A Centrality Based Multi-Objective Disease-Gene Association Approach Using Genetic Algorithms

Tyler K. Collins

Submitted in partial fulfillment
of the requirements for the degree of

Master of Science

Department of Computer Science
Faculty of Mathematics and Science
Brock University
St. Catharines, Ontario

Abstract

The Disease Gene Association Problem (DGAP) is a bioinformatics problem in which genes are ranked with respect to how involved they are in the presentation of a particular disease. Previous approaches have shown the strength of both Monte Carlo and evolutionary computation (EC) based techniques. Typically these past approaches improve ranking measures, develop new gene relation definitions, or implement more complex EC systems.

This thesis presents 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. In an effort to explore the effectiveness of the technique compared to past work, multiple objective settings and different EC parameters are studied including the development of a new exchange methodology, safe dealer-based (SDB) crossover. Successful results with respect to breast cancer and Parkinson's disease compared to previous EC techniques and popular known databases are shown. In addition, the newly developed methodology is also successfully applied to Alzheimers, further demonstrating the flexibility of the technique.

Across all three cases studies the strongest results were produced by the shortest path-based measures stress and betweenness in a single objective parameter setting. When used in conjunction in a multi-objective environment, competitive results were also obtained but fell short of the single objective settings studied as part of this work. Lastly, while SDB crossover fell short of expectations on breast cancer and Parkinson's, it achieved the best results when applied to Alzheimers, illustrating the potential of the technique for future study.

Acknowledgements

I would like to begin by thanking my supervisor Dr. Sheridan Houghten for being consistently awesome throughout my entire time at Brock University, as well as understanding my need to take a break and play video games sometimes.

I would also like to thank all of the professors and staff in the computer science department at Brock University for putting up with me all of these years. Everyone taught me so much, and I will always be grateful.

A big thank you to Dr. Dan Ashlock (University of Guelph) for his help in formulating the early stages of Safe Dealer-Based crossover.

I would also like to thank Dave Bockus, for not only doing an excellent job of teaching me the fundamentals in this discipline, but for being an even better friend. One day we will get the guild back together.

Thanks to Adam Tonet for also being a great friend, as well as for being someone who I could rant to when stress levels got too high. One day you will pay me back for this crazy adventure, I am sure.

Thank you to Nik Unger, who got me started on computer science. Your infinite patience teaching me when I started is something I will never forget.

I would like to thank James Desjardins for being a great boss, as well as a patient boss while I finished this thesis. Your support these past few months has been invaluable to me.

Thank you to my Mother, Father, Omi, as well as Nana and Grandpa. I could not ask for a more supportive family. Everyone has always been so patient and understanding while I worked hard on this thesis. I hope I can always make you guys proud. This would not have been at all possible without each of you.

Lastly, I would like to thank Sammi Laffin. Your love and support made things feel possible when I was struggling the most. Thank you again, from the bottom of my heart.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Understanding the link between genes and how they pertain to the presentation of various diseases is important in extending the life expectancy of humans. This problem is primarily referred to as the Disease Gene Association Problem. Information gathered as part of this problem is often used in two ways. Firstly, it is used for *gene prioritization*, the ranking of known problematic genes based on their contribution to the presentation of a disease in hopes of cementing known disease causing genes. Secondly, it is used to identify new genes for future study that have yet to be investigated in relation to a given disease, in the hopes of gaining further insight into how the disease functions. This second use is based on the principle of *guilt by association*, which states that genes that are found to be highly interacting with already known contributors are also likely to be "guilty". For example, in the case of breast cancer, it is known that BRCA1 is one of the strongest contributors to the presentation of the disease. Genes that are found to be frequently interacting with BRCA1 (through, for example, protein-protein interaction networks) should thus be considered for future study in relation to breast cancer.

A significant challenge of the Disease Gene Association Problem is that the modelling of relationships between genes can often contain hundreds of thousands of interactions. As such, presenting this data in a manner suitable for study often becomes a problem for computer scientists to tackle.

Recently, there have been efforts to tackle the Disease Gene Association Problem using artificial intelligence techniques from the field of computer science. In addition to this style of approach, biologists often take the route of developing new ways to relate genes together, refining previous relations, or applying techniques to new diseases. As such, it stands to reason that a methodology that combines the recent advances of both of these styles of approaches is a strong candidate for improving on

current techniques.

The goal of this thesis is to improve upon past techniques and their application to breast cancer and Parkinson's disease. Additionally, the methodology developed in this thesis will be applied to Alzheimer's disease in order to broaden the number of diseases studied.

1.1 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 introduces necessary computer science background, including formal definitions for graph theory, approximation algorithms and evolutionary computation. Chapter 3 outlines the biological principles that make up the Disease Gene Association Problem, as well as defining several biological evidence types relating genes together. This chapter also describes three different approaches to scoring the effectiveness of a disease gene association problem methodology. Chapter 4 reviews three families of literature related to this topic. Chapter 5 defines the specific details of the methodology presented in this work as well as data generation procedures. Additionally, as part of this work, three case studies are investigated. Chapter 6 contains the methodology applied to breast cancer, Chapter 7 explores Parkinson's disease, while Chapter 8 contains the case study of Alzheimer's disease. Lastly, Chapter 9 concludes this thesis and provides avenues for future work.

Chapter 2

Computer Science Background

Bioinformatics and specifically the Disease Gene Association Problem (DGAP) requires a broad knowledge of fundamental computer science concepts. This is due to the complexity required to model the relationships in the input data of the DGAP. As such, this chapter will introduce the fundamentals of graph theory and algorithms to build a knowledge base sufficient for modeling and studying the DGAP as interaction between genes.

2.1 Graph Theory: A Formal Definition

A graph G is typically defined as $G = (V, E)$, where V is a set of *vertices* (also known as *nodes*) and E is a set of pairs of vertices implying some sort of relationship between the two nodes. This relationship in terms of graphs is often called an *edge*. Edges come with two implicit properties, *weight* and *direction*. The weight of an edge is a numerical value representing the cost to traverse that edge. For example, consider cities as nodes connected by highways as edges. Toll-free highways would possess a weight of zero, while toll based highways would have a weight matching their cost. A graph with weights for each edge is called *weighted*, while a graph without weights is called *unweighted*; in this case, all edges have the same cost or weight. Continuing with the example of cities connected by highways, *directed* edges can be thought of as one-way streets while undirected edges are two way streets. Graphs are directed if all edges are directed otherwise they are *undirected*. Mathematically this is represented by the notion of a particular relation between nodes being symmetric or not. Figure 2.1 contains an example of a directed and weighted graph, while Figure 2.2 contains an example of an undirected and unweighted graph.

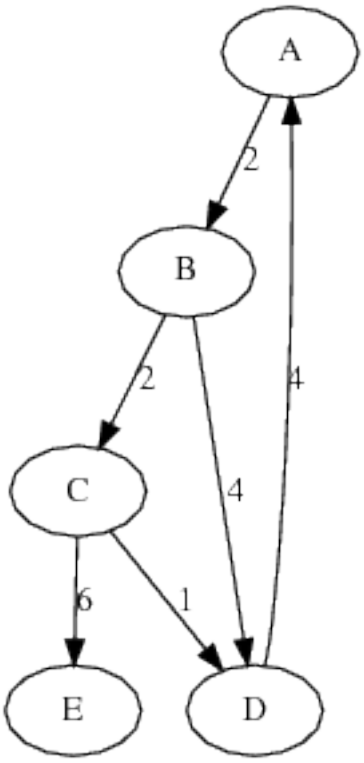

Figure 2.1: An Example of a directed and weighted graph.

2.1.1 Topological Features of Graphs

In addition to the properties defined in the previous section, graphs also include several different topological properties. This section will list and define several other fundamental graph properties and further topological features.

- The *degree* of a node is the number of connections a node possesses. In directed graphs, this is often separated into in-degree and out-degree.
- The *neighbour* of a node V is any other node that is directly connected to V via an edge. For the purpose of this thesis, $N(V)$ will be defined as the set of nodes neighbouring V.
- A *traversal* (also known as a *path*) is an ordered list of nodes that implies visiting each node in the list via edge connections acorrding to the order specified by the list.
- The *shortest path* is defined as a traversal through the graph between any two given nodes such that the path cost is minimal. For the purpose of this thesis, the shortest path between two nodes will be defined as $dist(u, v)$. For example,

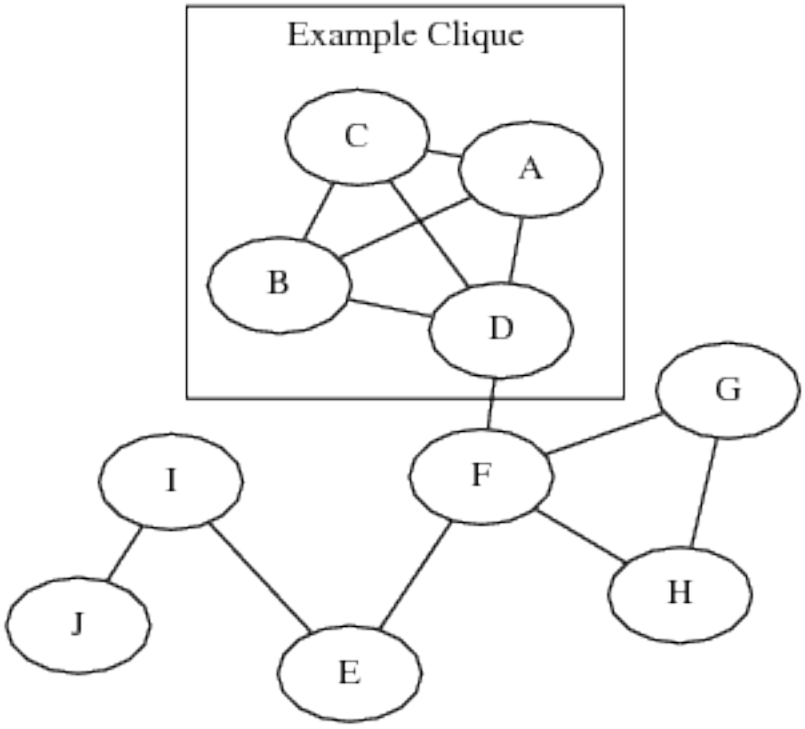

Figure 2.2: An Example of an undirected and unweighted graph with an example clique.

in Figure 2.1, $dist(A, D)$ is equal to five. This is achieved through the traversal A, B, C, D.

- The *diameter* of a graph is defined as the maximal shortest path. In the case of Figure 2.2, this is represented by $dist(A, J) = 6$ and the traversal *J, I, E, F, D, A*.

- A graph *cut* is a particular partition of the nodes in a graph that separates vertices into two disjoint subsets. A cut can be made in Figure 2.2 between the nodes D and F to separate the graph into the subsets of vertices $\{A,\ B,\ C,\ D\}$ and $\{E,\ F,\ G,\ H,\ I,\ J\}$.

- Graphs that are *complete* possess edges between every pair of nodes.

- A *clique* is a subset of nodes in a graph which, with respect to the group, form a complete graph as shown in Figure 2.2 with the bounded box.

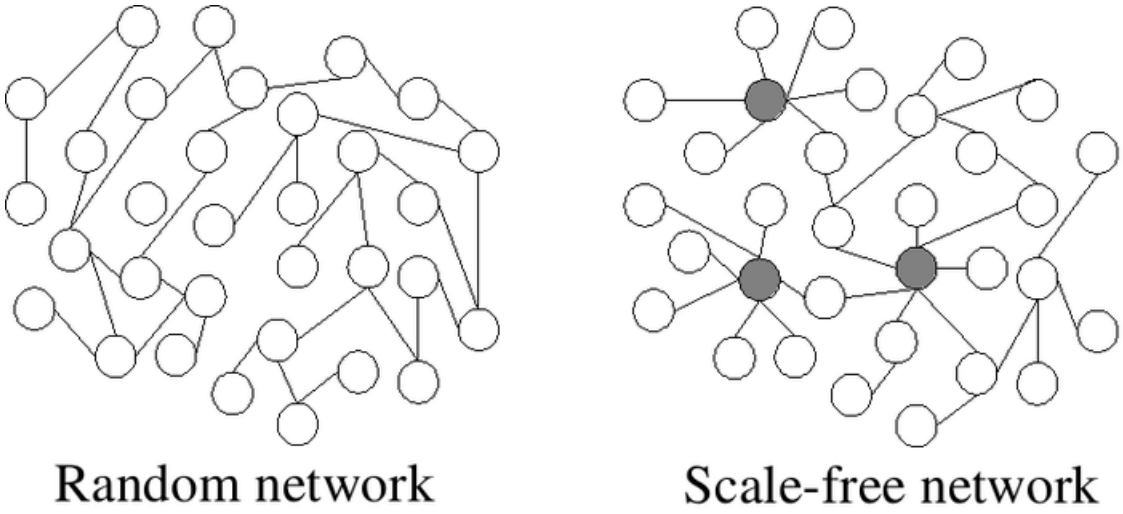

Figure 2.3: Visual comparison of a random network versus a scale-free network. Image taken from [11].

2.1.2 Network Types

Building on the notion of topological features, there exist families of graphs often also called *networks*. Belonging to these families requires adhering to several additional definitions and constraints [7]. The trade-off, however, is that several additional assumptions and properties about the network become true. Often, modeling real world data tends towards falling into these families. For examples, road networks in cities, maps of the Internet, social interaction networks, epidemic modeling and other biologically inspired graphs all fall into the category of complex networks [9][27]. The following sections outline three network types studied in current literature, namely, random graphs, scale-free networks, and complex networks. Figure 2.3 contains a visual example of a random network versus a scale-free network.

Random Graphs

Random graphs are often used as benchmarking tools for testing new approaches to solving graph problems. They are created by generating a list of n nodes, and iteratively adding edges following a particular sampling distribution with respect to n. The goal of this approach is to iterate until certain graph properties such as community structures become true or are discovered [8].

Complex Networks

A complex network is most commonly defined as a network in which there are non-trivial topological features. However, this definition is not entirely strong. Additionally, this definition does not exclude the appearance of some of the features introduced in Section 2.1.1. It does however imply that they are perhaps difficult both visually and computationally to discover. This behaviour often presents itself most heavily when the data used to generate a graph is real-world inspired [47].

Complex networks differ from random graphs in the sense that they are not necessarily governed by a probability function, and cannot be inspected at any point during the creation process. As such, it can be difficult to extract information and meaning from them. Difficulty aside, complex networks still tend to have inherent properties. The most notable of these properties is that complex networks tend to form small highly interconnected areas known as communities [60].

Scale-Free Networks

An extension to the notion of complex networks is that of scale-free networks. Scale-free networks inherit all of the implicit properties of a complex network but also add a new constraint. Namely, that the connectivity of each node in a scale-free network follows a power law based distribution. This implies that there is no function that neatly maps to the sorted distribution of degree of nodes in the graph. Formally, the probability distribution of the number of edges originating from a node k is defined as $P(k) \sim k^{-\gamma}$, where γ is taken to be a constant $2 < \gamma < 3$ [54].

2.1.3 Centrality Measures

As stated in the previous section, it is difficult to extract meaning from complex networks. Due to this, researchers studying complex networks have introduced and defined the notion of *centrality measures*. Centrality measures provide a measure indicating how important or central a node is in a graph. These measures allow for meaning to be extracted from extremely large networks and afford the ability of ranking of nodes by relative importance. Measures vary not only in computation, but also in their approach. For example, one measure may reward a node with a higher score for being isolated, while another favours nodes with higher degree. These measures may of course be adapted and combined, as what provides an effective means of measuring centrality in one graph problem, may not generalize to others.

An additional important consideration when working with centrality measures applied to large networks is that of pre-computation. If the network is to stay static throughout the duration of an experiment, computation time can often be exponentially reduced by pre-computing network statistics and centrality measures. This is especially true for the more computationally complex measures. A further optimization could include not loading the graph as a data structure at all, and instead simply loading attributes of node centrality. In the case of representing a graph as a matrix, this can reduce the memory requirements significantly.

The following subsections will provide definitions of centrality measures and the properties used to compute them based on [62]. However, Section 2.1.12 and the definition of bridging is instead based on [34]. 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], [67]).

2.1.4 Degree

As defined previously, degree is the number of nodes in direct connection to a given node n. This can be broken down into two forms. In-degree is the number of incoming connections to node n, while out-degree is the number of outgoing connection from node n. The sum of these two numbers is the total degree.

