Small peripheral structures in unlabelled trees and the evolution of polyploids

Fatemeh Pouryahya

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Many angiosperms have undergone some series of polyploidization events over the course of their evolutionary history. In these genomes, especially those resulting from multiple autopolyploidization, it may be relatively easy to recognize all the ξ sets of n homeologous chromosomes, but it is much harder, if not impossible, to partition these chromosomes into n subgenomes, each representing one distinct genomic component of ξ chromosomes making up the original polyploid. Thus, if we wish to infer the polyploidization history of the genome, we could make use of all the gene trees inferred from the genes in one set of homeologous chromosomes to construct a consensus tree, but there is no evident way of combining the trees from the ξ different sets, because we have no labelling of the chromosomes that is known to be consistent across these sets. We suggest here that lacking a consistent leaf-labelling, the topological structure of the trees may display sufficient resemblance so that a higher level consensus could be revealing of evolutionary history. This would be especially true of the peripheral structures of the tree, likely representing events that occurred more recently and have thus been less obscured by subsequent evolutionary processes. Here, we present a statistical test to assess whether the subgenomes in a polyploid genome could have been added one at a time. The null hypothesis is that the accumulation of chromosomes follows a stochastic process in which transition from one generation to the next is through randomly choosing an edge, and then subdividing this edge in order to link the new internal vertex to a new external vertex. We analyze the probability distributions of a number of peripheral tree substructures, namely leaf- or terminal-pairs, triples and quadruples, arising from this stochastic process, in terms of some exact recurrences. We propose some conjectures regarding the asymptotic behaviours of these distributions. Applying our analysis to a sugarcane genome, we demonstrate that it is unlikely that the accumulation of subgenomes has occurred one at a time in this genome.

Dedications

This thesis is dedicated to my father, who has always believed in me and has supported me throughout my life. It is also dedicated to my mother for her unconditional love and endless kindness.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Prof. David Sankoff, for his invaluable guidance and support. He helped me to improve my research skills and prepared for the future. I feel so honored and grateful for having the opportunity to join his lab and to work on this interesting project. There is nothing I can do to truly thank him. In addition, I would like to thank my friends and colleagues, Alma Oladi and Mona Meghdari for their kindness and support during my master's program. I also gratefully thank members of my thesis committee, Prof. Gary Walsh and Prof. Dave Amundsen for taking time to provide insightful comments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Virtually all known flowering plants[1] evidence at least one polyploidization event in their lineage, many of them two, three or more. These may have been autopolyploidizations, where a new genome is created consisting of two or more exact copies - subgenomes - of the parent genome, leading to high rates of multivalent pairing of chromosomes during meiosis, or allopolyploidizations, where the subgenomes come from distinct parental species, and where normal bivalent pairing is more general. In both processes, the entire genome, including each chromosome and each gene, is present in two or more copies or versions. It is not uncommon for a present-day flowering plants to have undergone two, three or even more of these events. For example the canola (*Brassica napus*) genome was formed by the tetraploidization of cabbage (*Brassica oleracea*) and Napa cabbage (*Brassica rapa*), which are divergent species originating in an earlier *Brassica* hexaploid. Furthermore, the preceding history of this lineage contains two successive tetraploidization events and a hexaploidization, known as α, β and γ, respectively, the latter reaching back some 120 Mya [16]. Maize (*Zea mays*) is the product of three or four tetraploidizations, the most recent one some 12 Mya [19], as well as the earlier ρ, σ and τ events dating as far back as 50 Mya[17]. These examples illustrate the important fact that two polyploids with $m \geq 1$ and $n \geq 1$ subgenomes may combine to form a polyploid with $m + n$ subgenomes.

As a consequence of polyploidization, the genomes of these plants will contain multiple versions of each of the ξ chromosomes. In autopolyploids, these "homeologous" chromosomes are initially identical copies, while in allopolyploids they will be similar but may have small differences. In any case the homeologous chromosomes eventually diverge due to random point mutation, gene loss or gain and intra- or inter-chromosomal rearrangements.

[1]Except *Amborella trichopeda* [6], the single extant species that diverged first from all the other angiosperms 200 Mya.

In polyploids, it is important to the study of their evolution to understand the sequence of polyploidization events that gave rise to the current genome. This is not a trivial question. For instance, in the octoploid strawberry genome, the sequence of polyploidy events is controversial [11, 7]. In another example, the alloploidization origin of the octoploid Alpine Grey Willow is a matter of speculation [8].It is in the context of this general question that we investigate the evolutionary history of an octoploid species of sugarcane.

In an autopolyploid like sugarcane, where there is little or no exchange between non-homeologous chromosomes, each of the ξ sets of n homeologous chromosomes could be studied as if it were a set of n independent species related by a phylogenetic tree T, descending from a common ancestor and diverging through random mutation. For present purposes, we assume tree T has a binary branching form. And if we were to observe the ξ sets of n homeologous chromosomes, these could be studied as ξ independent observations of T. More specifically, for each of the genes contained in the ξ original chromosomes, the n present day descendants of each of these genes could be considered to be related by the same phylogeny T.

A crucial difficulty in this approach is that the there is usually no clear correspondence connecting the non-homeologous chromosomes deriving from each of the ξ constituent subgenomes, no labelling of the leaves in one set of n homeologs that has a given relationship with a labelling in any of the other sets. Thus we can only compare unlabelled trees, and see whether they share common structural characteristics derived from a common evolutionary history.

As a result of substantial variability in the rates of evolutionary divergence of the chromosomes, the resemblance between two of the n homeologs no longer perfectly reflects the time elapsed since their common ancestor. Nevertheless, the effects of this divergence would affect the most closely related chromosomes the least. Even if the topology of a tree inferred from the resemblances among the n homeologous chromosomes is no longer statistically reflected the earliest branching events, in particular the root of T, there might still be enough signal to detect the most recent events. In general the recent events should be reflected in the peripheral subtrees of T, such as the pairs or triples of the leaves. The goal here is to devise a test of whether the number of these subtrees differs from those predicted by a null hypothesis.

Perhaps the best known pathway to polyploidy of degree greater than tetraploid, is the formation of a tetraploid followed closely in evolutionary time by the recruitment of a third subgenome. This appears to be the process underlying the γ hexaploidy as well as the *Brassica* hexaploidy mentioned above. And this may be a more general way for polyploids to accumulate subgenomes. Our statistical tests for the number of terminal pairs, triples, etc., is based on the null hypothesis that the subgenomes are added one at a time to a vertex inserted on some existing branch of the tree.

Under this hypothesis, we derive a recurrence for $P_{2i,n}$, the probability of i terminal pairs in a binary tree after n generations. We explore a number of avenues to find a closed form for $P_{2i,n}$. Though in applications, n is not a very large number, there is mathematical interest in studying the asymptotic behaviour of the model. We compute the probabilities for a large range of n. It appears that the expected values, normalized by the number of terminals, are asymptotically equal to $\frac{1}{4}$. Furthermore, the normalized variances approach $\frac{1}{16}$, so that the limiting distribution is under-dispersed. In addition, we recursively calculate the probability distributions of other peripheral structures in a binary tree generated after n steps under the null hypothesis. The peripheral structures explored in this thesis are triples and the two possible types of quadruples. We can conjecture that the normalized expected values of the number of triples, type I quadruples, and type II quadruples asymptotically equal to $\frac{1}{8}$, $\frac{1}{64}$, and $\frac{1}{16}$, respectively. Similar to the distributions of terminal pairs, the limiting distribution of the number of triples and of the numbers of two types of quadruples are under-dispersed as well.

We apply our analysis to data on sets of homeologous genes from the sugarcane species *Saccharum officinarum* [14], where $n = 8$ and $\xi = 10$. We find that the necessary resemblance data from the genes in sets of homeologous chromosomes are usually incomplete, either because one or more genes have been lost over the course of evolution, or because of other experimental and measurement difficulties. This leaves us with very few sets of genes as samples to infer a consensus T in any of the ξ subgenomes. Thus we adapt our methods to cases where there are only 7 or only 6 homeologous genes. We apply the same phylogenetic procedure and derive appropriate weightings for the tree bipartitions obtained as evidence for T. We discuss the implications for alternative hypotheses after rejecting the null hypothesis.

This thesis composed of six chapters. Chapter 2 provides a discussion of the related works. In Chapter 3, we present the new model to analyze polypoloidization events, formulate the distribution of paired terminals under the model, and give the analytical results. Chapter 4 analyzes other peripheral structures of the trees including triples and two types of quadruples. In Chapter 5, we describe the framework developed to analyze polyploidization events in *S. officinarum*, and discuss the result under the model. Chapter 6 gives a summary of the remarks and conclusions.

Chapter 2

Background

This chapter sheds light on some approaches used to find recurrence relations for the number of some type of binary trees. This includes the recurrence relation for the number of unlabelled unordered unrooted full-binary trees. Our one-branch-at-a-time model induces a distribution over these type of trees. Then, it describes some basic concepts in polyploidization and differnt types of polyploidy. Furthermore, this chapter briefly introduces *Saccharum officinarum*, and explains the necessity of conducting research in this realm. It, also, describes the difference among cultivars, and last but not least, in this chapter, you will find some remarks on the techniques used in establishment of consensus trees.

2.1 Binary trees

Binary trees have been the subject of numerous studies due to their substantial applications, especially in engineering and computer science [21], [22].

Binary trees are classified in many types based on specific characteristics. A tree in which the degree of each vertex is at most three is called a binary tree. A binary tree where the degree of each internal vertex is exactly three is a full-binary tree. If some labels have been assigned to every vertex of a tree, the tree is called labelled tree. In phylogentic analysis and hierarchical classification studies usually the terminally-labelled trees are a subject of interest. In these trees only terminal vertices have labels. In this thesis by labelled tree, we refer to terminally-labelled tree. An unlabelled tree is a tree where no labels have been assigned to any of its vertices. If a vertex of a tree has been distinguished to be the root of the tree, the tree is rooted. Otherwise, it is called an unrooted (or free) tree. Note that in computer science, roots are usually considered as a vertex of degree two, while in classification studies roots are of degree one similar to terminals. If a fixed order (e.g. left and right child) is imposed upon the

edges incident to each vertex, the tree is called an ordered (or plane) tree. Otherwise, if the relative order of the edges is not important, the tree is known as an unordered tree.

Counting trees may become a subject of interest since the 17^{th} century when Euler counted the number of unlabeled ordered rooted full-binary trees which are the so-called Catalan numbers. The number of this type of binary trees with n terminals is equal to the n^{th} Catalan number, satisfying the following recurrence relation:

$$C_n = \sum_{i=1}^{n-1} C_i C_{n-i} \quad for \quad n \geq 2 \tag{2.1.1}$$

with initial values: $C_0 = 1$ and $C_1 = 1$.

To see this, by definition of this type of tree, it is equal to recursively have this type of trees with i vertices in the left subtree and with the remaining $n - i$ vertices in the right subtree, for $1 \leq i \leq n - 1$. The closed-form solution can be drawn using a generating function $U(z) = C_0 + C_1 z + C_2 z^2 + ...$

Now, for $(U(z))^2$, we have

$$\begin{aligned}(U(z))^2 &= C_0C_0 + (C_1C_0 + C_0C_1)z + (C_2C_0 + C_1C_1 + C_0C_2)z^2 + ... \\ &= C_1 + C_2 z + C_3 z^2 + ...\end{aligned} \tag{2.1.2}$$

This leads to the following quadratic equation which can be easily solved.

$$U(z) = C_0 + z(U(z))^2 \tag{2.1.3}$$

Note that from two solutions, only $U(z) = \frac{1-\sqrt{1-4z}}{2z}$ is acceptable due to the initial condition $U(0) = C_0 = 1$. We can then manipulate the binomial expansion of $(1-4z)^{\frac{1}{2}}$ and use it to simplify the formula to $C_n = \frac{1}{n}\binom{2n-2}{n-1}$.

Ordered trees are more important to the computer science community, and the majority of the previous work found in literature that revolved around enumeration of binary trees focused on the ordered ones. There have been several attempts to facilitate random generation of some type of binary trees by developing some methods with better computational complexity in time and space. Note that this problem can be broken down into developing methods to carry out some general steps: enumeration of non-isomorphic binary trees of desired type, listing and numbering elements (e.g. using ranking/unranking functions), and then generating random samples usually from a uniform distribution over the trees (See e.g. [24, 25, 26, 28, 27, 29]).

In contrast, in hierarchical classification and phylogenetic studies the order of the subtrees is not important, and considering an underlying uniform distribution is not applicable/informative in some analyses.

