Modelling Amylopectin Biosynthesis with Evolved Stigmergic Building Algorithms

by

Asena Goren

A Thesis

Presented to

The University of Guelph

In partial fulfilment of requirements

for the degree of

Master of Science

in

Bioinformatics

Guelph, Ontario, Canada

ABSTRACT

Modelling Amylopectin Biosynthesis with Evolved Stigmergic Building Algorithms

Asena Goren
University of Guelph, 2017

Advisors:
Daniel Ashlock, Ian Tetlow

This thesis describes the investigation of the use of evolutionary computation to evolve stigmergic building algorithms to simulate the interplay of biosynthetic enzymes that is necessary to produce the amylopectin polysaccharides that are the major component of starch. Amylopectins are large, branched, semi-crystalline glucose polymers that are synthesized by a diverse group of enzymes with different functionalities. The structure and synthesis of amylopectin is not yet fully understood. This study uses an evolutionary computational model combined with a stigmergic model of wasp behaviour to elucidate unanswered questions about amylopectin biosynthesis. The interplay of three functional groups of enzymes, synthases, branchers, and debranchers, are modelled and analysed to determine their relative activities, context sensitivities, and roles in amylopectin synthesis.

Acknowledgments

I would like to express my gratitude to my advisers Daniel Ashlock and Ian Tetlow for their continued support and contribution to this project. I would also like to thank Lewis Lukens for serving on my advisory committee. I would like to extend thanks to Eric Bertoft and Fushan Liu for generously providing data that was fundamental to this project. I would also like to thank the faculty, staff and my peers in the Department of Mathematics and Statistics, Department of Molecular and Cellular Biology, and the Bioinformatics program who were supportive and helpful throughout my degree.

Table of Contents

Abstract ii

Acknowledgements iii

List of Figures and Tables vi

List of Terms and Abbreviations x

1 Introduction 1

1.1 Background on Starch Structure and Formation 2

1.2 Background on Evolutionary Computation 16

1.3 Background on the Stigmergic Model 21

1.4 Research Questions and Objectives 23

1.5 Outline of Thesis 25

2 Materials and Methods 26

2.1 Model Design 26

2.2 Experimental Design 33

2.3 Data Analysis and Other Tools 35

3 Results 37

3.1 Model Design and Refinement 37

3.1.1 Fitness Measurement Strategy 38

3.1.2 Number of Enzymes 44

3.1.3 Number of Generations 47

3.1.4 Breeding Parameters 48

3.1.5 Population Size 50

3.1.5 Population Size 50

3.2 Results Regarding Amylopectin Enzymes 51

3.2.1 Field of View 54

3.2.2 Modifications of Starch Synthase 54

3.2.3 Modifications of Branching Enzyme 55

3.2.4 Modifications of Debranching Enzyme 59

3.2.5 Comparison Between Maize and Barley CLD Profiles 65

3.3 Comparison of the Model to Natural Mutations of the Starch Biosynthetic … 66

3.3.1 Comparison Between sugary-2 and Experimental CLD 67

3.3.2 Addition of SS isoforms 68

3.4 Evolving Enzyme Contexts 71

4 Discussion **82**

4.1 Conclusions and Accomplishments 82

4.2 Directions for Future Research 85

Bibliography **88**

List of Figures and Tables

Figure 1.1.1: An illustration of the different scales of starch organization 5

Table 1.1.1: A summary of the starch biosynthetic enzymes described in Chapter 1.1. Enzyme action refers to the catalytic function of the enzyme, and enzyme context refers to substrate specificity of the enzyme 14

Figure 2.1.1: A two dimensional representation of the tree represented by the string "(GGGG(GGGG(GGGGG)GGG)GGG(GGGG(GGG)GG)GG)". The red circle marks the initial open parenthesis 27

Table 2.1.1 The alphabet of the custom regular expressions used to represent enzyme contexts 28

Table 2.2.1: Outline of the different sets of runs conducted and their objectives 33

Figure 3.1.1.1: The average NRMSE of the trees produced by experiments with different weightings of entropy in the fitness function. Error bars represent standard deviation 39

Figure 3.1.1.2: Different binning strategies are compared in terms of resultant NRMSE. The error bars represent standard deviation. The different columns represent testing the different binning strategies with different numbers of enzymes, from left to right they are 50,000, 75,000, and 100,000 42

Figure 3.1.1.3: Different binning strategies are compared in terms of resultant tree length. Tree length is defined as the total number of characters in the tree. The error bars represent standard deviation. The different columns represent testing the different binning strategies with different numbers of enzymes, from left to right they are 50,000, 75,000, and 100,000 43

Figure 3.1.2.1: A graph showing the improvement in NRMSE with an increasing number of enzymes. The plateau around 75,000 wasps shows where the model has enough enzymes to maximize the evolved enzyme activities. Error bars represent standard deviation 45

Figure 3.1.2.2: A graph of the runtime increasing proportional to the number of enzymes 46

Figure 3.1.3.1: A graph of the median NRMSE over 5000 generations showing stabilization around generation 900 48

Figure 3.1.4.1: The impact of mutation rate on NRMSE. The error bars represent standard deviation 49

Figure 3.1.5.1: A graph of the impact of population size on average, median, and minimum NRMSE 51

Figure 3.2.1: Four example trees produced by the model using the default parameters established in Section 3.1. The trees represent some of the diversity of structures achieved by the model. Branch lengths are proportional between the string and the image. The red circles mark the seed tree 53

Figure 3.2.1.1: The impact of enzyme field of view on NRMSE. Error bars represent standard deviation 54

Figure 3.2.3.1: Drawings of two example trees from the experiment with two BE isoforms where BE1 added branches of DP 8-12 and BE2 added branches of DP 13-17. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 57

Figure 3.2.3.2: Drawings of three example trees from the experiment with two BE isoforms where BE1 added branches of DP 6-8 and BE2 added branches of DP 11-13. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 58

Figure 3.2.3.3: The impact of adding a second BE of different range capabilities on NRMSE. The labels on the x-axis show the branch length ranges of BE1 and BE2, respectively, for each trial. Error bars represent standard deviation 58

Figure 3.2.4.1: The impact of different DBE controls on mean NRMSE. Error bars represent standard deviation. The controls included restricting DBE to cleaving branches a maximum of 6, 10, 12, 14, 18, 22, 28, or 40 long, restricting DBE to branches that are "crowded", i.e within 6 characters of another branch, or restricting DBE to branches that are both crowded and less than 10 long 60

Figure 3.2.4.2: Enzyme activities for SS, BE, and DBE with different DBE controls applied. The controls included restricting DBE to cleaving branches a maximum of 6, 10, 12, 14, 18, 22, 28, or 40 long, restricting DBE to branches that are "crowded", i.e within 6 characters of another branch, or restricting DBE to branches that are both crowded and less than 10 long 62

Figure 3.2.4.3: Drawings of four example trees from the run where DBE could only cleave branches a maximum of 10 long. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 63

Figure 3.2.4.4: Drawings of four example trees from the run where DBE could only cleave branches a maximum DP six. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 64

Figure 3.2.5.1: Drawings of three example trees from the experiment using barley CLD instead of maize. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 66

Figure 3.3.1.1: A difference plot comparing the CLDs of the model and *sugary-2/cgr04* mutant amylopectin against waxy maize amylopectin 68

Figure 3.3.2.1: A difference plot comparing the CLDs of the experiment using three SS isoforms and *sugary-2/cgr04* mutant amylopectin against waxy maize amylopectin . 69

Figure 3.3.2.2: Drawings of three example trees from experiment with three SS isoforms and 1-3 G elongation capacity. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 70

Figure 3.4.1: Drawings of three example trees from the experiments with evolvable enzyme contexts. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds 72

Figure 3.4.2: A comparison of the character length of the evolvable context strings in terms of average and minimum NRMSE. Error bars represent standard deviation 73

Figure 3.4.3: Debranching enzyme activities across runs with different enzyme context lengths. Error bars represent standard deviation 74

Figure 3.4.4: Branching enzyme activities across runs with different enzyme context lengths. Error bars represent standard deviation 75

Figure 3.4.5: Starch synthase activities across runs with different enzyme context lengths. Error bars represent standard deviation 75

Figure 3.4.6: A word cloud depicting the diversity of DBE contexts that evolved in an experiment where the maximum context length was 12. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts 78

Figure 3.4.7: A word cloud depicting the diversity of DBE contexts that evolved in an experiment where the maximum context length was 25. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts 79

Figure 3.4.8: A word cloud depicting all the enzyme contexts that evolved for starch synthase, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts 79

Figure 3.4.9: A word cloud depicting all the enzyme contexts that evolved for branching enzyme, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts 80

Figure 3.4.10: A word cloud depicting all the enzyme contexts that evolved for debranching enzyme, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts 81

List of Terms and Abbreviations

BE: branching enzyme - a category of glucan branching enzymes involved in amylopectin biosynthesis

CLD: chain length distribution - the lengths of glucans in an amylopectin sample that has been debranched by isoamylases

DBE: debranching enzyme - a category of glucan debranching enzymes involved in amylopectin biosynthesis

Domain knowledge: validated information regarding an environment or system

DP: degree of polymerization – the length of a glucan

Enzyme action: The outcome of a reaction that an enzyme catalyses

Enzyme activity: The extent to which an enzyme is active and functional. This is a black box model parameter representative of a variety of factors impacting overall catalytic capacity of an enzyme, including enzyme concentration, catalytic activity, post-translational modifications, etc.

Enzyme context: Any feature of the substrate that an enzyme can recognize and can trigger an enzyme action

Enzyme Number: Also referred to as number of enzymes, this is a model parameter that is the maximum possible enzyme actions that can be used to build a tree. This is not the same as enzyme concentration *in vivo* because in the model enzymes can catalyse only one reaction, while in reality a single enzyme can catalyse many reactions.

Fitness landscape: The composition of all possible solutions and their fitnesses

Human readable: Any representation of information that can be reasonably interpreted by most people. This is in contrast to machine readable information that must be processed by a computer before a human can interpret it.

SQL: A computer programming language used for managing data in databases

SS: starch synthase - a category of glucan polymerizing enzymes involved in amylopectin biosynthesis

String: A computer science term for a one-dimensional character array

SVG: scalable vector graphic – an image file format

Chapter 1

Introduction

Evolution is the driving force behind all biological innovations and is the inspiration for evolutionary computation, a branch of mathematics that uses the theory of evolution to guide algorithms that can solve problems. Evolutionary computation has been a useful tool in an abundance of fields, and has been implemented for addressing a wide variety of biological problems, modelling such diverse systems as epidemiology (Ashlock and Jafargholi 2007), primer design (Ashlock et al 2002), and even nutrient transfer in mycorrhizal networks (Ashlock and Goren 2014). In this project, evolutionary computation is being applied to the study of starch structure for the first time, specifically to the branching patterns of the amylopectin component of starch and the enzymes involved in amylopectin synthesis.

An evolutionary computational model is a good fit for understanding amylopectin synthesis for several reasons. Evolutionary algorithms are particularly well suited to problems where limited domain knowledge (subject knowledge) is available for incorporation into the algorithm. This said, the more information you can embed into the design of an evolutionary algorithm, the better the success of the model will be (Ashlock and Lathrop 1998; Hughes et al 2014; Ashlock and McNicholas 2012; Wolpert and Macready 1997). For this reason, it is common practice to iteratively develop evolutionary algorithms by incrementally incorporating domain knowledge as it is gained from the simulation, a technique that was employed in this study. In the case of amylopectin synthesis, there is currently insufficient domain knowledge to create a non-

evolutionary algorithm that can simulate the process by which enzymes construct an amylopectin polymer. While it is known that starch synthases, branching enzymes, and debranching enzymes are major players in the synthesis of amylopectin, it is not yet known how the enzymes coordinate their actions to build the intricately branched amylopectin structures that are required for the specific physical and chemical properties of amylopectin, such as semi-crystallinity and water insolubility. While the enzyme kinetics of several starch enzymes have been studied *in vitro*, the true kinetics *in vivo* can be very different, and are very non-trivial parameters to deduce. For all these reasons, an evolutionary algorithm is an ideal starting place for modelling the biosynthesis of amylopectin to further our understanding of how the enzymes achieve the coordinated construction of amylopectin.

1.1 Background on Starch Structure and Formation

Starch is essential to much of life on Earth. Starch is a key component of plant metabolism because it is the primary method plants use to store energy and is found across the Kingdom Plantae in everything from red and green algae to angiosperms, and even in non-photosynthetic (myco-heterotrophic) plants (Leake 1994). Animals, including humans, that rely on plants nutritionally derive a lot of energy from digesting the starch stored in various plant tissues. Humans are highly dependent on starch for energy because we lack the enzymes required to digest the beta-linked glucose found in cellulose, the other major glucose polymer synthesized by plants. Starch is also used

extensively for a wide variety of industrial applications because of it's unique material properties and relative abundance in an agricultural society (Doane 1994).

From a plant biology perspective, starch has several attributes that are important to it's function. Starch must be a dense and efficient way of storing glucose in an osmotically neutral state. It must also be semi-crystalline to allow for water insolubility, but not be fully crystalline so that it is accessible by degradation enzymes when needed. Transient starch, which is typically located in plant leaves, differs from storage starch, typically located in roots, stems, seeds, and other storage tissues. It is worthwhile to note that some form of starch exists in nearly all plant tissues (Pérez et al 2009). Transient and storage starch have different properties to match their differing functions, primarily that they have different rates of turnover (Stitt and Zeeman 2012). Transient starch is used to store energy over the plant's day/night cycle to maintain a consistent energy supply during the night when photosynthesis can not take place, thus playing an important role in plant productivity (Graf and Smith 2011). Storage starch is intended to provide energy over longer duration life cycle events, such a over seasons and even generations. The different properties of different types of starches are determined by their chemical composition and structure. Glycogen is the animal and fungal analogue of starch, that, due to its high branch frequency at 10%, which is double amylopectin's branch frequency at 5%, is water soluble and non-crystalline (Deng et al 2016).

The study of starch spans many orders of magnitude, from the molecular level to the plant tissue level (see Figure 1.1.1). The smallest unit of structure is the glucose molecule. Glucose molecules are polymerized into the two major components of starch:

amylose and amylopectin. Amylose and amylopectin are both α(1,4) glucose polymers with primary and secondary, and possibly tertiary structures. It is known that the amylopectin organizes radially in layers as is evidenced by microscopy using polarized light (Gallant et al 1997). The radial layering occurs on at least two size scales. On the smaller scale is what is commonly called the 9 nm repeat, clusters, or lamellae, which refers to the tightly packed amylopectin branches alternating with the more loosely packed linkages between them. The 9 nm periodicity has been observed through X-ray diffraction and electron microscopy studies, and is widely conserved in starches from different botanical sources (Pérez et al 2009). A second alternating feature occurs at the 100-300nm scale. These features are often termed growth rings because of their visual similarity to the rings in a tree trunk, but are not thought to have any similarity in function or ontology. The amylose and amylopectin, along with minor components of proteins, lipids, and other polyglucans, form starch granules which are the predominant way starch is stored in plant cells, specifically in the plastids. The size and shape of starch granules can vary considerably based on botanical source. Granule size measurements range from 1 μm to more than 100 μm. Electron and confocal microscopy have been used to make collections of the different shapes and sizes of starches found in different botanical sources (Jane et al 1994, van de Velde et al 2002). In this study we will focus on the polysaccharide level of organization, specifically on the synthesis of amylopectin.

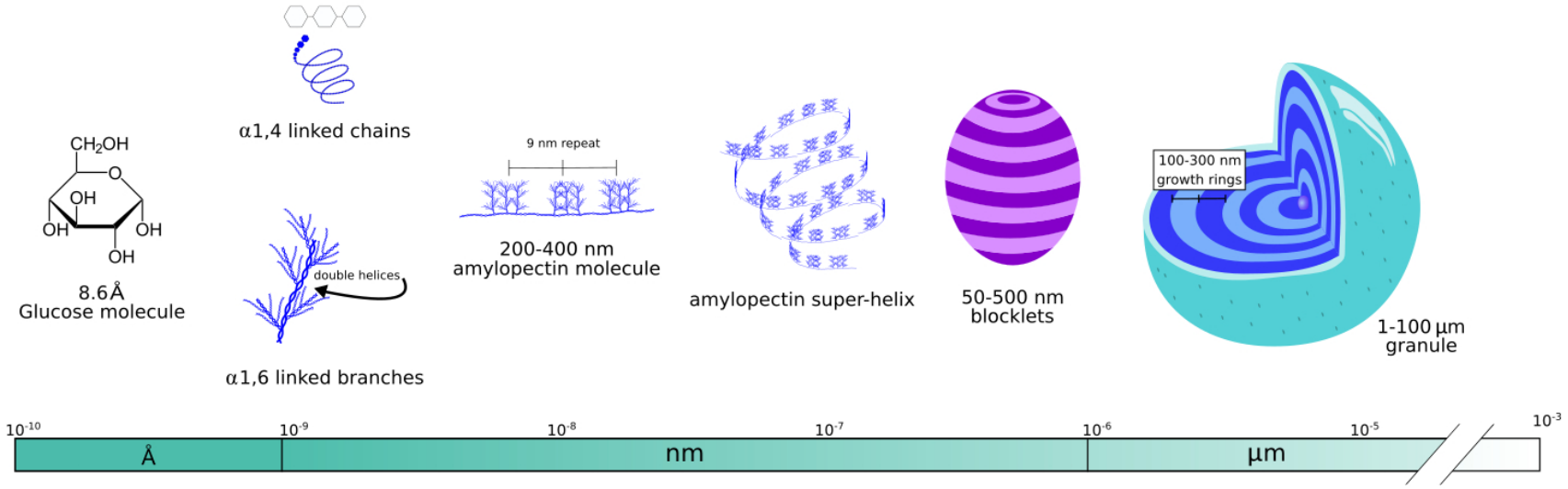

Figure 1.1.1: An illustration of the different scales of starch organization.

Starch is comprised of two different glucose polymers: amylose and amylopectin. Both amylose and amylopectin have primary chains of glucose linked with α(1,4) linkages. Branches off the primary chain are attached by an α(1,6) linked glucose. The main difference between amylose and amylopectin is that amylopectin is highly branched while amylose is sparsely branched, which has both chemical and biological significance. Starch is created by the interweaving of amylose and amylopectin, and can only be identified as containing two separate polymers after separation following solubilization of a starch granule (Pérez et al 2009). The localization and role of amylose in the starch is unclear, however it has been hypothesized that amylose acts as a glue because of the reduced mechanical strength of starch lacking amylose (Bettge et al 2000). Starch is thus not a single substance but a category of materials with different properties and attributes based on the ratio of amylose to amylopectin in the starch and the details of the structure and organization.

While the composition of starch can vary by source, e.g. the species and tissue of origin, amylopectin is generally the major component of starch, making up 70-80% of the granule (Pérez and Bertoft 2010). Amylopectin is one of the largest known natural polymers, with the average diameter of a single amylopectin molecule estimated at 200-400 nm (Martin and Smith 1995). Amylopectin has approximately one branching point for every 20 glucose residues (Manners 1962). It's unique chemical and physical properties that arise from the branching patterns are thought to be to a large extent responsible for the supramolecular structure of starch (Pérez et al 2009). It is worthwhile to note that, as with starch, amylopectin describes not a single uniform polymer but rather

a category of polysaccharides that can vary in branching patterns. In other words, not all amylopectin has the same branching pattern. Bertoft et al (2008) classified amylopectin from ten different botanical sources into four structural types based on internal unit chain profiles. The chain length distributions of amylopectin from different botanical sources has been studied in detail and will be used extensively in this study.

There are currently two physical models of amylopectin that are most widely recognized. Both models conform to some physical data, but neither does so perfectly. These models are known as the cluster model (Hizukuri 1986) and backbone model (Bertoft 2013). The cluster model represents amylopectin as a main chain along which there are clusters of extensive fractal branching. The backbone model represents amylopectin as a long backbone, from which radially arise branch chains, off which the fractal branching occurs. The cluster model has never been possible to achieve in computer simulation, which may suggest that it is physically impossible in three dimensions. Three dimensional modeling of the backbone model has not yet been completed. In this study, amylopectin is represented as a one dimensional string, which allows the results presented to conform to both the cluster and backbone models of amylopectin.

The overall crystal structure of amylopectin has been described has having multiple polymorphs categorized into A, B, and C types. Mathematical modeling of the branches of amylopectin has shown that they are likely to form double helices, which has been confirmed *in vitro* (Pérez et al 2009; Gidley and Bulpin 1987). X-ray diffraction patterns show parallel double helices with six glucose units per turn and a period of 2.1

nm. The different crystal polymorphs are often characteristic of the botanical origin of the starch, with the A-type being predominant in cereals. In the A-type crystal the double helices are densely aligned, whereas the B-type is less dense and more hydrated. Starches with a mixture of the two polymorphs are called C-type. The factors responsible for determining polymorph are not yet fully understood but it is thought that chain length distribution is an important factor (O'Sullivan and Pérez 1999). The double helical structure is consistent with the biosynthetic specificity of starch synthesis enzymes, in that these enzymes are predicted to have the three dimensional specificity to build non-random branching structures.