A high degree can imply that a node is highly connected in a network, thus inferring that it is central and important. Conversely, a node with low degree can be thought of as less important and a node on the fringes of a network.

2.1.5 Eigenvector

Eigenvector is a recursive measure of the reachability of a node in a graph. A Node n with a high eigenvector will have many neighbours. These neighbours will also have many connections, thus contributing to the overall score of n. If n were to have a low eigenvector, its local neighbourhood would be sparsely connected with few overall nodes. Therefore, if a node has a high eigenvector relative to the rest of the network, it can be thought of as a central node.

The following recursive equation is used to compute eigenvector for a given node, 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.

$$Eig(v) = \frac{1}{\lambda} \sum_{w \epsilon N(v)} Eig(w) \tag{2.1}$$

2.1.6 Closeness

The closeness of a node n is computed by finding the shortest path from n 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.

It follows simply that nodes with high closeness are central to a graph as it takes them less time to reach all other nodes in the graph on average. Conversely nodes that must travel further would be taken to be less central under this measure.

2.1.7 Eccentricity

Eccentricity is much like closeness, as it takes a given node n and computes shortest paths from n to all other nodes in the graph. Next, instead of summation, the maximum distance path is found, and the reciprocal value of this distance is the eccentricity value.

Eccentricity can be a difficult measure to work with. Consider the case of an extremely central node that has one node far from it. The resulting eccentricity value would then be overall very low. To combat these problems eccentricity should only be used when the topology of the graph can be guaranteed to be fairly regular and without trails of nodes branching off from the main body of the graph.

2.1.8 Radiality

Radiality is a centrality measure computed respective to the diameter (Δ_G) of a graph. As such it is defined as follows:

$$C_r(v) = \frac{\sum_{w \epsilon N(v)} (\Delta_G + 1 - dist(v, w))}{n - 1} \tag{2.2}$$

Subtracting shortest paths from the longest shortest path, this centrality measure will return a high value if the node is involved in shortest path traversals.

2.1.9 Centroid Value

The centroid value of a node in a network is much like eigenvector, that is, the measure is highly influenced by neighbours that are more than one edge away. This increases

the complexity drastically. Centroid value is defined as follows:

$$C_c(v) = min\{f(v,w) : w\epsilon N(v)\} \tag{2.3}$$

where

$$f(v,w) = \gamma_v(w) - \gamma_w(v) \tag{2.4}$$

and $\gamma_v(w)$ refers to the number of nodes in the graph that are closer in distance to v than w.

Note that similar to eccentricity, centroid value must be compared to the overall average to obtain a meaningful interpretation. A node with a comparatively high centroid value relative to the network can be interpreted as a central node.

2.1.10 Stress

The stress of a given node n in a graph is computed by referencing all shortest paths in the network and counting how many of these paths contain the node n.

Stress differs from the previous measures as it does not necessarily reward a higher value based on connectivity relative to the graph. It instead rewards higher score based on how "travelled" the edges are. In a complex network based on communication, stress can reward nodes that connect highly dense areas such as cliques. Note that comparison to the average stress value of the network as a benchmark can provide a more accurate comparison between nodes.

2.1.11 Betweenness

Betweenness functions much like stress, but scales the number of shortest paths that pass through a given node n. 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 n by all other shortest paths of equal length. Thus if betweenness returns a high value, not only is the node providing an important connection, but it is a critical non-redundant connection.

2.1.12 Bridging

This centrality measure also functions like stress as it rewards nodes that occur in shortest paths rather than direct connectivity. However, it differs vastly in computation. Simply put, 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 a high degree typically exist in cliques, and bridging as a measure rewards nodes that "bridge" gaps between highly connected areas. By minimizing this measure, it is possible to find the opposite. That is, minimizing bridging searches for nodes that are part of highly connected areas that are non-bridging.

2.1.13 Limitations of Centrality Measures

Though centrality measures provide an effective way paired with pre-computation to analyze complex networks, they are not without flaws. In the previous section, eccentricity, centroid value, and stress all required reference to the overall average values of the network to be effective as raw comparison between two numerical values. With this in mind, recall that not every centrality measure or combination of measures will work for all graph types. When searching for important nodes in a telecommunication network, bridging may be useful. This may not be the case for a biologically generated network. These two types of networks often differ in both size and connectivity [27]. Individual measures are also being iterated on and new functions are being created all the time. Choosing correct measures for complex network analysis is eventually reduced to empirical study.

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. This problem is illustrated by the question of if a node v in a network has the highest bridging value, and a node w has the second most bridging, how much less "bridging" is it in comparison? What would that mean? This problem becomes exacerbated at low scores [9]. The meaning is easy to discover when working with something as simple as degree for example, but when looking at centroid value it is not so clear.

On a final note, most if not all centrality measures are related. That is not to say that these measures are ineffective, instead it is the case that most centrality rankings are based on neighbours and as such are highly influenced by a node's degree. All the measures discussed in this report except for stress and betweenness take into account neighbours in some way. However, stress as a measure is not without its own problems as outlined in the definition of betweenness.

2.1.14 Community Detection

Often, real world data tends to have several highly connected nodes (*hubs*) while the remaining system is sparsely interconnected in comparison [69]. In terms of the cities and roads in North America as an example, hubs could be capital cities or suburbs and sparse regions could be inhospitable areas. While finding large cities on an unfamiliar map may seem trivial and unimportant, consider telecommunication networks. In these systems hubs may be areas of the network under significant stress and in need of infrastructure upgrades and failsafe mechanisms.

However, if these hubs are so important, how does automatic detection work? 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 [69]. 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 [69]. A proof sketch of this can be found in Fortunato's work in [22] where community detection is reduced to that of a search for a maximally "weak-clique" based on the notion that a search for a maximal standard clique is already known to be NP-complete [38]. Currently popular methodologies in the field are discussed in Section 4.3 of this thesis.

2.2 Algorithms

The study of algorithms is a fundamental part of computer science. Optimality of these algorithms is then of course a further focus. However, it is known that there exists problems that for large enough inputs, can take lifetimes to solve optimally, specifically, the set of problems that are known as NP-Complete [23]. To combat this, there is a class of methodologies known as *approximation algorithms*. These algorithms attempt to create solutions that are nearly optimal, but do not necessarily result in solutions that are actually optimal. Approximation algorithms come in many forms, but this section will focus on *artificial intelligence* (AI) techniques, specifically those inspired by biological principles.

2.2.1 Random Search

While not biologically inspired, random search is a fundamental theoretical starting point for understanding approximation algorithms. Implementation is done by randomly generating numerous candidate solutions and keeping track of the best found.

Note that this does not attempt to use regression, learn, or exploit any portion of the solution landscape. Thus this technique is most often used to measure the minimal results necessary to be achieved by an approach to be accepted as useful. However, this benchmarking tool begins to fall apart when as part of a problem definition, a relative ranking between candidates must be produced, i.e. to determine how much better one item is in comparison to another.

2.2.2 Principles of Evolutionary Computation

Evolutionary computation (EC) is a subfield of AI in which populations of candidate solutions are permuted subject to selection inspired by biological principles such as Darwinian evolution [32]. The effectiveness of these candidate solutions is typically measured by a heuristic often referred to as a *fitness function* (see Section 2.2.8). This differs from the typical definition of biological fitness which is defined by how much a given individual influences the next generation. Instead, fitness as a heuristic represents how successful a candidate is at solving its respective problem. As such, the motivation of Darwinian evolution in the context of EC is to exploit the structure of successful candidates and distribute beneficial traits throughout the population until genetic diversity becomes stale. This final state is referred to as *convergence*.

Early techniques in this field include work by Fogel [21] and Schwefel [6], who developed *evolutionary programming* and *evolutionary strategies* respectively. EC quickly expanded to include further techniques such as, particle swarm optimization [39], ant colony optimization [17], self-organizing maps [40], and more. In general, EC techniques possess several common properties and are listed below in no particular order.

- Population of candidates
- Biologically inspired exchange or reproduction property
- Fitness or heuristic measure of success
- Convergence criteria: success level or maximum iterations reached

2.2.3 Genetic Algorithms

In addition to the early techniques listed in the previous section, Holland popularized a technique he named the *genetic algorithm* (GA) [32]. GAs have been shown to be

C1	1	0	1	0	0	1	0	1	1	0
C2	1	1	1	1	1	1	1	1	1	1

Table 2.1: Examples individuals for the Max Ones problem of size ten. C1 is a random candidate solution, while C2 is the optimal solution.

able to solve numerous challenging problems and remain an active area of study in the field.

The structure of a standard GA and associated definitions are listed below, while further specification can be found in the coming sections.

1. Initialization: Generate an initial population of candidate solutions. Typically taken to be random unless specified otherwise.

2. Fitness: Score each individual in the population in accordance with the fitness function.

3. Stopping criteria: Determine if the best solution, a sufficient solution, or other stopping criterion has been met. Otherwise, continue.

4. Selection: Randomly select pairs of individuals to undergo *crossover* and *mutation* operations.

5. New Population: Repeat step 4 until an entirely new population of individuals has been made of the same size.

6. Continue: Go to step 2.

2.2.4 Individuals

Individuals in a GA are candidate solutions to the problem definition. These candidates have an internal data structure representation known as a *chromosome*. The set of individuals being evolved in a GA are known as the *population*. The size of this set, and the amount of evolutionary *epochs* (iterations of the algorithm) are both parameters of the GA methodology. Table 2.1 contains two example individuals of chromosomes for the Max Ones problem of size ten. The Max Ones problem attempts to take a binary string (of a given variable size) containing zeros and ones and generate an individual containing only ones.

C1	1	0	1	**0**	0	1	0	1	1	0
C2	0	1	0	**1**	0	1	1	1	1	1

Table 2.2: Example of initial state of one-point crossover applied to the Max Ones problem. See Table 2.3 for finished state. Emphasized entry is defined as the start position of the exchange.

C1'	1	0	1	**1**	**0**	**1**	**1**	**1**	**1**	**1**
C2'	0	1	0	**0**	**0**	**1**	**0**	**1**	**1**	**0**

Table 2.3: Example result of one-point crossover applied to the Max Ones problem. Emphasized entries represent new genetic material obtained as result of crossover.

2.2.5 Selection

Selection is the process by which the GA rewards strong individuals for possessing successful "genetic material". A standard approach to this segment of a GA is that of *tournament selection*. This begins by randomly selecting k individuals from the population. Next, the strongest of this subset is selected, as measured by which individual has the strongest fitness value. This individual can be said to have won the "tournament", and often proceeds to exchange information with another tournament winner. Typically k is kept quite small, i.e. three or four. This is due to the fact that increasing k greatly increases the selection pressure in the population but can quickly lower the diversity of genetic material. As such, as k grows too large, a GA's results may fall into a local optimum instead of a global optimum.

2.2.6 Crossover

Crossover refers to the process that allows individuals to distribute genetic material throughout the population via an exchange process. These new individuals, called *children*, make up the next generation of candidate solutions. Since crossover is an operator that is highly representation dependent, studies often define their own custom crossover operators. Further, the rate at which individuals in the population undergo crossover is subject to choice via parameter at run time. Tables 2.2 and 2.3 contain an example of one-point crossover on the previously defined Max Ones problem. One-point crossover in the case of the Max Ones problem is accomplished by taking two individuals, selecting a random index, and swapping all genetic material between the two chromosomes after that point.

C1	1	0	1	0	**0**	1	0	1	1	0
C1'	1	0	1	0	**1**	1	0	1	1	0

Table 2.4: Example of single-bit mutation applied to the Max Ones problem. C1 is the original chromosome, while C1' is the result of a beneficial mutation.

2.2.7 Mutation

Mutation is an operator which attempts to duplicate the micro-level changes that happen upon creation of new individuals between generations. As such, the rate of mutation is a parameter to be defined at run time. Table 2.4 contains a worked example of single-bit mutation applied again to the Max Ones problem. Single-bit mutation in this case is defined as taking a single random index in the chromosome, and flipping the bit at that position.

2.2.8 Fitness Functions and Principles

As stated previously, fitness in relation to EC differs from biological fitness. This new notion of fitness only refers to how adept a chromosome is at solving a particular problem. Just as there exist numerous schemes for encoding problems into chromosome structures, there exists just as many fitness function choices. The choice of fitness function is often left to what is best able to describe the fitness landscape. A fitness landscape can be thought of as the possible solution space spanned by all chromosomes and their resulting fitness. Figure 2.4 contains a visual representation of a fitness landscape where the EC technique is trying to find global minima. The fitness for this particular problem would be distance of a given coordinate to the known minimum and thus minimizing that distance. This differs significantly from the notion of the Max Ones fitness function as in that case it is simply the summation of the chromosome being maximized.

Something these examples have in common is that they are optimizing a single value. This family of fitness functions is called single objective. It should come as no surprise that there exists another family of functions called multi-objective. Consider the example of a factory producing some product. Inside of the factory they would like to both maximize profit as well as worker safety. How should the success of an individual then be measured? An entry level technique is that of weighted sum. This in effect compresses the multi-objective problem back down to a single objective i.e. $\sum_{i=1}^{\infty} w_i o_i$ where i is the id of each objective, o_i is the raw value of an objective, and w_i is a weight value. The difficulty of this approach is that for every problem or

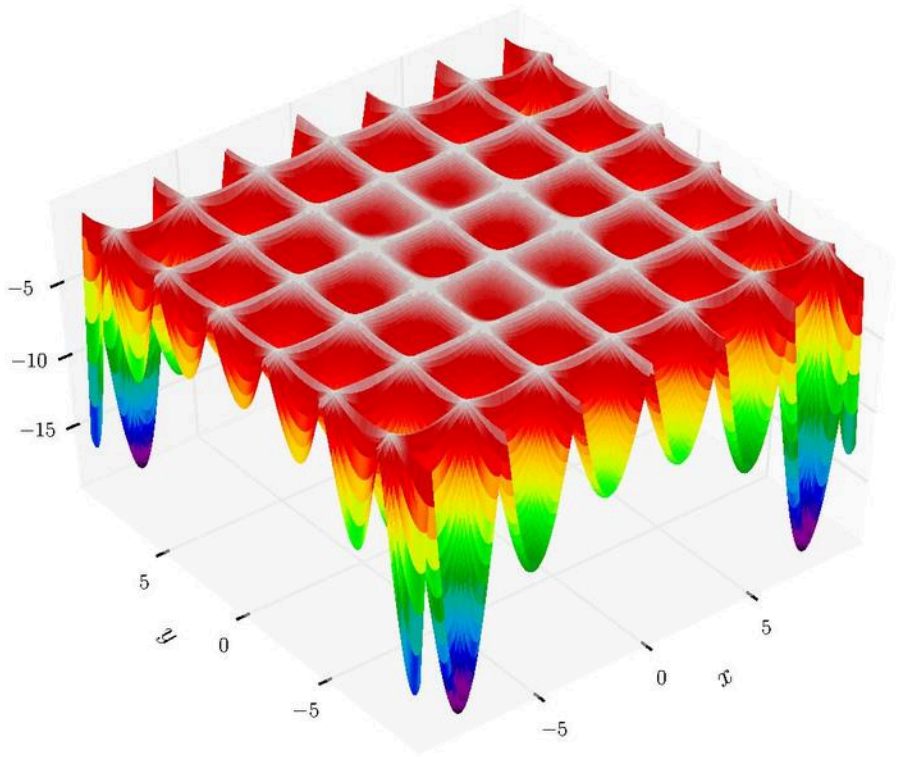

Figure 2.4: An Example of a fitness landscape: The Holder Table function [36]

parameter setting, new weight values must be examined empirically.

A second more advanced technique adapted from economic principles, is the idea of maintaining *Pareto fronts*. The motivation behind Pareto approaches is to keep a set of distinct solutions which for each item in the set, are "non-dominated" [16]. An individual x is said to be strictly dominated by an individual y for the following equation:

$$x \prec y \leftrightarrow \forall i(x_i \leq y_i) \wedge \exists i(x_i < y_i) \tag{2.5}$$

where x_i and y_i are various objectives of an optimization problem. As illustrated by Figure 2.5, Pareto fronts often contain multiple solutions. These solutions are then examined for exceptional individuals and can be passed to other researchers as candidate solutions for their own potential work.

