Counting Subnetworks Under Gene Duplication in Genetic Regulatory Networks

by
Ashley Scruse, Jonathan Arnold, Robert Robinson

Abstract

Gene duplication is a fundamental evolutionary mechanism that contributes to biological complexity and diversity [6]. Traditionally, research has focused on the duplication of gene sequences [21]. However, evidence suggests that the duplication of regulatory elements may also play a significant role in the evolution of genomic functions [8; 19]. In this work the evolution of regulatory relationships belonging to gene-specific-substructures in a GRN are modeled. In the model, a network grows from an initial configuration by repeatedly choosing a random gene to duplicate. The likelihood that the regulatory relationships associated with the selected gene are retained through duplication is determined by a vector of probabilities. That is to say that each gene family has its own probability of retaining regulatory relationships. Occurrences of gene-family-specific substructures are counted under the gene duplication model. In this thesis gene-family-specific substructures are referred to as subnetwork motifs. These subnetwork motifs are motivated by network motifs which are patterns of interconnections that recur more often in a specialized network than in a random network [15]. Subnetwork motifs differ from network motifs in the way that subnetwork motifs are instances of gene-family-specific substructures while network motifs are isomorphic substructures. These subnetwork motifs are counted under Full and Partial Duplication, which differ in the way in which regulation relationships are inherited. Full duplication occurs when all regulatory links are inherited at each duplication step, and Partial Duplication occurs when regulation inheritance varies at each duplication step. Note that Full Duplication is just a special case of Partial Duplication. Moments for the number of occurrences of subnetwork motifs are determined in each model. In the end, the results presented offer a method for discovering gene-family-specific substructures that are significant in a GRN under gene duplication.

1 INTRODUCTION

Traditionally the study of gene duplication has a primary focus on sequence duplication, which involves discovering similar regions of DNA that contain a homologous gene [21]. Of equal importance is how the regulation evolves during the duplication process. This evolution of function has yet to be studied in the same depth as duplication of sequence. In this paper we will use combinatorial probability to investigate the duplication of gene regulation inside of a genetic regulatory network (GRN).

GRNs are collections of genes and their products that interact with one another to control a specific cell function and they play a vital role in different cellular processes. These genes and their regulatory elements create complex networks with unexplained design properties [15]. An approach to discovering some of the structural design properties is to look for network motifs, which are defined as patterns of interconnections that recur more often in the complex network than in a special randomized network [15]. These motifs can be found in complex networks, but in this paper we will consider motifs within eukaryotic transcriptional networks [10] [9].

Network motifs in transcriptional networks constitute the building blocks of these networks [15] [14]. They can be constructed by identifying most or all of the transcription factors in a genome and identifying their binding sites to other genes in the genome [17]. The result is that the transcription factors with the binding sites link genes into a transcriptional network for an entire eukaryotic genome [10] [9]. A key challenge is to identify these network motifs within a eukaryotic transcriptional network. The usual tool for doing so is simulation and envisioning that instances of a particular motif as the product of randomized networks of similar structure[15]. However in this paper we examine the evolution of gene-family-specific network motifs, which will be referred to as subnetwork motifs, under gene duplication.

Subnetwork motifs are distinguished from network motifs by being specific substructures associated with particular gene families, whereas network motifs represent substructures that are isomorphic across different gene families. The difference can also been seen in Figure 1. While network motifs are applicable to complex networks in general, this paper aims to further explore the concept of network motifs by focusing on the

occurrences of subnetwork motifs in a GRN. By focusing on subnetwork motifs found in GRNs, this research intends to help researchers determine which gene families and regulation relationship are significant enough to explore.

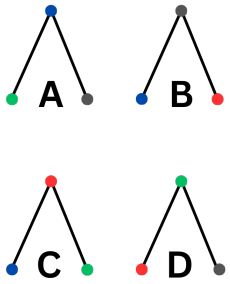

Figure 1: Each color (blue, green, red, grey) represents a different gene family. Then under the definition of network motifs found in [15] A, B, C, D would be grouped under 3-node network motifs, since they are isomorphic subgraphs. However, under the subnetwork motifs definition presented in this thesis A, B, C, D would be considered 4 different subnetwork motifs, since the specified gene families are different in each subgraph.

The framework of subnetwork motifs for identifying building blocks assumes that there are functional labels on the families in the GRN. This is not the case for network motifs [15]. The discovery of these important subnetwork motifs will depend on the how the regulation matrix was inferred and what data was used to infer it. For example, in the case of the *Arabidopsis* clock, a feedforward network motif was fruitfully identified [16]. On the other hand a transcriptional network for yeast was not fruitful in identifying some of the regulatory links (particularly post-transcriptional links) in several well studied GRNs [12]. What is also interesting about subnetwork motifs is that they incorporate evolutionary information, which network motifs do not do [3]. This can be used to further validate subnetwork motifs as over or under-represented. Also, some healthy skepticism should be maintained when applying subnetwork motifs to identifying important building blocks that are based on both the data specifying the regulatory links and the evolutionary information available on the gene families in the motif.

Identifying subnetwork motifs in GRNs gives clues to the function of genetic networks. Network motifs in GRNs can help us to understand the functions of networks and allow us to compare how the regulation of networks has evolved or can be evolved by engineering [20]. At the heart of synthetic biology is altering regulation rather than sequence, and it is now possible to create libraries of regulators with different functions and carry out directed evolution on the wiring of genetic networks for improved enzyme activity [20]. In order for this program of directed evolution through the regulation to work, it is necessary to identify "the selected motifs" driving the evolution of new functions. Being able to identify "significant" network motifs is central to identifying the function of regulatory networks from their components, how they evolve, and how synthetic biology can be used in protein engineering through selection on regulation rather than sequence. The results of this paper will provide a way to identify the "significant" network motifs.

The gene duplication and inheritance model that is presented in this paper is adapted from a gene duplication model for biological networks presented in [2]. In the model presented in [2], each inheritance mode, Full Duplication and Partial Duplication, are controlled by a single probability respectively. However, in our model each gene family has its own probability of inheriting regulatory links through duplication. That is to say that the gene duplication and inheritance model uses a vector of probabilities to determine the inheritance of regulatory relationships and the vector is determined by the mode of inheritance. In the end, we find that our model is a generalization of the model presented in [2].

In this paper we will use combinatorial probability to study the occurrences of subnetwork motifs in our gene duplication and inheritance model. We begin by defining a stochastic gene duplication process that

governs the gene duplication and inheritance model under all variations of the inheritance vector. Then we define in detail a subnetwork motif and observe their occurrences in both Full and Partial Duplication.

To create the framework for carrying out significance tests for subnetwork motifs we will calculate two moments: the mean and the variance. We will begin with defining the gene duplication process that applies to both Full Duplication and Partial Duplication. Then we will differentiate the models of inheritance before calculating the moments for each model. In the end we will present exact results and some asymptotic results for the moments.

2 CONSTRUCTING THE GENE DUPLICATION AND INHERITANCE MODEL

2.1 THE GENE DUPLICATION PROCESS

The *gene duplication process* is a stochastic process that begins with m individual genes, where $n \geq m \geq 1$. Initially, each of these genes is the sole member of its gene family, called the i^{th} family for $i = 1, 2, 3, ..., m$ in some arbitrary but fixed order. At each step, a random gene is selected to be duplicated. If a gene in the i^{th} family is duplicated then the new duplicated gene belongs to the i^{th} family. After d duplications the total number of genes will be $n = m + d$. Often n will be referred to as the *stage* of the gene duplication process. This gene duplication model

Proposition 1. *Suppose there are c_i genes in the i^{th} family, where $c_i \in \mathbb{Z}^+$ for $i = 1, ..., m$. Then $\vec{c} = (c_1, c_2, ..., c_m)$ is a composition of n into m parts where $\sum c_i = n$. From now on we will discuss the results of a series of duplications in terms of compositions. It turns out that for given m and n all such compositions are equally likely.*

This result characterizes the uniformity of the duplication process that governs both Full and Partial Duplication inheritance modes.

Proof. We induct on $n \geq m$. For the base case $n = m$ the only possible composition is $(1, ..., 1)$. Since $\binom{m-1}{m-1} = 1$ the base case is verified.

For the induction step, assume that that $n > m$, and let p be the probability that after $n-m$ duplications the composition is $(c_1, ..., c_m)$. Here $c_i \geq 1$ for $i = 1, ..., m$ and $c_1 + \cdots + c_m = n$. Now $p = p_1 + \cdots + p_m$, where p_i is the probability that the given composition is the result of duplicating a member of the i^{th} family at the $(n-m)^{th}$ duplication step. In order for the i^{th} family to be duplicated at the $(n-m)^{th}$ step the composition after the previous step must have been $\vec{\gamma} = (c_1, ..., c_{i-1}, c_i - 1, c_{i+1}, ..., c_m)$. We claim that $p_i = (c_i - 1)/\binom{n-2}{m-1}$. If $c_i = 1$ this yields $p_i = 0$, which is correct since $\vec{\gamma}$ is not a possible composition in this case. If $c_i \geq 2$ then the probability of $\vec{\gamma}$ is $1/\binom{n-2}{m-1}$ by the induction hypothesis, and the probability given $\vec{\gamma}$ that the next duplication is in the i^{th} family is $(c_i - 1)/(n-1)$. Taking the product we find

$$p_i = (c_i - 1)/\left((n-1)\tbinom{n-2}{m-1}\right).$$

Since

$$(n-1)\tbinom{n-2}{m-1} = (n-m)\tbinom{n-1}{m-1}$$

and

$$\textstyle\sum_{i-1}^{m}(c_i - 1) = n - m$$

we have that

$$p_{=}\textstyle\sum_{i=1}^{m} p_i = (n-m)/\left((n-m)\tbinom{n-1}{m-1}\right) = 1/\tbinom{n-1}{m-1}.$$

Finally, if we sum over all the possible compositions of n into m parts, then the total probability must be 1, so there are $\binom{n-1}{m-1}$ equally likely compositions. □

It is important to note that there are multiple ways to arrive at $\binom{n-1}{m-1}$ as the total number of equally likely compositions of n into m parts. From a combinatorial point of view let a family be a bin for gene markers and each family is understood to initially contain 1 gene marker which will not be shown explicitly, leaving $n-m$ gene markers represented by 0, and $m-1$ dividers represented by $|$, to be arranged in linear order. Note that the combinatorial symbols for gene markers are identical and the dividers are identical. Then the linear arrangement has $n-1$ locations that have $m-1$ dividers.

Consider the case where $m=3$ and $n=6$ such that $\vec{c}=(2,1,3)$. Then the combinatorial representation is

$$0||00$$

where the two dividers creates three blocks of zeros with lengths $1, 0$, and 2 representing family sizes $2, 1$, and 3 respectively.

The number of gene markers and dividers that are going to be arranged in linear order is the number of dividers plus the additional gene markers that need to be place and $m-1+n-m=n-1$. Thus there are $\binom{n-1}{m-1}$ possible linear arrangements for n gene markers to be placed in m families.

An ordinary generating function is an alternative way of counting the compositions of n into m parts. Using a generating function an infinite sequence of numbers can be expressed by allowing those numbers to be the coefficients of a formal power series [7]. We find that the use of generating functions will be a convenient way to calculate some of the numbers we are going to need when analyzing subnetwork motifs since generating functions can be represented as Taylor Series of explicit rational functions.

If you let an infinite series of numbers be denoted by $g_0, g_1, ..., g_j$ where $g_i \geq 0$ then its *generating function* is defined as the infinite series

$$g(x) = (g_0x^0 + g_1x^1 + ... + g_jx^j + ...) = (g_0 + g_1x + ... + g_jx^j + ...).$$

Let α be a real number and $j \geq 0$ be an integer then

$$\binom{\alpha}{j} \equiv \frac{(\alpha)_j}{j!}$$

where $(\alpha)_j$ is a falling factorial. Note that this is an extension of the standard binomial definition since taking α to be a non-negative integer would yield the standard binomial coefficient.

Lemma 1. *Let $\alpha \geq 0$ be a real number and $j \geq 0$ be an integer. Then,*

$$[x^j](1-x)^{-\alpha} = \binom{\alpha+j-1}{j}.$$

Proof. This follows from the Taylor Series at $x=0$. Note that for any integer $j \geq 0$, the operator $[x^j]$ returns the coefficient of x^j in a power series. Thus $[x^j]g(x) = g_j$ in the infinite power series. Notice that

$$[x^j]xg(x) = 0$$

if $j=0$ and

$$[x^j]xg(x) = g_{j-1}$$

if

$$j \geq 1.$$

In general,

$$[x^j]x^\ell g(x) = 0$$

if $j < \ell$ and

$$[x^j]x^\ell g(x) = g_{j-\ell}$$

if $j \geq \ell$. □

Lemma 2. *Let $\ell \geq 0$ be a non-negative integer. Then the ordinary generating function for the number of compositions into ℓ parts is*

$$x^{\ell}(1-x)^{-\ell}.$$

Proof. When $\ell \geq 1$ the Lemma follows directly Lemma 1 and Proposition 1. However, when $\ell = 0$ the Lemma still stands since there is only 1 composition of 0 into 0 parts and there are no compositions of n into 0 parts when $n \geq 0$. □

There is also an alternative derivation of the generating function that we will find useful. For compositions of n into 1 part we know the generating function to be
$g(x) = (x + x^2 + x^3 + ...) = \frac{x}{1-x}$. In order to obtain the generating function for the number of compositions of n into ℓ parts we can allow each of the ℓ parts be denoted by $g(x)$. Therefore the generating function for the compositions of n into ℓ parts is

$$g(x)^{\ell} = (x + x^2 + x^3 + ...)^{\ell} = (\frac{x}{1-x})^{\ell}.$$

Let $m \geq 1$ be fixed with $n \geq m$, and consider the process of starting from the composition $(1, ..., 1)$ (of dimension m) and performing $n - m$ duplications. The resulting vector $(X_{m,n}^{(1)}, ..., X_{m,n}^{(m)})$ is a composition of n into m parts. From Proposition 1 we know that there are $\binom{n-1}{m-1}$ such vectors, all with the same probability of occurring.

We will find the following special notation for vectors helpful in the remainder of the thesis. For vectors $\vec{x} = (x_1, ..., x_h)$ and $\vec{y} = (y_1, ..., y_h)$ of the same dimension h we let $\vec{y} \leq \vec{x}$ denote the conjunction of the h inequalities $y_1 \leq x_1, ..., y_h \leq x_h$. We define $<, \geq$, and $>$ for vectors similarly. We also define a special operator $|| \vec{x} ||$ such that $|| \vec{x} || = \sum x_i$, and special constant vectors $\vec{0} = (0, ..., 0)$ and $\vec{1} = (1, ..., 1)$. For the latter the dimension is to be made clear by context.

In general, the gene duplication process should be seen as a random markov process. To make calculating the expectations more convenient we present the following lemma.

Lemma 3. *Let m and n be integers such that $1 \leq m \leq n$, and for $i = 1, ..., m$ let $\mathbb{U}_i$ be a real valued function over the positive integers and $\mathbb{U}_i(x) = \sum_{j=1}^{\infty} U_i(j)x^j$. Then the sum of $\prod_{i=1}^{m} U_i(c_i)$ over the compositions $\vec{c} = (c_1, ..., c_m)$ of n into m parts is*

$$\sum_{\vec{c}} \Big(\prod_{i=1}^{m} U_i(c_i) \Big) = [x^n] \prod_{i=1}^{m} \mathbb{U}_i(x).$$

Proof. Suppose $(c_1, ..., c_m)$ is a partition of n into m parts. Then $\prod_{i=1}^{m} U_i(c_i)$ arises as $[x^n](U_1(c_1)x^{c_1} \cdot ... \cdot U_m(c_m)x^{c_m})$ since $c_1 + ... + c_m = n$. That is one monomial that is obtained from expanding the product into a single generating function. Each partition of n into m parts similarly contributes its own share to $[x^n] \prod_{i=1}^{m} \mathbb{U}_i(x)$.

To obtain the result of the Lemma we expand the product of these generating functions into a sum of monomials. Then to arrive at a single generating function, we collect the monomials that correspond to each power of x. Note that since the exponents sum to n, the only monomials that can correspond to x^n must have exponents that sum to n and therefore arise as one of the partitions of n into m parts. □

3 INTRODUCTION TO SUBNETWORK MOTIFS

The subnetwork motifs discussed in this thesis are gene-family-specific network motifs. A network motif is said to be a pattern of interconnections that occur more often in a complex network than a special random network [15]. From a biological point of view, a subnetwork motif is a gene-family-specific network motif that occurs more in real data than expected from the moments calculated in the coming sections. It is important to note that the gene-family-specific-substructures discussed in this thesis will be called *subnetwork motifs* although for biological applications *subnetwork motif candidate* would be more apposite.

Subnetwork motifs are going to be built from the genes that arise from the gene duplication process. For the purposes of this thesis, a subnetwork motif $\mathbb{M}$ is characterized by k gene families and the chances of

creating a new subnetwork motif instance in duplication. Given a subnetwork motif $\mathbb{M}$, we will assume the indices of the families that belong to $\mathbb{M}$ are 1 to k for notational convenience. Then by definition the set of k original genes forms the original instance of $\mathbb{M}$. At any stage n during the duplication process the set of instances of $\mathbb{M}$ will be denoted $\mathbb{M}(n)$. When $n = m$, the only subnetwork motif instance present is the original instance of $\mathbb{M}$; therefore $|\mathbb{M}(m)| = 1$. Since instances of $\mathbb{M}$ will always have dimension k we will denote the vector of family sizes that belong to $\mathbb{M}$ as $\vec{s} = (s_1, ..., s_k)$, where $s_i = |X^{(i)}_{m,n}|$. For any $n \geq m$ new instances of the subnetwork motif are possible at stage $n + 1$. Suppose there is a subnetwork motif instance $\mathscr{I} = (a_1, ..., a_i, ...a_k) \in \mathbb{M}(n)$. If a gene a'_i is duplicated from a_i at stage $n + 1$ then $\mathscr{I}' = (a_1, ..., a'_i, ...a_k)$ is a potential new instance of $\mathbb{M}$ and if $\mathscr{I}'$ is inherited then $\mathscr{I}' \in \mathbb{M}(n+1)$. At any stage $n \geq m$, $\mathbb{M}(n)$ consists of the original instance of $\mathbb{M}$ and the additional instances of $\mathbb{M}'$ that were inherited through the duplication process.

