks is computed. Finally, each individual is then re-ranked by this summed value. The resulting ranking is taken to be the output of the SoR technique. SoR is perfect for the proposed technique as no extra post processing must be done to determine which centrality measures to possibly sacrifice for a "good" choice to be made. Further value can be found in that no adjustment to standard methods of fitness selection must be done to fit SoR into a GA. It is entirely safe to return to basic premises such as tournament selection.

The variant known as normalized SoR differs from standard SoR only by normalizing the rank values obtained from the previous steps. The advantage of this optimization lies in that of eliminating issues of individuals having widely varying yet highly clustered scores resulting in ties. Note the implication that all objectives are balanced fairly despite potential diversity in ranks. Table 2 shows normalized SoR on the same population as Table 1.

With these properties in mind, (normalized) SoR becomes a very strong and clear choice for use in the proposed technique as it is easily extended to more dimensions, allows for the use of other standard techniques (selection, etc), and most importantly accounts for the wide variety of raw values. As such, this study will employ the use of normalized SoR in all experiments. Raw values

Indiv	Ranks				Sum of Ranks	Re-ranked
	A	*B*	*C*	*D*		
1	2	1	2	2	7	1
2	3	2	1	3	9	2
3	4	1	4	4	13	4
4	5	2	3	2	12	3
5	5	1	5	1	12	3
6	1	3	6	5	15	5

Table 1: Sum of ranks example

Indiv	Ranks				Sum Norm	Re-ranked
	A	*B*	*C*	*D*		
1	0.4	0.33	0.33	0.4	1.46	1
2	0.6	0.66	0.16	0.6	2.03	2
3	0.8	0.33	0.66	0.8	2.6	5
4	1	0.66	0.5	0.4	2.56	4
5	1	0.33	0.83	0.2	2.36	3
6	0.2	1	1	1	3.2	6

Table 2: Normalized sum of ranks example with same six individuals as in Table 1

for each objective will be computed by the summation of each gene's respective centrality measures that are being maximized if they are inside the candidate community.

4.2. Datasets and External Tools

4.2.1. Dataset Generation Process

For each disease studied, fixed known disease genes are selected via Genotator [46], a tool used to aggregate data and make it easier to search and index. Some number of the top genes are always selected as known; the number is specified separately for each case study. Note that these also become the benchmarking criteria for leave one out analysis (see Section 2.5.1). Next, these known genes are input into Cytoscape [10], open source software that allows for the study of genetically inspired networks. GeneMANIA [47] is a plugin for Cytoscape which predicts relationships between genes. It accomplishes this by building a complex network via the various available evidence types (PPI, pathway, etc). In this study, given the known disease genes from the disease being studied, we use GeneMANIA to query the next N most frequently interacting genes based on evidence types and store the resulting network in a Cytoscape structure, with the size of this network specified separately for each disease studied. Evidence types for this work include physical interactions (PPI), co-expression evidence, phenotype based relations, functional annotations, and predicted relationships via text mining. Lastly, the CentiScaPe [40] plugin is used. This allows the Cytoscape software to compute centrality measures based on

Parameter	HighCross	HighMut	Balance
Population	8000	8000	8000
Generations	2500	2500	2500
Selection: Tournament Size	5	5	5
Elitism	1	1	1
Crossover Rate	75%	30%	50%
Mutation Rate	25%	70%	50%
Runs	30	30	30
Community Size	100	100	100

Table 3: Summary of Parameter Settings

the currently open graph and export the results, which are then fed as input into the GA.

4.3. Evaluation Criteria

Biological parameters as well as various GA parameters are chosen to allow for comparison with previous work. In addition, all three testing procedures from Section 2.5 are implemented. Particular attention is given to the results of Leave-One-Out validation during empirical testing as successful runs can be examined individually. The results are compared to two other DGAP methodologies: CIPHER[49] for breast cancer and Endeavour[1] for Parkinson's disease.

5. Case Studies

We apply our methodology to three diseases: breast cancer, Parkinson's disease and Alzheimer's disease. In this section we first describe the overall experimental design, including parameter settings and fitness objectives common to all three case studies. The results for each of the case studies are described in Sections 5.3, 5.4, and 5.5 respectively, and discussion relevant to all three case studies is found in Section 5.6.

5.1. Experimental Design

Three individual parameter settings were empirically chosen in line with previous work to test the effectiveness of the methodology presented by this paper. Table 3 summarizes these settings. The fitness method used was normalized sum of ranks. Both one-point crossover and safe-dealer based (SDB) crossover were explored, with the same crossover rate applied to both.

5.2. Fitness Objectives

The purpose of this section is to illustrate the decision making process for selecting fitness objectives. A correlation study and preliminary fitness testing are presented in Sections 5.2.1 and 5.2.2 respectively. Although the fitness objectives are common to all three case studies, they were initially developed

Measures	Radiality	Betweenness	Bridging	Centroid	Closeness	Degree	Eccentricity	Eigenvector	Stress
Radiality	-	0.85	-0.48	0.99	0.99	0.94	0.31	0.92	0.88
Betweenness	0.85	-	-0.31	0.89	0.89	0.91	0.28	0.84	0.98
Bridging	-0.48	-0.31	-	-0.50	-0.50	-0.58	-0.21	-0.63	-0.41
Centroid	0.99	0.89	-0.50	-	0.99	0.95	0.32	0.94	0.90
Closeness	0.99	0.89	-0.50	0.99	-	0.98	0.33	0.95	0.92
Degree	0.94	0.91	-0.58	0.95	0.98	-	0.34	0.97	0.95
Eccentricity	0.31	0.28	-0.21	0.32	0.33	0.34	-	0.35	0.32
Eigenvector	0.92	0.84	-0.63	0.94	0.95	0.97	0.35	-	0.90
Stress	0.88	0.98	-0.41	0.90	0.92	0.95	0.32	0.90	-

Table 4: Correlation study from Breast Cancer data

with respect to the breast cancer dataset and thus these sections show the results for breast cancer. This disease was chosen due to its highly studied nature, allowing for greater possibility of drawing accurate conclusions.

Data generation for the breast cancer case study was performed as described in Section 4.2.1. The known genes for breast cancer are BRCA1, BRCA2, TP53, AR, ATM, CHEK2, STK11, RAD51, PTEN, BARD1, RB1CC1, NCOA3, PIK3CA, PPM1D and CASP8. These are kept equivalent to previous work to allow for easy comparison. Including these 15 known genes, the overall number of genes in the network is 2015.