For unordered trees, the number of labeled rooted full-binary trees denoted by A_n can be drawn recursively as follows (see [32, 31, 30] for more details). Note that a rooted full-binary tree with $n-1$ terminals has $2n-3$ edges. Each of these edges can be subdivided to link the new internal vertex to a new terminal (external) vertex. Thus, for $n \geq 4$,

$$\begin{aligned} A_n &= (2n-3)A_{n-1} \\ &= \frac{(2n-3)!}{2^{n-2}(n-2)!} \end{aligned} \tag{2.1.4}$$

With initial $A_3 = 3$ (the number of way to place 3 distinct labels on the only possible tree topology on 3 terminals ), and in this case, the number of unrooted ones with n terminals would be simply equal to rooted ones with $n-1$ terminals as we may swap the position of the root in the rooted case for the position of additional terminal in the unrooted case.

It is clear that the enumeration of unlabelled trees is a harder problem than the labelled ones due to the isomorphic trees with different layouts. The number of full-binary unlabeled unordered rooted trees satisfies the following recurrence relation [33, p. 54].

$$w_n = \begin{cases} \displaystyle\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor - 1} w_i w_{n-i} + \binom{w_{\frac{n}{2}} + 1}{2} & n \geq 1 \text{ and n is even} \\ \displaystyle\sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} w_i w_{n-i} & n \geq 2 \text{ and n is odd} \end{cases} \tag{2.1.5}$$

with initial values: $w_1 = 1$ and $w_2 = 1$

This formula came from the same logic used to get the recurrence for the Catalan numbers, as we explained. Here, the difference is due to the fact that for unordered trees, there is no difference between left and right subtrees, counting up to the mid-point would suffice. No closed-form solution has ever been established for the recurrence relation 2.1.5 (See [34, 35] for the work on its generating function). If we let $\Delta(k) = \binom{k+1}{2}$ (and $w_x = 0$ for non-integer x), we can write the recurrence 2.1.5 in one line as follows:

$$w_n = \sum_{i=1}^{<\frac{n}{2}} w_i w_{n-i} + \Delta(w_{\frac{n}{2}}) \tag{2.1.6}$$

In [23], Furnas used this recurrence to recursively find the number of unlabelled unordered unrooted full-binary trees. To do so, he introduced a root and an ordering to these type of trees, and then enumerated them. He used the concept of a center

of a graph to root the trees. We recall that the center of a graph is a vertex that has minimal eccentricity[1]. Based on bicentroidal theorem, a tree is either a centered, having only one center, or bicentered, having two adjacent centers, but not both. If the tree is centered, let the center be the root. Otherwise, subdivide the incident edge of the two centers, and let the new internal vertex be the root of the tree. Then, in order to introduce an ordering, note that in any layout of a rooted tree, we can drag and drop each vertex together with its incident edges such that for each internal vertex, subtrees are ordered from lightest to heaviest from left to right, and being lighter means not including more number of terminal vertices. To put it more formally, denote the number of terminal vertices of a tree T with $|T|$, and denote its left and right subtrees with T_L and T_R, respectively. We say a tree T is lighter than T', $T < T'$, if and only if

1. $|T| < |T'|$ or
2. $|T| = |T'|$ and $|T_L| < |T'_L|$ or
3. $|T| = |T'|$ and $|T_L| = |T'_L|$ and $|T_R| < |T'_R|$

Furnas refers to this layout as left-light-rooted trees. It is clear that there is a one to one correspondence between the left-light-centrally-rooted trees and the original unrooted unordered ones. If we denote the number of the centered and bicentered ones with Ψ_n and Υ_n, respectively, the total number of the trees would be

$$P_n = \Psi_n + \Upsilon_n \tag{2.1.7}$$

Using the recurrence relation 2.1.5, it is straightforward to see that the number of centered ones are

$$\Psi_n = \sum_{\substack{0<i\leq j\leq k<\frac{n}{2} \\ i+j+k=n}} F(i,j,k) \tag{2.1.8}$$

Where $F(i,j,k)$ are

$$F(i,j,k) = \begin{cases} w_i w_j w_k & \text{if } i < j < k \\ \Delta(w_i) w_j & \text{if } i = j < k \\ w_i \Delta(w_j) & \text{if } i < j = k \\ \Delta_3(w_i) & \text{if } i = j = k \end{cases} \tag{2.1.9}$$

[1]In a graph G(V,E), the eccentricity of a vertex $v \in V$ denoted by $e(v)$ is the maximum distance from vertex v to any other vertices of the tree. i.e.,

$$e(v) = \max_{v' \in V} \text{distance}(v, v')$$

Where $\Delta_3(k) = \binom{k+2}{3}$ (choosing 3 objects from k objects with replacement), and the number of bicentered ones are

$$\Upsilon_n = \Delta(w_{\frac{n}{2}}) \tag{2.1.10}$$

Furnas depicted a scheme for ranking and unranking functions. Although one can manipulate his approach to develop a method for non-uniform sampling, but the paucity of an explicit formula for equation (2.1.7) makes it generally not efficient to quickly determine the corresponding tree topology in the random sampling.

In this thesis, our one-branch-at-a-time model induces a non-uniform distribution over the unlabelled unordered unrooted full-binary trees that can be recursively enumerated by equation (2.1.7). Enumeration, listing, or more precisely determining the probability of occurrence of each non-homomorphic tree topology based on our one-branch-at-a-time model seems to be a prerequisite to our study in order to make an assessment of whether the constructed phylogenetic trees of the data followed our model or not. However, we got around this problem by narrowing down our focus to peripheral structures of the tree. We successfully determined recurrence relations for some peripheral substructures of the trees, which is sufficient to make a decision about the data with high confidence.

2.2 Polyploidization

Each cell of an organism (somatic cell[2]or gamete) has a certain number of homologous chromosomes. The number of homologous chromosome sets that a cell possesses is called ploidy level. In the literature, the ploidy level and the basic chromosome number (the number of chromosomes in the monoploid set) are shown with the letter N and X, respectively. For example in Figure 2.1, the first one (the haploid cell) is cytologically designated as $N = 4$ or $X = 4$, the second one (the diploid) is designated as $2N = 8$ or $2X = 8$.

Cells may undergo chromosomal variation which can be generally classified into two groups of euploidy and aneuploidy. In euploids, the number of all of the chromosomes is a multiple of basic chromosome numbers. On the other hand, in aneuploids, the total number of chromosomes is abnormal. They may have some extra chromosomes (ascending aneuploidy) or some chromosomes may be missing (descending aneuploidy) [50]. Aneuploidy is a genetic disorder. It can happen to a polyploid or a diploid species. Note that ascending aneuploidy can be tolerated better than the descending one, and although the loss of chromosomes in diploids can be lethal, it often can be tolerated by polyploids due to the duplicated genetic materials [51].

Euploids can be monoploids, diploids, or polyploids. Haploids are individuals or

[2]All cells of individuals except their gametes (sex cells) are called somatic cells.

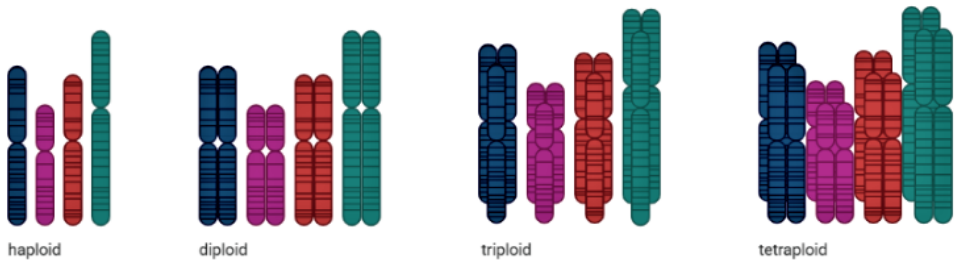


Figure 2.1: **An illustration that shows the number of homologous chromosomes in a haploid(ploidy number of one), diploid (ploidy number of two), triploid (ploidy number of three), and tetraploid (ploidy number of four) cell.**

gametes who have the same number of chromosomes as the gametes. The term is often applied to species that are derived from diploids. A well-known example of haploids is a species of *Drosophila*. There are more examples of animal haploids such as some species of frog, salamanders, chicken, etc. Haploidy in normally diploid animal individuals usually leads to a detrimental physiological condition that results in miscarriage. On the other hand, monoploid plants can survive, but they are usually sterile. They are especially important in genetic improvement. Note that monoploidy paves the way to distinguish recessive/dominant alleles, for example in testing the plant's resistance to particular herbicides.

The majority of eukaryotes are diploids which means in their somatic cells they have two sets of chromosomes each inherited from one of their parents. The number of chromosomes in gametes of diploids is the same as the number of their basic chromosome number.

Polyploidy is a condition in which there appear more than two subgenomes in somatic cells. Although few animal species are polyploids, and no mammals, polyploidy is widespread among plants. It is a key component of adaptation and speciation. Thus, it plays an important role in plant evolution and diversity. Polyploids can arise naturally and artificially. They can form through the whole genome duplication of embryonic somatic cells but more importantly through developments of diploid gametes in meiosis owing to the non-disjunction of homologous chromosomes [48]. They can also be generated artificially through drug treatment of organism's cells that prevent cell division (e.g. via colchicine), see [50] for more examples.

Polyploids can be classified based on the evolutionary path to their formation. Although there has been some variation in the definition of different types of polyploids [53], the classification of Stebbins in [52] is generally acceptable. He decently categorized polyploids into three major types of autopolyploids, allopolyploids, and

segmental allopolyploids.

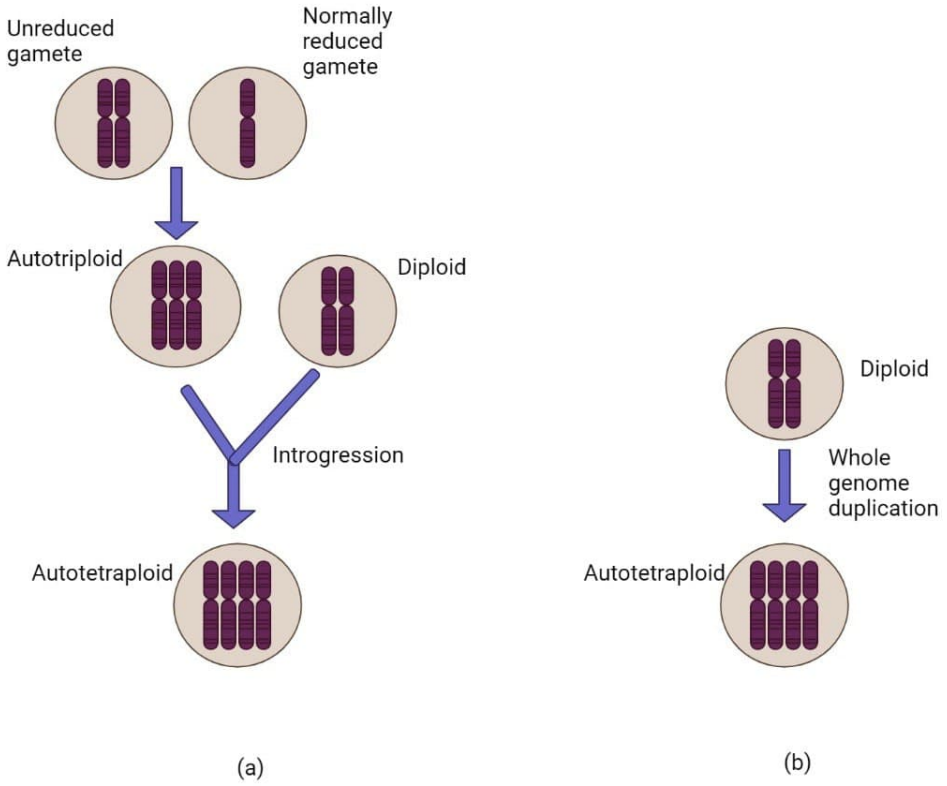


Figure 2.2: **An illustration that describes two evolutionary paths leading to the formation of autotetraploid: (a) Fusion of unreduced gamete (diploid) with a normally reduced gamete (haploid) within a single species. (b) Spontaneous whole genome duplication within a single species.**

Autopolyploidy happens by whole genome duplication within a zygote or fusion of unreduced gamete and normally reduced gamete, see Figure 2.2. Note that autopolyploids arise within or between populations of a single species. Thus, subgenomes in an autoploid have a high level of similarity, which leads to having a complete pairing in meiosis [55]. In comparison to their progenitors, autopolyploids show usually significantly different phenotypes such as morphological characteristics and ecological behaviors [49]. They usually possess a larger overall body mass as well as a bigger nucleus and cytoplasm. These phenotypes make the autoploids more favorable for agricultural purposes. Therefore, there has been a certain number of attempts over the past years to provide genetic improvement by focusing on varieties of them.