Each subnetwork motif $\mathbb{M}$ has an associated vector of probabilities, $\vec{\pi} = (\pi_1, ..., \pi_k)$. Here π_i is the probability that an instance $\mathscr{I} = (a_1, ..., a_i, ...a_k) \in \mathbb{M}(n)$ gives rise by inheritance to the new instance $\mathscr{I}' = (a_1, ..., a'_i, ...a_k)$ when the gene in the i^{th} family is duplicated. The general concept of subnetwork motif inheritance is central to the thesis. Simpler results are obtained when $\vec{\pi} = \vec{1}$ which is called Full Duplication. A more general case is also analyzed where $\vec{0} \leq \vec{\pi} \leq \vec{1}$ and this is called Partial Duplication.

To study the expected size of $\mathbb{M}(n)$, the gene duplication process and the subnetwork motif inheritances process are combined into one process. The *random duplication and inheritance process* is a stochastic process that begins with m individual genes and a k sized subnetwork motif that has an associated vector of probabilities $\vec{\pi} = (\pi_1, ..., \pi_k)$, where $n \geq m \geq k \geq 1$ and $\vec{0} \leq \vec{\pi} \leq \vec{1}$. Initially, each of these genes are the sole member of its gene family called the i^{th} family and $\mathbb{M}$ is the original subnetwork motif instance in $\mathbb{M}(n)$. At each step, a random gene is selected for duplication. If a gene is duplicated from the i^{th} family, then the new gene belongs to the i^{th} family and there is a π_i chance that the regulations are inherited at that step if $1 \leq i \leq k$. After $n - m$ duplications we are interested in the size of $\mathbb{M}(n)$.

3.1 FULL DUPLICATION

This section will cover the *Full Duplication* model, which is the inheritance model that is controlled by the probability vector $\vec{\pi} = \vec{1}$. Suppose $k \geq 1$ and $\mathbb{M}$ is a subnetwork motif of size k. Every possible instance is duplicated since $\vec{\pi} = \vec{1}$. Therefore $\mathbb{M}(n) = X^{(1)}_{m,n} \times ... \times X^{(k)}_{m,n}$, so that $|\mathbb{M}(n)| = |X^{(1)}_{m,n}| \cdot ... \cdot |X^{(k)}_{m,n}|$. We will start by analyzing the expected value of $|\mathbb{M}(n)|$ given m, n, and k.

Theorem 1. *Assume $\mathbb{M}$ is a subnetwork motif of size k and $1 \leq k \leq m \leq n$. Then the expected number of instances of $\mathbb{M}$ given the random duplication process is*

$$\mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) = \frac{\Gamma(n+k)\Gamma(m)}{\Gamma(n)\Gamma(m+k)}.$$

This result allows us to construct a significance test for subnetwork motifs using the mean of the number of instances of $\mathbb{M}$ under Full Duplication.

Proof. Let $\vec{c} = (c_1, ..., c_m)$ where c_i is the size of the i^{th} family, that is $c_i = |X^{(i)}_{m,n}|$. Then $|\mathbb{M}(n)| = c_1 \cdot ... \cdot c_k = c_1 \cdot ... \cdot c_k \cdot 1^{m-k}$. To calculate the expectation we use Lemma 3 to evaluate $\sum |\mathbb{M}(n)|$ over all the compositions of n into m parts then divide the result by $\binom{n-1}{m-1}$. In order apply Lemma 3 we let $\mathbb{U}_i(x) = (x + 2x^2 + ...)$ when $1 \leq i \leq k$ and $\mathbb{U}_i(x) = (x + x^2 + ...)$ when $k + 1 \leq i \leq m$. We will call the series $a(x)$ and $b(x)$ for convenience. When $1 \leq i \leq k$ the i^{th} family contributes to $|\mathbb{M}(n)|$ and $a(x) = (x + 2x^2 + ...) = \frac{x}{(1-x)^2}$. Furthermore, when $k + 1 \leq i \leq m$ the i^{th} family does not contribute to $|\mathbb{M}(n)|$ and $b(x) = (x + x^2 + ...) = \frac{x}{(1-x)}$.

It is easy to see the direct correlation of the above generating functions as rational functions of x. However, it can also be seen by applying Lemma 1 and its proof when $\alpha = 2$ for $a(x)$ and $\alpha = 1$ for $b(x)$. Thus, applying Lemma 3 with $\mathbb{U}_i(x) = a(x)$ if $k + 1 \leq i \leq m$ and $\mathbb{U}_i(x) = b(x)$ if $1 \leq i \leq k$ we obtain

$$\sum_{\vec{c}} |\mathbb{M}(n)| = x^{[n]}\big(\frac{x}{1-x}\big)^{m-k}\big(\frac{x}{(1-x)^2}\big)^{k}$$
$$= x^{[n]}\big(\frac{x^m}{(1-x)^{m+k}}\big)$$
$$= \binom{n+k-1}{m+k-1},$$

where the third equality follows from Lemma 1 when $\alpha = m + k$.

Note that the result of the Theorem is expressed in terms of the gamma function as it will be convenient in the Partial Duplication section. The gamma function $\Gamma(z) = (z-1)!$ is defined on all of the complex plane except for $z \leq 0$ an integer [1]. Therefore, $\Gamma(n) = (n-1)!$ as long as $n \geq 1$ is an integer.

Recall from Proposition 1 that there are exactly $\binom{n-1}{m-1}$ compositions of n into m parts and they are equally likely. Thus

$$\mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) = \frac{\binom{n+k-1}{m+k-1}}{\binom{n-1}{m-1}}$$
$$= \frac{\Gamma(n+k)\Gamma(m)}{\Gamma(n)\Gamma(m+k)}.$$

□

Similar to the combinatorial explanation after Proposition 1, $\binom{n+k-1}{m+k-1}$ can also be derived combinatorially. We are now interested in the linear arrangements considering k sized subnetwork motifs. Given a particular $\vec{c} = (c_1, ..., c_k, c_{k+1}, ..., c_m)$, we add a selector to the first k families such that there are k selectors represented by $|$. The selectors indicate which gene in that family has been selected for the subnetwork motif so that the k selected genes form a subnetwork motif. This leaves $n - m$ gene markers represented by 0 and $m + k - 1$ selectors and dividers represented by $|$ to be arranged in linear order. Note that the first k odd placements for pipes are selectors and the first k even placements for pipes are dividers.

Consider the case where $k = 2, m = 3$ and $n = 6$ such that $\vec{c} = (3, 1, 2)$. Let g_i be the original gene for the i^{th} family for $1 \leq i \leq m$ and g_i' be the first duplicated gene in the i^{th} family where the number of primes represents the order in which the gene was duplicated. Then the combinatorial representation for (g_1'', g_2, g_3') is

$$00||||0$$

where the first pipe is selecting the second duplicated gene marker for the first family, the second pipe is the first divider, the third pipe is selecting the original gene marker from the second family, and the fourth pipe is the second divider.

The number of genes, dividers, and selectors that are going to be arranged in linear order is the number of dividers and selectors plus the number of additional genes that need to be placed and $m+k-1+n-m = n+k-1$. Thus the total number of linear arrangements considering a k sized subnetwork motifs is $\binom{n+k-1}{m+k-1}$.

Corollary 1. *Suppose $1 \leq k \leq m \leq n$ and $n \geq k^2$. Then the expected number of instances of subnetwork motifs of size k in Full Duplication satisfies*

$$\frac{n^k}{(m+k-1)_k} \leq \mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) \leq \frac{n^k}{(m+k-1)_k}\Big(1 + \frac{3k^2}{4n}\Big).$$

Proof. From Theorem 1 we know that $\mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big)$ is

$$\frac{\Gamma(n+k)\Gamma(m)}{\Gamma(n)\Gamma(m+k)} = \frac{(n+k-1)_k}{(m+k-1)_k}.$$

The bounds for the expectation fall directly from Lemma 14 (occurs in Section 6). □

We are going to use the same general approach to evaluating the second moment of the number of instances of $\mathbb{M}$ in Full Duplication which we will denote $\mathbb{E}\Big(|\mathbb{M}(n)|^2 \ ; \ k,m,n\Big)$.

Theorem 2. *Assume $\mathbb{M}$ is a subnetwork motif of size k and $1 \leq k \leq m \leq n$. Then the second moment of the number of instances of $\mathbb{M}$ in Full Duplication is evaluates to*

$$\frac{\Gamma(n+k)\Gamma(m)\Gamma(n-m+1)}{\Gamma(n)}\sum_{i=0}^{k}\binom{k}{i}2^i\big(\Gamma(m+k+i)\Gamma(n-m-i+1)\big)^{-1}.$$

This result allows us to calculate the variance of the number of instances of $\mathbb{M}$ under Full Duplication, which would be required for a significance test.

Proof. Let $\vec{c} = (c_1, ..., c_m)$ where c_i is the size of the i^{th} family, that is $c_i = |X^{(i)}_{m,n}|$. Thus $|\mathbb{M}(n)|^2 = (c_1 \cdot ... \cdot c_k)^2 = (c_1^2 \cdot ... \cdot c_k^2) \cdot 1^{m-k}$. In order to calculate the second moment we can proceed as in the proof of Theorem 1 and use Lemma 3 to evaluate $\sum |\mathbb{M}(n)|^2$ over all the compositions of n into m parts then divide the result by $\binom{n-1}{m-1}$. Recall from the proof of Theorem 1 the ordinary generating functions used to evaluate $\sum \mathbb{M}(n)$ over the compositions of n into m parts are:

$$a(x) = (x + 2x^2 + ...) = \Big(\frac{x}{(1-x)^2}\Big) \text{ and } b(x) = (x + x^2 + x^3 + ...) = \Big(\frac{x}{1-x}\Big).$$

In order to apply Lemma 3 we must modify $a(x)$ since the product first k family sizes has been squared. To evaluate the second moment we will replace $a(x)$ with another series we denote $y(x)$ where

$$y(x) = x \cdot \frac{d}{dx}\big(a(x)\big) = (1 + 4x + 9x^2 + ...).$$

When $1 \leq i \leq k$ the i^{th} family contributes to $|\mathbb{M}(n)|$, $y(x) = \frac{2x^2}{(1-x)^3} + \frac{x}{(1-x)^2}$ and when $k+1 \leq i \leq m$ the family does not contribute to $|\mathbb{M}(n)|$, $b(x) = \frac{x}{1-x}$. Therefore, we apply Lemma 3 with $\mathbb{U}_i(x) = y(x)$ when $1 \leq i \leq k$ and $\mathbb{U}_i(x) = b(x)$ when $k+1 \leq i$. By the binomial theorem $y(x)^k$ can be expressed as

$$y(x)^k = \sum_{i=0}^{k}\binom{k}{i}\Big(\frac{2x^2}{(1-x)^3}\Big)^i\Big(\frac{x}{(1-x)^2}\Big)^{k-i},$$

which gives the following

$$\begin{aligned}
\sum_{\vec{c}} |\mathbb{M}(n)|^2 &= x^{[n]}(\frac{x}{1-x})^{m-k}\sum_{i=0}^{k}\binom{k}{i}(\frac{2x^2}{(1-x)^3})^i(\frac{x}{(1-x)^2})^{k-i} \\
&= [x^n]\sum_{i=0}^{k}\binom{k}{i}\frac{2^i x^i x^m}{(1-x)^{k+m+i}} \\
&= \sum_{i=0}^{k}2^i\binom{k}{i}\binom{n+k-1}{n-m-i} \\
&= \sum_{i=0}^{k}2^i\binom{k}{i}\binom{n+k-1}{m-1+k+i}.
\end{aligned}$$

Since the $\binom{n-1}{m-1}$ compositions of n into m parts are equally likely

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2 \ ; \ k,m,n\Big) = \binom{n-1}{m-1}^{-1} \cdot \sum_{i=0}^{k}2^i\binom{k}{i}\binom{n+k-1}{m-1+k+i},$$

which evaluates to

$$\frac{\Gamma(n+k)\Gamma(m)\Gamma(n-m+1)}{\Gamma(n)}\sum_{i=0}^{k}2^i\binom{k}{i}\big(\Gamma(m+k+i)\Gamma(n-m-i+1)\big)^{-1},$$

where $\Gamma(z)$ is the special function that is noted in the proof of Theorem 1. □

Similar to Theorem 1, we will find a combinatorial explanation for the binomial coefficient portion of the second moment calculation. Consider an ordered pair of subnetwork motifs $(\mathscr{I}_1, \mathscr{I}_2) \in \mathbb{M}(n)^2$ where $\mathscr{I}_1$ is inherited first and $\mathscr{I}_2$ is inherited second. Since we are observing ordered pairs of subnetwork motifs there will be $0 \leq i \leq k$ families that have two gene indicators selecting for two different gene markers and $k-i$ families that will have one gene indicator selecting for the same gene marker. Consider the case where $k = 3, m = 4, n = 6$, and $i = 2$ where $\vec{c} = (2, 2, 1, 1)$. Suppose the first two families have two gene indicators, $\mathscr{I}_1 = (g_1, g_2, g_3)$ comes first, $\mathscr{I}_2 = (g_1', g_2', g_3)$ come second, g_1 is duplicated earlier, and g_2 is duplicated later. Then the combinatorial representation of the ordered pair of subnetwork motifs is

$$|0|||0||||.$$

To visually show the distinction between the two types of selectors and dividers we show an annotated version of the combinatorial representation

$$\underline{|}0\underline{|}\hat{|}0\underline{|}\underline{|}\underline{|}$$

where $\underline{|}$ represents selectors that select for different genes and $\hat{|}$ represents selectors that select for the same gene.

Given an ordered pair of subnetwork motifs, the process begins with m original genes that are not shown explicitly, there are $n-m$ additional gene markers that need to be placed. Note that for each of the i families with double gene selectors there is a gene indicator that is not shown explicitly to ensure separation for each pair of selectors. Therefore, there are $m-1+k+i$ dividers and selectors and the number of gene markers, dividers, and selectors that are going to be arranged in linear order is the number of dividers and selectors plus the number of additional gene markers and $m-1+k+i+n-m+i = n+k-1$. Thus the linear arrangement has $n+k-1$ locations that have $m-1+k+i$ selectors and dividers.

Note that there are $\binom{k}{i}$ ways the families with 2 selectors could have been chosen and 2^i ways the selectors could be arranged, since the selector that is selecting for the first subnetwork motif can come before or after the selector that is selecting for the second subnetwork motif.

Corollary 2. *Suppose $1 \leq k \leq m \leq n$. Then the variance of the number of instances of subnetwork motifs of size k in Full Duplication is*

$$\binom{n-1}{m-1}^{-1}\cdot\Big(\sum_{i=0}^{k}\binom{k}{i}2^i\binom{n+k-1}{m+k-1+i}\Big)-\Big(\frac{(n+k-1)_k}{(m+k-1)_k}\Big)^2.$$

This result allows us to construct a significance test for subnetwork motifs under Full Duplication.

Proof. The corollary follows immediately from Theorems 1 and 2. Furthermore, when $m = 1$ or when $m = n$ the variance is 0 and non-zero in all other cases. □

For the rest of the section we will present the inequalities in the following form

$$n^{rk}\mathbb{C}_i\Big(1-\mathbb{L}_i/n\Big)\leq\mathbb{Y}\leq n^{rk}\mathbb{C}_i\Big(1+\mathbb{U}_i/n\Big)$$

where the power of n depends on the r^{th} moment of $\mathbb{Y}, \mathbb{C}$ is the constant for the main term and $\mathbb{L}$ and $\mathbb{U}$ are the error terms for the upper and lower bounds respectively. Using this notation to present Corollary 1 the inequalities will take the following form

$$n^k\mathbb{C}_1\Big(1-\mathbb{L}_1/n\Big)\leq\mathbb{E}\Big(|\mathbb{M}(n)|\ ;\ k,m,n\Big)\leq n^k\mathbb{C}_1\Big(1+\mathbb{U}_1/n\Big)$$

where $\mathbb{C}_1 = \Gamma(m)/\Gamma(m+k), \mathbb{L}_1 = 0,$ and $\mathbb{U}_1 = 3k^2$.

Corollary 3. *Suppose $1 \le k \le m \le n$ and $n \ge 3m^2$. Then the second moment of the number of instances of subnetwork motifs of size k in Full Duplication is*

$$n^{2k}\mathbb{C}_2\Big(1 - \mathbb{L}_2/n\Big) \le \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big) \le n^{2k}\mathbb{C}_2\Big(1 + \mathbb{U}_2/n\Big)$$

where $\mathbb{C}_2 = 2^k/(m+2k-1)_{2k}, \mathbb{L}_2 = 4m^2,$ and $\mathbb{U}_2 = 9m^2$.

Proof. From the proof of Theorem 2 we know that

$$\mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big) = \sum_{i=0}^{k} \binom{k}{i} 2^i \binom{n+k-1}{m-1+k+i} \binom{n-1}{m-1}^{-1}$$

$$= \sum_{i=0}^{k} \binom{k}{i} 2^i \frac{(n-1+k)!(m-1)!(n-m)!}{(n-1)!(m-1+k+i)!(n-m-i)!}$$

$$= \sum_{i=0}^{k} \binom{k}{i} 2^i \frac{(n-1+k)_k (n-m)_i}{(m-1+k+i)_{k+i}}. \tag{1}$$

In order to obtain bounds for the second moment in Full Duplication we will be bounding Equation (1) above and below.

