Network Centralities and the Retention of Genes Following Whole Genome Duplication in *Saccharomyces cerevisiae*

by

Matthew J. Imrie
B.Sc., University of Victoria, 2010

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Interdisciplinary Studies

Network Centralities and the Retention of Genes Following Whole Genome Duplication in *Saccharomyces cerevisiae*

by

Matthew J. Imrie
B.Sc., University of Victoria, 2010

Supervisory Committee

Dr. Ulrike Stege, Supervisor
(Department of Computer Science)

Dr. John Taylor, Co-Supervisor
(Department of Biology)

Dr. Alex Thomo, Co-Supervisor
(Department of Computer Science)

Supervisory Committee

Dr. Ulrike Stege, Supervisor
(Department of Computer Science)

Dr. John Taylor, Co-Supervisor
(Department of Biology)

Dr. Alex Thomo, Co-Supervisor
(Department of Computer Science)

ABSTRACT

The yeast *Saccharomyces cerevisiae* genome is descendant from a whole genome duplication event approximately 150 million years ago. Following this duplication many genes were lost however, a certain class of genes, termed ohnologs, persist in duplicate. In this thesis we investigate network centrality as it relates to ohnolog retention with the goal of determining why only certain genes were retained. With this in mind, we compare physical and genetic interaction networks and genetic and protein sequence data in order to reveal how network characteristics and post-duplication retention are related. We show that there are two subclasses of ohnologs, those that interact with their duplication sister and those that do not and that these two classes have distinct characteristics that provide insight into the evolutionary mechanisms that affected their retention following whole genome duplication. Namely, a very low ratio of non-synonymous mutations per non-synonymous site for ohnologs that retain an interaction with their duplicate. The opposite observation is seen for ohnologs that have lost their interaction with their duplicate. We interpret this in the following way: ohnologs that have retained their interaction with their duplicate are functionally constrained to buffer for the other ohnolog. For this reason they are retained; ohnologs that have lost their interaction with their duplicate are retained because they are functionally divergent to the point of being individually essential.

Additionally we investigate small scale duplications and show that, generally, the mechanism of duplication (smale scale or whole genomes) does not affect the distribution of network characteristics. Nor do these network characteristics correlate to the selective pressure observed by retained paralogous genes, including both ohnologs and small scale duplicates. In contrast, we show that the network characteristics of individual genes, particularly the magnitude of their physical and genetic network centralities, do influence their retention following whole genome duplication.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

For me, it is far better to grasp the Universe as it really is than to persist in delusion, however satisfying and reassuring.
Carl Sagan

DEDICATION

To K.M. for convincing me to finish things left unfinished.

Chapter 1

Introduction

Approximately 150 million years ago a whole genome duplication (WGD) event occurred within an ancestor of *Saccharomyces cerevisiae* [32, 22, 5]. In 1970 Susumo Ohno postulated that this type of duplication is a major contributing factor to the generation of novel genetic material [44] and contributes to "genomic redundancy, specialization, degeneration, innovation and speciation" [7]. This process has been called a revolutionary event [13], where following duplication the genome goes through a period of genetic upheaval as genes [56, 23] and entire chromosomes are lost [41, 3, 8].

While this process of genome duplication and gene loss results in most duplicate genes losing one of both copies within a few million years after the duplication event [39], there is a class of genes, termed *ohnologs* [58], that survive in duplicate.

What determines gene fate? Recently, large sets of interaction data for *S. cerevisiae* have bene made available. These interactions, coupled with the evolutionary history of duplicates may provide additional understanding of what determines gene fate following duplication.

1.1 Motivation

Our motivation is to understand why, following whole genome duplication (WGD), some genes are retained to the present day, whereas others are not. All ohnolog duplicates have survived millions of years within the yeast genome and compose a large fraction of all yeast genes today (18%) [5]. This also implies that a large number of duplicate genes have been lost during this time. This leads to the inevitable question: *why have some ohnologs been retained while other genes no longer have a* *duplicate originating from the whole genome duplication event?* Furthermore, there also exists duplicates independent of WGD termed small scale duplicates (SSDs), which are pairs or groups of genes that have been duplicated. Identifying and utilizing these in addition to WGDs may provide insight to why differential retention exists.

Of particular assistance is that relatively recent literature has identified many genes arising from the WGD of *S. cerevisiae* [32, 5]. Additionally, there is a well annotated database of the *S. cerevisiae* genome [16] that contains both nucleotide and amino acid sequences for genes and their protein products, respectively.

1.2 Research Questions and Contributions

Our novel approach is to utilize social network analysis techniques to investigate retention. These techniques, specifically centrality measures, will be applied to maps of interactions between genes in *S. cerevisiae* that have recently been published or updated. Both physical interactions [53] and genetic interactions [9] will be considered.

In order to answer the overarching question *why have some ohnologs been retained while other genes no longer have a duplicate originating from the whole genome duplication event?*, while using a network centrality specific methodology, we derived the following four questions about *S. cerevisiae* paralog evolution:

1. Does the mechanism of duplication, whole genome or small scale duplication, correlate to distribution of network centrality measures for duplicated genes?

2. Is there a correlation between the change in network centrality measures between paralogous pairs and the selective pressure experienced after duplication?

3. If there is a correlation, is this correlation different for small scale duplicates compared to ohnologs?

4. Does the retention of ohnologs correlate with network centrality measures?

Our first contribution is a set of software tools, called "**Ne**twork **C**entrality and **Pa**ralog **D**ivergence **I**ntegrator" (NeC-PaDI). Nec-PADI integrates network, sequence and paralog data—for any organism, contingent on a standardized input format. We utilized these tools to associate centrality measures, nucleotide and amino acid sequence data, and duplication relationships to each gene in our yeast gene set. These data were then queried to produce our further contributions. It can be downloaded from `http://webhome.csc.uvic.ca/~imrie/necpadi`.

Our main contribution shows that there are two classes of ohnologs and that each have a different profile of retention following duplication. The first of these two classes are those that genetically interact with their duplication sister. The second class are those ohnologs that have do not. We show that ohnolog pairs that do have a genetic interaction with their duplication sister have higher duplicate pair retention at higher centrality values. We also show that they exhibit a lower ratio of non-synonymous mutations per non-synonymous site which indicates their sequences have changed relatively little. The fact that they share a genetic interaction with one another shows that they have retained an ability to buffer one another as if there was no buffering existed then no genetic interaction would be observed.

For ohnologs that have lost a genetic interaction with their duplication sister we find higher retention at lower centrality values and that they exhibit a higher ratio of non-synonymous mutations per non-synonymous site. With this loss of genetic interaction we interpret their retention as being due to functional divergence resulting in indispensability. This, based on previous literature, we interpret as subfunctionalization[26, 12].

Secondary contributions show that evolutionary pressure does not correlate to centrality measures nor does the distribution of centrality measures change significantly based on the type of duplication. We also show that the type of interaction (genetic or physical) has an effect on the distribution of centrality measures. This is due to the physical and genetic interaction networks describing different characteristics of the same genes.

Our central hypothesis is that ohnolog pairs will be retained at higher centrality measures compared to lower centrality measures. We base this hypothesis on the fact that centrality positively correlates with essentiality[46]. By extension if a gene is essential it must be retained. Therefore, at higher centrality measures there should be a higher fraction of retained duplicates.

1.3 Thesis Overview

The subsequent chapters of this thesis are structured as follows. In Chapter 2 we provide an overview of the related topics required to fully understand our methodology, results and interpretations, including definitions from biology and computer science. Here we introduce much of the terminology used through this thesis, but provide expanded detail in the appendices. We present a review of related research in

Chapter 3. We then move on to our methodology and first contribution in Chapter 4, which describes the software tools and methods we developed to pursue our research questions. Our second contribution of this thesis is presented in Chapter 5, where we begin our analyses, presenting results to answer each of our four research questions. We conclude in Chapter 6 with a summary of our findings and provide avenues for future work. Throughout this thesis we use concepts from biology, statistics and social network analysis that may be unfamiliar to the read. We provide detailed explanations of the most fundamental in the appendices which follow Chapter 6: Appendix A is devoted to an in depth overview of the biological concepts that we mention in Chapter 2; Appendix B provides a review of the statistical methods we utilized in our analyses; finally, Appendix C contains examples on calculating network centralities as used in this thesis.

Chapter 2

Related Topics

Our research covers two very different realms of science. On one hand we are concerned with evolutionary biology, and on the other, with computer science, specifically network analysis. In order to understand our reasoning, methodology, results and interpretation we review the most necessary of topics in order to proceed.

We begin by discussing gene duplication and then whole genome duplication. Our goal in this thesis is to investigate whether certain network traits of genes indicate why both duplicates are retained. Since we chose to investigate this topic with a hypothesis that the interactions between genes play a direct role in retention, we discuss both physical and genetic interaction networks. We chose to investigate using both network types as they show different aspects of the same genes. We explain the differences between these two network types, and how their data are generated. Finally, we introduce and explain the topics of graphs and the centrality measures we will be using in our analyses.

2.1 Gene Duplication

Gene duplication is a major contributing factor in the development of novel genetic material and genetic redundancy [44]. The basis of this concept is that a single ancestral gene duplicates and thus two related genes are created. These are called paralogs. At the time immediately following duplication, each gene in a duplcate pair is identical in their nucleotide sequence. Over time, and depending on the evolutionary forces to which each is subjected, these nucleotide sequences will change. This change is known as sequence divergence. As sequences diverge, function also diverges such that over sufficient time two homologs by have very different functional roles. Eventually, any signature of their relatedness may also disappear.

The term *homolog* is an all encompassing word for genes that are related. There are two possible way in which genes may be related: speciation, and duplication, termed paralogs. Paralogs can be further categorized by whether the duplication was at a small scale, involving individual or groups of genes, or at a large scale, where an entire genome is duplicated. These relationships each have visualized in Figure 2.1

Orthologs are genes that are related due to speciation. Individual genes in two different species are orthologs if the ancestry of each gene can be traced to a single gene in the most recent common ancestor of the two species.

Paralogs arose from a physical duplication event within a species: that is, some biological process that produced two identical copies of a gene within a species. Within this thesis, paralogs arising from a single gene duplication are termed *small scale duplicates*(SSDs). Paralogs that arose from a whole genome duplication are termed *ohnologs*, in honour of Susumu Ohno[58]. Ohno was not the first to ponder the relevance of gene duplication in general as the idea has existed since the early 20th century [55]. However, he brought the idea of genome duplication being essential for higher eukaryote evolution to the forefront.

Fate of Duplicate Genes

Broadly, there are four possible fates for duplicate genes. Each either subfunctionalizes, in relation to their common ancestor, neofunctionalizes, or is lost [17]. We will focus on explaining the terms subfunctionalization and neofunctionalization.

The term *subfunctionalization* [17] is a phenomenon that occurs when both duplicates partition the function of their common ancestor between themselves. Some of these functions may be common, where as each may retain different functions. Those functions that are retained by both are dosage sensitive in the opposite way than those functions differentially retained to one or the other. When both duplicates retain an ancestral function or interaction, a reduction in the relative amount of this function in the cell is detrimental. Therefore, both duplicates retain the function. Where there is a differential distribution of functions between duplicates, there is a gene dosage that is too high and is therefore detrimental to the cell.

The process of neofunctionalization occurs when one of a pair of duplicates is unrestrained by selection and able to mutate without detrimental effects to the or-

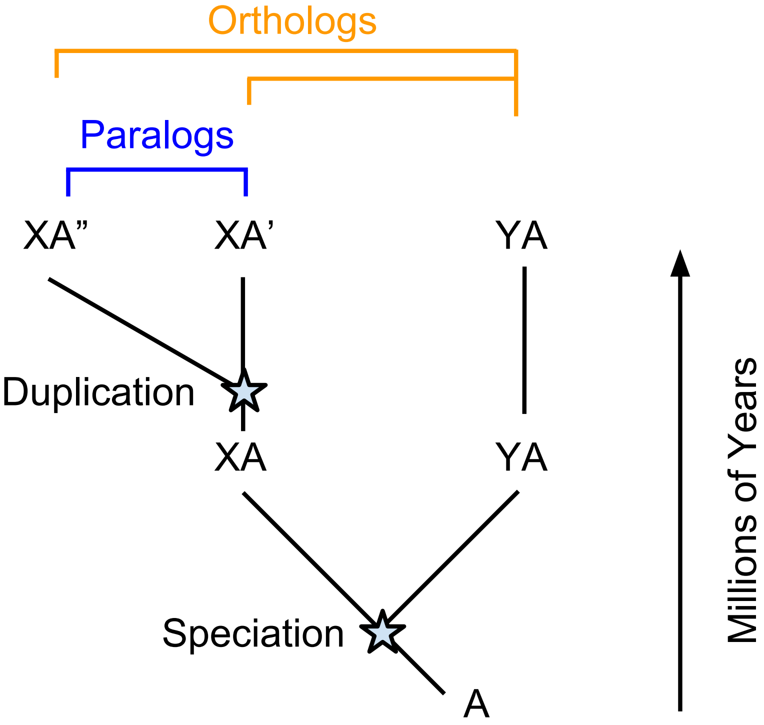


Figure 2.1: Evolutionary history of related genes *XA'*, *XA'*, and *YA*. A common ancestor gene *A* existed some million years in the past. A speciation event created two different lineages for the descendants of A: the lineage of *XA* and the lineage of *YA*. After some period of time the species harbouring gene *XA* had a duplication event that duplicated *XA* into *XA'* and *XA"*. Therefore, *YA* and either *XA'* or *XB"* are orthologs. *XA'* and *XA"* are paralogs. If *XA'* and *XA"* are the result of a WGD they are ohnologs. Alternatively, if *XA'* and *XA"* are the result of a single duplication they are termed SSDs. In addition, both *XA'* and *XA"* are co-orthologs of *YA* whereas *YA* is a pro-orthlog of *XA'* and *XA"*. All extant genes are homologs to one another.

ganism. Using dosage for our explanation, this implies that the required dosage for the ancestral function can be met by one of the duplicate pairs. This allows the other of the pair to mutate and possibly acquire novel function, while the other of the pair maintains its ancestral function. This is the original theory of gene fate proposed by Susumo Ohno when discussing WGD. However, the prevalence of neofunctionalization has been questioned in recent literature [20] as neofunctionalization can be explained using iterative rounds of subfunctionalization and loss. and subfunctionalization is simply because interactions arising from subfunctionalization are obfuscated due to the age of duplicates and subsequent loss of interactions between individual genes–providing the signature of neofunctionalization not subfunctionalization.

2.1.1 Whole Genome Duplication

WGD is gene duplication on a massive scale: that of the entire genome. WGD physically doubles the number of chromosomes in an organism, with the effect of doubling (or nearly doubling) the number of genes in the genome. As each gene interacts with some other combination of genes, this doubling produces a intricate network of interacting genes and regulatory components, with some interactions beneficial and some detrimental. This results in an evolutionary state termed "genomic resolution" [38] where the genome undergoes an extensive restructuring in order to regain stability and return to a state where massive gene loss has ceased. From this point, and subsequently over the eons, this new genomic landscape is shaped and sculpted by a multitude of evolutionary, genetic, epigenetic and molecular forces.

Previous literature has explained that the most important non-random determinant of individual gene survival is gene dosage [7, 11, 12, 23, 26, 39, 45, 49, 57]. Gene dosage is the concept that the relative amounts of individual genes are in a specific balance with one another—too much or too little of one will have a cascading effect, possibly detrimental, on the expression patterns of many other genes. Being subjected to a WGD results in a massive change in the genome but little change in gene dosage as the relative amounts of genes are unchanged. However, the act of duplication results in dramatic gene loss and rearrangement [13, 38] as well as epigenetic silencing [37, 40] in order to return the genome to a reproductively and genomically stable state.

WGD in *S. cerevisiae*

The theory that the *S. cerevisiae* genome is the result of a whole genome duplication was initially proposed in 1997 [59]. This idea proved contentious until 2004 when Kellis *et al.* provided definitive proof of a paleo-duplication event approximately 100-150 million years ago [32]. This proof was found by using the genome of the related organism *Kluyveromyces waltii* to search for regions of doubly conserved synteny with *S. cerevisiae.* Conserved synteny is the physical feature between two chromosomal regions that homologous genes between each region follow the same order [32]. When the order of genes on one chromosome matches the order of homologous genes in two separate chromosomal regions, we call this doubly conserved synteny (DCS).

Using DCS to identify homologous regions, Kellis *et al.* found that the 16 *S. cerevisiae* chromosomes mapped to eight *K. waltii* chromosomes. This mapping covered 82% of *S. cerevisiae* genes and 75% of *K. waltii* genes. These mappings are not contiguous, not all of *K. waltii* chromosome 1 maps to the entirety of *S. cerevisiae* chromosome 4 and 12. Rather, portions of each ancestral chromosome are scattered through the 16 *S. cerevisiae* chromosomes. This is due to the divergence of the *K. waltii* and *S. cerevisaie* genomes as they have been subjected to very different combinations of chromosomal reordering and rearrangement.

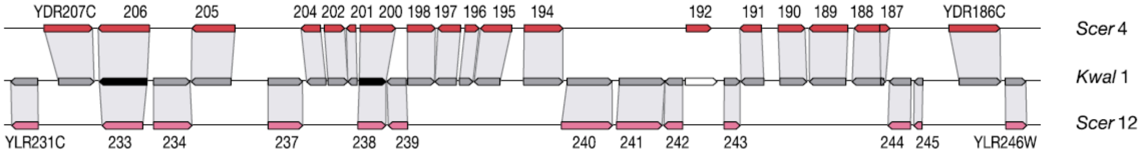


Figure 2.2: Double conserved synteny between *K. waltii* Chromosome 1 and *S. cerevisiae* Chromosome 4 and 12. Image from [32]

Following the identification of regions with DCS, Kellis *et al.* identified paralogs between them. These paralogs are genes retained in duplicate and derived from the WGD event. As noted above, these paralogs are called ohnologs because of their WGD origin. In total 457 ohnolog pairs were found by Kellis *et al.* However, differences in similar studies [10, 60] have increased the total number of unique ohnologs to a current total of 523 ohnologs [5].

Considering the state of a genome following WGD, the fact that 523 ohnologous pairs persist in the *S. cerevisiae* genome is a very interesting phenomenon and the vast majority of duplicates are lost very shortly after duplication [39]. Using gene dosage as a basis, there are three reasons why these genes may persist following duplication [12, 26]:

1. gene duplication provides a selective advantage, such as increased redundancy (the fact that pre-duplication functionality now is covered by two genes instead of one);

2. selection for an increased dosage; or,

3. duplicates subfunctionalize in order to remove dosage effects.

In this thesis we investigate whether network centrality measures are an addition to this list.

2.2 Physical and Genetic Interaction Networks

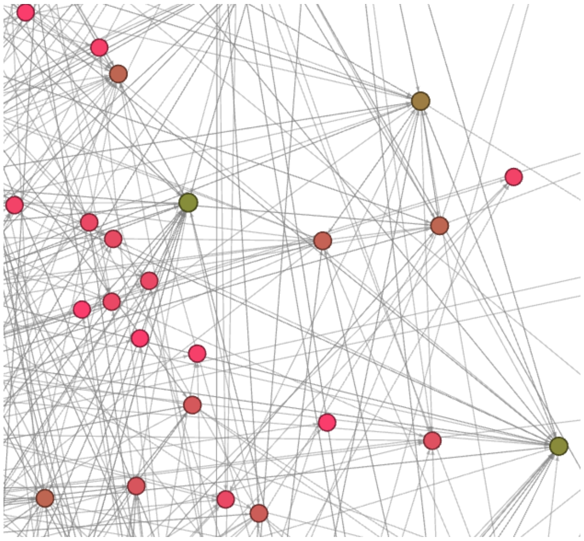


Figure 2.3: An example portion of an interaction network. Nodes are individual genes. Edges are interactions between genes. Depending on the network type these interactions can either be physical or genetic in nature.

In this thesis we utilize two different interaction network types for our analyses: physical and genetic. Physical interaction data was obtained from The BioGrid [53] (version 3.2.106) and genetic interaction data from Costanzo *et al.* [9].

Within *S. cerevisiae* there are approximately 6000 genes [16] and within this set of genes there are those that physically interact and those that interact genetically. It is these interactions, physical or genetic, that are collected into a global view that describe either a physical or genetic interaction network, respectively.

Each interaction type is fundamentally different and each shows different characteristics for the genes under study. Physical interactions occur when protein products physically bind to one another. Genetic interactions occur when the deletion of two genes causes an aggravated change in an observable phenotype when compared to that phenotype for single deletion mutants. These two interaction types are obtained in very different ways. Here we present a brief overview of how these are discovered, first physical interactions and then genetic interactions.

2.2.1 Physical Interactions

The physical interaction network is a collection of individual *in vitro* experiments that report physical interactions between two gene products. Although there are many different methods to identify physical interactions, three methods account for over 80% of reported physical interactions collated by "The BioGrid" [53]: affinity capture-MS, affinity capture-Western and Yeast Two Hybrid.