On the other hand, allopolyploidy sometimes occurs during the hybridization of different species followed by whole-genome doubling or fusion of unreduced gametes of different species. Allopolyploids consist of two or more distinct subgenomes from different evolutionary lineage [48]. These chromosomes differ in DNA sequence (allelic differences), gene order (chromosomal rearrangements), and their total numbers. Although the formation of both autopolyploids and allopolyploidy involves duplicated subgenomes, hybridization of different species in allopolyploids results in their separate behavior in chromosomal pairing during meiosis. In autopolyploids, multivalent pairing is generally formed during meiosis. However, in allopolyploids, bivalent chromosomal pairing within the same subgenomes is common.

Segmental allopolyploids contain more than two partially distinct subgenomes. In this case, some of the chromosomes within a set may be homologous with some of the chromosomes in another set or sets, but the rest of the chromosomes in the set are non-homologous or homeologous to those chromosomes in those sets. This type of polyploidy may occur through hybridization of two species followed by partial genome duplication, see Figure 2.3, and it covers all polyploids within the spectrum between autopolyploids and allopolyploids. The fact that segmental alloploids composed of more than two different subgenomes can result in formation of both bivalents and multivalents in chromosomal pairing [54].

2.3 The autoploid octoploid *S. officinarum*

Sugarcane is an indispensable crop. It is one the most cultivated crops in tonnage. In 2017, its cultivation even exceeds major grass species including rice, wheat, and maize[4]. It provides 80% of the world-wide sugar and it plays a pivotal role in bioenergy production. Bagasse, sugarcane biomass, and also bioethanol that comes from the sugar in sugarcane are the most efficient source of biofuels, great replacements for oil which contributes to reducing the greenhouse gas emissions and diminishing the related air pollution[5]. Developing genetic improvements is the central key to optimize its sugar yield and decent biomass for bio-energy production. This requires sufficient awareness of sugarcane genomics. Despite all the advancement in biotechnology, a thorough knowledge of the sugarcane genome is a significant barrier.

Sugarcane has one of the most complex genomes on the planet. High level of ploidy, heterozygosity, and the genome size of around > 10 Gb pose a great challenge in its sequencing and consequently a delay in the process of developing its genome reference in comparison with other grass species such as rice, maize, and sorghum[3].

The genus *Saccharum* includes *Saccharum spontaneum*, *Saccharum robustum*, *Saccharum officinarum*, *Saccharum barberi*, *Saccharum sinense*, and *Saccharum edule*.[1] The last three species are hybrid of two of the first three ones. *Saccharum officinarum*

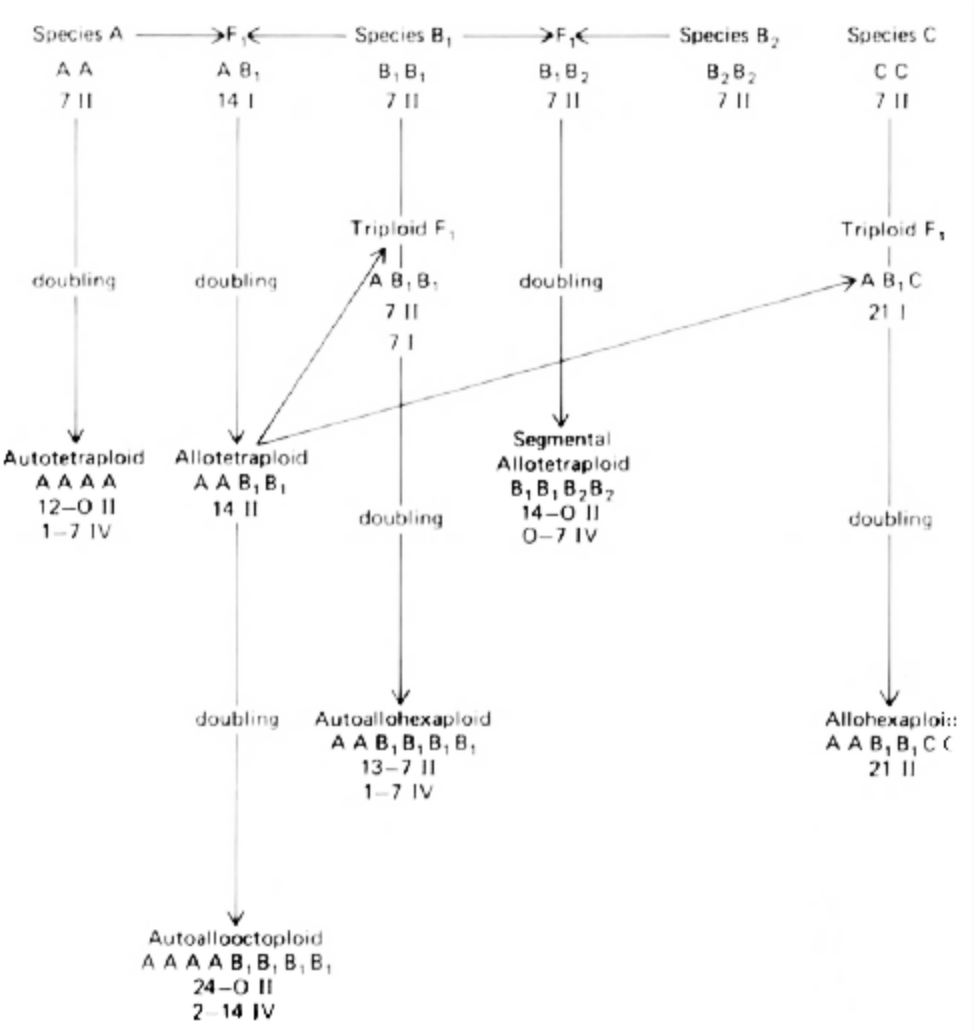


Figure 2.3: **Different types of polyploids**: autopolyploids, allopolyploids, segmental allopolyploids, and autoallopolyploids (from [55]).

is high in sugar content whereas *Saccharum spontaneum* is harder and more resistant to diseases. Thus, the modern cultivars of sugarcane are hybrid of these two varieties[2]. While, other *Saccharum* species have different levels of ploidy and different number of chromosomes in their monoploid chromosome set, *S. officinarum* accessions possess 10 chromosomes in its monoploid chromosome set and 8 level of ploidy; *S. officinarum* genome comprises 80 chromosomes overall.

In 2018, a team of CIRAD researchers led by Angélique D'Hont made use of the collinearity of sugarcane (*Saccharum spontaneum*) and sorghum genomes to produce its reference sequence [20]. In the same year, a team led by Ray Ming sequenced the *Saccharum spontaneum* [12], and recently they sequenced *Saccharum officinaum*, from which the data used in our study is drawn[14].

2.4 Consensus trees

In 1972, Edward N. Adams III drew attention to the need for some methods to amalgamate information in a set of phylogenetic trees and to represent it with a single tree, named a consensus tree [37]. He developed a method to compute the consensus tree for a collection of fully labelled and terminally-labelled trees, and later, he also showed that the consensus trees of his method preserve the nesting properties of the clusters[3] (see [39] for the proof). A straightforward version of the algorithm can be found in [40]. Adam's consensus tree is not necessarily a binary tree for a set of binary trees. Besides, it cannot be used in case of unrooted trees (see [38] for a proof), which makes it unsuitable for our study.

Since then, more methods has been developed to capture the agreements among the members of a set of phylogenetic trees. Among these methods, strict consensus method and majority rule are also well-known. These methods can be used for both rooted and unrooted trees. Strict consensus trees discard all the clusters/bipartitions which are not in all the input trees, and only keep those that exist in all the input trees [41]. Majority rule consensus trees keep all clusters/bipartitions that exist in more than half of the input trees [40]. It must be obvious that the clusters/bipartitions drawn with these methods are compatible and the result would be a tree. The problem is that these two methods do not necessarily give a binary tree as an output. Indeed, it is not rare to get a non-binary tree as an output even for a small set of binary trees.

Other methods are not as well-known as the three methods mentioned above, and it depends on the purpose of the study which method is the proper one. For instance, in [44], Nelson's consensus trees are built on the replicated components (clusters that appears in at least one of the input trees) and all the clusters that are compatible with the replicated components. The idea behind this approach is repeated clusters

[3]Clusters for rooted trees are analogous to bipartitions for unrooted ones.

could not be random. In [46], Neumann's method takes an intersection of maximal clusters of each input tree for each height. The goal of the Neumann's method is to retain as much information as possible. There are various versions of this method (interested readers can see [47, 45] for other methods.)

Up to now, all methods that have been developed construct a consensus tree on a set of phylogenetic trees of the same set of terminals. In 5, we describe our framework that is applicable to a set of phylogenetic trees so that the terminal sets of some trees can be proper subsets of largest one. The method used in this thesis bears a resemblance to the extention of majority rule consensus tree in [43, 42] where a greedy strategy is applied to extract the consensus tree. Note that all the bipartitions in a strict consensus tree would appear in a majority rule tree, and all the bipartitions in a majority rule consensus tree would show up in a greedy consensus tree. Perhaps greedy consensus trees are not as popular as majority rule due to the computational complexity or the possibilities of having ties (some bipartitions are compatible with all the bipartitions that already selected and they have the same frequency), but as we explain in chapter 5, our framework differs in that instead of counting frequencies of bipartitions, we sum over their weights. Thus, it is highly unlikely that we encounter ties.

Moreover, at first glance, this problem might appears as finding the maximum weight k-clique problem where employing other optimization algorithms may be preferable. However, it is better to focus on including the maximum weight bipartitions than on an objective function which maximizes the sum over the weights as in maximum weight k-clique problem. This is due to the fact that this approach guarantees an agreement of a larger number of input trees in the worse case. We explain this further in chapter 5.

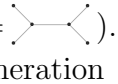
Chapter 3

A mathematical model for accumulation of subgenomes

In this section, we introduce and analyze a model for the accumulation of subgenomes in a polyploid. This model is built on full-binary unlabelled unordered trees.

3.1 Distribution of paired terminals

Consider a stochastic process defined on a full-binary free unlabelled unordered trees such that transition from one generation to another is by subdividing a randomly chosen edge and linking the new internal vertex to a new terminal vertex.

In this section, the quantity of interest is the number terminal pairs and the initial state is a tree on $n = 4$ terminals($b^4 =$ ). Let the random variables β_n be the number of terminal pairs in n^{th} generation with the probability distributions $\{Pr\{\beta_n = i\} := P_{2i,n}\ ; i = 2, 3, ..., \lfloor \frac{n}{2} \rfloor\}$. The distributions can be obtained through the following procedure.

Proposition 3.1.1. *For $i = 2, 3, ..., \lfloor \frac{n}{2} \rfloor$ and $n = 4, 5, ...$, the probability distributions $P_{2i,n}$ satisfy the following bivariate recurrence relation:*

$$P_{2i,n} = \begin{cases} 1, & \text{if } n = 4 \text{ and } i = 2 \\ 0, & \text{if } i < 2 \text{ or } \dfrac{n}{2} < i \\ \dfrac{(n+2i-4)}{2n-5} P_{2i,n-1} + \dfrac{(n-2i+1)}{2n-5} P_{2(i-1),n-1}, & \text{otherwise} \end{cases} \tag{3.1.1}$$

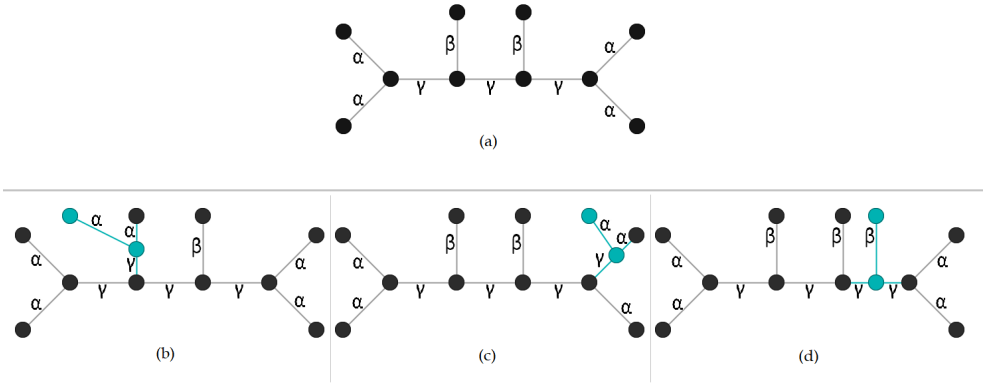


Figure 3.1: **An illustration to depict possible changes in cardinality of subsets A, B, and Γ as a result of subdividing an edge labelled by β, α, and γ and linking the new internal vertex to a new terminal vertex.** The binary tree in Figure (a) has $N = 7$ terminals and we have $|A| = 4$, $|B| = 2$, and $|\Gamma| = 3$. Figures (b) is the result of subdividing an edge labelled by β where the $|A|$ increases by 2 and $|B|$ decreases by 1, while Figure (c) and (d) are the result of subdividing an edge labelled by α and γ, respectively, the cardinality of subset A remains the same and the cardinality of subset B increased by 1. In all figures (b), (c). and (d), $|\Gamma|$ increases by 1.