Since all the terms in the summation are positive we can take the $i = k$ term as the lower bound since any one of the $k+1$ is a lower bound for the second moment.

Doing this we obtain

$$2^k \cdot \frac{(n+k-1)_k (n-m)_k}{(m+2k-1)_{2k}} \le \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big).$$

In order to derive a simpler lower bound we will bound the term below. Since m and k are both fixed the denominator remains the same and n^k is a trivial lower bound for $(n+k-1)_k$. Therefore, the only non-trivial bound comes from the term $(n-m)_k$. Note that

$$(n-m)_k = \frac{(n)_{m+k}}{(n)_m}$$

which we can bound using Lemma 16 and Lemma 17 (both occur in Section 6),

$$n^k\Big(1 - \frac{(m+k)^2}{n}\Big) \le \frac{(n)_{m+k}}{(n)_{(m)}}.$$

Therefore the lower bound for the second moment is

$$\frac{2^k n^{2k}}{(m+2k-1)_{2k}}\Big(1 - \frac{4m^2}{n}\Big) \le \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big).$$

We will now find the upper bound of the second moment. The upper bound is a summation of the upper bounds of the individual terms. The most important term when bounding the second moment above is the $i = k$ term. We know from the calculation of the lower bound that the $i = k$ term is as follows

$$2^k \cdot \frac{(n+k-1)_k (n-m)_k}{(m+2k-1)_{2k}}.$$

Now we will bound the $i = k$ term above. Since n^k is the trivial upper bound for $(n-m)_k$ we will only use Lemma 14 (occurs in Section 6) to bound the term $(n+k-1)_k$ above,

$$2^k \cdot \frac{(n+k-1)_k (n-m)_k}{(m+2k-1)_{2k}} \leq \frac{2^k \cdot n^{2k}}{(m+2k-1)_{2k}} (1 + \frac{3k^2}{4n}).$$

It is important to note that each of the $k+1$ terms has the same error bound that can be bounded above

$$(1 + \frac{3k^2}{4n}) \leq \frac{5}{4}$$

since $3k^2/4n \leq 1/4$.

When $i \leq k$ let $j = k-1$ so that $j = 0, 1, ...k$. Then we can bound each of the $k+1$ terms above with the following expression

$$\begin{aligned}
\mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big) &\leq \frac{2^k n^{2k}}{(m+2k-1)_{2k}} (1 + \frac{3k^2}{4n} + \frac{5}{4}(\frac{k(m+2k-1)}{2n} + ... + \frac{(m+2k-2)_{2k}}{2^k n^k (m+k-1)_k})) \\
&\leq \frac{2^k n^{2k}}{(m+2k-1)_{2k}} (1 + \frac{3m^2}{4n} + \frac{3m^2}{2n} + \frac{5}{4}(\frac{9m^4}{8n^2} + ... \frac{(3m)^{2m}}{(2n)^m (2m)^m})) \\
&\leq \frac{2^k n^{2k}}{(m+2k-1)_{2k}} (1 + \frac{3m^2}{4n} + \frac{5}{4} \sum_{j=1}^{k} \frac{m^j (3m)^j}{2^j j! n^j}) \\
&\leq \frac{2^k n^{2k}}{(m+2k-1)_{2k}} (1 + \frac{9m^2}{2n}).
\end{aligned}$$

□

We will now begin to discuss the variance of subnetwork motif occurrences in the Full Duplication mode. Recall that $k \geq 1$ and $\mathbb{M}$ represents a subnetwork motif of size k after $n-m$ duplications. In order to calculate the variance we will use the expressions of the first and second moments. Throughout the rest of the section the variance will be denoted $\mathbb{VAR}\Big(|\mathbb{M}(n)| \ ; \ k, m, n)\Big)$, where $\vec{s} = (s_1, s_2, ..., s_k)$ is the vector of the k family sizes and $\sum \vec{s} = h$.

As usual we will calculate the variance using the second moment and the square of the first moment. In order to simplify the error of the square of the first moment we will find bounds for $e^{2\delta}$ where δ is as presented in the proof of Lemma 14 (occurs in Section 6).

It is important to note that the variance is identically 0 when $m = 1$ since the number of subnetwork motifs is a point distribution in this case. For $m \geq 2$ we obtain the following inequalities which will be useful for large n.

Corollary 4. *Let $1 \leq k \leq m \leq n$, $m \geq 2$, and $n \geq 3m^2$. Then the variance of the number of instances of subnetwork motifs of size k in Full Duplication is*

$$n^{2k} \mathbb{C}_\nu \Big(1 - \mathbb{L}_\nu / n\Big) \leq \mathbb{VAR}\Big(|\mathbb{M}(n)| \ ; \ k, m, n)\Big) \leq n^{2k} \mathbb{C}_\nu \Big(1 + \mathbb{U}_\nu / n\Big)$$

where $\mathbb{C}_\nu = \mathbb{C}_2 - \mathbb{C}_1^2, \mathbb{L}_\nu = (\mathbb{C}_2 \mathbb{L}_2 - 2\mathbb{C}_1^2 \mathbb{U}_1)/(\mathbb{C}_2 - \mathbb{C}_1^2)$, *and*
$\mathbb{U}_\nu = \mathbb{C}_2 \mathbb{U}_2 / (\mathbb{C}_2 - \mathbb{C}_1^2)$.

Proof. In order to calculate the variance we will take advantage of the inequalities from the first and second moments. Using the standard definition of variance we know

$$\mathbb{VAR}\Big(|\mathbb{M}(n)|\ ;\ k,m,n)\Big) = \mathbb{E}\Big(|\mathbb{M}(n)^2|\ ;\ k,m,n\Big) - \mathbb{E}\Big(|\mathbb{M}(n)|\ ;\ k,m,n\Big)^2.$$

It is important to note that when $m = 1$ the variance is identically 0. Therefore we will calculate the variance for $m \geq 2$.

Recall from Corollary 3 that

$$n^{2k}\mathbb{C}_2\Big(1 - \mathbb{L}_2/n\Big) \leq \mathbb{E}\Big(|\mathbb{M}(n)^2|\ ;\ k,m,n\Big) \leq n^{2k}\mathbb{C}_2\Big(1 + \mathbb{U}_2/n\Big)$$

and from Corollary 2

$$n^{2k}\mathbb{C}_1^2\Big(1 - \mathbb{L}_1/n\Big) \leq \Big(\mathbb{E}\Big(|\mathbb{M}(n)|\ ;\ k,m,n\Big)\Big)^2 \leq n^{2k}\mathbb{C}_1^2\Big(1 + 2\mathbb{U}_1/n\Big)$$

by applying Lemma 17 (occurs in Section 6).

Thus the variance is bounded above by

$$\begin{aligned}
\mathbb{VAR}\Big(|\mathbb{M}(n)|\ ;\ k,m,n)\Big) &\leq n^{2k}\mathbb{C}_2 + n^{2k-1}\mathbb{C}_2\mathbb{U}_2 - n^{2k}\mathbb{C}_1^2 + n^{2k-1}\mathbb{C}_1^2\mathbb{L}_1 \\
&\leq n^{2k}(\mathbb{C}_2 - \mathbb{C}_1^2) + n^{2k-1}(\mathbb{C}_2\mathbb{U}_2 + \mathbb{C}_1^2\mathbb{L}_1) \\
&\leq n^{2k}(\mathbb{C}_2 - \mathbb{C}_1^2)\Big(1 + \frac{\mathbb{C}_2\mathbb{U}_2 - \mathbb{C}_1^2\mathbb{L}_1}{n(\mathbb{C}_2 - \mathbb{C}_1^2)}\Big).
\end{aligned}$$

and bounded below by

$$\begin{aligned}
\mathbb{VAR}\Big(|\mathbb{M}(n)|\ ;\ k,m,n)\Big) &\geq n^{2k}\mathbb{C}_2 - n^{2k-1}\mathbb{C}_2\mathbb{L}_2 - n^{2k}\mathbb{C}_1^2 - n^{2k-1}\mathbb{C}_1^2\mathbb{U}_1 \\
&\geq n^{2k}(\mathbb{C}_2 - \mathbb{C}_1^2) - n^{2k-1}(\mathbb{C}_2\mathbb{L}_2 - \mathbb{C}_1^2\mathbb{U}_1) \\
&\geq n^{2k}(\mathbb{C}_2 - \mathbb{C}_1^2)\Big(1 + \frac{\mathbb{C}_2\mathbb{L}_2}{n(\mathbb{C}_2 - \mathbb{C}_1^2)}\Big).
\end{aligned}$$

since $\mathbb{L}_1 = 0$. Recall that

$$\mathbb{C}_1^2 = \Big(\frac{1}{(m+k-1)_k}\Big)^2$$

and

$$\mathbb{C}_2 = \frac{2^k}{(m+2k-1)_{2k}}.$$

Then

$$\frac{2^k}{(m+2k-1)_{2k}} - \frac{1}{((m+k-1)_k)^2} > 0$$

since

$$2^k((m+k-1)_k)^2 > (m+2k-1)_{2k}.$$

□

Corollary 5. *Let $k \geq 1$ be fixed, $n \geq m \geq k^2$, then the expected number of instances of subnetwork motifs of size k in Full Duplication is*

$$(\frac{n}{m})^k\Big(1 - \frac{k^2}{2m}\Big) \leq \mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) \leq (\frac{n}{m})^k\Big(1 + \frac{3k^2}{4n}\Big).$$

Proof. From the result of Corollary 4 we know that

$$\frac{n^k}{(m+k-1)_k} \leq \mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) \leq \frac{n^k}{(m+k-1)_k}(1 + \frac{3k^2}{4n})$$

when m is fixed. We can apply Lemma 15 (occurs in Section 6) to approximate the bounds for $1/(m+k-1)_k$ when m is growing. Doing so we obtain

$$\frac{1}{m^k}(1 - \frac{k^2}{2m}) \leq \frac{1}{(m+k-1)_k} \leq \frac{1}{m^k}.$$

□

Corollary 6. *Let $k \geq 1$ be fixed, $n \geq m \geq k^2$, then the second moment of the number of instances of subnetwork motifs of size k in Full Duplication is*

$$2^k(\frac{n}{m})^{2k}(1 - \frac{4m^2}{n})(1 - \frac{2k^2}{m}) \leq \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big) \leq 2^k(\frac{n}{m})^{2k}(1 + \frac{9m^2}{n})$$

Proof. We know from Corollary 3 that the second moment is

$$\frac{n^{2k}2^k}{(m+2k-1)_{2k}}(1 - \frac{4m^2}{n}) \leq \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, m\Big) \leq \frac{n^{2k}2^k}{(m+2k-1)_{2k}}(1 + \frac{9m^2}{n})$$

when m is fixed. Using Lemma 15 (occurs in Section 6) we can bound $1/(m+2k-1)_{2k}$ as m grows. Doing so we obtain

$$\frac{1}{m^{2k}}(1 - \frac{2k^2}{m}) \leq \frac{1}{(m+2k-1)_{2k}} \leq \frac{1}{m^{2k}}.$$

□

Corollary 7. *Let $k \geq 1$ be fixed, $n \geq m \geq k^2$, then the variance of the number of instances of subnetwork motifs of size k in Full Duplication is*

$$\Big(\frac{n}{m}\Big)^{2k}\Big(2^k - 1\Big)\Big(1 - \frac{2^{k+3}+3}{(2^k-1)m} - \frac{m^2}{(2^k-1)n}\Big) \leq \mathbb{VAR}\Big(|\mathbb{M}(n)| \ ; \ k, m, n)\Big)$$

$$\mathbb{VAR}\Big(|\mathbb{M}(n)| \ ; \ k, m, n)\Big) \leq \Big(\frac{n}{m}\Big)^{2k}\Big(2^k - 1\Big)\Big(1 + \frac{k^2}{(2^k-1)m} + \frac{2^k 9m^2}{(2^k-1)n}\Big).$$

Proof. In order to calculate the variance we will use the standard expression of variance as stated below

$$\mathbb{VAR}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big) = \mathbb{E}\Big(|\mathbb{M}(n)^2| \ ; \ k, m, n\Big) - \mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big)^2.$$

Using the result of Corollary 1 we will calculate the bounds of the square of the first moment

$$\Big(\frac{n}{m}\Big)^{2k}\Big(1 - \frac{k^2}{m}\Big) \leq \mathbb{E}\Big(|\mathbb{M}(n)| \ ; \ k, m, n\Big)^2 \leq \Big(\frac{n}{m}\Big)^{2k}\Big(1 + \frac{3k^2}{2n}\Big)$$

where the lower bound error follows from Lemma 15 (occurs in Section 6) and the upper bound error follows from Lemma 18 (occurs in Section 6).

Using the result of Corollary 6 (occurs in Section 6) we can calculate bounds for the variance as follows,

$$\mathbb{VAR}\Big(|\mathbb{M}(n)|\ ;\ k,m,n\Big) \leq 2^k\Big(\frac{n}{m}\Big)^{2k}\Big(1+\frac{9m^2}{n}\Big)-\Big(\frac{n}{m}\Big)^{2k}\Big(1-\frac{k^2}{m}\Big)$$

$$\leq \Big(\frac{n}{m}\Big)^{2k}\Big(2^k-1+\frac{k^2}{m}+\frac{9m^2 2^k n^{2k}}{n}\Big)$$

$$\leq \Big(\frac{n}{m}\Big)^{2k}\Big(2^k-1\Big)\Big(1+\frac{k^2}{(2^k-1)m}+\frac{2^k 9m^2}{(2^k-1)n}\Big),$$

$$\mathbb{VAR}\Big(|\mathbb{M}(n)|\ ;\ k,m,n\Big) \geq 2^k\Big(\frac{n}{m}\Big)^{2k}\Big(1-\frac{4m^2}{n}\Big)\Big(1-\frac{2k^2}{m}\Big)-\Big(\frac{n}{m}\Big)^{2k}\Big(1+\frac{3k^2}{2n}\Big)$$

$$\geq 2^k\Big(\frac{n}{m}\Big)^{2k}\Big(1-\frac{2k^2}{m}-\frac{4m^2}{n}\Big)-\Big(\frac{n}{m}\Big)^{2k}\Big(1+\frac{3k^2}{2n}\Big)$$

$$\geq \Big(\frac{n}{m}\Big)^{k}\Big(2^k-1\Big)\Big(1-\frac{2^k 2k^2}{m(2^k-1)}-\frac{2^k 8m^2+3k^2}{2n(2^k-1)}\Big)$$

$$\geq \Big(\frac{n}{m}\Big)^{k}\Big(2^k-1\Big)\Big(1-\frac{2^{k+3}+3}{(2^k-1)m}-\frac{m^2}{(2^k-1)n}\Big).$$

□

4 PARTIAL DUPLICATION

The Partial Duplication mode occurs under the *random duplication and inheritance process*, which is an extension of the gene duplication process that involves random inheritance at every step that is controlled by a vector of probabilities, $\vec{\pi}$. At the end of the random duplication and inheritance process there is a set of subnetwork motifs given $m, n, k, \vec{\pi}$ and this section will cover the expectation of that.

A *Partial Duplication* mode is an inheritance model that covers the expectation at some stage n of a subnetwork motif $\mathbb{M}$ for some arbitrary k. The significance of this section is that the expected number of instances of $\mathbb{M}$ depends on a vector of inheritance probabilities $\vec{0} \leq \vec{\pi} \leq \vec{1}$. Recall that each $\mathbb{M}$ has an associated vector of probabilities $\vec{\pi} = (\pi_1, ..., \pi_k)$. However, the previous section covers a special case of this inheritance mode where $\vec{\pi} = \vec{1}$. It'll turn out that $\vec{\pi}$ completely determines the expected number of subnetwork motifs for any given stage n in duplication given the random duplication and inheritance process.

4.1 FIRST MOMENTS

To begin evaluating the expected number of subnetwork motif instances of any size k, we will look at instances of single gene subnetwork motifs where $k = 1$. Let s be the size of the family at any stage n and p be the probability of inheritance. In this case, we define $f(p, s)$ as the expectation of a single gene subnetwork motif $\mathbb{M}$. The expected number of instances of these single gene subnetwork motifs only depends on p and s which follows from the proof of the following lemma. We will show that the expectation can be expressed as a ratio of gamma functions and we will later show that we can express the expectation as a generating function which will be useful when we let $k \geq 1$.

Lemma 4. *Assume $\mathbb{M}$ is a single gene subnetwork motif. Let s be the size of the gene family and p be the probability of inheritance. Then the expected number of instances of $\mathbb{M}$ under Partial Duplication is*

$$f(p,s) = \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)}.$$

Proof. We will prove the statement of the lemma by induction on s.

Assume there have been no duplications in the family belonging to $\mathbb{M}$; then the only possible single gene subnetwork motif is the original instance of the subnetwork motif. If $s = 1$ then

$$f(p,1) = \frac{\Gamma(p+1)}{\Gamma(1)\Gamma(p+1)} = 1$$

since $\Gamma(1) = 1$.