5.2.1. Correlation Study

In an effort to fully understand which centrality measures to use as fitness objectives, a brief correlation study was performed on the breast cancer dataset as a precursor to potentially using principal component analysis. Samples were formed by grouping together numerical results from each centrality measure found in Section 2.2. Correlation was then computed using a standard Pearson correlation coefficient [3]. Table 4 contains the results of this study.

Understanding why eccentricity possessed a low correlation coefficient yet was unfavoured by previous methods can be found in the raw data. Across all 2015 nodes in the graph, only four values appeared (0.3, 0.5, 0.25, and 0.75). This illustrates why previous studies found eccentricity to be lacking, as well as accounting for the lower correlation coefficients across the table. Placing the actual complex network into visualization software reveals why eccentricity failed to report varied values. Several nodes exist as outliers to the network. This gives each node in the network a far distant node to travel to, producing practically static values. This is a practical example of the limitations of centrality measures as specified in Section 2.2.10.

Particular note should be made of the correlation existing via stress and betweenness. With stress and betweenness possessing a correlation of 0.98, it presents as counter-intuitive that previous work [20] identified that these two measures working in conjunction produced the most successful results. As a result, the following subsection will include an effort to replicate the success of these two measures in a purely multi-objective environment. Additionally, bridging will also be a focus of preliminary testing as it presents fairly low negative correlations with respect to all other centrality measures.

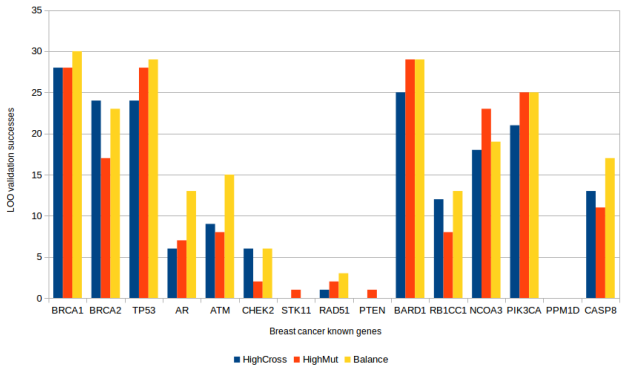

Figure 3: Bridging and betweenness LOO validation parameter setting comparison on one-point crossover

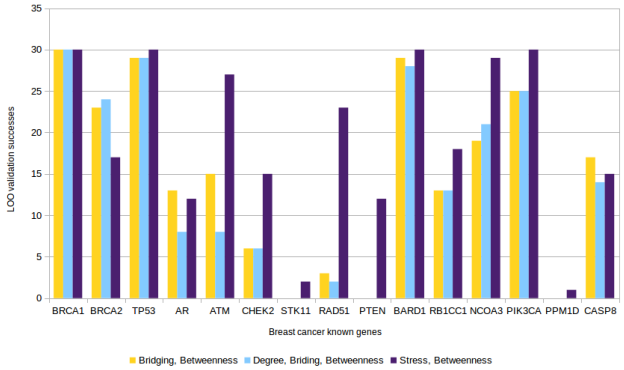

Figure 4: Comparison of LOO validation successes on breast cancer using one-point crossover on the Balance parameter setting

5.2.2. Preliminary Fitness Testing

The breast cancer dataset was used to perform preliminary fitness testing, using the information from the correlation study. This was then used to establish the fitness objective settings to be investigated.

Figure 3 illustrates the success rates of Leave-One-Out (LOO) validation based on the GA parameters given in Table 3 on the hypothesized multi-objective setting of bridging and betweenness. Unsurprisingly, the HighCross parameter setting performs worse in all validation cases due to the destructive nature of one-point crossover for this methodology's representation. Going forward, results in following experiments will also include Safe Dealer-Based (SDB) crossover (see Section 4.1.5) for comparison.

Several fitness objective settings were investigated that combined different centrality measures. From among these, three produced completely dominant results; as a consequence the remainder are excluded from this paper. Figure 4

Label	Objective(s)
Base	Bridging, Betweenness
Three	Degree, Bridging, Betweenness
AK	Stress, Betweenness
sBet	Betweenness
sStress	Stress

Table 5: Fitness objective labeling scheme

contains the comparison of these top three schemes using the Balance parameter setting. This shows that the objective setting of stress and betweenness produces the strongest results. Note that despite the differences in approach, these two path based measures working together have been replicated to be highly successful despite their extremely high correlation.

For the sake of brevity, Table 5 defines labels for the objective settings that will be explored as part of this work. These include the top three multi-objective settings, as well as two single-objective settings. The labels are designed to be used in conjunction with the parameter settings defined in Table 3. For example, results defined by the label "AK-Balance" imply a fitness objective setting of stress and betweenness combined with the Balance parameter setting.

5.3. Breast Cancer Case Study

This section contains detailed results from the application of our methodology applied to breast cancer. Breast cancer typically begins with tumors growing in breast tissue. In general, the prognosis for the disease is highly dependent on the nature of the cancer, extent of spreading to other organs, as well as the individual's age [16]. Despite this, recent scientific advances have lead to a significantly improved prognosis of this disease.

5.3.1. Experimental Results

Tables 6 and 7 contain study-wide summaries of each experiment type based on the evaluation criteria defined in Section 4.3. Note that fold enrichment (FE) measures are computed on one set of thirty samples, one gene at a time, before being aggregated into a list to find the best and median values and are then averaged with higher values representing better performance. Single objective settings do not include HighCross and HighMut due to the Balance setting possessing completely dominant results.

From these results, it is apparent that the SDB crossover is significantly less effective for this problem as even the worst one-point setting out performs the best SDB approach in terms of sensitivity. However, it still interesting to note that some SDB settings produce a higher FE. This is likely due to the fact that SDB crossover, while less effective, is considerably more stable and converges more quickly.

