A stochastic simulation approach for improving response in genomic selection

by

Saba Moeinizade

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Guiping Hu, Major Professor
Lizhi Wang
Thomas Lubberstedt

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

I would like to dedicate this thesis to my fiance, Sina, who has been a constant source of support and encouragement during the last year. I am truly thankful for having you in my life. This work is also dedicated to my parents, Mahin and Asad, and my sister, Sama who have always loved me unconditionally and whose good examples have taught me to work hard for the things that I aspire to achieve. I would also like to thank all my friends for being so supportive during the writing of this work.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Guiping Hu and Dr. Lizhi Wang for their guidance, patience and support throughout this research and the writing of this thesis. Their insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee member, Dr. Thomas Lubberstedt for his effort and contribution to this work. Additionally, I would like to thank Hieu Pham for proofreading my manuscript.

This work is supported by Agriculture and Food Research Initiative Grant no. 2017-67007-26175/Accession No. 1011702 from the USDA National Institute of Food and Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture. This work is also supported by the Plant Sciences Institutes Faculty Scholars program at Iowa State University.

ABSTRACT

The world population is increasing rapidly and is projected to hit 9.1 billion by 2050. As the demand for food increases, agriculture production will continue to play a significant role. As a method to maintain and increase agriculture production, plant breeding is critical. To improve efficiency in the plant breeding process, an interdisciplinary effort is needed. Operations research as a discipline focuses on decision making and efficient and effective strategy design. In this thesis, operations research tools of simulation, optimization and mathematical modeling are applied to plant breeding, specifically Genomic Selection (GS). GS techniques allow breeders to select the best plants to make crosses by predicting, for example, the heights of the plants using the genotypic data at an early stage of the plant growth cycle, saving both time and cost that would otherwise be necessary to grow the plants to maturity before their heights can be measured. A major limitation of existing GS approaches is the trade-off between short-term genetic gains and long-term growth potential. Some approaches focus on achieving short-term genetic gains at the cost of losing genetic diversity for long-term gains, and others aim to maximize the long-term genetic gains but are unable to achieve it by the breeding deadline. Our contribution is to define a new look-ahead method for assessing a selection decision, which evaluates the probability to achieve both genetic diversity and breeding deadline. Moreover, we propose a heuristic algorithm to find an optimal selection decision with respect to the new method. Our new selection method outperforms the other selection methods in the literature.

CHAPTER 1. GENERAL INTRODUCTION

1.1 Introduction

Humans have been breeding plants for food since the dawn of agriculture. Today, we know that the impact of agriculture is profound on humanity. The world population is growing fast and is projected to hit 9.1 billion by 2050. Feeding the growing population is a daunting challenge. Producing new crop varieties that offer higher yields but require less water, fertilizers or other inputs would greatly help. Plant breeding as a discipline has been instrumental in this area. National Association of plant breeders defines plant breeding as the science driven creative process of developing new plant varieties which involves crossing parental plants to obtain the best characteristics for the future generation.

To improve the efficiency in the breeding process an interdisciplinary effort is needed. Operations research (OR) as a discipline focuses on decision and strategy design. OR is an analytical method of problem-solving to achieve efficient and effective decisions. Analytical methods used in OR include mathematical modeling, simulation, optimization, and statistics.

The gap between engineering and plant breeding brings several opportunities and challenges for operations researchers. Trait introgression and genomic selection are two existing challenges. In recent years, operations research tools have been applied to multi-allelic introgression Han et al. (2017), and genomic selection Goiffon et al. (2017). This thesis explores the application of operations research to improve response in genomic selection by designing a new method, Look-ahead selection with emphasis on optimizing selection, mating strategies, and resource allocation given a breeding timeline.

1.2 Literature review

Since the late 19th century, plant breeders have been relying on phenotypic selection to improve plant verieties. Plants with desirable phenotypes were selected as the breeding parents. With the advent of molecular markers in the late 1970s, advances have been made in the plant breeding techniques (Brumlop and Finckh, 2011). Today, with the wider availability and reduced cost of molecular marker technology, marker assisted selection of genotypes has become viable (Mcdowell et al., 2016). Marker assisted selection (MAS) is an indirect selection process that aims to incorporate genotypic information into selection decisions (Lande and Thompson, 1990). MAS has been a useful tool for plant breeders but has some limitations in improving complex traits as it cannot capture small effect quantitative trait loci (QTL) (Heffner et al., 2010). MAS becomes less effective when selections are made for traits with many contributing genes distributed widely across a genome that have small effect (Mcdowell et al., 2016). Genomic selection aims at addressing this limitation of MAS due to its effectiveness for traits controlled by many genes with small effects.

Genomic selection (GS) is a form of marker assisted selection first proposed by Meuwissen et al. (2001) that uses phenotypic and genotypic data from past trials of individuals to build a prediction model. The model is then used to predict the value of individuals that have not been phenotyped. GS's ability to increase genetic gain has been validated through a number of simulation and empirical studies. Bernardo and Yu (2007) has compared the response resulting from GS with MARS by simulation in a bi-parental maize breeding program. They showed that GS leads to a larger response than MARS. Similarly, empirical studies in wheat populations have showed that GS results in greater prediction accuracy than MAS (Lorenz et al., 2011).