Assume the induction hypothesis is true for some $s \geq 1$. We know that when the family size is s the expected number of instances are $f(p,s)$ and the probability of adding one gene to the family belonging to $\mathbb{M}$ upon duplication is $\frac{f(p,s)}{s} \cdot p$ since each gene has an equal chance of being selected for duplication. Thus,

$$f(p,s+1) = f(p,s) + f(p,s) \cdot \frac{p}{s} = f(p,s)\Big(1 + \frac{p}{s}\Big) = \frac{p+s}{s} f(p,s)$$

Using the induction hypothesis the expectation is as follows

$$\frac{p+s}{s} f(p,s) = \frac{p+s}{s} \cdot \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)} = \frac{\Gamma(p+s+1)}{\Gamma(s+1)\Gamma(p+1)}.$$

□

Suppose $\mathbb{M}$ is a subnetwork motif of arbitrary size $k \geq 1$. Assume as in section 3.1 that for notational convenience the families that belong to $\mathbb{M}$ are indexed from 1 to k. To begin evaluating the expected number of subnetwork motif instances of size $k \geq 1$, let $\vec{s} = (s_1, ..., s_k)$ where s_i is the size of the i^{th} family at any stage n for $i = 1, ..., k$ and $\vec{\pi} = (\pi_1, ..., \pi_k)$ gives the inheritance probabilities associated with $\mathbb{M}$.

Theorem 3. *Suppose $\mathbb{M}$ is a subnetwork motif of arbitrary size $k \geq 1$. Let $k \leq m \leq n$ and $\vec{0} \leq \vec{\pi} \leq \vec{1}$ be fixed. Let $\vec{s} = (s_1, ..., s_k)$ be the sequence of family sizes belong to $\mathbb{M}$ at some stage n. Then the expected number of instances of subnetwork motifs conditioned on the sequence of family sizes is*

$$\mathbb{E}\Big(|\mathbb{M}(\vec{s})|\Big) = \prod_{i=1}^{k} f(\pi_i, s_i).$$

Proof. In order to prove the theorem we will be inducting on $h = ||\vec{s}||$. Assume that $h = k$. Then $\vec{s} = (1, ..., 1)$, so the only possible instance is the original instance and $|\mathbb{M}(\vec{1})| = 1$. Since we know from the proof of Lemma 4 that $f(\pi_i, 1) = 1$ for $i = 1, ..., k$ we see that

$$\prod_{i=1}^{k} f(\pi_i, 1) = 1.$$

Thus the base case is verified.

Assume that the theorem holds true for some $h \geq k$. We will now show that it holds true for $h+1$. Let $\vec{s'} = (s_1, ..., s_{j-1}, s'_j, s_{j+1}, ..., s_k)$ be a sequence of family sizes such that $||\vec{s'}|| = h + 1$. Thus at least one duplication has occurred. Without loss of generality assume that the most recent duplication occurred in the j^{th} family. Just prior to the last duplication $\vec{s} = (s_1, ..., s_{j-1}, s_j, s_{j+1}, ..., s_k)$ is the sequence of family sizes such that $||\vec{s}|| = h$, $s_j = s'_j - 1$, and $\mathbb{E}(|\mathbb{M}(\vec{s})|) = \prod_{i=1}^{k} f(\pi_i, s_i)$ by the induction hypothesis. When the gene is duplicated from the j^{th} family each instance of $\mathbb{M}(\vec{s})$ has probability $\frac{\pi_j}{s_j}$ of giving rise to a new instance therefore the expected number of new instances is $\Big(\prod_{i=1}^{k} f(\pi_i, s_i)\Big)\Big(\frac{\pi_j}{s_j}\Big)$. Given the recursion we proved in Lemma 4 and $s'_j = s_j + 1$ the total expectation is as follows,

$$\begin{aligned}\mathbb{E}\Big(|\mathbb{M}(\vec{s'})|\Big) &= \mathbb{E}\Big(|\mathbb{M}(\vec{s})|\Big) + \mathbb{E}\Big(|\mathbb{M}(\vec{s})|\Big)\cdot\frac{\pi_j}{s_j}\\ &= \mathbb{E}\Big(|\mathbb{M}(\vec{s})|\Big)\Big(1+\frac{\pi_j}{s_j}\Big)\\ &= \Big(\prod_{i=1}^{k} f(\pi_i, s_i)\Big)\Big(1+\frac{\pi_j}{s_j}\Big)\\ &= \prod_{i=1}^{k} f(\pi_i, s_i')\end{aligned}$$

since $s_i = s_i'$ if $i \neq j$.

□

We will now evaluate what happens to the expectation of the number of instances of $\mathbb{M}(n)$ when the set is not conditioned on the vector of family sizes. For the purposes of this calculation it is useful to begin with the following Corollary.

Corollary 8. *Assume $\mathbb{M}$ is a single gene subnetwork motif. Let s be the size of the gene family and p be the probability of inheritance. Then the expected number of instances of $\mathbb{M}$ under Partial Duplication can be expressed as*

$$f(p,s) = [x^s]\Big(\frac{x}{(1-x)^{p+1}}\Big)$$

Proof. We know from Lemma 4 that

$$f(p,s) = \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)}$$

Now we can apply Lemma 1 we obtain the generating function,

$$\begin{aligned}f(p,s) &= \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)}\\ &= \frac{(p+s-1)_{s-1}}{(s-1)!}\\ &= [x^{s-1}]\Big(\frac{1}{(1-x)^{p+1}}\Big)\\ &= [x^s]\Big(\frac{x}{(1-x)^{p+1}}\Big).\end{aligned}$$

□

Theorem 4. *Suppose $\mathbb{M}$ is a subnetwork motif of size k, $1 \leq k \leq m \leq n$, $\vec{0} \leq \vec{\pi} \leq \vec{1}$ is fixed, and $\hat{\pi} = \pi_1 + ... + \pi_k$. Then the expected number of instances of $\mathbb{M}$ in Partial Duplication is*

$$\mathbb{E}\Big(|\mathbb{M}(n) \mid m, n, \vec{\pi}, k\Big) = \frac{\Gamma(\hat{\pi}+n)\Gamma(m)}{\Gamma(\hat{\pi}+m)\Gamma(n)}.$$

This result allows us to construct a significance test for subnetwork motifs using the mean of the number of instances of $\mathbb{M}$ under Partial Duplication.

Proof. In order to prove the statement of the theorem we will take a similar approach from Theorem 1 and utilize generating functions. We know that

$$\mathbb{E}\Big(\prod_{i=1}^{k} f(\pi_i, s_i)\Big) = \binom{n-1}{m-1}^{-1}\Big(\sum_{h=k}^{n-m+k}\binom{n-h-1}{m-k-1}\sum_{h}\prod_{i=1}^{k} f(\pi_i, s_i)\Big),$$

where $s_1, ..., s_k$ are the family sizes, $h = s_1 + ... + s_k$, and $\vec{s}$ is a composition of h into k parts. From Theorem 3 we know the generating function associated with $\sum_{h=k}^{n-m+k}\binom{n-h-1}{m-k-1}$ is

$$a(x) = (x + x^2 + ...)^{m-k} = \Big(\frac{x}{1-x}\Big)^{m-k}.$$

The generating function associated with $\sum_h \prod_{i=1}^{k} f(\pi_i, s_i)$ is

$$\prod_{i=1}^{k}\Big(\frac{x}{(1-x)^{\pi_i+1}}\Big). \tag{2}$$

We will now utilize the generating function to obtain the expectation by dividing the coeffecient of x^n in Expression (2) by the total number of equally likely compositions of n into m parts. Thus,

$$\begin{aligned}
\mathbb{E}\Big(\prod_{i=1}^{k} f(\pi_i, s_i)\Big) &= \Big([x^n](\frac{x^m}{(1-x)^{\hat{\pi}+m}})\Big)\binom{n-1}{m-1}^{-1} \\
&= \Big([x^{n-m}](\frac{1}{(1-x)^{\hat{\pi}+m}})\Big)\binom{n-1}{m-1}^{-1} \\
&= \frac{\binom{\hat{\pi}+n-1}{n-m}}{\binom{n-1}{m-1}} \\
&= \frac{\Gamma(\hat{\pi}+n)\Gamma(m)}{\Gamma(\hat{\pi}+m)\Gamma(n)}.
\end{aligned}$$

□

In order to derive bounds for the expected number of subnetwork motifs in Partial Duplication the following lemmas will be useful.

Lemma 5. *Suppose $z \geq 0, y \geq 1$ and $y \geq 2z^2$ then*

$$z\ln(y) - \frac{1}{y}(\frac{z}{2} + \frac{1}{12}) \leq \ln\Big(\frac{\Gamma(y+z)}{\Gamma(y)}\Big) \leq z\ln(y) + \frac{z^2}{y}$$

Proof. In order to derive bounds for the ratio we will use the fact that

$$\ln(\Gamma(x)) = (x - \frac{1}{2})\ln(x) - x + \frac{1}{2}\ln(2\pi) + \sum_{a=1}^{\infty}\frac{\mathrm{B}_{2a}}{2a(2a-1)x^{2a-1}}$$

for $x > 0$ [1], where B_{2a} are Bernoulli numbers and $B_2 = \frac{1}{6}$. The above expression can be bounded by truncating the summation at the first term neglected such that when the first term neglected is negative an upper bound is obtained and when the first term neglected is positive a lower bound is obtained. Then

$$(x-\frac{1}{2})\ln(x)-x+\frac{1}{2}\ln(2\pi)\leq \ln(\Gamma(x))\leq (x-\frac{1}{2})\ln(x)-x+\frac{1}{2}\ln(2\pi)+\frac{1}{12x}.$$

To obtain an upper bound for $\ln\left(\frac{\Gamma(y+z)}{\Gamma(y)}\right)$ we derive an upper bound for $\ln(\Gamma(y+z))$, a lower bound for $\ln(\Gamma(y))$, and then combine them. We can apply the asymptotic relationship when $x=y+z$ and when $x=y$. The upper bound calculation is as follows:

$$\begin{aligned}
\ln\Big(\frac{\Gamma(y+z)}{\Gamma(y)}\Big) &= \ln(\Gamma(y+z))-\ln(\Gamma(y))\\
&\leq (z+y-\frac{1}{2})\ln(z+y)-z-y+\frac{1}{2}\ln(2\pi)+\frac{1}{12(y+z)}-(y-\frac{1}{2})\ln(y)+y-\frac{1}{2}\ln(2\pi)\\
&= (z+y-\frac{1}{2})\ln(z+y)-z+\frac{1}{12(y+z)}-(y-\frac{1}{2})\ln(y)\\
&\leq z\ln(z+y)+(y-\frac{1}{2})\big(\ln(z+y)-\ln(y)\big)-z-\frac{1}{12(y+z)}\\
&= z\ln(z+y)+(y-\frac{1}{2})\big(\ln(1+\frac{z}{y})\big)-z-\frac{1}{12(y+z)}\\
&= z\ln(z+y)+(y-\frac{1}{2})(\frac{z}{y}-\frac{z^2}{2y}+\frac{z^3}{3y}...)-z-\frac{1}{12(y+z)}\\
&\leq z\ln(z+y)+(y-\frac{1}{2})(\frac{z}{y})-z-\frac{1}{12(y+z)}\\
&= z\ln(y(\frac{z}{y}+1))+(y-\frac{1}{2})(\frac{z}{y})-z-\frac{1}{12(y+z)}\\
&\leq z\ln(y)+\frac{z^2}{y}-\frac{z}{2y}+\frac{1}{12(y+z)}\\
&\leq z\ln(y)+\frac{z^2}{y}.
\end{aligned}$$

To derive a lower bound we derive a lower bound for $\ln(\Gamma(y+z))$, an upper bound for $\ln(\Gamma(y))$, and then combine them. Then the lower bound is as follows:

$$\begin{aligned}
\ln\Big(\frac{\Gamma(y+z)}{\Gamma(y)}\Big) &= \ln(\Gamma(y+z))-\ln(\Gamma(y))\\
&\leq (z+y-\frac{1}{2})\ln(z+y)-(y-\frac{1}{2})\ln(y)-z-\frac{1}{12y}\\
&= z\ln(y+z)+(y-\frac{1}{2})\ln(z+y)-(n-\frac{1}{2})\ln(n)-z-\frac{1}{12y}\\
&= z\ln(y+z)+(y-\frac{1}{2})\big(\ln(1+\frac{z}{y})\big)-z-\frac{1}{12y}\\
&\leq z\ln(y+z)+(y-\frac{1}{2})(\frac{z}{n}-\frac{z^2}{2y^2})-z-\frac{1}{12y}\\
&= z\ln(y)+z(\frac{z}{y}-\frac{z^2}{2y^2})+(y-\frac{1}{2})(\frac{z}{y}-\frac{z^2}{2y^2})-z-\frac{1}{12y}\\
&= z\ln(y)+\frac{1}{n}\big(z^2-\frac{z^3}{2n}-\frac{z^2}{2}-\frac{z}{2}+\frac{z^2}{4}-\frac{1}{12}\big)\\
&\geq z\ln(y)-\frac{1}{y}\big(\frac{3z^2}{4}+\frac{z^3}{2n}+\frac{z}{2}+\frac{1}{12}\big)\\
&\geq z\ln(y)-\frac{1}{y}\big(\frac{z}{2}+\frac{1}{12}\big).
\end{aligned}$$

□

Lemma 6. *Suppose* $z \geq 0, y \geq 1,$ *and* $y \geq 2z^2$ *then*

$$y^z(1 - \frac{1}{y}(\frac{z}{2} + \frac{1}{12})) \leq \frac{\Gamma(y+z)}{\Gamma(y)} \leq y^z(1 + \frac{3z^2}{y}).$$

Proof. In order to obtain the result of the Lemma we begin by exponentiating the result of Lemma 5 to obtain

$$y^z e^{-(1/y)(z/2+1/12)} \leq \frac{\Gamma(y+z)}{\Gamma(y)} \leq y^z e^{z^2/y}.$$

By applying Lemma 12 (occurs in Section 6) with $\zeta = \frac{z^2}{y}$ to the upper bound and Lemma 13 with $\zeta = \frac{1}{y}(\frac{z}{2} + \frac{1}{12})$ to the lower bound we obtain the result of the Lemma. Note that the conditions for Lemma 12 and 13 (both occur in Section 6) follow directly from the hypothesis. □

Corollary 9. *Let* $k \geq 1$ *be fixed,* $n \geq m \geq 1, n \geq 2k^2,$ *and* $0 \leq || \vec{\pi} || \leq k$. *Then the expected number of instances of* $\mathbb{M}(n)$ *in Partial Duplication is*

$$\frac{n^{||\vec{\pi}||}\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)}\Big(1 - \frac{1}{n}(\frac{k}{2} + \frac{1}{12})\Big) \leq \frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(n)\Gamma(||\vec{\pi}|| + m)} \leq \frac{n^{||\vec{\pi}||}\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)}\Big(1 + \frac{3k^2}{2n}\Big).$$

Proof. We know from Theorem 4 that $\mathbb{E}\Big(|\mathbb{M}(n)|\ m, n, \vec{\pi}, k\Big)$ is

$$\frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)\Gamma(n)}.$$

We derive the indicated bounds for the expectation by applying Lemma 6 with $y = n$ and $z = ||\vec{\pi}||$.

Note that we obtain the following more informative bounds by applying Lemma 6 to $\mathbb{E}\Big(|\mathbb{M}(n)|\ m, n, \vec{\pi}, k\Big)$ when $0 \leq ||\vec{\pi}|| < \frac{1}{2}$

$$\frac{n^{||\vec{\pi}||}\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)}\Big(1 - \frac{31(||\vec{\pi}||)^2}{32n}\Big) \leq \frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(n)\Gamma(||\vec{\pi}|| + m)} \leq \frac{n^{||\vec{\pi}||}\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)}\Big(1 - \frac{(||\vec{\pi}||^2}{n} + \frac{||\vec{\pi}||^4}{2n^2}\Big).$$

□

Lemma 7. *Suppose* $z \geq 0, y \geq 1,$ *and* $y \geq 2z^2$ *then*

$$-z\ln(y) + \frac{1}{y}(\frac{z}{2} + \frac{1}{12}) \leq \ln\big(\frac{\Gamma(y)}{\Gamma(y+z)}\big) \leq -z\ln(y) - \frac{z^2}{y}.$$

Proof. The result of the Lemma follows directly from multiplying the result of Lemma 5 by -1. □

Lemma 8. *Suppose* $z \geq 0, y \geq 1,$ *and* $y \geq 2z^2$ *then*

$$y^{-z}(1 + \frac{1}{y}(\frac{z}{2} + \frac{1}{12})) \leq \frac{\Gamma(y)}{\Gamma(y+z)} \leq y^{-z}(1 - \frac{z^2}{y} + \frac{z^4}{2y^2}).$$

Proof. To obtain the result of the Lemma we begin by exponentiating the result of Lemma 7 to obtain

$$y^{-z}e^{(1/y)(z/2+1/12)} \leq \frac{\Gamma(y)}{\Gamma(y+z)} \leq y^{-z}e^{-z^2/y}.$$

We obtain the result of the Lemma by applying Lemma 13 (occurs in Section 6) with $\zeta = \frac{z^2}{y}$ to the upper bound and Lemma 12 (occurs in Section 6) with $\zeta = \frac{1}{y}(\frac{z}{2} + \frac{1}{12})$ to the lower bound. □

Corollary 10. *Let $k \geq 1$ be fixed, $n \geq m \geq 1, n \geq 2k^2$, $\frac{m}{n^2} \to \infty$, and $0 \leq ||\vec{\pi}|| \leq k$. Then the expected number of instances of $\mathbb{M}(n)$ in Partial Duplication is*

$$\frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(n)\Gamma(||\vec{\pi}|| + m)} \geq \Big(\frac{n}{m}\Big)^{||\vec{\pi}||}\Big(1 - \frac{1}{n}(\frac{k}{2} + \frac{1}{12})\Big)\Big(1 + \frac{1}{m}(\frac{||\vec{\pi}||^2}{2} + \frac{1}{12})\Big)$$

$$\frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(n)\Gamma(||\vec{\pi}|| + m)} \leq \Big(\frac{n}{m}\Big)^{||\vec{\pi}||}\Big(1 + \frac{3k^2}{2n}\Big)\Big(1 - \frac{||\vec{\pi}||^2}{m} + \frac{||\vec{\pi}||^4}{2m^2}\Big).$$

Proof. The indicated bounds follow directly from applying Lemma 8 with $y = m$ and $z = ||\vec{\pi}||$ to the result of Corollary 9. □

4.2 SECOND MOMENTS, $k = 1$

Now we turn our attention to studying the second moments of the number of instances of a given subnetwork motif. We will begin the study by considering the case of single gene subnetwork motifs. In order to calculate the variance we will need to analyze the expected number of ordered pairs of single gene subnetwork motifs.