Based on Table 6, the most successful measures for this case study are the sBet-Balance setting and the AK fitness scheme due to their result of recover-

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	12/15	0.80	678.29	35.13
Three-HighCross	13/15	0.87	712.45	18.43
Three-HighMut	12/15	0.80	578.91	66.23
Base-Balance	13/15	0.87	729.33	32.08
Base-HighCross	12/15	0.80	736.54	18.43
Base-HighMut	12/15	0.80	578.01	72.65
AK-Balance	15/15	1.00	1092.99	87.96
AK-HighCross	15/15	1.00	1137.39	46.68
AK-HighMut	15/15	1.00	1062.76	126.35
sBet-Balance	15/15	1.00	1243.21	91.26
sStress-Balance	14/15	0.93	831.25	72.49

Table 6: Breast cancer with one-point crossover experiment summary

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	8/15	0.53	875.56	57.88
Three-HighCross	8/15	0.53	837.33	47.86
Three-HighMut	8/15	0.53	790.59	58.17
Base-Balance	8/15	0.53	832.42	61.62
Base-HighCross	8/15	0.53	870.22	58.00
Base-HighMut	8/15	0.53	827.00	61.02
AK-Balance	11/15	0.73	1333.39	83.79
AK-HighCross	10/15	0.67	1333.33	75.37
AK-HighMut	10/15	0.67	1035.56	94.50
sBet-Balance	11/15	0.73	1352.38	127.24
sStress-Balance	9/15	0.60	1133.33	82.08

Table 7: Breast cancer with SDB crossover experiment summary

Gene	Best Rank	Median Rank	Mean Rank	Std Rank	LOOs
BRCA1	15	19.50	20.73	6.65	30
BRCA2	23	76.50	83.27	37.44	16
TP53	1	23.00	29.40	18.49	30
AR	27	88.50	176.27	146.14	9
ATM	15	72.00	111.83	119.40	25
CHEK2	31	74.50	104.70	94.81	22
STK11	68	213.50	254.83	139.50	1
RAD51	15	56.50	76.73	97.57	27
PTEN	30	188.00	221.47	134.91	11
BARD1	15	39.50	52.73	62.25	29
RB1CC1	15	60.50	96.37	100.57	20
NCOA3	15	34.00	42.60	24.37	29
PIK3CA	1	22.50	23.27	7.51	30
PPM1D	59	258.50	266.87	142.39	2
CASP8	15	83.50	130.00	130.03	21

Table 8: AK-Balance parameter setting with one-point crossover individual gene statistics on breast cancer

Gene	Best Rank	Median Rank	Mean Rank	Std Rank	LOOs
BRCA1	1	25.00	25.53	9.92	30
BRCA2	15	65.50	71.07	36.91	15
TP53	1	20.50	19.17	5.54	30
AR	1	85.50	100.47	87.10	13
ATM	1	40.50	46.23	26.39	30
CHEK2	20	83.00	103.17	86.08	21
STK11	33	220.00	255.23	150.55	7
RAD51	15	50.00	83.97	104.83	26
PTEN	32	112.50	174.47	142.91	12
BARD1	1	35.00	36.30	15.33	30
RB1CC1	1	76.50	115.00	128.79	19
NCOA3	1	31.00	32.93	15.59	30
PIK3CA	15	21.00	22.53	8.61	30
PPM1D	22	214.50	245.87	151.37	3
CASP8	1	56.00	102.17	107.86	21

Table 9: sBet-Balance parameter setting with one-point crossover individual gene statistics on breast cancer

Approach	Avg. Median FE	LOO	Sensitivity
CIPHER	25	10/15	0.67
Past GA Approach	30	12/16	0.75
GP Approach	24	9/15	0.60
AK-Balance	88	15/15	1.00
sBet-Balance	91	15/15	1.00

Table 10: Breast cancer DGAP methodology comparison.

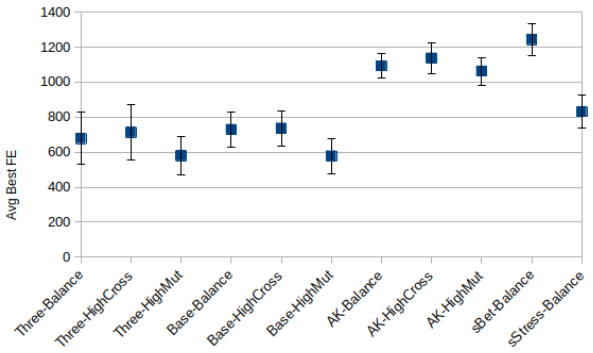

Figure 5: Breast cancer fold enrichment across all settings using one-point crossover

ing all fifteen known LOO validation genes. Tables 8 and 9 contain the individual gene success rates and statistics for both AK-Balance and sBet-Balance respectively, with the "LOOs" column representing the number of total LOO validation successes. AK-Balance has been selected from its family of settings as it possesses both the second best average FE, and second best median FE. In these tables, if a tie exists between two or more genes, the next place gene is incremented accordingly. For example, if there is a 14-way tie for first place, second place is then technically awarded 15th place. Surprisingly, despite the multi-objective scheme possessed by AK-Balance, sBet-Balance appears to successfully rank the known genes higher in the best cases. Both settings appear to struggle with similar genes, namely STK11 and PPM1D.

Table 10 compares the results with a non-EC methodology (CIPHER[49]) as well as the two EC techniques on which this study is based [43][20]. This shows that the technique proposed in this paper is an improvement on past techniques as all fifteen known LOO validation genes have been recovered. Additionally, average median FE has more than doubled in both cases.

Figures 5 and 6 show 95% confidence intervals of fold enrichment for one-point and SDB crossover respectively. These figures demonstrate that for this case study, when considering fold enrichment the "AK" settings that combine stress and betweenness outperform all other multi-objective settings; also, these were in turn slightly outperformed by the single objective setting of betweenness.

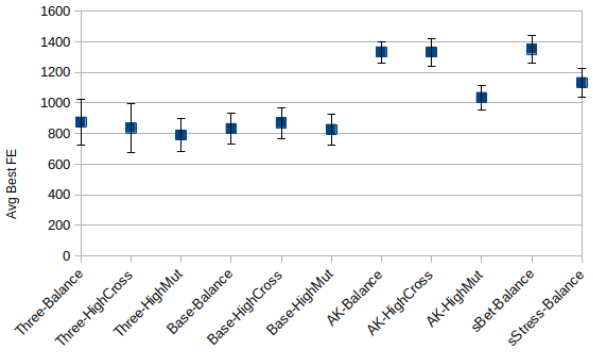

Figure 6: Breast cancer fold enrichment across all settings using SDB crossover

Gene	Description
APP	Amyloid beta precursor protein
PIK3R1	Phosphoinositide-3-kinase regulatory subunit 1
EP300	E1A binding protein p300
CHEK1	Checkpoint Kinase 1
RFC3	Replication Factor C Subunit 3
PCNA	Proliferating Cell Nuclear Antigen

Table 11: Predicted genes for future breast cancer study.