Proof: For arbitrary $n = 4, 5, ...$, we partition edge set $E(b^n)$ of a tree b^n at n^{th} generation into three subsets A, B, and Γ corresponding to the set of paired terminals, non-paired terminals, and internal edges, respectively. We label every edge with one of α, β, and γ according to being a member of subsets A, B, or Γ , respectively. Suppose b^{n-1} at $(n-1)^{\text{th}}$ generation has $|A| = 2i$, $|B| = j$, and $|\Gamma| = k$ where $2i + j = n - 1$ and $k = (n-1) - 3$, then for b^n either of the following cases is possible:

- Subdivision occurs on an edge labeled by α or γ. Then b^n has $|A| = 2i$, $|B| = j+1$, and $|\Gamma| = k + 1$
- Subdivision occurs on an edge labeled by β. Then for b^n, we have $|A| = 2i + 2$, $|B| = j - 1$, and $|\Gamma| = k + 1$

This can be proved by induction. See Figure 3.1 for an example. A tree b^n with $2i$ paired terminals can be derived from a tree b^{n-1} with either $2i$ or $2i - 2$ paired terminals. Regardless of number of paired terminals, a tree b^{n-1} has $|E(b^{n-1})| = 2n - 5$ and $|\Gamma| = n - 4$. The probability of having a tree b^{n-1} with $2i$ paired terminals is $P_{2i,n-1}$. In order for this tree lead to a b^n with same number of paired terminals, an edge labeled by either α or γ from all the edges in b^{n-1} must be picked for subdivision and linking to a new terminal. Thus, the probability of this case is $\frac{2i+n-4}{2n-5} P_{2i,n-1}$. On the other hand, a tree b^{n-1} with $2i - 2$ paired terminals has $(n-1) - (2i-2)$ edges

labeled by β that can be picked to raise the number of paired terminals. Therefore the probability of this case is $\frac{n-2i+1}{2n-5}P_{2i-2,n-1}$. Overall, the probability of having $2i$ paired terminals for b^n is $\frac{n+2i-4}{2n-5}P_{2i,n-1} + \frac{n-2i+1}{2n-5}P_{2i-2,n-1}$.
A tree on n terminals has at least 2 and at most $\lfloor \frac{n}{2} \rfloor$ terminal pairs. Therefore, if $i < 2$ or $\frac{n}{2} < i$, we have $P_{2i,n} = 0$. At initiation of the process, the probability of having 2 terminal pairs in a tree on $n = 4$ terminals is one, since it is the only possible tree topology. This completes the proposition. ∎

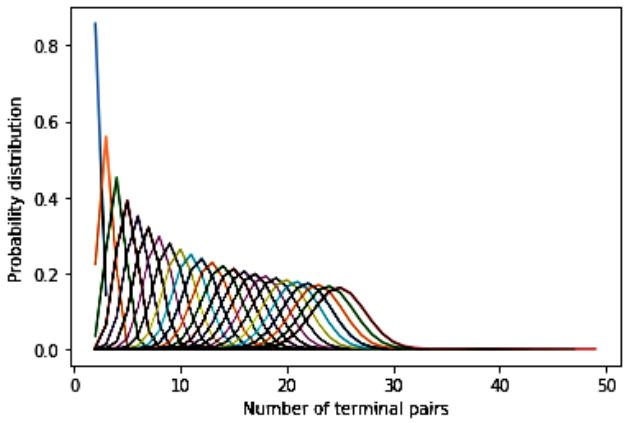


Figure 3.2: **Probabilities of number of terminal pairs in trees generated by the stochastic process of linking to a new terminal vertex.** Trees are from 6^{th} to 100^{th} generation with two gaps in between generations.

We computed the probabilities of the random variable β_n for the first 5,000 generations. The results, as shown in Figure 4.3, lead us to the following conjectures.

Conjecture 3.1.2. *Let $\bar{\beta}_n$ be the expected value of the number of terminal pairs in the n^{th} generation of the process of randomly subdividing an edge in order to link the new internal vertex to a new external vertex. Then,*

$$\frac{\bar{\beta}_n}{n} \to \frac{1}{4} \quad as \quad n \to \infty \tag{3.1.2}$$

Conjecture 3.1.3. *Let $\bar{\beta}_n$ and σ_n^2 be respectively the expected value and variance of the number of terminal pairs in the n^{th} generation of the stochastic process of randomly subdividing an edge in order to link the new internal vertex to a new external vertex. Then, based on the index of dispersion, the limiting distribution is under-dispersed.*

$$\frac{\sigma_n^2}{\bar{\beta}_n} \to \frac{1}{4} \quad as \quad n \to \infty \tag{3.1.3}$$

Although we can use the recurrence relation to compute the probability distributions exactly, finding its closed-form solution is challenging. In what follows, we explain our attempt regarding its closed-form solution.

3.1.1 On the closed-form solution

Let us first clarify the distributions mentioned in the Proposition 3.1.1. The recurrence relation for the probability distribution of $P_{2i,n}$ is equivalent to picking any weighted lattice path from $(4,4)$ to $(2i,n)$ while moving along an arrow to north or if it is possible to north-east as it is shown in the Figure 3.3. Each point $(2i,n)$ on the lattice corresponds to some tree b^n with $|A| = 2i$, $|B| = n-2i$, and $|\Gamma| = n-3$, and the probability of choosing a north or north-east step is $\frac{|A|+|\Gamma|}{|E(b^n)|}$ or $\frac{|B|}{|E(b^n)|}$, respectively.

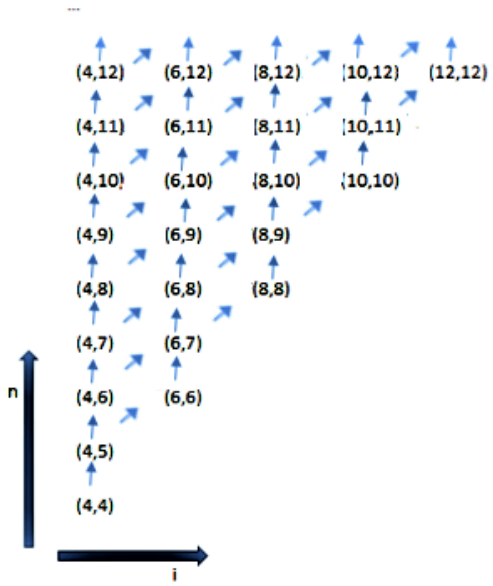


Figure 3.3: **Lattice paths of the recurrence relation in proposition 3.1.1**. Point $(2i,n)$ on the lattice corresponds to some tree at n^{th} generation with i terminal pairs.

A general approach to draw a closed-form solution of a recurrence relation is to rely on its generating function [9]. If we define $G(x,y) := \sum_{4\leq n} \sum_{2\leq i\leq \lfloor \frac{n}{2} \rfloor} P_{2i,n} x^{2i} y^n$ with boundary conditions $G(x,0) = 0$ and $G(0,y) = 0$, then by summing over two variables of the recurrence relation, we have:

$$\sum_{3\leq i} \sum_{2i+1\leq n} (2n-5) P_{2i,n} x^{2i} y^n = \sum_{3\leq i} \sum_{2i+1\leq n} (n+2i-4) P_{2i,n-1} x^{2i} y^n + \sum_{3\leq i} \sum_{2i+1\leq n} (n-2i+1) P_{2i-2,n-1} x^{2i} y^n \tag{3.1.4}$$

For the LHS of equation 3.1.1, we have:

$$\sum_{3\le i}\sum_{2i+1\le n}(2n-5)P_{2i,n}x^{2i}y^n = \sum_{2\le i}\sum_{2i\le n}(2n-5)P_{2i,n}x^{2i}y^n - [\sum_{5\le n}(2n-5)P_{4,n}x^4y^n + \sum_{3\le i}(4i-5)P_{2i,2i}x^{2i}y^{2i} + 3P_{4,4}x^4y^4] \tag{3.1.5}$$

where

$$\begin{aligned}\sum_{2\le i}\sum_{2i\le n}(2n-5)P_{2i,n}x^{2i}y^n &= 2y\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}nP_{2i,n}x^{2i}y^{n-1} - 5\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}P_{2i,n}x^{2i}y^n \\ &= 2y\frac{\partial G(x,y)}{\partial y} - 5G(x,y).\end{aligned} \tag{3.1.6}$$

For the first term on the RHS of equation 3.1.1, we have:

$$\sum_{3\le i}\sum_{2i+1\le n}(n+2i-4)P_{2i,n-1}x^{2i}y^n = \sum_{2\le i}\sum_{2i+1\le n}(n+2i-4)P_{2i,n-1}x^{2i}y^n - \sum_{5\le n}nP_{4,n-1}x^4y^n \tag{3.1.7}$$

where

$$\begin{aligned}&\sum_{2\le i}\sum_{2i+1\le n}(n+2i-4)P_{2i,n-1}x^{2i}y^n \\ &= \sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i+n-4)P_{2i,n-1}x^{2i}y^n \\ &= xy\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i)P_{2i,n-1}x^{2i-1}y^{n-1} + y^2\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(n-1)P_{2i,n-1}x^{2i}y^{n-2} \\ &\qquad - 3y\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}P_{2i,n-1}x^{2i}y^{n-1} \\ &= xy\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i)P_{2i,n}x^{2i-1}y^n + xy\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i)P_{2i,3}x^{2i}y^3 \\ &\qquad + y^2\sum_{3\le n-1}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(n-1)P_{2i,n-1}x^{2i}y^{n-2} - 3yG(x,y) \\ &= xy\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i)P_{2i,n}x^{2i-1}y^n + y^2\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}nP_{2i,n}x^{2i}y^{n-1} - 3yG(x,y) \\ &= xy\frac{\partial G(x,y)}{\partial x} + y^2\frac{\partial G(x,y)}{\partial y} - 3yG(x,y).\end{aligned} \tag{3.1.8}$$

And for the second term on the RHS of equation 3.1.1, we have:

$$\sum_{3\le i}\sum_{2i+1\le n}(n-2i+1)P_{2i-2,n-1}x^{2i}y^n = \sum_{3\le i}\sum_{2i\le n}(n-2i+1)P_{2i-2,n-1}x^{2i}y^n - \sum_{3\le i}P_{2i-2,2i-1}x^{2i}y^{2i} \tag{3.1.9}$$

Where

$$\begin{aligned}\sum_{3\le i}\sum_{2i\le n}(n-2i+1)P_{2i-2,n-1}x^{2i}y^n &= \sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(n-1)P_{2i-2,n-1}x^{2i}y^n - \sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i-2)P_{2i-2,n-1}x^{2i}y^n\\ &= x^2y^2\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(n-1)P_{2i-2,n-1}x^{2i-2}y^{n-2}\\ &\qquad - x^3y\sum_{4\le n}\sum_{2\le i\le\lfloor\frac{n}{2}\rfloor}(2i-2)P_{2i-2,n-1}x^{2i-3}y^{n-1}\\ &= x^2y^2\frac{\partial G(x,y)}{\partial y} - x^3y\frac{\partial G(x,y)}{\partial x}\end{aligned} \tag{3.1.10}$$

From the recurrence relation formula, we have:

$$\sum_{5\le n}[(2n-5)P_{4,n} - nP_{4,n-1}]x^4y^n + \sum_{3\le i}[(4i-5)P_{2i,2i} - P_{2i-2,2i-1}]x^{2i}y^{2i} = 0 \tag{3.1.11}$$

Now overall, we have:

$$\begin{aligned}2y\frac{\partial G(x,y)}{\partial y} - 5G(x,y) - 3x^4y^4 = xy\frac{\partial G(x,y)}{\partial x} + y^2\frac{\partial G(x,y)}{\partial y} - 3yG(x,y)\\ + x^2y^2\frac{\partial G(x,y)}{\partial y} - x^3y\frac{\partial G(x,y)}{\partial x}\end{aligned} \tag{3.1.12}$$

Simplifying this, we get the following partial differential equation:

$$(xy - x^3y)G_x(x,y) + (y^2 + x^2y^2 - 2y)G_y(x,y) = (3y-5)G(x,y) - 3x^4y^4 \tag{3.1.13}$$

Alternatively, we can also define the generating function of the recurrence relation with $f(x,y) := \sum_{2\le i}\sum_{2i\le n} P_{2i,n}x^{n+2i-8}y^{n-2i+1}$ with boundary conditions $f(x,0) = 0$ and $f(0,y) = y$. A similar procedure leads to the following partial differential equation.

$$(x^2y^2 - xy)f_x(x,y) + (x^3y - y^2)f_y(x,y) = (2y + x^3 - 5xy^2)f(x,y) - 3y^2 \tag{3.1.14}$$

Chapter 4

On small peripheral structures

In the previous chapter, we discussed the probability distributions of the number of terminal pairs in the trees generated by our one-branch-at-time model. This chapter presents the distribution of other peripheral structures including the number of triples and two types of quadruples.