Although each method varies in how interactions are reported, they similarly rely on a "bait" protein and a "prey" protein. The two most similar methods, affinity capture-MS and affinity capture-Western, differ only in how their prey proteins are identified. These methods use a known bait protein affixed to a permeable substrate that is then washed with cellular extract. Those prey proteins within the cellular extract that interact sufficiently with the bait protein will be affixed to the substrate via the bait. Following this, the prey proteins are identified using mass-spectrometry (MS) or Western blotting. Although prey proteins can be identified by the mass-spectrography method using the protein's molecular weight, Western blotting requires antibodies specific to each prey protein [1]. This requires a great deal more time and effort, which is reflected in the relative proportion of Western identified interactions: 12% of all physical interactions compared to 56% for mass-spectrometry methods [53].

The third of the most common methods for identifying physical interactions is yeast two hybrid (YTH). Accounting for 15.5% of known physical interactions [53], this method is markedly different from the previous two in that it relies on an *in vivo* reporter fused to the query proteins [1]—this means that YTH isn't affect by an *in vitro* bias.

The fundamental YTH approach utilizes a protein, GAL4, which is responsible for activating genes required for galactose utilization [15]. GAL4 is composed of two parts, or domains, a DNA binding domain and an activation domain. In order for *S. cerevisiae* to metabolise galactose the DNA binding domain must be in close proximity to the activation domain. Typically, since both domains are on the same protein, this is not an issue and galactose metabolism is activated as required. However, the fundamental YTH method [15] creates *S. cerevisiae* mutants that lack GAL4. If these mutants are grown on a substrate with galactose as the only source of food, they will die–the ability to metabolise galactose has been removed and no other food source exists. In order to return the ability of galactose metabolism the two domains of GAL4, the DNA binding domain and the activation domain, are split apart. A "bait" protein is fused to the DNA binding domain and a "prey" protein fused to the activation domain. These are then added to the *S. cerevisiae* mutants lacking GAL4. If the "bait" and "prey" interact, they will bring the activation domain into close proximity of the DNA binding domain. The close proximity of the two domains will rescue galactose metabolism. Whether two proteins interact is directly based on whether the *S. cerevisiae* mutants survive, or not.

2.2.2 Genetic Interactions

Genetic interactions are an *unexpected phenotype* that cannot be explained by the multiplicative effect of individual genetic variants. Essentially, a genetic interaction exists if the observed phenotype of a double mutant is significantly greater or less than the multiplied fitness of single mutants. To explain in more detail, this method relies on measuring the effect of individual genetic variants on a particular phenotype, such as number of viable offspring or colony size. From this measurement a fitness metric is constructed for each variant; one, a query, i; and the other, a subject, j. Costanzo *et al.* used this concept as a basis for a multiplicative null model to predict the fitness of double mutants,

$$f_{ij} = f_i f_j + \epsilon_{ij} \tag{2.1}$$

where ϵ_{ij} is the deviation from the model, in this case $f_j f_j$, by the double mutant, f_{ij}. Therefore, the fitness of a double mutant f_{ij} is a product of the fitness of f_i multiplied by the fitness of f_j and unexpected results are described by ϵ_{ij}.

Colony size was then represented as a function of fitness and additional variables to compensate for time, noise and experimental effects,

$$C_{ij} = f_{ij} \cdot t \cdot s_{ij} \cdot e \tag{2.2}$$

where t is time, s_{ij} is a correction factor to account for systematic experimental biases that affect colony growth and e is an estimation of log-normally distributed random noise. By substituting Equation 2.1 into Equation 2.2 and rearranging,

$$\epsilon_{ij} = \frac{C_{ij}}{t s_{ij} e} - f_i f_j \tag{2.3}$$

Therefore, ϵ_{ij} is an measure of genetic interaction strength, which can be either negative or positive. A negative ϵ_{ij} indicates a decrease in fitness over the null model whereas a positive ϵ_{ij} indicates an increase in fitness over the null model. What determines the "unexpectedness" of this phenotype is the threshold of an interaction's ϵ_{ij}. This determines whether an interaction is included or excluded in the network. Costanzo *et al.* provided four thresholds ranging from any interactions which a p-value < 0.05 to a "stringent cutoff" of $\epsilon < -0.12$ or $\epsilon > 0.16$, p-value < 0.05. These p-values were calculated based on four replicates per double mutant and an estimate of the log-normal error distributions for the each of the query and subject (i.e single) mutants (details within the supplemental material of [9]).

2.3 Graphs

Definition of a Graph

A *graph* (*network*), $G = (V, E)$, consists of a set of *vertices* (*nodes*), V, and *edges*, $E \subseteq V \times V$, that connect pairs of vertices. For each $(u, v) \in E$ an edge (u, v) is said to be *incident* to its *end points* u and v. In the course of this thesis we use the terms "network" and "graph", as well as "node" and "vertex" interchangeably. An *undirected graph* $G = (V, E)$ is a graph where $(u, v) \in E \Leftrightarrow (v, u) \in E$. For the sake of simplicity, the networks we are analyzing will be considered undirected and unweighted. This means that all edges have an edge weight of 1, $\{w_{(u,v)} = 1 : \forall(u, v) \in E\}$. Additionally we allow *self-loops*, that is edges $(u, v) \in E$ with $u = v$.

Degree

The degree of a node, v, is defined as $deg(v)$ and is simply the number of nodes incident to it.

Shortest Path

A path from u to v in a network is a sequences of nodes $u, u_1, ...u_n, v$ such that each node in the sequence is connected to the next node in the sequence by an edge. Depending on the topology of the network there may be more than one path between any u and v.

The shortest path, or a geodesic, between a node u and v is the number of edges on a path, $d_G(u, v)$, between u and v such that $d_G(u, v)$ is minimized. There may be more than one shortest path between vertices u and v, but the number of edges in each shortest path between u and v is identical. The total number of different shortest paths from u to v is then σ_{uv}.

2.4 Network Analysis

Some of most powerful network analysis tools are measures of centrality as they provide an unbiased characterization of all nodes in a network. When given a network we can utilize one of many centrality measures to rank each node. We then use this ranking to determine the nodes' relative importance. There are a number of these centrality measures however we chose to utilize those that are most common in biological network analysis. Those being degree, closeness and betweenness centrality[34, 35, 25, 30, 29, 28, 46, 52]. Each of these where formalized by Freeman in [18] and will be the focus of our attention in this thesis. These topics are expanded on, with examples, in Appendix A.2.2.

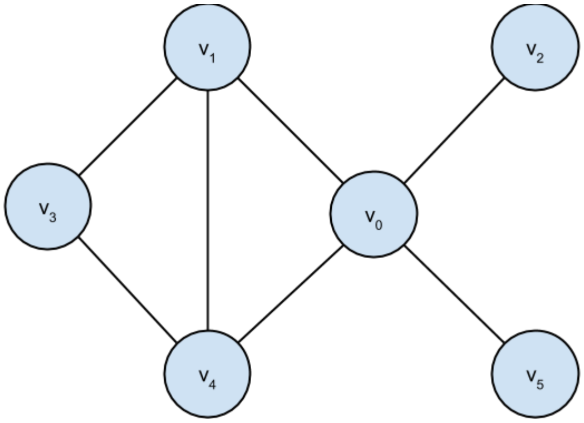


Figure 2.4: An example graph with six nodes and seven edges.

2.4.1 Degree Centrality

Conceptually, the simplest centrality measure for a node v is degree centrality, $C_D(v)$. This measure scores all nodes in a network by their degree. The degree, $C_D(v)$ of a node v, is simply the number of edges incident to it. It is important to note that Degree Centrality is a local metric, it is only a measure of the immediate neighborhood: those nodes directly adjacent to v. This contrasts to the other two metrics, betweenness and closeness, as they are global metrics, dependent on the topology of the entire graph.

$$C_D(v) = deg(v) \tag{2.4}$$

2.4.2 Betweenness Centrality

Betweenness centrality, $C_B(v)$, or more accurately *shortest path betweenness centrality*, for a node v, is a measure that scores the importance of v based on the number of shortest paths between any two nodes u and w of which v is a member. This is computed as a ratio of the number of shortest paths that pass through v and the total number of shortest paths in the network. More formally,

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.5}$$

where $s, t, v \in V$ and wehre $\sigma_{st}(v)$ is the number of shortest paths from s to t of which v is a member and where σ_{st} is the total number of shortest paths from s to t.

The above metric returns a value between 0 and $(|V| - 2)(|V| - 1)/2$. To see this, note that the maximum number of paths that can contain v is the total number of pairs of nodes, excluding v. This is precisely $(|V| - 2)(|V| - 1)/2$. We use this information in order calculate the normalized betweenness, $0 \leq C'_B(v) \leq 1$, as found in Equation 2.6.

$$C'_B(v) = \frac{C_B(v)}{(|V| - 2)(|V| - 1)/2} \tag{2.6}$$

2.4.3 Closeness Centrality

Closeness centrality is a measure of how close any node v is to any other node $t \in V$. It is the inverse of the related metric *farness*. Farness, $F(v)$, for a node v is the sum of geodesic distances to all other nodes $t \in V$,

$$F(v) = \sum_{t \in V} d_G(v, t) \tag{2.7}$$

Then, closeness centrality, $C_C(v)$, is defined as,

$$\begin{aligned} C_C(v) &= F(v)^{-1} \qquad (2.8) \\ &= \frac{1}{\sum_{t \in V} d_G(v,t)} \qquad (2.9) \end{aligned}$$

Unfortunately, the above metric only produces valid results when the (entire) graph is connected. For non-connected graphs there exists a pair with $d_G(uv) = \infty$. To circumvent this limitation, where graphs are not connected, we modify this metric and ignore all nodes, t, not reachable from some v by setting $d_G(v,t) = 0$.

However, the resultant values of Equation 2.9 can become quite large so it is useful to normalize them so that $0 \leq C'_C(v) \leq 1$. To do this we must consider the maximum closeness a node can have. Using Equation 2.9 we can see that the maximum closeness is when a node is connected to every other node, $|V| - 1$.

Therefore, to normalize Equation 2.9,

$$C'_C(v) = \frac{|V| - 1}{\sum_{t \in V} d_G(v,t)} \qquad (2.10)$$

Chapter 3

Methodology and Software Tools

3.1 Introduction

In this chapter we describe the methodology used to generate the data that we used to answer our research questions:

1. Does the mechanism of duplication, whole genome or small scale duplication, correlate to distribution of network centrality measures for duplicated genes?

2. Is there a correlation between the change in network centrality measures between paralogous pairs and the selective pressure experienced after duplication?

3. If there is a correlation, is this correlation different for small scale duplicates compared to ohnologs?

4. Does the retention of ohnologs correlate with network centrality measures?

To get the broadest understanding of how genes and their interaction change following duplication we chose to analyze both physical and genetic interaction networks. The physical interaction network is a network of qualifiable physical interactions between pairs of gene products (proteins). This differs from the genetic interaction network which is a network of quantifiable genetic effects. It is important to note that just because a gene product interacts in the physical interaction network, it does not mean that the two genes interact in the genetic interaction network, and *vice versa*.

We used these two networks to assist us in answering our questions above. However, in order to make our results from these two networks relatable (i.e., an "apples-to-apples" comparison), we processed the networks by removing all nodes that were not common to both (and by default, interactions with these deleted nodes). Secondly, within this common set of nodes, we identified pairs that are paralogous, and sorted these pairs into both whole genome duplicates (ohnologs) and small scale duplicates (SSDs). Thirdly, we calcuated the network characteristics of ohnologs and SSDs within each network type: degree, closeness and betweenness. Finally, we measured the selective constraints affecting SSDs and ohnologs in each type of network. These steps are detailed below.

3.2 Constructing Comparison Networks

3.2.1 Network Data

We were interested in the differences, if any, in the physical and genetic interactions of ohnologs and SSDs. As such, we utilized physical and genetic interaction network data as detailed below.

Physical Interaction Network Data

Physical interaction network data was downloaded from *The BioGrid* [53]. We first cleaned the data of all genetic interactions, labeled “genetic”, similar to [26] as well as any interactions involving anything other than *S. cerevisiae* proteins including RNA interactions arising from Affinity Capture-RNA and Protein-RNA. The remaining interactions were solely between proteins or protein components, and we refer to this cleaned data set as the *Protein Interaction Network*.

Genetic Interaction Network Data

We obtained our genetic interaction data from the supplementary material of Costanzo *et al.* [9]. To minimize spurious interactions we chose to use the “stringent cutoff” synthetic genetic array analysis data-set ($\epsilon < -0.12$, p-value < 0.05 or $\epsilon > 0.16$, p-value < 0.05). This data set had an overall precision of 0.89 for negative interactions and 1 for positive interactions. The total number of interactions were decreased due to a reduction in sensitivity but the reduction in sensitivity was much less than the gain in precision [9].

Genes represented by more than one allele (334 alleles), either “TSQ” (Thermo-Sensitive Query, 214 alleles) or “DAmP” (decreased abundance by mRNA perturba-

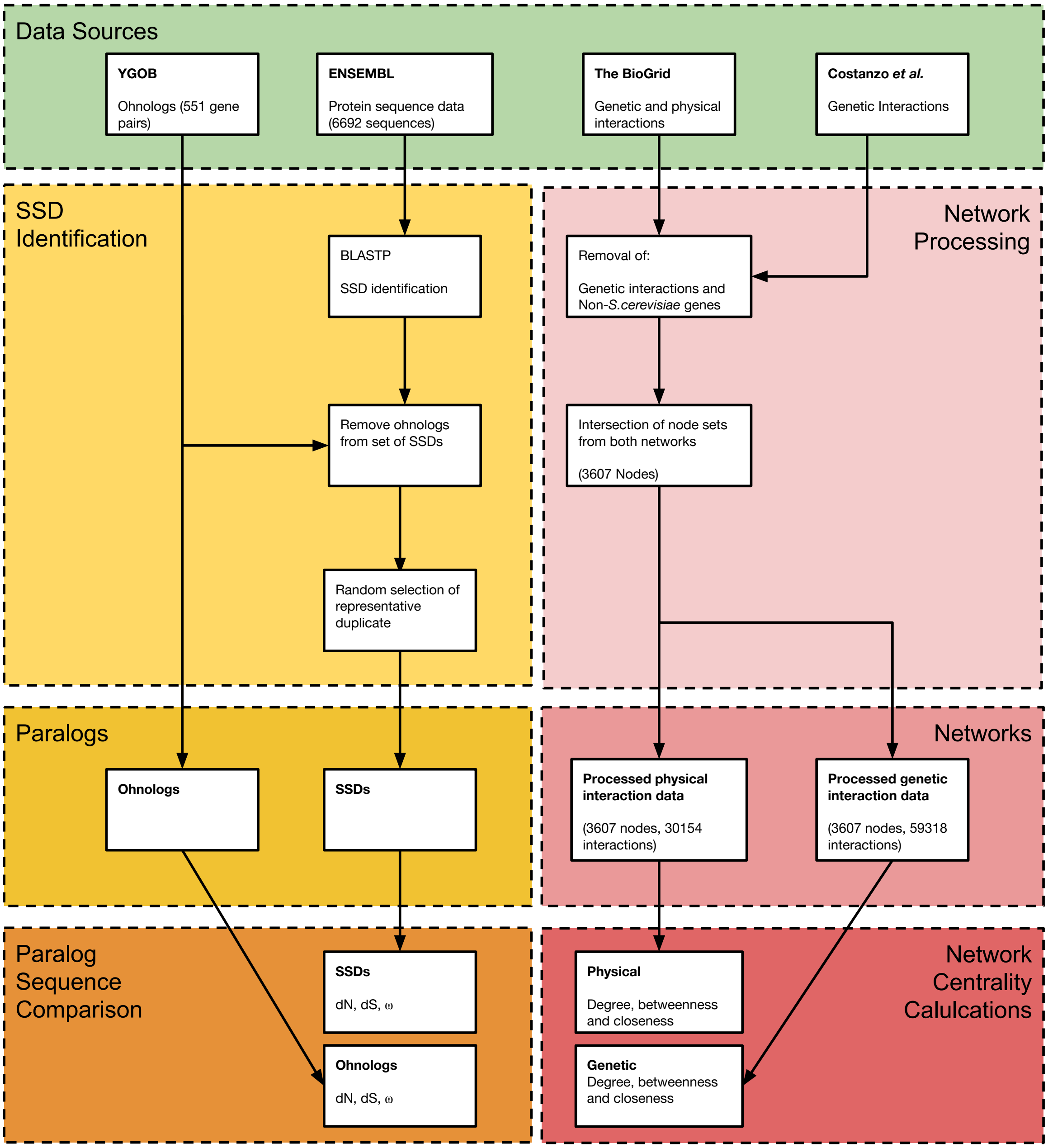

Data Sources
YGOB
Ohnologs (551 gene pairs)
ENSEMBL
Protein sequence data (6692 sequences)
The BioGrid
Genetic and physical interactions
Costanzo et al.
Genetic Interactions
SSD Identification
BLASTP
SSD identification
Remove ohnologs from set of SSDs
Random selection of representative duplicate
Network Processing
Removal of:
Genetic interactions and Non-S.cerevisiae genes
Intersection of node sets from both networks
(3607 Nodes)
Paralogs
Ohnologs
SSDs
Networks
Processed physical interaction data
(3607 nodes, 30154 interactions)
Processed genetic interaction data
(3607 nodes, 59318 interactions)
Paralog Sequence Comparison
SSDs
dN, dS, ω
Ohnologs
dN, dS, ω
Network Centrality Calulcations
Physical
Degree, betweenness and closeness
Genetic
Degree, betweenness and closeness

Figure 3.1: Work-flow to generate experiment data for subsequent analysis.

tion, 120 alleles), had their interactions combined and “TSQ” and “DAmP” removed from their name. These 334 alleles represented 300 unique genes. This produced a network composed of 4273 unique genes with and a total of 73825 interactions.

3.2.2 Constructing and Processing Networks

To construct comparable networks composed of the same genes/nodes we utilized the “systematic name” found in the data of Costanzo *et al.* and The BioGrid. This resulted in a total of 3607 genes common to both networks. For each network we retained only those interactions between these 3607 genes. For example, consider that the gene YPR203W interacts with YDR039C and YGL200C. If YDR039C does not exist in the set of 3607 common genes, then the interaction between YPR203W and YDR039 is removed from the network. However, since YGL200C does exist in the set of common genes, the interaction between YPR203W and YGL200C is retained.

This yielded 54833 interactions within the physical interaction network and 66193 interactions in the genetic interaction network.

3.3 Identifying Paralogs

We wished to identify protein-encoding nucleotide sequences as either ohnologs or SSDs. As a quick reminder, those sequences that are identified as ohnologs are pairs of paralogous sequences that can attribute their common ancestry to the *S. cerevisiae* whole genome duplication event. SSDs, are pairs of paralogous sequences that share a common ancestry independent of the *S. cerevisiae* whole genome duplication.

Identification of ohnologs was simple as the set of *S. cerevisiae* whole genome duplicates has been pre-identified and this set is curated by the Yeast Genome Order Browser (YGOB) [5]. YGOB utilizes a multiple alignment method between different species’ chromosomal regions to identify homologous regions between the genomes of seven yeast species, which cover both pre- and post-whole genome duplication ancestry: *S. cerevisiae*, *S. castelli*, *C. glabrata*, *K. lactis*, *A. gossypii*, *K. waltii* and *S. kluyveri*. The algorithm incorporates allowances for inversions and utilizes gaps between contiguous genes on any particular region to maximize the multiple alignment score. This resulted in 551 ohnolog pairs being identified, 28 more than the union of previous studies (523 ohnolog pairs) by Deitric *et al.* [10], Kellis *et al.* [32] and Wong *et al.* [60]. Of important note is that all ohnologs identified in each of the

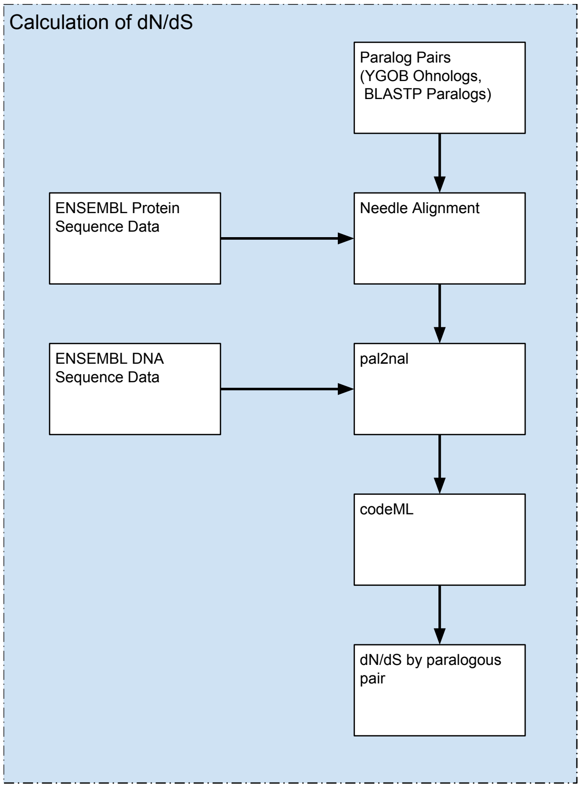

Calculation of dN/dS
Paralog Pairs (YGOB Ohnologs, BLASTP Paralogs)
ENSEMBL Protein Sequence Data
Needle Alignment
ENSEMBL DNA Sequence Data
pal2nal
codeML
dN/dS by paralogous pair

Figure 3.2: Identifying small scale duplicates

previous studies are also found in YGOB. We identified 326 of these ohnolog pairs in our processed networks.