Adaptation into a fitness technique can be done naively based on Algorithm 1. In this scheme, Pareto rank becomes the new fitness value for the purposes of selection operators. While an innovative solution, Pareto front management introduces large amounts of comparisons into fitness evaluations, increasing execution time. Diversity of a population also decreases as the number of objectives increases. This is due

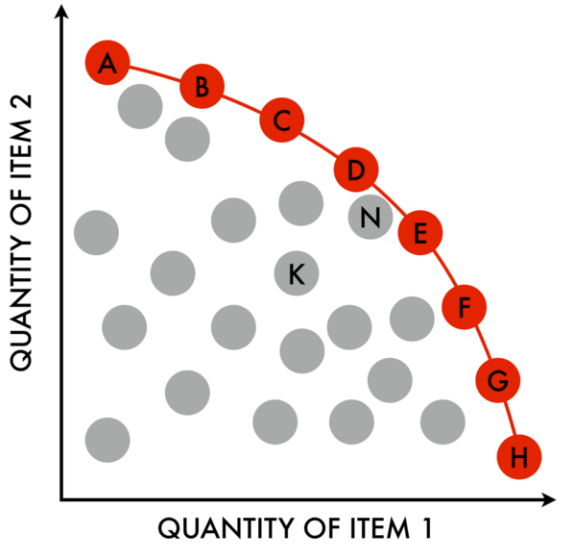

Figure 2.5: An Example of a Pareto front taken from [72]

to the quality of solutions in middle tier fronts fluctuating wildly. More advanced implementations that attempt to preserve diversity of solutions and reduce execution time are NSGA-II [16] and SPEA2 [76].

Algorithm 1 Naive Pareto Fitness Function

```
1: procedure PARETO(Pop)                    ▷ Returns a rank for each individual
2:     Rank ← 1
3:     while unranked individuals exist do
4:         for each unranked p_i ∈ Pop do
5:             if no p_j(j ≠ i) dominates p_i then
6:                 rank(p_i) ← Rank
7:         Rank ← Rank + 1
```

A third technique for multi-objective problems without the disadvantages of Pareto front management is Sum of Ranks, presented in Section 5.1.6 for use in this thesis.

2.2.9 Genetic Programming

Genetic Programming (GP) is an EC technique developed by Koza [41] which also exploits the principles of Darwinian evolution and natural selection. It differs from a GA in that the internal representation of individuals are typically tree-based and as such have adapted crossover and mutation operators associated with standard methods. These tree-based individuals can be thought of as functions or programs,

making GP suited for evolving solutions to problems with which a typical GA would struggle.

To illustrate these differences, consider the Artificial Ant problem [42]. In this problem, an environment containing a path with rewards is defined. Standard GAs at most could be expected to recover the path through evolution, whereas GP has been shown to be able to evolve a program which traverses areas systemically looking for rewards [42].

Chapter 3

Biological Background

As the Disease Gene Association Problem (DGAP) is a biologically inspired problem, the purpose of this chapter is to introduce the fundamental definitions required to understand the basics of genetics and how small changes in genetic information can lead to the presentation of malicious diseases. This chapter also gives a formal definition to the DGAP, its input data structure, and methods of measuring success of a methodology.

3.1 Genetics and DNA

Genetics is based upon the study of deoxyribonucleic acid (DNA). Of particular interest is how small changes in the overall structure of DNA can lead to such drastic changes in the health of cells and organisms [19]. DNA is made up of two long strings called *nucleotides*, which bond with each other and form a *double helix*. These four nucleotides are known as guanine, cytosine, adenine, and thymine, typically denoted by G, C, A, and T respectively. Bonds formed in the double helix model are formed between the pairs G,C and A,T due to the chemical structure of each nucleotide and its potential for hydrogen bonding. Figure 3.1 contains a visualization of the double helix model where the bridges between strands are the bonds between G, C, A, and T nucleotides.

3.2 Genes

The long strands of nucleotides inside of DNA can be thought of as a list of approximately three billion characters. However, some portions of this list are considered

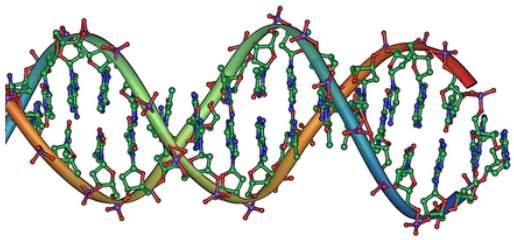

Figure 3.1: DNA Double Helix visualization taken from [37]

more important than others. Subsets of the DNA sequence that are actually responsible for distinct behaviours are known as *genes* and are typically inheritable [13]. Every cell in a human, for example, contains two complete sets of genes, each set found lying on one of the two strands of DNA. If one gene on one strand is somehow deactivated, damaged, or incapable of functioning, the other set of genes can potentially take over [28].

Despite this, what is it that makes these genes actually able to do anything? How do they represent functionality much less physical attributes? Essentially a gene can be thought of as a blueprint or algorithm for developing proteins (see Section 3.3) to accomplish various tasks. This process is defined as *gene expression*. Consider the case of something akin to the generation of a human nerve cell, muscle cell, or bone cell. Of the entirety of information contained in human DNA, these cells are "expressing" very little of the potential information at their disposal.

3.3 Proteins

Proteins in the context of DNA and genetics differ from the more typically known protein nutrient, in that a protein is a complex molecule which accomplishes low level tasks for cells as dictated by its respective gene. These tasks may include: providing cell structure, assisting in DNA replication, transport of lower level structures (atoms), communication between other proteins, and more.

3.4 Phenotypes and Disease Presentation

A *phenotype* is defined as the total set of visible and observable characteristics or traits of some organism. These individual traits result from at least one and sometimes many more genes being expressed. Some examples of phenotypes include blood type, hair colour, and for birds, nesting habits. This leads to the definition of a *high level phenotype* and a *low level phenotype*. Examples of high level phenotypes are hair colour, eye colour, and various behaviours. However, low level phenotypes can be properties of a cell. For example, these might include type or shape. In summary, phenotypes result from genes being expressed, which is an interaction of proteins.

Of particular interest to biologists is when genes are expressed and manifest into some sort of genetic disorder via a *single nucleotide polymorphism* (SNP). An SNP is when a single G, C, A, or T is permuted and replaced by another nucleotide. These changes can sometimes be far reaching and quite harmful, for example in the case of Parkinsons disease and the SNCA gene [64]. However, not all undesirable phenotypes are necessarily bad. An example of a possibly undesirable phenotype that is not particularly harmful is that of male pattern baldness.

3.5 Bioinformatics

Bioinformatics is a hybrid field of study which typically combines biologically inspired problems with large scale computer science algorithmic approaches. Contributions to the field of bioinformatics typically focus on the methodology surrounding the processing of large amounts of biological data. For example, shotgun sequencing of DNA takes short stands of DNA, and using substring matching (tiling), attempts to construct larger strings. This process is recursively repeated until a chosen length sequence is found. As an approach, shotgun sequencing was a fundamental part of the initial study of human genome data generation [35]. The following section introduces an additional bioinformatics problem that is the focus of this thesis.

3.6 The Disease Gene Association Problem

The Disease Gene Association Problem (DGAP) is a bioinformatics problem in which search strategies inspired by computer science are used to find and rank various genes based on their involvement in the presentation of a given disease. Results from the DGAP typically are of the form of either a ranking of each gene, representing how

likely they are to be involved in disease presentation, or a reduced subset of genes which are considered to be highly involved. The motivation for this result format is to be able to pass this ranked list or highly involved subset to a biologist for focused study on specific genes. Use of this list would help to provide an ordering by priority, so that the deadliest genes are studied most rapidly.

Typically, known information about the disease being studied is included in the computational approach and is used in hope of increasing the accuracy of methodologies. The following two subsections are extended assumptions to the DGAP problem and serve as known principles that can optimize DGAP search strategies.

3.6.1 Modularity Principle

The Modularity principle states that there does not exist a one-to-one relationship between genes and genetic diseases [55]. Take for example the real world gene BRCA1. This gene is frequently present in cases of breast cancer. However, with this gene highly interacting on the protein level with other genes is not 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 deadly.

3.6.2 Guilt-by-Association Principle

Consider the other genes interacting with BRCA1 in the previous example. If it is found that the interaction between this 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 [26]. Strong search strategies should take advantage of this and examine frequent gene interactions.

3.7 DGAP Input Data

Unsurprisingly, the inputs to the DGAP are various gene interaction measures. However, even a small subset of interactions contains a large amount of relations between numerous genes. Representing all of these relations at once in a graph easily forms a complex network [2] (see Section 2.1.2). Interaction measures that can be used in the DGAP are explored in the following subsections. Note that all of these form complex networks.

3.7.1 Protein-Protein Interaction

Perhaps one of the most widely used and most effective evidence types for gene prioritization type problems is that of Protein-Protein Internation (PPI) networks [53],[12]. In this evidence type a graph is formed by considering each gene as a node, and physical protein interactions between genes as edges. Creation of these networks is often done via physically observing the movement of proteins by a method called *signal transduction*. Considering the previously discussed DGAP principle of guilt-by-association, PPI based methodologies allow for computational based techniques to search for areas of interest based on things such as density. Specific use of PPIs in the DGAP are explored in Chapter 4.

3.7.2 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 [57].

3.7.3 Phenotype

Phenotypical evidence is based on pairs of genes leading to the presentation of the same phenotypes, i.e. hair colour or blood type. 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 [57].

3.7.4 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 evidence type cane be though of as a low-level hybrid of both phenotypical and co-expression based relations.

3.7.5 Text Mining

Text mining 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 [29] via natural language methodologies. Of

particular note for text mining is that it is among the first large scale approaches taken for the DGAP. Modern approaches however, often use text mining in conjunction with the previously discussed evidence types.

3.8 Analysis and Benchmarking

As the success of a DGAP methodology is not directly related to the best relative "score" produced, various post processing and benchmarking techniques have been defined in the literature. The following subsections introduce three of the most common approaches.

3.8.1 Leave-One-Out Validation

Leave-One-Out (LOO) validation is the most common [43] approach to benchmarking and focuses on whether a methodology is able to recover known genes during execution. Formally, given N known disease related genes, fix *N-1* of these to be in all candidate solutions. A LOO validation test is defined to be successful if the left out gene is recovered by the method upon completion. This process is to be repeated for all N genes [43].

3.8.2 Fold Enrichment

Fold enrichment 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 [74]. A fold enrichment is counted as a success if the previously defined ratio is less than a defined threshold. This threshold often differs between studies and methodologies and is left up to the choice of researchers.

3.8.3 Receiver-Operating Characteristic

Receiver-Operator Characteristic (ROC) is a second extension to LOO analysis and defines a sensitivity measure 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. The following equation describes

the result of a ROC analysis:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.1}$$

This type of analysis is often studied in classifiers and presented in results to combat algorithms which are easily fooled.

Chapter 4

Literature Review

The purpose of this chapter is to provide a sample of the current study of the Disease Gene Association Problem (DGAP) and its associated literature. Section 4.1 discusses current biologically inspired techniques. Section 4.2 introduces a hybrid bioinformatics approach which seeks to merge community detection and typical biological approaches. Finally, Section 4.3 introduces two modern community detection approaches and discusses their relative strengths and weaknesses.

4.1 Biological Approaches

In previous biological studies of the DGAP and related problems, biologists have commonly taken the route of increasing the accuracy of input datasets. Biologists have also frequently applied methods to entirely new types of data sets, or attempted to automate the entire process via a construction of a pipeline. An example of this can be found in the paper by Lysenko *et al.* [48]. The system proposed in this paper attempts to improve the accuracy of results by including additional evidence types other than only protein-protein interactions. These additional evidence types include but are not limited to co-expression relations and phenotype relations. While results in the study by Lysenko *et al.* were favourable, the main technique used was that of a random walk algorithm. Random walk algorithms in computer science are known for being a first step in testing whether a fitness landscape has structure. As such, the authors do mention that while their technique was not necessarily computationally the strongest, that techniques in the future should include multiple evidence types to build as much relational evidence as possible.

A second study by Sakthivel *et al.* [61] includes a more advanced computer science technique than Lysenko *et al.* while remaining firmly biologically inspired. This

comparative study tests 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 did struggle compared to the other standard 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 to build from. The authors discuss that this relative difficulty can come from how highly studied the already known disease genes are. For example, breast cancer is more highly studied than generic cerebral vascular diseases. To combat this, HGPEC [44] is a recently developed application for Cytoscape (see Section 5.2.1) which seeks to refine the decision making process for choosing known genes on a disease by disease basis. Though 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.

Wu *et al.* [73] introduce an additional approach which seeks to refine PPI based evidence via aggregating large amounts of standard PPI data together and building a weighted network. This weighted network is 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 is reported not only in accuracy, but also stability. Due to this, the authors speculate that their approach to aggregation of PPI data into a weighted network is a notion that should be adopted by future techniques.

Additional biological literature review efforts not contained in this section can be found in the previous definition of the DGAP (Section 3.6), DGAP input data and their various evidence type definitions (Section 3.7), and DGAP benchmarking techniques (Section 3.8). Lastly, Table 5.4 contains several biological DGAP techniques that will be used to measure the effectiveness of the methodology proposed by this thesis.

4.2 Evolutionary Computation Approaches

In a study by Tahmasebipour and Houghten [66], the DGAP was adapted to an evolutionary computation (EC) approach. The technique used by the authors was that of a genetic algorithm (GA) based on the principles of community detection (Section 2.1.14) problems. As such, candidate solutions in the GA took the form of communities of potentially highly interacting genes. As part of the study both breast

cancer and Parkinson's disease were investigated. Input data for these case studies however only included PPI based evidence types. Favourable results were achieved in comparison to the popular disease gene databases found in Table 5.4.

As a continuation of the previous study, Heravi and Houghten present a genetic programming (GP) approach to the DGAP [30]. Additionally, the study adds two new aspects to the family of EC approaches. The first addition is the inclusion of multiple evidence types in the data generation stage of the DGAP, while the second is the introduction 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 decision 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 is also briefly explored in the study. The authors conclude that the measures of stress and betweenness when used in conjunction produced the strongest results. These results include a marked 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. It is important to note that both case studies performed by the authors did improve on popular known DGAP databases.

4.3 Community Detection Algorithms

As stated previously, there potentially exists merit in the EC based approach to the DGAP for the introduction of advanced community detection techniques. One such technique is that of the locus-based adjacency representation (LAR) [56]. This chromosome implementation focuses on the ability of the candidate solution to possess multiple communities, as well as vary in size. This ability not only comes from encoding and decoding of the chromosome, but from modified genetic operators (i.e. crossover). With this in mind, LAR has been shown to be an effective methodology on benchmarking problems [58].

An extension to the LAR representation proposed by Liu *et al.* [46] introduces a local search based genetic algorithm (GALS). This approach focuses on adapting the typical mutation operator in a LAR based scheme to be more focused on making local optimizations. Experimental results show that the technique is efficient as well as highly effective.

Though these techniques are successful in their own right, the biological assumptions made as part of the DGAP nullify the main strength of these methodologies,

namely, the ability of the candidate community size to grow and shrink arbitrarily during evolution.

A second type of methodology is that of a multiobjective optimization approach. As outlined in Section 2.2.8, the principle is to be maximizing two (or more) different evaluations of individuals at the same time. This has been shown to be a successful approach by Shi *et al.* [63], who employed a multiobjective approach to the community detection problem. The objectives in the authors' approach were the in-degree and out-degree of nodes in the network. With this in mind, the DGAP could potentially show more accurate results if used in conjunction with advanced multiobjective community detection approaches.

Chapter 5

Methodology

As outlined in Section 4.2, there exists a chance to produce a novel system for predicting genes involved in various diseases. This novel system would be a hybrid of biological and computer science approaches that takes into account multiple evidence types and advanced multi-objective AI techniques. This new technique, however, will still be inspired by the initial work of Tahmasebipour and Houghten [66]. Unlike the work of Heravi *et al.* [30], this novel approach will use a genetic algorithm.

This proposed GA will be multi-objective, where each objective is the sum of the centrality measures of each node across the community (see Section 5.1.6). Typical multi-objective fitness techniques like that of weighted sum require a value to be placed on an objective. In the case of centrality measures as objectives this poses a problem. For example, how can you decide if a higher "eigenvector" value is better for a community than higher "betweenness" value? Surely values can be found experimentally, but this must be done for each graph, as not all graphs share the same topology. This problem also eliminates the possibility of using a Pareto approach. The argument against Pareto is constructed much the same as the one against a weighted sum: what does it mean when looking at the Pareto front for an experiment if one candidate community has more of one attribute than another? To be able to construct a post processing of the Pareto front to give solutions to a biologist, these questions need to be able to be answered.