5.3.2. Predicted Genes

As a goal of disease gene association is predict new genes for study, the top 1% of non-fixed genes across the AK-Balance and sBet-Balance experiments were examined. Table 11 contains a subset of these selected genes confirmed by Genotator to have known interactions with breast cancer.

Of particular note from these genes are APP and EP300. APP is typically thought to be a gene involved in the presentation of Alzheimer's disease as it is often found in brain and spine tissues. Furthermore, even small mutations in this gene often lead to early onset Alzheimer's disease [31]. Interestingly, recent studies have begun to examine its relevance to the onset of breast cancer [26] confirming the previous prediction. EP300 is often seen regulating cell growth and division and as such has already been confirmed to be linked to various different types of cancer [31].

5.4. Parkinson's Disease Case Study

Parkinson's disease is a genetic disease which leads to the degeneration of the nervous system [37]. Typical symptoms of Parkinson's include tremors, impaired balance, and communication difficulties [22]. As a result of the degeneration of the nervous system, individuals with Parkinson's have a reduced life expectancy. Parkinson's is also known to be a particularly complex disease with numerous genetic interactions. This section applies our methodology to

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	8/15	0.53	677.81	409.69
Three-HighCross	9/15	0.60	449.29	404.39
Three-HighMut	8/15	0.53	656.16	426.45
Base-Balance	8/15	0.53	723.99	405.51
Base-HighCross	8/15	0.53	647.72	404.35
Base-HighMut	8/15	0.53	586.14	417.82
AK-Balance	12/15	0.80	1276.49	508.05
AK-HighCross	13/15	0.87	1427.95	486.99
AK-HighMut	12/15	0.80	1203.65	664.49
sBet-Balance	12/15	0.80	1402.37	512.99
sStress-Balance	13/15	0.87	1150.58	500.15

Table 12: Parkinson's with one-point crossover experiment summary

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	6/15	0.40	977.78	429.36
Three-HighCross	6/15	0.40	906.00	420.88
Three-HighMut	6/15	0.40	591.03	430.63
Base-Balance	5/15	0.33	805.96	417.31
Base-HighCross	6/15	0.40	716.83	317.04
Base-HighMut	5/15	0.33	820.00	424.29
AK-Balance	8/15	0.53	1234.29	365.71
AK-HighCross	8/15	0.53	1087.33	463.43
AK-HighMut	8/15	0.53	1106.98	474.71
sBet-Balance	6/15	0.40	1200.00	467.83
sStress-Balance	8/15	0.53	1333.33	490.63

Table 13: Parkinson's with SDB crossover experiment summary

Parkinson's disease. The parameter settings and fitness objectives are as defined in Sections 5.1 and 5.2 respectively.

Input data for this case study is generated as defined in Section 4.2.1. The known genes for Parkinson's disease are LRRK2, SNCA, PARK2, MAPT, APOE, GBA, GAK, BST1, DRD2, PINK1, MAOB, BDNF, CYP2D6, PON1, and COMT. These are the same as those in previous work to allow for easy comparison. Including these 15 known genes, the overall number of genes in the network is 3015.

5.4.1. Experimental Results

Tables 12 and 13 contain experiment summaries for both one-point and SDB crossover respectively. It is evident that, despite the overall success of the methodology, Parkinson's disease presents a greater challenge than breast cancer

Gene	Best Rank	Median Rank	Mean Rank	Std Rank	LOOs
CYP2D6	197	370.00	355.33	43.50	1
GAK	15	357.50	268.20	134.67	5
MAOB	15	22.50	30.83	31.75	26
PON1	15	123.00	160.80	130.72	22
COMT	28	358.50	264.43	141.04	8
GBA	287	370.50	368.87	18.45	0
BST1	15	48.00	84.10	100.20	27
DRD2	15	19.00	18.50	3.17	30
LRRK2	67	364.00	347.93	58.56	0
MAPT	15	303.00	218.83	157.94	14
PINK1	15	52.50	105.87	107.57	26
APOE	22	119.50	197.40	148.49	9
BDNF	22	98.50	149.63	130.19	16
PARK2	1	15.00	11.20	6.88	30
SNCA	1	15.00	12.57	6.47	30

Table 14: AK-HighCross parameter setting with one-point crossover individual gene statistics on Parkinson's

for the problem of disease gene association. However, the relative performance between settings mirrors that of breast cancer. In particular, the hypothesized performance of bridging is lower than expected, and also the single-objective settings perform extremely well. Furthermore, SDB crossover is again, even in the best case, out performed by some of the worst case one-point crossover parameter settings. One interesting difference between case studies is that the average median FE values appear to be higher across all parameter settings in comparison to those for breast cancer. This is likely tied to individual gene rankings, with some being significantly easier to identify than others. Tables 14 and 15 contain individual gene rankings and performance measures. One-point crossover versions of AK-HighCross and sStress-Balance have been selected as they recover the highest number of the known LOO validation genes.

As with the previous case study, some of the genes are more difficult than others to recover. For Parkinson's disease, these include CYP2D6, GAK, COMT, and APOE. In the case of LRRK2 and GBA, the methodology did not recover them at all. The remaining LOO validation genes however appear to be extremely stable. Interestingly, the approximate difficulty of these genes appears to be mirrored in the sStress-Balance parameter results.

Table 16 contains a comparison between past EC approaches [21][20] as well as a popular biologically inspired DGAP approach, Endeavour[1]. As in the previous case study, the proposed technique out performed past methodologies in terms of sensitivity and FE. Note that a large portion of the success of this methodology in terms of FE is due to the overall stability obtained due to the ease of ranking several of the known LOO validation genes.

Gene	Best Rank	Median Rank	Mean Rank	Std Rank	LOOs
CYP2D6	83	425.50	377.13	103.39	1
GAK	81	410.50	324.43	136.42	2
MAOB	15	21.50	37.87	30.77	21
PON1	26	94.00	165.07	150.37	16
COMT	36	421.50	331.30	138.53	3
GBA	336	430.00	429.07	23.56	0
BST1	1	31.50	53.67	81.84	29
DRD2	1	18.50	17.63	4.16	30
LRRK2	348	427.50	428.50	21.41	0
MAPT	25	265.50	246.00	177.65	10
PINK1	1	39.50	65.87	101.67	28
APOE	18	206.00	253.37	167.85	5
BDNF	33	152.50	187.93	129.63	5
PARK2	1	15.00	13.73	4.68	30
SNCA	1	15.00	13.17	6.06	30

Table 15: sStress-Balance parameter setting with one-point crossover individual gene statistics on Parkinson's

Approach	Avg. Median FE	LOO	Sensitivity
Endeavour	11.11	3/15	0.20
Past GA Approach	18.33	5/15	0.33
GP Approach	22.22	6/15	0.40
AK-HighCross	486.99	13/15	0.87
sStress-Balance	500.15	13/15	0.87

Table 16: Parkinson's DGAP methodology comparison. Balance method implements the one-point crossover technique.