To identify our set of SSDs we obtained protein sequence data from ENSEMBL (2012 database, 5635 protein encoding genes) [16] and performed an all-against-all BLAST [2] query for all pairwise combinations of *S. cerevisiae* protein sequences using the same criteria used by Hakes *et al.* [26]: an e-value $\leq 10^{-8}$, an alignment length ≥ 100, a translation length ≥ 100 residues and a percent identity threshold ≥ 40. Our reasoning for a 40% identity cutoff comes from the results of Hakes *et al.* They found that smaller percent identities introduced increasingly larger numbers of highly divergent pairs, their interpretation being that paralogs were being falsely identified. We removed any of the ohnolog pairs previously identified, which results in an intermediate total of 326 SSDs. Since we are interested in unique pairs of genes, if there was more than one duplicate per gene we selected a duplicate of a gene at random. This paring resulted in a final set of 187 SSD pairs within our processed networks.

3.4 Measuring Centrality

To measure the centralities (degree, closeness, betweenness) for each gene in the two network types we used the network analysis program Gephi 0.8.2 [4]. The centrality measures for each network was exported to a single file for subsequent analysis. Each of these files contained the same number of genes: 3607.

3.5 Measuring Selection

To measure the selective pressure experienced by paralogs following their duplication we used the paralogs previously identified, both ohnologs and SSDs. Python was used to link input and output of each of the following steps. We first aligned pairs of homologous protein sequences that we obtained from ENSEMBL (v. 70) using the Needleman-Wunsch global alignment algorithm as implemented by "needle" in the EMBOSS analysis package [47]. The pairwise protein alignments were used in conjunction with their corresponding nucleotide sequences, also obtained from ENSEMBL (v. 70), and the "pal2nal" program [54]. Utilizing our optimum matchings of paralog sequences, "pal2nal" produced alignments of the corresponding nucleotide sequences with on a codon by codon basis. The pairwise codon alignments were used as input to the codeML program of the PAML package [61]. This process is summarized in Figure 3.2 and produced values for number of non-synonymous mutations per non-synonymous site (dN), synonymous mutations per synonymous (dS) and dN/dS. The value of dN/dS, for each paralogous pair, is a measure and identifier of selection. We use this value in our analyses where dN/dS less than one indicate negative, or purifying, selection; values greater than one indicate positive selection; and, finally, a value equal to one indicates neutral selection. For a detailed explanation of how dN, dS and dN/dS are calculated see Appendix A.1.

3.6 Summary

This chapter has explained the methodology on how we generated our data. The result of each of the steps outlined above is a collection of data for each gene common to the two network types. In summary, for each gene we have calculated network centrality values for both network types, whether the gene has a small scale or whole genome duplicate, and, when a duplicate exists, dN and dS (Figure 3.3).

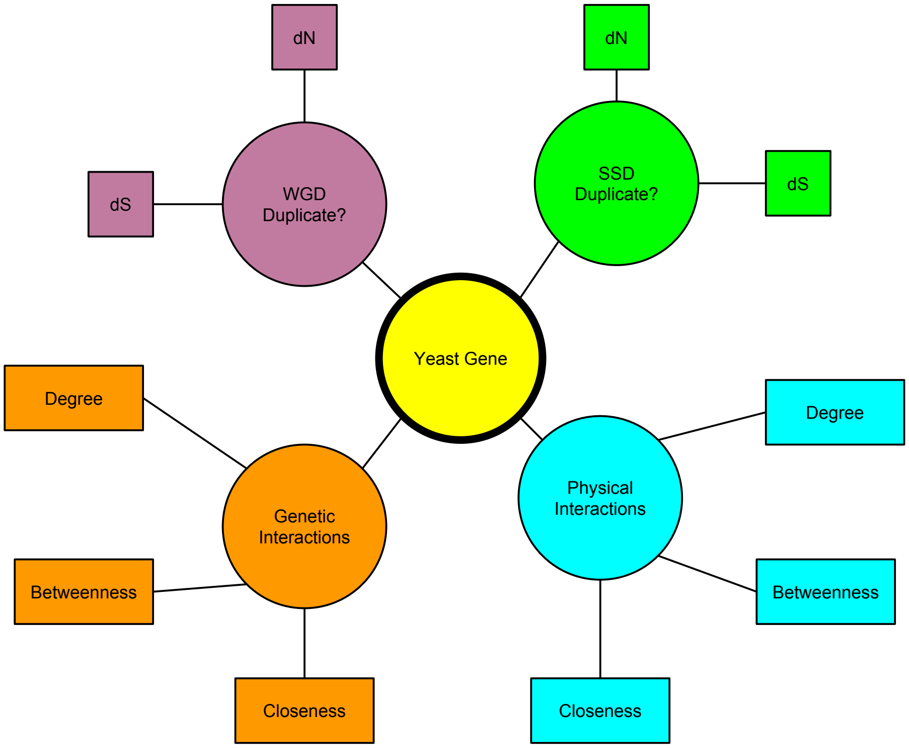


Figure 3.3: Attributes calculated for each gene common to the two network types under investigation.

Chapter 4

Network Characteristics and Analysis of *S. cerevisiae* Paralogs

In this chapter, composed of two sections, we introduce our experimental observations derived from the data that we integrated in our methods.

The first section introduces an overview of the network characteristics for the processed genetic and physical interaction networks. We analyze these networks with the goal of producing observations that will assist us in ascertaining whether the network differences between paralogs are *attributable to differences in the method of duplication, or to fundamental characteristics of the networks.* We first present an overview of the general interaction profile for each of our processed networks, showing the distribution of degree within each network. We then provide similar observations for betweenness and closeness.

Our second section describes our experimental observations that directly relate to answering our research questions. We start this second section similarly to the previous by providing an overview of the interaction profile for each type of duplicate within each network, beginning with degree and then followed by betweenness and closeness. Following this, we show the correlations between these centrality measures and describe the distribution of the relative differences in centrality measures between pairs of paralogs. Next, using our calculated values of dN/dS for each duplicate pair we map dN/dS values to centrality measures to show relationships between them. Finally, we use sliding windows of ordered centrality values to demonstrate possible correlations of these values to ohnolog retention.

4.1 Network Characteristics of *S. cerevisiae* Genetic and Physical Interaction Networks

4.1.1 Distribution of Centrality Measures

We begin our analysis by introducing the general characteristics of each network. Since the construction of our networks, detailed in Chapter 3, removed a number of nodes that were not common to both the physical and genetic interaction network, we first wanted to confirm that this did not dramatically change the distributions of degree, closeness and betweenness.

For the genetic network we calculated the kernel density for degree, closeness and betweenness both for the original, unmodified network (i.e. directly downloaded from the source) and the processed common-node network (Figure 4.1). There was little observable difference between each characteristic for both network data sets. By removing nodes from the network the density of higher degree nodes decreased after processing. This resulted in a small positive shift in closeness and near indistinguishable difference in betweenness—except at higher values. Each of these changes are moderate and the distribution, although shifted, retains the same relative distributions as their related unprocessed characteristics.

These trends are not the same for the physical interaction network (Figure 4.2) where the differences are pronounced. Degree is the least affected by the removal of nodes during our pre-processing. The trends between our experimental data and the original pre-processed data are similar, although a large number of high degree values have been removed by our data processing. This can be seen in Figure 4.2a: nodes with a degree greater than 785 are removed by our processing. The effect of our data processing on betweenness values is not obvious, but it appears that the removal of some of the high degree nodes resulted in a net increase of higher betweenness values.

The most dramatic difference is observed between the closeness values of the pre- and post-processed networks. The pre-processed network has a three-peaked distribution, with the majority of values above closeness values of 0.4. Following processing, the highest closeness values have been removed resulting in a clear bimodal distribution.

To account for the dramatic change in the density distributions for closeness values we must consider two things. The first, unlike the genetic interaction network, is that the nodes in the physical interaction network and their interactions are inherently

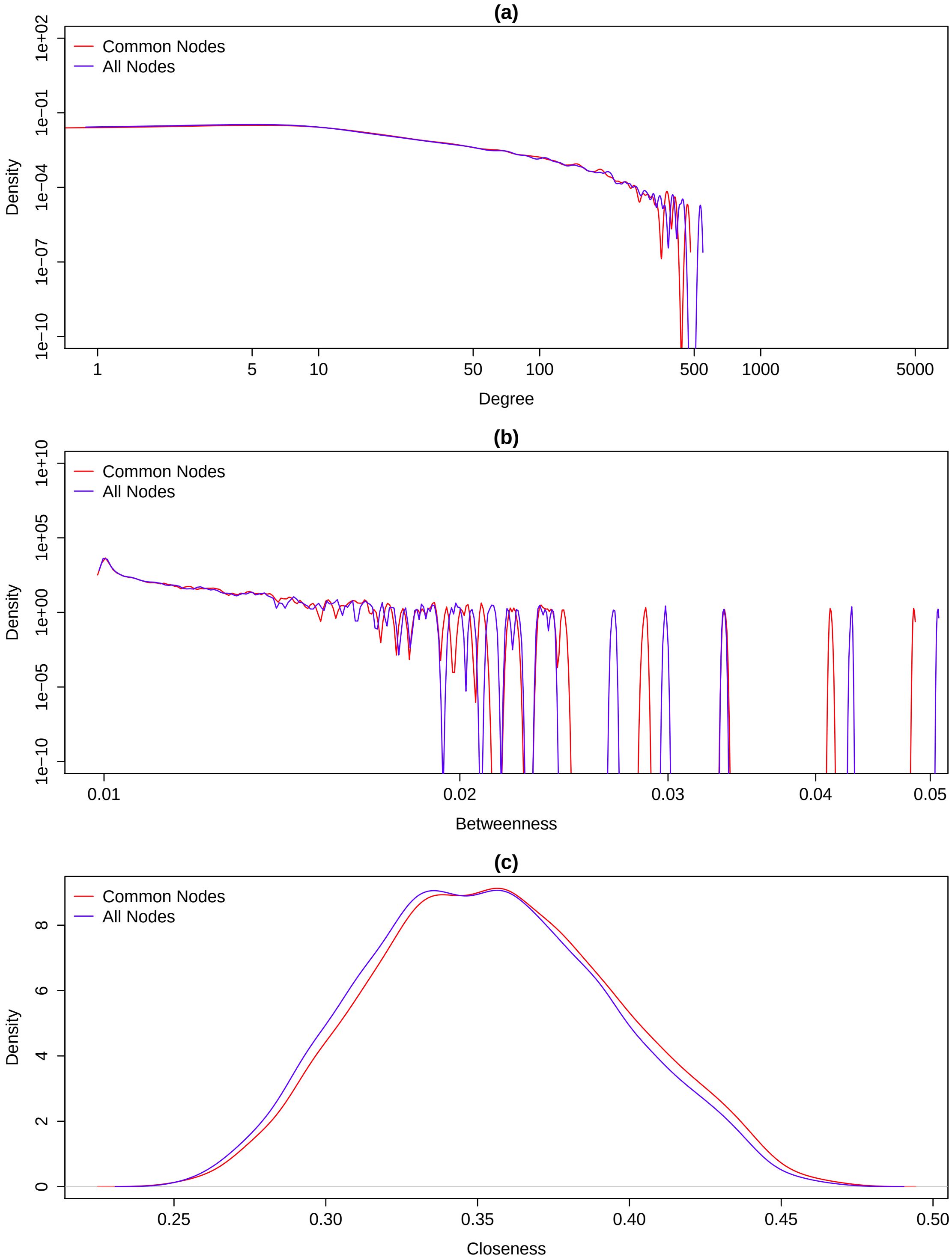


Figure 4.1: Kernel density distributions of genetic degree (a), betweenness (b) and closeness (c) for all network nodes (blue) and nodes in common with the physical network (red).

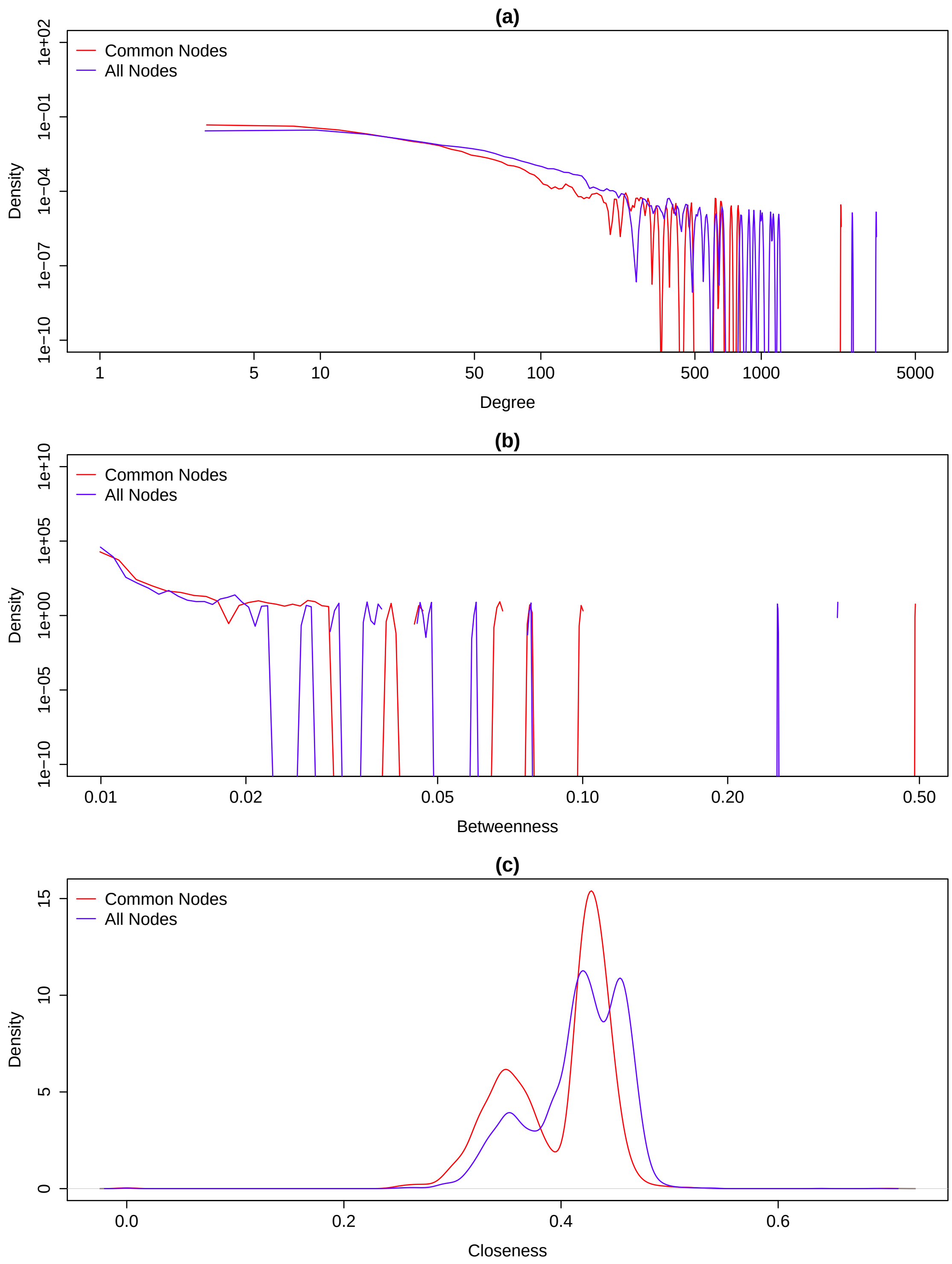


Figure 4.2: Kernel density distributions of physical degree (a), betweenness (b) and closeness (c) for all network nodes (blue) and nodes in common with the genetic network (red).

biased—a single experiment is devised to confirm the physical interaction between any two gene products. If no experiment is used to confirm the interaction between two gene products no knowledge about it can exist. High degree nodes in the physical interaction network are nodes that are of high interest as each interaction indicates a single experiment. These nodes would be of low degree if there were only one or two experiments investigating their interacting parters. Additionally, the lower closeness values are likely due to poor experimental representation of some gene products. There is a large density in both the pre- and post-processed data for low closeness. This observation is likely due to the opposite of what we just described—a subset of gene products that are not central to any experiments. Both of these are the opposite of the genetic interaction network whose experimental design was entirely unbiased: every query gene was given the ability to interact with every subject gene.

The second consideration is that physical interaction data does not consider the promiscuity of gene products when outside their unique cellular context. It is entirely possible that two nodes in the physical interaction network may share an interaction that would never exist inside the cell. For example, an interaction is unlikely when one gene is confined to the nuclear envelope and the other to the cellular membrane. Unfortunately, the sheer number of nodes and interactions within the physical interaction network, as well as incomplete knowledge of gene product location, makes it very difficult to compensate for this possibility.

With these two considerations in mind, we hypothesize that the shift in the closeness distribution is not an invalidation of our processed physical data. The common node set between the genetic and physical interaction network removes some of the bias inherent in the physical data by using an external determinant on what nodes are included and excluded. The query genes within the genetic interaction network are selected at random, which, when we determined the nodes common to both networks, imparted this randomness to the physical interaction data. Furthermore, by removing high degree nodes from the network we likely reduced the effect of overly promiscuous physical interactions. Ultimately, it is important that we note how our pre-processing has affected the network, and its possible reasons. However, it is beyond the scope of this thesis to explain the exact biological reasoning for the discrepancy seen between the pre- and post-processed physical interaction.

4.2 Network Characteristics of *S. cerevisiae* Ohnologs and SSDs

In this section we discuss the analysis of our experimental results, specifically to answer our research questions. The sections below are organized by research questions they address.

4.2.1 Distributions of Ohnolog and SSD Centrality Measures

We approach our first research question,

> *Does the mechanism of duplication correlate to the distribution of network centrality measures for duplicated genes?*

by determining the distribution of each centrality measure for both types of duplicates, ohnologs and SSDs, and for each network interaction type (Figure 4.3).

For both degree and betweenness, the majority of paralogs for each network interaction type follow the same trend. The large majority have low degree or low betweenness with a few sporadic higher value centralities composing no discernable pattern. Closeness values follow similar trends within each network type, but not between networks reflecting the closeness differences of the networks seen in Figures 4.1 and 4.2 and described in Section 4.1.1. These plots show that, generally, there are few differences in the distributions of network centrality measures, considering both physical interaction and genetic interactions, for either paralog type. The only differences appear to occur when closeness values are compared, but these differences may be due to the topology of the physical interaction network as explained in Section 4.1.1.