This chapter seeks to discuss the various motivations of decisions made in this thesis and define the technique created completely. As such, genetic algorithm details are found in Section 5.1, dataset choice and generation in Section 5.2, and finally various evaluation criteria and benchmarking are discussed in Section 5.3.

5.1 Genetic Algorithm Details

This section and relevant subsections detail the internal structure of the Genetic Algorithm created for the purposes of this thesis. As a whole, this technique was implemented in the programming language Java version 10.0.1. All algorithms developed in this thesis were implemented directly, i.e. without the use of a package.

5.1.1 Individual Structure

The internal representation of a candidate solution will be that of a bitstring. Each index in the bitstring will represent a gene, and a value of one at that index signals that the gene being referenced is inside the community. The number of indices with a value of one (true) in a chromosome is equivalent to the maximum community size selected. Chromosomes will also possess a number of "fixed genes" for LOO testing (see Section 3.8.1). These genes, and as such their indices in the bitstring, will always be made to be true. Implementation will be achieved through the Java BitSet data structure available in the Collections library.

5.1.2 Self Correction

If through genetic operators a chromosome ends up not possessing the maximum community size, the individual must undergo self correction. Self correction occurs in three total cases. The following subsections introduce each respective correction procedure. Note that these operations can lead to destructive behaviour and are brought about entirely through genetic operators. This is discussed further in Section 5.1.5.

Missing Fixed Genes

Through exchange operators it is possible in numerous cases that the known disease genes are missing from the children. Self correction is simply achieved by setting the fixed gene indices to be true. Note that this can lead to the community size being larger than the maximum and as such performs the procedure in the next subsection immediately.

Over Community Maximum Size

If the chromosome possesses more genes in the candidate community than the maximum, the process of self correction will randomly remove genes from the community

until the size is correct. Note that the randomly removed genes cannot be the fixed genes.

Under Community Maximum Size

If a candidate community is under the maximum community size limit, random new genes are added until the fixed size is again reached.

5.1.3 Selection

For the purposes of this thesis, all selection is implemented via tournament selection. For full definition see Section 2.2.5.

5.1.4 Mutation

Original implementations of a mutation operator were inspired by that of single-bit mutation. This technique selects a random index and flips the state found there. However, this can lead to numerous amounts of self corrections for an individual. For example, if a known gene is set to false, a self correction is incurred. As a result, the mutation operator used in this thesis is an extension of a basic single-bit mutation, *exchange mutation*. Exchange mutation is achieved by randomly selecting one non-fixed true index in the bitstring, and randomly selecting one false index in the bitstring. These two indices then have their states inverted thus preserving the maximum community size and incurring no self corrections as a result of mutation.

5.1.5 Crossover

This thesis incorporates the use of two crossover methods. The following two subsections define their procedure and discuss their strengths and weaknesses.

One-Point Crossover

The first method is that of a standard one-point crossover. For this method two individuals are chosen via the selection operator, as well as a random index i. From indices $[0, i)$, the individuals remain the same. Indices from $[i, N]$ (where N is the bitstring length) have their values swapped between individuals. While this method encourages strong exploration of the fitness landscape, it can be quite destructive to the chromosome. There are numerous cases in which random selection of i can incur

self correction and damage strong genetic material meant to be passed on to future generations.

Safe Dealer-Based Crossover

To combat self correction, a second crossover technique named Safe Dealer-Based Crossover (SDB) has been created. SDB is achieved by first selecting two individuals and finding the intersection of their true indices inside of the bitstrings. The two children of the method 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 data structure and 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 maximum 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 is due to new indices never being created, only "passed" between individuals. As such, for SDB to be successful it relies on large population.

5.1.6 Fitness Methodology

In Section 2.2.8), various issues and reasoning for the exclusion of a Pareto based fitness scheme were presented. This section introduces another multi-objective fitness technique known as Sum of Ranks (SoR) [4][15]. SoR differs from other popular multi-objective techniques such as NSGA-II [16] and SPEA2 [76] as SoR does not create or maintain a Pareto front. Instead, the returned values of executing the SoR method on a population are 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 Tables 5.1, 5.2, and 5.3. Table 5.1 contains the raw values of six individuals and their associated Pareto ranking. Table 5.2 contains SoR taking place. 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.

Clearly as stated previously, this does not create a Pareto front of solutions. This is perfect for the proposed technique as no extra post processing must be done about what centrality measures to possibly sacrifice for a "good" choice to be made. Further

Indiv	**Raw Fitness** A	B	C	D	**Pareto Rank**
1	1	9	5	4	1
2	2	100	4	8	1
3	10	9	9	10	2
4	16	100	8	4	2
5	16	9	500	0	1
6	0	1000	1000	1000	1

Table 5.1: Pareto ranking example with six individuals

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.

Like most modern techniques in evolutionary computation, there exist numerous small adjustments that can be made to SoR to improve performance and results. For example, Table 5.3 shows a variant known as normalized SoR. This 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.

With these properties in mind, 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 that may be found in a multi-objective problem. As such, this thesis will employ the use of normalized SoR in all experiments. Raw values for each objective will be computed by the summation of each gene's respective centrality measures that are being maximized if they are inside of the candidate community. The generation of centrality measures for each gene is described in Section 5.2.

5.2 Datasets and External Tools

In an effort to remain consistent with previous work, both Breast Cancer and Parkinson's disease will be studied in this thesis. Additionally, used datasets will extend to that of Alzheimer's disease. Generation of these data sets is outlined in Section 5.2.1 and remains true to Heravi et al. [30].

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

Table 5.2: Sum of ranks example with same six individuals as in Table 5.1

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

Table 5.3: Normalized sum of ranks examples with same six individuals as in Tables 5.1 and 5.2

5.2.1 Dataset Generation Process

Initially, fixed known disease genes are selected via the Genotator [70] tool. Genotator is a tool used by biologists to aggregate data and make it easier to search and index. In the case of this thesis, the top fifteen genes are always selected as known. Note that these also become the benchmarking criteria for LOO (leave one out) analysis (see Section 3.8.1). Next, these known genes are input into Cytoscape [14]. Cytoscape is open source software which allows for the study of genetically inspired networks. GeneMANIA [71] 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 thesis, given fifteen known disease genes from Breast Cancer, GeneMANIA allows the querying of the next N most frequently interacting genes based on evidence types and stores the resulting network in a Cytoscape structure. 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 [62] plugin is used. This allows the Cytoscape software to compute various centrality measures based on the currently open graph and export the results. Exporting is done simply to a csv

Name	Evidence Types
CIPHER	phenotype, PPI [74]
ENDEAVOUR	expression, PPI, functional, text mining [1]
GFINDER	phenotype, expression [50]
CAESAR	expression, functional, text mining [24]
CGPRIO	sequence, functional, PPI [74]
GENESEEKER	expression, functional, text mining, phenotype [68]

Table 5.4: Popular DGAP methodologies and their respective evidence type usage

file, and is then fed as input into the GA.

5.3 Evaluation Criteria

Biological parameters as well as various GA parameters will be chosen such that comparison can be made with previous bodies of work. In addition, all three testing procedures from Section 3.8 will be implemented so as to review the proposed methodology as thoroughly as possible. Particular attention will be given to the results of LOO validation during empirical testing as successful runs can be examined individually. Table 5.4 contains additional popular DGAP methodologies and will be used to compare with the new technique outlined in this thesis.

Chapter 6

Breast Cancer Case Study

Breast Cancer is a potentially hereditary type of cancer which typically begins with tumors growing in breast tissue. In general, the prognosis for the disease is highly dependant on the nature of the cancer, extent of spreading to other organs, as well as the individual's age [25]. Despite this, recent scientific advances have lead to a significantly improved prognosis of this disease. This chapter contains a case study of Chapter 5's methodology applied to the disease of Breast Cancer. This disease was chosen due to its highly studied nature, allowing for greater possibility of drawing accurate conclusions.

6.1 Data Generation

Data generation for this case study is performed as described in Section 5.2.1. Table 6.1 contains the known genes for breast cancer. These known genes are kept equivalent to previous work so as to allow for easy comparison.

6.2 Experimental Design

For the purposes of this case study, three individual parameter settings were empirically chosen in line with previous work to test the effectiveness of the methodology presented by this thesis. Table 6.2 contains the "HighCross" (high crossover) parameter setting, while Tables 6.3 and 6.4 contain the settings for "Balance" (balanced crossover and mutation) and "HighMut" (high mutation) respectively. Note that Tables 6.3 and 6.4 contain only the changes from the default parameter setting of HighCross. Fitness objective choice is discussed in the following section.

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

Table 6.1: Breast Cancer Known Disease Genes

Parameter	Value
Population	8000
Generations	2500
Selection	Tournament, k=5
Elitism	1
Crossover Method	One-point
Crossover %	75%
Mutation %	25%
Runs	30
Community Size	100
Fitness Method	SOR

Table 6.2: HighCross parameter setting

Parameter	Value
Crossover %	50%
Mutation %	50%

Table 6.3: Balance parameter setting

Parameter	Value
Crossover %	30%
Mutation %	70%

Table 6.4: HighMut parameter setting

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 6.5: Correlation study from Breast Cancer data

6.3 Fitness Objectives

The purpose of this section is to illustrate the decision making process for selecting fitness objectives.

6.3.1 Correlation Study

In an effort to fully understand which centrality measures should be used as fitness objectives, a brief correlation study was performed on the breast cancer dataset as a precursor to potentially using principle component analysis. Samples were formed by grouping numerical results from each centrality measure together as found in Section 2.1.3. Correlation was then computed using a standard Pearson correlation coefficient [3]. Table 6.5 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 eccentricity much like is specified in Section 2.1.13.

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 (see Section 4.2) would identify that these two measures working in conjunction produce 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

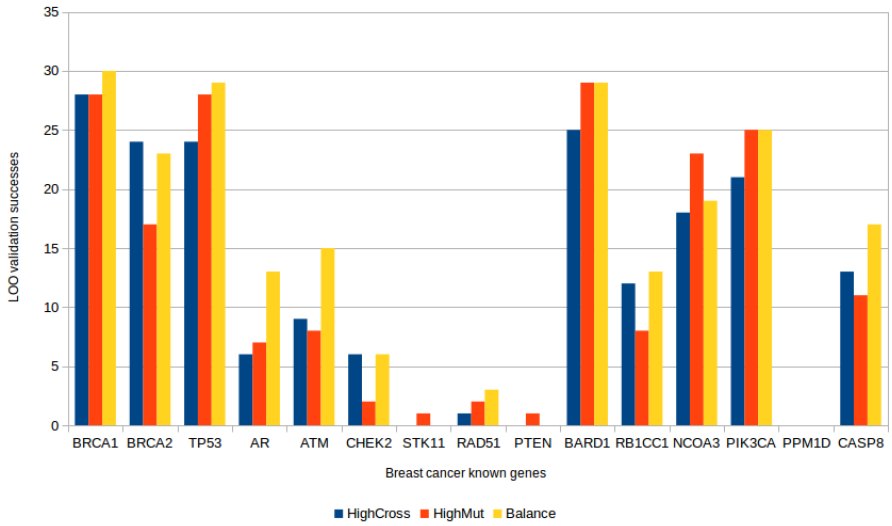

Figure 6.1: Bridging and betweenness LOO validation parameter setting comparison on one-point crossover

respect to all other centrality measures.

6.3.2 Preliminary Fitness Testing

Figure 6.1 illustrates the success rates of Leave-One-Out (LOO) validation based on GA parameters defined in Section 6.2 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 include Safe Dealer-Based (SDB) crossover (see Section 5.1.5) for comparison with the same parameter settings defined in Section 6.2.

As such, Figure 6.2 contains the comparison of three different objective schemes using the Balance parameter setting. Based on this figure, it is straightforward to see that the objective setting of stress and betweenness produces the strongest results. Note that despite the differences in approach (see Section 4.2), these two path based measures working together have been replicated to be highly successful despite their extremely high correlation. Other fitness objective settings were investigated, but are left out of this thesis due to the completely dominant results given by the objectives found in Figure 6.2. However, raw summary results of these experiments can be found in Section A.1.

For the sake of brevity, Table 6.6 defines labels for the objective settings that will be explored as part of this work. These labels are designed to be used in conjunction

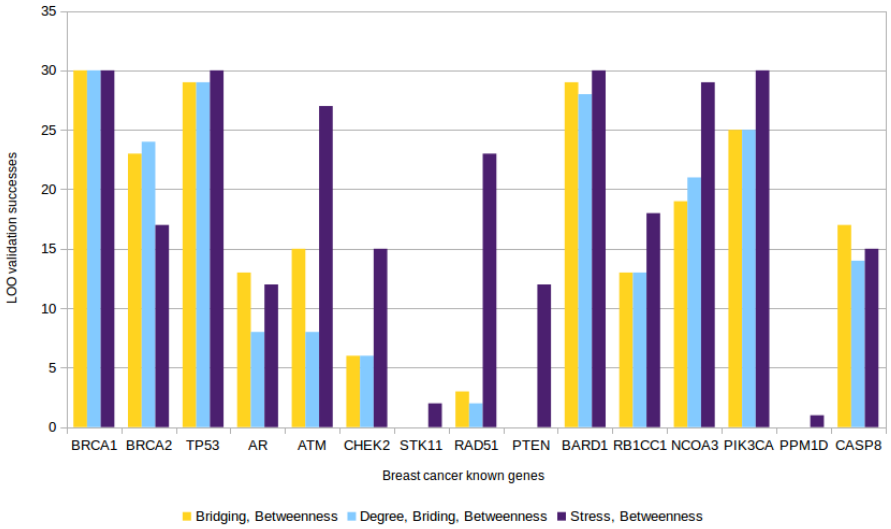

Figure 6.2: Comparison of LOO validation successes on breast cancer using one-point crossover on the Balance parameter setting

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

Table 6.6: Fitness objective labeling scheme

with the parameter settings defined in Section 6.2. For example, results defined by the label "AK-Balance" imply a fitness objective setting of stress and betweenness with EC parameters specified in Table 6.3. The inclusion of the single-objective settings sBet and sStress is in hopes of validating the use of a multi-objective methodology.

6.4 Experimental Results

Figures 6.3 and 6.4 contain example convergence curves for AK-Balance experiments using one-point and SDB crossover respectively. 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.

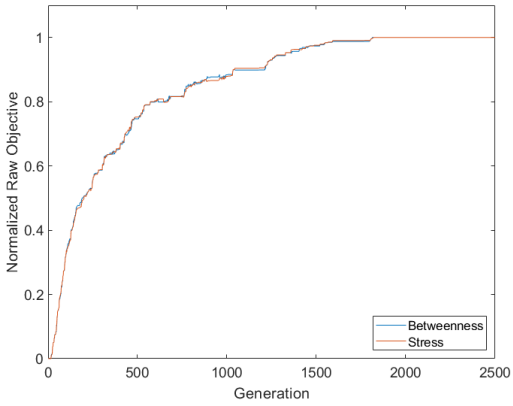

Figure 6.3: Breast cancer AK-Balance with one-point crossover experiment fitness curve

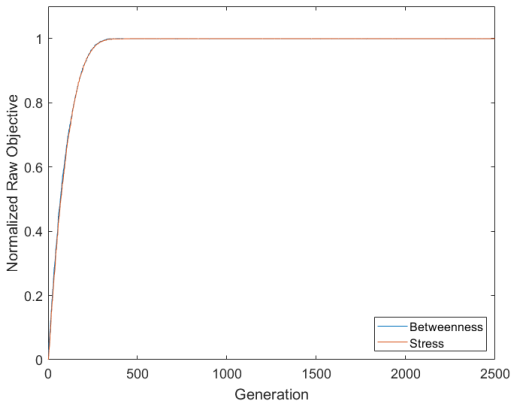

Figure 6.4: Breast cancer AK-Balance with SDB crossover experiment fitness curve

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 6.7: Breast cancer with one-point crossover experiment summary

Figure 6.3 illustrates a fairly standard convergence for an EC technique, while Figure 6.4 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. This issue is further explored in Section 6.7 with reference to specific performance measures.