Let $\mathbb{M}$ be an arbitrary single gene subnetwork motif with inheritance probability $p = \pi_1$ and family size s. For the purposes of this subsection, we will be conditioning on s so we let $\mathbb{M}(s)$ be the set of instances of $\mathbb{M}$ when the family size is s. The members of the of the subnetwork motif's gene family are indexed in order of duplication and denoted $\{b_1, ..., b_s\}$. Here, (b_1) is the original instance of $\mathbb{M}$ and $\{(b_2), ..., (b_s)\}$ are the potential members of $\mathbb{M}(s)$.

The expected size of $\mathbb{M}(s)$ is $f(p, s)$ by Lemma 10. Let $g(p, s)$ be the expected size of $\mathbb{M}(s)^2$. We now prove a recurrence satisfied by $g(p, s)$.

Corollary 11. *We have $g(p, 1) = 1$ and for all $s \geq 1$*

$$g(p, s + 1) = g(p, s) + \frac{2p}{s} \cdot g(p, s) + \frac{p}{s} \cdot f(p, s).$$

Proof. When the family size is $s = 1$ no duplications have occurred. At that point the only pair of subnetwork motif instances is the original instance of $\mathbb{M}$ and itself, so $g(p, 1) = 1$.

We will now show that the recurrence relationship is true for any $s \geq 1$. Any member $< (b_i), (b_j) > \in \mathbb{M}(s + 1)^2$ belongs to one of three categories: the pair falls into the first category if $1 \leq i, j \leq s$, the second category if $1 \leq i \leq s$ and $j = s + 1$ or $i = s + 1$ and $1 \leq j \leq s$, and the third category if $i = j = s + 1$. In order to calculate $g(p, s + 1)$ we will find the expected number of ordered pairs of each category.

In the first category the pair consists of two instances of $\mathbb{M}$ when the family size is s. Thus the expected number of such pairs is $g(p, s)$ by definition.

In the second category the pair consists of an old subnetwork motif instance and a new subnetwork motif instance. A pair from the second category can only be generated from a pair from the first category by duplicating one the of elements of the pair. Suppose $< (b_i), (b_j) > \in \mathbb{M}(s)^2$. To generate a pair in the second category b_i can be duplicated and the subnetwork motif must be inherited in duplication. There is a $\frac{1}{s}$ chance that b_i is selected for duplication to generate b_{s+1} and probability p the new subnetwork motif instance (b_{s+1}) is inherited in duplication. We sum over all the pairs in $\mathbb{M}(s)^2$ to calculate the expected number of category 2 pairs in $\mathbb{M}(s + 1)^2$. Since the expected number of pairs in $\mathbb{M}(s)^2$ is $g(p, s)$ the expected number of pairs in category 2 where the first element is the new instance is $\frac{p}{s} \cdot g(p, s)$. Note that the expected number of pairs in category 2 where the second element is the new instance is also $\frac{p}{s} \cdot g(p, s)$. Since the events are disjoint, $\frac{2p}{s} \cdot g(p, s)$ is the expected number of category 2 pairs in $\mathbb{M}(s + 1)^2$.

In the third category the only possible pair is the new instance and itself. The new instance of $\mathbb{M}$ must be generated from a member of $\mathbb{M}(s)$. Suppose $(b_i) \in \mathbb{M}(s)$. Then there is a $\frac{1}{s}$ chance that b_i is selected for duplication to generate the new gene b_{s+1} and the new gene has probability p of becoming a new instance (b_{s+1}). To find the expected number of category 3 pairs we sum over all the members in $\mathbb{M}(s)$. By Lemma 4 the expected number of members in $\mathbb{M}(s)$ is $f(p, s)$ so the expected number of category 3 pairs in $\mathbb{M}(s + 1)^2$ is $\frac{p}{s} \cdot f(p, s)$.

The corollary follows from the fact that the sum of expectations is the expectation of the sum. □

Corollary 12. *For $s \geq 1$ we have*

$$g(p,s) = \big(\frac{2\Gamma(s+2p)}{\Gamma(s)\Gamma(2p+1)} - \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)}\big).$$

Proof. Let $h(p,s) = g(p,s) + f(p,s)$. Then $h(p,1) = 2$. For $s \geq 1$ we have the following:

$$\begin{aligned} h(p,s+1) &= g(p,s+1) + f(p,s+1) \\ &= g(p,s) + \frac{2p}{s}g(p,s) + \frac{p}{s}f(p,s) + f(p,s)(1+\frac{p}{s}) \\ &= h(p,s) + \frac{2p}{s}g(p,s) + \frac{2p}{s}f(p,s) \\ &= h(p,s) + \frac{2p}{s}h(p,s) \\ &= h(p,s)\big(1+\frac{2p}{s}\big). \end{aligned}$$

Therefore by induction on s

$$h(p,s) = 2\prod_{i=1}^{s-1}\big(1+\frac{2p}{i}\big) = \frac{2\Gamma(s+2p)}{\Gamma(s)\Gamma(2p+1)} \text{ for all } s \geq 1.$$

Thus

$$g(p,s) = \big(\frac{2\Gamma(s+2p)}{\Gamma(s)\Gamma(2p+1)} - \frac{\Gamma(p+s)}{\Gamma(s)\Gamma(p+1)}\big)$$

since we know $f(p,s)$ from Lemma 4 and $g(p,s) = h(p,s) - f(p,s)$. □

Corollary 13. *For $s \geq 1$ we have*

$$g(p,s) = [x^s]\Big(\frac{2x}{(1-x)^{2p+1}} - \frac{x}{(1-x)^{p+1}}\Big).$$

Proof. We know from the proof of Corollary 12 that $h(p,s) = \frac{2\Gamma(s+2p)}{\Gamma(s)\Gamma(2p+1)}$. Thus

$$\begin{aligned} h(p,s) &= \frac{2(s+2p-1)_{s-1}}{(s-1)} \\ &= \binom{2(s+2p-1)}{s-1} \\ &= [x^{s-1}]\big(\frac{2}{(1-x)^{2p+1}}\big) \\ &= [x^s]\big(\frac{2x}{(1-x)^{2p+1}}\big), \end{aligned}$$

using Lemma 1. Then

$$g(p,s) = [x^s]\Big(\frac{2}{(1-x)^{2p+s}}\Big) - [x^s]\Big(\frac{x}{(1-x)^{p+1}}\Big)$$

from Lemma 4 and the fact that $g(p,s) = h(p,s) - f(p,s)$. □

Notice that the value of $g(p,s)$ depends only on the values of p and s. That is to say the expected number of pairs of subnetwork motifs of size $k = 1$ depends only on p and s. In the next section we will show by example that when $k \geq 2$ the size of the gene families and corresponding inheritance probabilities do not necessarily determine the expected number of pairs of subnetwork motifs.

4.3 MAXIMIZING THE SECOND MOMENT

We will now study the second moment when $k \geq 2$. In order to do so we must specify how subnetwork motifs are inherited when they share a common gene which is duplicated. Assume as in previous sections that for notational convenience the families that belong to $\mathbb{M}$ are indexed from 1 to k. The following inheritance modes are two of the possible refinements of the basic Partial Duplication mode.

Previously we aggregated our random duplication process over the k family sizes that contribute to $\mathbb{M}$. Now we need a finer analysis and we will aggregate according to the sequence of duplications. It will be useful to select an arbitrary sequence of duplications for the purposes of computation. At the start of the duplication process there are m genes to choose from. After the first duplication there are $m+1$ genes to choose from, and the options for genes increment after every duplication. Given $n-m$ duplications there are $(m)(m+1)(m+2)\cdot ... \cdot (m+n-1) = (n-1)_{n-m}$ possible sequence of duplications, and they are all equally likely since the duplication at each stage is chosen at random.

The genes are indexed as follows. For $j = 1, ..., m$, u_j is the original gene in the j^{th} family. For $m < j \leq n$, u_j is the gene that results from the $(j-m)^{th}$ duplication. For $m \leq j \leq n$ let $U_j = \{u_\ell \mid 1 \leq \ell \leq j\}$. Then U_j is the set of genes which exist after the $(j-m)^{th}$ duplication. For $1 \leq i \leq k$ let S_i be the set of genes in U_n which belong to the i^{th} family which contributes to $\mathbb{M}$. Then $\mathbb{M}(n) \subseteq S_1 \times ... \times S_k$. That is to say that the set of potential subnetwork motifs is a subset of the product of the S_i's.

Suppose $\mathbb{M}$ is a subnetwork motif of arbitrary size $k \geq 2$, $\mathscr{I}$ and $\mathscr{J}$ are instances of $\mathbb{M}(n)$ that share a particular gene b from the i^{th} family for $1 \leq i \leq k$, and b is duplicated to generate b'. In the first refinement when b' is generated the inheritance events $\mathscr{I} \in \mathbb{M}(n)$ and $\mathscr{J} \in \mathbb{M}(n)$ are fully correlated, so that π_i is the probability that both inheritance events occur. In the second refinement when when b' is generated the probability that the inheritance events $\mathscr{I}' \in \mathbb{M}(n)$ and $\mathscr{J}' \in \mathbb{M}(n)$ occur are mutually independent, so that π_i^2 is the probability that both inheritance events occur.

It can be seen in Figure 2 that given a particular sequence of duplications, these two refinements give different results for the second moment. Take the simple case of $\vec{s} = (2,2)$ and $\vec{\pi} = (\frac{1}{2}, \frac{1}{2})$. The second moment in the first refinement is $6\frac{1}{4}$ and the second moment in the second refinement is $5\frac{3}{4}$.

We choose to focus on the second refinement, defined below as *Binary Inheritance*, because it is tractable and gives us the maximum second moment value of any refinement of the Partial Duplication mode. The latter will be proved in Theorem 5.

Definition 1. *Binary inheritance is a refinement of the Partial Duplication mode which acts as follows. Suppose $\mathbb{M}$ is a subnetwork motif of size $k, 1 \leq k \leq m \leq n$, $\mathscr{I}$ and $\mathscr{J}$ are instances of $\mathbb{M}$ that share a common gene b in the i^{th} family for $1 \leq i \leq k$, and $\vec{0} \leq \vec{\pi} \leq \vec{1}$ is fixed. For any arbitrary duplication step, if b is selected for duplication to generate b' then with probability π_i the inheritance events $\mathscr{I}' \in \mathbb{M}(n)$ and $\mathscr{J}' \in \mathbb{M}(n)$ both occur, and with probability $1-\pi_i$ neither of the inheritance events occurs.*

Note that if the common gene b is contained in r instances then all r possible inheritance events occur or none of them occur. Also if $i > k$ there are no possible inheritance events that can occur at that step.

Theorem 5. *Suppose $\mathbb{M}$ is a subnetwork motif of size $k, 1 \leq k \leq m \leq n$, and $\vec{0} \leq \vec{\pi} \leq \vec{1}$ is fixed. Then Binary Inheritance gives the maximum value for $\mathbb{E}\Big(|\mathbb{M}(n)^2|; m, n, \vec{\pi}, k\Big)$ over all refinements of Partial Duplication.*

This result allows us to obtain the variance of the number of instances of $\mathbb{M}$ under Partial Duplication using the Binary Inheritance mode.

Proof. Let Challenge Inheritance be an arbitrary refinement of Partial Duplication which we denote by $\mathbb{C}$. Then let $\mathbb{E}(|\mathbb{M}(n)^2| \ ; \ \mathbb{C})$ be the expected number of pairs of subnetwork motif instances under Challenge Inheritance for the given $m, n, \vec{\pi},$ and k. Let Binary Inheritance be denoted by $\mathbb{B}$ and let $\mathbb{E}(|\mathbb{M}(n)^2| \ ; \ \mathbb{B})$ be the expected number of pairs of subnetwork motif instances under Binary Inheritance. We will show that

$$\mathbb{E}(|\mathbb{M}(n)^2| \ ; \ \mathbb{C}, m, n, \vec{\pi}_i, k) \leq \mathbb{E}(|\mathbb{M}(n)^2| \ ; \ \mathbb{B}, m, n, \vec{\pi}_i, k). \tag{3}$$

In order to prove the statement of the theorem we fix on one sequence of duplications and calculate the expected number of ordered pairs in $\mathbb{M}(n)$ for refinements $\mathbb{C}$ and $\mathbb{B}$. We are going to consider the

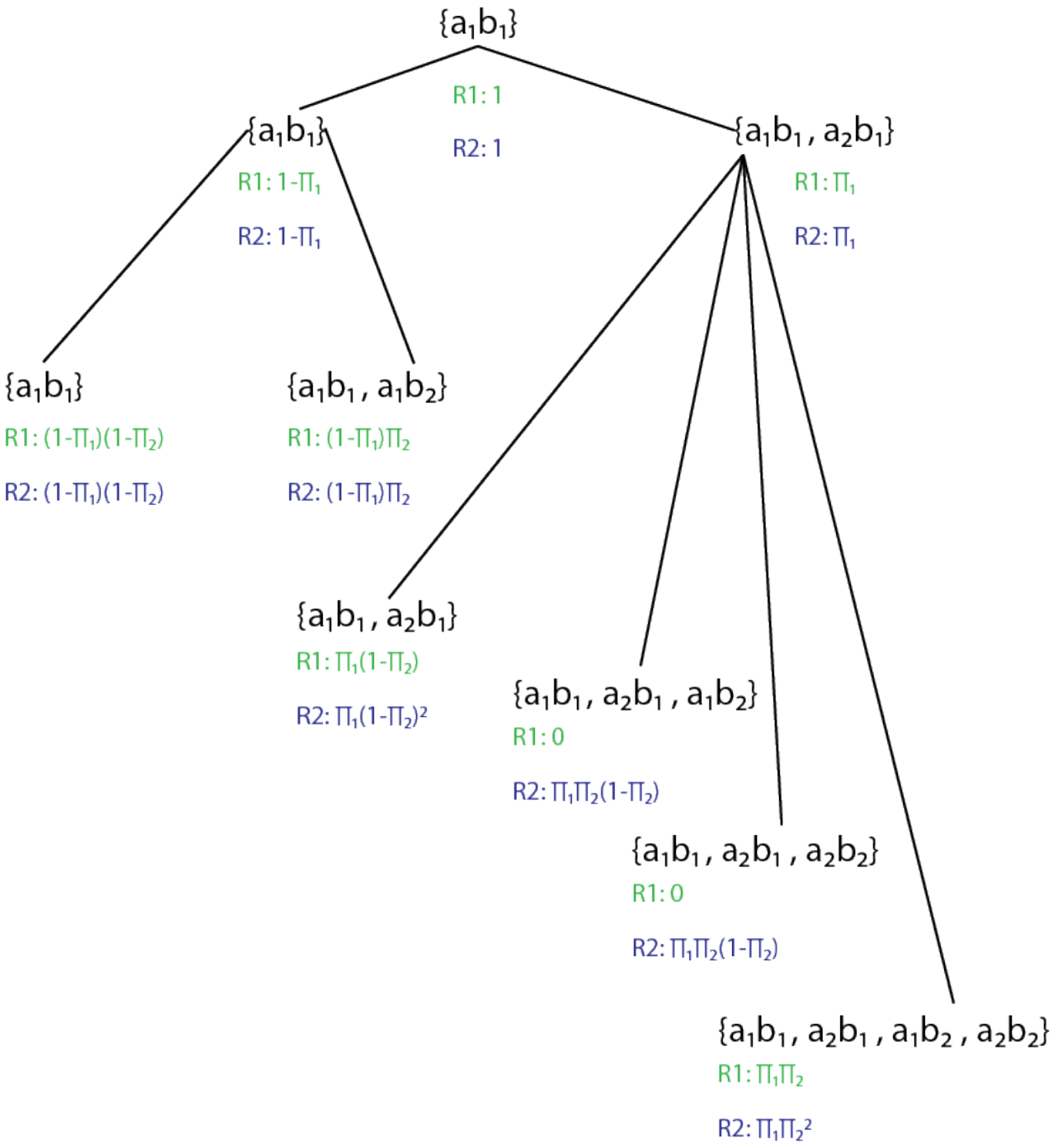

Figure 2: The duplication process begins at stage 0 with the original subnetwork motif $\{a_1, b_1\}$. If the sequence of genes after the duplication process is $\{a_1, b_1, a_2, b_2\}$ then all possible subnetwork motif events are depicted with probabilities for each. The probabilities in green (on top) are from refinement 1 and the probabilities in blue (on the bottom) are from refinement 2.

probability that pairs of potential instances are contained in $\mathbb{M}(n)$ conditioned on the chosen sequence of $n-m$ duplications.

We choose an arbitrary ordered pair of potential subnetwork motif instances $(\mathscr{I},\mathscr{H})$, *i.e.,* $\mathscr{I},\mathscr{H} \in S_1 \times ... \times S_k$. Suppose $\mathscr{I} \in \mathbb{M}(n)$. For $m \leq j \leq n$ there is a predecessor to $\mathscr{I}$ which we call $\mathscr{I}_j$, which must belong to $\mathbb{M}(j)$ in order for $\mathscr{I}$ to be in $\mathbb{M}(n)$. We will define this sequence by inducting on $j=n$ down to $j=m$.