4.2.2 Differences in the Centrality Measures of Ohnolog and SSD Pairs

Considering there were very few differences in the distribution of centrality measures for each paralog type as whole, we wished to determine whether there were any differences in centrality measure between each pair of ohnologs or pair of SSDs. To accomplish this we computed the percent difference of each centrality measure

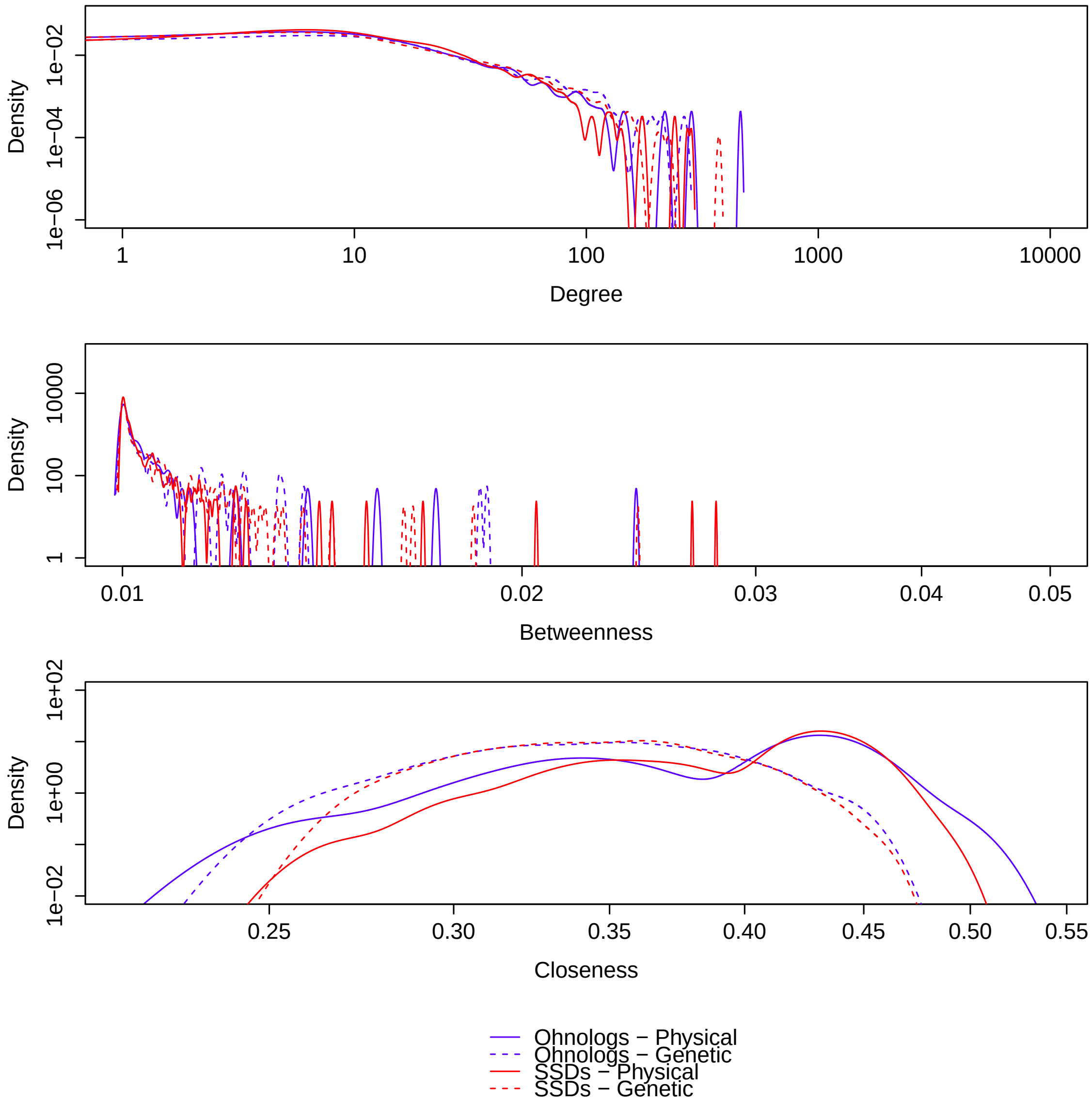


Figure 4.3: Kernel density estimates for the distribution of degree (top), betweenness (middle) and closeness (bottom) for ohnologs (blue) and SSDs (red) using both physical interaction data (solid line) and genetic interaction data (dashed line). Individual peaks in the density plot are individual genes isolated from the majority of centrality values. This creates discontinuous plots.

between pairs of ohnologs and pairs of SSDs, covering both genetic and physical interactions. We grouped these results by network centrality measure: Degree, Figure 4.4; Betweenness, Figure 4.5; and, Closeness, Figure 4.6.

The percent difference between degree values of paralogs is similar within each network interaction type for both ohnologs and SSDs (Figure 4.4), but different between each type of network. We see that the distribution of percent differences in physical degree values is a wide smooth curve for both ohnologs and SSDs. This contrasts with the left skewed distribution exhibited by the percent differences in genetic degree. This left skew indicates that a large proportion of paralogous pairs, both ohnologs and SSDs, have large differences in their genetic interaction degree.

Statistically, we find that the differences in degree data for both ohnologs and SSDs in physical interaction network is not significant (Mann Whitney U = 35001, p=0.642, $\alpha = 0.05$). Similarly, the two distributions of genetic interaction degree are not significantly different (Mann Whitney U = 41385.5, p = 0.075, $\alpha = 0.05$).

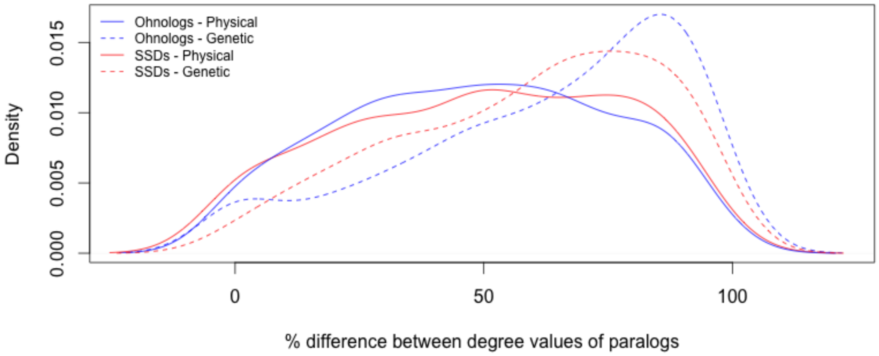


Figure 4.4: Kernel density estimate of the percent difference in degree for all pairs of ohnologs and SSDs for both the genetic interaction network and physical interaction network.

When comparing the percent difference in betweenness values of paralogs in each network (Figure 4.5) we see a left skew for each combination of pairs considered. Each set of data shows a similar distribution: there is an increase in the relative number of pairs as the percent difference between them increase—a large proportion of paralogs have high percent differences in their betweenness. This trend is amplified for the genetic betweenness differences of ohnologs, as an even larger proportion of these paralogs have large differences in their betweenness.

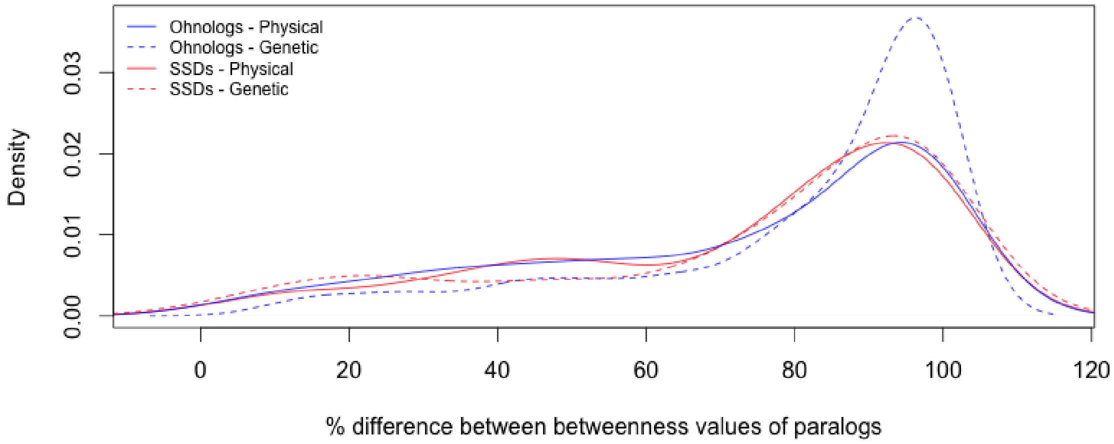


Figure 4.5: (Kernel density estimate of the percent difference in betweenness for all pairs of ohnologs and SSDs for both the genetic interaction network and physical interaction network.

Similar to degree values, there is no statistical difference between ohnolog and SSD differences in the physical interaction data (Mann Whitney U = 25880, p=0.94, $\alpha = 0.05$). However, unlike our findings for ohnolog genetic degree difference, there is statistically significant difference between the distributions of the differences in genetic betweenness of ohnologs and differences between SSDs (Mann Whitney U = 35001, p = 0.0056, $\alpha = 0.05$). These findings show that a large proportion of paralogous pairs have large percent differences in both their physical and genetic betweenness, with high genetic betweenness differences being an even larger proportion of ohnologs.

Finally, we compare the physical and genetic closeness values for both types of paralogs. Here we find that the distributions are different for the genetic and physical interaction networks, but, as seen in each other example, similar within each network. The difference in physical closeness have two distinct peaks: a large peak and small peak. The large peaks having very little percent difference in physical closeness between pairs (peak approximately 5% difference). The smaller peaks for each paralog type have slightly higher percent differences in closeness (peak approximately 2% difference). These two distributions are visually very similar, which reflect the fact that any differences are not statistically significant (Mann Whitney U = 35259, p = 0.16, $\alpha = 0.05$)

Our three observations show that, within each network, the mechanism of duplica-

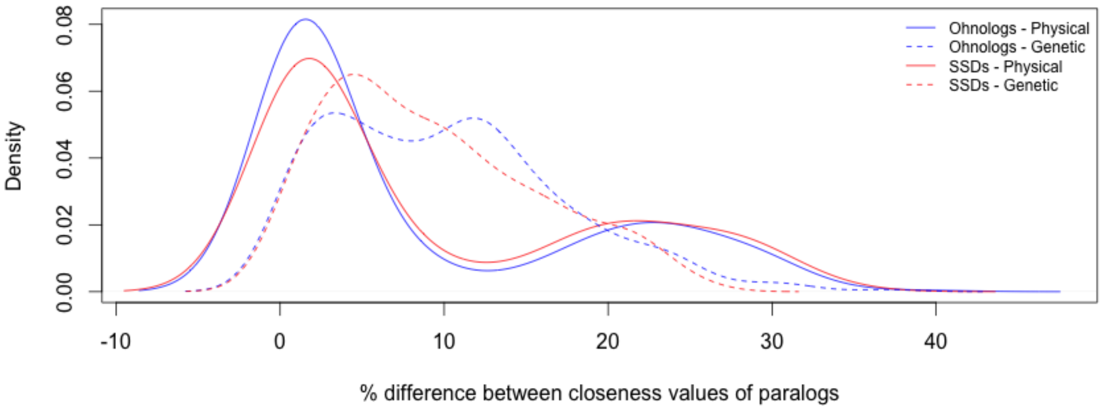


Figure 4.6: Kernel density estimate of the percent difference in closeness for all pairs of ohnologs and SSDs for both the genetic interaction network and physical interaction network.

tion is not an indicator for the percent differences between the centralities of pairs of paralogous genes. However, the differences between centrality measures for paralogs can be dramatic between networks: a larger proportion of paralogs have greater differences in genetic degree than physical degree; a larger proportion of paralogs have small differences in physical closeness than genetic closeness.

These findings are interesting since the genes in each network are identical but we observe the characteristics of two different features, physical and genetic interactions. Returning to an answer for our research question, these observations indicate that the mechanism of duplication has little bearing on the centrality differences of duplicates pairs. However, centralities are affected differently depending on the type of interaction, whether physical or genetic. To what amount is dependent on the centrality measure and the type of interaction.

4.2.3 Correlations Between Differences in Centrality Measures and Selective Pressure for Ohnolog and SSD Pairs

In this section we provide our observations to answer two of our research questions:

> *Is there a correlation in the difference of network centrality measures between paralogous pairs and the selective pressure experienced after duplication?* and,

Is this correlation different for small scale duplicates compared to ohnologs?.

We combined the results of our dN/dS calculations (Measuring Selection in Chapter 4) with the absolute percent change in each measured network centrality score (Section 4.2.2, above) between each paralogous pair. Therefore, each pair has three pairs of values: the dN/dS value matched with each of the percent difference in degree (Figure 4.7), betweenness (Figure 4.8), and closeness (Figure 4.9).

In all of our dN/dS observations we found evidence of strong purifying selection (dN/dS < 0.6) for both ohnologs and SSDs. We did not observe any pairs with dN/dS values at or above unity, which shows that there were no duplicates that were subject to neutral or positive selection. We did not investigate individual regions between aligned sequences, only the gene alignment in their entirely.

In the first set of our observations, pairing dN/dS with absolute percent change in degree between pairs (Figure 4.7), we see that, as the percent difference increases, there is little change in the values of dN/dS. This trend is similar for the same analysis using betweenness (Figure 4.8) and closeness (Figure 4.9). Using Spearman's ρ confirms that no linear correlation can be assumed in any of these sets of data: ohnolog genetic degree $\rho = -0.010895$; SSD genetic degree $\rho = 0.001780887$; ohnolog physical degree $\rho = -0.05474795$; SSD physical degree $\rho = 0.003078439$. This finding holds for dN/dS and differences in betweenness: ohnolog genetic betweenness $\rho = 0.05626678$; SSD genetic betweenness $\rho = -0.07640954$; ohnolog physical betweenness $\rho = -0.1416087$; SSD physical betweenness $\rho = -0.1269569$. And, for correlations between dN/dS and differences in closeness (ohnolog genetic closeness $\rho = -0.03029416$; SSD genetic closeness $\rho = -0.08331598$; ohnolog physical closeness $\rho = -0.0326225$; SSD physical closeness $\rho = 0.01535$).

These findings show that there is no statistically significant correlation between the difference in network centrality measures of paralogous pair and the selective pressure experienced by them since duplication: regardless of differences in centrality, surviving duplicates have been subject to strong purifying selection.

4.2.4 Correlations Between Centrality Measures and Retention Following Whole Genome Duplication

Finally, we turn to our last research question,

Does the retention of ohnologs correlate with network centrality measures?

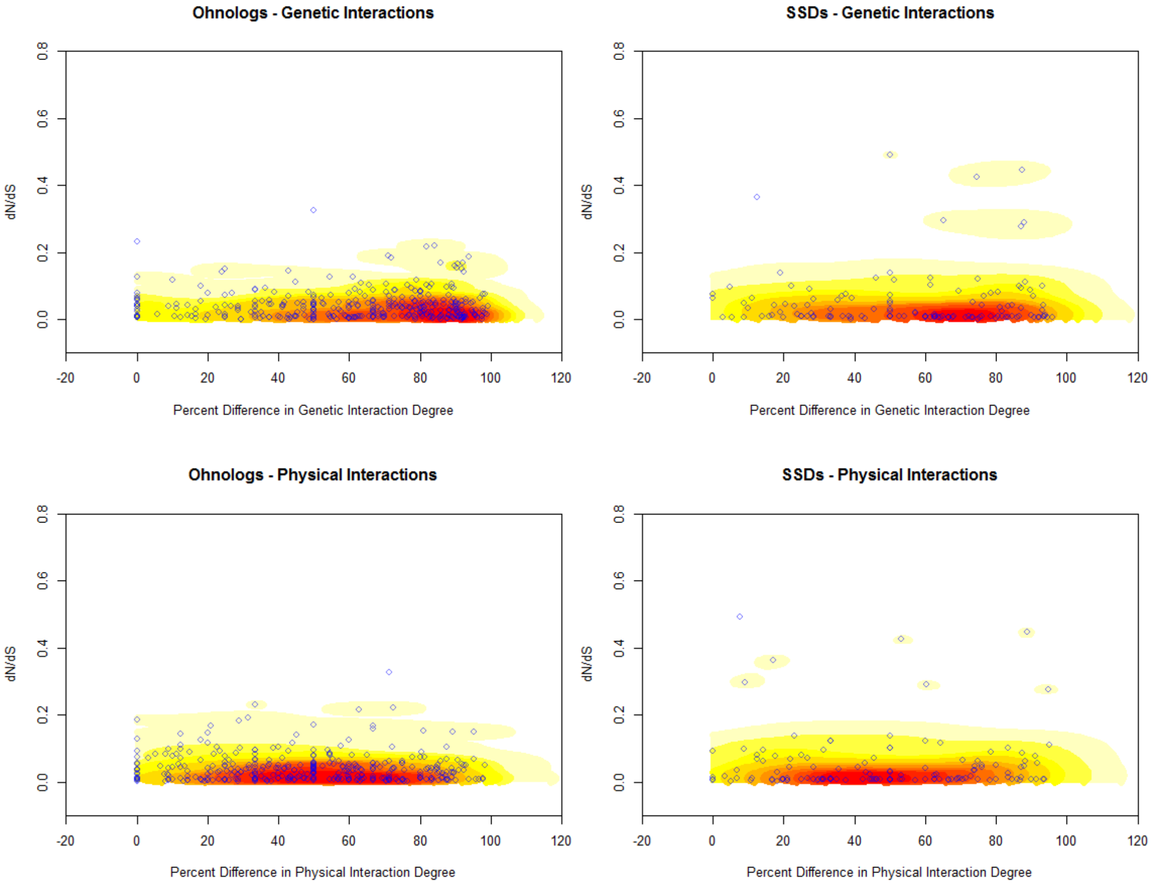


Figure 4.7: Density heat maps of dN/dS and degree for ohnologs and SSDs in both the genetic and physical interaction networks.

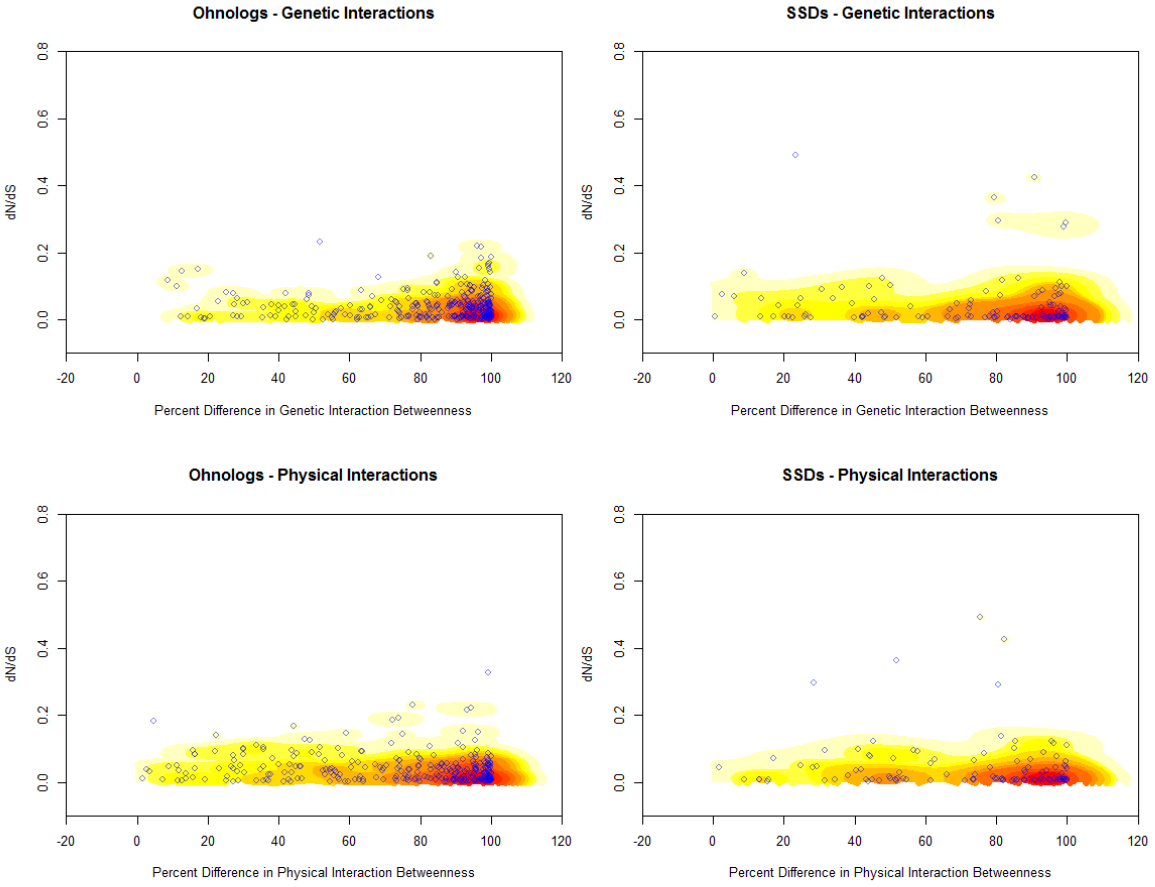


Figure 4.8: Density heat maps of dN/dS and betweenness for ohnologs and SSDs in both the genetic and physical interaction networks.