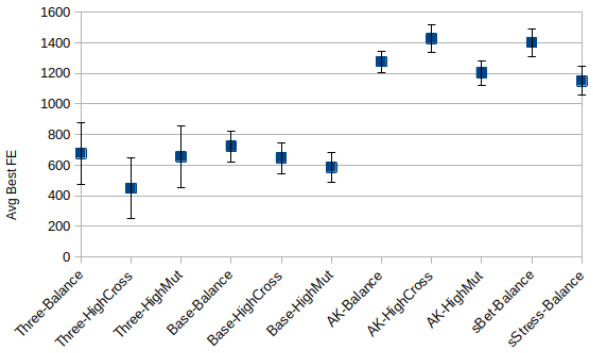

Figure 7: Parkinson's disease fold enrichment across all settings using one-point crossover

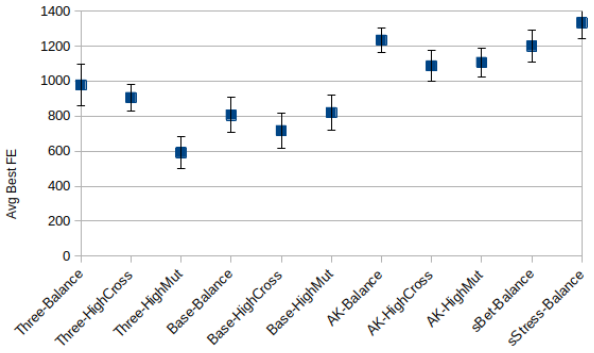

Figure 8: Parkinson's disease fold enrichment across all settings using SDB crossover

Gene	Description
APP	Amyloid beta precursor protein
NTRK2	Neurotrophic Receptor Tyrosine Kinase 2
UBC	Ubiquitin C
C8A	Complement C8 Alpha Chain
LECT2	Leukocyte Cell Derived Chemotaxin 2

Table 17: Predicted genes for future Parkinson's study.

Figures 7 and 8 show 95% confidence intervals of fold enrichment for one-point and SDB crossover respectively. For this case study, the "AK" settings outperform all other multi-objective settings even more clearly than for the breast cancer case study. When considering fold enrichment, the "AK-HighCross" setting outperforms all settings when one-point crossover is used, however the single-objective setting of stress outperforms all other settings when SDB crossover is used.

5.4.2. Predicted Genes

The top 1% of non-fixed genes across both the AK-HighCross and sStress-Balance parameter settings were investigated. Table 17 contains the noteworthy genes confirmed to have known interactions with Parkinson's by Genotator. Interestingly, the APP gene is again identified by the methodology as particularly guilty. As stated previously, the APP gene is often found in brain and spine tissues so it comes as no surprise that there already exist links between this gene and the neurodegenerative properties of Parkinson's disease [13]. Second, NTRK2 is typically responsible for maturation and development of various central nervous system properties such as synapse formation. This gene is often associated with epilepsy, however recent studies have confirmed its involvement with Parkinson's disease [42]. Additionally, the C8A and UBC genes have been previously identified by past EC approaches as being guilty by association [21].

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	9/9	1.00	1181.42	30.24
Base-Balance	9/9	1.00	1272.37	24.97
AK-Balance	9/9	1.00	1828.56	901.28
sBet-Balance	9/9	1.00	1814.30	1170.43
sStress-Balance	9/9	1.00	1784.29	706.91

Table 18: Alzheimer's with one-point crossover experiment summary

Setting	LOO	Sensitivity	Avg. Best FE	Avg. Median FE
Three-Balance	9/9	1.00	1791.36	650.30
Base-Balance	9/9	1.00	1909.37	1187.79
AK-Balance	9/9	1.00	1899.23	1303.82
sBet-Balance	9/9	1.00	1942.23	1509.97
sStress-Balance	9/9	1.00	1838.50	1205.16

Table 19: Alzheimer's with SDB crossover experiment summary

5.5. Alzheimer's Disease Case Study

Alzheimer's disease is a long term degenerative disease which affects the brain. Typical symptoms include loss of short-term memory, mood swings, language deficiencies, and eventual sustained dementia. Long term effects of the disease typically result in loss of bodily functions, and eventually death. Despite the deadly nature of the disease, it is still largely poorly understood [8].

Data generation for this case study is based on the procedure outlined in Section 4.2.1. However, Genotator does not possess database entries for Alzheimer's disease. As such, the selection process for finding known genes must be accomplished another way. The Online Mendelian Inheritance in Man (OMIM) database is an online resource for aggregating human genome data with a focus on genetic disorders [19]. The known Alzheimer's genes as defined by OMIM are HFE, NOS3, PLAU, A2M, MPO, APP, PSEN1, PSEN2, and APOE. These genes are used as input to Cytoscape, and the data generation then continues as normal. Note that as this is a less rigorous process than the previous case studies, the known genes have been selected in a conservative manner. Including the 9 known genes, the overall number of genes in the network is 2009.

The parameter settings and fitness objective labels are as defined in Sections 5.1 and 5.2 respectively. However, for the sake of brevity, only the Balance parameter setting is reported for this case study. This is due to the apparent difficulty of the problem being significantly lowered.