4.1 Distribution of triples

In this section, the quantity of interest is the occurrence of a substructure, which we refer to as a triple, consisting of a pair of terminals and an unpaired terminal in addition to two internal edges, as shown in Figure 4.1.

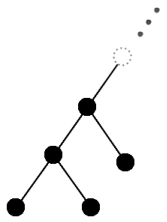


Figure 4.1: **The layout of a triple.**

Let τ_n be the random variable defined by the number of triples in n^{th} generation of our one-branch-at-a-time model with probability distributions $\{Pr\{\tau_n = m\}; m = 0, 1, ..., \lfloor \frac{n}{3} \rfloor\}$. By marginalizing over the number of terminal pairs, we have:

$$Pr\{\tau_n = m\} = \sum_{k=max(m,2)}^{\lfloor \frac{n}{2} \rfloor} Pr\{\tau_n = m, \beta_n = k\} \tag{4.1.1}$$

We denote $Pr\{\tau_n = m, \beta_n = k\}$ by $P_{m,k,n}$. Through the following procedure, we obtain a recurrence relation for the the number of triples:

Proposition 4.1.1. *For $0 \leq m \leq min(k, \lfloor \frac{n}{3} \rfloor)$ and $2 \leq k \leq \lfloor \frac{n}{2} \rfloor$, the following recurrence relation holds for the distribution $P_{m,k,n}$:*

$$P_{m,k,n} = \frac{n-4+3m-k}{2n-5} P_{m,k,n-1} + \frac{3k-3m+3}{2n-5} P_{m-1,k,n-1} + \frac{n-2k-m+1}{2n-5} P_{m,k-1,n-1} + \frac{m+1}{2n-5} P_{m+1,k-1,n-1} \tag{4.1.2}$$

Without loss of generality, we let the initial condition be the probabilities for $n = 6$: $P_{0,2,6} = \frac{1}{7}$, $P_{1,2,6} = 0$, and $P_{2,2,6} = \frac{6}{7}$.

Proof: Assume the tree b^n at n^{th} generation contains m triples and k terminal pairs. This tree can be derived from a tree b^{n-1} with $m+1$, m or $m-1$ triples and with k or $k-1$ terminal pairs.

If the tree b^{n-1} also had k terminal pairs, then either a paired terminal α or an internal edge γ has been chosen for the subdivision. In this case, the tree b^{n-1} could not have $m+1$ triples, since the number of triples decreases only if a non-paired terminal β of a triple is picked for subdivision. If the tree b^{n-1} had $m-1$ triples, then there are $k-(m-1)$ substructures, each of which consists of a terminal pair α in addition to an adjacent internal edge γ such that the substructure is not a subgraph of any triples in the tree. There are $3(k-(m-1))$ edges involved in these substructures. Hence, the probability of this case is $\frac{3k-3m+3}{2n-5} P_{m-1,k,n-1}$. However, if the tree b^{n-1} had the same number of m triples as the current tree b^n, then any edge of the tree b^{n-1} could have been picked unless through its subdivision and then linking to a new external vertex, there would be an increase or a decrease in the number of triples. Hence, there would be $(2n-5) - \Big(3(k-m) + \big((n-1)-2k\big)\Big)$ edges which can be picked in order that the number of triples remains the same. Thus, the probability of this case is $\frac{n-4+3m-k}{2n-5} P_{m,k,n-1}$.

On the other hand, if the tree b^{n-1} had $k-1$ terminal pairs, then a non-paired terminal β was picked for subdivision. In this case, the tree b^{n-1} could not have $m-1$ triples, since the number of triples increases only if the subdivision occurs on any of three edges forming a substructure, consisting of a terminal pair α and the adjacent internal edges γ, which is not a subgraph of a triple. If the tree b^{n-1} had $m+1$ triples, then a non-paired terminal of a triple was picked for subdivision. Therefore, the probability of this case is $\frac{m+1}{2n-5} P_{m+1,k-1,n-1}$. However, if the tree b^{n-1} had the same number of m triples, then there are $(n-1) - \Big(2(k-1)+m\Big)$ edges can be picked so that the number of triples does not change. Hence, the probability of this

case is $\frac{n-2k-m+1}{2n-5} P_{m,k-1,n-1}$. This completes the proposition. ∎

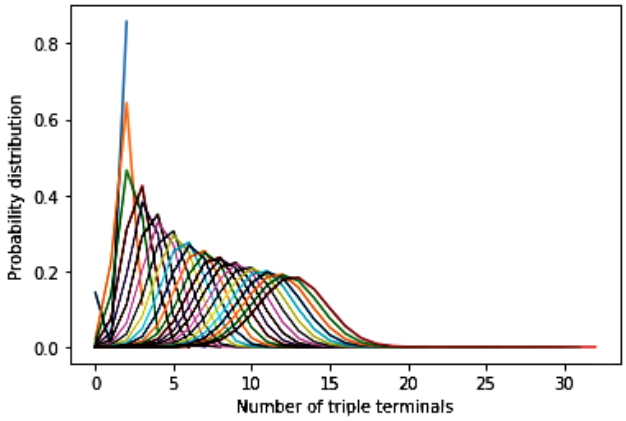


Figure 4.2: **Probabilities of number of triples in trees generated by the stochastic process of linking to a new terminal vertex.** Trees are from 6^{th} to 100^{th} generation with three gaps in between generations.

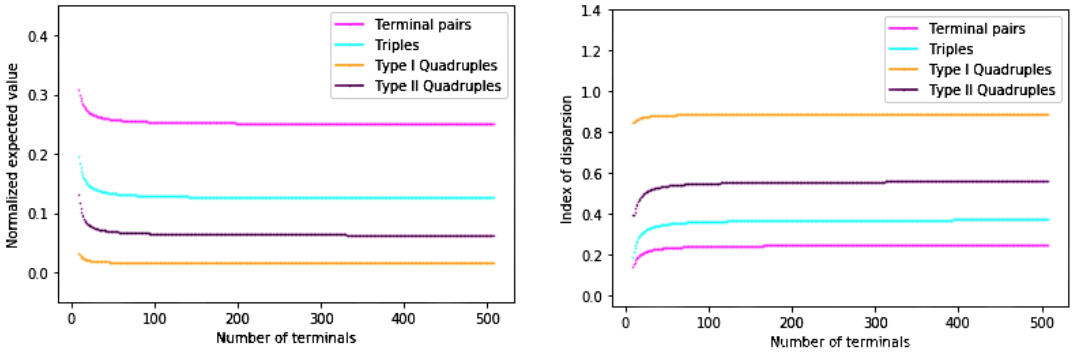


Figure 4.3: **(a) Normalized expected value of number of terminal pairs, triples, and quadruples. (b) Indices of dispersion of number of terminal pairs, triples, and quadruples.**

We computed the probabilities of the random variable τ_n for the first 1000 generations. The results lead us to the following conjectures.

Conjecture 4.1.2. *Let $\bar{\tau_n}$ be the expected value of the number of triples in the n^{th} generation of the process of randomly subdividing an edge so as to link the new internal*

vertex to a new external vertex. Then,

$$\frac{\bar{\tau_n}}{n} \to \frac{1}{8} \quad as \quad n \to \infty \tag{4.1.3}$$

Conjecture 4.1.3. *Let $\bar{\tau_n}$ and σ_n^2 be respectively the expected value and variance of the number of triples in the n^{th} generation of the stochastic process of subdividing a random edge in order to link the new internal vertex to a new external vertex. Then, based on the index of dispersion, the limiting distribution is under-dispersed.*

$$\frac{\sigma_n^2}{\bar{\tau_n}} \to \frac{3}{8} \quad as \quad n \to \infty \tag{4.1.4}$$

4.2 Distribution of quadruples

4.2.1 Type I quadruples

In what follows, we analyze the distribution of another peripheral substructure of the trees. This substructure which we refer to as a type I quadruple comprises of a pair of terminal pairs descending from a single ancestor along with two internal edges which are adjacent to the ancestor and the pairs' ancestral nodes, as shown in Figure 4.4.

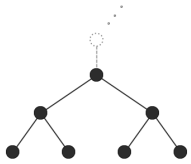


Figure 4.4: **The layout of a type I quadruple.**

Let ϕ_n be the random variable defined by the number of type I quadruples in n^{th} generation with probability distributions $\{Pr\{\phi_n = f\}; f = 0, 1, ..., \lfloor \frac{n}{4} \rfloor\}$. By marginalizing over the number of terminal pairs and of triples, we have:

$$Pr\{\phi_n = f\} = \sum_{m=0}^{\lfloor \frac{n}{3} \rfloor} \sum_{k=max(m,2)}^{\lfloor \frac{n}{2} \rfloor} Pr\{\phi_n = f, \tau_n = m, \beta_n = k\} \tag{4.2.1}$$

We denote $Pr\{\phi_n = f, \tau_n = m, \beta_n = k\}$ by $P_{f,m,k,n}$. The following procedure yields a recurrence relation for the distribution $P_{f,m,k,n}$.

Proposition 4.2.1. *For $0 \leq f \leq min(\lfloor \frac{k}{2} \rfloor, \lfloor \frac{n}{4} \rfloor)$, $0 \leq m \leq min(k, \lfloor \frac{n}{3} \rfloor)$, and $0 \leq k \leq \lfloor \frac{n}{2} \rfloor$, the following recurrence relation holds for the distribution $P_{f,m,k,n}$:*

$$P_{f,m,k,n} = \frac{m+1}{2n-5}P_{f-1,m+1,k-1,n-1} + \frac{6f+6}{2n-5}P_{f+1,m-1,k,n-1} + \frac{3k-6f-3m+3}{2n-5}P_{f,m-1,k,n-1} + \frac{n-4-k+3m}{2n-5}P_{f,m,k,n-1} + \frac{n-2k-m+1}{2n-5}P_{f,m,k-1,n-1} \tag{4.2.2}$$

Without loss of generality, let the initial conditions be the probabilities for $n = 7$: $P_{0,2,2,7} = \frac{42}{63}$, $P_{1,1,3,7} = \frac{21}{63}$, otherwise $P_{f,m,k,7} = 0$, for $n = 7$.

Proof: Consider a tree b^n containing f type I quadruples, m triples, and k terminal pairs. A tree b^n having f type I quadruples can be derived from a tree b^{n-1} with $f-1$, f, or $f+1$ type I quadruples.

The number of type I quadruples increases only if the subdivision occurs on a non-paired terminal β of a triple. As a result, an increment of the number of type I quadruples is always accompanied by a reduction of the number of triples and by an increment of the number of terminal pairs. Therefore, if the tree b^{n-1} had $f-1$ type I quadruples, it must have had $m+1$ triples and $k-1$ terminal pairs, and the subdivision must have occurred on a non-paired terminal β of a triple. The probability of this case is $\frac{m+1}{2n-5}P_{f-1,m+1,k-1,n-1}$.

On the other hand, If the tree b^{n-1} had $f+1$ type I quadruples, then each of the six edges involved in any of the type I quadruples of the tree b^{n-1} could have been picked for the subdivision; thus, the probability of this is $\frac{6(f+1)}{2n-5}P_{f+1,m-1,k,n-1}$.