The synthesis of amylopectin is achieved through the coordinated interactions of several enzymes. A summary of the enzymes important to starch synthesis can be found in Table 1.1.1. The amino acid sequences of these enzymes are highly conserved between species and tissues (Jespersen et al 1993). However, while the sequences are conserved the relative contribution of the different enzymes varies between botanical sources, which is believed to account for the structural variation between starches (Wang et al 2014). There are three main categories of enzymes involved in starch biosynthesis: the starch synthases (SS), the starch branching enzymes (SBE) and the starch debranching enzymes (DBE). Several other enzymes, *e.g.* starch phosphorylase, disproportionating enzyme (D-enzyme), and others, are also known to play a role in starch synthesis, but their precise roles are unclear. Much of our current understanding of the specific roles of the different enzymes and enzyme isoforms involved in starch biosynthesis has been derived from the study of mutants. It is important to exercise caution when interpreting the results of

mutant studies because many of the genes have demonstrated pleiotropic effects. These pleiotropic effects may arise from or be exacerbated by the fact that there is increasing evidence of several starch biosynthetic enzymes functioning in complexes, so separating the individual functions of the enzymes may be a more complex problem than would be initially assumed. Additionally, several classes of the enzymes have overlapping functions and are able to compensate for each other in some mutants (Zhang et al 2008; Wattebled et al 2008). What is known about the different categories and functions of starch biosynthetic enzymes will be summarized below, because of their importance to the development of the model used in this study.

The first feature of amylopectin to discuss is the synthesis of different DP glucan chains which is accomplished by starch synthases. The first committed step in the biosynthesis of starch is the production of ADP-glucose (adenosine 5’ diphosphate glucose) by AGPase (ADP-glucose pyrophosphorylase). ADP-glucose serves as a substrate for the starch synthases, which attach the glucosyl moiety in ADP-glucose in an α(1,4) configuration to the non-reducing end of a glucan chain. There exist multiple isoforms of SS, five isoforms have been categorized based on conserved amino acid sequences. Some isoforms of SS, namely the granule-bound starch synthases (GBSSI and GBSSII), are thought to specifically function in elongating amylose and not amylopectin. The remaining SS's, SSI, SSII, SSIII, and SSIV, are thought to function specifically on amylopectin. The SSs share a common C-terminal catalytic domain, but have variable N-terminal extensions (Leterrier et al 2008). The N-terminal regions contain features such as carbohydrate binding modules that could confer substrate specificity and functional

differences between the SS categories (Valdez et al 2008). SSI is thought to be primarily responsible for the synthesis of short glucan chains in the degree of polymerization (DP) range of 10 or less (Commuri and Keeling 2001). SSII, which has two tissue-specific isoforms in most monocots, SSIIa and SSIIb, is thought to elongate chains of DP 10 or less to mid-length chains in the range of 12-24 (Imparl-Radosevich et al 2003). SSIII is thought to be involved in the synthesis of the longest DP glucans in the intercluster region, but little is known about it's function (Commuri and Keeling 2001; James et al 2003). It has also been proposed that SSIII has a regulatory function (Edward et al 1999; Zhang et al, 2008). While SSIII does seem to play an important role in amylopectin synthesis, the impact of the loss of SSIII changes with different genetic backgrounds (Mangelsdorf 1947; Inouchi et al 1983; Fujita et al 2007; Ryoo et al 2007; Tetlow 2011). SSIV is phylogenetically related to SSIII, and has two tissue-specific isoforms SSIVa and SSIVb. It has been proposed from structural analysis that SSIV and SSIII have different glucan specificities (Leterrier et al 2008; Szydlowski et al 2009). In recent years the possibility of a fifth starch synthase group, SSV, has been proposed. It is also worthwhile to note that while the functions described above are generally accepted, there do exist sizeable gaps in the fundamental understanding of how each SS class performs its proposed role on the molecular level (Pfister and Zeeman 2016). The model proposed in this study can be used to aid in the further discovery of starch synthetic enzymes and their mechanisms of action and lends itself to continual refinement as domain knowledge about the starch synthases improves.

The next feature of amylopectin is the intricate branching. Branch points are achieved by the starch branching enzymes which affix glucan chains in an α(1,6) configuration. The specific placement of branches by SBEs is believed to be a major determinant of the cluster structure of amylopectin (Gérard et al 2000). Branching enzymes also have hydrolytic capability as typically they first cleave an α(1,4) bond to acquire a glucan chain that is then affixed in the α(1,6) configuration. The branching enzymes are thought to both transfer glucan segments within the same molecule as well as between different molecules, which is called inter-chain transfer (Borovsky et al 1979). It is worthwhile to note that this branching activity creates more non-reducing ends for the starch synthases to elongate. There is also evidence that at least some SBEs may be able to transfer branched segments (Nakamura et al 2010; Sawada et al 2014). There are two major classes of SBE, SBEI and SBEII, categorized based on their amino acid sequences. The two classes are thought to have different functions in terms of substrate specificity and the DP of the glucan chain that is attached at the branching point (Takeda et al 1993; Guan and Preiss 1993). More specifically, it is thought that SBEII transfers shorter chains and has a higher affinity for amylopectin, while SBEI transfers longer chains and has more affinity for amylose (Rydberg et al 2001). SBEII mutants have a distinct starch phenotype known as the *amylose-extender* phenotype (Li et al 2008; Yun and Matheson 1993). Interestingly, Gregor Mendel's famous wrinkled peas are in fact SBEII mutants (Bhattacharyya et al 1990). A deep understanding of BE activity on the molecular level would require understanding where branches are placed, the DP of

the glucans added, the source of the branch, and the degree of branching in the branch. Control of these features could yield a huge diversity in amylopectin structures.

Debranching enzymes are able to cleave the α(1,6) linkages synthesized by branching enzymes and thus remove branches from amylopectin. Two groups of DBE exist in plants, the isoamylase type, which includes at least three isoforms (isoamylase-1, isoamylase-2 and isoamylase-3), and the pullulanase type (Deschamps et al 2008). It has been shown in mutants that DBE is important to the formation of amylopectin but the mechanism of action is unknown. Isoamylase mutants accumulate a glucose polymer known as phytoglycogen in place of amylopectin (Zeeman et al 1998; Mouille et al 1996). Phytoglycogen is much more densely branched than amylopectin and is water soluble. One hypothesis about DBE's role in amylopectin synthesis is that DBE cleaves any excess branching that prevents the semi-crystalline structure of amylopectin to form, and that this trimming of glucan chains is required to form an insoluble granule (Ball et al 1996; Myers et al 2000). Another hypothesis in that DBE's function more indirectly by debranching small glucans in the plastidial stroma, preparing them for degradation by amylases. By helping to clear away excess glucans, DBE prevents starch synthases and branching enzymes from futilely extending and branching small glucans that will not turn into amylopectins (Zeeman et al 1998). In this study the model DBE functions under the first hypothesis, directly trimming the amylopectin to help create the correct branching pattern. The output of the model is thus capable of either furthering the first hypothesis or suggesting that DBEs are more likely to play an indirect role.

While the starch synthases, branching enzymes, and debranching enzymes are considered the main suite of enzymes involved in amylopectin synthesis and were used alone as the basis for the model in this study, it is worth noting that there are other enzymes that may play an important role in amylopectin synthesis, and could be incorporated into the presented model. For example, starch phosphorylase is capable of elongating a glucan chain by adding glucose 1-phosphate to a non-reducing end. This is a reversibe reaction that starch phosphoylase can catalyse in either direction, resulting in the addition or removal for a glucose from the glucan. The direction of the reaction is thought to be determined by substrate concentration (Mu et al 2001). Another example is D-enzyme, which is capable of transferring glucose from maltotriose to longer glucan chains. While the role of D-enzyme is not fully understood, D-enzyme lacking mutants do show changes in starch metabolism (Colleoni et al 1999; Bresolin et al 2005). Protein Targeting To Starch (PTST) is another protein that may be involved in starch synthesis and is thought to complex with GBSS allowing it to bind to starch granules (Seung et al 2015). While GBSS is involved in the synthesis of amylose and not amylopectin, the discovery of PTST demonstrates the possibility of more enzymes and proteins that could be playing unexpectedly important roles in amylopectin biosynthesis even without direct catalytic functions. Non-catalytic proteins could have relevance to the interpretation of results from the model presented. For example, if an enzyme is found to have a certain substrate specificity in the model which does not match any of the enzyme's known carbohydrate binding domains, this could indicate that a non-catalytic protein involved in enzyme targeting is responsible for the enzyme context.

Enzymes Involved in Amylopectin Synthesis		
Enzyme Name	**Action**	**Context**
Starch Synthases **(SS)**	• Elongate in α(1,4) configuration using ADP-glucose	• Non-reducing ends of existing glucan chains
SSI	• Synthesize short chains DP ≤ 10	• Maltooligosaccharides
SSII	• Synthesize intermediate chains DP 12-24	• SSI products
SSIII	• Synthesize long intercluster chains	• SSII products
SSIV	• Elongation • Granule initiation	• Maltotriose
Starch Branching Enzymes **(SBE)**	• Attach glucans in α(1,6) configuration • Cleave α(1,4) linkages	• SS products
SBEI	• Transfer longer chains	• Amylose
SBEII	• Transfer shorter chains	• Amylopectin
Debranching Enzymes **(DBE)**	• Cleave α(1,6) linkages	• Branch points
Isoamylases	• Cleave α(1,6) linkages	• Amylopectin
Pullulanase	• Cleave α(1,6) linkages	• Amylopectin, Pullulan
Other Enzymes		
Starch Phosphorylase	• Elongate in α(1,4) configuration using glucose 1-phosphate (also reverse reaction)	• Non-reducing ends of existing glucan chains
D-enzyme	• Transfer glucose from maltotriose	• Non-reducing ends of existing glucan chains

Table 1.1.1: A summary of the starch biosynthetic enzymes described in Chapter 1.1. Enzyme action refers to the catalytic function of the enzyme, and enzyme context refers to substrate specificity of the enzyme.

The regulation of starch biosynthetic enzymes is another important topic with relatively little domain knowledge available. It is known that many of the enzymes possess a carbohydrate binding module (CBM) that may confer substrate specificity that could control the enzymes' activities. Effector molecules, which are often metabolic intermediates, have also been shown to impact the enzymes' activities. One example is the allosteric regulation of AGPase, which is stimulated by the substrate 3-phospho-glycerate and inhibited by inorganic orthophosphate (Preiss 1991). Additionally, branching enzyme has demonstrated modulation in catalytic activity by phosphorylated intermediates and inorganic orthophosphate (Morell et al 1997), and maize endosperm pullulanase-type DBE has been shown to be inhibited by soluble sugars (Wu et al 2002). There is also evidence that some of the starch biosynthetic enzymes, specifically AGPase, SSI, and DBE, may be redox sensitive (Glaring et al 2012; Tiessen et al 2002; Schindler et al 2001). Some of the starch biosynthetic enzymes, specifically branching enzymes, are also regulated by phosphorylation by plastidial protein kinases. The phosphorylation can directly impact catalytic activity, but more importantly is involved in the formation of protein complexes, which could have both functional and regulatory significance (Tetlow et al 2004).

The *sugary-2* mutant is one of particular importance to this study, and as such will be introduced briefly here. The *sugary-2* mutant lacks SSII activity and has one of the most striking mutant starch phenotypes (Zhang et al 2004). The SSII mutation has been characterized in maize, wheat, barley, japonica rice, pea, potato, sweet potato, and *Arabidopsis thaliana*, with all the species sharing similar starch phenotypes. The mutant

phenotype can be readily measured and demonstrated by the altered chain length distribution (CLD), and altered crystallinity (Liu et al 2012). CLD is a measurement of the lengths (DP) of the glucan chains in amylopectin after debranching with isoamylase. In the *sugary-2* mutant, CLD analysis shows an increase in short chains of DP 6-10 and decreased intermediate chains of DP 12-30. This would suggest that SSII has a role in the synthesis of intermediate length glucan chains, possibly by elongating short chains (Morell et al 2003). The CLDs of normal and *sugary-2* mutant maize starches were used extensively in this study.

Despite the vast body of knowledge that has been gained about starch structure and biosynthesis, definite gaps exist in our understanding of this biologically complex and economically important biopolymer. The main reactions in each of the known biosynthetic pathways do not in themselves explain how the intricate amylopectin or granular structures are formed, nor account for the diversity of starches and amylopectins that exists across species and tissues, and *in vitro* attempts at synthesizing amylopectin have been unsuccessful. Understanding the coordination of enzymes required to synthesize (and degrade) an architecturally complex polymer is not only a worthy problem to solve in and of itself, but also a great model system for the application of evolutionary algorithms to enzymatic systems.

1.2 Background on Evolutionary Computation

Evolutionary computation describes algorithms that use evolution to solve a problem. In both nature and evolutionary computation, there are three basic components

necessary for evolution: variability, selection, and heritability. Variability simply means that there are differences between individuals. It is readily apparent that a uniform population can not evolve. Selection and heritability will be described in more detail below.

Selection is the force that drives evolution because without it there would be no requirement to improve fitness. Selection is what rewards high fitness solutions and removes low fitness solutions, causing the population as a whole to improve. Selection improves the existing population within a single generation, but can not extend over generations without heritability.

Heritability is the ability for individuals to pass on traits to their offspring. In nature this is accomplished through genes. Without genes, organisms with higher fitness would not be able to pass on the traits that improved their fitness to their offspring, resulting in no generational improvements. When designing an evolutionary algorithm it is necessary to design a data structure that has heritable components so that evolution can be accomplished. Selection and heritability often function together to create generational improvements. For example, selecting higher fitness individuals to breed improves the likelihood that the offspring will inherit beneficial traits. Evolutionary algorithms gain their advantage over random sampling techniques because evolutionary sampling is directed by selection and heritability.

Evolutionary algorithms share common features regardless of the problem they are being applied to. A generalized evolutionary algorithm can be presented as follows:

Step 1: Generate an initial population of solutions

Step 2: Assess the fitness of each individual in the population

Step 3: Use fitness biased selection to choose individuals to reproduce

Step 4: Use a breeding algorithm to combine features of the individuals selected in Step 3 to generate a new population. Apply any mutations or other variation operators.

Step 5: Repeat steps 2-4 until termination criteria are met

The first step, generating an initial population, can be accomplished by design or by random generation. In order to be able to generate a population the designer must have decided upon a representation. The representation is how the members of the population are encoded and what type of data structure is used. Some examples of representations include strings, arrays, instruction sets, values, functions, images, or almost anything else that can be assigned a fitness value and bred. The choice of representation is often not a trivial decision to make, and can greatly impact the outcome (Ashlock and Kim 2005; Ashlock et al 2006).

Step 2 is where each member of the population is assigned some fitness value based on the fitness function. The fitness function is a specified figure of merit by which an individual can be scored. The fitness function can take a multitude of forms, such as a function, simulation, stochastic simulation, or game. The fitness function could be be a very simple metric, such as simply looking for a maximum or minimum value, or something much more complicated such as a complex formula of weighted attributes or even performance on a standardized driving test. Deciding on the fitness function to use could be very easy or very challenging depending on the problem being addressed.

Step 3 is the selection step, where individuals with higher fitness are selected to reproduce. This is called fitness biased selection. Many methods for selection exist, and the method selected can have a big impact on the outcome. It is also worthwhile to note that different representations could make different selection methods function differently and be a better or worse choice for a specific problem. For example, knowing what percent of the population should be selected is non-trivial. Selecting only the very elite may have benefits, but may also reduce diversity and yield poor results in the long run. A representation with many components might have enough diversity in 10% of the population, while a different representation with fewer components might suffer from lack of diversity when using 50% of the population.

If the fitness function in Step 2 is very computationally intensive, efficiency may be gained by combining steps 2 and 3, iteratively selecting sub-populations of individuals, scoring them, and filling the selection buffer until you have enough individuals above a fitness threshold with which to proceed. This method improves efficiency because not every member of the population has to be scored, but has the drawback of possibly excluding the highest fitness individuals. While this method is only practical in scenarios with a computationally intensive fitness function, it is useful as an example to demonstrate the flexibility that exists even in the generalized evolutionary algorithm which must be creatively tailored to suit the problem at hand.

After reproductive individuals have been selected, it is time for them to breed (Step 4) to form a new population. It is typical to term the generations as in biology as parent, child/offspring, etc. There are countless ways breeding can be accomplished.

Many questions must be answered by the designer when deciding on a breeding algorithm, for example: will any individuals from the parental generation continue directly into the next generation? How many parents will combine features to generate a child? How many children will a set of parents generate? Can parents breed more than once? How will parental features be selected and combined into an offspring? It is important to note how, once again, the breeding function is very dependent on the representation selected.

During Step 4, in addition to breeding, other variation operators can be applied. For example, mutation is a commonly used variation operator. Mutation means making a small, random change to the offspring before including it in the new population. For example, if the data structure is a string, a mutation could be randomizing a single character in the string. The rate and impact of mutations needs to be optimized on a case by case basis. Mutation is beneficial because it improves diversity and allows features that are not in the starting population to enter over generations, meaning that not as large of a starting population is necessary. The drawback of mutation is that the offspring inherit less from their parents, so some of the benefits of heritability are lost.

Lastly, a termination criteria for the algorithm must be determined. In an ideal situation, the algorithm can terminate when an optimal solution is reached. However, this is not often feasible in the context of the problem being addressed, so another termination criteria is selected, and can be based on something as simple as elapsed runtime or elapsed generations. Depending on how many generations and how big of a population must be examined, evolutionary algorithms could take a very long time to run. However,

it is worthwhile to maintain the perspective that an evolutionary algorithm is much more efficient and time saving than a brute force or exhaustive algorithm, and can often produce better results more quickly than a random sampling method.

1.3 Background on the Stigmergic Model

In 1995, Guy Theraulaz and Eric Bonabeau published a model for stigmergic building algorithms, initially proposed by Grassé in 1959 (Grassé 1959; Theraulaz and Bonabeau 1999), to describe how individual wasps, each possessing a limited behavioral repertoire, can coordinate their activities to build the complex architecture of a wasp nest (Theraulaz and Bonabeau 1995). Stigmergy describes indirect coordination by agents that enables the production of complex structures by individuals which communicate only through the local environment they perceive. Indirect coordination is achieved by environmental signals that result from each agent's actions, which are sensed by other agents to determine and incite subsequent actions (Bonabeau 1999). Building actions are directed by the dynamically evolving shape under construction (Theraulaz and Bonabeau 1995). Theraulaz and Bonabeau's model has been expanded to design optimization strategies based on the behaviors of ant colonies, and has had applications in communication networks and many other sectors (Bonabeau et al 2000). The model is a very good fit for starch biosynthesis because enzymes, like wasps, have limited behaviors and communication ability, and the end product, starch, has structural intricacies that can only be accomplished through coordination (Wang et al 2014). Despite the apparent

similarity, this is the first time the concept of stigmergy has been applied to an enzymatic process.

Theraulaz and Bonabeau developed a computer simulation of wasp nest construction using artificial agents. The agents moved randomly on a cubic lattice and deposited elementary bricks of two types depending on the local configurations of the neighboring 26 cells. The size of the neighborhood is dependent on the radius of perception of the agents. All actions by agents were predetermined in the form of a lookup table with an equal number of entries for actions and stimulating configurations (Theraulaz and Bonabeau 1995). This idea gave rise to the action-context model for enzymes used in this study, where each enzyme has a prescribed action (or set of actions) in response to a specific context (or set of contexts). The term context in this study is intended to have the same meaning as stimulating configuration, with only the method of stimulation differing between insects and enzymes.

Coherent architectures were found to be generated by specific building algorithms which were all within a subdomain of possible building algorithms. One requirement for successful building of a structure by the stigmergic model was that the desired end product could be decomposed into modular subshapes. The subshapes are observed in wasp nests in nature (Theraulaz and Bonabeau 1995). This could be a possible explanation of the periodicity, such as clusters, growth rings, and blocklets, observed in starch.

This project is also the first time evolutionary computation has been combined with the stigmergy model. A novel metaheuristic based on stigmergy called Ant Colony

Optimization was described and compared to evolutionary computation algorithms (Dorigo et al 2000). However, there was no attempt to combine the two strategies. Evolutionary computation would be beneficial to include in the wasp colony simulation described earlier because of its similarity to other agent based evolutionary systems. In essence, evolving the building algorithms that would follow the stigmergic building process would provide a more effective way to understand the fitness landscape of the building algorithms that produce coherent architectures. Combining evolutionary computation with a stigmergic building process is the method applied to amylopectin synthesis in this project.

1.4 Research Questions and Objectives

This project is centred on the design and development of new computational tools for understanding and exploring amylopectin synthesis. The methods presented are not limited to amylopectin and could be tailored to other systems, as could different models and methods be developed to answer the fundamental questions presented about amylopectin biosynthesis. The following research questions and objectives will help to outline the scope of this project and focus on specific areas of interest.

Starch is a complex material composed to two polymers as well as trace amounts of proteins, lipids, and inorganic components (*e.g.* phosphate). This project is focused only on the biosynthesis of the amylopectin component of starch. While a lot is known about starch biosynthetic enzymes, a concise understanding of which enzymes are necessary for amylopectin synthesis and how these enzymes coordinate their actions is

still lacking. The model presented can help address these questions but by no means entirely solve them. The following research questions provide a realistic picture of what the model presented can accomplish within the scope of this project.