Tables 6.7 and 6.8 contain study-wide summaries of each experiment type based on the evaluation criteria defined in 5.3. Note that fold enrichments (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. Individual gene success rates for SDB crossover are found in Section A.1.

Based on Table 6.7, the most successful measures for this case study are the sBet-Balance setting, and the AK fitness scheme due to their result of recovering all fifteen known LOO validation genes. Tables 6.9 and 6.10 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-

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 6.8: Breast cancer with SDB crossover experiment summary

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. However, both settings appear to struggle with similar genes, namely STK11 and PPM1D. Further discussion and comparison can be found in Section 6.7.

6.5 Comparison to Previous Work

Table 6.11 shows the comparison between two non-EC methodologies as well as the two EC techniques on which this thesis is based.

Clearly the technique proposed in this thesis 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. Section 6.7 contains discussion as to why this may be the case.

6.6 Predicted Genes

As the overall goal of the DGAP is predict new genes for study, the top 1% of non-fixed genes across the AK-Balance and sBet-Balance experiments were examined.

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 6.9: 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 6.10: 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	0.80
GP Approach	24	9/15	0.60
AK-Balance	88	15/15	1.00
sBet-Balance	91	15/15	1.00

Table 6.11: Breast cancer DGAP methodology comparison. Balance methods implement the one-point crossover technique.

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 6.12: Predicted genes for future breast cancer study.

Table 6.12 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 [52]. Interestingly, recent studies have begun to study its relevance to the onset of breast cancer [45] 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 [52].

6.7 Case Study Discussion and Conclusions

Originally, the hypothesis proposed in Chapter 5 and then refined in Section 6.3.2 of this chapter was that a multi-objective methodology containing the centrality measure of bridging (Section 2.1.12) would produce the strongest results. The motivation for this decision came from examining the correlation between pairs of measures and determining that bridging in conjunction with any other measure provided at least some amount of not highly correlated data. Despite these facts, this case study has shown and confirmed that the previous literature was correct to use stress and betweenness together (AK fitness schemes) when studying breast cancer. Furthermore,

using just stress or betweenness on their own in a single-objective environment was, for this case study, able to produce improved results. Another contradiction given by the results of this study was the overall predicted performance of SDB crossover. The original motivation behind this crossover was to avoid the numerous self corrections behind breeding new individuals if one-point crossover was in use. Use of SDB resulted in lowered sensitivity rates across all testing conditions. However, it was found to produce higher FE values for this case study. This is 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 (see Figure 6.4) in comparison to a one-point crossover approach even with such a large population size.

As stated previously, the path based measures of stress and betweenness used both in conjunction or on their own yielded significant improvements over past approaches (see Section 6.5). These improvements include the recovery of each known LOO validation gene. This success is likely to do with integration of multiple evidence types, as well as the exchange mutation approach used by this thesis being a strong hybrid of both exploration and exploitation. In the future to improve the recovery rate of the challenging genes in this problem (STK11 and PPM1D), more shortest path based measures should be explored. Furthermore, SDB crossover as a methodology should be rigorously empirically tested and iterated upon to balance its exploitative nature.

Chapter 7

Parkinson's Disease Case Study

Parkinson's disease is a genetic disease which leads to the degeneration of the nervous system [59]. Typical symptoms of Parkinson's include tremors, impaired balance, and communication difficulties [33]. 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 chapter applies the methodology defined in Chapter 5, as well as the refined changes from Chapter 6 to Parkinson's disease in order to measure the effectiveness of the introduced methodology.

7.1 Data Generation

Input data for this case study is generated as defined in Section 5.2.1. Known genes, listed in Table 7.1, are taken from previous work to allow for various comparisons.

7.2 Experimental Design

Parameter settings as well as fitness objective labels are kept consistent with the naming convention outlined in Section 6.3.2. For the purposes of this case study, all settings are repeated.

7.3 Experimental Results

Figures 7.1 and 7.2 contain convergence curves for AK-Balance experiments on one-point and SDB crossover respectively. Data points are computed corresponding to the

Gene	NCBI ID
LRRK2	120892
SNCA	6622
PARK2	5071
MAPT	4137
APOE	348
GBA	2629
GAK	2580
BST1	683
DRD2	1813
PINK1	65018
MAOB	4129
BDNF	627
CYP2D6	1565
PON1	5444
COMT	1312

Table 7.1: Parkinson's Known Disease Genes

previous case study. As such, each curve represents the change of the best individual's raw objective values over time, then normalized.

Similarly to the previous chapter, the one-point crossover experiment's curve behaves much as a typical EC technique tends to behave, with a gradual build-up over time. SDB crossover possesses the behaviour of an experiment that either has excessive selection pressure, or potentially a methodology that favours exploitation too heavily. As discussed in Section 6.7, this behaviour is almost certainly the result of how SDB crossover functions as a result of not creating "new" information strictly

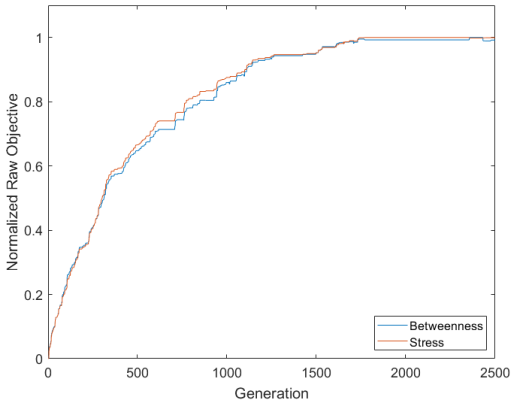

Figure 7.1: Parkinson's AK-Balance with one-point crossover experiment fitness curve

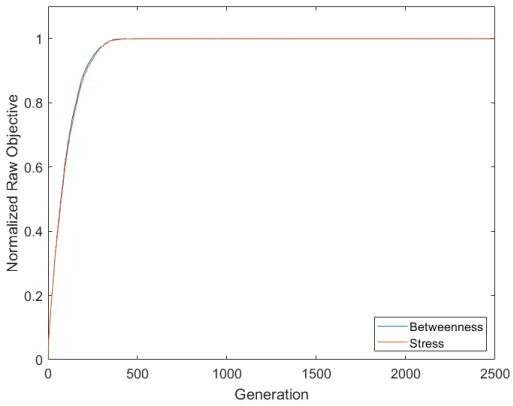

Figure 7.2: Parkinson's AK-Balance with SDB crossover experiment fitness curve

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 7.2: Parkinson's with one-point crossover experiment summary

through the breeding process.

Tables 7.2 and 7.3 contain experiment summaries for both one-point and SDB crossover respectively. Measures are computed as outlined in the previous case study, described in Section 6.4.

It is immediately evident based on Tables 7.2 and 7.3 that, despite the overall success of the methodology, Parkinson's disease presents a greater challenge than breast cancer 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 ap-

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 7.3: Parkinson's with SDB crossover experiment summary

pear 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 7.4 and 7.5 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. Summary statistics for the remaining objectives are available in Appendix A.2.

Much like the previous case study, Tables 7.4 and 7.5 illustrate that 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. Section 7.6 further discusses difficulty comparisons between case studies, and known gene difficulty within case studies.

7.4 Comparison to Previous Work

Table 7.6 contains a comparison between past EC approaches as well as a popular biologically inspired DGAP approach, ENDEAVOUR.

Much like in the previous case study, the technique proposed by this thesis out performed past methodologies in terms of sensitivity and FE. Note, a large portion of the success of this methodology in terms of FE is due to the overall stability obtained

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

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 7.5: 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 7.6: Parkinson's DGAP methodology comparison. Balance method implements the one-point crossover technique.

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 7.7: Predicted genes for future Parkinson's study.

due to the ease of ranking several of the known LOO validation genes. Section 7.6 contains further discussion as to how the problem of disease gene association differs in difficulty for Parkinson's relative to the previous case study.

7.5 Predicted Genes

Maintaining the overall goal of the DGAP, the top 1% of non-fixed genes across both the AK-HighCross and sStress-Balance parameter settings were investigated. Table 7.7 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 exists links between this gene and the neurodegenerative properties of Parkinson's disease [18]. 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 [65]. Additionally, the C8A and UBC genes have been previously identified by past EC approaches as being guilty by association [31].

7.6 Case Study Discussion and Conclusions

Parkinson's presents a unique challenge due to the relatively balanced difficulty of the known genes. Despite this more challenging nature, the results presented in this case study mirror that of the breast cancer case study, namely, the weak results of SDB crossover. As stated previously, the weaker performance of SDB crossover appears to be due to the lack of exploration and heavy exploitation principles in the approach.

A second similarity is that of the strength of the path based objective settings of AK and sStress. However, despite the strength of these methodologies not all genes were recovered. In an effort to investigate this, the raw values of the troublesome genes (i.e. GBA) 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. This inspection was also repeated for the troublesome genes from the previous case study's dataset, and confirmed the same behaviour. This implies that stress and betweenness in some cases is not enough to see how important a gene's role is in a network. Future studies should thus investigate more shortest path based measures.

A significant difference between case studies is the highly increased FE values across all parameter settings. One potential reason for this increase is that the methodology is very stable on the genes it manages to recover. This can be seen in Table 7.5 and inspecting the standard deviation of the ranks for the genes which are recovered all thirty times. Even in the case of some of the difficult genes, the standard deviation is quite low. In the future, more performance metrics should be introduced to measure stability.

Chapter 8

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 [10]. As part of a continued effort to test the methodology defined in this work, this chapter contains a case study of Alzheimer's disease based on principles learned from Chapters 5, 6, and 7.

8.1 Data Generation

Data generation for this case study is based on the procedure outlined in Section 5.2.1. 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 [29]. Table 8.1 contains the known Alzheimer's genes as defined by OMIM. These genes are then used as input to Cytoscape, and the data generation 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.

8.2 Experimental Design

Parameter settings and fitness objective labels are kept consistent with the naming convention outlined in Section 6.3.2 and are repeated as part of this case study.

Gene	NCBI ID
HFE	3077
NOS3	4846
PLAU	5328
A2M	2
MPO	4353
APP	351
PSEN1	5663
PSEN2	5664
APOE	348

Table 8.1: Alzheimer's Known Disease Genes

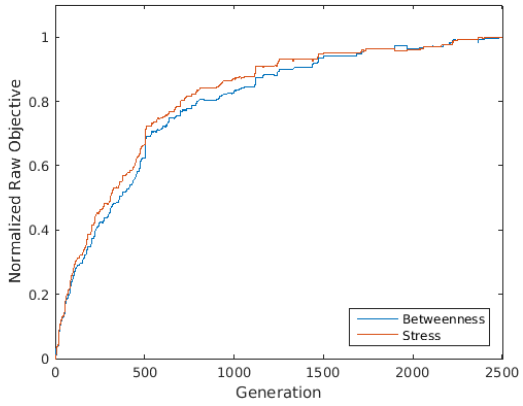

Figure 8.1: Alzheimer's AK-Balance with one-point crossover experiment fitness curve

However, for the sake of brevity, only the Balance parameter setting in this case study is reported. This is due to the apparent difficulty of the problem being significantly lowered. This is further discussed in Sections 8.3 and 8.5.

8.3 Experimental Results

Figures 8.1 and 8.2 contain single example convergence curves for two families of experiments, namely, AK-Balance with one-point crossover and AK-Balance SDB crossover respectively. Data for these figures consist of the best individual's raw objective values at each generation that have then been normalized.

These figures illustrate that the convergence properties of this problem correspond to those of the previous two case studies. Namely, Figure 8.1 demonstrates a fitness curve typically associated with EC studies and Figure 8.2 shows a method that is converging quickly either due to selection pressure or lack of diversity in the popu-

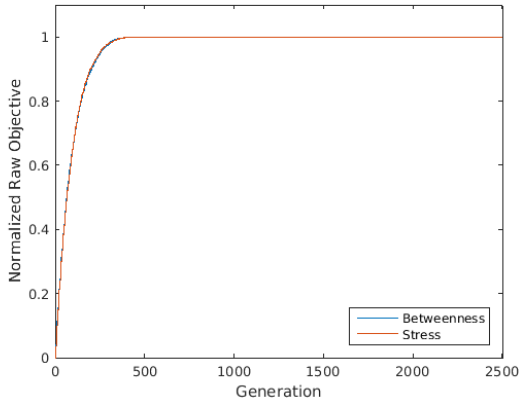

Figure 8.2: Alzheimer's AK-Balance with SDB crossover experiment fitness curve

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 8.2: Alzheimer's with one-point crossover experiment summary

lation. As discussed previously, it is certainly the case that SDB crossover possesses extremely high exploitative properties in its current form.

Tables 8.2 and 8.3 contain experimental summary statistics computed as previously defined in Section 6.4.

Simply by inspection, it is clear that this problem is significantly easier for this methodology 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 8.4 contains the individual LOO successes 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

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 8.3: Alzheimer's with SDB crossover experiment summary

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 8.4: sBet-Balance parameter setting with SDB crossover individual gene statistics on Alzheimer's

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

Table 8.5: Predicted genes for future Alzheimer's study.

known genes and produced the highest FE values. Remaining results are contained in Section A.3.

As expected, Table 8.4 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 the genes MPO and PSEN1. However, all thirty test cases were recovered easily. Further discussion on this set of known genes and future direction for this case study is outlined in Section 8.5.

8.4 Predicted Genes

Table 8.5 contains potential new genes for study based on the top 1% most frequently found non-fixed genes that Genotator has confirmed to be linked to Alzheimer's disease.

The top gene for future study reported by this methodology is that of CTNNB1. This gene 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 [52]. Despite this, multiple studies have shown interaction between CTNNB1 and Alzheimer's [49][20].

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 in genetic studies of Alzheimer's.

Upon further investigation, both the CD59 and CDH5 genes appear to already be known as influential genes with respect to Alzheimer's [75][51]. 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.

8.5 Case Study Discussion and Conclusions

Unlike the previous case studies, the known genes for Alzheimer's were chosen in an extremely conservative manner. As stated previously, this was done in part because tools like Genotator did not natively possesses database entries for this disease. Additionally this was done to provide a base point for comparison when using future EC techniques, as Alzheimer's has been rarely studied in this manner. Due to this, the difficulty of the case study is significantly reduced in comparison to the previous case studies, especially Parkinson's as seen by the performance of the previously underwhelming SDB crossover.

A further complication with this case study is the lack of available comparative databases (e.g. ENDEAVOUR) as tools using similar known gene setups. As outlined in Section 8.4, 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 seek to be less conservative when selecting the known LOO genes, as this case study illustrates that the problem has sufficient structure for EC to exploit.

A second important conclusion previously outlined was the success of SDB crossover. Now that the problem difficulty has reduced, the highly exploitative nature of the technique is significantly more desirable. This implies that a modification to the exchange process of SDB to include more explorative properties could drastically improve the applicability to the previous case studies.

Chapter 9

Conclusion

In sum, 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 relating genes together. 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 results were the shortest path based measures, stress and betweenness, matching previous literature. Contrary to the original hypothesis, the single-objective parameter settings for this study showed significant success. This confirms the notions that despite the widespread use of multi-objective approaches in evolutionary computation, single-objective techniques still have their strengths. However, the weakness of the 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.

Furthermore, 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 of the Disease Gene Association Problem. 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.

Additional future work should include the refinement of the safe dealer-based

(SDB) crossover technique. Originally 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, safe dealer-based crossover could be a worthwhile crossover approach. As balancing exploration and exploitation is part of every evolutionary computation technique, it is also potentially worth trying different mutation techniques, selection techniques, as well as fitness methodologies in the future. This could also include the notion of switching techniques during the search procedure, for example, switching to more explorative techniques if the population begins to converge during execution. Another important consideration would be to vary the elitism parameter to more carefully study selection pressure in both crossover techniques. In order to compare these numerous potential settings in the future, more rigorous statistical techniques such as ANOVA testing should be used.

Another potential expansion would be the adaptation of the technique to use weighted networks. This would also include exploration of various new path based centrality measures in the hopes of detecting the troublesome genes in breast cancer and Parkinson's as well as less understood diseases in the future. As this changes the overall structure of the input networks, it would be potentially beneficial to further study the properties of the graphs generated by the generation process of this work.

In addition, despite the overall success of this methodology it should be noted that this technique 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 in the future, application of the methodology only requires changes to the known disease genes, and as such the data generation procedure.

Lastly, each predicted gene from all three case studies as part of the Disease Gene Association Problem present potentially interesting future studies for biologists to investigate as part of further understanding deadly diseases.