Finally, if the tree b^{n-1} embraced f type I quadruples, then it must contain either $m-1$ or m triples. Provided that the tree b^{n-1} had $m-1$ triples, the subdivision must have occurred on any of three edges involved in a substructure consisting of a terminal pair and their adjacent internal edge such that the substructure is not a subgraph of any triples or of any type I quadruples in b^{n-1}. In this case, there are $k-2f-(m-1)$ of such substructures; therefore, with a probability of $\frac{3k-6f-3m+3}{2n-5}P_{f,m-1,k,n-1}$, the tree b^{n-1} give rise to the traced tree b^n. On the other hand, if the tree b^{n-1} also had m triples, then the number of terminal pairs either remained the same or increased through the transition. If the number of terminal pairs has remained the same as k terminal pairs, then there exist $(2n-5) - \Big(\big((n-1)-2k\big) + 3(k-m)\Big)$ edges, none of which is a non-paired terminal β or is one of the three edges involved in the substructures on which the occurrence of subdivision contributes to an increment of number of triples; hence, the probability of this case is $\frac{n-4-k+3m}{2n-5}P_{f,m,k,n-1}$. On the other hand, if the number of terminal pairs has been increased by the transition, then the subdivision must have occurred on an non-paired terminal which is not involved in the construction of any triple. The number of such edges is $(n-1)-\big(2(k-1)+m\big)$. Therefore, the probability of this case is $\frac{n-2k-m+1}{2n-5}P_{f,m,k-1,n-1}$. This completes the proof of the proposition. ∎

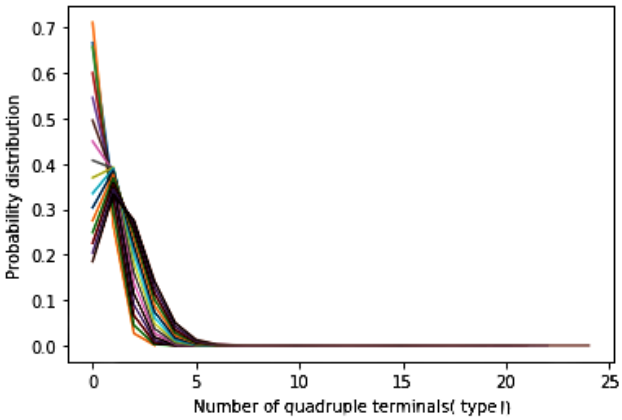


Figure 4.5: **Probabilities of number of type I quadruples in trees generated by the stochastic process of linking to a new terminal vertex.** Trees are from 7^{th} to 100^{th} generation with three gaps in between generations.

We computed the probabilities of the random variable ϕ_n for the first 500 generations. The results lead us to the following conjectures.

Conjecture 4.2.2. *Let $\bar{\phi_n}$ be the expected value of type I quadruples in the n^{th} generation of the process of randomly subdividing an edge so as to link the new internal vertex to a new external vertex. Then,*

$$\frac{\bar{\phi_n}}{n} \to \frac{1}{64} \quad as \quad n \to \infty \tag{4.2.3}$$

Conjecture 4.2.3. *Let $\bar{\phi_n}$ and σ_n^2 be the expected value and variance of the number of type I quadruples in the n^{th} generation of the stochastic process of randomly subdividing an edge in order to link the new internal vertex to a new external vertex. Then, based on the index of dispersion, the limiting distribution is under-dispersed.*

$$\frac{\sigma_n^2}{\bar{\phi_n}} \to \frac{57}{64} \quad as \quad n \to \infty \tag{4.2.4}$$

4.2.2 Type II quadruples

In this section, we shed light on the distribution of another type of quadruples consisting of a terminal pair, and two non-paired terminals in addition to three internal edges, one of which is adjacent to the terminal pair and one of the non-paired terminal, and the other two internal edges are adjacent to the non-paired terminals in consecutive order. as shown in Figure 4.6. We refer to this substructure as a type II quadruple.

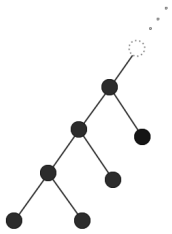


Figure 4.6: **The layout of a type II quadruple.**

Let θ_n be the random variable of the number of type II quadruples in n^{th} generation with probability distributions $\{Pr\{\theta_n = e\}; e = 0, 1, ..., \lfloor \frac{n}{4} \rfloor\}$. By marginalizing over the number of terminal pairs and of triples, we have:

$$Pr\{\theta_n = e\} = \sum_{m=0}^{\lfloor \frac{n}{3} \rfloor} \sum_{k=max(m,2)}^{\lfloor \frac{n}{2} \rfloor} Pr\{\theta_n = e, \tau_n = m, \beta_n = k\} \tag{4.2.5}$$

We denote $Pr\{\theta_n = e, \tau_n = m, \beta_n = k\}$ by $P_{e,m,k,n}$.

Proposition 4.2.4. *For $0 \leq e \leq min(k, \lfloor \frac{n}{4} \rfloor)$, $0 \leq m \leq min(k, \lfloor \frac{n}{3} \rfloor)$, and $0 \leq k \leq \lfloor \frac{n}{2} \rfloor$, the following recurrence relation holds for the distribution $P_{e,m,k,n}$:*

$$\begin{aligned} P_{e,m,k,n} =& \frac{4m - 4e + 4}{2n - 5} P_{e-1,m,k,n-1} + \frac{e+1}{2n-5} P_{e+1,m,k-1,n-1} + \frac{e+1}{2n-5} P_{e+1,m+1,k-1,n-1} \\ &+ \frac{m+1-e}{2n-5} P_{e,m+1,k-1,n-1} + \frac{3k - 3m + 3}{2n-5} P_{e,m-1,k,n-1} + \frac{n-k-m+4e-4}{2n-5} P_{e,m,k,n-1} \\ &+ \frac{n-2k-m+1-e}{2n-5} P_{e,m,k-1,n-1} \end{aligned} \tag{4.2.6}$$

Without loss of generality, let the initial conditions be the probabilities for $n = 8$: $P_{0,0,4,8} = \frac{21}{693}$, $P_{0,2,3,8} = \frac{168}{693}$, $P_{1,1,3,8} = \frac{168}{693}$, $P_{2,2,2,8} = \frac{336}{693}$, otherwise $P_{e,m,k,8} = 0$, for $n = 8$,

Proof: Suppose the tree b^n contains e type II quadruples, m triples, and k terminal pairs. A tree b^n with e type II quadruples can be derived from a tree b^{n-1} with $e-1$, $e+1$, or e type II quadruples.

The tree b^n can rise from a tree b^{n-1} with $e-1$ type II quadruples only if a triple which is not a subgraph of a type II quadruple contributes to the formation of a type II quadruple. In other words, the tree b^{n-1} must also have had m triples and k terminal pairs, and the subdivision must have occurred on any of four edges,

consisting of a terminal pair and two internal edges, involved in a triple which is not a subgraph of a type II quadruple; there are $4\big(m-(e-1)\big)$ edges which can be subdivided such that by linking the new internal vertex to a new external vertex, the number of type II quadruples would increase; therefore, the probability of this case is $\frac{4m-4e+4}{2n-5}P_{e-1,m,k,n-1}$.

On the other hand, if the tree b^{n-1} had $e+1$ type II quadruples, the subdivision must have occurred on either of the two non-pair terminals β of a type II quadruple, and therefore, the tree b^{n-1} must have had $k-1$ terminal pairs. If the subdivision has occurred on the closest non-paired terminal to the terminal pairs of a type II quadruple, then the tree b^{n-1} must have had $m+1$ triples. Thus, the probability of this case is $\frac{e+1}{2n-5}P_{e+1,m+1,k-1,n-1}$. Otherwise, if the subdivision has occurred on the second closest non-paired terminal to the terminal pairs, then the tree b^{n-1} must have had m triples; thus, the probability of this case is $\frac{e+1}{2n-5}P_{e+1,m,k-1,n-1}$.

Finally, suppose the number of type II quadruples has not been altered through the transition. The tree b^{n-1} could have had $m+1$, $m-1$, or m triples. If b^{n-1} had $m+1$ triples, it must have had $k-1$ terminal pairs, and the subdivision must have occurred on a triple which is not a subgraph of a type II quadruples; the probability of this case is $\frac{m+1-e}{2n-5}P_{e,m+1,k-1,n-1}$. Alternatively, if the tree b^{n-1} had $m-1$ triples, then there are $k-(m-1)$ substructures, each of which are consist of a terminal pair in addition to an adjacent internal edge such that the substructure is not a subgraph of any triples in the tree. There are $3(k-(m-1))$ edges included in these structures that can be chosen for subdivision. Hence, the probability of this case is $\frac{3k-3m+3}{2n-5}P_{e,m-1,k,n-1}$. Otherwise, if the tree b^{n-1} had m triples, then, based on its number of terminal pairs, there are two possible cases. If the tree b^{n-1} also had k terminal pairs, then there are $(2n-5)-\Big(\big((n-1)-2k\big)+3(k-m)+4(m-e)\Big)$ edges which can be picked in order that the number of terminal pairs, triples, and type II quadruples remain the same. Therefore, the probability of this case is $\frac{n-k-m+4e-4}{2n-5}P_{e,m,k,n-1}$. However, if the tree b^{n-1} had $k-1$ terminal pairs, then a non-paired terminal which is not involved in formation of a triple or of a type II quadruple must have been subdivided to link to a new external vertex. There are $(n-1)-\Big(2(k-1)+2e+(m-e)\Big)$ of such edges. Hence, the probability of this case is $\frac{n-2k-m+1-e}{2n-5}P_{e,m,k-1,n-1}$. This completes the proof of the proposition. ∎

The probabilities of the random variable θ_n for the first 500 generations are computed. The results lead us to the following conjectures.

Conjecture 4.2.5. *Let $\bar{\theta_n}$ be the expected value of type II quadruples in the n^{th} generation of the process of randomly subdividing an edge so as to link the new internal*

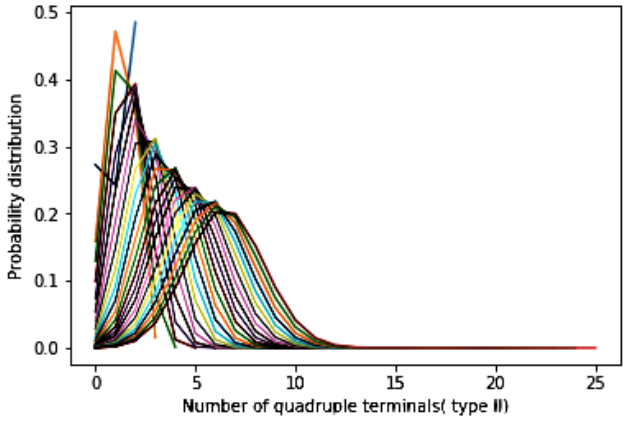


Figure 4.7: **Probabilities of number of type II quadruples in trees generated by the stochastic process of linking to a new terminal vertex.** Trees are from 8^{th} to 100^{th} generation with three gaps in between generations.

vertex to a external new vertex. Then,

$$\frac{\bar{\theta_n}}{n} \to \frac{1}{16} \quad as \quad n \to \infty \tag{4.2.7}$$

Conjecture 4.2.6. *Let $\bar{\theta_n}$ and σ_n^2 be respectively the expected value and variance of type II quadruples in the n^{th} generation of the process of randomly subdividing an edge so as to link the new internal vertex to a new external vertex. Then, the indices of dispersion shows that for the limiting distribution, we have:*

$$\frac{\sigma_n^2}{\bar{\theta_n}} \to \frac{9}{16} \quad as \quad n \to \infty \tag{4.2.8}$$

Chapter 5

Looking for evidence of polyploidization events in *S. officinarum*

The preceding combinatorial analyses were carried out in the service of a null hypothesis that each increment in genome ploidy is the result of the addition of a complete complement of distinct chromosomes to the $n-1$ copies already present, and that this addition is described by the attachment of a new edge to any branch of the existing phylogeny with equal probabilities.

5.1 The framework

Here we illustrate how to adapt our analysis to study the evoluton of a variety of sugarcane, *Saccharum officinarum*, (L.A. Purple) [14]. The genome has 80 chromosomes in all, partitioned into 10 sets C_i ($1 \leq i \leq \xi = 10$), each of which consists of 8 homeologous chromosomes. There is no given or known subgenome structure, i.e., for any choice of 10 chromosomes, one from each set of 8 homeologs, there is no information whether these chromosomes were, or were not, added to the genome during the same polyploidization event.

We first calculated the sequence similarities (percent identity) in the set of all homologous gene pairs found by SynMap in syntenic blocks [15, 13].