5.5.1. Experimental Results

Tables 18 and 19 contain experimental summary statistics. Simply by inspection, it is clear that this problem is significantly easier for this methodology

Gene	Best Rank	Median Rank	Mean Rank	Std Rank	LOOs
HFE	1	10.00	11.07	8.63	30
NOS3	1	1.00	3.97	6.34	30
PLAU	1	10.50	10.57	7.98	30
A2M	1	1.00	1.50	3.42	30
MPO	1	47.00	42.80	23.75	30
APP	1	1.00	3.80	6.05	30
PSEN1	1	23.00	22.47	12.37	30
PSEN2	1	4.00	5.93	6.71	30
APOE	1	1.00	4.20	6.75	30

Table 20: sBet-Balance with SDB crossover individual gene statistics on Alzheimer's

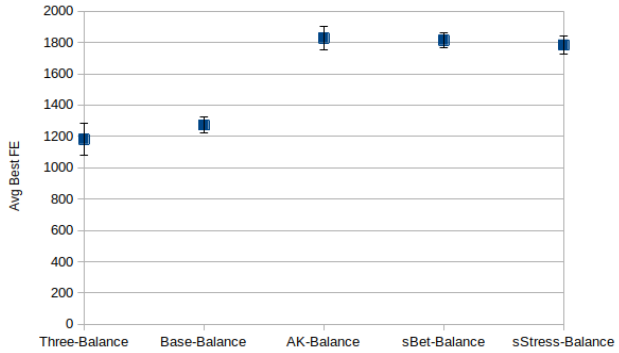

Figure 9: Alzheimer's disease fold enrichment across all settings using one-point crossover

than the previous two case studies. Even in the case of the previously under performing SDB crossover, all nine known genes are recovered with particularly high FE values across all objective settings.

Table 20 contains the individual LOO success rates on a gene by gene basis for the sBet-Balance parameter setting with SDB crossover. This objective setting is selected as it successfully recovered all nine known genes and produced the highest FE values. This table contains extremely stable and highly ranked results for each known gene. The only two genes that the methodology did not have in the top ten of its ranking were MPO and PSEN1. However, all thirty test cases were recovered easily.

Figures 9 and 10 show 95% confidence intervals of fold enrichment for one-point and SDB crossover respectively. In comparing these two figures, one can see that SDB crossover both outperforms and has more consistent results than one-point crossover in terms of fold enrichment. The only exceptions are the "AK-Balance" and "sStress-Balance" settings, in which for both cases they are effectively tied.

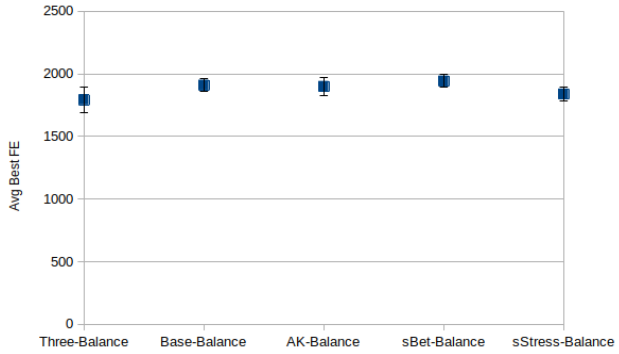

Figure 10: Alzheimer's disease fold enrichment across all settings using SDB crossover

Gene	Description
CTNNB1	Catenin Beta 1
BGN	Biglycan
CD59	CD59 Molecule (CD59 Blood Group)
CDH5	Cadherin 5

Table 21: Predicted genes for future Alzheimer's study.

5.5.2. Predicted Genes

Table 21 contains potential new genes for study based on the top 1% most frequently found non-fixed genes. The top such gene reported by this methodology is CTNNB1, which is known to be responsible for protein creation that regulates connections between cells. Typical health concerns associated with this gene are tumors and various different types of cancers [31]. Despite this, multiple studies have shown interaction between CTNNB1 and Alzheimer's [29][14]. The BGN gene, which is known to be responsible for bone growth and muscle development, does not necessarily have any explicit links to Alzheimer's in recent studies. As such, this gene is a perfect candidate for future investigation of this disease.

Upon further investigation, both the CD59 and CDH5 genes appear to already be known as influential genes with respect to Alzheimer's [50][30]. Going forward, the known genes list should probably be expanded to include this pair as the data generation for this case study was quite conservative.

5.6. Discussion

5.6.1. Centrality Measures

Originally, our hypothesis was that a multi-objective methodology containing the centrality measure of bridging would produce the strongest results. This hypothesis came from examining the correlation between pairs of measures and determining that bridging in conjunction with any other measure provided at least some quantity of data that was not highly correlated.

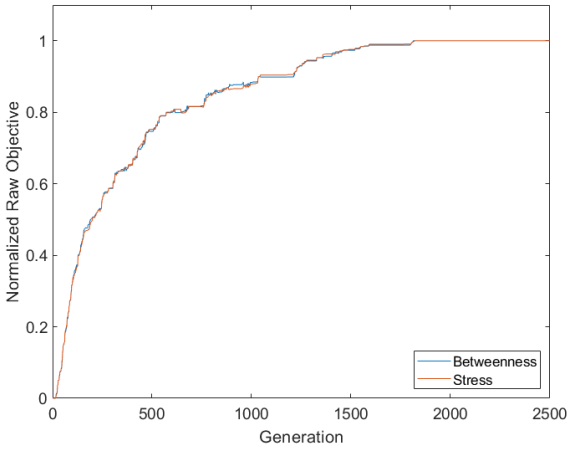

Figure 11: Breast cancer AK-Balance with one-point crossover experiment fitness curve

The breast cancer case study showed that the previous literature was correct to use stress and betweenness together (AK fitness schemes). Furthermore, using betweenness on its own in a single-objective environment produced improved results. For breast cancer, the path based measures of stress and betweenness used in conjunction or on their own yielded significant improvements over past approaches in terms of the recovery of each known LOO validation gene. This success is likely due to integration of multiple evidence types, as well as the exchange mutation approach being a strong hybrid of both exploration and exploitation.

A similar situation occurred with the Parkinson's dataset, with strong results for the path based objective settings, AK and sStress. However, not all genes were recovered and hence the raw values of these genes were inspected. In the case of each missed gene, the raw centrality measure values of stress and betweenness were at best approaching the mean or lower. In fact, a similar inspection of the troublesome genes from the breast cancer case study confirmed the same behaviour. This implies that stress and betweenness in some cases are not enough to see the importance of a gene's role in a network. Future studies should thus investigate more centrality measures.

The conservative manner of choosing the known genes for Alzheimer's disease was in part because tools like Genotator did not natively possesses database entries for this disease. As a result, the difficulty of this case study is significantly reduced in comparison to the others, and all known genes were recovered for all objective settings.