To begin suppose $\mathscr{I}_n = \mathscr{I}$ then $\mathscr{I}_n \in \mathbb{M}(n)$. For $m \leq j < n$ let $\hat{u}$ be the gene from which u_{j+1} is duplicated and assume that $\mathscr{I}_{j+1} \in \mathbb{M}(j+1)$. In the case that u_{j+1} does not appear in $\mathscr{I}_{j+1}$, let $\mathscr{I}_j = \mathscr{I}_{j+1}$; thus it can be seen that $\mathscr{I}_{j+1} \in \mathbb{M}(j)$. In the case that u_{j+1} does appear in $\mathscr{I}_{j+1}$, let $\mathscr{I}_j$ be an instance that is obtained by replacing u_{j+1} with $\hat{u}$ and $\mathscr{I}_j \in \mathbb{M}(j)$.

From the inductive definition of the sequence of subnetwork motif instances it can be seen that the members of $\mathscr{I}_j$ are contained in U_j for $m \leq j \leq n$. In particular the members of $\mathscr{I}_m$ must be contained in U_m; thus it consists only of original genes and must be the original instance of $\mathbb{M}$.

It is easy to see by induction on $j=n$ to $j=m$ that the definition and results for $(\mathscr{I}_m,...,\mathscr{I}_n)$ apply equally to $(\mathscr{H}_m,...,\mathscr{H}_n)$. We can now express the probability that $\mathscr{I}$ and $\mathscr{H}$ are both in $\mathbb{M}(n)$ as

$$Pr\big(\mathscr{I}_m,\mathscr{H}_m \in \mathbb{M}(m)\big) \prod_{j=m}^{n-1} Pr\big(\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1) \mid \mathscr{I}_j,\mathscr{H}_j \in \mathbb{M}(j)\big). \tag{4}$$

Since $\mathscr{I}_m$ and $\mathscr{H}_m$ are the original instance of $\mathbb{M}$ the probability that $\mathscr{I}_m$ and $\mathscr{H}_m$ are contained in $\mathbb{M}(n)$ is 1. For any $m \leq j \leq n-1$, we are interested in the conditional probability that $\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1)$ given $\mathscr{I}_j,\mathscr{H}_j \in \mathbb{M}(j)$. So assume that $\mathscr{I}_j$ and $\mathscr{H}_j$ are contained in $\mathbb{M}(j)$. At stage j, let $\hat{u}$ be selected for duplication to produce the new gene u_{j+1} in the i^{th} family. If u_{j+1} does not appear in $\mathscr{I}_{j+1}$ nor $\mathscr{H}_{j+1}$ then the probability that both $\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1)$ is 1. If u_{j+1} is in only in $\mathscr{I}_{j+1}$ or $\mathscr{H}_{j+1}$ then the probability that both $\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1)$ is π_i. Note that if a duplication occurs outside of the first k families then $\mathbb{M}(j+1) = \mathbb{M}(j)$ and the probability is 1. If u_{j+1} appears in both $\mathscr{I}_{j+1}$ and $\mathscr{H}_{j+1}$ then the probability that both $\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1)$ under $\mathbb{C}$ is between 0 and π_i, whereas under $\mathbb{B}$ the probability is exactly π_i. Note that the only time the probability that $\mathscr{I}_{j+1},\mathscr{H}_{j+1} \in \mathbb{M}(j+1)$ can differ between $\mathbb{C}$ and $\mathbb{B}$ is in this case. It follows that

$$Pr\big(\mathscr{I},\mathscr{H} \in \mathbb{M}(n)\ ;\ \mathbb{C}\big) \leq Pr\big(\mathscr{I},\mathscr{H} \in \mathbb{M}(n)\ ;\ \mathbb{B}\big). \tag{5}$$

Summing Equation 5 over $\mathscr{I},\mathscr{H} \in S_1 \times ... \times S_k$ we obtain

$$\sum_{\mathscr{I},\mathscr{H}} Pr\big(\mathscr{I},\mathscr{H} \in \mathbb{M}(n)\ ;\ \mathbb{C}\big) \leq \sum_{\mathscr{I},\mathscr{H}} Pr\big(\mathscr{I},\mathscr{H} \in \mathbb{M}(n)\ ;\ \mathbb{B}\big)$$

so that

$$\mathbb{E}(|\mathbb{M}(n)^2|\ ;\ \mathbb{C}) \leq \mathbb{E}(|\mathbb{M}(n)^2|\ ;\ \mathbb{B}).$$

Recall that there are $(n-1)_{n-m}$ possible sequences of duplications. The results above are conditioned on the arbitrarily chosen sequence of duplications. In order to extend the results to the overall second moment we average overall all $(n-1)_{n-m}$ sequences to obtain Equation (3), which is the desired result.

From now on we will confine our study of Partial Duplication to the Binary Inheritance refinement. □

4.4 SECOND MOMENTS FOR BINARY INHERITANCE

We will now study second moments for Binary Inheritance. For the purposes of this section we will adopt the gene indexing and terminology from the introduction to Theorem 5. Thus we will be aggregating our gene duplication process over the sequence of duplications rather than the family sizes that contribute to $\mathbb{M}$. In addition, we will consider properties of particular realizations of the duplication and inheritance process

that includes not only the selection of what to duplicate but also which particular instances are inherited under Binary Inheritance.

Suppose $\mathbb{M}$ is a subnetwork motif of size $1 \leq k \leq m \leq n$ and $\vec{\pi} = (\pi_1, ..., \pi_k)$ is the vector of inheritance probabilities. As before, $\vec{S} = (S_1, ..., S_k)$ is the sequence of the sets of genes that belong to $\mathbb{M}$ where $S_i \subseteq U_n$ consist of the genes in the i^{th} family. We define a gene $\hat{u} \in U_n$ to be *viable* if and only if $\hat{u}$ appears in a member of $\mathbb{M}(n)$. If $\hat{u} = u_j$ for $1 \leq j \leq k$, then $\hat{u}$ is viable since it appears in the original instance of $\mathbb{M}$. Alternatively, if $\hat{u} = u_j$ for $m \leq j \leq n$ then $\hat{u}$ will be viable if and only if u_j appears in a member of $\mathbb{M}(j)$. When $j \geq m$ a new instance of $\mathbb{M}$ can only be inherited from an instance of $\mathbb{M}$ and a viable gene can only be inherited from a viable gene. Note that if a gene is viable at the stage when it appears then it remains viable throughout the duplication process.

Lemma 9. *A k-tuple of $S_1 \times ... \times S_k$ belongs to $\mathbb{M}(n)$ if and only if every gene in it is viable.*

Proof. Every instance in $\mathbb{M}(n)$ consists of viable genes by the definition of viability.

It remains to prove that if we have a k-tuple of viable genes $\mathscr{L} \in S_1 \times ... \times S_k$ then it forms an instance of $\mathbb{M}$. Let j be the smallest value such that $j \geq m$ and all the members of $\mathscr{L}$ are in U_j. The proof proceeds by induction on j.

Suppose $j = m$. Then there have been no duplications. Thus the only possibility for $\mathscr{L}$ is the original instance of $\mathbb{M}$. Suppose $j > m$ and let $\mathscr{L} = (a_1, ..., a_{i-1}, a_i, a_{i+1}, ..., a_k)$. By the minimality of j we know that $u_j = a_i$ for some $i \in \mathbb{K}$ where $\mathbb{K} = \{1, ..., k\}$. In order for a_i to be viable it must be inherited from a viable gene which we denote $\hat{a}_i$. Since $\hat{a}_i$ is viable then $\mathscr{I} = (a_1, ..., a_{i-1}, \hat{a_i}, a_{i+1}, ..., a_k)$ must be an instance of $\mathbb{M}$ and every member of $\mathscr{I}$ must belong to U_{j-1}.

By the induction hypothesis $\mathscr{I}$ is an instance of $\mathbb{M}$. A viable gene can only be inherited from a viable gene, therefore a_i must appear in some member of $\mathbb{M}$. Since a_i appears in an instance of $\mathbb{M}$ it must have been inherited from $\hat{a}_i$ which appears in another instance of $\mathbb{M}$. Under the Binary Inheritance Mode, when $\hat{a}_i$ is selected for duplication to generate a_i, either every instance that contains $\hat{a}_i$ results in a new instance or there are no new instances. Therefore, it follows that $\mathscr{L}$ must be inherited from $\mathscr{I}$. □

Corollary 14. *Suppose $\mathbb{M}$ is a subnetwork motif of size $k \geq 1$ under the Binary Inheritance Mode, $(S_1, ..., S_k)$ is the sequence of the gene families belonging to $\mathbb{M}$, $s_i = |S_i|$ for $i \in \mathbb{K}$, and $\vec{\pi} = (\pi_1, ..., \pi_k)$ is the vector of inheritance probabilities. Then in family S_i $f(\pi_i, s_i)$ is the expected number of viable genes and $g(\pi_i, s_i)$ is the expected number of ordered pairs of viable genes.*

Proof. Within S_i a viable gene can only be inherited from a viable gene and has chance π_i of inheriting viability. Therefore, viability is inherited by the same probabilistic process as single gene subnetwork motifs and we can apply Lemma 4 and Corollary 12 to obtain the expected number of viable genes and the second moment for viable genes respectively. □

Corollary 15. *Assume the same hypothesis as Corollary 14. Then the mean number of ordered pairs of instances of $\mathbb{M}$ conditioned on the vector of family sizes is*

$$\mathbb{E}\big(|\mathbb{M}(n)|^2 \ ; \ s_1, ..., s_k\big) = \prod_{i=1}^{k} g(\pi_i, s_i).$$

Proof. Let V_i be the set of viable genes in S_i for $1 \leq i \leq k$. By Lemma 9 we know that $\mathbb{M}(n) = V_1 \times ... \times V_k$. Then $\mathbb{E}\big(|\mathbb{M}(n)|^2 \ ; \ s_1, ..., s_k\big) = \mathbb{E}\big(\prod_{i=1}^{k} |V_i^2|\big)$. We know by Corollary 14 that the expectation of $|V_i^2|$ is $g(\pi_i, s_i)$ for $i = 1, ..., k$. The expected number of ordered pairs of $\mathbb{M}$ is

$$\mathbb{E}\big(|\mathbb{M}(n)|^2 \ ; \ s_1, ..., s_k\big) = \mathbb{E}\big(\prod_{i=1}^{k} |V_i|^2\big) = \prod_{i=1}^{k} \mathbb{E}\big(|V_i|^2\big) = \prod_{i=1}^{k} g(\pi_i, s_i),$$

since the probability of viability for any gene is independent from the probability of viability for any gene in a different family. □

Theorem 6. *Suppose $\mathbb{M}$ is a subnetwork motif of size k, $1 \leq k \leq m \leq n$ and $\vec{0} \leq \vec{\pi} \leq \vec{1}$ is fixed. Then the second moment of the number of instances of $\mathbb{M}$ in binary Partial Duplication is*

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) = \frac{\Gamma(m)}{\Gamma(n)} \sum_{\mathbb{A}\subseteq\mathbb{K}} \frac{(-1)^{|\mathbb{A}|} 2^k \Gamma(n + ||\vec{\pi}|| + \sum_{i\notin\mathbb{A}} \pi_i)}{2^{|\mathbb{A}|}\Gamma(m + ||\vec{\pi}|| + \sum_{i\notin\mathbb{A}} \pi_i)}.$$

This result allows us to calculate the variance of the number of instances of $\mathbb{M}$ under Partial Duplication using Binary Inheritance, which would be required for a significance test.

Proof. Our approach to prove the statement of the Theorem is similar to our approach to proving Theorem 1 by utilizing generating functions to calculate the value of $\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big)$. Corollary 14 gives the following expression for the second moment

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) = \binom{n-1}{m-1}^{-1} \Big(\sum_{h=k}^{n-m+k} \binom{n-h-1}{m-k-1} \sum_{||\vec{s}||} \prod_{i=1}^{k} g(\pi_i, s_i)\Big),$$

where the inner summation is over $||\vec{s}|| = h$. From the proof of Lemma 3 we know that

$$[x^h] \sum_{h=k}^{n-m+k} \binom{n-h-1}{m-k-1} = (x + x^2 + ...)^{m-k} = (\frac{x}{(1-x)})^{m-k}.$$

From Corollaries 13 and 15 we obtain

$$\begin{aligned}\sum_{h\geq 1} x^h \sum \prod_{i=1}^{k} g(\pi_i, s_i) &= \prod_{i=1}^{k} \Big(\frac{2x}{(1-x)^{2\pi_i+1}} - \frac{x}{(1-x)^{\pi_i+1}} \Big) \\ &= \sum_{\mathbb{A}\subseteq\{\mathbb{K}\}} \Big(\prod_{i\in\mathbb{K}-\mathbb{A}} \frac{2x}{(1-x)^{2\pi_i+1}} \Big) \Big(\prod_{j\in\mathbb{A}} -\frac{x}{(1-x)^{\pi_j+1}} \Big)\end{aligned}$$

the second sum is over $||\vec{s}|| = h$ and $\mathbb{K} = \{1, ..., k\}$.

As in Theorem 1 we can apply the generating function to obtain the expectation by dividing the coefficient of x^n in the product of the generating functions by the total number of equally likely compositions of n into m parts. That is,
$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) \times \binom{n-1}{m-1}^{-1}$ is

$$[x^n]\Big((\frac{x}{(1-x)})^{m-k} \sum_{Q\subseteq\mathbb{K}} \Big(\prod_{i\in\mathbb{K}-\mathbb{A}} \frac{2x}{(1-x)^{2\pi_i+1}} \Big) \Big(\prod_{j\in\mathbb{A}} -\frac{x}{(1-x)^{\pi_j+1}} \Big) \Big)$$

$$= [x^n] \sum_{\mathbb{A}\subseteq\mathbb{K}} \frac{x^{m-k}(2x)^{|\mathbb{K}-\mathbb{A}|}(-x)^{|\mathbb{A}|}}{(1-x)^{m-k}(1-x)^{2(\sum_{i\notin\mathbb{A}}\pi_i)+|\mathbb{K}-\mathbb{A}|}(1-x)^{\sum_{j\in\mathbb{A}}\pi_j+|\mathbb{A}|}}$$

$$= [x^{n-m}] \sum_{\mathbb{A}\subseteq\mathbb{K}} \frac{(-1)^{\mathbb{A}} 2^k}{(1-x)^{m+2(\sum_{i\notin\mathbb{A}}\pi_i)+\sum_{j\in\mathbb{A}}\pi_j}}$$

$$= \sum_{\mathbb{A}\subseteq\mathbb{K}} \frac{(-1)^{|\mathbb{A}|} 2^k \Gamma(n + ||\vec{\pi}|| + \sum_{i\notin\mathbb{A}} \pi_i)}{2^{|\mathbb{A}|}\Gamma(m + ||\vec{\pi}|| + \sum_{i\notin\mathbb{A}} \pi_i)}.$$

We obtain the result of the theorem by solving for the expectation and simplifying. □

Corollary 16. *Suppose $\mathbb{M}$ is a subnetwork motif of size k, $1 \leq k \leq m \leq n$ and $\vec{0} \leq \vec{\pi} \leq \vec{1}$ is fixed. Then the variance of the number of instances of $\mathbb{M}$ in binary Partial Duplication is $\mathbb{VAR}\Big(|\mathbb{M}(n)|; m, n, \vec{\pi}, k\Big)$ which evaluates to*

$$\frac{\Gamma(m)}{\Gamma(n)} \sum_{Q \subseteq \mathbb{K}} \frac{(-1)^{k-|Q|} 2^{|Q|} \Gamma(n + ||\vec{\pi}|| + \sum_{i \in Q} \pi_i)}{\Gamma(m + ||\vec{\pi}|| + \sum_{i \in Q} \pi_i)} - \Big(\frac{\Gamma(||\vec{\pi}|| + n)\Gamma(m)}{\Gamma(||\vec{\pi}|| + m)\Gamma(n)}\Big)^2.$$

This result allows us to construct a significance test for $\mathbb{M}$ under Partial Duplication using Binary Inheritance.

Proof. The Corollary follows immediately from Theorems 4 and 6. □

Lemma 10. *Suppose $a_i > b_i \geq 0$ for $1 \leq i \leq k$, and for $0 \leq j \leq k$ let*

$$\mathbb{S}_j = \sum_{\mathbb{A} \subseteq \mathbb{K}} \Big(\prod_{i \in \mathbb{K} - \mathbb{A}} a_i \prod_{i \in \mathbb{A}} b_i \Big)$$

where the summation is over $|\mathbb{A}_i| = j$. Then

$$\mathbb{S}_0 - \mathbb{S}_1 + \mathbb{S}_2 - ... - \mathbb{S}_{2\ell+1} \leq \prod_{i \in \mathbb{K}} (a_i - b_i) \leq \mathbb{S}_0 - \mathbb{S}_1 + ... + \mathbb{S}_{2\ell},$$

for $0 \leq \ell \leq \frac{k}{2}$.