For any labelling of the 8 chromosomes in each set of homeologs, say A,B,...,H, it would thus be futile to try to consider the 10 phylogenies constructed from the 10 sets as samples from some distribution of trees with labelled terminals, since there is no known one to one correspondence between the chromosomes in different sets. On the

other hand, were there some well-supported, well-resolved phylogeny in common in some or all of the sets, though these consist of unlabelled trees, we might be justified in concluding that this gene tree represented the evolutionary history of the genome, solely on the basis of a common topology. Thus we will analyze each of the 10 sets of homeologous chromosomes, determine the tree that best fits the data on that set, and finally compare the 10 phylogenies, without labels, to see to what extent they share common structural features.

For each set of homeologous chromosomes, we first compile all gene families, sets of 8 paralogous genes one on each chromosome, that have no other paralogous links within the set or elsewhere. Because of the extensive pattern of missing genes and ambiguous paralogy, however, there are actually few complete gene families with 8 replicates. To address this problem, we extend our study to gene families with only 7 or 6 replicates. The numbers of families of each size are presented in Table 5.1.

Table 5.1: **The number of extracted gene families of homeologous chromosomes in *Saccharum officinarum*.**

Homeology set	8-genes	7-genes	6-genes	Total
1	0	3	15	18
2	0	2	36	38
3	0	7	23	30
4	0	8	43	51
5	0	6	12	18
6	0	5	28	33
7	0	1	16	17
8	5	6	17	28
9	1	6	24	31
10	0	3	19	22
Total	6	47	233	286

Since our phylogenetic reconstruction in each gene family is based on bipartitions shared by the largest number of gene trees, we pay careful attention to the statistically fair weighting of bipartitions coming from the 6,7 and 8-replicate cases.

For each gene family, we calculate its maximum likelihood phylogeny via the

RAxML wrapper contained in the Python package Dendropy. We carry out the rest of our analysis based on the non-trivial bipartitions of these phylogenies.

We recall that the compatibility of a pair of bipartitions means that the two bipartitions can co-exist in a tree. For instance, consider two bipartitions $d^1 = \{X, R\setminus X\}$ and $d^2 = \{Y, R\setminus Y\}$,where R is the set of all the leaves' labels. The two bipartitions are compatible if and only if at least one of the intersections $X \cap Y$, $X \cap (R \setminus Y)$, $(R\setminus X)\cap Y$, and $(R\setminus X)\cap(R\setminus Y)$ is empty [36]. Constructing a tree on eight terminals entails inferring five compatible non-trivial bipartitions. While we are combining evidence from all the gene families in a homeology set, however, we should weight more heavily bipartitions that discriminate among the set of possible trees than those bipartitions that are compatible with a larger set of trees. Thus {A B C D}{E F G H}, which is compatible with only 225 unrooted binary branching trees, should be weighted more heavily than {A B}{C D E F G H}, which is compatible with 945 trees.

Moreover, bipartitions derived from 7-gene families and 6-gene families will be compatible with (i.e. constitute evidence for) many more 8-terminal trees, and should thus carry much less weight in constructing the "consensus" tree of all the gene families in the homeology set.

5.1.1 Weighting the compatible bipartitions

For each homeologous chromosome set C_i, let $m^{(i)}$ be the number of gene families that has been extracted. Let phylogenetic tree T^{ij} corresponds to gene family G^{ij} where $1 \leq i \leq \xi = 10$ and $1 \leq j \leq m^{(i)}$. Suppose R is the set of labels of chromosomes of n replicates(here, $R = \{A, B, ..., H\}$), $b^{ij} = \{b^{ij}_1, ..., b^{ij}_{2n'-3}\}$ is a bipartition set of T^{ij} where $n' \leq n = 8$, for some $1 \leq i \leq \xi$ and $1 \leq j \leq m^{(i)}$. Let R' be the set of missing terminal labels of T^{ij} , and $d^{ij} = \{d^{ij}_1, ..., d^{ij}_{\rho_j}\}$ be the new non-trivial bipartition set consisting of all non-trivial bipartitions of n terminals compatible with bipartitions in b^{ij}. Now, we assign weight ω^{ij}_k to d^{ij}_k where $d^{ij}_k = \{S^{ij}_k, R \setminus S^{ij}_k\}$ for some $S^{ij}_k \subset R$, for $1 \leq k \leq \rho_j$, such that:

1. The new set of compatible non-trivial bipartitions must support all possible bipartitions fairly. We provide this balance through the following criteria:

$$\omega^{ij}_{k_1}\Lambda^{ij}_{k_1} = \omega^{ij}_{k_2}\Lambda^{ij}_{k_2} \qquad \forall 1 \leq k_1 \neq k_2 \leq \rho_j \tag{5.1.1}$$

   where Λ^{ij}_k is the number of all full-binary trees with n' terminals with which d^{ij}_k is compatible. That is the number of all full-binary trees with n' terminals compatible with bipartition $\{S^{ij}_k \setminus (S^{ij}_k \cap R'), R\setminus(R\cap(S^{ij}_k \cup R'))\}$, for $1 \leq k \leq \rho_j$. More explanation on this and counting Λ^{ij}_k is provided later.

2. The new set of bipartitions d^{ij} must contribute a total weight of 1 to the formation of the consensus tree; thus, we let

$$\sum_{k=1}^{\rho} \lambda_k^{ij} \omega_k^{ij} = 1 \tag{5.1.2}$$

where λ_k^{ij} is the number of all bipartitions in the set d^{ij} such that there is same number of Λ_k^{ij} full-binary trees compatible with them.

Hence, we have a system of ρ equations involving ρ weights that can be easily solved to get the weights.

For each d_k^{ij}, we count Λ_k^{ij} through the following procedure. Note that the bipartition d_k^{ij} arises from the bipartition $b_{k'}^{ij} = \{S_k^{ij} \setminus (S_k^{ij} \cap R'), R \setminus (R \cap (S_k^{ij} \cup R'))\}$, for some k', on n' terminals corresponding to the phylogenetic tree of the gene family G^{ij}. The bipartition d_k^{ij} can arise from all the binary trees with co-existing bipartition $b_{k'}^{ij}$. We know Λ_k^{ij} is the number of all these trees. Now, suppose $b_{k'}^{ij} = \{R_1, R_2\}$ where $|R_1 \cup R_2| = n'$. Then, Λ_k^{ij} is

$$\frac{(2|R_1| - 3)\,!(2|R_2| - 3)\,!}{2^{(|R_1|+|R_2|-4)}(|R_1| - 2)\,!(|R_2| - 2)\,!} \tag{5.1.3}$$

The bipartiton $b_{k'}^{ij}$ corresponds to an edge linking R_1 and R_2. Therefore, the number of all full-binary trees compatible with bipartition $b_{k'}^{ij}$ is product of the number of full-binary terminally labelled trees with labels in R_1 and with labels in R_2. Each of these two trees is rooted at one of the two incident vertices to the edge linking R_1 and R_2. There are $\frac{(2n-3)\,!}{2^{n-2}(n-2)\,!}$ full-binary rooted labelled trees with n terminals. See Table 5.2 for an example.

5.1.2 Consensus trees

After assigning weights to each new non-trivial compatible bipartition, let $D^i = \{d_1^i, ..., d_{l^i}^i\}$ be the set of all distinct bipartitions appearing in at least one of the new non-trivial compatible bipartition sets of the i^{th} homeologous chromosome set. Note that l^i is at most equal to the number of all possible non-trivial bipartition on a set of n labels (e.g., if $n = 8$, then $l^i \leq 119$). Now, for each $d_l^i \in D^i$, we let the weight ω_l^i of d_l^i be the sum of the weights of all the same bipartition appearing in a new non-trivial compatible bipartition set of the i^{th} homeologous chromosome set. That is,

$$\omega_l^i = \sum_{j=1}^{m^{(i)}} \omega_{kj}^{ij} \qquad \text{s.t.} \quad d_l^i = d_{k_j}^{ij}, \quad 1 \leq k_j \leq \rho_j \tag{5.1.4}$$

Table 5.2: **An example of computed ω^{ij} and Λ^{ij} for a 7-gene family.** The original bipartition set corresponding to the phylogenetic tree of one of the gene families is given in the first column. Second column indicates the new non-trivial bipartitions compatible with the phylogeny shown in the first column. Computed ω^{ij} and Λ^{ij} shown in 3^{th} and 4^{th} columns corresponding to the new compatible bipartition in the same row in the second column.

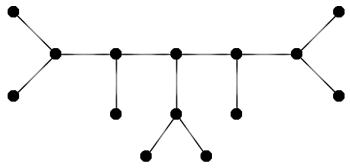
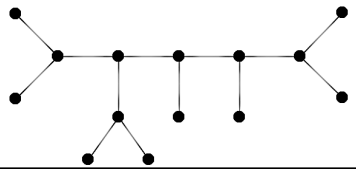
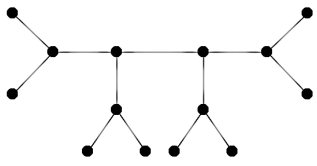
bipartition set b^{ij}	**Non-trivial bipartiton set** d^{ij}	ω_k^{ij}	Λ_k^{ij}
$\{\{\{E,G\},\{F,A,D,C,B\}\}$,	$\{\{E,G,H\},\{F,A,D,C,B\}\}$	0.07087	105
	$\{\{E,G\},\{H,F,A,D,C,B\}\}$	0.07087	105
$\{\{E,A,G\},\{F,D,C,B\}\}$,	$\{\{E,A,G,H\},\{F,D,C,B\}\}$	0.16535	45
	$\{\{E,A,G\},\{H,F,D,C,B\}\}$	0.16535	45
$\{\{E,A,D,G\},\{F,C,B\}\}$,	$\{\{E,A,D,G,H\},\{F,C,B\}\}$	0.16535	45
	$\{\{E,A,D,G\},\{H,F,C,B\}\}$	0.16535	45
$\{\{E,A,D,C,G\},\{F,B\}\}$,	$\{\{E,A,D,C,G,H\},\{F,B\}\}$	0.07087	105
	$\{\{E,A,D,C,G\},\{H,F,B\}\}$	0.07087	105
$\{\{E\},\{A,B,C,D,F,G\}\}$,	$\{\{E,H\},\{A,D,C,G,F,B\}\}$	0.00787	945
$\{\{A\},\{E,D,C,G,F,B\}\}$	$\{\{A,H\},\{E,D,C,G,F,B\}\}$	0.00787	945
$\{\{D\},\{A,B,C,E,F,G\}\}$,	$\{\{D,H\},\{E,A,C,G,F,B\}\}$	0.00787	945
$\{\{C\},\{A,B,D,E,F,G\}\}$,	$\{\{C,H\},\{E,A,D,G,F,B\}\}$	0.00787	945
$\{\{G\},\{A,B,C,D,E,F\}\}$,	$\{\{G,H\},\{E,A,D,C,F,B\}\}$	0.00787	945
$\{\{F\},\{A,B,C,D,E,G\}\}$,	$\{\{F,H\},\{E,A,D,C,G,B\}\}$	0.00787	945
$\{\{B\},\{A,C,D,E,F,G\}\}\}$	$\{\{B,H\},\{E,A,D,C,G,F\}\}$	0.00787	945

Afterwards, for $1 \leq i \leq \xi = 10$, we let the non-trivial bipartition set $CST^{(i)}$ corresponding to the consensus tree of the i^{th} homeologous chromosome set be a subset of D^i with cardinality of $n-3$ (here, 5) such that the bipartitons have the highest weights among the bipartitions in D^i that are compatible with one another. This consensus was estimated by applying a greedy algorithm, at each step selecting the highest weight remaining bipartition compatible with all the bipartitions already selected. As opposed to maximizing the sum over the weights, this approach promises a better consensus tree when the subsequent compatible bipartitions belong to the same former gene-trees(e.g. in the worse case, all the compatible bipartitions belong to the gene-trees corresponding to the bipartition with the highest weight in the new bipartition set).

5.2 Results

The consensus trees of homeology sets of *Saccharum officinarum* revealed overall three tree topologies. Seven out ten homeology sets had three terminal pairs. Both possible tree shapes with three terminal pairs appears among the trees, but with ratio of 6:1. The rest of the homeology sets had four terminals. Tree topologies are shown in Table 5.3, and the frequency of each peripheral configuration is shown in Table 5.4.

Table 5.3: **Tree topologies and frequencies.**

Consensus tree topology	**Frequency**
	6
	1
	3

We examined the probability of observing such phylogenetic trees under the null

hypothesis, the assumption that the phylogeny was generated by the one-branch-at-a-time model. The results , as shown in Table 5.3, are statistically significant for all the peripheral substructures discussed in this thesis based on the Multinomial Exact Test (P-value < 0.05). Thus, we reject the null hypothesis that the homeologous chromosomes were added one at a time.