1. What enzyme actions are necessary to create the chain length distribution seen in amylopectin? What is the the minimum suite of enzymes required to synthesize the different length glucan chains found in amylopectin in a single branched molecule?
2. What relative activities of the minimum suite of enzymes are required to create a single branched polymer with the CLD of amylopectin?
3. Is a stigmergic building algorithm representative of amylopectin biosynthesis? Does substrate recognition play an important role in the coordination of enzyme activities?

Additionally, some research questions not directly related to amylopectin synthesis are also within the scope of this project. They are summarized as follows.

1. Can the stigmergic building model be improved by combination with evolutionary computation methods?
2. Can an evolvable representation of enzymes be designed to simulate amylopectin biosynthesis?
3. Can a simple molecular representation be designed for a lightweight tool which can quickly evolve and still emulate the behaviour of a physical polymer?

Ultimately, the hope of this project is to venture into new frontiers of starch modelling that have not yet been embarked on. It will hopefully serve as the first steps

into further investigation of the applications of evolutionary computation and stigmergic building algorithms in the field of biopolymer synthesis and enzymatic systems. Uncovering surprising similarities between insect and enzyme behaviour can provide insight into different types of intelligence found in nature, and aid in the expansion of modelling diverse natural processes.

1.5 Outline of Thesis

The organization of this thesis will be presented in the manner of a typical scientific paper. The next chapter, Chapter 2, will give an overview of the materials and methods used in this project, including details of the model's architecture and other tools used. Chapter 3 will discuss the results, showing the iterative process of model refinement, any conclusions that could be reached about amylopectin synthesis, and the application of the model to the well-characterized *sugary-2* mutant. The final chapter, Chapter 4, will summarize what this project has accomplished, discuss the lessons learned in model development and point out directions for future research.

Chapter 2

Materials and Methods

This section covers the materials and methods used in this study. First, the overall design of the model is described, including the acquisition of the chain length distribution data embedded in the model. The following section outlines the experiments conducted using the model. The last section describes all other tools used in this project.

2.1 Model Design

The model was created *de novo* using entirely free and open source tools. The code was written entirely in C++ using Gedit, an open source text editor, and compiled with the G++ compiler on a GNU/Linux based operating system. It is intended that the model and all supplementary tools in this study can be run on any personal computer with only basic hardware requirements using free and widely available software.

The goal of the model developed in this study is to be able to simulate the construction of an amylopectin polymer. The amylopectin polymer is represented as a string of glucoses (abbreviated as the letter G) with branch points represented by open parentheses and non-reducing ends represented by closing parentheses. Each string is referred to as a tree because it represents a single branched polymer. An example tree could look like "(GGGG(GGGG(GGGGG)GGG)GGG(GGGG(GGG)GG)GG)", which could be represented two dimensionally as in Figure 2.1.1 below. This representation works well because it is lightweight and easy to handle.

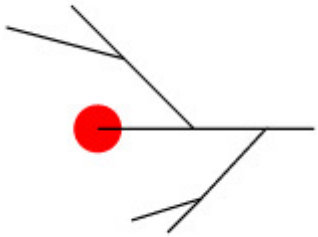

Figure 2.1.1: A two dimensional representation of the tree represented by the string "(GGGG(GGGG(GGGGG)GGG)GGG(GGGG(GGG)GG)GG)". The red circle marks the initial open parenthesis.

The trees are built by stigmergic building algorithms called programs, each representing a suite of enzymes. Each enzyme has built-in actions, context sensitivities and activity levels, each stored as a separate, evolvable component. The programs make up the evolving population. In other words, stigmergic building algorithms are what evolves. The simplest program component is enzyme activity level, so it was the first component targeted for evolution. Enzyme activity levels are stored in a two dimensional array, with one value for each enzyme in the program, and the number of programs as determined by the population size. The value is a whole number in the range 0-100 and represents what percent of the time the enzyme is functional. This effectively groups any forms of biological enzyme control, be it enzyme concentration, catalytic activity level, or post-translational modifications such as phosphorylation or complexation, into a single value that is the net of all these factors.

In accordance with the stigmergy model, enzymes each have an action and context. Enzyme actions were varied but never targeted for evolution within the scope of this project. If they were to be targeted for evolution, enzyme actions would be stored in a

lookup table but, for simplicity, were kept as single instruction sets directly in the body of the code. The last component, context sensitivity, is representative of any feature in amylopectin that an enzyme is able to recognize, and thus triggers an action. This likely occurs in nature through carbohydrate binding domains and enzyme active sites. In the model, the enzyme contexts are stored in a two dimensional array and represented as strings of regular expressions in a customized regular expression language. The enzyme contexts were either kept fixed for simplicity in early experiments, or allowed to evolve in later experiments. The regular expression language contains 9 characters and any numerals, and is tabulated in Table 2.1.1 below.

Regular Expression Alphabet	
Character	**Meaning**
G,), (	Match a G,), or (
b	A complete branch *e.g.* (GGGG)
o	Either a branch point or a complete branch
N	Any character (wildcard)
+, >, <	Accompanies a numeral (n), means n or more, greater than n, less than n, respectively.
Numerals	Indicates a number of feature repeats, typically a G, can be combined with the symbols +,>,<. *e.g.* G3 means GGG

Table 2.1.1 The alphabet of the custom regular expressions used to represent enzyme contexts.

At this point we will step through the evolutionary algorithm the model uses, elaborating on each step of the generalized evolutionary algorithm presented in the first chapter.

- **Step 1:** Generate an initial population of programs (stigmergic building algorithms)
    - Enzyme activity levels are stored as an array of randomized values 0-100.
    - Enzyme actions are hardcoded in the program before the population is initialized
    - Enzyme contexts are created as random strings of a maximum preset length of the characters in the customized regular expression alphabet and stored in an array.
- **Step 2:** Assess the fitness of each program in the population
    - Have each program execute a building simulation (*i.e.* create a tree)
        - Initialize a tree as a string, *e.g.* "(G)"
        - Repeat the following for a predetermined number of times (number of enzymes):
            - Randomly select a location in the tree for the enzyme
            - Randomly select an enzyme category from the suite being investigated. For now we will assume a basic three enzyme suite containing one elongation (polymerizing), one branching, and one debranching enzyme.
            - Test against the enzyme activity level by randomly generating a number from 0-100. If the generated number is less than the enzyme activity level stored in the array, proceed with the following steps.

Otherwise, skip the remaining steps for this enzyme and continue on to the next enzyme.

- Locate the nearest occurrence of a feature in the tree that matches the enzyme's context sensitivity, searching only within a predetermined field of view. A custom regular expression matcher determines if a region of the tree in the enzyme's field of view fits the enzyme's context sensitivity. If a match is found proceed to the following steps. If no match is found, skip the remaining steps for this enzyme and continue on to the next enzyme.
- Apply the enzyme's action. For example, if the enzyme is a branching enzyme, create a branch.

- Determine the chain length distribution (CLD) of the tree and any other desired metrics. The CLD is calculated by simply counting how many G's are in each branch of the tree, and calculating a percent frequency for each length found.
- Calculate the fitness of the program by comparing the tree metrics against experimental data. The primary fitness function used was normalized root mean square error (NRMSE) comparing the CLD of the tree against amylopectin CLD data. Store the fitness values, called scores, in an array.

- **Step 3 & 4:** Breed a new generation of programs to form an updated program population
  - Fitness biased selection of parent programs

    - The selection method used is a novel method called threshold selection. It is a relatively soft selection technique that was specifically designed for this model.
    - Each child program in the new population is assigned a random number in the range of the scores of the parent population. The random number is the child's parental score threshold, meaning that the child will only select parents with a score equal or better than their score threshold.
    - Each child selects parents by randomly selecting individuals in the parent population and comparing the parent's scores to their score threshold. The first two programs that pass the child's score threshold are selected as the parents. This systems allows programs with higher scores a better likelihood of reproduction.
- Combine parental features to create children
    - For each evolving component the child randomly takes on one of their parents attributes or a combination of the two where applicable. For example, for the enzyme context component, the child randomly takes one of their parents' enzyme context for each enzyme. For the enzyme activity component, the child can either directly take one of their parents' enzyme activities or average their parents' enzyme activities, for each enzyme in the suite.
- Apply the mutation operator

- At a predetermined rate, any inherited feature of a child may be randomly reinitialized. For example, if the mutation rate is set to 1 in 100, then 1 in 100 times that a child inherits any feature from their parents, instead of the feature getting passed on a random attribute (of the correct data type) is assigned to the child.

- **Step 5:** Repeat steps 2-4 for a prescribed number of generations. Collect, summarize, and analyze data.

The five step algorithm described above is the base algorithm, or overall model architecture used in this study. For individual experiments, different steps were modified to asses their impacts. Also, as the majority of this project was the process of developing the algorithm, the description above represents the final state of the algorithm and differs from earlier experiments. The details of the different experiments and the model's progression over time will be elaborated on in the following section.

It is important to take a moment to discuss the chain length distribution data used in the model. The fitness of the building algorithms is assessed by the similarity of the CLD in the simulation trees to experimental amylopectin CLD data. Amylopectin CLD data is derived experimentally by careful enzymatic partial digestion of isolated amylopectin, followed by chain length analysis using electrophoresis for separation and chromatography for glucan length determination. The CLD data used in this study was largely from maize endosperm and occasionally from barley endosperm for comparison. This data was generously provided by Dr. Eric Bertoft (Åbo Akademi University, Finland) and acquired using the methods described in his paper (Bertoft et al 2008). For

the *sugary-2* mutant studies which will be discussed later, CLD data was from maize endosperm of *sugary-2/cgr04* and *sugary-2/cgx33* mutants, and was generously provided by Dr. Fushan Liu (J.R. Symplot, Idaho, USA) after being collected using the methods detailed in his paper (Liu et al. 2012). The *cgr04* and *cgx33* backgrounds of the *sugary-2* mutants are local maize varieties used in breeding programs in Ontario.

2.2 Experimental Design

The model in this study was developed using an iterative design methodology common to the development of evolutionary algorithms. In this method, once a basic algorithm has been designed, features are isolated and tested to refine the model, determine which features are successful, and eventually incorporate increasing amounts of domain knowledge. Each time the simulation is executed is referred to as a run. The runs can be categorized into sets based on the target or goal of the runs so that each one doesn't have to be enumerated individually. The sets conducted for this project are described in Table 2.2.1 below.

Sets of Runs Conducted		
Set #	**Goal**	**Run Details**
1	Determine fitness function	• Tested different ways of binning the CLD into DP groups, *e.g.* even bins, short/medium/long bins, and no bins • Tested different fitness formulas that include entropy, average branch length, and total branch number with different weightings • Co-variance with different numbers of enzymes

2	Determine default BE action	• Tested different default settings until determining that BE must branch DP 6 or more
3	Adjust number of enzymes	• Tested enzyme number in the range 20,000 to 150,000
4	Adjust field of view (FOV)	• Tested FOV in range 0-40 • Tested DBE having a smaller FOV
5	Adjust number of generations	• Compared fitness over generations in range 0-5000
6	Determine the effect of children averaging parental enzyme activities as an inheritance option	• Tested the effect of removing the option to average parental enzyme activity values, only allowing direct inheritance
7	Adjust mutation rate	• Tested mutation rate in range 0.005% - 1.5%
8	Adjust population size	• Tested the program population size in the range 250-1000 programs per generation
9	Adjust elitism in fitness biased selection	• Tested if eliminating the lowest scoring half or quarter of programs from the breeding pool improves fitness
10	Adjust starting tree lengths	• Tested seed trees in DP range 1-10
11	Make elongation more powerful	• Bolster SS with different strategies, for example: ◦ Make SS twice as likely to be selected compared to BE, DBE ◦ Simulate an SS/BE and SS/DBE complex by having SS act before each BE and DBE action ◦ Increase the DP of elongation
12	Modify branching actions, including adding a second branching enzyme	• Tested allowing BE to only add branches if there are no other branches in a neighbourhood of six characters • Tested having BE randomly add

		glucans of different lengths within specified ranges so that it is not always a fixed length branch being added • Tested adding a second branching enzyme BE2 which adds glucans in a different specified DP range from BE1
13	Modify debranching actions	• Tested allowing DBE to only debranch small glucans by setting a maximum length, in the range 6-40 • Tested allowing DBE to only debranch if there are nearby branches
14	Use barley instead of maize	• Tested using the barley endosperm CLD instead of maize
15	Comparison to *sugary-2* mutant	• Tested adding a suite of 3 SS enzymes, each controlled with their own activity level, SS1 being responsible for substrates of DP <10, SS2 DP 10-24, SS3 DP >24
16	Add enzyme contexts as evolving regular expression strings	• Compare predefined contexts against evolving contexts • Tested different lengths of the regular expression string in the range 6-25

Table 2.2.1: Outline of the different sets of runs conducted and their objectives.

2.3 Data Analysis and Other Tools

Data from the runs was collected and analysed in multiple steps. Raw data from the model was fed into several custom made C++ and Bourne shell scripts to create summaries of fitness values, tree lengths, and number of branches per tree for each generation and aggregate the data between experiments. The summarized data was

further analysed in LibreOffice Calc, a free and open source spreadsheet tool. The model also generated SQL formatted raw data that can be viewed with any SQL database tool. In the database, each run, generation, and individual are uniquely coded so that it is easy to query. Novel tools were written in C++ for generating SVG formatted vector images of the trees created by the model. The drawing tool could have broad applications for creating crisp two-dimensional drawings of complex polymers, and was useful for visually summarizing data which, in its raw format, is a barely human readable string. Word cloud images presented in Chapter 3 for graphical representation of the diversity of enzyme contexts were generated using the Word Cloud Generator written by Jason Davies (https://www.jasondavies.com/wordcloud/), which is based on free and open source code available on GitHub.

Chapter 3

Results

This chapter summarizes the results of the experiments outlined in the previous chapter. The first section focuses on results regarding the development and refinement of the model including parameter testing. The second section focuses on results from the experiments that are about the enzymes involved in amylopectin biosynthesis. The third section presents the results of the comparison with the *sugary-2* mutant. The last section shows the results of evolving enzyme contexts as strings of regular expressions.

The success of a each experiment conducted using the model is judged primarily by the CLD of the trees produced. The experimental (model) CLD is compared to actual CLD data using NRMSE (normalized root mean square error), typically the average NRMSE of all the trees produced in the last generation. The lower the NRMSE value, the closer the tree produced resembles an actual amylopectin molecule's branching pattern.

3.1 Model Design and Refinement

The most important step in the refinement of the model was the incorporation of the domain knowledge that branching enzyme must attach branches a minimum of six long for it to be possible to evolve branched trees. This is because in the actual CLD data there are no branches less than six glucoses long, and biochemical characterization of branching enzymes from a variety of sources indicates the minimum DP for branching is in this region (Tetlow and Emes 2014). If BE attaches branches less than six DP, the model always evolves a zero activity for BE to avoid creating any short branches that

would be scored against the desired value of zero. With zero BE activity, only straight chains without branches form.

3.1.1 Fitness Measurement Strategy

In the early stages of model development it was unclear if the fitness function should be based on CLD alone. Different fitness functions were tested before it was decided that the best fitness function is directly the NRMSE of the CLD. One feature that was strongly considered for inclusion in the fitness function is entropy. Entropy is basically a factor that would encourage an even distribution of branch lengths. It was expected that by weighting entropy appropriately the model could yield better results in terms of CLD, especially when combined with different methods of binning the CLD. Regardless of weighting and binning, entropy did not prove to be a useful addition to the fitness function. Figure 3.1.1.1 below shows the negative impact of entropy on CLD accuracy using a six bin CLD. A similar trend was seen with every CLD binning strategy tested and remained unchanged if enzyme number was modified. After thorough testing it was determined that entropy being included in the fitness function does not improve CLD similarity.

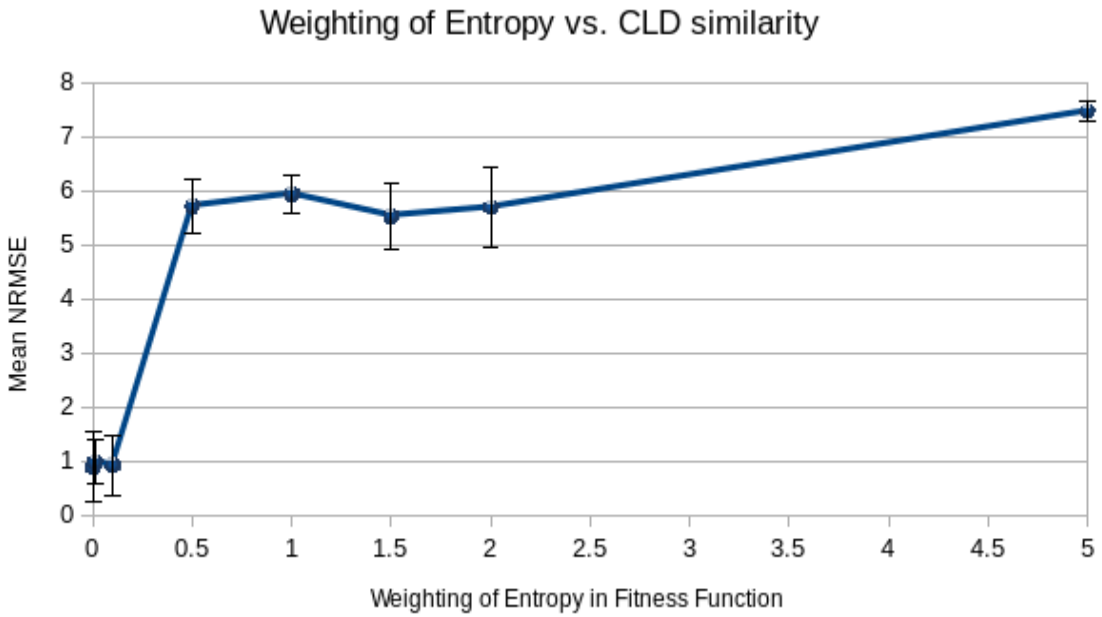

Figure 3.1.1.1: The average NRMSE of the trees produced by experiments with different weightings of entropy in the fitness function. Error bars represent standard deviation.

The representation of the actual CLD data used by the model is a non-trivial factor in model accuracy. The CLD can be binned by different methods, strategically grouping similar DPs to make the CLD more or less difficult for the algorithm to match. Different methods of binning were tested during development of the model. Typically data was rounded to the nearest 0.1% unless otherwise specified. The initial method used was six characteristic bins of short, medium, long, and very long chains, along with head and tail bins for glucans longer and shorter than those typically found in amylopectins. The advantage of the six bin method is that it is representative of the botanical source and enzymatically relevant while remaining simple. Another method used was to have twelve bins of equal sizes. The advantage of equal sized bins is that there is no fitness advantage to scoring better in one bin than another. The twelve bin method was tested as both near

equal bins using the true data, rounded to the nearest 0.1%, and exactly equal bins made from rounding the near equal bins to the nearest 10%. Lastly, binless data was used by simply not grouping DP's. This is an ideal way of representing the CLD data because it lacks modification and shows similarity between simulated and actual amylopectins at the highest resolution.

The results of the different binning experiments were highly counter intuitive, and thus very interesting. The six bin system, despite being the simplest and most closely tied to enzyme activities, proved the most difficult to evolve to and yielded the highest NRMSE. The twelve bin method yielded the lowest NRMSE. Surprisingly, the almost even bins outperformed the exactly even rounded bins. However, the twelve bin method resulted in surprisingly short trees not representative of real amylopectin that did not increase in size proportional to the number of enzymes provided, a pattern which did not occur in any other binning system. The twelve bin method was discarded because the small and unpredictable tree sizes are indicative that, while the NRMSE may show a small improvement over the other methods in the short term, the ability of the system to improve with the addition of new features may be limited because the trees are too small to be representative of amylopectin structure and thus may not respond appropriately to the incorporation of domain knowledge. Ultimately, the binless method proved the best and most reliable, because it yielded adequate CLD results, produced large trees, and responded in a consistent and predictable way to different numbers of enzymes. The binless method also yielded surprisingly fast runtimes, essentially equivalent to runs with fewer bins. The binless method is also ideal because it has the least potential for bias and

is the highest resolution representation of the real CLD. Upon discovery that the binless method is the best representation of the CLD, all subsequent experiments were conducted using binless CLD data. The figures 3.1.1.2 and 3.1.1.3 below show the results of the different binning methods.