5.6.2. Crossover

Figures 11 and 12 contain example convergence curves for AK-Balance experiments using one-point and SDB crossover respectively on the breast cancer

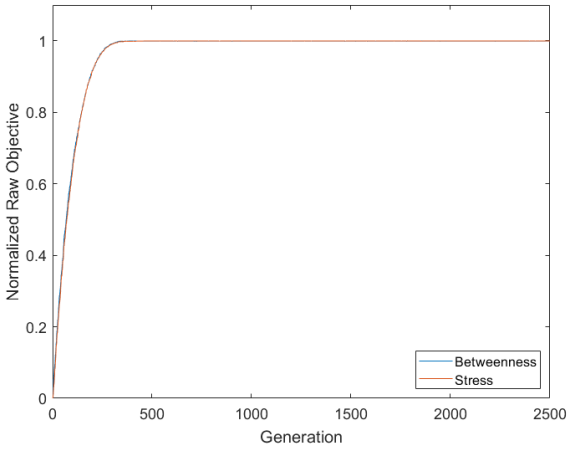

Figure 12: Breast cancer AK-Balance with SDB crossover experiment fitness curve

dataset; although not shown, the corresponding convergence curves for both Parkinson's disease and Alzheimer's disease are very similar to these. Note that regular convergence curves using the Sum of Ranks (SOR) methodology are not possible due to the fitness of individuals always being relative to each epoch. As a result, data points in these figures are the raw objective scores of the best individual at each generation. These raw values are then normalized for easier visualization as the numerical values of various centrality measures can vary widely with respect to one another.

Figure 11 illustrates a fairly standard convergence for an EC technique, while Figure 12 shows an extremely fast convergence rate. While not necessarily always a detriment, a convergence of this type can often mean lack of diversity in a population due to either selection pressure being too high, or the actual exchange of genetic material between individuals being limited during evolution.

The original motivation behind SDB was to avoid the numerous self corrections possible when one-point crossover was used. Use of SDB resulted in lowered sensitivity rates across all testing conditions. However, it generally produced higher FE values, likely due to the highly exploitative nature of the technique. Since it does not create "new" information and only exchanges data between chromosomes, the technique is considerably less explorative. This is reflected in how quickly the technique converges in comparison to a one-point crossover approach even with such a large population size.

The reduction in difficulty for the Alzheimer's data set greatly affects the performance of the previously underwhelming SDB crossover. With this reduction in problem difficulty, the highly exploitative nature of the technique is much more desirable. This implies that a modification to the exchange process of SDB to include more explorative properties could drastically improve its applicability to the other case studies.

6. Conclusions and Future Work

This work presented a multi-objective genetic algorithm which attempts to evolve candidate communities of the most highly influential genes related to a particular disease's presentation. This is achieved through aggregating several different types of biological evidence. Through these numerous relations, complex networks are formed. These networks are then investigated via centrality measures, with their values passed to the genetic algorithm as potential fitness objectives. As part of this work, three case studies were explored, namely, breast cancer, Parkinson's, and as part of a new contribution, Alzheimer's. In each case study, the most successful multi-objective setting combined the shortest path based measures, stress and betweenness, matching previous literature. The single-objective parameter settings also showed significant success, confirming the notion that despite the widespread use of multi-objective approaches, single-objective techniques still have their strengths. However, the weakness of Sum of Ranks could be due to the lack of diversity in populations due to ties in ranks. Future use should include the notion of a diversity penalty to combat this.

Significantly increased performance in both fold enrichment and sensitivity metrics were shown in the case of breast cancer and Parkinson's relative to past methodologies. Much of this success can be attributed to the inclusion of multiple evidence types in the data generation stage. This is echoed by the rise of weighted protein-protein interaction networks and various graph refining techniques used in recent studies.

In the case of Alzheimer's, the success of this work implies that the problem has sufficient structure for this technique and others of a similar nature to exploit in the future. A complication with this case study is the lack of available comparative tools (e.g. Endeavour) using similar known gene setups. Future comparisons to this case study should potentially include both the CD59 and CDH5 genes as they are already known Alzheimer's contributors. Future methods could also be less conservative when selecting the known LOO genes.

Additional future work should include the refinement of the safe dealer-based (SDB) crossover technique. On breast cancer and Parkinson's, the technique was shown to be extremely exploitative. When presented with the easier problem of Alzheimer's, due to the conservative known gene selection it was able to produce the best results while remaining very stable. By introducing more variation into the exchange portion of the operator, thus including more explorative properties, SDB crossover could be a worthwhile approach. As balancing exploration and exploitation is part of every evolutionary computation technique, it is also potentially worth trying different mutation techniques, selection techniques, and fitness methodologies in the future. This could also include the notion of switching techniques during the search procedure, e.g. switching to more explorative techniques if the population begins to converge. Another consideration would be to vary the elitism parameter to more carefully study selection pressure in both crossover techniques. To compare these numerous potential settings, more rigorous statistical techniques such as ANOVA testing should be used.

Another potential expansion would be adaptation of the technique to use weighted networks. This would also include study of various new path based centrality measures. As this changes the overall structure of the input networks, it may be beneficial to further study the properties of the graphs created by the generation process.

Despite the overall success of this methodology it should be noted that it does not automatically generalize to all diseases, much like one evolutionary computation technique does not generalize to all problems. However, the approach has been crafted such that future application of the methodology requires changes only to the known disease genes, and as such the data generation procedure.

Lastly, each predicted gene from all case studies indicate potentially interesting future directions for biologists to investigate as part of further understanding of deadly diseases.

Acknowledgements

The authors would like to thank Daniel Ashlock for helpful discussions, especially with respect to SDB Crossover. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537, 2006.

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

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, pages 1–4. Springer, 2009.

[4] P.J. Bentley and J.P. Wakefield. Finding acceptable solutions in the pareto-optimal range using multiobjective genetic algorithms. In *Soft Computing in Engineering Design and Manufacturing*. Springer Verlag, 1997.

[5] Michele Benzi and Christine Klymko. On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, 36(2):686–706, 2015.

[6] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[7] Stephen P Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[8] Alastair Burns, Robin Jacoby, and Raymond Levy. Psychiatric phenomena in alzheimer's disease. iv: Disorders of behaviour. *The British Journal of Psychiatry*, 157(1):86–94, 1990.

[9] 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.

[10] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, et al. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10):2366, 2007.

[11] D. Corne and J. Knowles. Techniques for highly multiobjective optimisation: Some nondominated points are better than others. In *Proc. GECCO 2007*, pages 773–780. ACM Press, 2007.