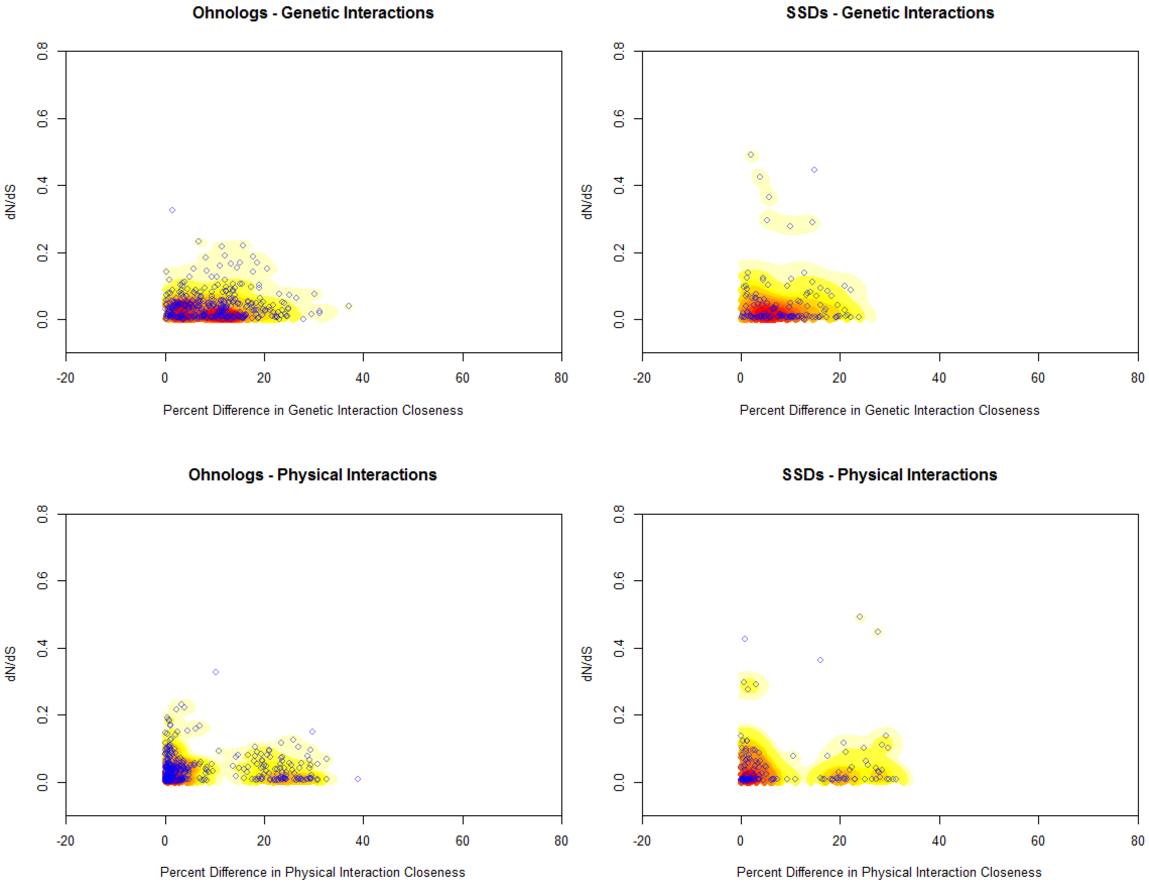


Figure 4.9: Density heat maps of dN/dS and closeness for ohnologs and SSDs in both the genetic and physical interaction networks.

We approach this question on the assumption that following WGD every gene existed in duplicate. Using this as a basis it is clear that extant ohnologs are those genes that retained their whole genome duplicate partner. All other genes have lost their duplicate. By observing those pairs of genes that are ohnologs, we can attempt to determine what, if any, centrality measure correlate to their retention.

Our analysis began using each of the three centrality measures. We sorted all the genes in our analysis using each of our three centrality measures, resulting in three sets of data. Then, using a sliding window, we found the fraction of ohnologs per 1000 gene window. The results of these sliding windows for each centrality measure are seen in Figure 4.10.

The reason for the sliding window of 1000 genes was to minimize variation between consecutive windows. For example, a window size of one results in an ohnolog fraction of zero for each gene that isn't an ohnolog, and a value of one for each gene that is an ohnolog. This would create noisy data—a change between zero and one continuously, which provides no information on retention trends. By using a much larger window size, such as 1000, large variations between windows are reduced and global trends can be visualized. Therefore, since we assume that immediately following whole genome duplication all genes exist in duplicate, the sliding window plot is a visualization of a trend where increases or decreases in the ohnolog fraction indicate relative retention or loss dependent on increasing centrality measure.

We calculated Spearman's ρ values for each centrality measure and network type (Table 4.1). In each of our findings centrality is strongly positively correlated with the percentage of ohnologs in each window for the physical interaction network. Contrastingly, we see a strong negative correlation in the genetic interaction network.

We thought it surprising that our findings were opposite between the physical and genetic networks. Intuitively, we assumed that the genetic network would follow similar trends as the physical network: higher centrality nodes indicate higher essentiality [25] and would therefore indicate that there are more ohnologs at higher centrality values. We proceeded to answer the question: why do the observations between physical and genetic interaction networks appear to be contradictory?

To pursue an explanation to our question, we began by considering the state of the genome immediately following whole genome duplication. In the generation following the duplication event the stoichiometric balance of genes has not been disturbed—the relative amounts of any one gene to any other remains the same. Our question now becomes: how do the physical and genetic interactions differ following the duplication

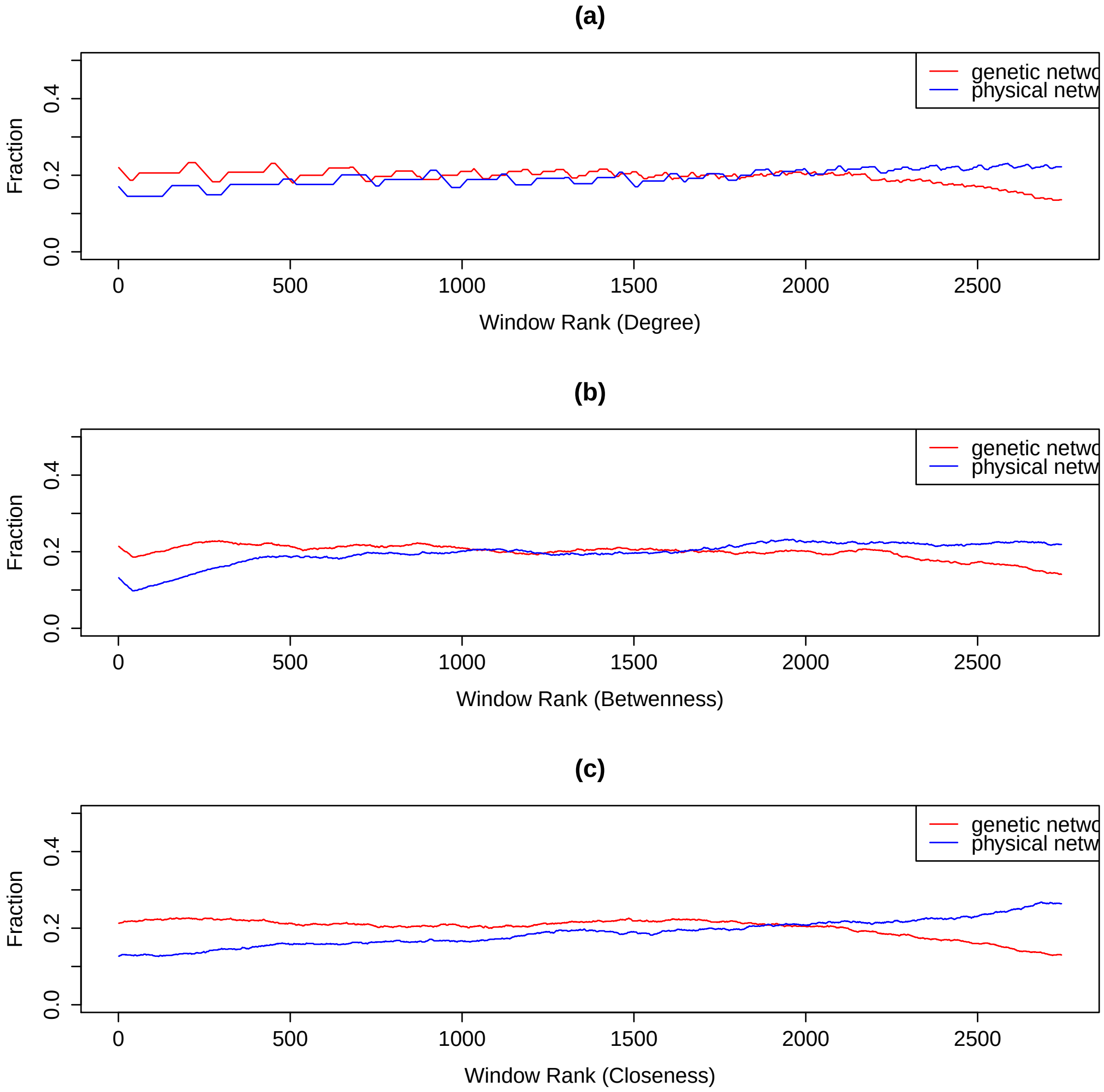


Figure 4.10: Fraction of ohnologs per sliding window of 1000 genes sorted by increasing centrality value for genetic (red) and physical (blue) interaction networks: (a) degree; (b) betweenness; and, (c) closeness.

event?

Let us first consider the physical interaction network, which now has a second, identical, copy of all nodes and interactions. Conceptually, we can consider these two identical networks operating in parallel, with the caveat that interactions connect these two parallel networks such that each gene can now interact with either of both duplicates of its ancestral interacting partners. We can see this in our trivial example of Figure 4.11.

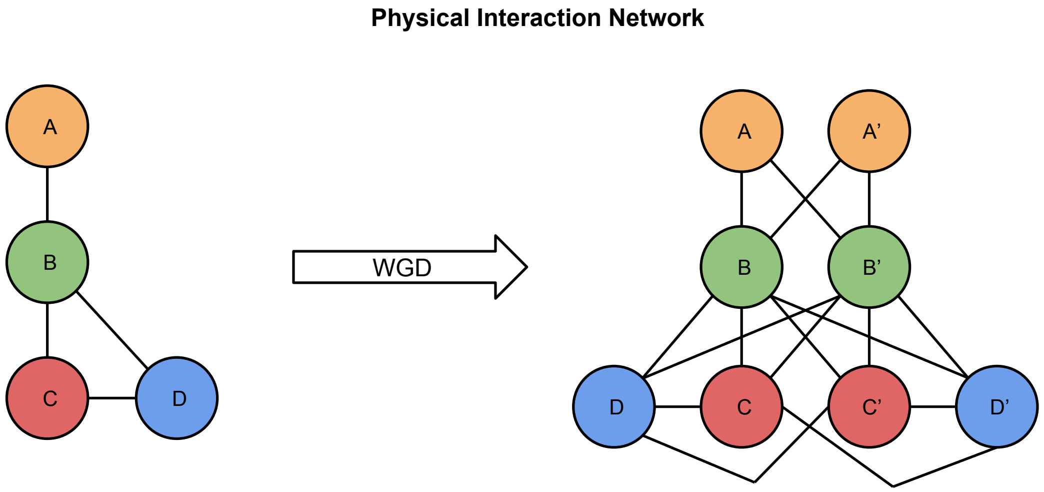


Figure 4.11: A physical interaction network with four nodes pre-duplication and eight nodes post-duplication.

This situation is very different for the genetic interaction network (Figure 4.12). Now that every node has been duplicated, a genetic interaction can only be found if both of these duplicates are removed. To understand why, consider what would happen if, following a WGD, we selected two genes from the ancestral network, selected one duplicate for each in the duplicated network, and deleted each of our selected pairs. Since their identical duplicate copies still exist, and have not diverged, little if no phenotypic response would be observed (assuming their basal dosage requirements were still met); *therefore, genetic interactions initially existed either between duplicate genes (if they were not dosage sensitive) or between any two genes (if either or both were dosage sensitive)* [57]. Only if duplicates are dosage sensitive will genetic interactions outside the pair be observable.

With this in mind, we hypothesized that ohnolog pairs that have maintained their genetic interaction to one another, yet interact with others, are a combination of the two extremes following WGD: they are dosage sensitive to some interactions yet re-

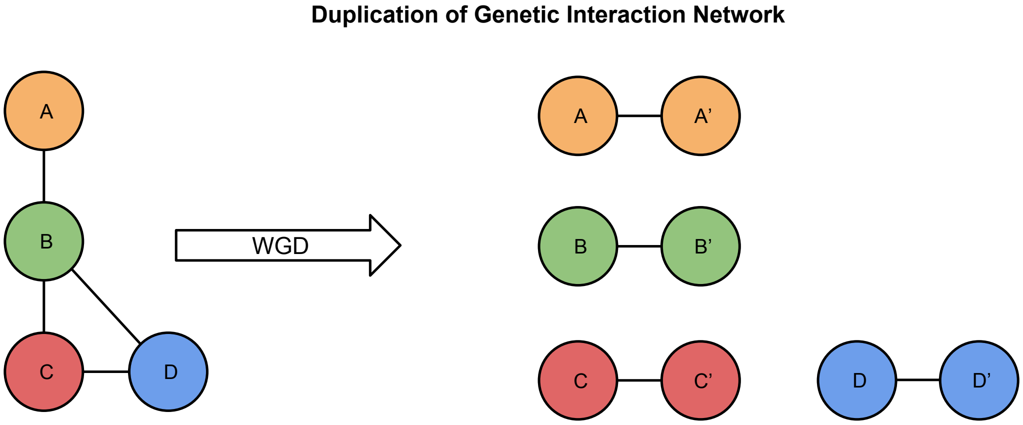


Figure 4.12: A genetic interaction network with four nodes pre-duplication and eight nodes post-duplication. Following duplication each duplicate pair is disconnected from every other duplicate pair, forming a disconnected network. The only genetic interactions are those between duplicates.

Network	Centrality Measure	ρ	p
Genetic	Degree	-0.8270584	$< 10^{-9}$
	Betweenness	-0.9331068	
	Closeness	-0.853526	
Physical	Degree	0.9783577	$< 10^{-9}$
	Betweenness	0.936841	
	Closeness	0.9724897	

Table 4.1: Spearman's ρ values and associated p-values for the fraction of ohnologs per window rank for genetic and physical interactions.

tain an ability to buffer other interactions, explained by their retained intra-duplicate interaction. The fact that they have retained this intra-duplicate interaction, yet diverged sufficiently to have genetic interactions with other genes, led us to hypothesize that there is evolutionary constraint to this intra-duplicate interaction. We predicted that this means that the subclass of ohnologs that interact with their duplicate are subject to low rates of non-synonymous substitutions per non-synonymous site. Conversely, ohnologs that do not interact with their duplicate should observe relative higher rates.

To find test our hypothesis, we first classified our two types of ohnologs: those that interact with their duplicate, and those that do not interact with their duplicate. Of the 714 ohnologs present in both networks, we found that, within the genetic interaction network, 69 ohnolog pairs interacted (out of 357 ohnolog pairs) and 288 ohnolog pairs did not interact. Within the physical interaction network we found that 110 ohnolog pairs interacted and 247 ohnolog pairs did not. Non-ohnologs comprised the remaining 3029 genes within the network. We then calculated dN values for the pairs of ohnologs within each of the two classes found in our two networks. These data were plotted as density estimate (Figure 4.13) to show their distribution. We found ohnologs that interact with their duplicate observe low dN and ohnologs that do not interact with their duplicate observe higher dN, for both genetic and physical interactions. There is a strong right skew of distribution for the genetically interacting ohnologs to where the highest density of non-genetically interacting ohnologs is at a dN of 0.04. This is much lower than the dN at the highest density of non-genetically interacting ohnologs (dN = 0.42). The median dN value for non-genetically interacting ohnologs was 0.164 compared to a median dN value of 0.432.

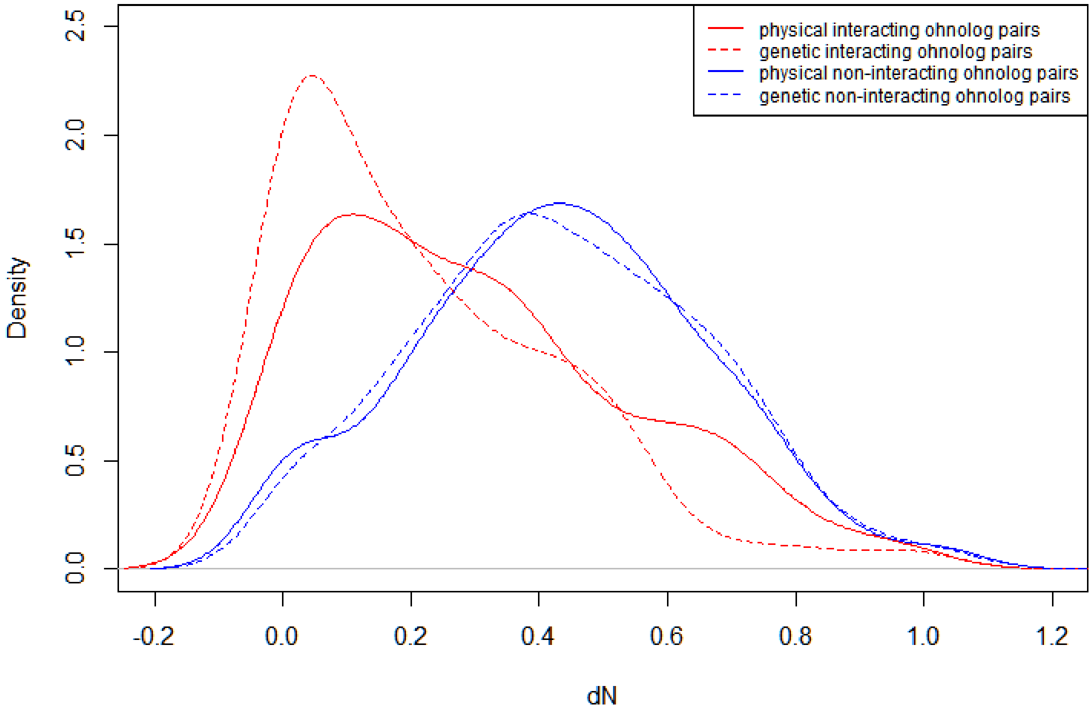


Figure 4.13: Density plot of non-synonymous mutations per non-synonymous site for ohnologs that interact with their WGD duplicate and ohnologs that do not interact with their WGD duplicate

A similar trend of low dN interacting ohnologs and higher dN non-interacting ohnologs is seen for the physical network data. In fact, the non-physically interacting ohnologs have a median dN of 0.436 very similar to the dN of the non-genetically interacting ohnologs (median dN=0.432). However, there is a marked difference in the media dN value of 0.277 for physically interacting ohnologs compared to genetically interacting ohnologs (dN = 0.164).

These observations led us to modify our theory above (that ohnolog pairs are dosage sensitive to some interactions yet retain an ability to buffer other interactions). Instead, it appears that these pairs have changed little when compared to one another and the dosage sensitivity and retention of the intra-duplicate interaction reflects the repertoire of interactions for this class of ohnologs is essential.

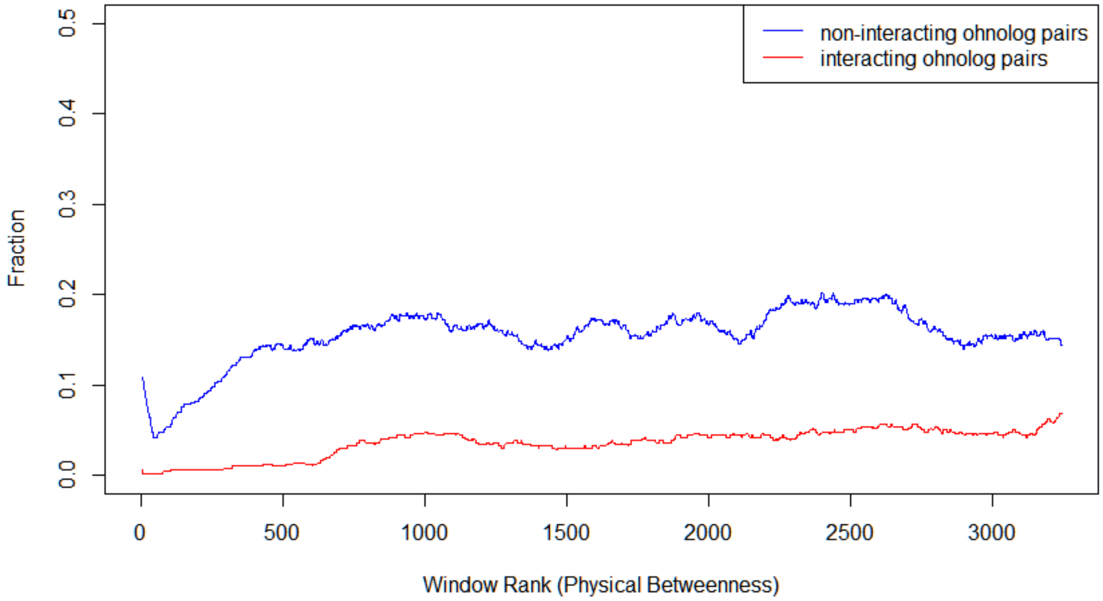


Figure 4.14: Fraction of ohnologs that do not interact with their duplicate and ohnologs that do interact with their duplicate over a sliding window of 500 genes. All genes in the network, including non-ohnologs, were sorted by physical betweenness centrality value in ascending order. The trend is similar for degree and closeness.