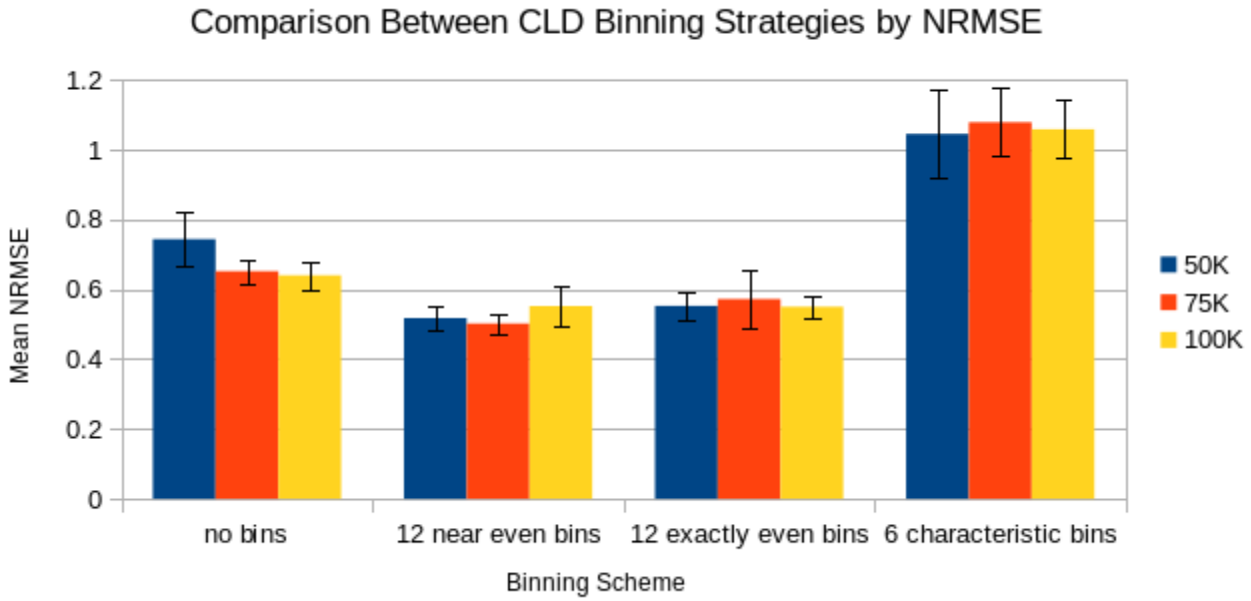

Figure 3.1.1.2: Different binning strategies are compared in terms of resultant NRMSE. The error bars represent standard deviation. The different columns represent testing the different binning strategies with different numbers of enzymes, from left to right they are 50,000, 75,000, and 100,000.

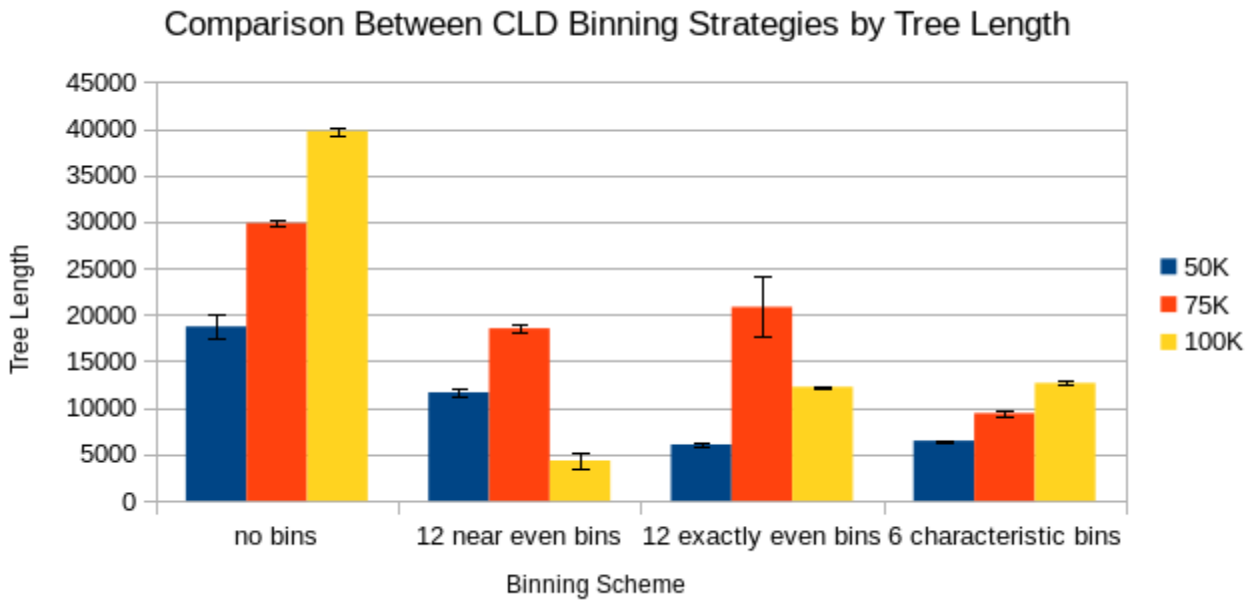

Figure 3.1.1.3: Different binning strategies are compared in terms of resultant tree length. Tree length is defined as the total number of characters in the tree. The error bars represent standard deviation. The different columns represent testing the different binning strategies with different numbers of enzymes, from left to right they are 50,000, 75,000, and 100,000.

3.1.2 Number of Enzymes

The number of enzymes in the simulation has a huge impact on accuracy, but is also a readily optimizable parameter because it is directly correlated to runtime and enzyme catalytic activity. In essence, the goal is to provide the maximum number of enzymes that does not result in the simulation simply evolving lower enzyme activities to compensate, while maintaining a feasible runtime. Figure 3.1.2.1 shows how NRMSE improves with an increasing number of enzymes until it plateaus around 75,000 enzymes. This plateau is the optimal number of enzymes because it means that further increasing the enzymes just results in evolving lower enzyme activities to compensate. Figure 3.1.2.2 shows the direct correlation between runtime and number of enzymes. All subsequent enzymes used 75,000 enzymes unless otherwise stated.

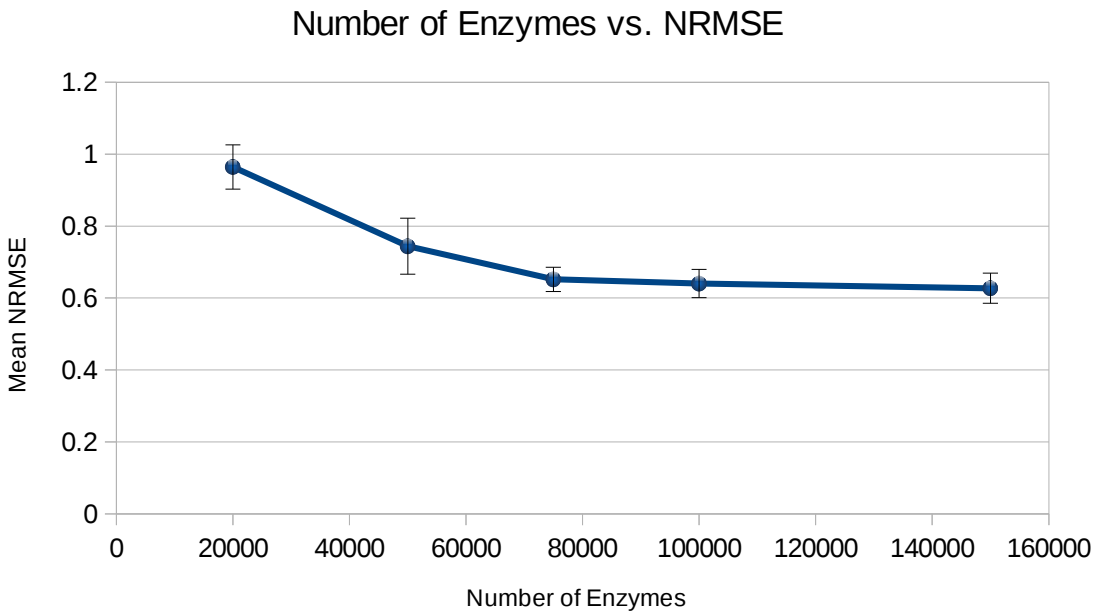

Figure 3.1.2.1: A graph showing the improvement in NRMSE with an increasing number of enzymes. The plateau around 75,000 wasps shows where the model has enough enzymes to maximize the evolved enzyme activities. Error bars represent standard deviation.

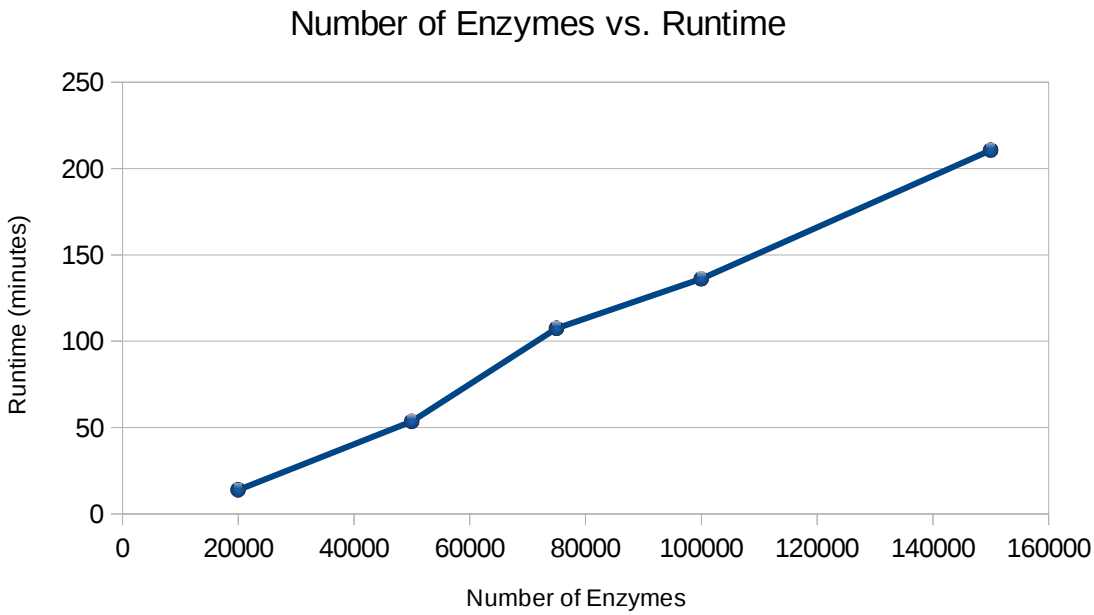

Figure 3.1.2.2: A graph of the runtime increasing proportional to the number of enzymes.

A parameter related to the number of enzymes is the length of the seed tree provided at the start of the algorithm. The length of the seed tree was tested in the range of DP 1-10 but found to have no impact on fitness and so was left at DP 1 for simplicity. This is likely because even a 10 long seed is only a tiny fraction of the 25,000 elongation enzymes provided and quickly becomes irrelevant as the tree grows.

3.1.3 Number of Generations

As discussed in Chapter 2, every evolutionary algorithm needs a stopping point. In this model, the termination criteria is a set number of generations. It was observed in this model that the fitness plateaus over time, as is typical for evolutionary algorithms. Figure 3.1.3.1 shows the NRMSE over 5000 generations. Fitness improvements occur in waves as new innovations arise until around generation 900 where the fitness stabilizes. The average fitness at generation 1000 was equivalent to the average fitness at generation 5000. Due to these results, 1000 generations was determined to be a good termination criteria and used consistently for all subsequent experiments.

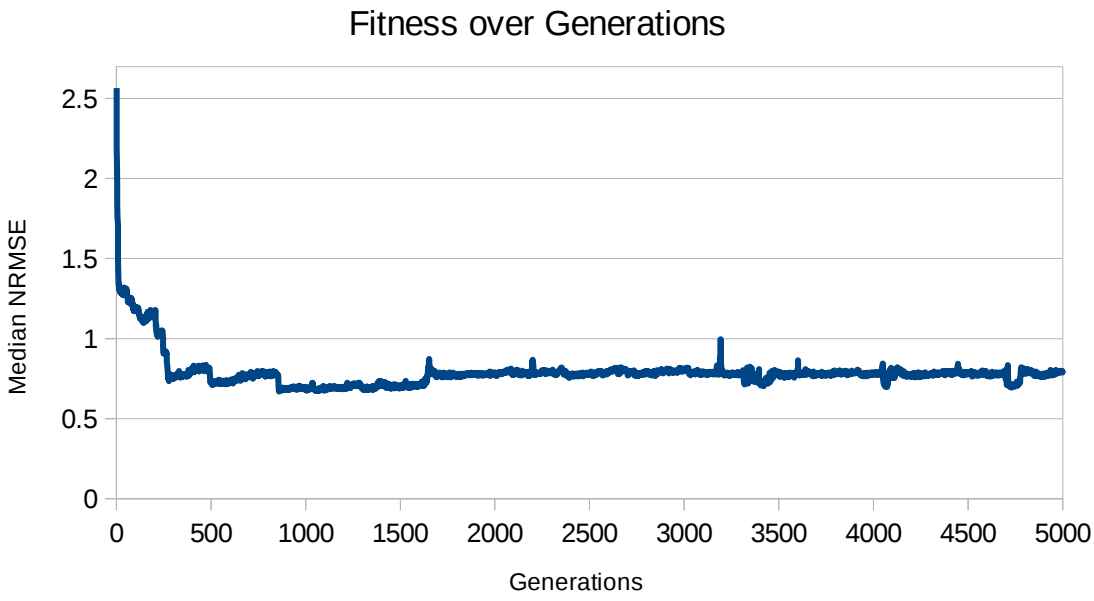

Figure 3.1.3.1: A graph of the median NRMSE over 5000 generations showing stabilization around generation 900.

3.1.4 Breeding Parameters

An unusual feature in this model was the inclusion of the option to average parental enzyme activity levels as a form of inheritance, in addition to the option of directly inheriting the activity level from one parent or the other. A traditional inheritance system would be where one of the parents is selected at random and the enzyme activity is taken from that parent. The averaging system creates a third option of averaging the two parents, and is weighted double so that it is is equally likely that the child inherits an averaged or parental value. Both the traditional and averaging methods were compared for this model before the averaging model was adopted as the default. The traditional method performed on average 30% worse than the averaging method in terms of NRMSE. The averaging system had an average NRMSE of 0.70 ± 0.04 while the

traditional system had an average NRMSE of 1.05 ± 0.15. The averaging system may confer an advantage because it allows for partial preservation of synergistic relations between enzyme activities, resulting in fewer offspring that have very poor enzyme coordination due to the inheritance of disparate values from parents.

A very important parameter in an evolutionary algorithm is the mutation rate. Both high and low mutation rates can be detrimental to fitness and cause excessive variability between experiments. For this reason, mutation rates in the range 0.005% - 1.5% were tested to determine the optimal mutation rate for the model. Figure 3.1.4.1 shows the impact of different mutation rates on NRMSE. The optimal mutation rate was determined to be 0.5% because it was the value that not only yielded the lowest NRMSE but also had the smallest variance between individuals in the population.

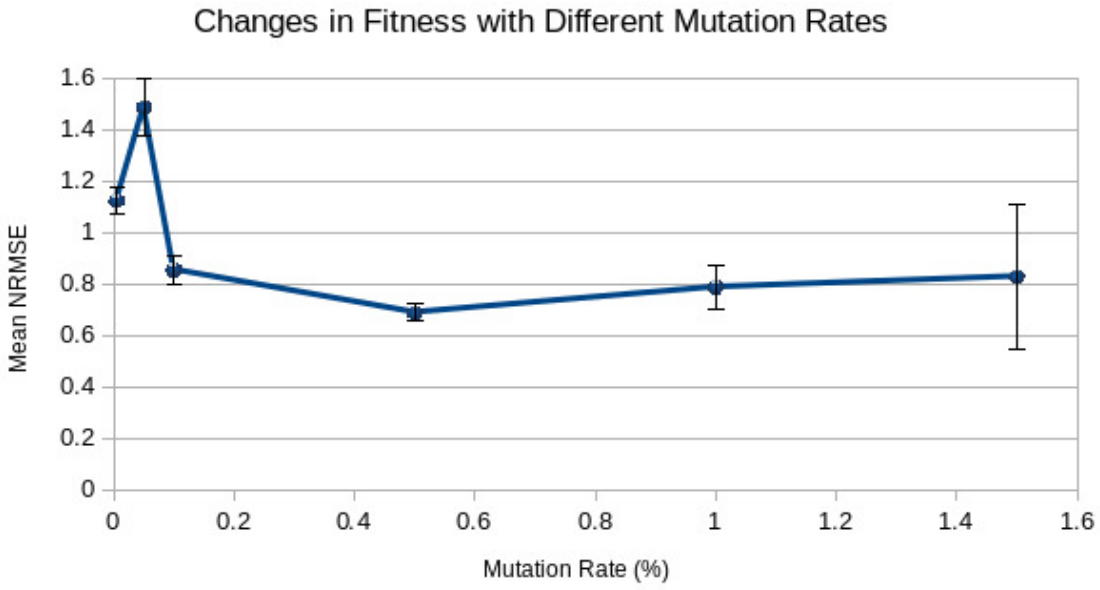

Figure 3.1.4.1: The impact of mutation rate on NRMSE. The error bars represent standard deviation.

The fitness biased selection method used in this model (described in detail in Chapter 2) is a relatively soft selection method. In order to quickly test if a more elitist selection algorithm would be beneficial for the model without having to make major adjustments, the method of simply eliminating the bottom half or quarter of the population from the breeding pool was used. Eliminating part of the breeding pool resulted in a slight but not statistically significant NRMSE increase, suggesting that the selection method used is already sufficiently selective to create the best possible stable population at generation 1000. It was determined from these experiments that the threshold selection method is sufficiently selective and that reducing the size of the breeding population is not a feature to retain because it poses the risk of reducing diversity with no evidence of benefit.

3.1.5 Population Size

Population size is an important parameter in an evolutionary algorithm because a population too small does not have sufficient diversity for a good solution to be found, but a population too large can take a very long time to evolve and become computationally expensive. Population sizes in the range of 100-1000 were tested to determine the optimal size for this model. Figure 3.1.5.1 shows the impact of population size on average, median, and minimum NRMSE after 1000 generations. The smallest population of 100 programs had the lowest NRMSE and also the least discrepancy between the average and median NRMSE. The smallest population also had the shortest runtime. All subsequent experiments used a population size of 100 as the result of this investigation.

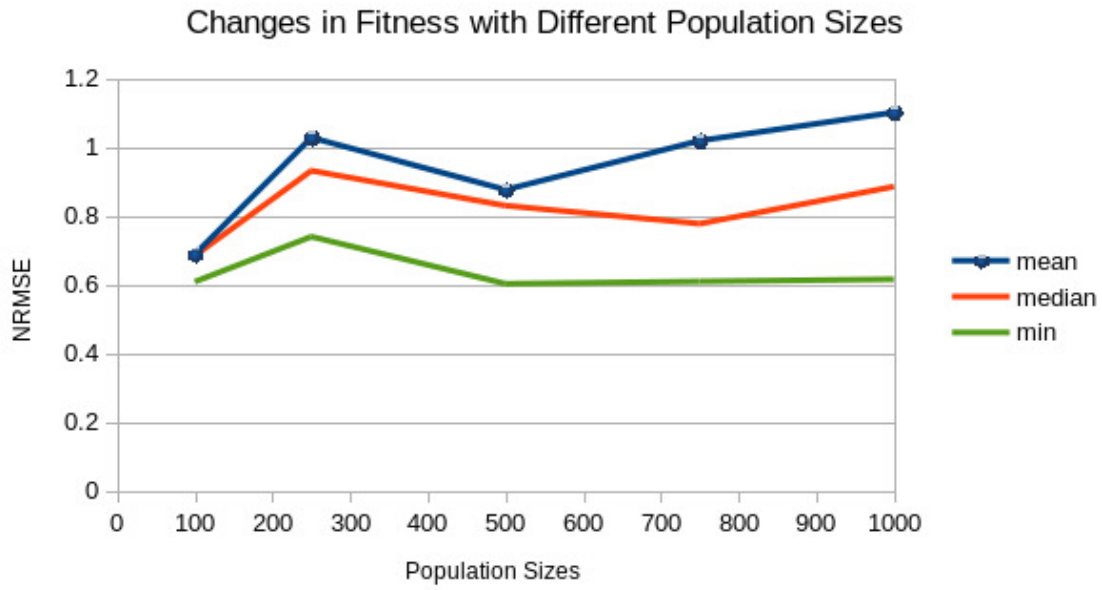

Figure 3.1.5.1: A graph of the impact of population size on average, median, and minimum NRMSE.

3.2 Results Regarding Amylopectin Enzymes

The result discussed earlier that branched trees can only form when branching enzyme attaches branches of a minimum six DP matches biochemical data on branching enzymes (Tetlow and Emes 2014). It is an interesting result that not only did the six DP branch minimum allow the creation of branched trees, it also created trees that have a branch frequency that matches amylopectin at 5% (Manners 1962). This was highly consistent between experiments and resilient to all modifications the model underwent. All of the model trees described in this chapter maintained a branch frequency of 5% ± 2%, with the vast majority (over 90%) falling between 5.0% and 5.6% branch frequency. Real amylopectin does not have much variance in branch frequency, so the consistency of the branch frequency in the experimental trees serves as validation of the model.

An interesting feature of the basic starch biosynthetic enzyme suite (SS, BE, DBE) is that DBE is inherently much more powerful than the other two enzymes; *i.e.* if DBE is able to cleave anywhere on the amylopectin, a single action can remove a sub-tree requiring many SS and BE actions to build. The catalytic power of DBE is probably kept under control in many ways *in vivo*. For example, in a real plant system the cleaved subtrees do not simply disappear as they do in the model, and could function as building blocks. DBE could be controlled by having a lower concentration or activity level than SS and BE, posttranslational modification, a more stringent substrate specificity, physical blocking by hydrogen bonding or steric hindrance, and many other possible methods. In the first rendition of this model, the only control on DBE is enzyme activity. Early results form this model made it clear that enzyme activity alone is insufficient to control DBE, because early versions of the model consistently evolved very low (below 5%) and typically zero DBE enzyme activity. Many of the experiments discussed below feature methods for specific suppression of DBE relative to the other enzymes which showed varying degrees of success. The discovery of insights such as the need for DBE control and potential methods of control that may be used by plants is exactly what this model aims to accomplish. The immediate result of DBE activity approaching zero in the absence of sufficient controls functioned as excellent validation of the model.