[12] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[13] Christopher B Eckman, Nitin D Mehta, Richard Crook, Jordi Perez-tur, Guy Prihar, Eric Pfeiffer, Neill Graff-Radford, Paul Hinder, Debra Yager, Brenda Zenk, et al. A new pathogenic mutation in the app gene (i716v) increases the relative proportion of aβ42 (43). *Human Molecular Genetics*, 6(12):2087–2089, 1997.

[14] Nicola S Fearnhead, Jennifer L Wilding, Bruce Winney, Susan Tonks, Sylvia Bartlett, David C Bicknell, Ian PM Tomlinson, Neil J McC Mortensen, and Walter F Bodmer. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proceedings of the National Academy of Sciences*, 101(45):15992–15997, 2004.

[15] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[16] Asma Ghafoor, Ahmedin Jemal, Elizabeth Ward, Vilma Cokkinides, Robert Smith, and Michael Thun. Trends in breast cancer by race and ethnicity. *CA: A Cancer Journal for Clinicians*, 53(6):342–355, 2003.

[17] Jesse Gillis and Paul Pavlidis. "guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3):e1002444, 2012.

[18] 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.

[19] 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.

[20] Ashkan Entezari Heravi and Sheridan Houghten. A methodology for disease gene association using centrality measures. In *Evolutionary Computation (CEC), 2016 IEEE Congress on*, pages 24–31. IEEE, 2016.

[21] Ashkan Entezari Heravi, Koosha Tahmasebipour, and Sheridan Houghten. Evolutionary computation for disease gene association. In *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–8. IEEE, 2015.

[22] Andrew J Hughes, Susan E Daniel, Linda Kilford, and Andrew J Lees. Accuracy of clinical diagnosis of idiopathic parkinson's disease: a clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(3):181–184, 1992.

[23] Woochang Hwang, Taehyong Kim, Murali Ramanathan, and Aidong Zhang. Bridging centrality: graph mining from element level to group level. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 336–344. ACM, 2008.

[24] 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.

[25] Duc-Hau Le and Van-Huy Pham. Hgpec: a cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network. *BMC Systems Biology*, 11(1):61, 2017.

[26] Seunghwan Lim, Byoung Kwon Yoo, Hae-Suk Kim, Hannah L Gilmore, Yonghun Lee, Hyun-pil Lee, Seong-Jin Kim, John Letterio, and Hyounggon Lee. Amyloid-β precursor protein promotes cell proliferation and motility of advanced breast cancer. *BMC Cancer*, 14(1):928, 2014.

[27] Dayou Liu, Di Jin, Carlos Baquero, Dongxiao He, Bo Yang, and Qiangyuan Yu. Genetic algorithm with a local search strategy for discovering communities in complex networks. *International Journal of Computational Intelligence Systems*, 6(2):354–369, 2013.

[28] Artem Lysenko, Keith Anthony Boroevich, and Tatsuhiko Tsunoda. Arete–candidate gene prioritization using biological network topology with additional evidence types. *BioData Mining*, 10(1):22, 2017.

[29] B Mann, M Gelos, A Siedow, ML Hanski, A Gratchev, M Ilyas, WF Bodmer, MP Moyer, EO Riecken, HJ Buhr, et al. Target genes of β-catenin–t cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. *Proceedings of the National Academy of Sciences*, 96(4):1603–1608, 1999.

[30] James D Mills, Thomas Nalpathamkalam, Heidi IL Jacobs, Caroline Janitz, Daniele Merico, Pingzhao Hu, and Michael Janitz. Rna-seq analysis of the parietal cortex in alzheimer's disease reveals alternatively spliced isoforms related to lipid metabolism. *Neuroscience Letters*, 536:90–95, 2013.

[31] Joyce A Mitchell, Jane Fun, and Alexa T McCray. Design of genetics home reference: a new nlm consumer health resource. *Journal of the American Medical Informatics Association*, 11(6):439–447, 2004.

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

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

[34] YoungJa Park and ManSuk Song. A genetic algorithm for clustering problems. In *Proceedings of the Third Annual Conference on Genetic Programming*, volume 1998, pages 568–575, 1998.

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

[36] Clara Pizzuti. Community detection in social networks with genetic algorithms. In *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation (GECCO)*, pages 1137–1138. ACM, 2008.

[37] Mihael H Polymeropoulos, Christian Lavedan, Elisabeth Leroy, Susan E Ide, Anindya Dehejia, Amalia Dutra, Brian Pike, Holly Root, Jeffrey Rubenstein, Rebecca Boyer, et al. Mutation in the α-synuclein gene identified in families with parkinson's disease. *science*, 276(5321):2045–2047, 1997.

[38] 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.

[39] NK Sakthivel, NP Gopalan, and S Subasree. A comparative study and analysis of dna sequence classifiers for predicting human diseases. In *Proceedings of the International Conference on Informatics and Analytics*, page 107. ACM, 2016.

[40] Giovanni Scardoni, Michele Petterlini, and Carlo Laudanna. Analyzing biological network parameters with centiscape. *Bioinformatics*, 25(21):2857–2859, 2009.

[41] Chuan Shi, Zhenyu Yan, Yanan Cai, and Bin Wu. Multi-objective community detection in complex networks. *Applied Soft Computing*, 12(2):850–859, 2012.

[42] Markus Storvik, Marie-Jeanne Arguel, Sandra Schmieder, Audrey Delerue-Audegond, Qin Li, Chuan Qin, Anne Vital, Bernard Bioulac, Christian E Gross, Garry Wong, et al. Genes regulated in mptp-treated macaques and human parkinson's disease suggest a common signature in prefrontal cortex. *Neurobiology of Disease*, 38(3):386–394, 2010.

[43] Koosha Tahmasebipour and Sheridan Houghten. Disease-gene association using a genetic algorithm. In *2014 IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 191–197. IEEE, 2014.

[44] Dane Taylor, Sean A Myers, Aaron Clauset, Mason A Porter, and Peter J Mucha. Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574, 2017.

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

[46] 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.

[47] 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.

[48] Chao Wu, Jun Zhu, and Xuegong Zhang. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13(1):182, 2012.

[49] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(1):189, 2008.

[50] Li-Bang Yang, Rena Li, Seppo Meri, Joseph Rogers, and Yong Shen. Deficiency of complement defense protein cd59 may contribute to neurodegeneration in alzheimer's disease. *Journal of Neuroscience*, 20(20):7505–7509, 2000.

[51] Eckart Zitzler, Marco Laumanns, and Lothar Thiele. Spea2: Improving the strength pareto evolutionary algorithm. *TIK-Report*, 103, 2001.