Proof. Since

$$\mathbb{S}_0 = \prod_{i \in \mathbb{K}} a_i > 0$$

the following inequalities,

$$1 - \frac{\mathbb{S}_1}{\mathbb{S}_0} + \frac{\mathbb{S}_2}{\mathbb{S}_0} - ... - \frac{\mathbb{S}_{2\ell+1}}{\mathbb{S}_0} \leq \prod_{i \in \mathbb{K}} (1 - \frac{b_i}{a_i}) \leq 1 - \frac{\mathbb{S}_1}{\mathbb{S}_0} + ... + \frac{\mathbb{S}_{2\ell}}{\mathbb{S}_0},$$

are equivalent to the inequalities in the statement of the the lemma and are implied by the Bonferroni inequalities [4]. To see the latter, simply apply the Bonferroni inequalities to k mutually independent events $E_1, ..., E_k$ where E_i has probability $\frac{b_i}{a_i}$ for $i = 1, ..., k$. □

Corollary 17. *Suppose $\mathbb{M}$ is a subnetwork motif of size k, $1 \leq k \leq m \leq n$, and $\vec{0} \leq \vec{\pi} \leq \vec{1}$. Then the second moment of the number of instances of subnetwork motifs of size k in partial binary duplication satisfies*

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) \geq \frac{\Gamma(m) 2^k \Gamma(n + 2||\vec{\pi}||)}{\Gamma(n)\Gamma(m + 2||\vec{\pi}||)} - \frac{\Gamma(m)}{\Gamma(n)} \sum_{j=1}^{k} \frac{2^{k-1}\Gamma(n + 2||\vec{\pi}|| - \pi_j)}{\Gamma(m + 2||\vec{\pi}|| - \pi_j)}$$

and

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) \leq \frac{\Gamma(m) 2^k \Gamma(n + 2||\vec{\pi}||)}{\Gamma(n)\Gamma(m + 2||\vec{\pi}||)}.$$

Proof. Our approach to prove the statement of the Corollary is to find bounds for the second moment given a particular $\vec{s}$ and then use them to bound the result of Theorem 6.

Recall from the proof of Theorem 6 that

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) = \binom{n-1}{m-1}^{-1} \Big(\sum_{h=k}^{n-m+k} \binom{n-h-1}{m-k-1} \sum \prod_{i=1}^{k} g(\pi_i, s_i) \Big)$$

where the second sum is over $||\vec{s}|| = h$.

To obtain bounds for $\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big)$ we start by applying Lemma 10 to $\prod_{i=1}^{k} g(\pi_i, s_i)$ where $a_i = h(\pi_i, s_i)$ and $b_i = f(\pi_i, s_i)$ since
$g(\pi_i, s_i) = h(\pi_i, s_i) - f(\pi_i, s_i)$. Thus the simplest bounds for $\prod_{i=1}^{k} g(\pi_i, s_i)$ are,

$$\mathbb{S}_0 - \mathbb{S}_1 \leq \prod_{i=1}^{k} g(\pi_i, s_i) \leq \mathbb{S}_0$$

where $\mathbb{S}_0 = \prod_{i=1}^{k} h(\pi_i, s_i)$ and $\mathbb{S}_1 = \sum_{j=1}^{k} f(\pi_j, s_j) \prod_{i=1, i\neq j}^{k} h(\pi_i, s_i)$.

By replacing $\binom{n-h-1}{m-k-1}$ with its generating function and using the bounds for $\prod_{i=1}^{k} g(\pi_i, s_i)$ we see

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) > \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\big(\mathbb{S}_0 - \mathbb{S}_1\big)\Big)$$

and

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) < \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\mathbb{S}_0\Big).$$

Thus we can use the generating functions for $h(\pi_i, s_i)$ and $f(\pi_j, s_j)$ to calculate upper and lower bounds for the second moment. The upper bound is as follows

$$\begin{aligned}
\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) &< \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\big(\prod_{i=1}^{k} h(\pi_i, s_i)\big)\Big) \\
&= \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\Big(\prod_{i=1}^{k} \frac{2x}{(1-x)^{2\pi_i+1}}\Big)\Big) \\
&= \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\Big(\prod_{i=1}^{k} \frac{2x}{(1-x)^{2\pi_i+1}}\Big) \\
&= \binom{n-1}{m-1}^{-1} [x^{n-m}]\frac{2^k}{(1-x)^{m+2||\vec{\pi}||}} \\
&= \frac{\Gamma(m)2^k\Gamma(n+2||\vec{\pi}||)}{\Gamma(n)\Gamma(m+2||\vec{\pi}||)}.
\end{aligned}$$

To obtain results for the lower bound we calculate the second term in the lower bound since the first term is the upper bound. Thus,

$$\begin{aligned}
&\binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\Big(\sum_{j=1}^{k} f(\pi_j, s_j) \prod_{i=1, i\neq j}^{k} h(\pi_i, s_i)\Big) \\
&= \binom{n-1}{m-1}^{-1} [x^n]\Big((\frac{x}{(1-x)})^{m-k}\Big(\sum_{j=1}^{k} \frac{x}{(1-x)^{\pi_i+1}} \prod_{i=1, 1\neq j}^{k} \frac{2x}{(1-x)^{2\pi_i+1}}\Big) \\
&= \binom{n-1}{m-1}^{-1} \sum_{j=1}^{k} [x^{n-m}]\frac{2^{k-1}}{(1-x)^{m-1+(2||\vec{\pi}||-\pi_j)}} \\
&= \frac{\Gamma(m)}{\Gamma(n)} \sum_{j=1}^{k} \frac{2^{k-1}\Gamma(n+2||\vec{\pi}||-\pi_j)}{\Gamma(m+2||\vec{\pi}||-\pi_j)}.
\end{aligned}$$

Therefore the lower bound is as follows

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) > \frac{\Gamma(m)2^k\Gamma(n+2||\vec{\pi}||)}{\Gamma(n)\Gamma(m+2||\vec{\pi}||)} - \frac{\Gamma(m)}{\Gamma(n)} \sum_{j=1}^{k} \frac{2^{k-1}\Gamma(n+2||\vec{\pi}||-\pi_j)}{\Gamma(m+2||\vec{\pi}||-\pi_j)}.$$

□

Bounds are now obtained for the second moment when n is large in order to display explicitly the growth rate with respect to n. If $\vec{\pi} = \vec{0}$ the inheritance process is trivial since no there is no chance that an inheritance event can occur and $|\mathbb{M}(n)| = 1$. If $\pi_i = 0$ for some i, $1 \leq i \leq k$, then the only member of the family that occurs in instances of $\mathbb{M}$ is the original gene, and the i^{th} family is non-reproductive.

Now consider a modified version of the inheritance and duplication process. Suppose k' is the size of a subnetwork motif and $\vec{\pi}'$ is the vector of inheritance probabilities. Then $\mathbb{M}'$ is a subnetwork motif of size k' where all non-reproductive families have been removed from $\vec{\pi}$. If $\mathscr{I}$ is a potential instance of $\mathbb{M}$ then $\mathscr{I}'$ is a potential instance of $\mathbb{M}'$ which is obtained by dropping all non-reproductive genes. Then the probability of the modified inheritance event $\mathscr{I}' \in \mathbb{M}'(n)$ is the same as the probability of the original inheritance event $\mathscr{I} \in \mathbb{M}(n)$. Thus it suffices to study $\mathbb{M}'$ since the probability distributions of $|\mathbb{M}(n)|$ and $|\mathbb{M}'(n)|$ are the same. Therefore without loss of generality we can assume that $\vec{0} < \vec{\pi} \leq \vec{1}$. That is, $\mathbb{M}'$ has no family with a zero probability of inheritance.

Corollary 18. *Suppose $\mathbb{M}$ is a subnetwork motif of size k, $1 \leq k \leq m \leq n, n \geq k^2$ and $\vec{0} < \vec{\pi} \leq \vec{1}$. Then the second moment of the number of instances of subnetwork motifs of size k in partial binary duplication satisfies*

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) \geq \frac{n^{2||\vec{\pi}||}2^k\Gamma(m)}{\Gamma(m+2||\vec{\pi}||)} - \frac{1}{\Gamma(m)}\sum_{j=1}^{k}\frac{2^{k-1}n^{2||\vec{\pi}||-\pi_j}}{\Gamma(m+2||\vec{\pi}||-\pi_j)}$$

and

$$\mathbb{E}\Big(|\mathbb{M}(n)|^2; m, n, \vec{\pi}, k\Big) \leq \frac{n^{2||\vec{\pi}||}2^k\Gamma(m)}{\Gamma(m+2||\vec{\pi}||)}\Big(1 + \frac{6||\vec{\pi}||^2}{4n}\Big).$$

Proof. Since

$$\frac{\Gamma(n+2||\vec{\pi}||)}{\Gamma(n)} = (n+2||\vec{\pi}||-1)_{2||\vec{\pi}||}$$

and

$$\frac{\Gamma(n+2||\vec{\pi}||-\pi_j)}{\Gamma(n)} = (n+2||\vec{\pi}||-\pi_j-1)_{2||\vec{\pi}||-\pi_j}$$

the bounds for the corollary fall directly from applying Lemma 14 (occurs in Section 6) to the result of Corollary 17. □

5 FULL DUPLICATION FOR MULTIPLE SUBNETWORK MOTIFS

In this section of the thesis we will study the number of instances of subnetwork motifs in Full Duplication when there are r individual subnetwork motifs before the start of the duplication. We will only explore mutliple motifs in Full Duplication in this thesis. Results for multiple subnetworks in Partial Duplication can be obtained. However, this is something that will be explored in the future.

In this scenario, the duplication process begins with m genes and ends with n total genes. Initially, there are r subnetwork motif instances denoted $\mathbb{M}_i$ for $i = 1, ...r$ and $\mathbb{M}_i(n)$ is the set of subnetwork motif instances after $n - m$ duplications for $i = 1, ..., r$. As before, prior to any duplications each $|\mathbb{M}_i(m)| = 1$. We will let k_i be the size of $\mathbb{M}_i$ for $i = 1, ..., r$ such that $\vec{k} = (k_1, ..., k_r)$ is the vector of r subnetwork motif sizes. If we let $\mathbb{X} = \sum_1^r |\mathbb{M}_i(n)|$ then the expected number of subnetwork motif instances can be denoted $\mathbb{E}\Big(\mathbb{X}|m, n\Big)$.

Theorem 7. *Let $r \geq 1$ and $k_i \geq 1$ for $i = 1, ..., r$ be as above and let $m \leq n$ such that $k_i \leq m$ for $i = 1, ...r$. Then the expected total number of subnetwork motif instances in Full Duplication is*

$$\mathbb{E}\Big(\mathbb{X}|m, n\Big) = \sum_{i=1}^{r}\frac{\Gamma(n+k_i)\Gamma(m)}{\Gamma(n)\Gamma(m+k_i)}.$$

Proof. The lemma follows from Theorem 1 and the linearity of the expectation. □

Now we are going to derive the second moment of $\mathbb{X}$. In order to do so we will apply linearity of expectation to $\mathbb{X}^2$ which gives

$$\mathbb{E}\Big(\mathbb{X}^2; m, n\Big) = \sum_{i=1}^{r} \mathbb{E}\big(|\mathbb{M}_i(n)| \cdot |\mathbb{M}_i(n)|\big) + 2\sum_{i=1}^{r-1}\sum_{j=i+1}^{r} \mathbb{E}\big(|\mathbb{M}_i(n)| \cdot |\mathbb{M}_j(n)|\big). \tag{6}$$

Note that the summand in the first summation is given by Theorem 2. In order to illustrate what may occur in the second summation we will consider the example $r = 2, k_1 = 2$, and $k_2 = 2$. Then there are three different scenarios for the joint distribution of $\mathbb{M}_1(n)$ and $\mathbb{M}_2(n)$. Each scenario differs based on how the 2 original subnetwork motifs are related. The two individual subnetwork motif instances can present as disjoint such that $\mathscr{I}_1 = (a_1, a_2) \in \mathbb{M}_1(n)$ and $\mathscr{I}_2 = (a_3, a_4) \in \mathbb{M}_2(n)$. In the second scenario the subnetwork motifs can share one gene such that $\mathscr{I}_1 = (a_1, a_2) \in \mathbb{M}_1(n)$ and $\mathscr{I}_2 = (a_1, a_3) \in \mathbb{M}_2(n)$. Lastly, the subnetwork motifs can share the same genes with differing mechanisms where $\mathscr{I}_1 = (a_1, a_2) \in \mathbb{M}_1(n)$ and $\mathscr{I}_2 = (a_1, a_2) \in \mathbb{M}_2(n)$ are original instances of these genes. In order to calculate the results for $\mathbb{E}\big(|\mathbb{M}_1(n)| \cdot |\mathbb{M}_2(n)|\big)$ we will use $a(x), b(x),$ and $c(x)$ from the proofs of Theorems 1 and 2. The results for each case are as follows:

Shared gene case where $m = 3$:

$$\begin{aligned}
\mathbb{E}\big(|\mathbb{M}_1(n)| \cdot |\mathbb{M}_2(n)|\big) &= [x^n]\Big(a(x)^{m-3} \cdot b(x)^2 \cdot c(x)\Big)\binom{n-1}{m-1}^{-1} \\
&= \Big([x^{n-m-1}]2(1-x)^{-(m+4)} + [x^{n-m}](1-x)^{-(m+3)}\Big)\binom{n-1}{m-1}^{-1} \\
&= \Big(\frac{2\Gamma(n+3)\Gamma(m)\Gamma(n-m+1)}{\Gamma(m+5)\Gamma(n-m-1)\Gamma(n)} + \frac{\Gamma(n+3)\Gamma(m)\Gamma(n-m+1)}{\Gamma(m+4)\Gamma(n-m)\Gamma(n)}\Big).
\end{aligned}$$

Same genes, different mechanism case where $m = 2$:

$$\begin{aligned}
\mathbb{E}\big(|\mathbb{M}_1(n)| \cdot |\mathbb{M}_2(n)|\big) &= [x^{n-m-i}]\sum_{i=0}^{2} 2^i\binom{2}{i}(1-x)^{-(m+i+2)}\binom{n-1}{m-1}^{-1} \\
&= \sum_{i=0}^{2} 2^i\binom{2}{i}\frac{\Gamma(n+2)\Gamma(n-m+1)\Gamma(m)}{\Gamma(n-m-i+1)\Gamma(m+i+2)\Gamma(n)}.
\end{aligned}$$

We will now look at the expectation of the number of instances of r subnetwork motifs. Initially there are r original subnetwork motif instances $\mathscr{I}_1, ..., \mathscr{I}_r$ with size $k_1, ..., k_r$. Suppose for original instances $\mathscr{I}_i$ and $\mathscr{I}_j$ there exist an overlap in the genes that make up each instance, then we denote the overlap size $\ell_{i,j}$. Then the summand in the first summation of Equation 6 is a function of m, n, k_i, k_j and $\ell_{i,j}$. Thus

$$\mathbb{E}\Big(|\mathbb{M}_i(n)| \cdot |\mathbb{M}_j(n)|\Big) = \binom{n-1}{m-1}^{-1}[x^n]\Big(a(x)^{m-k_i-k_j}b(x)^{k_i+k_j-\ell_{i,j}}(c(x)^{\ell_{i,j}}\Big).$$

Lemma 11. *Suppose $\mathbb{M}_i$ and $\mathbb{M}_j$ are subnetwork motifs of size k_i and k_j respectively, $l_{i,j}$ is the overlap in genes between $\mathbb{M}_i$ and $\mathbb{M}_j$, and $1 \le k_i \le k_j \le m \le n$. Then $\mathbb{E}\Big(|\mathbb{M}_i(n)| \cdot |\mathbb{M}_j(n)|\Big)$ is the expected number of instances of $\mathbb{M}_i$ and $\mathbb{M}_j$, which evaluates to*

$$\frac{\Gamma(n+k_i+k_j)\Gamma(m)\Gamma(n-m+1)}{\Gamma(n)}\sum_{u=0}^{\ell_{i,j}} 2^u\binom{\ell_{i,j}}{u}\Big(\Gamma(m+k_i+k_j+u)\Gamma(n-m-u+1)\Big)^{-1}.$$

Proof.

$$[x^n]\Big(\frac{x}{1-x}\Big)^{m-k_i-k_j}\Big(\frac{x}{(1-x)^2}\Big)^{k_i+k_j-\ell_{i,j}}\Big(\sum_{i=1}^{\ell_{i,j}}\binom{\ell_{i,j}}{i}(\frac{2x^2}{(1-x)^3)^i}(\frac{x}{(1-x)^2})^{\ell_{i,j}}\Big)$$

□

In the following theorem we define $R(m,n)$ as $\frac{\Gamma(m)\Gamma(n-m+1)}{\Gamma(n)}$.

Corollary 19. *Let $r \geq 1$ and $k_i \geq 1$ be the size of $\mathbb{M}_i$ for $i = 1, ..., r$. Then the expected number of ordered pairs of subnetwork motifs in Full Duplication is $\mathbb{E}\big(\mathbb{X}^2 \,|\, m, n\big)$, which evaluates to $R(m,n)$ times*

$$\sum_{i=1}^{r}\Big(\Gamma(n+k_i)\sum_{u=0}^{k_i}2^u\binom{2}{u}\big(\Gamma(m+2+u)\Gamma(n-m-u+1)\big)^{-1}$$
$$+2\sum_{i=1}^{r-1}\sum_{j=i+1}^{r}\Gamma(n+k_i+k_j)\sum_{u=1}^{\ell_{i,j}}2^u\binom{\ell_{i,j}}{u}\big(\Gamma(m+k_i+k_j+u)\Gamma(n-m-u+1)\big)^{-1}\Big).$$

Proof. The corollary follows from applying Theorem 2 and Lemma 11 to Equation 6. □

6 USEFUL FORMULAE FOR APPROXIMATING MOMENTS

We will now derive approximations that are both rapid in computation and maintain accuracy for large n. Note that large n corresponds in biology to an organism with a large genome. The following lemmas were useful for exponentiating approximations to logarithms for various moments in the previous section.