Bibliography

[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] Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies–a comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002.

[7] 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.

[8] Béla Bollobás. Random graphs. 1985. *Academic, London*, 2001.

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

[10] 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.

[11] Carlos Castillo. Effective web crawling (ph. d. thesis). university of chile. Technical report, Retrieved 2010-08-03, 2004.

[12] 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.

[13] Caroline St Clair and Jonathan E Visick. *Exploring Bioinformatics*. Jones & Bartlett Publishers, 2013.

[14] 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.

[15] 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.

[16] 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.

[17] Marco Dorigo and Luca Maria Gambardella. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation*, 1(1):53–66, 1997.

[18] 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.

[19] Anton J Enright, Stijn Van Dongen, and Christos A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[20] 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.

[21] David B Fogel and Lawrence J Fogel. An introduction to evolutionary programming. In *European Conference on Artificial Evolution*, pages 21–33. Springer, 1995.

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

[23] Michael R Garey and David S Johnson. Crossing number is np-complete. *SIAM Journal on Algebraic Discrete Methods*, 4(3):312–316, 1983.

[24] Kyle J Gaulton, Karen L Mohlke, and Todd J Vision. A computational system to select candidate genes for complex human traits. *Bioinformatics*, 23(9):1132–1140, 2007.

[25] 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.

[26] 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.

[27] 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.

[28] Shuba Gopal, Anne Haake, Rhys Price Jones, and Paul Tymann. Bioinformatics: a computing perspective. 2008.

[29] 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.

[30] 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.

[31] 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.

[32] John H Holland. Genetic algorithms. *Scientific American*, 267(1):66–73, 1992.

[33] 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.

[34] 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.

[35] Sorin Istrail, Granger G Sutton, Liliana Florea, Aaron L Halpern, Clark M Mobarry, Ross Lippert, Brian Walenz, Hagit Shatkay, Ian Dew, Jason R Miller, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proceedings of the National Academy of Sciences*, 101(7):1916–1921, 2004.

[36] Momin Jamil and Xin-She Yang. A literature survey of benchmark functions for global optimization problems. *arXiv preprint arXiv:1308.4008*, 2013.

[37] Jerome Walker. Double helix visualization. `https://stq.wikipedia.org/wiki/Bielde:DNA_double_helix_horizontal.png`, 2006. [Online; accessed 27-November-2018].

[38] Richard M Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.

[39] J Kennedy and RC Eberhart. Particle swarm optimization. *Piscataway December*, 1995.

[40] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[41] John R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, 1992.

[42] John R Koza. *Genetic programming II, automatic discovery of reusable subprograms*. MIT Press, Cambridge, MA, 1994.

[43] 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.

[44] 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.

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

[46] 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.

[47] Jinhu Lü, Guanrong Chen, Maciej J Ogorzalek, and Ljiljana Trajković. Theory and applications of complex networks: Advances and challenges. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 2291–2294. IEEE, 2013.

[48] 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.

[49] 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.

[50] Marco Masseroli, Osvaldo Galati, and Francesco Pinciroli. Gfinder: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Research*, 33(suppl_2):W717–W723, 2005.

[51] 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.

[52] 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.

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

[54] J-P Onnela, Jari Saramäki, Jorkki Hyvönen, György Szabó, David Lazer, Kimmo Kaski, János Kertész, and A-L Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

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

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

[64] Javier Simon-Sanchez, Claudia Schulte, Jose M Bras, Manu Sharma, J Raphael Gibbs, Daniela Berg, Coro Paisan-Ruiz, Peter Lichtner, Sonja W Scholz, Dena G Hernandez, et al. Genome-wide association study reveals genetic risk underlying parkinson's disease. *Nature Genetics*, 41(12):1308, 2009.

[65] 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.

[66] 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.

[67] 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.

[68] MA Van Driel, K Cuelenaere, PPCW Kemmeren, Jack AM Leunissen, Han G Brunner, and Gert Vriend. Geneseeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Research*, 33(suppl_2):W758–W761, 2005.

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

[70] 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.

[71] 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.

[72] Wikipedia contributors. Pareto efficient frontier. `https://en.wikipedia.org/wiki/File:Pareto_Efficient_Frontier_1024x1024.png`, 2014. [Online; accessed 7-November-2018].

[73] 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.

[74] 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.

[75] 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.

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

Appendix A

Additional Experimental Analysis

This appendix contains subsets of the fully parsed output files for each disease studied.

A.1 Breast Cancer Summary

```
-----------------------------------------------
Disease: bc
Cross: base
3Balance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,25.000000,29.466667,14.868929,29
BRCA2,20,54.500000,59.500000,35.032743,20
TP53,14,32.500000,47.766667,59.236337,29
AR,26,161.000000,232.800000,158.895411,8
ATM,20,259.000000,257.533333,171.863164,7
CHEK2,21,169.500000,230.200000,159.152150,11
STK11,69,407.000000,354.300000,109.370976,1
RAD51,104,415.000000,371.833333,108.263435,0
PTEN,111,414.000000,379.000000,94.920657,0
BARD1,15,60.000000,81.666667,74.682470,27
RB1CC1,26,135.500000,207.366667,154.770238,7
NCOA3,0,62.500000,76.700000,80.062971,22
PIK3CA,14,32.000000,80.066667,121.556836,25
PPM1D,186,420.500000,403.500000,60.768498,0
CASP8,14,145.500000,196.733333,145.721021,12
3HighCross:
```

Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,41.000000,39.600000,21.740793,29
BRCA2,0,50.500000,68.166667,65.561281,25
TP53,0,58.000000,87.633333,94.863417,25
AR,39,346.000000,262.500000,120.512912,4
ATM,22,344.000000,245.400000,137.262547,7
CHEK2,23,346.500000,259.000000,133.357957,8
STK11,56,349.500000,297.066667,102.054730,1
RAD51,19,349.500000,298.066667,106.660251,4
PTEN,95,350.500000,312.133333,82.860803,0
BARD1,14,108.500000,110.866667,71.192664,21
RB1CC1,24,342.000000,249.400000,129.055615,10
NCOA3,14,70.500000,121.900000,121.057225,22
PIK3CA,0,42.000000,82.066667,96.104086,27
PPM1D,100,352.500000,322.333333,75.946593,0
CASP8,14,346.500000,284.166667,115.006472,7
3HighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,22.500000,23.533333,8.696941,25
BRCA2,19,45.500000,46.700000,19.285782,21
TP53,14,21.000000,23.566667,9.291017,29
AR,42,182.000000,257.600000,184.403905,8
ATM,23,78.500000,141.933333,165.553603,13
CHEK2,52,130.000000,227.900000,198.276601,5
STK11,171,539.000000,462.233333,160.694116,0
RAD51,67,551.000000,397.400000,231.337133,1
PTEN,91,464.500000,421.733333,185.345332,0
BARD1,14,34.000000,51.000000,44.773761,28
RB1CC1,34,112.500000,198.433333,179.935177,5
NCOA3,17,45.000000,57.533333,43.905646,18
PIK3CA,15,23.000000,29.466667,17.363722,28
PPM1D,172,592.000000,536.100000,130.898315,0
CASP8,18,103.500000,164.066667,170.115160,11
Balance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,24.500000,28.233333,11.813736,30

BRCA2,18,49.500000,57.433333,32.735811,26
TP53,14,36.000000,66.566667,78.306948,26
AR,21,132.500000,174.333333,133.687317,13
ATM,25,364.500000,269.566667,170.089973,9
CHEK2,37,176.500000,247.733333,150.565585,4
STK11,134,404.000000,355.233333,103.650169,2
RAD51,27,405.000000,326.666667,138.560166,2
PTEN,86,411.000000,349.066667,113.968820,0
BARD1,14,62.500000,76.033333,50.214357,24
RB1CC1,14,158.000000,209.900000,161.435233,10
NCOA3,14,54.500000,114.466667,135.672713,19
PIK3CA,14,64.500000,89.500000,97.997449,27
PPM1D,149,408.000000,384.166667,74.183874,0
CASP8,14,69.000000,159.266667,159.803097,16
HighCross:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,43.500000,48.233333,33.282162,30
BRCA2,0,64.500000,71.433333,61.017060,23
TP53,0,55.500000,96.066667,100.280618,25
AR,33,299.000000,233.266667,124.512825,9
ATM,15,338.000000,280.433333,114.107839,5
CHEK2,23,330.500000,246.866667,129.747429,8
STK11,64,346.500000,323.966667,64.155782,0
RAD51,14,337.000000,270.033333,130.831158,4
PTEN,91,339.000000,303.033333,79.447004,0
BARD1,14,108.000000,115.600000,76.410236,25
RB1CC1,33,228.500000,205.333333,133.712335,13
NCOA3,14,59.500000,121.666667,117.498154,20
PIK3CA,14,38.000000,95.166667,121.181563,24
PPM1D,324,346.000000,345.300000,10.157891,0
CASP8,17,157.500000,188.400000,129.413691,15
HighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,19.500000,20.766667,6.775632,27
BRCA2,21,41.000000,43.666667,20.608558,22
TP53,0,23.500000,28.266667,24.621176,29

AR,30,153.000000,212.666667,178.917113,8
ATM,31,144.500000,312.300000,287.445725,8
CHEK2,22,162.000000,257.300000,237.055377,8
STK11,142,374.500000,462.166667,238.470281,0
RAD51,85,521.000000,486.800000,244.316556,3
PTEN,107,629.000000,521.633333,247.380969,0
BARD1,14,26.500000,44.566667,79.005318,26
RB1CC1,21,152.500000,203.666667,164.892093,5
NCOA3,15,41.000000,46.533333,29.534005,18
PIK3CA,14,23.000000,36.966667,23.417476,22
PPM1D,202,722.000000,668.500000,148.443360,0
CASP8,24,86.500000,159.500000,191.826745,12
akBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,19.500000,20.733333,6.648481,30
BRCA2,23,76.500000,83.266667,37.438624,16
TP53,0,23.000000,29.400000,18.485036,30
AR,27,88.500000,176.266667,146.139030,9
ATM,15,72.000000,111.833333,119.400827,25
CHEK2,31,74.500000,104.700000,94.805118,22
STK11,68,213.500000,254.833333,139.501349,1
RAD51,14,56.500000,76.733333,97.566576,27
PTEN,30,188.000000,221.466667,134.907052,11
BARD1,15,39.500000,52.733333,62.252508,29
RB1CC1,14,60.500000,96.366667,100.572322,20
NCOA3,15,34.000000,42.600000,24.368719,29
PIK3CA,0,22.500000,23.266667,7.506013,30
PPM1D,59,258.500000,266.866667,142.389881,2
CASP8,15,83.500000,130.000000,130.033417,21
akHighCross:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,25.000000,26.366667,8.640256,30
BRCA2,20,63.000000,77.633333,70.184723,23
TP53,14,28.000000,34.566667,18.421658,30
AR,14,142.500000,184.633333,135.882678,13
ATM,0,60.500000,120.233333,129.783290,23

CHEK2,24,125.500000,176.566667,131.334709,18
STK11,20,297.500000,244.200000,128.035394,5
RAD51,0,104.500000,148.433333,133.659806,20
PTEN,16,157.500000,202.633333,139.099687,10
BARD1,14,44.000000,58.866667,38.564530,30
RB1CC1,15,106.000000,177.200000,143.561449,15
NCOA3,0,48.500000,60.666667,40.893751,30
PIK3CA,0,37.500000,41.200000,26.058819,30
PPM1D,31,316.000000,257.000000,123.508202,5
CASP8,15,104.500000,166.666667,139.446772,17
akHighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,18.500000,19.100000,6.707433,30
BRCA2,24,65.500000,77.000000,39.709288,14
TP53,14,20.000000,19.066667,3.938609,30
AR,42,102.500000,164.233333,148.983147,12
ATM,15,57.500000,89.800000,112.044757,25
CHEK2,15,81.000000,158.333333,180.053313,16
STK11,80,488.000000,383.266667,173.238710,2
RAD51,14,65.500000,125.000000,149.490168,19
PTEN,38,173.500000,255.966667,184.831607,7
BARD1,14,24.000000,25.300000,7.391281,30
RB1CC1,17,62.500000,127.866667,150.496359,17
NCOA3,14,24.500000,31.066667,16.986675,30
PIK3CA,14,21.000000,21.333333,4.943357,30
PPM1D,22,233.500000,296.400000,177.252324,2
CASP8,16,69.500000,111.800000,106.620629,18
sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,25.000000,25.533333,9.919446,30
BRCA2,15,65.500000,71.066667,36.910828,15
TP53,0,20.500000,19.166667,5.540343,30
AR,0,85.500000,100.466667,87.101896,13
ATM,0,40.500000,46.233333,26.390546,30
CHEK2,20,83.000000,103.166667,86.084995,21
STK11,33,220.000000,255.233333,150.548579,7

```
RAD51,14,50.000000,83.966667,104.830399,26
PTEN,32,112.500000,174.466667,142.908153,12
BARD1,0,35.000000,36.300000,15.326673,30
RB1CC1,0,76.500000,115.000000,128.786645,19
NCOA3,0,31.000000,32.933333,15.591628,30
PIK3CA,14,21.000000,22.533333,8.617277,30
PPM1D,22,214.500000,245.866667,151.366131,3
CASP8,0,56.000000,102.166667,107.859450,21
sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,22.000000,22.400000,10.918507,30
BRCA2,21,62.000000,82.033333,53.871164,15
TP53,0,24.500000,31.133333,18.490880,30
AR,14,115.000000,178.366667,155.159869,11
ATM,20,67.500000,112.300000,121.290233,20
CHEK2,15,78.500000,139.100000,137.359420,16
STK11,134,424.500000,374.466667,101.607335,0
RAD51,31,119.000000,186.433333,154.259451,15
PTEN,29,182.500000,239.300000,154.011005,5
BARD1,0,28.500000,33.866667,18.687509,30
RB1CC1,20,95.000000,140.300000,128.704995,19
NCOA3,14,36.500000,41.900000,23.632751,27
PIK3CA,15,23.000000,27.500000,13.454265,30
PPM1D,112,416.000000,360.566667,114.200273,2
CASP8,27,120.500000,142.666667,99.117253,17
------------------------------------------------
Cross: newcross
3Balance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,25.000000,24.333333,11.414852,30
BRCA2,19,47.500000,55.533333,27.883666,24
TP53,0,25.500000,25.100000,10.639387,30
AR,108,333.000000,305.733333,123.695380,0
ATM,16,100.500000,160.400000,142.559752,13
CHEK2,108,307.000000,302.433333,116.809802,0
STK11,124,373.000000,364.500000,113.274385,0
```

RAD51,112,369.000000,319.466667,135.075645,0
PTEN,155,378.000000,369.600000,80.357220,0
BARD1,0,24.500000,25.033333,13.435628,30
RB1CC1,102,220.000000,237.900000,117.013218,0
NCOA3,0,57.000000,58.866667,28.051349,28
PIK3CA,14,30.000000,31.300000,10.249138,30
PPM1D,150,430.000000,417.966667,108.392481,0
CASP8,16,81.500000,70.700000,27.545761,30
3HighCross:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,26.500000,29.700000,9.048338,30
BRCA2,20,71.500000,69.233333,28.876292,22
TP53,0,26.000000,25.066667,8.208210,30
AR,105,241.500000,240.766667,91.570468,0
ATM,16,145.500000,163.533333,98.620391,7
CHEK2,105,277.000000,262.866667,109.238977,0
STK11,138,338.500000,351.700000,84.738197,0
RAD51,110,331.500000,336.433333,109.624338,0
PTEN,119,355.000000,346.300000,115.014887,0
BARD1,0,30.000000,28.533333,12.610158,30
RB1CC1,121,298.500000,312.266667,107.598530,0
NCOA3,0,47.000000,55.466667,30.411810,27
PIK3CA,14,33.000000,36.000000,14.202962,30
PPM1D,131,328.000000,324.833333,98.476702,0
CASP8,14,77.500000,68.766667,23.992360,30
3HighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,25.500000,26.733333,8.216608,30
BRCA2,30,92.500000,83.100000,26.437108,17
TP53,0,23.500000,22.766667,9.375990,30
AR,107,227.000000,266.966667,141.829544,0
ATM,23,101.000000,150.233333,123.763034,14
CHEK2,107,366.000000,331.533333,143.624687,0
STK11,144,424.500000,393.233333,107.481733,0
RAD51,127,443.500000,431.033333,128.765888,0
PTEN,138,424.000000,417.900000,105.514944,0