Visualizations of the trees created by different experiments show the diversity of structures created by the model. Figure 3.2.1 below contains drawings of trees using the default parameters established in section 3.1. The drawings were created using a custom C++ script which could be used to create two dimensional representations of any large

branched polymers that can be represented by a string. The lengths of branches in the drawing are directly proportional to number of characters representing the branch in the string. The red circle highlights the starting point or seed tree for the algorithm.

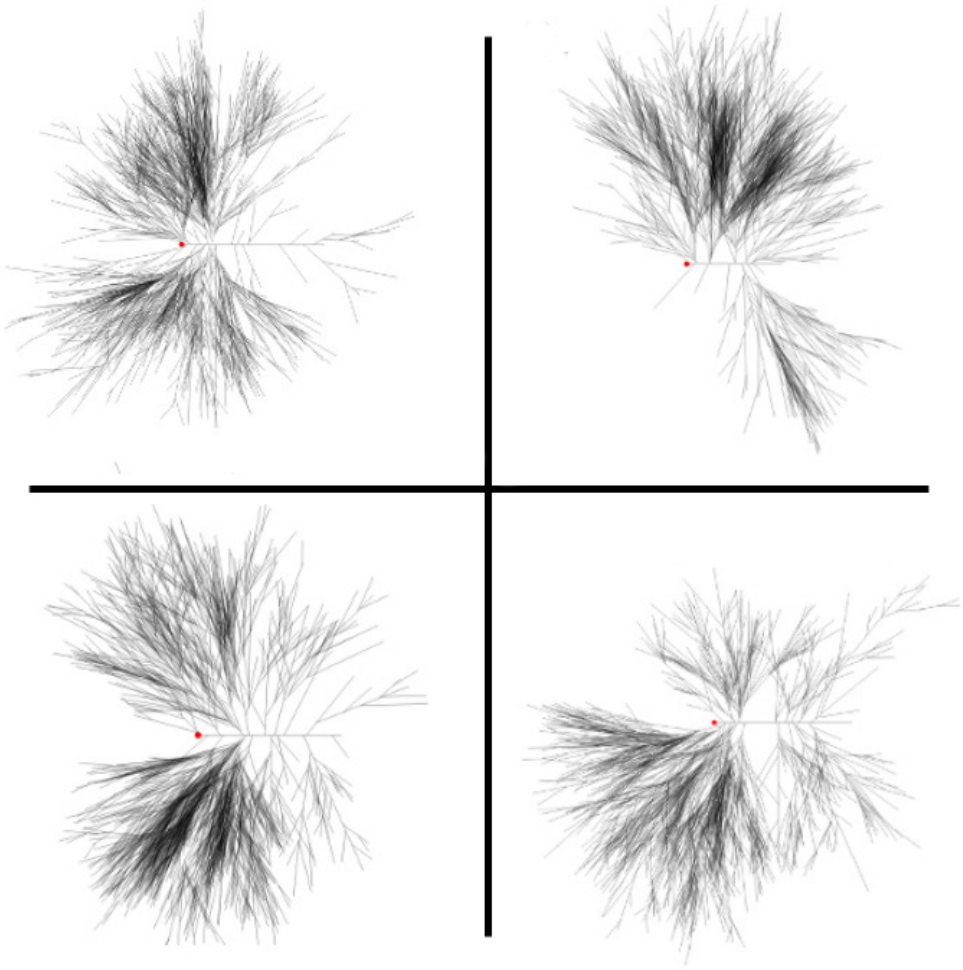

Figure 3.2.1: Four example trees produced by the model using the default parameters established in Section 3.1. The trees represent some of the diversity of structures achieved by the model. Branch lengths are proportional between the string and the image. The red circles mark the seed tree.

3.2.1 Field of View

An important enzyme parameter is field of view, *i.e.* how wide of an area each enzyme can scan. Figure 3.2.1.1 shows the impact of different fields of view on NRMSE. From this investigation, a field of view of 20 was selected as the default because it has both the lowest NRMSE as well as the smallest variance. Experiments with giving DBE a separate field of view of 0 or 4 while SS and BE have the default 20 resulted in a 2% fitness improvement.

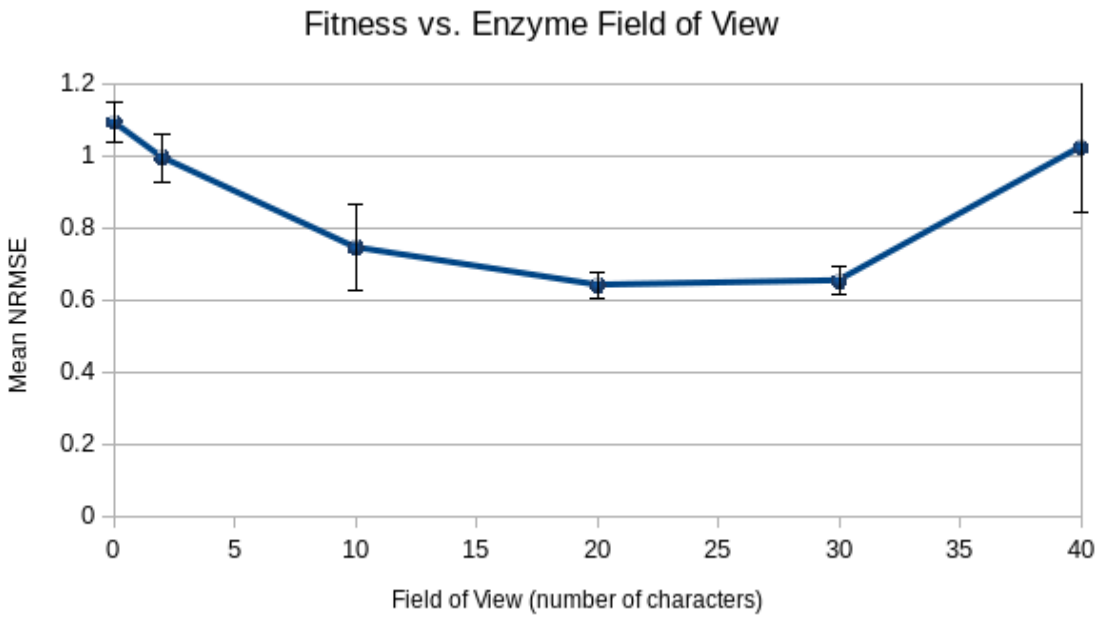

Figure 3.2.1.1: The impact of enzyme field of view on NRMSE. Error bars represent standard deviation.

3.2.2 Modifications of Starch Synthase

SS is inherently the weakest enzyme in the biosynthetic suite because it can only polymerize at a non-reducing end. Attempts at improving the CLD by making SS more powerful relative to BE and DBE were limited in their success. The model default is to provide an even distribution of enzymes to the three enzyme classes. One attempt at bolstering SS was to skew the relative enzyme distribution from 1:1:1 (SS:BE:DBE) to 2:1:1, and 3:1:1. Skewing the enzyme distribution further in SS's favor yielded no change in NRMSE despite doubling the average tree length.

An experiment where SS appends four G's to the non-reducing end instead of the default one G resulted in a 70% increase in NRMSE. This experiment also yielded an average branch frequency of 7%, the highest discrepancy in branch frequency of all the experiments from the desired 5%. This experiment also had a characteristic enzyme activity profile with a 20% decrease in SS activity, 24% increase in BE activity, and 3% increase in DBE activity. The average tree length was 2.5 times the average tree length without the modification. This experiment made it clear that while SS does need to be more powerful, it needs to be more controlled as well, which will be discussed further in section 3.3 in the context of the *sugary-2* mutant.

3.2.3 Modifications of Branching Enzyme

To encourage branches to be more spaced out, an experiment was conducted where branching enzyme could only act if there were no branches within three characters on either side of the desired branch point. The three character spacing was selected because a DP of six is the minimum length required for the formation of hydrogen bonded helices,

and so spacing branch points minimally six characters apart allows for the formation of tightly packed clusters. The experiment resulted in a 2% decrease in average NRMSE, but made the minimum NRMSE 10% higher. It appears that forcing branches to be more spaced improves the CLD for the worse scoring individuals in the population, but worsens the CLD for the better scoring individuals in the population. The interesting implication of this result is that controlling enzyme activity alone can yield better branch spacing than forcing branches to be a minimum distance apart.

One difference between the BE in the model and *in vivo* is control over branch length. In the first rendition of the model, BE would simply create the α(1,6) branch point of uniform six DP and leave SS entirely responsible for determining the branch's final length. In reality, BE attaches branches of different lengths depending on isoform and possibly other factors as well. In order to better replicate the *in vivo* function of BE, an experiment was conducted where BE could add a branch of a random length in the range 6-10. The results of this experiment were impressive, with a 14% improvement to the average NRMSE, and a 6% improvement to the minimum NRMSE.

One of the primary goals of this model is to determine the minimum suite of enzymes required to create the CLD of amylopectin. A series of experiments were conducted to see the impact of adding a second BE to the suite of enzymes. Figures 3.2.3.1 and 3.2.3.2 below contain drawings of example trees produced by the experiments with two BE isoforms. Adding a second BE affords the model independent control of adding branches of different lengths. Because of the success of the prior experiment in which the BE added branches in a range (6-10), the second BE was added with a range as well. Six

different sets of ranges were tested, and the results of each can be found in Figure 3.2.3.3 below. Most of the two BE configurations did not function as well as the single BE with a range of DP's. One configuration, with BE-1 having the range 6-8 and BE-2 having the range 11-13 did equally well to a single BE with a range of 6-10. This is interesting because in the CLD data there are similar frequencies of branches in the 6-8 range and similar frequencies of branches in the 11-13 range. One would assume that being able to control independently how many branches in each category are created would confer better conformity to the desired CLD. However, it appears that BE control is not the limiting factor to creating an improved CLD, and that, at this stage of the model, adding a second BE to the enzyme suite is inconsequential.

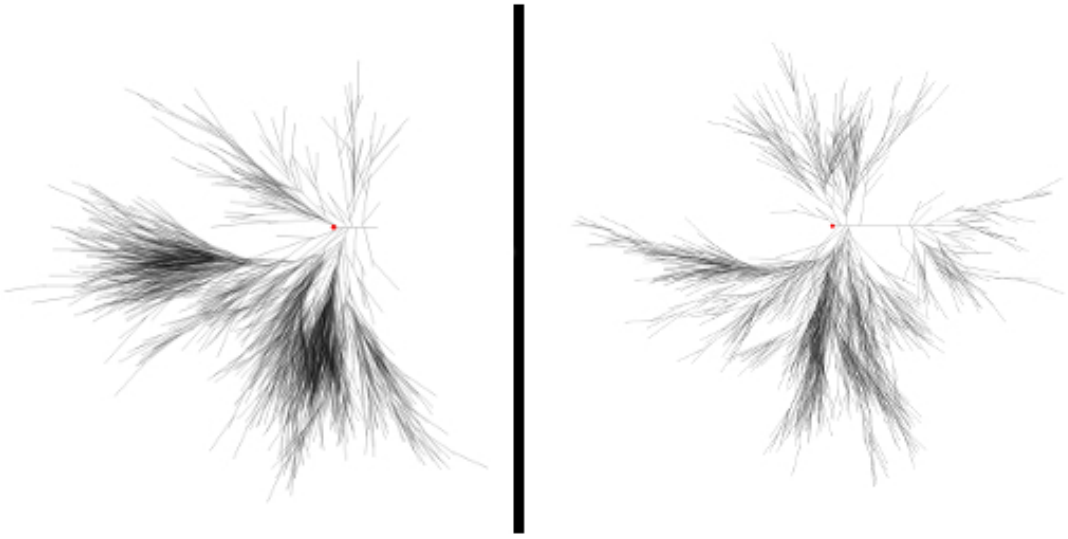

Figure 3.2.3.1: Drawings of two example trees from the experiment with two BE isoforms where BE1 added branches of DP 8-12 and BE2 added branches of DP 13-17. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

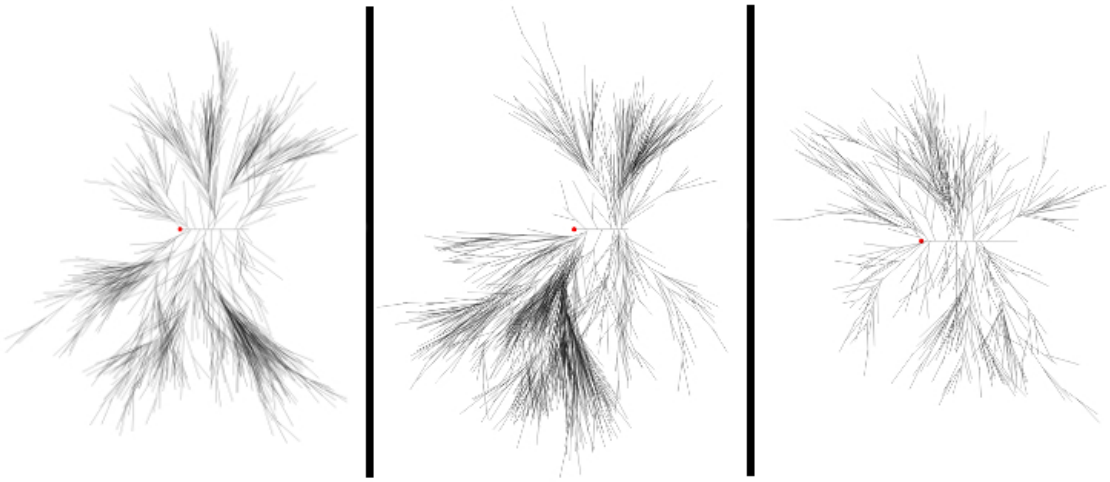

Figure 3.2.3.2: Drawings of three example trees from the experiment with two BE isoforms where BE1 added branches of DP 6-8 and BE2 added branches of DP 11-13. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

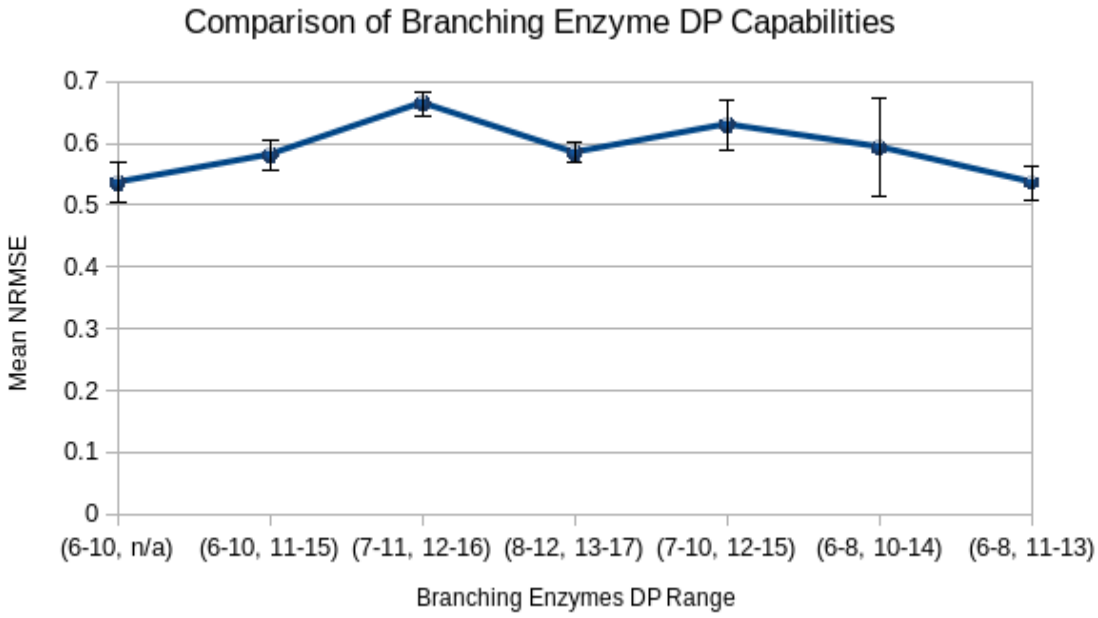

Figure 3.2.3.3: The impact of adding a second BE of different range capabilities on NRMSE. The labels on the x-axis show the branch length ranges of BE1 and BE2, respectively, for each trial. Error bars represent standard deviation.

3.2.4 Modifications of Debranching Enzyme

A series of experiments were conducted in which DBE could only cleave branches shorter than a set maximum length, or only cleave branches that are close to other branches within a set neighbourhood. Figure 3.2.4.1 below shows the results of these experiments, comparing the average NRMSE of a run without DBE modification, eight runs where DBE can cleave branches less than 6, 10, 12, 14, 18, 22, 28, or 40 long, one run where DBE can cleave branches only if they are within 6 characters of another branch, and one run where DBE can only cleave branches less than 10 long within 6 characters of another branch. Surprisingly, none of these imposed controls on DBE resulted in any significant improvement to the CLD. The most stringent controls, namely those where DBE could only cleave branches less than 6 or 10 long or within 6 characters of another branch yielded nearly identical results to having no additional controls on DBE. Meanwhile, the weaker controls resulted in worse CLDs in a manner not proportional to control stringency.

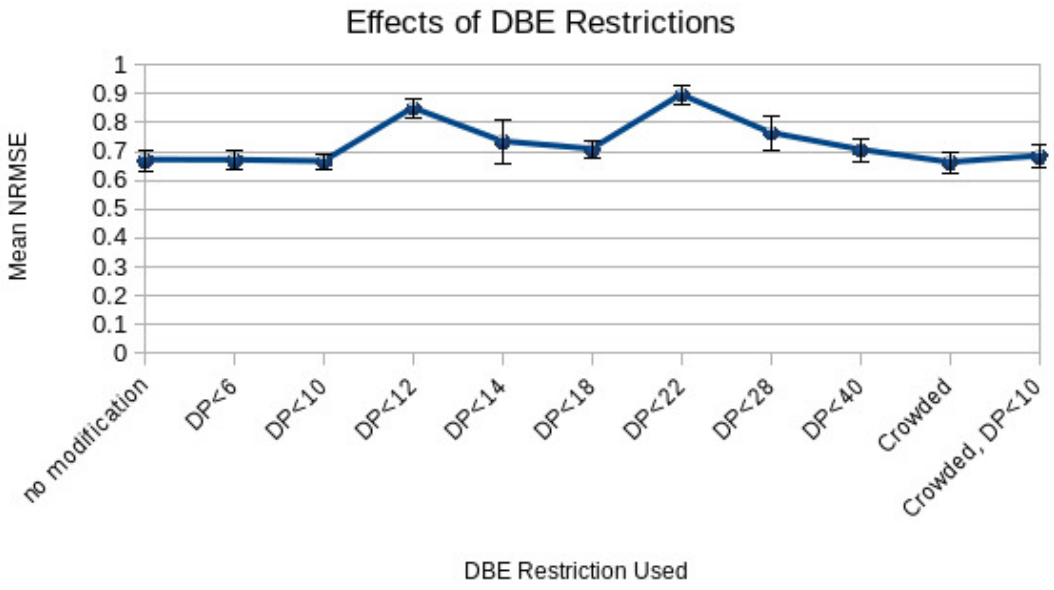

Figure 3.2.4.1: The impact of different DBE controls on mean NRMSE. Error bars represent standard deviation. The controls included restricting DBE to cleaving branches a maximum of 6, 10, 12, 14, 18, 22, 28, or 40 long, restricting DBE to branches that are "crowded", i.e within 6 characters of another branch, or restricting DBE to branches that are both crowded and less than 10 long.

The different DBE modifications yielded characteristic enzyme activity profiles, summarized in Figure 3.2.4.2 below. For example, restricting DBE to cleaving only branches a maximum of 6 long resulted in much higher average DBE activity. Since the NRMSE did not change with this DBE modification, the unique enzyme activity profile suggests that the model evolved a new strategy to compensate for the DBE modification. It is interesting that the DBE activities are in no way proportional to the stringency of the modification applied. Tree visualizations were created for the runs where DBE could cleave maximum 6 long and 10 long branches, to see if the change in DBE activity have any visually discernible impact on the trees. The drawings can be found in Figures 3.2.4.3 and 3.2.4.4 below. The trees with higher DBE activity do have more spaced out branches

and have visually interesting structures. It is important to note that the trees from both runs have the same branch frequency, so trees that appear more sparse do not have fewer branches but differently distributed branches.