Next we applied our sliding window methodology to our two classes of ohnologs: ohnologs that interact (genetic network, n=69) with their duplicate; and, ohnologs that do not interact (genetic network, n=288) with their duplicate (betweenness, Figure 4.15, similar trend for degree and closeness). When considering genetic inter-

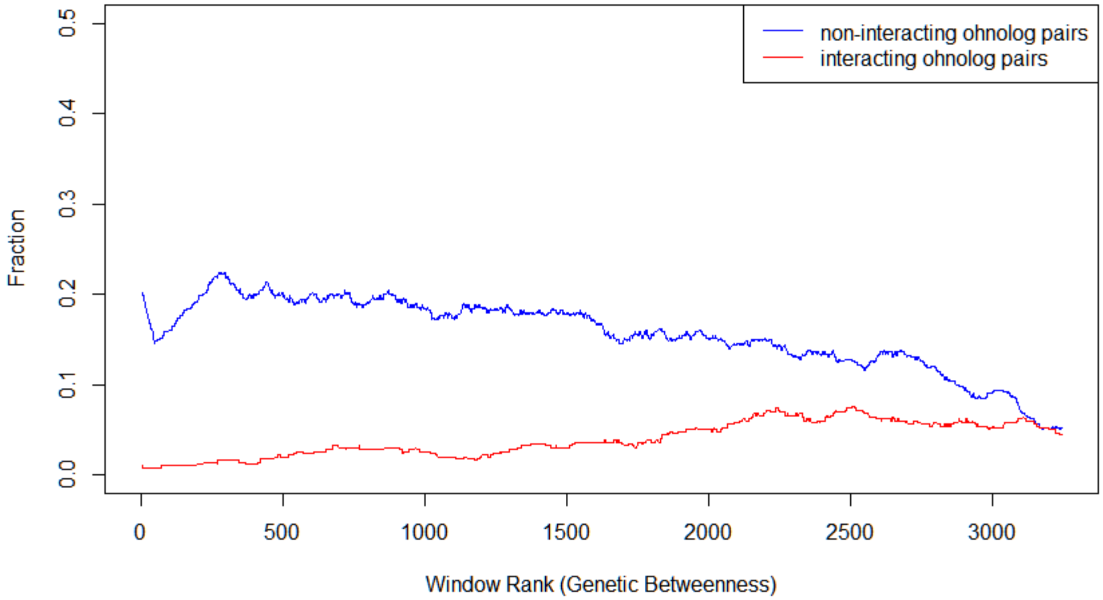


Figure 4.15: Fraction of ohnologs that do not interact with their duplicate and ohnologs that do interact with their duplicate over a sliding window of 500 genes. All genes in the network, including non-ohnologs, were sorted by genetic betweenness centrality value in ascending order. The trend is similar for degree and closeness.

actions, we find that those ohnologs that retain an intra-duplicate interaction also exhibit low dN as well as an increasing duplicate retention fraction as centrality values increases. Since essentiality and centrality are positively correlated [25], we can extrapolate that this class of genes is essential.

Conversely, those ohnologs that have lost the interaction to their duplicate have higher duplicate retention fraction at low centrality. This is perplexing as centrality and essentiality imply that these genes are non-essential, yet they are retained in duplicate. We will investigate this further below.

We performed the same analysis on the pairs of physically interacting ohnolog sisters (n=110) and non-interacting sisters (n=247) (Figure 4.14). Here we find that the fractions of non-interacting ohnolog sisters and interacting ohnolog sisters increase as centrality increases. The trend for the group of ohnolog sisters that do not interact is clearly the opposite.

Although it is unclear why this is the case we hypothesize that the increase in dN and higher retention at low centrality may indicate that the original set of genes that produced this class was dosage sensitive—but for a single copy. The relatively higher rate of non-synonymous mutations implies that there was a higher incidence of pseudogenization and gene loss for this class. We consider this as pointing towards the fact that those genes persisting in duplicate (within this class) do so because they have subfunctionalized sufficiently to offset dosage effects.

Conversely, as hypothesized by Vandersluis *et al.*, the retention of an interaction between ohnolog sisters does not need to be explained by selection on their dosage. In other words it does not need to be explained by redundancy or subfunctionalization. Instead, the retention can be explained by neofunctionalization of the less constrained ohnolog requiring, or allowing, the ancestral role to be maintained. However, this does not explain why a high density of low dN values was observed within this class of genes; this contradicts the possibility of neofunctionalization in the classical functional sense. Instead, the neofunctionalization may occur in transcription factor binding sites, which are not measured by dN. The physical proteins may be very similar, but due to regulatory difference, may be expressed under different conditions.

In summary, the evidence indicates that network centrality does correlate with the retention of ohnologs. However, whether the correlation is positive or negative depends on whether the ohnolog sisters interact in addition to the type of network being considered. In physical interaction networks we see positive correlation between centrality and retention. Similarly, in genetic interaction networks we see a positive correlation between centrality and retention for sister ohnologs that retain their genetic interaction. However, ohnolog sisters that have lost their genetic interaction between one another have a negative correlation between centrality and retention. In addition to this, ohnolog sisters that do not interact either physically or genetically, have a higher proportion of non-synonymous mutations compared to those sisters that do interact.

Therefore, we can conclude that ohnolog sisters that interact with one another, either genetically or physically, are unable to acquire large numbers of non-synonymous mutations. Our interpretation is that these interactions are functionally constrained since the underlying sequences are constrained. We can conclude from this that the retention of ohnologs within this class is due to dosage sensitivity of the pair. If either of the pair is lost there is an imbalance in the dosage requirements and cell fitness is affected. In the physical interaction network, we propose that each pair of ohnologs are a component of a protein complex. These complexes were likely homodimers prior to duplication. After duplication these homodimers became heterodimers composed of two sister ohnologs. Since both duplicates show little non-synonymous change, both have maintained their ancestral dimerization function and thus buffer one another.

We find further support for this hypothesis by recognizing the fact that knocking out both of a pair within this class of ohnologs produces an aggravated genetic interaction, compared to a single knock out of either. This is interpreted as each gene within the pair buffering the other to some degree, producing a homodime, when any single gene of a pair is knocked out. Although, if these genes are dosage sensitive this buffering may be insufficient to maintain organism health in the long term (reduction in the total number of complexes by 50%). However, knocking out both produces an even less fit organism (reduction in the total number of complexes by 100%).

On the other hand, ohnolog sisters that do not interact genetically with one another have acquired large numbers of non-synonymous mutations, and are preferentially retained at low centrality. Since centrality and essentiality are positively correlated, centrality can not explain their essentiality. This implies that for both to be retained each must acquire unique and novel function such that the fitness effect of deleting one is more or less detrimental than deleting both. In other words, for there to be no genetic interaction between ohnolog sisters the effect of both genes being knocked out must be statistically insignificant compared to the multiplicative effect of each being knocked out on its own. Therefore, ohnologs that do not interact with their sister are retained in duplicate due to divergence in function: each has become indispensable on its own in spite of their low centrality. We propose that this divergence is best described as a partitioning of ancestral function (i.e. subfunctionalization) to the point that each sister is incapable of buffering the other.

These observations expand on those previous that show WGDs tend towards subfunctionalization [12, 26]. However, subfunctionalization does not tell the full story. Our results indicate that the tendency towards subfunctionalization can be teased apart based on genetic interaction centrality and interaction profile. The overall trend towards subfunctionalization is unaffected as this partition of ancestral function is observed by the 80% of ohnologs that do not genetically interact with their sister. The remaining 20% of ohnologs that do genetically interact with their sister tend towards their ancestral function.

Chapter 5

Summary and Future Work

5.1 Summary

This thesis has investigated why some whole genome duplicates are retained in duplicate while others are not. We approached this problem from a network analysis perspective with the theory that network characteristics are important for determining retention.

Our first contribution is a set of software tools, called "**Ne**twork **C**entrality and **Pa**ralog **D**ivergence **I**ntegrator" (NeC-PaDI). Nec-PADI integrates network, sequence and paralog data—for any organism, contingent on a standardized input format. We utilized these tools to associate centrality measures, nucleotide and amino acid sequence data, and duplication relationships to each gene in our yeast gene set. These data were then queried to produce our further contributions. It can be downloaded from `http://webhome.csc.uvic.ca/~imrie/necpadi`.

Our main contribution shows that there are two classes of ohnologs and that each have a different profile of retention following duplication. The first of these two classes are those that genetically interact with their duplication sister. The second class are those ohnologs that have do not. We show that ohnolog pairs that do have a genetic interaction with their duplication sister have higher duplicate pair retention at higher centrality values. We also show that they exhibit a lower ratio of non-synonymous mutations per non-synonymous site which indicates their sequences have changed relatively little. The fact that they share a genetic interaction with one another shows that they have retained an ability to buffer one another as if there was no buffering existed then no genetic interaction would be observed.

For ohnologs that have lost a genetic interaction with their duplication sister we find higher retention at lower centrality values and that they exhibit a higher ratio of non-synonymous mutations per non-synonymous site. With this loss of genetic interaction we interpret their retention as being due to functional divergence resulting in indispensability. This, based on previous literature, we interpret as subfunctionalization[26, 12].

Secondary contributions show that evolutionary pressure does not correlate to centrality measures nor does the distribution of centrality measures change significantly based on the type of duplication. We also show that the type of interaction (genetic or physical) has an effect on the distribution of centrality measures. This is due to the physical and genetic interaction networks describing different characteristics of the same genes.

The tools we constructed also allow us to ask similar questions of organisms other than *S. cerevisiae*. The requirements are that network data, ohnolog data and sequence data are available for the species of interest. With comprehensive network data we can investigate other species whose ancestry contained whole genome duplication such as *Xenopus* species, *Arabidopsis thaliana*, *Zae mays* species, and some fish species [51, 55].

To summarize our results, we found that there was no correlation between the type of duplication and any single centrality measure for each gene we studied. We found that the mechanism of duplication does affect each centrality differently. We found that paralogs had a higher percent difference between genetic centralities, on average, than physical centralities. Assuming that each pair of paralogs had identical centralities immediately following duplication, our findings show that there is more variation in the genetic interactions than physical interactions of duplicate pairs. This may imply relaxed evolutionary constraint on genetic interactions over physical interactions.

However, our investigation found no significant correlation between the network centralities of paralogous pairs and selective pressure as currently observable. Due to the age of the whole genome duplicates and likely many of the small scale duplicates, this is not unexpected as dS may likely outnumber dN since the former is only subject to neutral evolution.

Our final question produced the most interesting findings of this thesis: retention of both duplicates is positively correlated with physical centralities but negatively correlated with genetic centralities. Due to the unique feature of genetic networks, that ohnologs exclusively interact with one another following duplication, we partitioned ohnologs into those that interact with their sister gene and those that do not. With this partitioning we found that physically interacting and genetic interacting ohnolog pairs had a very low rate of non-synonymous mutations per non-synonymous site. Ohnologs that did not retain a physical or genetic interaction had a much higher fraction of non-synonymous mutations per non-synonymous site. We then showed that the ohnolog pairs that did not interact with one another are the reason for the apparent trend of decreasing retention as centrality increases. This finding was interpreted in the following way: genes that have lost their genetic interaction are dosage sensitive and as centrality increases, dosage effects are amplified - retaining a duplicate would alter the dosage in a detrimental way.

To conclude, we found that genetic centralities had a greater percent differences for both SSDs and ohnologs compared to physical centralities; that there is no correlation between dN/dS and differences in centrality due to evolutionary age; and, that as physical centrality values increase, an increasing fraction of ohnologs are retained. However, as genetic centrality increases there is a positive correlation with the retention of ohnolog pairs that interact, but a negative correlation for ohnolog pairs that do not.

5.2 Future Work

The observations and discussion of this thesis provide many avenues for further investigation.

5.2.1 Investigating Post-Processed Physical Network Centrality

In our methodology we generated physical and genetic interaction networks using a common set of nodes. By doing this to the original physical interaction network we downloaded from The BioGrid, we produced an artefact where closeness had a two peaked distribution: a population of low centrality and a population of high centrality. Since the kernel density plot is a continuous function, the distribution of centrality measure implies that the physical interaction network was partitioned into two disconnected graphs by our processing. Therefore, the two subcomponents of the physical interaction network are held together by one or more genes that are not in the genetic interaction network. What are these genes and why are they so fundamental in connecting the network?

5.2.2 Correlating Expression Patterns with Ohnolog Types

One of our statements in our results and discussion indicated that neofunctionalization may occur in genetic interacting ohnolog sisters with low dN, but not in the classical sense of differential protein function. Instead, we proposed that neofunctionalization may be observed in regulatory sites resulting in differential expression. The function of the protein has changed, not from a physical perspective, but from a temporal or environmental one. There are plenty of data showing expression levels, both for developmental phases and environmental states. Combining these data with each ohnolog type would provide further insight into how the two ohnolog types differ.

5.2.3 Functional Divergence Between Ohnologs Types

It has been shown that ohnologs functionally diverge less than SSDs [12, 26]. However, do ohnologs that do interact diverge, functionally, more or less than those sisters that do not interact? Our observations show that there is less sequence divergence between ohnolog sisters that interact but what effect does this have on the interaction partners of each duplicate? A more thorough analysis into the similarity of interaction partners, considering each group of ohnologs separately, would provide insight to this.

5.2.4 Centrality and Chromosomal Location

An interesting avenue for inquiry is whether centrality can be correlated to chromosomal position. It has been shown previously that retention of genes is chromosome and chromosome region specific [56]. Particularly, one region of a chromosome may be lost entirely. What determines which of the pair is lost? What is more important, the location of a gene within the genome or its centrality? We have shown how network centrality correlates with ohnolog retention, how does this fit with the retention of genes being chromosome specific?

5.2.5 Evolutionary Analysis of Ohnnolog Types

For much of our analysis we considered ohnologs to be a single set of genes. However, it would be interesting to observe the selective pressure faced by each group of ohnolog types, those that interact with their sister and those that do not. In addition, network characteristics within each group may be very different. We showed that there is a tendency for ohnologs that interact with their sister to be more prevalent at higher centrality than lower. The opposite holds for ohnologs that did not interact with their sister. It would be interesting to repeat many of our analyses and use this information to compare these two types of ohnologs.

5.2.6 Confirm that Ohnologs with Genetic Interactions with their Sister are in Complexes

We made the hypothesis, based on both physical and genetic interaction data, that ohnologs that interacted with their sister are likely part of complexes and buffer their sister's functionality. It would interesting to see if this is actually the case, and how large these complexes. Are they simply dimers or are they large complexes composed of many different genes? Additionally, does the size of the complex affect centrality?

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. Molecular biology of the cell in cell 4e, 2002.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[3] A.A. Andalis, Z. Storchova, Styles C., T. Galitski, D. Pellman, and G.R. Fink. Defects arising from whole-genome duplications in *Saccharomyces cerevisiae*. *Genetics*, 167(3):1109–1121, 2004.

[4] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. 2009.

[5] K.P. Byrne and K.H. Wolfe. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, 15(10):1456–1461, 2005.

[6] G.M. Cannarozzi and A. Schneider. *Codon Evolution: Mechanisms and Models*. Oxford University Press, 2012.

[7] F.J.J. Chain, J. Dushoff, and B.J. Evans. The odds of duplicate gene persistence after polyploidization. *BMC Genomics*, 12(1):599, 2011.

[8] M. Chester, J.P. Gallagher, V.V. Symonds, A.V.C. da Silva, E.V. Mavrodiev, A.R. Leitch, P.S. Soltis, and D.E. Soltis. Extensive chromosomal variation in a recently formed natural allopolyploid species, tragopogon miscellus (asteraceae). *Proceedings of the National Academy of Sciences*, 109(4):1176–1181, 2012.

[9] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E.D. Spear, C.S. Sevier, H. Ding, J.L.Y. Koh, K. Toufighi, S. Mostafavi, et al. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010.

[10] F.S. Dietrich, S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pöhlmann, P. Luedi, S. Choi, et al. The ashbya gossypii genome as a tool for mapping the ancient saccharomyces cerevisiae genome. *Science*, 304(5668):304–307, 2004.

[11] P.P. Edger and J.C. Pires. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Research*, 17(5):699–717, 2009.

[12] M.A. Fares, O.M. Keane, C. Toft, L. Carretero-Paulet, and G.W. Jones. The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. *PLoS Genet*, 9(1):e1003176, 01 2013.

[13] M. Feldman and A.A. Levy. Genome evolution in allopolyploid wheat: A revolutionary reprogramming followed by gradual changes. *Journal of Genetics and Genomics*, 36(9):511–518, 2009.

[14] J. Felsenstein. Phylogenies from molecular sequences: Inference and reliability. *Annual review of genetics*, 22(1):521–565, 1988.

[15] S. Fields and O. Song. A novel genetic system to detect protein protein interactions. 1989.

[16] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, et al. Ensembl 2012. *Nucleic Acids Research*, 40(D1):D84–D90, 2012.

[17] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[18] L.C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.

[19] J.D. Gibbons and S. Chakraborti. *Nonparametric statistical inference*, volume 168. CRC press, 2003.

[20] T.A. Gibson and D.S. Goldberg. Questioning the ubiquity of neofunctionalization. *PLoS Computational Biology*, 5(1):e1000252, 2009.

[21] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular Biology and Evolution*, 11(5):725–736, 1994.

[22] J.L. Gordon, K.P. Byrne, and K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genetics*, 5(5):e1000485, 2009.

[23] J.F. Gout, L. Duret, and D. Kahn. Differential retention of metabolic genes following whole-genome duplication. *Molecular Biology and Evolution*, 26(5):1067–1072, 2009.

[24] D. Graur and W.H. Li. *Fundamentals of Molecular Evolution*, volume 2. Sinauer Associates Sunderland, 2000.

[25] M.W. Hahn and A.D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[26] L. Hakes, J.W. Pinney, S.C. Lovell, S.G. Oliver, and D.L. Robertson. All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biol*, 8(10):R209, 2007.

[27] M. Hasegawa. Phylogeny and molecular evolution in primates. *Idengaku zasshi*, 65(4):243, 1990.

[28] H. Jeong, S.P. Mason, A.L. Barabási, and Z.N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[29] R. Jovelin and P.C. Phillips. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol*, 10(4):R35, 2009.

[30] M.P. Joy, A. Brock, , D.E. Ingber, and S. Huang. High-betweenness proteins in the yeast protein interaction network. *BioMed Research International*, 2005(2):96–103, 2005.

[31] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. 1969.

[32] M. Kellis, B.W. Birren, and E.S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, 2004.

[33] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120, 1980.

[34] D. Koschützki and F. Schreiber. Comparison of centralities for biological networks. In *German Conference on Bioinformatics*, pages 199–206, 2004.

[35] D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2:193, 2008.

[36] W.H Li, C.I. Wu, and C.C Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2(2):150–174, 1985.

[37] B. Liu, J.F. Wendel, et al. Epigenetic phenomena and the evolution of plant allopolyploids. *Molecular Phylogenetics and Evolution*, 29(3):365–379, 2003.

[38] V.L. Louis, L. Despons, A. Friedrich, T. Martin, P. Durrens, S. Casarégola, C. Neuvéglise, C. Fairhead, C. Marck, J.A. Cruz, et al. *Pichia sorbitophila*, an interspecies yeast hybrid, reveals early steps of genome resolution after polyploidization. *G3: Genes— Genomes— Genetics*, 2(2):299–311, 2012.

[39] M. Lynch and J.S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.

[40] A. Madlung, R.W. Masuelli, B. Watson, S.H. Reynolds, J. Davison, and L. Comai. Remodeling of dna methylation and phenotypic and transcriptional changes in synthetic arabidopsis allotetraploids. *Plant Physiology*, 129(2):733–746, 2002.

[41] I. Mestiri, V. Chagué, A.M. Tanguy, C. Huneau, V. Huteau, H. Belcram, O. Coriton, B. Chalhoub, and J. Jahier. Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. *New Phytologist*, 186(1):86–101, 2010.

[42] T. Miyata and T. Yasunaga. Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1):23–36, 1980.

[43] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418–426, 1986.

[44] S. Ohno. *Evolution by Gene Duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag., 1970.

[45] B. Papp, C. Pál, and L.D. Hurst. Dosage sensitivity and the evolution of gene families in yeast. *Nature*, 424(6945):194–197, 2003.

[46] Keunwan Park and Dongsup Kim. Localized network centrality and essentiality in the yeast–protein interaction network. *Proteomics*, 9(22):5143–5154, 2009.

[47] P. Rice, I. Longden, and A. Bleasby. Emboss: The european molecular biology open software suite. *Trends in genetics*, 16(6):276–277, 2000.

[48] F. Rodriguez, J.L. Oliver, A. Marin, and J.R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142(4):485–501, 1990.

[49] D. Sankoff, C. Zheng, and Q. Zhu. The collapse of gene complement following whole genome duplication. *BMC Genomics*, 11(1):313, 2010.

[50] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC press, 1986.

[51] Lucy Skrabanek and Kenneth H Wolfe. Eukaryote genome duplication-where's the evidence? *Current opinion in genetics & development*, 8(6):694–700, 1998.