BARD1,0,27.000000,28.033333,11.903298,30
RB1CC1,102,225.000000,235.966667,105.044320,0
NCOA3,0,63.500000,63.300000,30.828950,25
PIK3CA,0,25.500000,27.466667,9.978367,30
PPM1D,180,425.500000,444.500000,111.613480,0
CASP8,0,82.000000,70.233333,26.958056,30
Balance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,23.000000,22.433333,10.849069,30
BRCA2,32,105.500000,104.300000,23.319594,5
TP53,0,23.500000,24.700000,9.074975,30
AR,104,249.000000,247.566667,114.211746,0
ATM,18,94.000000,125.566667,93.506291,21
CHEK2,106,322.000000,289.366667,129.424851,0
STK11,143,385.500000,398.433333,78.691315,0
RAD51,110,349.500000,342.733333,149.559568,0
PTEN,112,392.500000,383.000000,113.822911,0
BARD1,0,29.500000,29.633333,11.189537,30
RB1CC1,103,271.500000,275.233333,131.367787,0
NCOA3,0,59.500000,54.133333,26.497538,30
PIK3CA,0,25.000000,25.666667,8.687420,30
PPM1D,289,421.500000,422.766667,68.408854,0
CASP8,14,72.500000,64.600000,23.758265,30
HighCross:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,23.000000,24.800000,12.177990,30
BRCA2,25,101.500000,97.833333,28.399955,8
TP53,0,27.000000,26.900000,8.591696,30
AR,101,171.500000,193.900000,81.945461,0
ATM,0,99.500000,149.266667,116.842934,15
CHEK2,111,327.500000,305.966667,112.708193,0
STK11,140,372.000000,360.300000,92.758177,0
RAD51,108,350.000000,348.966667,109.391572,0
PTEN,242,370.500000,375.166667,78.665632,0
BARD1,0,25.500000,25.566667,12.883251,30
RB1CC1,101,245.500000,232.300000,83.920385,0

NCOA3,14,58.000000,56.700000,26.634370,30
PIK3CA,0,27.000000,31.833333,18.678927,30
PPM1D,147,356.500000,360.833333,84.476107,0
CASP8,0,78.000000,69.133333,22.905704,30
HighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,14,24.000000,24.766667,6.724548,30
BRCA2,68,102.000000,102.066667,9.291944,4
TP53,0,23.000000,23.033333,11.211089,30
AR,105,234.000000,244.400000,135.118747,0
ATM,0,96.500000,141.566667,96.555749,21
CHEK2,105,349.000000,339.133333,145.335221,0
STK11,140,410.000000,386.200000,123.276924,0
RAD51,116,452.000000,430.733333,127.908624,0
PTEN,136,456.500000,445.500000,121.505712,0
BARD1,0,31.000000,29.866667,13.927728,30
RB1CC1,103,161.000000,215.233333,128.225739,0
NCOA3,17,58.000000,55.466667,20.885126,30
PIK3CA,15,24.500000,28.366667,9.426607,30
PPM1D,207,476.500000,459.833333,126.156193,0
CASP8,14,64.000000,59.500000,25.825342,30
akBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,22.500000,19.500000,8.873634,30
BRCA2,101,130.000000,157.066667,57.033525,0
TP53,0,23.500000,19.566667,10.798095,30
AR,94,108.000000,146.700000,58.827715,1
ATM,0,40.500000,41.300000,19.044684,30
CHEK2,15,72.000000,65.100000,23.765159,30
STK11,109,323.000000,301.233333,109.439854,0
RAD51,0,63.500000,55.366667,22.748222,30
PTEN,101,275.000000,259.833333,115.673763,0
BARD1,0,24.000000,21.700000,13.365292,30
RB1CC1,16,66.000000,61.466667,22.992852,30
NCOA3,0,34.000000,33.600000,18.587352,30
PIK3CA,16,25.000000,25.000000,4.168850,30

PPM1D,108,339.000000,320.566667,93.999884,0
CASP8,0,76.000000,67.066667,24.244350,30
akHighCross:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,24.000000,21.233333,8.377652,30
BRCA2,102,159.500000,173.933333,62.259708,0
TP53,0,23.500000,19.733333,10.660702,30
AR,100,129.000000,157.366667,63.283970,0
ATM,0,48.000000,42.933333,15.813424,30
CHEK2,0,72.500000,68.000000,20.384240,30
STK11,109,305.500000,292.533333,83.216682,0
RAD51,0,50.000000,47.266667,24.343707,30
PTEN,107,235.500000,246.766667,83.223291,0
BARD1,0,27.000000,26.366667,9.488590,30
RB1CC1,0,73.500000,65.333333,24.196513,30
NCOA3,0,38.500000,35.200000,12.026981,30
PIK3CA,0,25.000000,23.266667,7.714444,30
PPM1D,121,301.500000,296.833333,83.974989,0
CASP8,0,73.000000,67.000000,22.685253,30
akHighMut:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,21.500000,17.466667,9.754692,30
BRCA2,101,117.000000,153.600000,63.858464,0
TP53,0,22.000000,18.633333,10.499535,30
AR,100,148.000000,154.600000,53.344424,0
ATM,17,43.500000,44.066667,14.975689,30
CHEK2,20,79.000000,70.433333,21.607284,30
STK11,119,414.000000,390.900000,128.583624,0
RAD51,30,63.500000,60.633333,15.075633,30
PTEN,102,224.000000,236.033333,108.394071,0
BARD1,0,22.000000,19.066667,10.859394,30
RB1CC1,0,60.500000,56.533333,25.856012,30
NCOA3,0,33.500000,31.933333,13.670843,30
PIK3CA,0,24.500000,21.066667,11.094992,30
PPM1D,118,394.000000,358.233333,153.715071,0
CASP8,14,75.500000,63.933333,24.133156,30

sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,20.000000,18.633333,9.880260,30
BRCA2,104,162.000000,170.200000,56.012560,0
TP53,0,19.000000,15.366667,9.129276,30
AR,22,86.500000,71.066667,27.268693,30
ATM,0,33.000000,34.033333,13.777752,30
CHEK2,0,61.500000,59.266667,22.726156,30
STK11,116,294.500000,303.600000,134.645793,0
RAD51,0,45.500000,42.933333,22.666903,30
PTEN,100,203.000000,186.900000,70.993127,0
BARD1,0,23.500000,19.933333,10.546885,30
RB1CC1,0,63.000000,55.733333,25.550401,30
NCOA3,0,28.000000,27.466667,13.778377,30
PIK3CA,0,22.000000,16.233333,10.496907,30
PPM1D,102,275.500000,271.100000,96.725830,0
CASP8,0,60.000000,52.966667,18.342495,30
sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
BRCA1,0,20.500000,18.133333,8.161023,30
BRCA2,102,183.500000,191.266667,70.753059,0
TP53,0,23.500000,21.533333,9.860172,30
AR,100,211.500000,221.866667,121.528125,0
ATM,0,65.000000,60.400000,19.340149,30
CHEK2,17,75.500000,68.166667,25.106783,30
STK11,115,350.000000,330.933333,107.377174,0
RAD51,0,69.500000,57.900000,24.050737,30
PTEN,119,276.000000,312.700000,131.214788,0
BARD1,0,24.000000,21.800000,11.943545,30
RB1CC1,0,76.500000,67.466667,25.146102,30
NCOA3,0,44.500000,46.100000,17.323196,30
PIK3CA,0,25.000000,22.966667,9.901178,30
PPM1D,107,375.500000,365.200000,108.029179,0
CASP8,100,140.000000,160.600000,64.380121,0

A.2 Parkinson's Summary

```
-----------------------------------------------
Disease: park
Cross: base
3Balance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,279,423.000000,420.066667,30.640218,0
GAK,337,422.500000,421.600000,20.978807,0
MAOB,17,92.000000,91.266667,46.249647,7
PON1,20,83.000000,176.500000,161.366097,17
COMT,323,426.000000,418.700000,27.050591,0
GBA,396,426.000000,426.700000,12.961880,0
BST1,14,51.000000,144.633333,159.825440,22
DRD2,37,57.000000,61.033333,15.744694,17
LRRK2,335,425.500000,424.900000,21.367636,0
MAPT,19,176.000000,223.133333,155.235690,12
PINK1,14,410.000000,301.400000,177.334401,9
APOE,86,413.000000,338.800000,123.060792,0
BDNF,210,424.000000,403.233333,60.984273,0
PARK2,0,14.000000,14.033333,2.760351,30
SNCA,0,14.000000,13.833333,2.666307,30
3HighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,128,366.000000,348.833333,56.991278,0
GAK,171,360.000000,350.500000,49.257067,0
MAOB,22,64.000000,74.366667,44.401369,10
PON1,30,253.000000,214.266667,139.055914,15
COMT,199,361.500000,352.533333,39.935926,0
GBA,222,360.000000,351.300000,32.531311,0
BST1,21,180.500000,193.533333,154.168232,15
DRD2,51,80.000000,83.433333,19.155774,3
LRRK2,130,364.000000,352.266667,51.827255,0
MAPT,22,79.500000,152.133333,134.766396,16
PINK1,28,361.000000,345.966667,64.871482,1
APOE,68,350.500000,263.266667,128.058697,1
```

BDNF,172,361.500000,345.733333,56.987557,0
PARK2,0,14.000000,13.133333,4.516127,30
SNCA,0,14.000000,14.000000,2.779233,30
3HighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,242,579.000000,537.966667,112.053092,0
GAK,237,578.000000,563.600000,65.410428,0
MAOB,18,32.000000,32.266667,9.013145,20
PON1,25,138.000000,220.766667,194.509406,15
COMT,272,560.500000,535.033333,80.462235,0
GBA,491,590.000000,587.933333,26.155943,0
BST1,17,44.500000,104.933333,149.225803,22
DRD2,18,28.500000,30.266667,6.872902,23
LRRK2,199,580.000000,564.866667,82.677910,0
MAPT,37,112.500000,204.066667,198.919310,13
PINK1,15,142.500000,290.066667,259.168023,11
APOE,121,339.500000,383.700000,190.874060,0
BDNF,213,552.000000,512.100000,100.967509,0
PARK2,0,14.000000,13.300000,3.640292,30
SNCA,0,14.000000,13.266667,3.628749,30
Balance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,173,418.000000,402.200000,55.853503,0
GAK,251,415.500000,412.900000,33.023868,0
MAOB,37,91.500000,116.300000,93.701341,5
PON1,16,66.500000,117.233333,117.043665,24
COMT,181,422.500000,409.000000,57.258639,0
GBA,259,412.500000,406.566667,37.927047,0
BST1,14,381.500000,275.400000,171.380923,12
DRD2,50,76.500000,80.500000,18.810764,13
LRRK2,296,419.000000,415.600000,29.460201,0
MAPT,14,204.500000,227.266667,156.084401,8
PINK1,14,415.500000,365.100000,130.770225,4
APOE,106,419.000000,358.866667,116.280201,0
BDNF,385,421.500000,420.000000,15.058507,0
PARK2,0,14.000000,14.566667,3.588135,30

SNCA,0,14.000000,13.333333,3.660915,30
HighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,228,359.000000,350.433333,30.853026,0
GAK,196,358.000000,346.166667,40.744861,0
MAOB,28,59.000000,69.766667,38.090349,13
PON1,14,104.000000,157.600000,124.528518,20
COMT,168,355.000000,347.700000,36.905424,0
GBA,187,353.000000,348.833333,33.223468,0
BST1,15,338.000000,263.333333,130.891725,7
DRD2,53,87.500000,83.533333,17.260146,6
LRRK2,335,356.000000,356.166667,10.923191,0
MAPT,14,341.500000,286.466667,110.710786,6
PINK1,14,352.000000,313.100000,103.282121,3
APOE,84,218.500000,236.066667,114.633128,0
BDNF,168,358.000000,350.000000,36.931437,0
PARK2,0,14.000000,13.500000,3.730120,30
SNCA,0,14.000000,13.900000,4.221047,30
HighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,209,515.000000,465.466667,134.758208,0
GAK,349,575.500000,559.366667,58.975438,0
MAOB,24,39.500000,45.100000,18.846659,22
PON1,25,100.500000,181.600000,186.039039,16
COMT,316,568.000000,563.100000,57.514226,0
GBA,425,586.500000,572.700000,45.681166,0
BST1,14,154.000000,268.666667,235.045019,16
DRD2,24,32.000000,36.300000,11.884937,24
LRRK2,286,579.500000,553.533333,73.620056,0
MAPT,53,169.000000,248.866667,212.472916,8
PINK1,14,569.000000,472.966667,221.078935,5
APOE,110,547.000000,473.466667,153.916174,0
BDNF,201,567.000000,535.700000,96.213179,0
PARK2,14,14.000000,14.533333,0.776079,30
SNCA,14,14.000000,14.400000,0.770132,30
akBalance:

```
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,140,430.000000,400.366667,91.638527,0
GAK,91,417.500000,324.666667,140.850699,1
MAOB,15,21.000000,31.300000,16.966803,26
PON1,21,75.500000,178.700000,180.400732,18
COMT,56,411.500000,313.333333,145.772930,3
GBA,333,433.000000,427.866667,22.504763,0
BST1,15,34.500000,52.333333,72.639086,30
DRD2,0,18.500000,17.700000,4.251977,30
LRRK2,124,431.500000,406.100000,74.783066,0
MAPT,15,100.500000,179.266667,158.570975,11
PINK1,16,41.000000,52.666667,34.855696,30
APOE,22,153.500000,211.566667,159.780169,6
BDNF,32,83.500000,132.933333,120.905188,9
PARK2,0,15.000000,13.300000,5.324958,30
SNCA,0,15.000000,11.866667,6.724804,30
akHighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,197,370.000000,355.333333,43.502147,1
GAK,15,357.500000,268.200000,134.670581,5
MAOB,15,22.500000,30.833333,31.751278,26
PON1,15,123.000000,160.800000,130.724084,22
COMT,28,358.500000,264.433333,141.041369,8
GBA,287,370.500000,368.866667,18.451677,0
BST1,15,48.000000,84.100000,100.204739,27
DRD2,15,19.000000,18.500000,3.170445,30
LRRK2,67,364.000000,347.933333,58.564985,0
MAPT,15,303.000000,218.833333,157.937588,14
PINK1,15,52.500000,105.866667,107.570581,26
APOE,22,119.500000,197.400000,148.487919,9
BDNF,22,98.500000,149.633333,130.190440,16
PARK2,0,15.000000,11.200000,6.880257,30
SNCA,0,15.000000,12.566667,6.468402,30
akHighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,138,504.500000,447.500000,132.574702,0
```

GAK,52,303.000000,325.900000,181.600234,2
MAOB,15,20.000000,20.700000,5.246674,28
PON1,26,74.500000,133.566667,146.240870,20
COMT,83,330.500000,324.233333,162.054415,3
GBA,368,520.500000,513.266667,42.973876,0
BST1,0,28.000000,34.566667,24.909053,30
DRD2,0,16.000000,16.566667,3.910052,30
LRRK2,95,521.500000,499.033333,83.672796,0
MAPT,44,149.500000,253.900000,203.537398,12
PINK1,15,40.500000,56.333333,87.366976,29
APOE,55,280.000000,318.500000,199.481181,2
BDNF,27,93.000000,146.366667,151.621462,14
PARK2,0,15.000000,14.333333,3.933353,30
SNCA,0,15.000000,9.633333,7.462473,30
sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,231,429.000000,412.133333,56.231500,0
GAK,47,383.000000,286.200000,159.804708,1
MAOB,0,23.500000,45.200000,58.191953,26
PON1,16,74.000000,173.766667,173.851638,18
COMT,19,351.500000,278.000000,167.333036,5
GBA,313,434.000000,426.400000,32.031881,0
BST1,15,29.500000,44.766667,35.093651,30
DRD2,15,18.000000,17.833333,2.018592,30
LRRK2,99,434.000000,411.000000,80.990847,0
MAPT,15,140.000000,201.833333,161.363390,12
PINK1,0,43.000000,65.000000,78.154136,29
APOE,34,160.500000,224.433333,159.681079,9
BDNF,20,74.000000,135.566667,144.351268,18
PARK2,0,15.000000,13.400000,5.385805,30
SNCA,0,15.000000,12.266667,6.263927,30
sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,83,425.500000,377.133333,103.394702,1
GAK,81,410.500000,324.433333,136.419152,2
MAOB,15,21.500000,37.866667,30.772506,21