Table 5.4: **Frequency of each peripheral substructure in the consensus trees and the expected value based on the null hypothesis.** The test result for each of the peripheral structure is statistically significant based on the Multinomial Exact Test (P-value < 0.05).

Peripheral structure	Configuration	Expectation	Observed frequency	P-Value
Terminal pairs	**2**	4.8	0	0.0001
	3	4.8	7	
	4	0.3	3	
Triples	**0**	0.3	3	0.0043
	1	2.4	1	
	2	7.3	6	
Type I Quadruples	**0**	7.3	6	0.0043
	1	2.4	1	
	2	0.3	3	
Type II Quadruples	**0**	2.7	9	0
	1	2.4	1	
	2	4.8	0	

Chapter 6

Discussion and conclusion

In this thesis, we have introduced an approach to deciphering the history of multiple polyploidizations of crops and other flowering plants. This is particularly pertinent for autopolyploids and for allopolyploids derived from closely related genomes. [12]

In contrast to most phylogenetic analyses, our calculations involved unlabelled unrooted trees, because there is no matching between the different sets of homeologous chromosomes. Enumeration involving unlabelled unrooted trees is notoriously difficult, but we have succeeded in developing easily computed recurrences for the expected number of occurrences of small structures on the periphery of the tree.

We apply our analysis to data on sets of homeologous genes from the sugarcane species *Saccharum officinarum*. Indeed, even though we have not specified any alternative hypothesis, the results are suggestive of one or two successive whole genome doublings underlying the gene similarity data. And although our null hypothesis is motivated by known examples of subgenomes being recruited on at a time, other null hypotheses may be equally reasonable in particular contexts.

Finally, we note that our data form but a very small subset of the homeologous gene sets in the *S. officinarum* genome, because of the stringent requirements on gene families for our tree construction method. Whether or not consideration of a larger data set may suggest different conclusions than ours, is an open question.

Glossary

gene family A set of closely related genes originating from the same ancestral gene by replication and mutation processes. They may either be clustered on the same chromosome or be dispersed throughout the genome. 31

heterozygosity Describes the condition in which the progeny inherited different alleles at a locus from its parents. 11

homeologous chromosomes Partially homologous chromosomes, as a result of ancestral homology. Homeologous chromosomes are created through chromosome duplication events. 30

paralogous genes Homologous DNA sequences in two organisms A and B that are descendants of two different copies of a sequence originally created by a gene replication event in the genome of a common ancestor. 31

polyploid A cell or organism that has more than two haploid sets of chromosomes. For instance, triploids have three sets, tetraploid have four, etc. 15

polyploidization The process of chromosome replication without subsequent nuclear division, that leads to polyploidy. 1

polyploidy The occurrence of more than two complete sets of chromosomes within a cell, a tissue, an organ, or an organism, resulting from chromosome replication without nuclear division or the recombination of two gametes with differing chromosome sets. The normal set is then diploid (diploidy), a triple set is triploid (triploidy), a quadruple set is tetraploid (tetraploidy), and so on. 2

Bibliography

[1] Zhang J, Nagai C, Yu Q, Pan YB, Ayala-Silva T, Schnell RJ, Comstock JC, Arumuganathan AK, Ming R. Genome size variation in three Saccharum species. Euphytica 185(3):511–519. 2012. doi: 10.1007/s10681-012-0664-6.

[2] Zhang JS, Zhang XT, Tang HB, Zhang Q, Hua XT, Ma XK, Zhu F, Jones T, Zhu XG, Bowers J, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nat Genet. 2018;50(11):1565–73.

[3] Thirugnanasambandam PP, Hoang NV, Henry RJ. The Challenge of Analyzing the Sugarcane Genome. Front Plant Sci. 2018;9:616.

[4] Food and Agriculture Organization of the United Nations. Production/Crops, Food and Agriculture Organization of the United Nations - Statistics Division. 2018. http://www.fao.org.

[5] Hoang NV, Furtado A, Botha FC, Simmons BA, Henry RJ. Potential for genetic improvement of sugarcane as a source of biomass for biofuels. Front Bioeng Biotechnol. 2015;3:182.

[6] Amborella Genome Consortium"The Amborella genome and the evolution of flowering plants." Science 342, 2013, 1241089. doi:10.1126/science.1241089.

[7] Liston, Aaron, Na Wei, Jacob A. Tennessen, Junmin Li, Ming Dong, and Tia-Lynn Ashman. "Revisiting the origin of octoploid strawberry." Nature Genetics 52, no. 1 . 2019, 2-4. doi:10.1038/s41588-019-0543-3.

[8] Wagner, Natascha D., He, Li and Hörandl, Elvira. "Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD sequencing data." Frontiers in Plant Science 11. 2020.

[9] H. S. Wilf, Generatingfunctionology, 2nd ed. Boston, MA: Academic Press. ISBN 0-12-751956-4. Zbl 0831.05001.

[10] Bondy, J. A.; Murty, U. S. R. Graph theory. Graduate Texts in Mathematics 244. Berlin: Springer (ISBN 978-1-84628-969-9/hbk). . 2008. Zbl 1134.05001

[11] Edger, P.P., Poorten, T.J., VanBuren, R. et al. Origin and evolution of the octoploid strawberry genome. Nature Genetics 51, 541–547. 2019. doi: 10.1038/s41588-019-0356-4

[12] Zhang, Jisen, Zhang, Xingtan, Tang, Haibao et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum L.* Nature Genetics 50, no. 11, 1565-1573. 2018.

[13] Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., Freeling M. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape. COGE with rosids, Plant Physiology 148, 1772 -1781. 2008b doi:10.1104/pp.108.124867.

[14] R. Ming et al. Autooctoploid sugarcane genome elucidates sucrose accumulation and female restitution. Manuscript submitted for publication. 2021.

[15] Lyons, E., Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53, 661- 673.(2008a). doi:10.1111/j.1365-313X.2007.03326.x.

[16] Jiao, Y., Li, J., Tang, H., Paterson, A. H. . Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. The Plant cell, 26(7), 2792–2802. 2014 doi: 10.1105/tpc.114.127597.

[17] Bowers, J. E., Chapman, B. A., Rong, J., Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature, 422(6930), 433–438. 2003 doi:10.1038/nature01521.

[18] Stanley, R., Fomin, S. Enumerative Combinatorics volume 2, Cambridge University Press (ISBN 978-1-107-60262-5). 1999. doi: 10.1017/CBO9780511609589.

[19] Gaut, Brandon S. and Doebley, John F. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences 94, no. 13. 6809-6814. 1997. doi: 10.1073/pnas.94.13.6809.

[20] Garsmeur, O., Droc, G., Antonise, R. et al. A mosaic monoploid reference sequence for the highly complex genome of sugarcane. Nat Commun 9, 2638. 2018. doi:10.1038/s41467-018-05051-5

[21] D. E. Knuth, The Art of Computer Programming. Fundamental Algorithms, Vol. 1, Third Edition, Addison-Wesley, Reading, MA. 1997.

[22] Deo, N. Graph Theory with Applications to Engineering and Computer Science. Englewood Cliffs, NJ: Prentice-Hall. 1974.

[23] Furnas, G.W. The generation of random, binary unordered trees. Journal of Classification 1, 187–233. 1984. doi:10.1007/BF01890123.

[24] Beyer, T., Hedetniemi, M.: Constant Time Generation of Rooted Trees. SIAM Journal on Computing 9. 1980. 706–712.

[25] Ruskey, F., and HU, T. C. (1977), "Generating Binary Trees Lexicographically," SlAM Journal of Computing, 6, 748-758.

[26] Rotem, D., and Varol, Y. L., "Generation of Binary Trees from Ballot Sequences," Journal of the Association for Computing Machinery, 25, 396-404. 1978.

[27] Zaks, S., and Ricards, D., "Generating Trees and other Combinatorial Objects Lexicographically," SIAM Journal of Computing, 8, 1979. 73-81.

[28] Knott, G. D. A numbering system for binary trees. Commun. ACM 20, 2. 113–115. 1977. doi: 10.1145/359423.359434.

[29] Pallo, J.M.: Enumerating, Ranking and Unranking Binary Trees. The Computer Journal 29 171–175. 1986.

[30] Phipps, J. B. The numbers of classifications. Canadian Journal of Botany. 54(8): 686-688. doi:10.1139/b76-073.

[31] Felsenstein., Joseph . The Number of Evolutionary Trees. Systematic Biology 27. 1. 27–33. 1978. doi:10.2307/2412810.

[32] Edwards, A. W. F. and L. L. Cavalli-Sforza. Reconstruction of Evolutionary Trees. In Phenetic and Phylogenetic Classification, W. H. Heywood and J. McNeil, pp. 67-76. Systematics Association Publication No. 6, London. 1964.

[33] COMTET, L., Advanced Combinatorics, Boston: D. Reidel Publishing Co. 1970.

[34] Otter, Richard. "The Number of Trees." Annals of Mathematics, Second Series, 49, no. 3. 583-99. 1948. doi:10.2307/1969046.

[35] Wedderburn, J. H. M. "The Functional Equation $g(x^2) = 2\alpha x + [g(x)]^2$." Annals of Mathematics, Second Series, 24, no. 2. 121-40. 1922. doi:10.2307/1967710.

[36] Buneman, P. The recovery of trees from measures of dissimilarity. Mathematics in the Archaeological and Historical Sciences. University Press; Edinburgh. 1971.

[37] Adams, Edward N. "Consensus Techniques and the Comparison of Taxonomic Trees." Systematic Zoology 21, no. 4. 1972. 390-97. doi:10.2307/2412432.

[38] Steel, Mike, Andreas W. M. Dress, and Sebastian Bocker. "Simple but Fundamental Limitations on Supertree and Consensus Tree Methods." Systematic Biology 49, no. 2. 2000. 363-68.

[39] Adams, E.N. N-trees as nestings: Complexity, similarity, and consensus. Journal of Classification 3, 299–317. 1986. doi:10.1007/BF01894192.

[40] T. Margush, F.R. McMorris, Consensusn-trees. Bulletin of Mathematical Biology 43. No. 2. 239-244. (ISSN 0092-8240). 1981 .doi:10.1016/S0092824081900197.

[41] Sokal, Robert R., and F. James Rohlf. "Taxonomic Congruence in the Leptopodomorpha Re-Examined." Systematic Zoology 30, no. 3. 309-25. 1981. doi:10.2307/2413252.

[42] Swofford D. Version 4. Sunderland (MA): Sinauer Associates; 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods).

[43] Felsenstein J. Version 3.5. Seattle (WA): University of Washington; 1995. Phylogenetic Inference Package (PHYLIP).

[44] Gareth Nelson, Cladistic Analysis and Synthesis: Principles and Definitions, with a Historical Note on Adanson's Familles Des Plantes. Systematic Biology 28, No. 1. 1763–1764. 1979. doi:10.1093/sysbio/28.1.1

[45] McMorris F.R., Meronk D.B., Neumann D.A. A View of Some Consensus Methods for Trees. In: Felsenstein J. (eds) Numerical Taxonomy. NATO ASI Series (Series G: Ecological Sciences), vol 1. Springer, Berlin, Heidelberg. 1983. doi:10.1007/978364269024218.

[46] Neumann, D. A. "Faithful consensus methods for n-trees." Bellman Prize in Mathematical Biosciences 63 . 271-287. 1983.

[47] Bryant, D. "A classification of consensus methods for phylogenetics." Bioconsensus. 2001.

[48] Ramsey, J., Schemske, D. Neopolyploidy in Flowering Plants. Annual Review of Ecology and Systematics, 33, 589-639. 2002.

[49] Levin DA. The Role of Chromosomal Change in Plant Evolution. Oxford, UK: Oxford Univ. Press. 2002.

[50] Dar, T. and R. Rehman. "Polyploidy: Recent Trends and Future Perspectives." Springer India 2017.

[51] Stace, CE. Plant taxonomy and biosystematics. Edward Arnold, Cambridge. (ISBN 9780713129557). 1989.

[52] Stebbins, G.L., Variation and evolution in plants. Columbia University Press. New York. 1950.

[53] Ramsey, J., Schemske, D. Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants. Annual Review of Ecology and Systematics, 29, 467-501. 1998.

[54] Madlung A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity (Edinb). 2013 Feb;110(2):99-104. doi: 10.1038/hdy.2012.79.

[55] Schulz-Schaeffer, J. Cytogenetics: plants, animals, humans. Springer Science and Business Media. 1980.