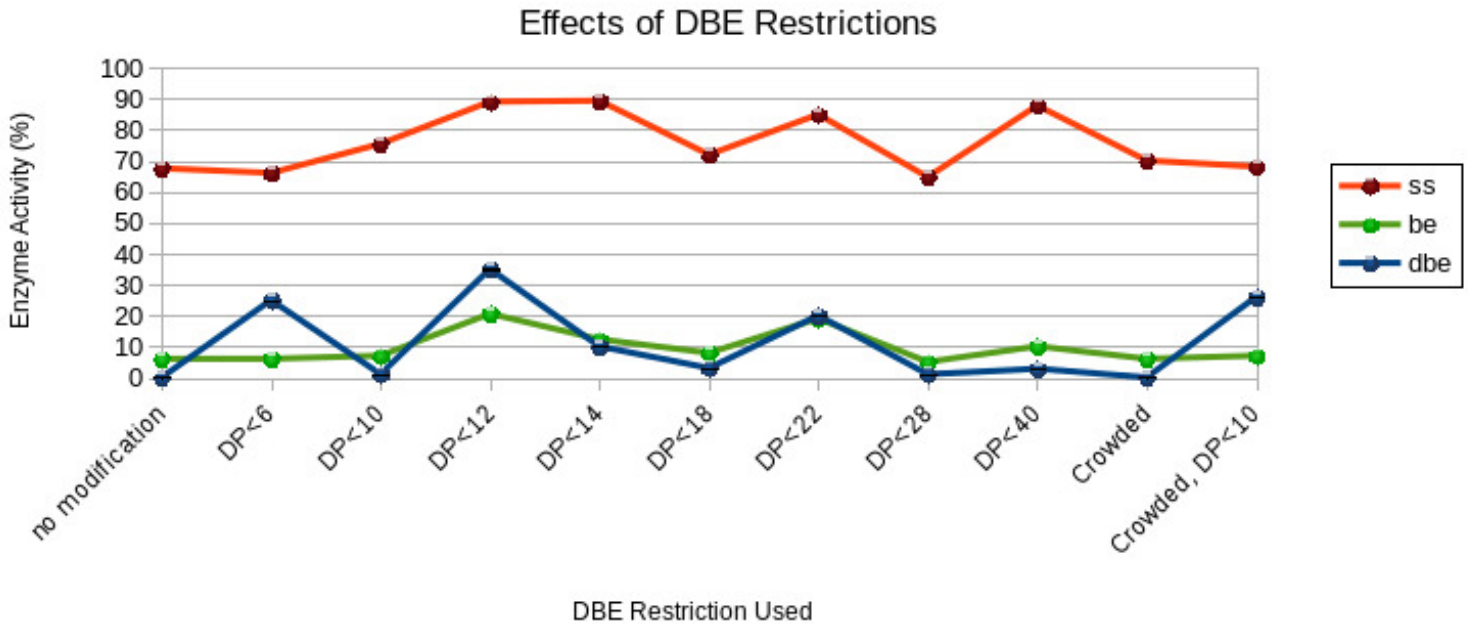

Figure 3.2.4.2: Enzyme activities for SS, BE, and DBE with different DBE controls applied. The controls included restricting DBE to cleaving branches a maximum of 6, 10, 12, 14, 18, 22, 28, or 40 long, restricting DBE to branches that are "crowded", i.e within 6 characters of another branch, or restricting DBE to branches that are both crowded and less than 10 long.

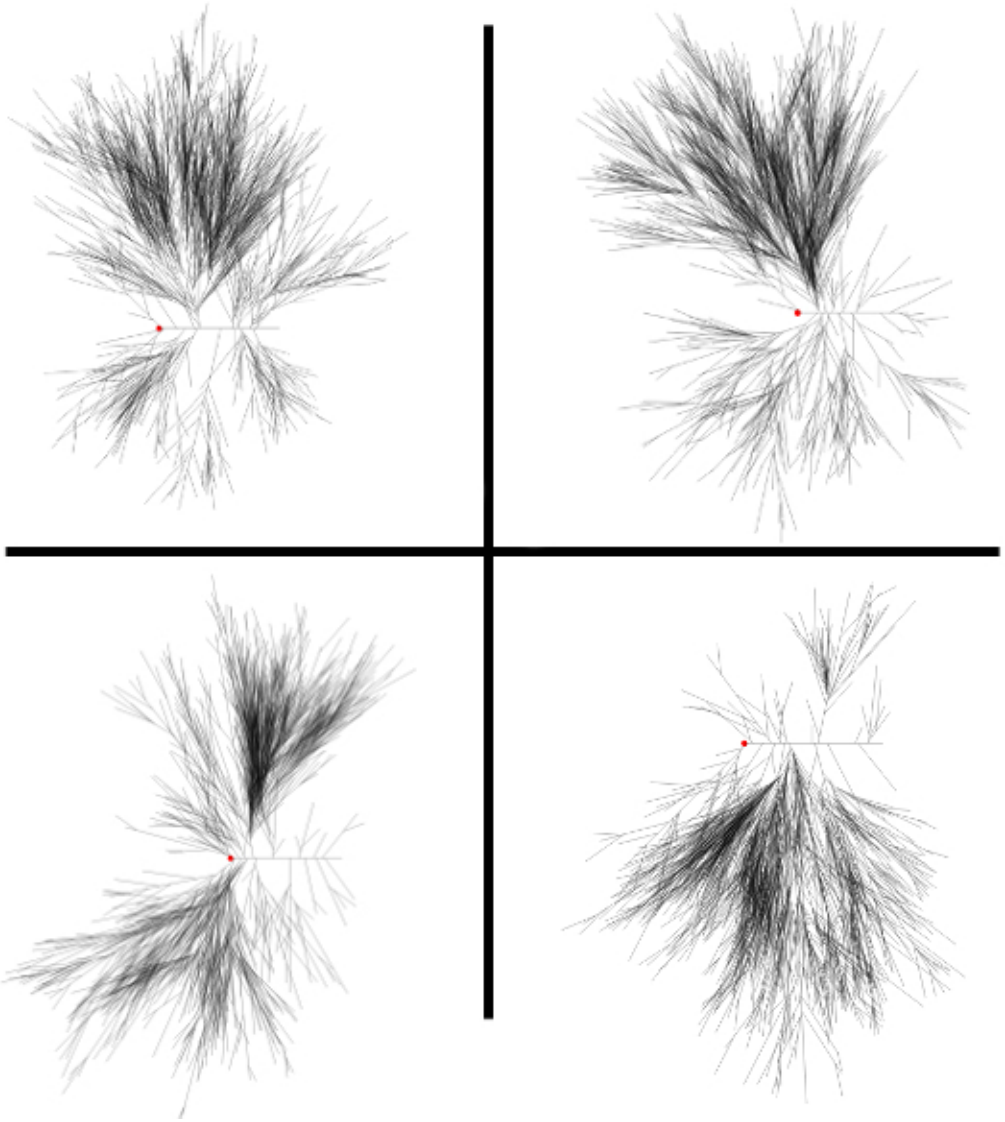

Figure 3.2.4.3: Drawings of four example trees from the run where DBE could only cleave branches a maximum of 10 long. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

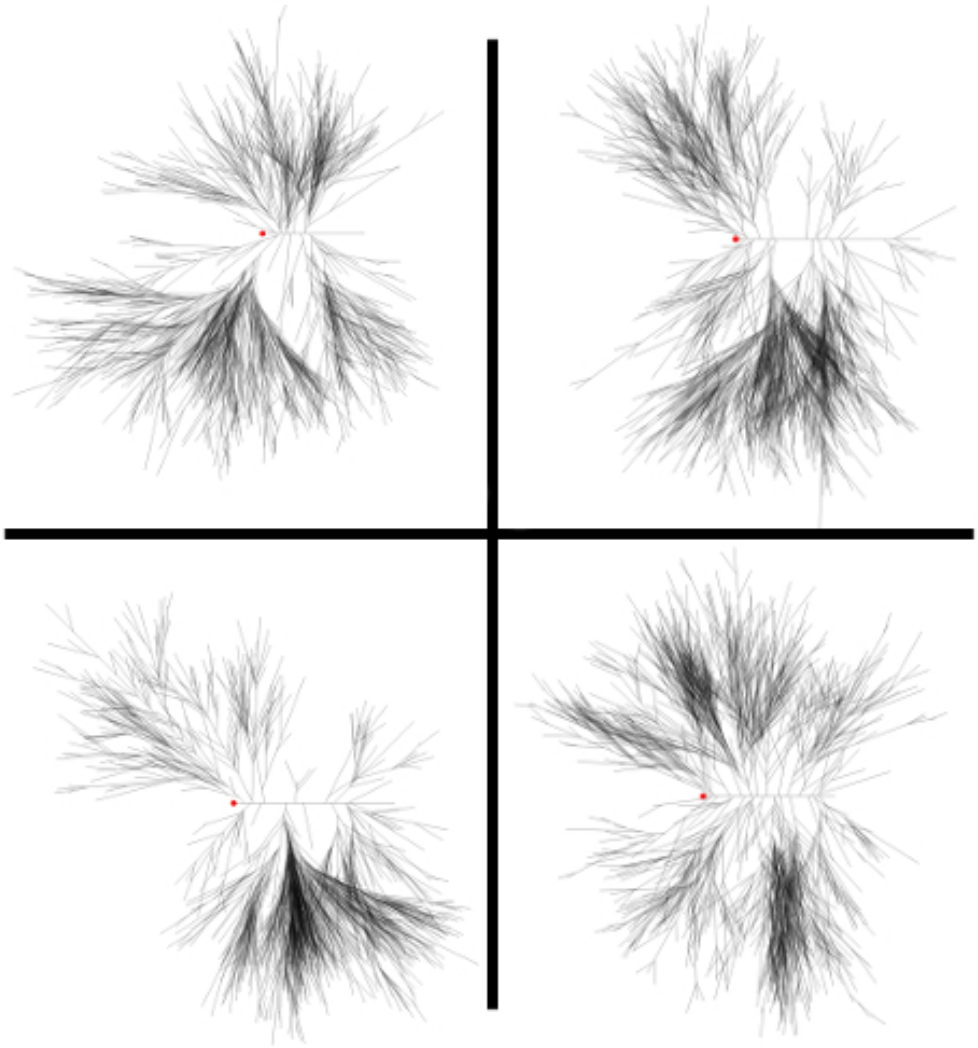

Figure 3.2.4.4: Drawings of four example trees from the run where DBE could only cleave branches a maximum DP six. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

3.2.5 Comparison Between Maize and Barley CLD Profiles

An experiment using CLD data from waxy barley instead of waxy maize was conducted to determine if any species specific trends in enzyme activity could be elucidated using the model. The mean NRMSE was very similar between barley and maize, the former being 0.70 ± 0.03 and the latter 0.67 ± 0.04, and the minimum NRMSE also similar at 0.61 for barley and 0.60 for maize. Analysis of the top ten scoring individuals' enzyme activities showed that the best strategies for maize and barley were very similar. BE and DBE had identical activities in maize and barley, while SS activity was 7% higher in barley than in maize. Figure 3.2.5.1 below contains drawings of example trees created using the barley CLD. Seeing the consistent difference in SS activity between species is an interesting start at using the model to better understand the biosynthetic reasons for differences in starches from different botanical sources, and repetition of this experiment with updated versions of the model with improved accuracy would be a very promising direction for future research.

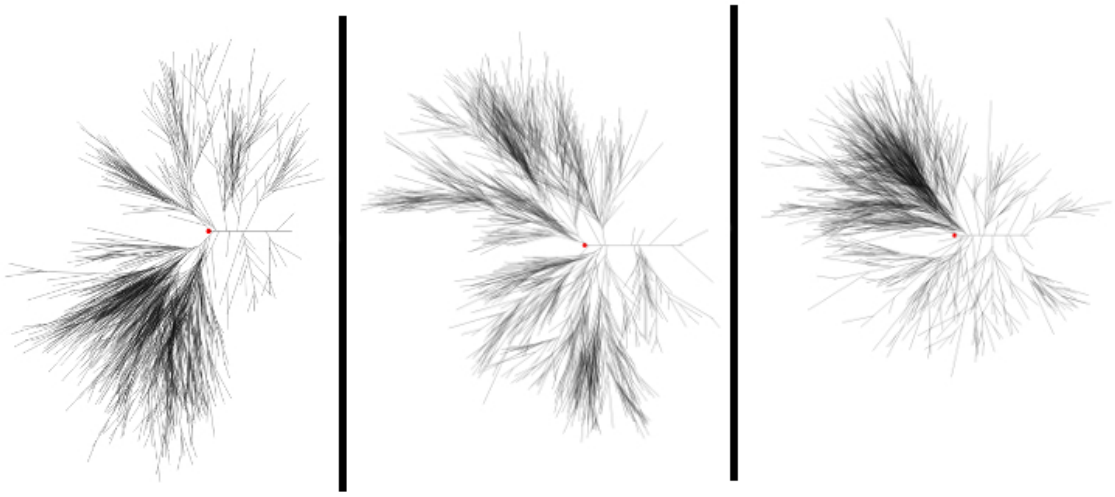

Figure 3.2.5.1: Drawings of three example trees from the experiment using barley CLD instead of maize. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

3.3 Comparison of the Model to Natural Mutations of the Starch Biosynthetic Pathway

The *sugary-2* mutant in maize lacks SSIIa (the endosperm specific SSII) activity and has characteristic CLD differences from normal maize (Liu et al 2012). This section describes the result of experiments involving the addition of multiple SS isoforms and comparisons with the *sugary-2* mutant CLD. There are several reasons to further investigate SS isoforms. As discussed earlier in section 3.2.2, modifications that make SS more powerful are of limited use without adding more control as well, suggesting that using additional starch synthases may be more beneficial than simply bolstering the existing SS. Additionally, the CLD of the amylopectin produced by the model showed surprising similarities to the CLD of the *sugary-2* mutant, both displaying an excess of short chains of DP 6-10 and decreased intermediate chains of DP 12-30. The addition of SSII functionality to the model is a way to both validate the model by demonstrating that

it is representative of a natural system, as well as provide insight on the importance of the SS isoforms and evidence that they are essential components of the minimum enzyme suite for amylopectin synthesis.

3.3.1 Comparison Between s*ugary-2* and Experimental CLD

The CLD of the experimental amylopectin trees produced by the model had similar features as the CLD of the maize *sugary-2* mutant. Both have an excess of short chains in the DP 6-10 range, and insufficient intermediate chains in the DP 12-30 range. An experimental tree that had an NRMSE of 0.664 when compared against waxy maize CLD, had an NRMSE of 0.648 when compared to *sugary-2/cgx33* and 0.568 when compared against *sugary-2/cgr04*, meaning that the CLD of the model trees is more similar to the *sugary-2* mutant than to the waxy maize. Visual inspection of a difference plot shows similar trends in the CLDs from the model and *sugary-2*, as can be seen in Figure 3.3.1.1 below.

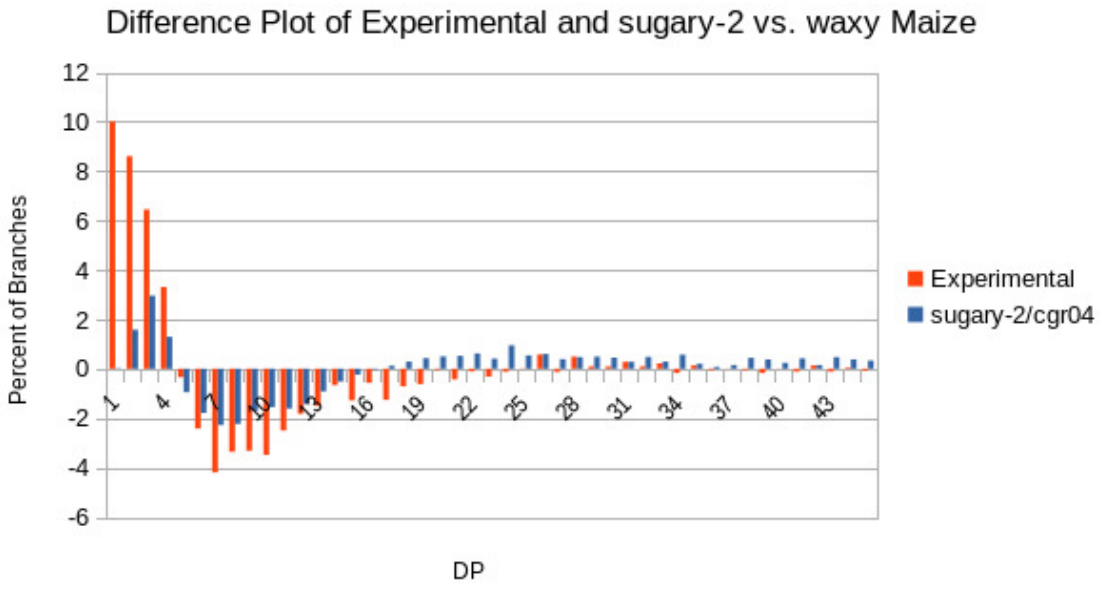

Figure 3.3.1.1: A difference plot comparing the CLDs of the model and *sugary-2/cgr04* mutant amylopectin against waxy maize amylopectin.

3.3.2 Addition of SS Isoforms

An experiment was conducted using three SS isoforms, SS1, SS2, and SS3, differentiated by substrate specificity deduced from biochemical and genetic analyses of cereal SS isoforms (Pfister and Zeeman 2016). SS1 elongated chains of DP less than 10, SS2 elongated DP 10-25, and SS3 elongated DP greater than 25, with each having independent control over enzyme activity. Two versions of this experiment were conducted, one where the SS's add only one G to the non-reducing end, and one where the SS's randomly add between one and three G's to the non-reducing end. SS's in nature can only append one glucose to the non-reducing end per catalytic event but are also processive enzymes, meaning that they can perform consecutive reactions before releasing the glucan. The addition of between one and three G's to the non-reducing end

in the model is a representation of SS processivity. Figure 3.3.2.1 below contains a difference plot comparing the results of the three SS isoform experiment and the *sugary-2* mutant to waxy maize. With the addition of SS2 and SS3 activity, the experimental and *sugary-2* CLDs no longer follow a similar visual trend.

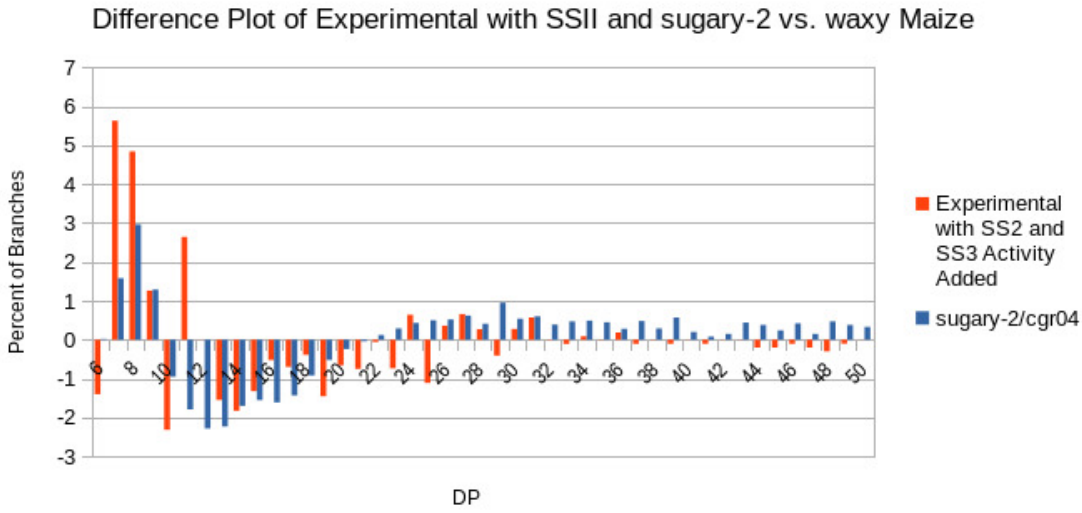

Figure 3.3.2.1: A difference plot comparing the CLDs of the experiment using three SS isoforms and *sugary-2/cgr04* mutant amylopectin against waxy maize amylopectin.

The addition of SS2 and SS3 isoforms to the experiment yielded a 5% NRMSE improvement in the version where only one G is appended to the non-reducing end, and an 11% NRMSE improvement in the version where 1-3 G's are appended. Overall the starch synthases had high activities, with SS1 consistently the highest. In the version where 1-3 G's are appended, SS1 had an activity of 72% ± 4%, SS2 had the lowest activity at 20% ± 2% and SS3 in the middle with 40% ± 0.5% activity. In the version where only one G could be appended, SS2 and SS3 switched rankings, with SS1 having 75% ± 1% activity, SS2 having 42% ± 1% activity, and SS3 having 40% ± 10% activity.

The high variability in SS3 activity exists because two apparent strategies in the population scored equally well, one strategy with high SS3 activities in the 70% range, and one strategy with lower SS3 activities in the 35% range. Figure 3.3.2.2 below contains drawings of example trees from the experiment with three SS isoforms and 1-3 G elongation capacity.