PON1,26,94.000000,165.066667,150.369873,16
COMT,36,421.500000,331.300000,138.533489,3
GBA,336,430.000000,429.066667,23.561999,0
BST1,0,31.500000,53.666667,81.841450,29
DRD2,0,18.500000,17.633333,4.156286,30
LRRK2,348,427.500000,428.500000,21.413418,0
MAPT,25,265.500000,246.000000,177.650762,10
PINK1,0,39.500000,65.866667,101.674509,28
APOE,18,206.000000,253.366667,167.848964,5
BDNF,33,152.500000,187.933333,129.630492,5
PARK2,0,15.000000,13.733333,4.675197,30
SNCA,0,15.000000,13.166667,6.063363,30

Cross: newcross
3Balance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,289,452.500000,460.766667,93.716993,0
GAK,343,499.500000,500.466667,88.368637,0
MAOB,115,139.500000,147.000000,30.326954,0
PON1,16,62.500000,56.766667,21.495281,30
COMT,159,483.000000,475.000000,101.654248,0
GBA,319,504.000000,504.133333,88.618412,0
BST1,16,28.000000,30.633333,9.034353,30
DRD2,124,132.000000,133.200000,6.784363,0
LRRK2,210,493.500000,504.366667,122.899870,0
MAPT,21,92.000000,98.600000,35.321772,22
PINK1,15,34.500000,101.666667,141.412659,23
APOE,179,525.500000,518.400000,112.357066,0
BDNF,229,467.000000,484.266667,91.701891,0
PARK2,14,15.000000,15.666667,1.787569,30
SNCA,0,15.000000,14.900000,4.451656,30
3HighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,147,431.000000,406.166667,97.059556,0
GAK,324,454.500000,464.166667,85.591284,0
MAOB,115,145.000000,192.133333,96.172522,0

PON1,15,65.000000,56.900000,19.511889,30
COMT,196,414.000000,432.100000,88.611648,0
GBA,179,404.500000,398.966667,110.538957,0
BST1,0,28.500000,28.166667,10.888410,30
DRD2,121,134.500000,135.833333,10.079283,0
LRRK2,140,443.500000,461.533333,115.707936,0
MAPT,25,92.500000,105.200000,61.701394,19
PINK1,0,234.000000,214.233333,187.877378,14
APOE,219,471.500000,464.466667,94.292185,0
BDNF,154,413.500000,421.433333,91.242641,0
PARK2,0,16.000000,16.333333,5.228129,30
SNCA,0,15.000000,14.533333,4.629242,30
3HighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,185,483.000000,484.866667,113.347721,0
GAK,390,514.500000,535.833333,99.751444,0
MAOB,117,137.000000,140.666667,21.598744,0
PON1,21,55.000000,56.466667,21.074144,30
COMT,324,504.000000,518.866667,103.818055,0
GBA,315,527.500000,523.066667,97.576472,0
BST1,18,27.000000,29.566667,9.320661,30
DRD2,120,130.000000,130.500000,5.876605,0
LRRK2,267,487.500000,497.333333,104.745779,0
MAPT,28,81.000000,79.000000,38.964175,28
PINK1,17,40.000000,106.700000,120.172907,23
APOE,151,470.000000,477.100000,100.218847,0
BDNF,398,540.500000,551.400000,92.926593,0
PARK2,0,15.000000,14.666667,5.516954,30
SNCA,14,15.000000,16.266667,2.066704,30
Balance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,199,436.500000,440.400000,88.427878,0
GAK,353,444.000000,451.200000,62.541242,0
MAOB,120,153.000000,198.233333,101.358253,0
PON1,16,62.500000,55.300000,19.738200,30
COMT,231,456.000000,453.100000,90.477412,0

GBA,309,452.000000,471.133333,83.154088,0
BST1,0,32.000000,32.133333,11.361561,30
DRD2,125,140.000000,143.700000,11.377988,0
LRRK2,346,441.000000,455.666667,81.505800,0
MAPT,100,205.500000,210.000000,89.342811,0
PINK1,34,100.000000,211.233333,178.355454,17
APOE,184,436.500000,439.966667,88.011160,0
BDNF,185,457.500000,451.666667,94.300815,0
PARK2,0,15.000000,15.666667,3.726266,30
SNCA,0,15.000000,15.633333,5.034251,30
HighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,320,369.000000,392.133333,57.895109,0
GAK,332,413.000000,409.400000,62.117575,0
MAOB,126,203.000000,245.233333,126.924817,0
PON1,15,49.000000,51.300000,19.248197,30
COMT,258,361.500000,389.266667,62.708374,0
GBA,165,403.000000,402.666667,91.185046,0
BST1,19,33.500000,35.266667,12.250076,30
DRD2,127,150.500000,149.300000,10.609527,0
LRRK2,327,400.000000,400.000000,55.702411,0
MAPT,98,130.000000,159.466667,60.783808,2
PINK1,15,302.500000,257.400000,177.907880,10
APOE,319,428.500000,433.933333,56.257985,0
BDNF,201,418.000000,419.966667,82.898976,0
PARK2,0,15.000000,15.533333,4.695804,30
SNCA,14,17.000000,17.500000,2.956582,30
HighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,316,479.500000,482.800000,85.758040,0
GAK,392,491.000000,502.633333,67.030864,0
MAOB,125,142.000000,152.200000,35.282505,0
PON1,15,60.500000,57.200000,20.988338,30
COMT,213,497.000000,502.666667,118.241813,0
GBA,399,483.000000,504.100000,68.840821,0
BST1,16,30.000000,33.400000,12.472314,30

DRD2,125,138.000000,136.666667,8.635665,0
LRRK2,407,516.000000,520.500000,79.233287,0
MAPT,100,187.500000,210.066667,102.796060,0
PINK1,25,45.500000,55.666667,26.290595,30
APOE,174,520.500000,518.800000,112.764142,0
BDNF,214,492.500000,497.866667,92.243581,0
PARK2,0,15.000000,16.000000,4.008611,30
SNCA,0,15.000000,15.266667,5.698961,30
akBalance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,120,424.500000,414.000000,102.098668,0
GAK,105,299.000000,317.700000,131.972346,0
MAOB,15,32.000000,31.033333,6.435266,30
PON1,0,67.000000,54.466667,22.955329,30
COMT,225,357.500000,368.000000,91.435298,0
GBA,259,390.000000,417.833333,83.749661,0
BST1,0,24.000000,23.300000,5.421223,30
DRD2,22,100.000000,76.633333,34.979288,9
LRRK2,194,419.500000,426.433333,85.215622,0
MAPT,22,98.000000,96.800000,31.678776,27
PINK1,0,23.000000,22.266667,5.105327,30
APOE,103,323.500000,300.733333,148.574823,0
BDNF,103,238.500000,222.800000,87.886055,0
PARK2,0,17.000000,14.500000,7.744854,30
SNCA,0,15.000000,11.600000,7.976690,30
akHighCross:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,229,390.000000,382.166667,85.200406,0
GAK,119,304.000000,302.266667,108.833290,0
MAOB,0,40.500000,43.200000,16.857210,30
PON1,30,67.000000,61.800000,13.246730,30
COMT,218,321.000000,339.033333,79.145646,0
GBA,280,361.000000,364.533333,65.016037,0
BST1,0,23.000000,21.866667,6.431246,30
DRD2,17,100.000000,70.600000,37.531228,11
LRRK2,107,384.500000,381.500000,90.958782,0

MAPT,20,94.500000,84.000000,21.706506,28
PINK1,0,23.500000,20.233333,9.313258,30
APOE,101,272.500000,285.266667,128.082122,0
BDNF,108,239.500000,240.933333,93.046126,0
PARK2,0,15.000000,13.633333,7.402159,30
SNCA,0,15.000000,12.433333,8.045917,30
akHighMut:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,152,443.500000,433.966667,121.539970,0
GAK,112,396.000000,372.500000,116.547090,0
MAOB,0,30.000000,28.266667,7.597338,30
PON1,17,65.000000,59.300000,17.356605,30
COMT,109,412.000000,409.700000,116.634753,0
GBA,192,445.000000,435.833333,92.027014,0
BST1,0,22.500000,21.633333,5.397850,30
DRD2,27,100.000000,97.733333,13.364733,1
LRRK2,103,455.000000,431.066667,108.051052,0
MAPT,42,99.000000,95.033333,13.137163,26
PINK1,0,22.000000,19.300000,8.816716,30
APOE,102,253.000000,258.966667,103.608175,0
BDNF,102,261.500000,272.566667,141.118850,0
PARK2,0,16.000000,15.766667,6.026856,30
SNCA,0,16.000000,14.500000,6.981503,30
sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,259,395.500000,412.933333,78.693615,0
GAK,143,332.500000,354.733333,120.965636,0
MAOB,0,32.000000,31.166667,8.952300,30
PON1,0,60.500000,56.133333,19.236699,30
COMT,101,307.500000,293.466667,121.676715,0
GBA,325,404.000000,412.800000,70.948451,0
BST1,0,23.000000,19.233333,9.186509,30
DRD2,100,100.000000,100.500000,0.629724,0
LRRK2,108,440.000000,440.100000,104.554012,0
MAPT,100,138.000000,168.700000,76.395590,0
PINK1,0,22.500000,20.566667,6.719418,30

```
APOE,100,225.000000,215.400000,102.976261,0
BDNF,100,115.000000,137.500000,43.697826,0
PARK2,0,16.000000,14.700000,7.139738,30
SNCA,0,16.000000,14.866667,6.306720,30
sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
CYP2D6,256,422.500000,413.133333,79.341572,0
GAK,201,358.000000,373.766667,102.303026,0
MAOB,20,33.000000,32.766667,5.418254,30
PON1,0,58.500000,56.666667,22.921656,30
COMT,109,352.000000,351.433333,114.137148,0
GBA,185,421.500000,421.366667,86.908387,0
BST1,0,24.000000,22.366667,5.281610,30
DRD2,0,22.000000,23.366667,15.182756,29
LRRK2,240,398.000000,400.333333,71.439549,0
MAPT,17,85.000000,73.133333,26.385384,30
PINK1,0,24.000000,23.666667,7.023769,30
APOE,105,340.500000,315.566667,119.705582,0
BDNF,111,243.500000,253.800000,111.605957,0
PARK2,0,16.000000,15.100000,5.554433,30
SNCA,0,16.000000,14.166667,7.719583,30
-----------------------------------------------
```

A.3 Alzheimer's Summary

```
-----------------------------------------------
Disease: alz
Cross: base
3Balance:
Gene,Best,Median,Mean,Std,LOOs
HFE,8,46.000000,77.666667,92.793182,26
NOS3,18,426.000000,313.200000,168.654595,7
PLAU,0,32.000000,64.933333,95.246299,28
A2M,8,48.500000,91.800000,100.335781,25
MPO,8,96.000000,173.533333,164.734422,11
APP,8,335.000000,263.866667,191.641500,11
```

PSEN1,8,112.500000,218.400000,200.534561,11
PSEN2,0,84.500000,169.433333,173.646824,20
APOE,0,36.000000,80.366667,111.122326,26
Balance:
Gene,Best,Median,Mean,Std,LOOs
HFE,9,97.500000,118.200000,91.697554,22
NOS3,0,362.000000,313.300000,169.104977,7
PLAU,8,33.500000,81.266667,116.292418,27
A2M,0,45.000000,138.900000,164.188989,22
MPO,8,129.000000,203.333333,171.480839,15
APP,0,357.500000,270.100000,181.387060,11
PSEN1,8,96.500000,194.733333,177.319881,19
PSEN2,0,51.000000,145.900000,166.821678,21
APOE,0,63.000000,98.933333,101.599258,27
akBalance:
Gene,Best,Median,Mean,Std,LOOs
HFE,0,24.500000,34.066667,46.284676,30
NOS3,0,9.000000,9.000000,3.107277,30
PLAU,0,19.000000,19.233333,15.021096,30
A2M,0,0.000000,2.400000,4.064989,30
MPO,13,72.000000,91.800000,84.220319,27
APP,0,12.000000,14.633333,9.041984,30
PSEN1,0,47.000000,66.600000,72.374314,29
PSEN2,0,11.000000,10.166667,5.669540,30
APOE,0,0.000000,4.000000,4.785034,30
sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
HFE,9,22.000000,33.366667,25.932915,30
NOS3,0,0.000000,3.166667,4.259540,30
PLAU,0,17.500000,20.600000,12.310915,30
A2M,0,0.000000,1.166667,3.040909,30
MPO,15,50.500000,82.466667,102.583770,23
APP,0,8.500000,7.566667,4.240066,30
PSEN1,9,42.000000,42.600000,34.139269,30
PSEN2,0,8.000000,6.733333,4.877346,30
APOE,0,0.000000,2.166667,4.161261,30

sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
HFE,9,21.000000,25.966667,14.782290,30
NOS3,0,11.000000,10.800000,3.872093,30
PLAU,8,15.500000,16.800000,7.694154,30
A2M,0,0.000000,3.300000,4.442351,30
MPO,19,53.500000,69.433333,48.239119,28
APP,9,16.500000,22.033333,12.092983,30
PSEN1,8,44.500000,84.066667,96.279741,27
PSEN2,0,13.000000,12.333333,5.591455,30
APOE,0,8.000000,6.133333,4.658943,30

Cross: newcross
3Balance:
Gene,Best,Median,Mean,Std,LOOs
HFE,0,20.500000,20.433333,11.437034,30
NOS3,0,18.500000,16.466667,9.420203,30
PLAU,0,21.000000,20.100000,9.848333,30
A2M,0,11.000000,10.633333,6.562607,30
MPO,17,100.500000,144.333333,90.128771,14
APP,0,18.000000,18.100000,7.734161,30
PSEN1,0,54.000000,56.566667,42.541406,29
PSEN2,0,10.000000,9.066667,8.270985,30
APOE,0,10.000000,11.500000,6.647530,30
Balance:
Gene,Best,Median,Mean,Std,LOOs
HFE,0,17.500000,16.200000,7.617584,30
NOS3,0,9.000000,10.400000,5.021334,30
PLAU,0,16.000000,15.233333,5.405510,30
A2M,0,0.000000,5.033333,6.435266,30
MPO,0,57.500000,52.300000,24.265983,30
APP,0,8.000000,9.366667,8.164192,30
PSEN1,0,34.000000,33.000000,10.888146,30
PSEN2,0,10.000000,10.166667,6.086069,30
APOE,0,9.000000,9.200000,7.485158,30
akBalance:

```
Gene,Best,Median,Mean,Std,LOOs
HFE,0,18.500000,15.133333,8.381973,30
NOS3,0,0.000000,5.466667,7.426575,30
PLAU,0,10.500000,9.666667,8.326664,30
A2M,0,0.000000,3.533333,5.981197,30
MPO,0,44.500000,44.100000,19.046857,30
APP,0,8.000000,8.566667,8.037341,30
PSEN1,0,33.500000,29.433333,14.928989,30
PSEN2,0,0.000000,6.133333,7.775884,30
APOE,0,0.000000,3.433333,6.162642,30
sBetBalance:
Gene,Best,Median,Mean,Std,LOOs
HFE,0,10.000000,11.066667,8.630073,30
NOS3,0,0.000000,3.966667,6.338080,30
PLAU,0,10.500000,10.566667,7.981372,30
A2M,0,0.000000,1.500000,3.421534,30
MPO,0,47.000000,42.800000,23.752169,30
APP,0,0.000000,3.800000,6.053782,30
PSEN1,0,23.000000,22.466667,12.372755,30
PSEN2,0,4.000000,5.933333,6.705290,30
APOE,0,0.000000,4.200000,6.748691,30
sStressBalance:
Gene,Best,Median,Mean,Std,LOOs
HFE,0,14.500000,11.466667,9.587108,30
NOS3,0,9.000000,9.566667,7.959741,30
PLAU,0,15.000000,12.866667,8.427105,30
A2M,0,0.000000,3.166667,5.711956,30
MPO,0,62.500000,53.766667,21.884782,30
APP,0,9.000000,9.066667,7.570056,30
PSEN1,0,40.500000,38.700000,15.778969,30
PSEN2,0,9.000000,9.800000,8.675769,30
APOE,0,0.000000,5.300000,6.555598,30
------------------------------------------------
```