[52] J. Song and M. Singh. From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS computational biology*, 9(2):e1002910, 2013.

[53] C. Stark, B.J. Breitkreutz, L. Reguly, T.and Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

[54] M. Suyama, D. Torrents, and P Bork. Pal2nal: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, 34(suppl 2):W609–W612, 2006.

[55] John S Taylor and Jeroen Raes. Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.*, 38:615–643, 2004.

[56] B.C. Thomas, B. Pedersen, and M. Freeling. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome research*, 16(7):934–946, 2006.

[57] B. VanderSluis, J. Bellay, G. Musso, B. Costanzo, M. Papp, F.J. Vizeacoumar, A. Baryshnikova, B. Andrews, C. Boone, and C.L. Myers. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Molecular Systems Biology*, 6(1), 2010.

[58] K.H. Wolfe. Robustness - it's not where you think it is. *Nature Genetics*, 25(1):3–4, 2000.

[59] K.H. Wolfe and D. C. Shields. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, 387(6634):708–712, 1997.

[60] S. Wong, G. Butler, and K.H. Wolfe. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proceedings of the National Academy of Sciences*, 99(14):9272–9277, 2002.

[61] Z. Yang. Paml 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591, 2007.

[62] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular biology and evolution*, 17(1):32–43, 2000.

[63] A. Zharkikh. Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, 39(3):315–329, 1994.

Appendix A

Additional Concepts and Examples

A.1 Observing Selective Pressure

Evolutionary analysis of the *S. cerevisiae* network would be incomplete without an analysis of the evolutionary forces affecting each sequence that compose a network element. We know that any two homologous nucleotide sequences have some similarities and some differences. These differences, where one nucleotide in one sequence does not match the aligned nucleotide in the aligned sequence, are classified as either *synonymous* or *non-synonymous*.

To understand what synonymous and non-synonymous mean, we fall back to basic molecular biology and defining a codon: a unique sequences of three nucleotides that encodes a single amino acid or a stop signal for transcription. The total number of possible codons is therefore dependant on the number of nucleotides of which there are four, Guanine (G), Cytosine (C), Adenine (A) and Thymine (T). Therefore, the total number of codons are 4^3, or 64. If we remove the three stop transcription codons we are left with 61 codons to encode 20 amino acids. This implies that one or more amino acids must be encoded by more than one codon. We can clearly see that this is the case in Table A.1 and is termed the *degeneracy* of the genetic code.

Now, considering the fact that the genetic code is degenerate, a change in only one nucleotide may not necessarily change the encoded amino acid. For example, if we consider TAC (Tyr) $\Leftarrow$ TAT (Tyr) the third position mutated from a C to a T, however TAC encodes tyrosine and TAT still encodes tyrosine. This change, which does not affect the encoded amino acid, is termed synonymous.

The majority ($\tilde{7}2\%$) of third position mutations are synonymous [43]. This is not the case with first and second position mutations. Of first position mutations nearly all (9̃5%) result in non-synonymous mutations [43]. A non-synonymous mutation being one that resulted in a different amino acid being encoded. The notable exclusions to this rule are arginine and leucine, which can partially tolerate a mutation in their first codon. We can see this in Table A.1 if we consider the four major rows, T C, A and G (which themselves contain rows of T, C, A, G) we can see there are only two examples that span two rows. The two rows indicating a change in the first position.

Mutations in the second position of a codon will always result in a different amino acid being encoded. We can see this in Table A.1 as a horizontal change between columns: there are no amino acids that span two columns. Therefore we can assume that mutations in the second position are always non-synonymous.

Knowing these terms, we wish to find what extent pairs of nucleotide sequences are subject to synonymous substitutions and non-synonymous substitutions. More accurately, we wish to estimate the rate of non-synonymous, dN, and synonymous, dS, substitutions for each pair of sequences. By estimating the substitution rate we can start to unravel the evolutionary forces that shaped the two sequences. Synonymous substitutions are generally thought to follow a *neutral* model of evolution. Very simply, under the neutral model synonymous mutations are “not seen” by natural selection. Therefore, their promulgation to sustenance within a species (fixation) is entirely by random genetic drift [6]. This is in contrast to non-synonymous mutations, which, due to the fact they alter the amino acid being encoded, are “seen” by natural selection since they alter the resultant protein. Ideally, this means that deleterious mutations will be subject to *purifying (negative) selection*, removing the mutation from the gene pool and beneficial mutations will be subject to *positive selection*, promulgating the gene in the gene pool to fixation. However, as with all things in biology, the opposite scenarios may occur, deleterious mutations may be promulgated and beneficial mutations may be lost. Fortunately, we are only concerned with what we can observe, and therefore we can categorize genes as being either subject to positive, negative or neutral selection.

This brings us back to asking what sort of selective forces have pairs of sequences been subjected to, positive or negative? In order to answer these questions we need to understand how dN and dS are calculated. To accomplish this we will use a small example based on the approximation method of Nei and Gojobori [43]. More recent approximation methods have been introduced [62] but they do not lend themselves as easily to example.

There are three basic steps, common to all the various methods, for estimating synonymous and non-synonymous substitution rates for two aligned sequences. First, count the number of synonymous and non-synonymous sites in each of the two sequences. Second, count the number of synonymous and non-synonymous differences between the two sequences. Finally and third, account for multiple substitutions at the same site to determine the rates, dN and dS [62].

We now present an example that covers each step outlined above and describes difficulties arising in each.

	2nd Position				
1st Position	T	C	A	G	3rd Position
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Table A.1: The universal genetic code. Adapted from [1].

Counting Synonymous and non-Synonymous Sites

We begin our estimation of dN and dS by calculating the number of synonymous and non-synonymous sites. Throughout this thesis we have used the nomenclature of dN and dS to symbolize non-synonymous mutations per non-synonymous site and synonymous mutations per synonymous site, respectively. In the discussion that follows we use d_n interchangeably with dN and for d_s likewise with dS.

Let's consider Table A.2 that contains two nucleotide sequences, NT_a and NT_b, as well as their associated translated amino acid sequences, AA_a and AA_b, respectively.

$NT_a \Rightarrow$	`ATG TAC AAA GTA CCC TTA ACT AGC`
	`--- *-- -** --* *-- *-- --* --*`
$NT_b \Rightarrow$	`ATG CAC ATA GTC ACC ATA ACG AGG`
$AA_a \Rightarrow$	`M Y K V P L T S`
	`- * * - - * - *`
$AA_b \Rightarrow$	`M H I V T I T R`

Table A.2: Two aligned nucleotide sequences and their associated aminoacid sequences

The number of synonymous sites can be calculated independent of non-synonymous sites, the sum of both must equal the number of nucleotides in the sequence. If we define the number of codons as L_c, as in [62], the number of synonymous (S) and non-synonymous (N) sites is therefore $N = (3L_c - S)$. However, in practice, the two values are calculated in the same step.

Let C_{ij} represent the nucleotide pair aligned at position $j, 0 \leq j \leq 2$ in codon $i, 0 \leq i \leq L_c - 1$. We determine the which proportion each site, j, of codon pair, i, is synonymous, S_{ij}, and non-synonymous, N_{ij}. To simplify matters somewhat, $C_{i2} : 0 \leq i \leq L_c - 1$ is always a non-synonymous site as we described above.

Returning to the sequences in Table A.2 we will first calculate the number of synonymous, S, and non-synonymous sites, N, by using the Universal Genetic Code in Table A.1. We begin by testing the first nucleotide of the first codon in each of the two sequences, $C_{a_{00}}$ and $C_{b_{00}}$. Since the first codon in both sequences is the same we have only one set of changes to consider:

ATG: M	⇒ CTG: L
	⇒ TTG: L
	⇒ GTG: V

Table A.3: Effects of mutations on the first base of the first codon from Table A.2

We can see that changing the first nucleotide from adenosine to any of the other three nucleotides results in a non-synonymous mutation of methionine to either leucine or valine. This means that the first position of the first codon is a non-synonymous site in both sequences.

For the sake of completeness, we must then consider the number of sequences being compared and calculate how the probability of effects the values of S_{ij} and N_{ij}. Even though both have ATG as their first codon and that the set of changes are the same this may not always be the case and in other comparisons the two codons may not be the same. To weight the result for each sequence we use equation (A.5):

$$\begin{aligned} 1 &= S_{ij} + N_{ij} && \text{(A.1)} \\ S_{ij} &= P_a S_{a_{ij}} + P_b S_{bij} && \text{(A.2)} \\ N_{ij} &= P_a N_{a_{ij}} + P_b N_{bij} && \text{(A.3)} \\ & && \text{(A.4)} \end{aligned}$$

Here P_a is the probably for sequence a, $S_{a_{ij}}$ is the synonymous value in sequence a, codon i, nucleotide j. N_a is the fraction of non-synonymous sites in sequence a, codon i, nucleotide j. P_b, $S_{b_{ij}}$ and $N_{b_{ij}}$ all follow similarly for sequence b.

Substituting A.2 and A.3 into A.1 and rearranging the terms gives us:

$$= P_a(S_{a_{ij}} + N_{a_{ij}}) + P_b(S_{b_{ij}} + N_{b_{ij}}) \qquad \text{(A.5)}$$

Therefore, if we use $S_{a_{ij}} = 0$, $S_{b_{ij}} = 0$, $N_{a_{ij}} = 1$ and $N_{b_{ij}} = 1$ for the current step in our example the question then becomes what is P_a and P_b?

Since we are using only two sequences, the average number of synonymous and non-synonymous sites are used. This means that P_a and P_b are, $P_a = \frac{1}{2}$ and $P_b = \frac{1}{2}$.

This means that for $i = 0$, $j = 0$:

$$\begin{aligned} 1 = S_{00} + N_{00} &= \frac{1}{2}(0+1) + \frac{1}{2}(0+1) \\ &= \frac{1}{2} + \frac{1}{2} \\ S_{00} &= 0 && \text{(A.6)} \\ N_{00} &= 1 \end{aligned}$$

The result in (A.6) gives us 0 synonymous sites and 1 non-synonymous site for the first position. Calculating the weight for each sequence is an important step to perform. However, in situations where all the compared codons are the same a weighted calculation is not necessary: the final values of S_{ij} and N_{ij} are simply the values for any $S_{k_{ij}}$ and $N_{k_{ij}}$.

A	`ATG TAC AAA GTA CCC TTA ACT AGC`
B	`ATG CAC ATA GTC ACC ATA ACG AGG`
synonymous	`000`
non-synonymous	`111`

Table A.4: Known values following calculation of synonymous and non-synonymous sites for the first aligned codon pair

Moving to the second nucleotide of the first codon we recall that the second position is always a non-synonymous site. We do not need to calculate the weighted values, they are simply $S_{01} = 0$ and $N_{01} = 1$.

Next, we consider the third position. We can see that the $S_a = 0$ and $N_a = 1$ and therefore, as described above, $S_{a+b} = 0$ and $N_{a+b} = 1$.

$$\begin{aligned} \text{ATG: M} &\Rightarrow \text{ATA: I} \\ &\Rightarrow \text{ATC: I} \\ &\Rightarrow \text{ATT: I} \end{aligned}$$

Table A.5: Effects of mutations on the third base of the first codon from Table A.2

Moving to the second codon our calculations become slightly more interesting. Here, the two sequences have two different codons so our calculations will not be as simple as taking arbitrary $S_{k_{ij}}$ and $N_{k_{ij}}$. Instead we must use (A.5). Let's look at mutations that can occur:

$$\begin{aligned} \text{TAC: Y} &\Rightarrow \text{AAC: I} & \text{CAC: H} &\Rightarrow \text{AAC: N} \\ &\Rightarrow \text{CAC: I} & &\Rightarrow \text{TAC: Y} \\ &\Rightarrow \text{GAC: M} & &\Rightarrow \text{GAC: D} \end{aligned}$$

Table A.6: Effects of mutations on the first base of the second codons of NT_a and NT_b from Table A.2

Here, $S_{a_{10}} = 0$, $N_{a_{10}} = 1$, $S_{b_{10}} = 0$ and $N_{b_{10}} = 1$. Although there many differences here the final result is still $S_{10} = 0$ and $N_{10} = 1$.

Moving two the second nucleotide of the second codon we once again set the second position to be a non-synonymous site. Therefore, $S_{11} = 0$ and $N_{11} = 1$.

The third positions of the second aligned codons in this case will not be simple one and zero-values. This is because we are now looking at changes in two-fold degenerate sites. That is, of the four possible nucleotides that can be in the third base position, two code for the same amino acid. Assuming, of course, that the first two nucleotides are the same. Mutations in the codon at two-fold degenerate sites have a one-in-three chance of coding for the same amino acid. This is similar for four-fold degenerate sites except any change in the third base position has no effect on the amino acid, it is a purely synonymous change.

TAC: Y ⇒ TAA: Stop	CAC: H ⇒ CAA: Q
⇒ TAT: Y	⇒ CAT: H
⇒ TAG: Stop	⇒ CAG: Q

Table A.7: Effects of mutations on the third base of the second codons of NT_a and NT_b from Table A.2

It is clear from the above mutations that $S_{a_{12}} = 1/3$ (Y→Y) $N_{a_{12}} = 2/3$ (Y→Stop x 2), $S_{b_{12}} = 1/3$ (H→H) and $N_{b_{12}} = 2/3$ (H→Q x 2). We now use Equation (A.5):

$$\begin{aligned}
1 = S_{12} + N_{12} &= P_a(S_{a_{12}} + N_{a_{12}}) + P_b(S_{b_{12}} + N_{b_{12}}) \\
&= \frac{1}{2}(\frac{1}{3} + \frac{2}{3}) + \frac{1}{2}(\frac{1}{3} + \frac{2}{3}) \\
&= \frac{1}{6} + \frac{2}{6} + \frac{1}{6} + \frac{2}{6} \\
S_{12} &= \frac{1}{3} \\
N_{12} &= \frac{2}{3}
\end{aligned} \tag{A.7}$$

The end result is that $S_{12} = \frac{1}{3}$ and $N_{12} = \frac{2}{3}$. Our progress thus far can be seen below in Table A.8

NT_a	ATG	TAC	AAA	GTA	CCC	TTA	ACT	AGC
NT_b	ATG	CAC	ATA	GTC	ACC	ATA	ACG	AGG
synonymous	000	$00\frac{1}{3}$						
non-synonymous	111	$11\frac{2}{3}$						

Table A.8: Known values following calculation of synonymous and non-synonymous sites for the first and seconds aligned codon pair

We continue in this manner, for the remaining pairs of codons in our two sequences until we determine the values for each synonymous and non-synonymous site. Our final result is tabulated in Table A.9.

	ATG	TAC	AAA	GTA	CCC	TTA	ACT	AGC
	ATG	CAC	ATA	GTC	ACC	ATA	ACG	AGG
synonymous	000	$00\frac{1}{3}$	$00\frac{1}{2}$	001	001	$\frac{1}{6}0\frac{1}{2}$	001	$00\frac{1}{3}$
non-synonymous	111	$11\frac{2}{3}$	$11\frac{1}{2}$	110	110	$\frac{5}{6}1\frac{1}{2}$	110	$11\frac{2}{3}$

Table A.9: Final synonymous and non-synonymous sites and their values for all base pairs in NT_a and NT_b from Table A.2

Each value for the synonymous sites are summed as in equation (A.8), and similarly for the non-synonymous sites as in equation (A.11). The result is the total number of synonymous and non-synonymous sites for our example, $S = 4.8333$ and $N = 19.1667$.

$$S = \sum_{i=0}^{L_c-1}\sum_{j=0}^{2} S_{ij}$$

$$= 3 + \frac{1}{3} + \frac{1}{2} + \frac{1}{6} + \frac{1}{2} + \frac{1}{3} \tag{A.8}$$

$$= 4\frac{5}{6} \tag{A.9}$$

$$= 4.8333 \tag{A.10}$$

$$N = \sum_{i=0}^{L_c-1}\sum_{j=0}^{2} N_{ij}$$

$$= 16 + \frac{2}{3} + \frac{1}{2} + \frac{5}{6} + \frac{1}{2} + \frac{2}{3} \tag{A.11}$$

$$= 19\frac{1}{6} \tag{A.12}$$

$$= 19.1667 \tag{A.13}$$

Counting Synonymous and non-Synonymous Differences

After calculating the number of synonymous and non-synonymous sites (S and N) we need to classify each difference between NT_a and NT_b as a synonymous difference, S_d, or a non-synonymous difference, N_d. This is trivial when only one nucleotide is different between two aligned codons: it is a synonymous difference if the two codons encode the same amino acid; and, it is a non-synonymous difference if the two codons encode different amino acids.

AA_a	M	Y	K	V	P	L	T	S
AA_b	M	H	I	V	T	I	T	R
Differences (*)	-	*	*	-	-	*	-	*

Table A.10: Amino acid differences between the nucleotide sequences in A.2

For example, AAA (Lys) ⇔ AAG (Lys) differ by one nucleotide in the third position but both codons code for lysine. This means that the number of synonymous differences is one, $S_d = 1$. A non-synonymous difference can be explained by AAA (Lys) ⇔ AAT (Asn). The two codons differ in the third position, like the previous example, but in this case one codon codes for lysine and the other for asparagine. Therefore, the number of non-synonymous differences is one, N_d. This is the case with our example. There are four non-synonymous differences, leaving four synonymous differences.

Generally, however, things are not so simple. When determining S_d and N_d between codons we must account for three possible types of differences between the ith codon of sequence a, C_{a_i}, and the ith codon of sequence b, C_{b_i}:

1. One nucleotide difference between C_{a_i} and C_{b_i}
2. Two nucleotide differences between C_{a_i} and C_{b_i}
3. Three nucleotide differences between C_{a_i} and C_{b_i}

The first case was covered in the example presented above. For the second case, where there are two nucleotide differences between C_{a_i} and C_{b_i}, we must consider that there are always two possible pathways that change C_{a_i} into C_{b_i} (or vice-a-versa). For example, consider AAA (Lys) ⇔ ATC (Ile)

Pathway I AAA (Lys) ⇔ ATA (Ile) ⇔ ATC(Ile)
Pathway II AAA (Lys) ⇔ AAC (Asn) ⇔ ATC (Ile)

In this example Pathway I has a non-synonymous substitution followed by a synonymous substitution. In Pathway II there are two sequential non-synonymous substitutions. Which pathway is the most likely pathway? Are they both equally likely? We must somehow incorporate the answer to these questions into our calculation of the number of synonymous and non-synonymous sites.

There are two general approaches to dealing with this situation, weighted and un-weighted. The un-weighted method was introduced in 1986 by Nei and Gojobri [43]. It is so named as it assumes that both pathways are equally likely. It is calculated in the following way, for the codon pair at position i in sequence a and sequence b, C_{a_i}, C_{b_i}, we sum the number of synonymous substitutions at codon position i, D_{s_i}, and divide by 2. We do the same for the non-synonymous substitutions, D_{n_i}. We do this for all codon pairs and sum all D_{s_i} and D_{n_i} to find D_s and D_n.

$$
\begin{aligned}
D_s = \sum_{i=0}^{L_c-1} \frac{D_{s_i}}{2} \\
= \frac{1}{2} \\
= 0.5
\end{aligned}
\tag{A.14}
$$

$$
\begin{aligned}
D_n = \sum_{i=0}^{L_c-1} \frac{D_{n_i}}{2} \\
= \frac{3}{2} \\
= 1.5
\end{aligned}
\tag{A.15}
$$

Substituting the number of synonymous substitutions from both of our example pathways, $D_{s_0} = 1$, into Equation A.14 we find that the weighted number of synonymous substitutions for our single codon alignment is $D_s = 0.5$. Similarity, substituting $D_{n_0} = 3$ into Equation (A.15) results in the number of non-synonymous substitutions, $D_n = 1.5$.

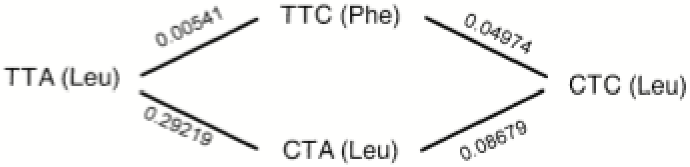


Figure A.1: Parsimonious pathways between TTA and CTC. Example from [62].

In reality, synonymous mutations are more likely to occur than non-synonymous mutations (Graur and Li, Chapter 4) [24]. This means that the pathway containing more synonymous mutations should be weighted more heavily to reflect this. The precise probability varies from sequence to sequence but empirically derived average values are typically used [36, 42]. The probability of each pathway is calculated and then used to determine the weight for each.