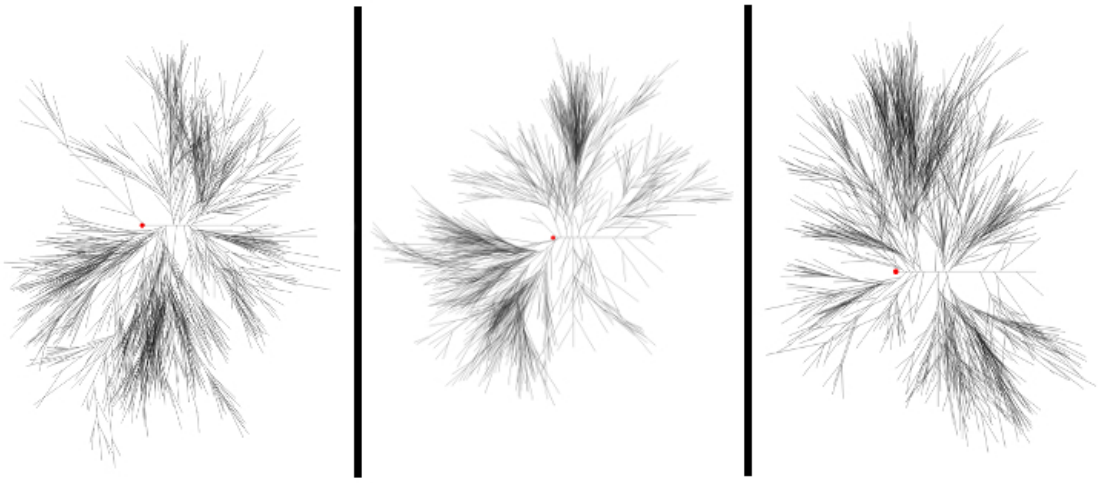

Figure 3.3.2.2: Drawings of three example trees from experiment with three SS isoforms and 1-3 G elongation capacity. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

Overall, these experiments demonstrated that incorporating multiple SS isoforms in the biosynthetic enzyme suite has surprising impacts on NRMSE. Incorporating SS2 and SS3 functionality into the model partially recovered the phenotype similar to the *sugary-2* mutant that was present in the model version with only one SS isoform. The fact that the CLD does not exactly match waxy maize after the incorporation of SS2 and SS3 is

likely because other important factors are missing from the model and are yet to be elucidated.

3.4 Evolving Enzyme Contexts

A set of experiments was conducted in which enzyme contexts, *i.e.* recognizable substrate features that trigger an action, were represented as evolvable structures rather than static, predefined variables. As described in Chapter 2, the evolvable enzyme contexts are represented as strings encoded in a custom regular expression alphabet. The first run of the model using the evolvable contexts had only a 10% increase in average NRMSE, which is quite impressive because of the massive increase in the complexity of the features being evolved. Additionally, the top scoring individuals in the population had NRMSE's equivalent or less than the average NRMSE in the version of the model with static enzyme contexts. Figure 3.4.1 below contains drawings of some example trees that resulted from the experiments with evolvable enzyme contexts.

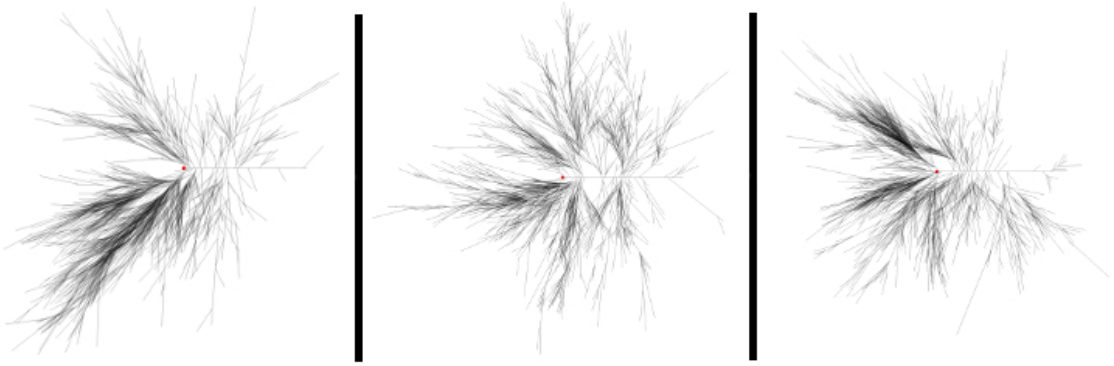

Figure 3.4.1: Drawings of three example trees from the experiments with evolvable enzyme contexts. The trees represent some of the diversity of structures achieved by this experiment. The red circles mark the tree seeds.

The first important question for designing the evolvable enzyme contexts is how long can the contexts be? This translates to the biological question, how large of a feature can the enzymes perceive and respond to? To address this question context lengths in the range of 6-25 characters long were tested to see the impact on NRMSE. The results of this exploration can be found in Figure 3.4.2 below.

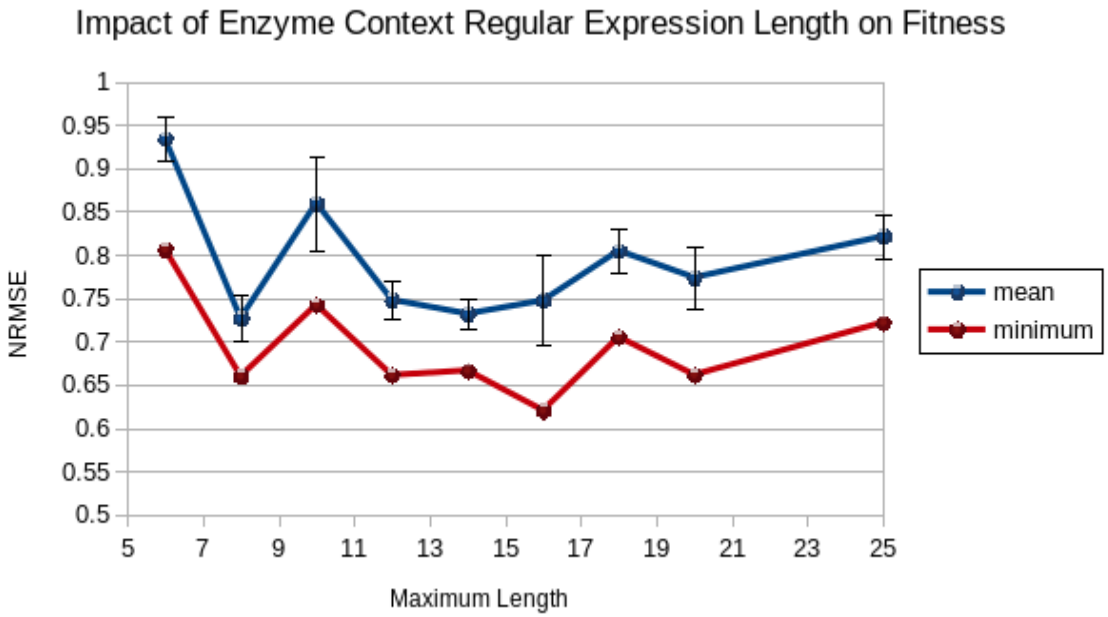

Figure 3.4.2: A comparison of the character length of the evolvable context strings in terms of average and minimum NRMSE. Error bars represent standard deviation.

The impact of context string length on NRMSE is both striking and trendless. A difference in length of a mere 2 characters resulted in a 20% decrease then a 13% increase in mean NRMSE. The minimum NRMSE matches the pattern of the mean, suggesting that differences are not due to the increased potential for variability between individuals.

Analysis of the enzyme activity trends between runs with different context lengths provided insight into the different strategies that evolved in response to different enzyme context string lengths. High variance in enzyme activities within a single experiment are an indicator of the existence of multiple strategies with similar scores, while low variance implies that only one strategy exists in the population. Figure 3.4.3 below charts the DBE

activities, Figure 3.4.4 the BE activities, and Figure 3.4.5 the SS activities across runs with different enzyme context lengths.

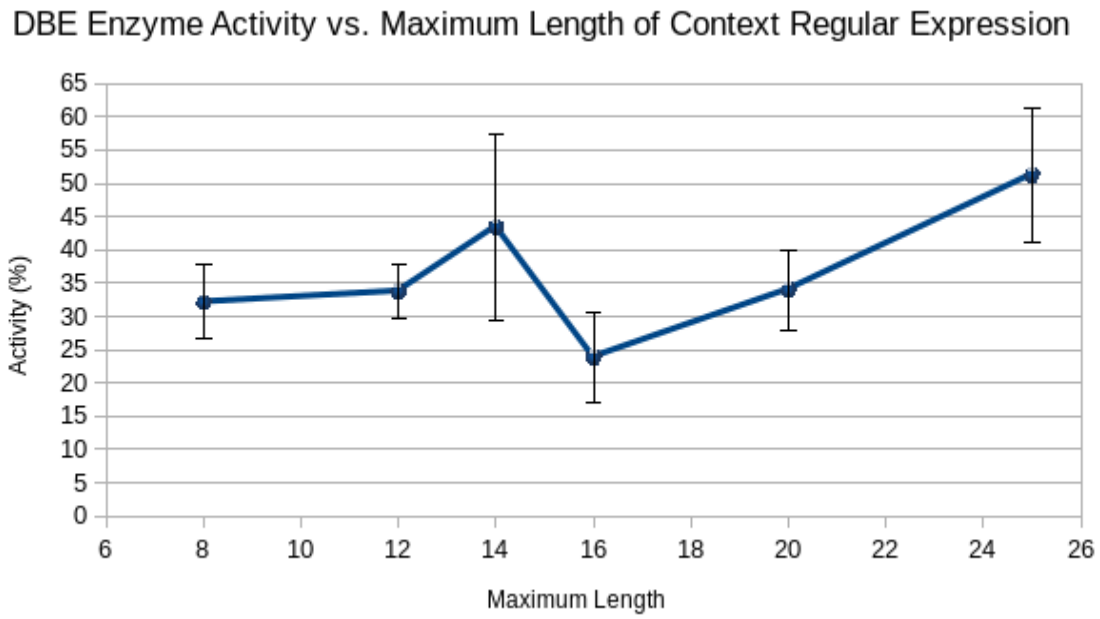

Figure 3.4.3: Debranching enzyme activities across runs with different enzyme context lengths. Error bars represent standard deviation.

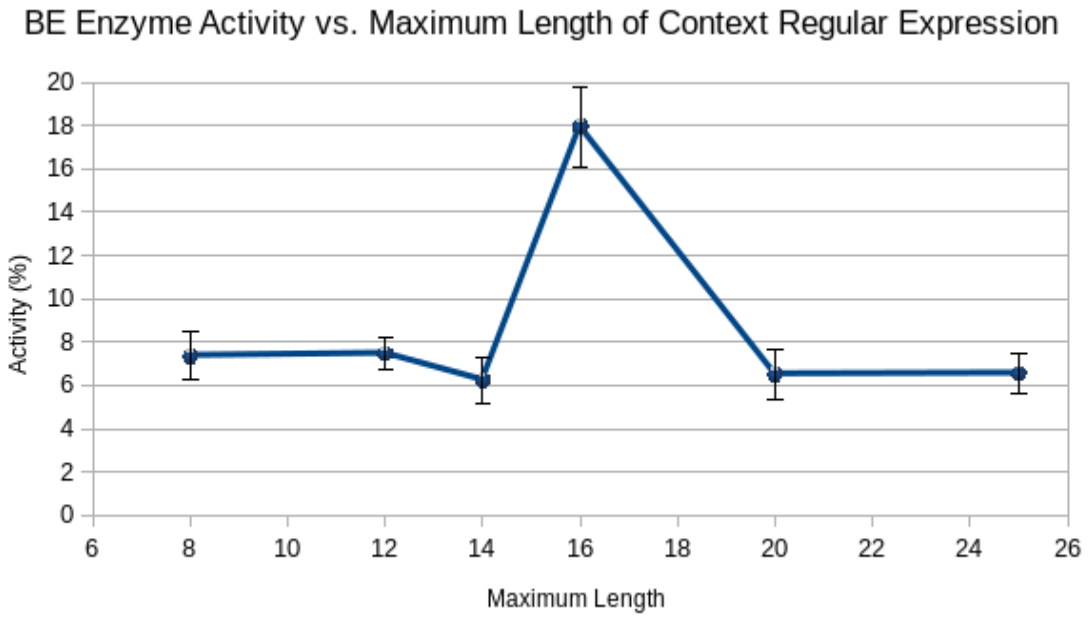

Figure 3.4.4: Branching enzyme activities across runs with different enzyme context lengths. Error bars represent standard deviation.

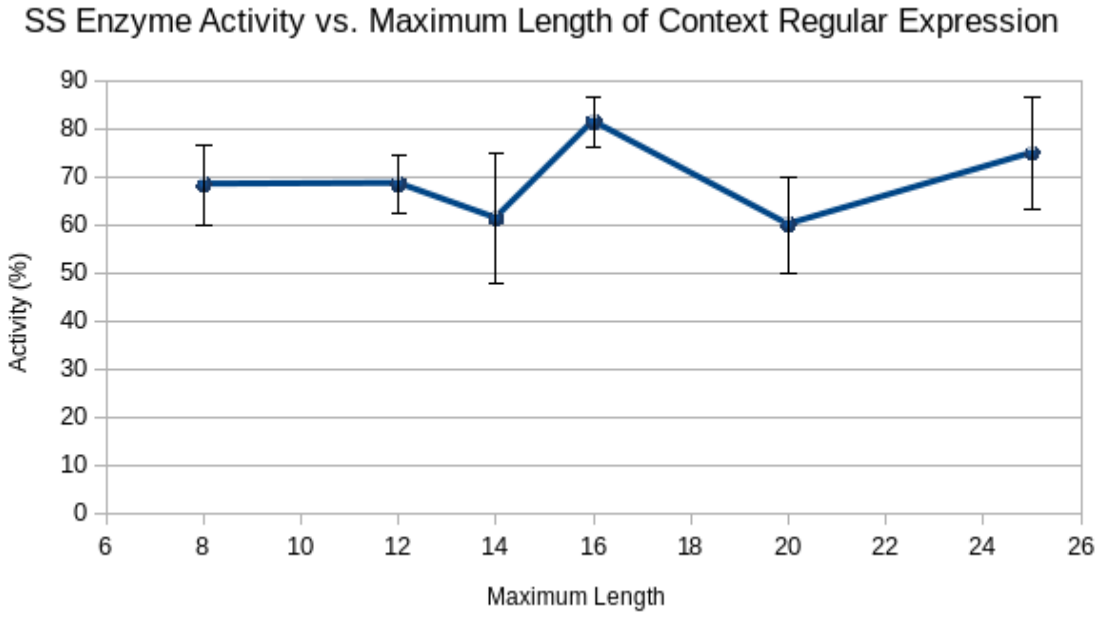

Figure 3.4.5: Starch synthase activities across runs with different enzyme context lengths. Error bars represent standard deviation.

Debranching enzyme showed the greatest change in activity levels in response to the addition of evolvable enzyme contexts of differing lengths. It appears that the control provided by the evolvable contexts is sufficient to allow DBE to have non-zero enzyme activities, which is a big accomplishment of this modification. While the pattern in DBE activity is somewhat erratic, the longest enzyme context strings did correlate with the highest enzyme activities, implying that the increased control conferred by a longer context string allows for higher enzyme activity.

Branching enzyme activity was the most consistent between runs. The addition of evolvable enzyme contexts did not cause a major change in activity level, which remained consistent across enzyme context lengths with the exception of length 16 where BE activity more than doubled. It is interesting that this coincides with the run with the lowest NRMSE, suggesting that increased BE activity may result in a more accurate CLD. While it is apparent that enzyme activity did not correlate with the maximum enzyme context string length allowed, examination of the enzyme context strings that resulted from each run showed that the branching enzyme context in the run with maximum length 16 was very unique. The length 16 run was the only run in which the dominant branching enzyme context searched for long unbranched regions, specifically those with DP >12. All other evolved branching contexts were much less stringent, matching extremely short unbranched regions that would occur very commonly in the tree. It is worthwhile to note that since average tree length and branching frequency did not decrease with the more stringent enzyme context, the strategies that evolved from this

run are simply best able to utilize the branching enzyme context for placement of branches. This promising result would merit further investigation in future work.

Starch synthase responded very little to changes in maximum length of context regular expressions. This is likely because the starch synthase context strings tended to be very short, and would therefore be unaffected by the maximum length available. Changes in SS activity between runs with different maximum context lengths was likely a response to the changing activity levels of the other enzymes more sensitive to the maximum length available, particularly DBE. It is interesting that, across runs, the most common starch synthase context to evolve was simply a non-reducing end, represented by a closing parenthesis, which is identical to the static context assigned in the experiments without evolvable contexts. The consistency of the trend was remarkable, that given any string of up to 25 characters long, composed of 9 possible characters plus numerals, that the one character long closing parenthesis consistently evolved suggests that there is a real fitness advantage associated with the context. From a biological perspective, it is quite plausible that starch synthases may be triggered simply by available non-reducing ends, since non-reducing ends are limited on the amylopectin molecule, and a lot of polymerization is required to accomplish the large size of amylopectin.

To visualize the diversity of enzyme context strings that evolved, word clouds were generated where the frequency of each context string is directly proportional to font size. Figures 3.4.6 and 3.4.7 display word clouds of evolved DBE contexts in experiments with the enzyme context string maximum length set to 12 and 25, respectively. Figures

3.4.8, 3.4.9, and 3.4.10 contain word clouds of the aggregated enzyme contexts evolved in all the runs for SS, BE, and DBE, respectively.

Figure 3.4.6: A word cloud depicting the diversity of DBE contexts that evolved in an experiment where the maximum context length was 12. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts.

Figure 3.4.7: A word cloud depicting the diversity of DBE contexts that evolved in an experiment where the maximum context length was 25. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts.

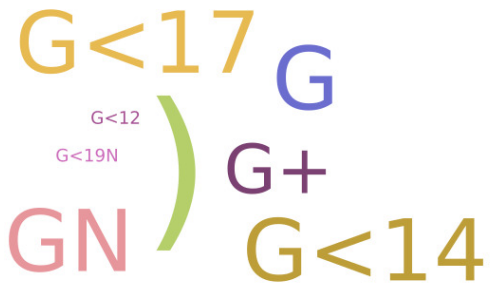

Figure 3.4.8: A word cloud depicting all the enzyme contexts that evolved for starch synthase, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts.

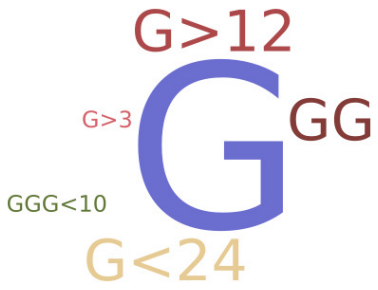

Figure 3.4.9: A word cloud depicting all the enzyme contexts that evolved for branching enzyme, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts.

Figure 3.4.10: A word cloud depicting all the enzyme contexts that evolved for debranching enzyme, across all runs, over the last 10 generations. Font size is directly proportional to frequency at which the string was evolved. See Table 2.1.1 in Chapter 2 for the alphabet used to represent enzyme contexts.

Chapter 4

Discussion

This chapter will highlight the accomplishments of the model presented and provide direction for further development of the model. The first section will summarize the results presented in Chapter 3 and present overarching conclusions that can be drawn from the experiments conducted. The second section will discuss directions for future research and development.

4.1 Conclusions and Accomplishments

The results presented in Chapter 3 represent three distinct phases in model development that took place within the scope of this project: parameter refinement, incorporation of domain knowledge, and target expansion. Parameter refinement categorizes all of the experiments that optimized any static values and features of the model. Static values, including the number of generations the algorithm runs for, the number of individuals in a population, the total number of enzymes available to each individual, and the field of view of the enzymes, were relatively simple to optimize because an ideal value could be found for each within the range of values that yield acceptable runtimes. Static features, including the fitness function used to score individuals, breeding algorithms for combining features to generate a new population, and enzyme actions are much more complicated to empirically optimize because of the boundless possibilities that exist for each, limited only by imagination rather than runtime. While the experiments directed at exploring each of these features yielded

improvements to the accuracy of the model and were executed to a logical break point, it is impossible to prove definitively that any feature used is the best possible.

While evolutionary computational models are specifically useful in situations where limited domain knowledge is available, the incorporation of domain knowledge can help refine and validate the model (Ashlock and Lathrop 1998; Hughes et al 2014; Ashlock and McNicholas 2012; Wolpert and Macready 1997). For example, the domain knowledge that a suite of starch synthases, branching enzymes, and debranching enzymes is highly conserved among starch synthesizing organisms provided a solid starting configuration for the model. In the absence of such information, every type of possible editing function would have to be tested. Enzyme behaviour, including actions and contexts (substrate specificities), were also initialized in the model based on domain knowledge. For example, the knowledge that starch synthases can only elongate at a non-reducing end was used as the static context for the model before enzyme contexts were evolved. The second phase of experimentation included further incorporation of domain knowledge by comparison to the *sugary-2* mutant. The similarities between the model CLD and the CLD of the *sugary-2* mutant guided model development by inspiring the addition of a second starch synthase to replicate SSII. The addition of multiple SS isoforms was an important step in model refinement as it improved model accuracy by 11%. Additionally, comparison to the *sugary-2* mutant CLD helped to validate that the model is truly representative of the amylopectin biosynthetic system because it responds to the absence of a necessary enzyme (SSII) in a similar way as a real plant does. Similarly, the use of barley CLD in comparison to maize CLD resulted in differences in

enzymatic activity. This further validated the model because it is known that the relative contribution of the different enzymes varies between botanical sources and is believed to account for the structural variation between starches (Wang et al 2014).