Lemma 12. *If $0 \leq \zeta \leq \frac{1}{2}$ then*

$$1 + \zeta \leq e^{\zeta} \leq 1 + \frac{3\zeta}{2}.$$

Proof. We know that for any ζ

$$e^{\zeta} = \sum_{i=0}^{\infty}\frac{\zeta^i}{i!}.$$

The lower bound is obvious for any $\zeta \geq 0$. For the upper bound we have

$$e^{\zeta} \leq 1 + \zeta + \zeta\sum_{i=2}^{\infty}\frac{(1/2)^{i-1}}{2} = 1 + \frac{3}{2}\zeta.$$

□

Lemma 13. *If $0 \leq \zeta \leq \frac{1}{2}$ then*

$$1 - \zeta \leq e^{-\zeta} \leq 1 - \zeta + \frac{1}{2}\zeta^2.$$

Proof. We know that

$$e^{-\zeta} = \sum_{i=0}^{\infty}\frac{(-1)^i\zeta^i}{i!}.$$

Since $|\zeta| \leq \frac{1}{2}$, the terms in the series expansion of $e^{-\zeta}$ are alternating in sign and their absolute value is strictly decreasing and tending to 0. The lower bound is obtained by truncating the series sum after a negative term and similarly the upper bound is obtained by truncating the series sum after a positive term. □

Lemma 14. *If $j \geq 1$ and $y \geq j^2$ then*

$$1 \leq \frac{(y+j-1)_j}{y^j} \leq 1 + \frac{3j^2}{4y}.$$

Proof. Since the lower bound is trivial we will focus on bounding the ratio above. Let

$$\delta = \ln(\frac{(y+j-1)_j}{y^j}) = \ln\Big(\prod_{r=1}^{j-1}(1+\frac{r}{y})\Big).$$

Since the log of the product is equal to the sum of the logs and $\ln(1+\frac{r}{y}) \leq \frac{r}{y}$ for $r \geq 0$ we obtain the following bounds

$$\delta \leq \sum_{r=1}^{j-1}\Big(\frac{r}{y}\Big) = \frac{\binom{j}{2}}{y} \leq \frac{j^2}{2y} \leq \frac{1}{2}. \tag{7}$$

Note that $e^\delta \leq 1+\frac{3}{2}\delta$ whenever $0 \leq \delta \leq \frac{1}{2}$ and $\frac{j^2}{2y} \leq \frac{1}{2}$, therefore

$$e^\delta = \sum_{i=0}^{\infty}\frac{\delta^i}{i!} \leq 1+\delta+\delta\sum_{i=2}^{\infty}\frac{(1/2)^i}{i!} \leq 1+\frac{3}{2}\delta$$

Since $\frac{j^2}{2y} \leq \frac{1}{2}$ we can apply Lemma 12 to obtain the result of the lemma. □

Lemma 15. *If $j \geq 1$ and $y \geq j^2$ then*

$$1-\frac{j^2}{2y} \leq \frac{y^j}{(y+j-1)_j} \leq 1.$$

Proof. Since the upper bound is trivial we will focus on bounding the ratio below. Using δ as it is in Lemma, note that $e^{-\delta} \geq 1-\delta$ for any real δ.

Now we know $\delta \leq \frac{j^2}{2y}$. Thus we will bound $e^{-\delta}$ to obtain the following

$$e^{-\delta} \geq 1-\frac{j^2}{2y} \text{ when } \delta \geq 0.$$

□

Lemma 16. *Suppose $j \geq 1$ and $y \geq 2j$ then*

$$1-\frac{j^2}{y} \leq \frac{(y)_j}{y^j} \leq 1.$$

Proof. Since the upper bound is trivial we will focus on bounding the expression below. Let

$$\delta = -\ln(\frac{(y)_j}{y^j}) = -\ln(\prod_{r=0}^{j+1}(1-\frac{r}{y}))$$

In order to prove the statement of the Lemma we must obtain bounds for the above expression. Since $|\frac{r}{y}| < 1$ we will use the Taylor Series to obtain bounds. Let

$$\delta = \sum_{r=1}^{j-1}\sum_{i=1}^{\infty}\frac{(r/y)^i}{i} = \sum_{i=1}^{\infty}\sum_{r=1}^{j-1}\frac{(r/y)^i}{i}.$$

When $i=1$ we know that

$$\sum_{r=1}^{j-1}\frac{r}{y} = \frac{1}{y}(1+2+...+j-1) = \frac{j(j-1)}{2y} \leq \frac{j^2}{2y}.$$

When $i \geq 2$ a bound for the inner summation is as follows

$$\sum_{r=1}^{j-1} \frac{(r/y)^i}{i} \leq \frac{j^{i+1}}{2y^i}.$$

Then

$$\sum_{i=2}^{\infty} \frac{j^{i+1}}{2y^i} = \frac{j^2}{2y} \sum_{i=2}^{\infty} (\frac{j}{y})^{i-1} \leq \frac{j^2}{2y} \sum_{i=2}^{\infty} (\frac{1}{2})^{i-1} = \frac{j^2}{2y}$$

since

$$\sum_{i=1}^{\infty} (\frac{1}{2})^{i-1} = 1.$$

Thus the lemma follows since

$$0 \leq \delta \leq \frac{j^2}{y}$$

so that

$$1 - \frac{j^2}{y} \leq 1 - \delta \leq e^{-\delta} \leq 1.$$

□

Lemma 17. *Suppose $j \geq 1$ and $y \geq 2j^2$ then*

$$1 \leq \frac{y^j}{(y)_j} \leq 1 + \frac{3j^2}{2y}.$$

Proof. Since the lower bound is trivial we will focus on bounding the expression above. Let

$$\frac{y^j}{(y)_j} = e^{\delta}$$

where δ is as is in the proof of Lemma 16.

From the proof of Lemma 16 we know that

$$\delta \leq \frac{j^2}{y}.$$

Therefore $\delta \leq \frac{1}{2}$ since $y \geq 2j^2$.

From proof of Lemma 14 we know that

$$e^{\delta} \leq 1 + \frac{3}{2}\delta \text{ when } \delta \leq \frac{1}{2}$$

Thus

$$1 \leq e^{\delta} \leq 1 + \frac{3}{2}\delta \leq 1 + \frac{3j^2}{2y}.$$

□

Lemma 18. *If $j \geq 1$ and $y \geq 2j^2$ then*

$$1 \leq \Big(\frac{(y+j-1)_j}{y^j}\Big)^2 \leq 1 + \frac{3j^2}{2y}.$$

Proof. In order to prove the statement of the lemma let

$$\delta = \ln\Big((\frac{(y+j-1)_j}{y^j})^2\Big) = 2\ln\Big(\prod_{r=1}^{j-1}(1+\frac{r}{y})\Big),$$

just as in the proof of Lemma 3. Therefore the ratio we are bounding is $e^{2\delta}$. Recall from Equation 7 that $\delta \leq \frac{j^2}{2y}$ and $2\delta \leq 1/2$ which mean $\delta \leq 1/4$. Thus,

$$e^{2\delta} \leq 1 + 3\delta \leq 1 + \frac{3j^2}{2y}$$

since $e^{\delta} \leq 1 + \frac{3}{2}\delta$. □

7 DISCUSSION

7.1 CONNECTION TO POLYA URNS

To study a more realistic model it is best to generalize the previous duplication model by generalizing the start of the duplication process. The *gene duplication process* discussed in Section is an extension of the random duplication process that begins with w individual gene families where each family has size $s_i \geq 1$ for $i = 1, ..., w$. We let $\vec{s} = (s_1, ..., s_w)$ and $m = s_1 + ... + s_w$. At each step, a random gene is selected to be duplicated. If a gene in the i^{th} family is duplicated then the new duplicated gene belongs to the i^{th} family. After $n - m$ duplications we let $t_i \geq 0$ be the number of duplicated genes that have been added to the i^{th} family and $\vec{t} = (t_1, ..., t_w)$. Note that the gene duplication process is a special case of this generalized process that begins with m single gene families.

Theorem 8. *Let $\vec{s}$ be the initial composition of family sizes and let $\vec{X}$ be the vector of numbers of individuals added to each family after $n - m \geq 0$ random duplications. Then for any weak composition $\vec{t}$ of $n - m$ into w parts we have*

$$P[\,\vec{X} = \vec{t}\,|\,\vec{s}\,] = \binom{n-m}{t_1, ..., t_w}\frac{(m-1)!}{(n-1)!}\prod_{j=1}^{w}\frac{(s_j+t_j-1)!}{(s_j-1)!}.$$

Proof. We need to calculate the likelihood that $\vec{X} = \vec{t}$ given $\vec{s}$. In order for $\vec{X} = \vec{t}$ at the end of a duplication sequence, the number of genes that have been duplicated in the j^{th} family must be x_j for $1 \leq j \leq w$. That is to say that $x_j = t_j$ for $j = 1, ..., w$ after $n - m$ duplications. There are exactly t_j locations in the sequence of $n - m$ duplications in which the duplication can occur in the j^{th} family for $1 \leq j \leq w$. Then a standard combinatorial fact is that the number of ways we can choose the locations so that $\vec{X} = \vec{t}$ is the multinomial coefficient $\binom{n-m}{t_1,...,t_w}$.

Given one of these patterns we can calculate the probability of its occurrence in a random duplication process as follows. Consider the i^{th} duplication in the series of $n - m$ duplications. For $\vec{X} = \vec{t}$, the given pattern determines the j such that the j^{th} family that must be duplicated at the i^{th} step. Then the probability that the i^{th} duplication occurs in the j^{th} family is the current size of the j^{th} family divided by $n + i - 1$, since the total number of genes at the i^{th} duplication step is $n + i - 1$. Therefore, the probability that $\vec{X} = \vec{t}$ is the product of the probabilities that the i^{th} duplication occurs in the j^{th} family for $i = 1, ..., n - m$ and $j = 1, ..., w$. Then for $i = 1, ..., n - m$, the denominators are as follows

$$\Big((m)\cdot(m+1)\cdot\cdots\cdot(m+(n-m-1))\Big) = \frac{(n-1)!}{(m-1)!}$$

and for $j = 1, ..., w$ the numerators are

$$\Big(s_j\cdot(s_j+1)\cdots(s_j+t_j-1)\Big) = \frac{(s_j+t_j-1)!}{(s_j-1)!}.$$

Therefore the probability that the given pattern is actually followed such that $\vec{X} = \vec{t}$ is

$$\frac{(m-1)!}{(n-1)!}\prod_{j=1}^{w}\frac{(s_j+t_j-1)!}{(s_j-1)!}.$$

Clearly the $\binom{n-m}{t_1,\ldots,t_w}$ different patterns of duplication are mutually exclusive. Therefore the total probability of $\vec{X}=\vec{t}$ given $\vec{s}$ is the product of this multinomial coefficient and the probability of duplication for each family, giving the result of the theorem. □

The formula from the conclusion of Theorem 8 shows up in the analysis of the Kriz polya urn model [11] in the case that the Kriz model parameter $s=1$. The Kriz model is a multi-urn model used to study the spread of a disease and the parameter s gives the number of people likely to to come into contact with the disease in a specific unit of time. The result of Theorem 8 can be expressed as

$$P[\,\vec{X}=\vec{t}\,|\,\vec{s}\,]=\binom{n-m}{t_1,\ldots,t_w}\frac{\prod_{j=1}^{w}(\prod_{l=0}^{t_j-1}(s_j+l))}{\prod_{i=m}^{n-1}i},$$

which is essentially the same formula as the one following the displayed expression [11, eq. (9.8)] after setting the Kriz model parameter s to 1.

The Kriz model is expressed in terms of urns, balls, and colors. To correspond to our model, the urn would be the entire genome, the balls would be the genes, and the colors would be the families. Our model is a special case of the Kriz model where there is one urn with n balls, m colors, and only one ball can be added at each step.

7.2 CONNECTION TO MITES

It has been expressed that transposable elements are one of the reasons a genome can make programmed responses to environmental challenges [13]. At its core, the genome uses transposable elements to rewire expressions in response to environmental challenges [13]. The rice genome has putative regulatory elements called MITES also know as miniature inverted transposable elements [5] that identify which subnetwork motifs are may be under selection to an environmental challenge. These MITES undergo a high rate of duplication and genes in the rice genome are sorted based on their MITE profile. This results in MITE-infested genes having a history of regulation by their MITES.

If two genes A and B share the same MITE, they are said to share a regulatory link. In the rice genome a simple subnetwork motif would be a connected collection of three genes and three MITES, where genes A and B share MITE 1, genes A and C share MITE 2, and genes B and C share MITE 3.

Genes in the rice genome can be sorted into gene families based on their DNA sequence or their MITE profile. The age of the links between genes can be characterized from the sequences of each MITE [18]. That is to say, the more divergence there is in the sequence between MITEs, the older the link between the genes. Moreover, the MITEs associated with genes can define regulatory links. Therefore, a history of different genes and their potential regulation by MITES can be reconstructed [18].

The modeling of MITE regulatory evolution is a direct application of the duplication models discussed in this thesis. Suppose we have the simple network discussed in the previous paragraph. If gene A gets duplicated to create gene A' and the MITE between A' and C is retained but the MITE between A' and B is not, then the simple regulatory network has undergone Partial Duplication.

Subnetwork motifs can be examined in the rice genome using MITES since it should be possible to consider a couple of gene families that have expanded over time and how the MITES evolved with them. To a limited extent the fate of the network can be examined.

7.3 MIXED DUPLICATION MODEL

The mixed duplication model defined in [2] is a duplication model where nodes in a network are duplicated along with their connections, which is a common occurrence in biological systems, especially in the context of gene duplication [2]. In the model the results of both Full Duplication and Partial Duplication on the

structure and connectivity of the network are examined. Note that node connectivity refers to the regulatory relation between genes in a genome. Under the mixed duplication model both full and Partial Duplication are considered at the same time. There are two probabilities, p and q, that define the behavior of the model. At each step in the duplication process a node is randomly selected for duplication. Then with probability $1-q$ all regulatory links are retained and with probability q each link is retained independently with probability p.

What we consider Full Duplication can be described as mixed duplication when $q = 0$, which is how it is referenced in [2]. However, our Partial Duplication mode is a vast generalization of the mixed duplication model in [2] such that each family has its own inheritance probability for regulatory links. One of the limitations of the mixed duplication model is that it is restrictive to situations where all regulation relationships have the same probability of inheritance through duplication. One case in which this model can be applied is when transcription factors are regulating a very high number of targets and have independently mutating binding sites. Our model extends the applications of the mixed duplication model and is also appropriate for transcription factors that are regulating several targets and the mutations occur in the regulator rather than the binding site.

The mixed duplication model is a special case of our Binary Inheritance mode where the inheritance probability for each π_i for $i = 1, ..., k$ is $\pi_i = 1 - q + qp$. Consider the mixed duplication model where the probability for link retention under Partial Duplication is $p = 0$. Then at each step of the node duplication process $1-q$ is the probability that all of the links where retained in Full Duplication and q is the probability that none of the links where retained in Partial Duplication. As already noted, the subnetwork motif results of this thesis apply to the mixed duplication model [2].

References

[1] Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York, ninth dover printing, tenth gpo printing edn. (1964)

[2] Chung, F., Lu, L., Gregory, D.T., Galas, D.J.: Duplication models for biological networks. Journal of Computational Biology (5), 677–687 (2003)

[3] Concant, G.C., Wagner, A.: Convergent evolution of gene circuits. Nature Genetics **34**(3), 264–266 (2003)

[4] Feller, W.: An Introduction to Probabillity Theory and its Applications, vol. 1. John Wiley & Sons, Incorporated, 3 edn. (1957)

[5] Feschotte, C., Zhang, X., Wessler, S.R.: Miniature inverted-repeat transposable elements and their relationship to established dna transposons. Mobile DNA II pp. 1147–1158 (2002)

[6] Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T., et al.: Lineage-specific gene duplication and loss in human and great ape evolution. PLoS biology **2**(7), e207 (2004)

[7] Graham, R.L., Knuth, Donald E.and Patashnik, O.: Concrete mathematics. Addison-Wesley **6**(1), 320–321 (1990)

[8] Hallin, J., Landry, C.R.: Regulation plays a multifaceted role in the retention of gene duplicates. PLoS Biology **17**(11) (2019)

[9] Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E., Young, R.A.: Transcriptional regulatory code of a eukaryotic genome. Nature **431**(7004), 99–104 (2004). https://doi.org/10.1038/nature02800, `https://doi.org/10.1038/nature02800`

[10] Lee, T.I., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Georg, G., Hannett, N., Harbison, C., Thompson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D.B., Ren, B., Wyrick, J., Jean-Bosco, T., Volkert, T., Fraenkel, E., Gifford, D., Young, R.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science **298**(5594), 799–804 (October 2002)

[11] Mahmoud, H.M.: Polya Urn Models. CRC Press, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742 (2009)

[12] Mazurie, A., Bottani, S., Vergassola, M.: An evolutionary and functional assessment of regulatory network motifs. Genome Biology **6**(4) (2005)

[13] McClintock, B.: The significance of responses of the genome to challenge. Science (226), 792–801 (1984)

[14] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., Alon, U.: Superfamilies of evolved and designed network. Science **303**(5663), 1538–1542 (2004)

[15] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)

[16] Polhilko, A., Fernandez, A.P., Edwards, K.D., Southern, M.M., Halliday, K.J., Miller, A.J.: The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops. Molecular Systems Biology **8**(574) (2012)

[17] Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., Young, R.A.: Genome-wide location and function of dna binding proteins. Science **290**(5500), 2306–2309 (Dec 2000). https://doi.org/10.1126/science.290.5500.2306

[18] SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L.: The paleontology of intergene retrotransposons of maize. Nature Genetics (20), 43–45 (1998)

[19] Teichmann, S.A., Babu, M.M.: Gene regulatory network growth by duplication. Nature Genetics **36**(5), 492–496 (2004)

[20] Yokobayashi, Y., Weiss, R., Arnold, F.H.: Directed evolution of a genetic circuit. PNAS (26), 16587–16591 (2002)

[21] Zhang, J.: Gene duplication. Genetics **9**, 938–950 (1914)