Figure A.1 gives an example where the total weight of the top pathway w_0 and the bottom pathway, w_1, from the probability for each, p_0 and p_1 is:

$$\begin{aligned}
p_0 &= (0.00541 \times 0.04974) \\
&= 0.00025 \\
p_1 &= (0.29219 \times 0.08679) \\
&= 0.02536 \\
w_0 &= \frac{p_0}{p_1} \\
&= \frac{0.00025}{0.02536} \qquad \text{(A.16)} \\
&= 0.011 \\
w_1 &= 1 - w_0 \\
&= 1 - 0.011 \\
&= 0.989
\end{aligned} \tag{A.17}$$

The above examples only consider codons with one or two nucleotide differences. The third, and final, case is when all three nucleotides between two codons are not the same. In this situation there are a total of 18 possible parsimonious pathways, such as in $AAA \Leftrightarrow TTT$, when changing one nucleotide at a time. The pathways themselves will be excluded for the sake of brevity, but the weighting (or lack-there-of) is analogous to the weighting for codons with two nucleotides.

Accounting for Multiple Substitutions at the Same Site

Returning to our example in Table A.2 we could naively calculate the rate of synonymous and non-synonymous mutations, d_s and d_n. We know that the number of synonymous sites, $S = 4.833$, and the number of non-synonymous sites $N = 19.167$. We also know from the second step the number of synonymous differences, $S_d = 4$, and non-synonymous difference $N_d = 4$. The values of d_s and d_n are calculated by:

$$d_s = \frac{S_d}{S} \tag{A.18}$$

$$d_n = \frac{N_d}{N} \tag{A.19}$$

Substituting our required values into Equation A.19 and Equations A.18 we find that $d_n = 0.209$ and $d_s = 0.828$. However, this would be underestimating both d_n and d_s. To understand why we would be underestimating these values it is important to understand that when we compare two sequences of nucleotides we are comparing the result of many thousands or millions of years of evolution: many mutations may have occurred that are not obvious by comparing the sequences. To accommodate this fact we must consider an additional parameter, the nucleotide substitution rate.

There are many other nucleotide substitution models [31, 33, 14, 27, 63, 48] that all have differing substitution rates and base frequencies. The earliest and simplest nucleotide substitution model was introduced by Jukes and Cantor in 1969 [31]. This model assumes that the substitution rates are equal and the base frequencies are equal. These assumptions result in the Jukes and Cantor formula (Equation A.20) which estimates the evolutionary distance, d, between two sequences based the number of nucleotide differences, p.

$$d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \tag{A.20}$$

In our case p is either $\frac{S_d}{S}$ or $\frac{N_d}{N}$ which results in d representing d_s or d_n, respectively.

Unfortunately, even using this model, we are under-estimating dS and over-estimating dS [62]. The only way to mitigate the over- or under-estimating of substitution rates, as much as possible, is to use a maximum-likelihood method [62]. This method incorporates the statistical likelihood of individual pathways, which we mentioned above. Thus, the generally preferred method for estimating these rates are maximum-likelihood, such as Goldman and Yang [21]. Goldman and Yang is implemented in the PaML[61] package which we utilized for our dN and dS calculations.

A.2 Statistical Methods

When applying any statistical method in our analyses we assume that our data is non-parametric univariate or non-parametric bivariate. Since some of the statistic methods we use herein may be foreign, or their interpretation non-intuitive, we provide brief description of the non-parametric statistical methods used in our analyses. These include Spearman's rank order correlation and kernel density estimation.

A.2.1 Spearman's Rank Order Correlations

Spearman's rank order correlation coefficient [19], ρ, is a non-parametric statistic that allows us to find monotonic trends within bivariate data. This differs from the related Pearson product-moment correlation coefficient in that it is a measure of linear correlation which is affected by non-linearity, non-normality and the presence of outliers: precisely those things that can arise in biological data sets.

The first step in calculating ρ for a particular sample $S = \{X = \{X_1, X_2, ..., X_n\}, Y = \{Y_1, Y_2, ..., Y_n\}\}$ is to rank X and Y independently. This ranking is equivalent to the index of the value if the set were ordered such that $x_i < x_j$ and $1 < i < j \leq |X|$. If the case arises where the same value appears more than once, it is not possible to rank each duplicate value individually. In this situation, the average rank among all the duplicates is assigned to each duplicate value.

Consider a bivariate population S that is composed of two paired data sets X and Y where $|X| = |Y| = n$,

$$\begin{aligned} S &= \{X, Y\} \\ X &= \{3, 1, 5, 5, 5, 7, 6, 6\} \\ Y &= \{5, 6, 7, 8, 9, 10, 11, 12\} \end{aligned} \tag{A.21}$$

The rank set, r of the values in each data set is then,

$$\begin{aligned} r_X &= \{2, 1, 4, 4, 4, 8, 6.5, 6.5\} \\ r_Y &= \{1, 2, 3, 4, 5, 6, 7, 8\} \end{aligned} \tag{A.22}$$

We then define a difference metric, d_i, between each pair (r_{X_i}, r_{y_i}) from $\{r_X, r_Y\}$,

$$d_i = x_i - y_i, \; 1 < i \leq n \tag{A.23}$$

which is used to compute ρ,

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{A.24}$$

There are three broad possible values for ρ, in each case the magnitude dictates the strength of the correlation:

- $0 < \rho \leq 1$, positive correlation between X and Y.
- 0, no correlation between X and Y.
- $-1 \leq \rho < 0$, negative correlation between X and Y.

Although we will not discuss the specifics of calculating an associated p-value for a correlation, we can then use it to provide a confidence interval for the correlation between X and Y.

A.2.2 Kernel Density Estimate

To introduce the idea of density estimation we refer to Silverman's overview *Density Estimation for Statistics and Data Analysis* [50]. Quite often, when trying to visualize the distribution of univariate data, a histogram is used. However, there are two major issues with this approach: the distribution of data in the histogram is entirely dependent on bin-sizing (Figures A.2a and A.2b); and the density distribution can be discontinuous if bin-sizes are too small. The second point may not always be important, but in a single population it is sometimes helpful that the data distribution is continuous as possible. This allows the inference of values outside the population being sampled. Regardless, a statistical method that provides a solution to both of these problems is the kernel density estimate (Figure A.2c).

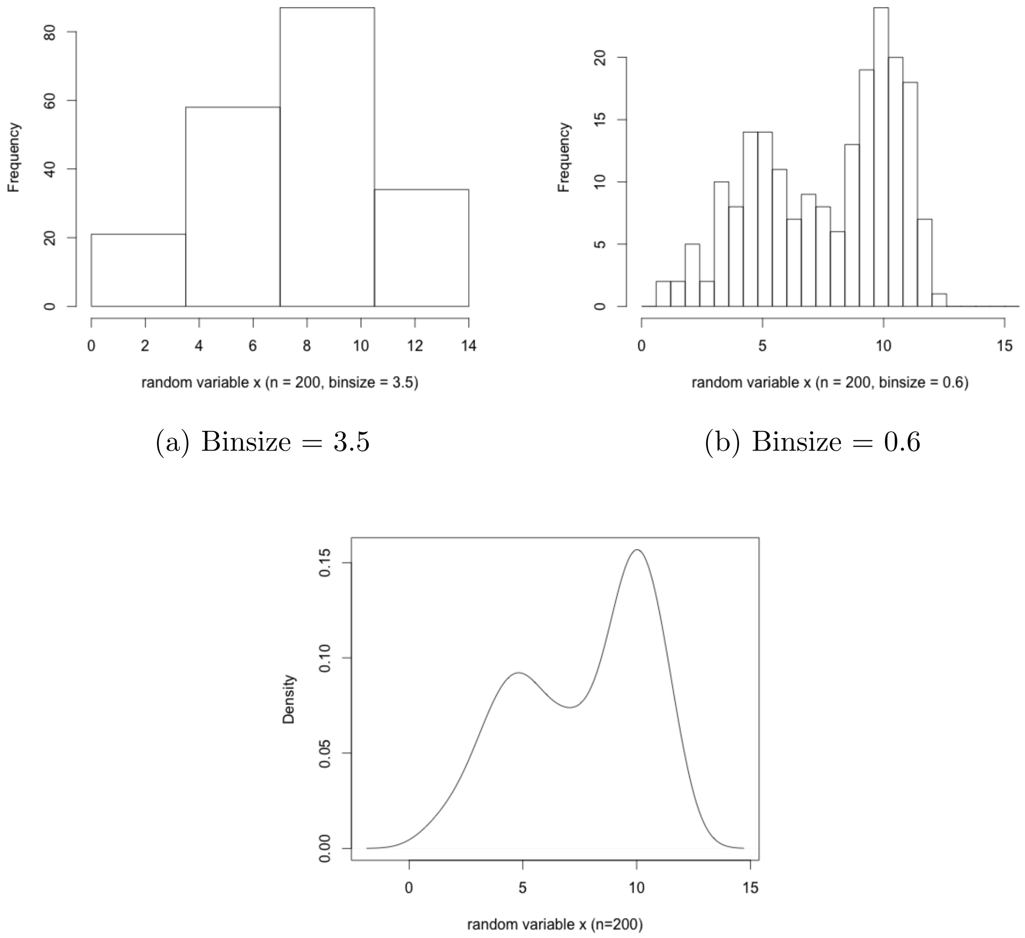
In parametric statistics, the well known univariate normal distribution, represented by a Gaussian bell curve, has a probability density function,

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{A.25}$$

Using the mean, μ and standard deviation σ for some population X we can estimate the density value of $x \in X$. We can then use our definition of f to find the probabilities associated with X within the interval (a, b),

$$P(a < X < b) = \int_a^b f(x)dx, \quad \forall a < b \tag{A.26}$$

The question now becomes: how do we compute the density function if we know nothing about the distribution of the data, i.e. it is non-parametric? In this situation



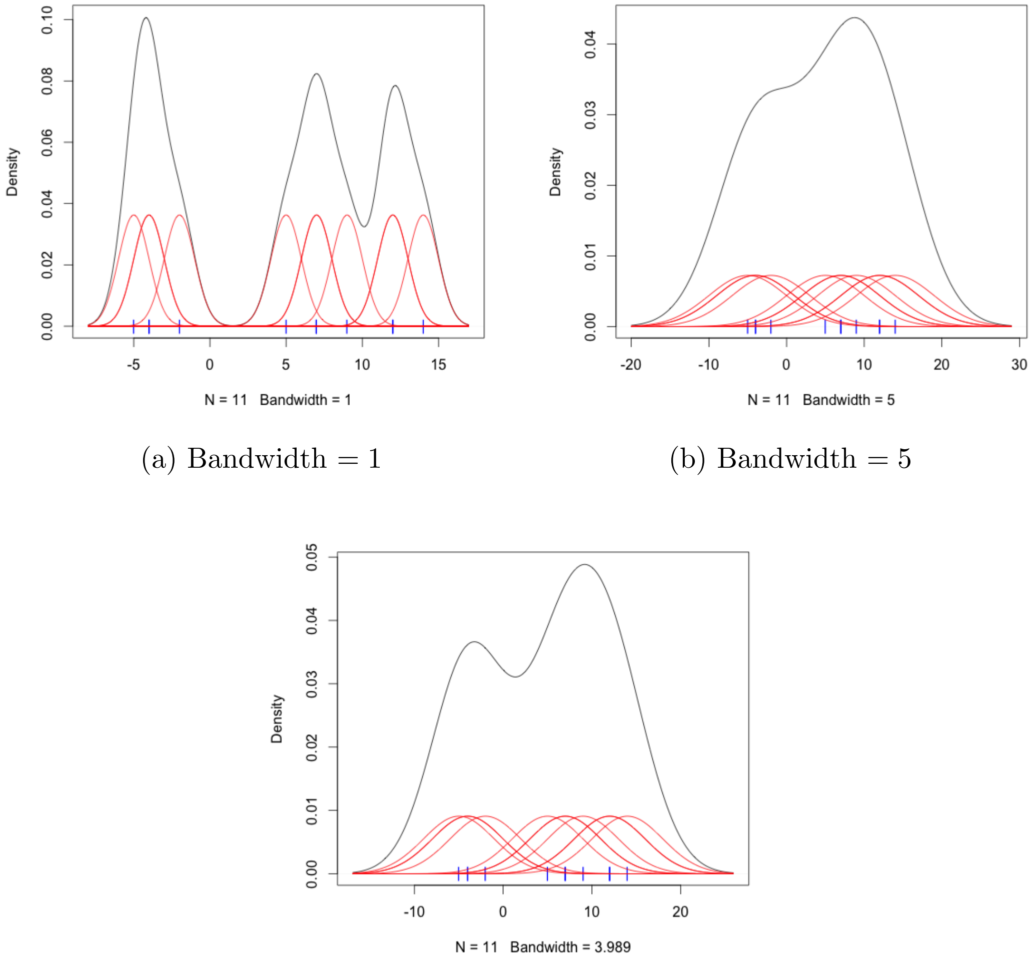
(a) Binsize = 3.5

(b) Binsize = 0.6

(c) Kernel Density Estimate

Figure A.2: Visualization of the distribution of the union of two normal data sets. Dataset 1: $\mu = 5$, $\sigma = 2$. Dataset 2: $\mu = 10$,$\sigma = 1$. (a), histogram with a bin-size of 3.5, the bimodal distribution of the data is not evident due to over-smoothing. (b), histogram with a bin-size of 0.6, the bimodal distribution of the data is clearly visualized. (c), kernel density estimate of the data clearly indicating a bimodal distribution.

we cannot use a parametric statistical function to calculate the probability density so we must estimate the density function.



(a) Bandwidth = 1

(b) Bandwidth = 5

(c) Kernel density estimate showing individual kernels (Gaussian, bandwidth = 3.989, $n = 11$)

Figure A.3: Kernel density plot of the values $\{-5, -4, -4, -2, 5, 7, 7, 9, 12, 12, 14\}$ using a bandwidth (smoothing) of 1 (a) and 5 (b). The near optimal bandwidth of 3.989, using "Sivermans's rule of thumb", is seen in (c). Kernels are coloured red. Kernel density estimate in black. Individual values in the population blue ticks on x-axis.

A univariate kernel density estimate is, unsurprisingly, a probability density estimate using a kernel. A univariate kernel is a function that describes the probability density about a single value X_i for all X_i in a non-parametric univariate population

$X = \{X_1, X_2, ..., X_n\}$. For our purposes we assume that $X \subset \mathbb{R}$. We use a kernel centered on each element $X_i \in X$ and then compute the probability of a value x within the kernel about this point. There are many different kernel shapes that can be used, however we will use the Gaussian kernel in both our explanation and our experimental analyses due to its simplicity. An example is visualized in Figure A.3c. The Gaussian kernel is defined in [50] as,

$$K(u) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \tag{A.27}$$

$$\int_{-\infty}^{+\infty} K(u) = 1 \tag{A.28}$$

We use the $K(u)$ to compute the probability density of a value $x \in \mathbb{R}$ [50],

$$f(x) = \frac{1}{|X|h} \sum_{i=1}^{|X|} K(u) \tag{A.29}$$

$$\tag{A.30}$$

where,

$$u = \frac{x - X_i}{h} \tag{A.31}$$

The fractional scalar in Equation A.30 ensures that $\int_{-\infty}^{+\infty} \sum_x f(x) = 1$. The *bandwidth*, h, is a smoothing factor analogous to σ and determines the width of the kernel. We can see the effects of bandwidth choice in Figure A.3.

The correct choice of bandwidth is key to the estimating the correct density function. There are multiple methods to choose from when calculating the optimal, or near optimal, bandwidth [50]: subjective choice, reference to a standard deviation, least-squares cross-validation, likelihood cross-validation, test graph method and integral estimation of density roughness. Without going into the details of each, we will simply state that our analyses use a method within the "reference to a standard deviation" category termed "Silverman's rule of thumb" (Equations 3.30 and 3.31 in [50]),

$$A = \min\{\sigma, (interquartile\ range/1.34)\} \tag{A.32}$$

$$h = 0.9A|X|^{-1/5} \tag{A.33}$$

The major benefits of this method is that it "will do very well for a wide range of densities and is trivial to evaluate"[50]. This contrasts with the other methods, which are much more computationally intensive and, depending on the method and data, may provide no better accuracy in bandwidth.

Of final note we refer to Equation A.28. Since density estimates, $f(x)$, must integrate to one, $f(x)$ from populations $\{X : 0 < X < 1\}$, must have values $f(x) > 1$. When $1 \leq X < +\infty$, $f(x) \leq 1$. Although it may be initially surprising to see a density estimate with a value much greater than one, it is important to note that the values of x may be much less than 1. The important factor in this that the area under the curve integrates to unity.

A.3 Network Centrality Examples

Here we present examples on how to calculate betweenness centrality and closeness centrality for the graph in Figure A.4.

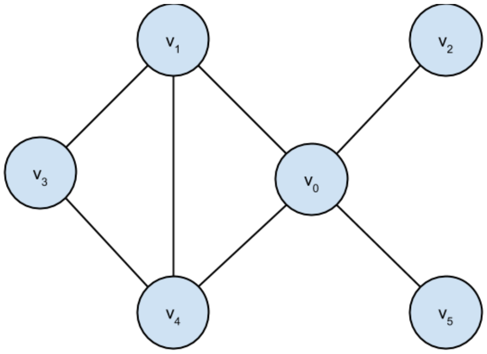


Figure A.4: A graph with six nodes and seven edges.

A.3.1 Calculating Betweenness Centrality

To calculate the normalized betweenness for Figure A.4 we first refer to Table A.11 which displays the combination of nodes that compose each unique shortest path for all pairs $(u, v) \in V \times V$.

	v_0	v_1	v_2	v_3	v_4	v_5
v_0	-	v_0v_1	v_0v_2	$v_0v_1v_3,\ v_0v_4v_3$	v_0v_4	v_0v_5
v_0	-	-	$v_1v_0v_2$	v_1v_3	v_1v_4	$v_1v_0v_5$
v_1	-	-	-	$v_2v_0v_1v_3,\ v_2v_0v_4v_3$	$v_2v_0v_4$	$v_2v_0v_5$
v_2	-	-	-	-	v_3v_4	$v_3v_1v_0v_5,\ v_3v_4v_0v_5$
v_3	-	-	-	-	-	$v_4v_0v_5$
v_4	-	-	-	-	-	-

Table A.11: Unique shortest paths between vertices in the graph of A.4. Pairs of vertices with more than one shortest path (both paths have equal length) have each path separated by a comma. In total there are 18 unique shortest paths.

We then use this information to calculate the weighted value of each path for which a node belongs in Table A.12. The values for each node are summed and their normalized values calculated in Table A.11.

$v \in V$	v_0v_1	v_0v_2	v_0v_3	v_0v_4	v_0v_5	v_1v_2	v_1v_3	v_1v_4	v_1v_5	v_2v_3	v_2v_4	v_2v_5	v_3v_4	v_3v_5	v_4v_5	$\sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$
v_0	0	0	0	0	0	1	0	0	1	$\frac{2}{2} = 1$	1	1	0	$\frac{2}{2} = 1$	1	7
v_1	0	0	$\frac{1}{2} = 0.5$	0	0	0	0	0	0	$\frac{1}{2} = 0.5$	0	0	0	$\frac{1}{2} = 0.5$	0	1.5
v_2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
v_4	0	0	$\frac{1}{2} = 0.5$	0	0	0	0	0	0	$\frac{1}{2} = 0.5$	0	0	0	$\frac{1}{2} = 0.5$	0	1.5
v_5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table A.12: Unnormalized weighted number of shortest paths that each $v \in V$ appears in for each s, t path in the graph of Figure A.4. Values were obtained from data in Table A.11. A value of 1 indicates that the particular v was in all s, t paths. Fractions indicate the number of paths s, t that contained vertex v divided by the total number of s, t paths, as described in the text. Unweighted betweenness values for each $v \in V$ are found in the last column: $v_0 = 7$, $v_1 = 1.5$, $v_4 = 1.5$, $_2, v_3, v_5 = 0$.

A.3.2 Calculating Closeness Centrality

We show the process of calculating closeness for Figure A.4. We first determine the geodesic between u and v, $d_G(v,t)$ for all pairs $(u,v) \in V \times V$ as shown in Table A.13. As a reminder, the shortest path, or a geodesic, between a node u and v in an undirected unweighed graph is the number of edges, $d_G(u,v)$, between u and v such that $d_G(u,v)$ is minimized (introduced in 2.3).

v	v_0	v_1	v_2	v_3	v_4	v_5	$\sum_{t\in V} d_G(v,t)$
v_0	-	1	1	2	1	1	6
v_1	1	-	2	1	1	2	7
v_2	1	2	-	3	2	2	10
v_3	2	1	3	-	2	3	11
v_4	1	1	2	2	-	2	8
v_5	1	2	2	3	2	-	10

Table A.13: Geodesic distances between nodes of the graph in A.4.

Using the summed values in Table

v	$\sum_{t\in V} d_G(v,t)$	$C'_C(v)$
v_0	6	5/6
v_1	7	5/7
v_2	10	5/10
v_3	11	5/11
v_4	8	5/8
v_5	10	5/10

Table A.14: Sum of geodesic distances for all nodes and the resultant normalized closeness centrality for the graph in A.4.