The third phase of model development was target expansion, *i.e.* the inclusion of enzyme contexts as an additional target for evolution. In the initial configuration of the model the enzyme contexts were static, predefined features, and the only feature evolved was enzyme activity. However, because there is insufficient domain knowledge on the details of substrate specificity and the impact it can have on enzyme coordination, targeting enzyme contexts for evolution was a natural next step in the model development process and yielded interesting results. The most striking result was for starch synthase, which consistently evolved a non-reducing end as it's context specificity, *i.e.* exactly the same context as the static context initially provided based on domain knowledge. Branching enzyme and debranching enzyme both demonstrated the potential for increased activity when enzyme context could be evolved and controlled. The model yielded improved accuracy when branching enzyme evolved to target regions that were not already branched. The evolution of enzyme contexts that target sparsely branched regions is of especial interest because of the biochemical evidence that densely branched regions in amylopectin may exclude enzymes through steric hindrance caused by hydrogen bonding. The model evolving enzyme contexts that reflect physical restrictions is both validative and indicative of directions for further research.

The model presented in this thesis is the first attempt ever made at modelling amylopectin biosynthesis using evolutionary computation and stigmergic building

algorithms. It is also the first time stigmergic building algorithms have been combined with evolutionary computational techniques. The biggest accomplishment of the model is demonstrating that a stigmergic building algorithm, adapted form insect behavioural studies, can be improved using evolutionary computation and applied to an enzymatic system in a way that is both realistic and informative. The success of the model was heavily dependant on the development of an evolvable representation of enzymatic systems and a lightweight representation of amylopectin as a one-dimensional string. The one-dimensional string representation of amylopectin proved exceedingly useful as it yielded fast runtimes and simplicity that allowed for easy modification and experimentation with the enzymes. The novel tools developed for visualizing and analysing data, particularly integration with SQL database tools, were also major accomplishments of this project, as they were both effective and would be useful for other projects, including those in different fields of research.

4.2 Directions for Future Research

The results presented in this thesis show a promising start to the further development of this model, as well as other models of amylopectin using evolutionary computational and stigmergic methods. The next big step that should take place in model development is further expansion of the evolutionary targets to include enzyme actions. Just as enzyme contexts were initially static features of the model and became targets for evolution, enzyme actions should also be made into evolvable features. Because enzyme actions have to be within the scope of what is plausible for starch synthases, branching

enzymes, and debranching enzymes, it would probably be best to store possible actions in a lookup table and have the evolvable feature be which of the entries in the lookup table is utilized. Enzyme actions have a huge impact on model accuracy, as was demonstrated in Chapter 3 by the impacts of different static branching enzyme actions that were tested, so it is very plausible that allowing enzyme actions to evolve could greatly improve model accuracy.

The selection and breeding algorithms used were designed specifically for this model, and would merit further testing and development. The model presented has unique challenges that the breeding algorithm must meet because the features being evolved are synergistic. The whole idea of the model is to show how enzymes coordinate behaviours, so shuffling enzymes between individuals at breeding events can yield offspring that have much worse fitness than their parents. The use of averaging in the breeding algorithm was one way the current model maintained some of the synergy between evolved enzyme activities in offspring. Designing and testing more novel breeding algorithms specficially suited to this model, as well as other techniques such as retaining successful individuals in the population, could be an interesting direction for future research and yield improvements to model accuracy. However it is worthwhile to note that, since the fitness level did plateau in each experiment because of the population size and number of generations, changes to the breeding algorithm are more likely to decrease runtime and improve accuracy in future experiments, rather than change any of the results of the experiments already completed.

The one-dimensional amylopectin model has great capacity for the inclusion of physical features of amylopectin. For example, the hydrogen bonding between closely spaced branches could be simulated by providing an additional ruleset that excludes enzymes from landing in locations with closely spaced branches. A small amount of experimentation on representation of physical features of amylopectin yielded promising results that such features are relatively simple to include in the model and would be an interesting target for future research.

One of the limitations of the model is that it simulates only a single tree, rendering inter-chain transfer by branching enzymes impossible. Also, the functions of starch synthases and other enzymes that are responsible for creating seed trees can not be represented in a single tree version of the model. Expansion of the model to allow for the creation of additional trees and determining the CLD based on multiple molecules would be a very interesting direction for future research as it might both improve the model and expand its scope.

Using the model to compare amylopectins from a wide variety of botanical sources would be very interesting, especially after the model has been further refined using the steps described above. Real CLD data is available for amylopectins from a huge variety of commerically significant species. Using the model to elucidate what differences in enzyme activity, action, and context yield the different CLDs found in nature would be a fantastic direction for future research.

Bibliography

1. Ashlock, D. A., & Jafargholi, F. (2007, April). Evolving extremal epidemic networks. In Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB'07. IEEE Symposium on (pp. 338-345). IEEE.
2. Ashlock, D., & Goren, A. (2014, May). Agent-based modelling of resource flow in plant networks. In Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on (pp. 1-7). IEEE.
3. Ashlock, D., & Kim, E. Y. (2005, June). The impact of cellular representation on finite state agents for prisoner's dilemma. In Proceedings of the 7th annual conference on Genetic and evolutionary computation (pp. 59-66). ACM.
4. Ashlock, D., & Lathrop, J. I. (1998, March). A fully characterized test suite for genetic programming. In International Conference on Evolutionary Programming, 537-546.
5. Ashlock, D., & McNicholas, S. (2012, June). Single parent generalization of cellular automata rules. In Evolutionary Computation (CEC), 2012 IEEE Congress on (pp. 1-8). IEEE.
6. Ashlock, D., Kim, E. Y., & Leahy, N. (2006). Understanding representational sensitivity in the iterated prisoner's dilemma with fingerprints. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 36(4), 464-475.

7. Ashlock, D., Wittrock, A., & Wen, T. J. (2002, May). Training finite state machines to improve PCR primer design. In Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on (Vol. 1, pp. 13-18). IEEE.
8. Ball, S., Guan, H. P., James, M., Myers, A., Keeling, P., Mouille, G., Buléon A., Colonna P., & Preiss, J. (1996). From glycogen to amylopectin: a model for the biogenesis of the plant starch granule. Cell, 86(3), 349-352.
9. Bertoft, E. (2013). On the building block and backbone concepts of amylopectin structure. Cereal Chemistry, 90(4), 294-311.
10. Bertoft, E., Piyachomkwan, K., Chatakanonda, P., & Sriroth, K. (2008). Internal unit chain composition in amylopectins. Carbohydrate Polymers, 74(3), 527-543.
11. Bettge, A. D., Giroux, M. J., & Morris, C. F. (2000). Susceptibility of Waxy Starch Granules to Mechanical Damage. Cereal chemistry, 77(6), 750-753.
12. Bhattacharyya, M. K., Smith, A. M., Ellis, T. N., Hedley, C., & Martin, C. (1990). The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell, 60(1), 115-122.
13. Bonabeau, E. (1999). Editor's introduction: stigmergy. Artificial Life, 5(2), 95-96.
14. Bonabeau, E., Dorigo, M., & Theraulaz, G. (2000). Inspiration for optimization from social insect behaviour. Nature, 406(6791), 39-42.
15. Borovsky, D., Smith, E. E., Whelan, W. J., French, D., & Kikumoto, S. (1979). The mechanism of Q-enzyme action and its influence on the structure of amylopectin. Archives of biochemistry and biophysics, 198(2), 627-631.

16. Bresolin, N. S., Li, Z., Kosar-Hashemi, B., Tetlow, I. J., Chatterjee, M., Rahman, S., Morell, M.K., & Howitt, C. A. (2006). Characterisation of disproportionating enzyme from wheat endosperm. Planta, 224(1), 20-31.
17. Colleoni, C., Dauvillée, D., Mouille, G., Buléon, A., Gallant, D., Bouchet, B., Morell, M., Samuel, M., Delrue, B., d'Hulst, C. and Bliard, C. (1999). Genetic and biochemical evidence for the involvement of α-1, 4 glucanotransferases in amylopectin synthesis. Plant Physiology, 120(4), 993-1004.
18. Commuri, P. D., & Keeling, P. L. (2001). Chain-length specificities of maize starch synthase I enzyme: studies of glucan affinity and catalytic properties. The Plant Journal, 25(5), 475-486.
19. Deng, B., Sullivan, M. A., Chen, C., Li, J., Powell, P. O., Hu, Z., & Gilbert, R. G. (2016). Molecular structure of human-liver glycogen. PloS one, 11(3), e0150540.
20. Deschamps, P., Haferkamp, I., d'Hulst, C., Neuhaus, H. E., & Ball, S. G. (2008). The relocation of starch metabolism to chloroplasts: when, why and how. Trends in plant science, 13(11), 574-582.
21. Doane, W. M. (1994). Opportunities and challenges for new industrial uses of starch. Cereal Foods World, 39(8), 556-563.
22. Edwards, A., Fulton, D. C., Hylton, C. M., Jobling, S. A., Gidley, M., Rössner, U., ... & Smith, A. M. (1999). A combined reduction in activity of starch synthases II and III of potato has novel effects on the starch of tubers. The Plant Journal, 17(3), 251-261.

23. Fujita, N., Yoshida, M., Kondo, T., Saito, K., Utsumi, Y., Tokunaga, T., Nishi, A., Satoh, H., Park, J.-H., Jane, J.-L., & Miyao, A. (2007). Characterization of SSIIIa-deficient mutants of rice: the function of SSIIIa and pleiotropic effects by SSIIIa deficiency in the rice endosperm. Plant Physiology, 144(4), 2009-2023.
24. Gallant, D. J., Bouchet, B., & Baldwin, P. M. (1997). Microscopy of starch: evidence of a new level of granule organization. Carbohydrate polymers, 32(3), 177-191.
25. Gérard, C., Planchot, V., Colonna, P., & Bertoft, E. (2000). Relationship between branching density and crystalline structure of A-and B-type maize mutant starches. Carbohydrate Research, 326(2), 130-144.
26. Gidley, M. J., & Bulpin, P. V. (1987). Crystallisation of malto-oligosaccharides as models of the crystalline forms of starch: minimum chain-length requirement for the formation of double helices. Carbohydrate Research, 161(2), 291-300.
27. Glaring, M. A., Skryhan, K., Kötting, O., Zeeman, S. C., & Blennow, A. (2012). Comprehensive survey of redox sensitive starch metabolising enzymes in Arabidopsis thaliana. Plant Physiology and Biochemistry, 58, 89-97.
28. Graf, A., & Smith, A. M. (2011). Starch and the clock: the dark side of plant productivity. Trends in plant science, 16(3), 169-175.
29. Grassé, P. P. (1959). La reconstruction du nid et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp. la théorie de la stigmergie: Essai d'interprétation du comportement des termites constructeurs. Insectes sociaux, 6(1), 41-80.

30. Guan, H. P., & Preiss, J. (1993). Differentiation of the properties of the branching isozymes from maize (Zea mays). Plant Physiology, 102(4), 1269-1273.
31. Hizukuri, S. (1986). Polymodal distribution of the chain lengths of amylopectins, and its significance. Carbohydrate research, 147(2), 342-347.
32. Hughes, J. A., Houghten, S., & Ashlock, D. (2014). Recentering and restarting a genetic algorithm using a generative representation for an ordered gene problem. International journal of hybrid intelligent systems, 11(4), 257-271.
33. Imparl-Radosevich, J. M., Gameon, J. R., McKean, A., Wetterberg, D., Keeling, P. L., & Guan, H. (2003). Understanding catalytic properties and functions of maize starch synthase isozymes. Journal of applied glycoscience, 50(2), 177-182.
34. Inouchi, N., Glover, D. V., Takaya, T., & Fuwa, H. (1983). Development changes in fine structure of starches of several endosperm mutants of maize. Starch-Stärke, 35(11), 371-376.
35. James, M. G., Denyer, K., & Myers, A. M. (2003). Starch synthesis in the cereal endosperm. Current opinion in plant biology, 6(3), 215-222.
36. Jane, J. L., Kasemsuwan, T., Leas, S., Zobel, H., & Robyt, J. F. (1994). Anthology of starch granule morphology by scanning electron microscopy. Starch-Stärke, 46(4), 121-129.
37. Jespersen, H. M., Ann MacGregor, E., Henrissat, B., Sierks, M. R., & Svensson, B. (1993). Starch-and glycogen-debranching and branching enzymes: prediction of structural features of the catalytic (β/α) 8-barrel domain and evolutionary

relationship to other amylolytic enzymes. Journal of protein chemistry, 12(6), 791-805.

38. Leake, J. R. (1994). The biology of myco-heterotrophic ('saprophytic') plants. New Phytologist, 127(2), 171-216.
39. Leterrier, M., Holappa, L. D., Broglie, K. E., & Beckles, D. M. (2008). Cloning, characterisation and comparative analysis of a starch synthase IV gene in wheat: functional and evolutionary implications. BMC Plant Biology, 8(1), 98.
40. Li, L., Jiang, H., Campbell, M., Blanco, M., & Jane, J. L. (2008). Characterization of maize amylose-extender (ae) mutant starches. Part I: Relationship between resistant starch contents and molecular structures. Carbohydrate polymers, 74(3), 396-404.
41. Mangelsdorf, P. C. (1947). The inheritance of amylaceous sugary endosperm and its derivatives in maize. Genetics, 32(5), 448.
42. Manners, D. J. (1963). Enzymic synthesis and degradation of starch and glycogen. Advances in carbohydrate chemistry, 17, 371-430.
43. Martin, C., & Smith, A. M. (1995). Starch biosynthesis. The plant cell, 7(7), 971.
44. Morell, M. K., Blennow, A., Kosar-Hashemi, B., & Samuel, M. S. (1997). Differential expression and properties of starch branching enzyme isoforms in developing wheat endosperm. Plant Physiology, 113(1), 201-208.
45. Morell, M.K., Kosar-Hashemi, B., Cmiel, M., Samuel, M.S., Chandler, P., Rahman, S., Buleon, A., Batey, I.L. and Li, Z., (2003). Barley sex6 mutants lack

starch synthase IIa activity and contain a starch with novel properties. The Plant Journal, 34(2), 173-185.

46. Mouille, G., Maddelein, M. L., Libessart, N., Talaga, P., Decq, A., Delrue, B., Ball, S. (1996). Phytoglycogen processing: a mandatory step for starch biosynthesis in plants. The Plant Cell 8, 1353–1366.
47. Mu, H. H., Yu, Y., Wasserman, B. P., & Carman, G. M. (2001). Purification and characterization of the maize amyloplast stromal 112-kDa starch phosphorylase. Archives of Biochemistry and Biophysics, 388(1), 155-164.
48. Myers, A. M., Morell, M. K., James, M. G., & Ball, S. G. (2000). Recent progress toward understanding biosynthesis of the amylopectin crystal. Plant Physiology, 122(4), 989-998.
49. Nakamura, Y., Utsumi, Y., Sawada, T., Aihara, S., Utsumi, C., Yoshida, M., & Kitamura, S. (2010). Characterization of the reactions of starch branching enzymes from rice endosperm. Plant and cell physiology, 51(5), 776-794.
50. O'Sullivan, A. C., & Perez, S. (1999). The relationship between internal chain length of amylopectin and crystallinity in starch. Biopolymers, 50(4), 381-390.
51. Pérez, S., & Bertoft, E. (2010). The molecular structures of starch components and their contribution to the architecture of starch granules: A comprehensive review. Starch-Stärke, 62(8), 389-420.
52. Pérez, S., Baldwin, P. M., & Gallant, D. J. (2009). Structural features of starch granules I. Starch: Chemistry and technology, 3.

53. Pfister, B., & Zeeman, S. C. (2016). Formation of starch in plant cells. Cellular and Molecular Life Sciences, 73(14), 2781-2807.

54. Preiss, J. (1991). Biology and molecular biology of starch synthesis and its regulation. Oxford surveys of cellular and molecular biology, Vol. 7., 59-114.

55. Rydberg, U., Andersson, L., Andersson, R., Åman, P., & Larsson, H. (2001). Comparison of starch branching enzyme I and II from potato. European Journal of Biochemistry, 268(23), 6140-6145.

56. Ryoo, N., Yu, C., Park, C. S., Baik, M. Y., Park, I. M., Cho, M. H., Bhoo, S.H., An, G., Hahn, F.R., & Jeon, J. S. (2007). Knockout of a starch synthase gene OsSSIIIa/Flo5 causes white-core floury endosperm in rice (Oryza sativa L.). Plant cell reports, 26(7), 1083-1095.

57. Sawada, T., Nakamura, Y., Ohdan, T., Saitoh, A., Francisco, P. B., Suzuki, E., Fujita, N., Shimonaga, T., Fujiwara, S., Tsuzuki, M., & Colleoni, C. (2014). Diversity of reaction characteristics of glucan branching enzymes and the fine structure of α-glucan from various sources. Archives of biochemistry and biophysics, 562, 9-21.

58. Schindler, I., Renz, A., Schmid, F. X., & Beck, E. (2001). Activation of spinach pullulanase by reduction results in a decrease in the number of isomeric forms. Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology, 1548(2), 175-186.

59. Seung, D., Soyk, S., Coiro, M., Maier, B. A., Eicke, S., & Zeeman, S. C. (2015). PROTEIN TARGETING TO STARCH is required for localising GRANULE-

BOUND STARCH SYNTHASE to starch granules and for normal amylose synthesis in Arabidopsis. PLoS Biol, 13(2), e1002080.

60. Stitt, M., & Zeeman, S. C. (2012). Starch turnover: pathways, regulation and role in growth. Current opinion in plant biology, 15(3), 282-292.
61. Szydlowski, N., Ragel, P., Raynaud, S., Lucas, M. M., Roldán, I., Montero, M., ... & Pozueta-Romero, J. (2009). Starch granule initiation in Arabidopsis requires the presence of either class IV or class III starch synthases. The Plant Cell, 21(8), 2443-2457.
62. Takeda, Y., Guan, H. P., & Preiss, J. (1993). Branching of amylose by the branching isoenzymes of maize endosperm. Carbohydrate research, 240, 253-263.
63. Tetlow, I. J. (2011). Starch biosynthesis in developing seeds. Seed Science Research, 21(01), 5-32.
64. Tetlow, I. J., & Emes, M. J. (2014). A review of starch-branching enzymes and their role in amylopectin biosynthesis. IUBMB life, 66(8), 546-558.
65. Tetlow, I.J., Wait, R., Lu, Z., Akkasaeng, R., Bowsher, C.G., Esposito, S., Kosar-Hashemi, B., Morell, M.K. and Emes, M.J., (2004). Protein phosphorylation in amyloplasts regulates starch branching enzyme activity and protein–protein interactions. The Plant Cell, 16(3), 694-708.
66. Theraulaz, G., & Bonabeau, E. (1999). A brief history of stigmergy. Artificial life, 5(2), 97-116.
67. Tiessen, A., Hendriks, J. H., Stitt, M., Branscheid, A., Gibon, Y., Farré, E. M., & Geigenberger, P. (2002). Starch synthesis in potato tubers is regulated by post-

translational redox modification of ADP-glucose pyrophosphorylase a novel regulatory mechanism linking starch synthesis to the sucrose supply. The Plant Cell, 14(9), 2191-2213.

68. Valdez, H. A., Busi, M. V., Wayllace, N. Z., Parisi, G., Ugalde, R. A., & Gomez-Casati, D. F. (2008). Role of the N-terminal starch-binding domains in the kinetic properties of starch synthase III from Arabidopsis thaliana. Biochemistry, 47(9), 3026-3032.
69. van de Velde, F., van Riel, J., & Tromp, R. H. (2002). Visualisation of starch granule morphologies using confocal scanning laser microscopy (CSLM). Journal of the Science of Food and Agriculture, 82(13), 1528-1536.
70. Wang, K., Henry, R. J., & Gilbert, R. G. (2014). Causal relations among starch biosynthesis, structure, and properties. Springer Science Reviews, 2(1-2), 15-33.
71. Wattebled, F., Planchot, V., Dong, Y., Szydlowski, N., Pontoire, B., Devin, A., Ball, S., & D'Hulst, C. (2008). Further evidence for the mandatory nature of polysaccharide debranching for the aggregation of semicrystalline starch and for overlapping functions of debranching enzymes in Arabidopsis leaves. Plant Physiology, 148(3), 1309-1323.
72. Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67-82.
73. Wu, C., Colleoni, C., Myers, A. M., & James, M. G. (2002). Enzymatic properties and regulation of ZPU1, the maize pullulanase-type starch debranching enzyme. Archives of biochemistry and biophysics, 406(1), 21-32.

74. Yun, S. H., & Matheson, N. K. (1993). Structures of the amylopectins of waxy, normal, amylose-extender, and wx: ae genotypes and of the phytoglycogen of maize. Carbohydrate Research, 243(2), 307-321.

75. Zeeman, S. C., Umemoto, T., Lue, W. L., Au-Yeung, P., Martin, C., Smith, A. M., & Chen, J. (1998). A mutant of Arabidopsis lacking a chloroplastic isoamylase accumulates both starch and phytoglycogen. The Plant Cell, 10(10), 1699-1711.

76. Zhang, X., Colleoni, C., Ratushna, V., Sirghie-Colleoni, M., James, M., & Myers, A. (2004). Molecular characterization demonstrates that the Zea mays gene sugary2 codes for the starch synthase isoform SSIIa. Plant molecular biology, 54(6), 865-879.

77. Zhang, X., Szydlowski, N., Delvallé, D., D'Hulst, C., James, M. G., & Myers, A. M. (2008). Overlapping functions of the starch synthases SSII and SSIII in amylopectin biosynthesis in Arabidopsis. BMC Plant Biology, 8(1), 96.