Genomic Estimated Breeding Value (GEBV) for individual plant has been adopted as the selection criterion in the original GS method and it selects individuals based on the sum of their estimated marker effects (Meuwissen et al., 2001). This approach has resulted

in genetic gain due to significant correlation between GEBVs and true breeding values. Since then, however, three extensions have been proposed to improve GS: weighted genomic selection (WGS) (Heffner et al., 2010), optimal haploid value (OHV) (Daetwyler et al., 2015), and optimal population value (OPV) (Goiffon et al., 2017). The conventional GS method, GEBV, can assure accelerating short-term gain, but doesn't guarantee achieving long-term gain (Jannink, 2010). To maximize long-term response, the first extension, WGS has been proposed as a variation of GS where marker effects are weighted to increase the frequency of rare favorable alleles (Goddard, 2009). The second extension, OHV calculates the breeding value of the best possible double haploid derived from an individual (Daetwyler et al., 2015). This method focuses selection on the haplotype and optimizes the breeding program toward its end goal of generating an elite fixed line (Daetwyler et al., 2015). GS, WGS, and OHV are truncation selection approaches as they rank individuals and select the top ones (typically a fraction of the population based on the available resources), but recently OPV proposed a different strategy that is population-based. OPV selects the best population based on the interactive population effects which calculates the breeding value of the best possible progeny produced after an unlimited number of generations (Goiffon et al., 2017). Like OPV, we focus on selecting sets of individuals as a unit by proposing an innovative method, look-ahead selection (LAS). Our new selection method can improve the genetic gain by maximizing the probability of producing outstanding progenies in the final targeted generation. The proposed method can focus on achieving both short-term and long-term genetic gains and has the flexibility of adjusting based on the deadline and resource availability.

1.3 Thesis organization

This thesis adheres to the Iowa State University Journal Paper format. Chapter 1 begins with a general background to plant breeding and literature review. Chapter 2 contains an article to be submitted to the *Journal of Genetics* which introduces a new selection method,

Look- ahead selection with a stochastic simulation approach. Chapter 3 outlines the results with a special emphasis on potential future work.

CHAPTER 2. LOOK-AHEAD SELECTION: A STOCHASTIC SIMULATION APPROACH FOR IMPROVING RESPONSE IN GENOMIC SELECTION

Abstract

Genotyping technologies unleashed a large amount of genotypic data for plant breeders to accelerate the rate of genetic gains. Genomic selection (GS) techniques allow breeders to select the best plants to make crosses by predicting, for example, the heights of the plants using the genotypic data at an early stage of the plant growth cycle, saving both time and cost that would otherwise be necessary to grow all plants to maturity before their heights can be measured. A major limitation of existing GS approaches is the trade-off between short-term genetic gains and long-term growth potential. Some approaches focus on achieving short-term genetic gains at the cost of losing genetic diversity for long-term gains, and others maximize the long-term genetic gains but are unable to achieve it by the breeding deadline. Our contribution is to define a new look-ahead method for assessing a selection decision, which evaluates the probability to achieve both genetic diversity and breeding deadline. Moreover, we propose a heuristic algorithm to find an optimal selection decision with respect to the new method. Our new selection method outperforms the other selection methods in the literature.

keywords: genetic gain; genomic selection; look-ahead selection; stochastic simulation; population-based selection

2.1 Introduction

The world population is expected to grow from 7.6 billion today to 9.1 billion by 2050. Feeding all the entire population remains a significant challenge. Producing new crop varieties that offer higher yields but require less water, fertilizers or other inputs would greatly help. Plant breeding discipline has been instrumental in this area. Classical plant breeding programs rely on the phenotyping of progenies in field trials to identify superior individuals. The number of phenotyped individuals is limited by high costs and time for relevant field evaluation (Rincent et al., 2017). This reduced number of selection candidates is a major limit to genetic progress. Genomic selection (GS) allows predicting the performance of unphenotyped individuals (Rincent et al., 2017; Meuwissen et al., 2001). GS refers to using the whole genome to estimate the breeding value of selection candidates for a quantitative trait (Goddard, 2009). Genomic Estimated Breeding Value (GEBV) for individual plant has been adopted as the selection criteria in the original GS method and it selects individuals based on the sum of their estimated marker effects (Meuwissen et al., 2001). This approach has resulted in great genetic gain due to significant correlation between GEBVs and true breeding values. Since then three extensions have been proposed to improve GS: weighted genomic selection (WGS) (Heffner et al., 2010), optimal haploid value (OHV) (Daetwyler et al., 2015), and optimal population value (OPV) (Goiffon et al., 2017). The conventional GS method, GEBV, can assure accelerating short-term gain, but doesn't guarantee achieving long-term gain (Jannink, 2010). To maximize long-term response, the first extension, WGS, has been proposed as a variation of GS where marker effects are weighted to increase the frequency of rare favorable alleles (Goddard, 2009). The second extension, OHV, calculates the breeding value of the best possible double haploid derived from an individual (Daetwyler et al., 2015). This method focuses selection on the haplotype and optimizes the breeding program toward its end goal of generating an elite fixed line (Daetwyler et al., 2015). GS, WGS, and OHV are truncation selection approaches as they rank individuals and select the top ones (typically a fraction of the population based on the available re-

sources), but recently OPV proposed a different strategy that is population-based. OPV selects the best population based on the interactive population effects which calculates the breeding value of the best possible progeny produced after an unlimited number of generations (Goiffon et al., 2017). Like OPV, we focus on selecting sets of individuals as a unit by proposing an innovative method, look-ahead selection (LAS). Our new selection method can improve the genetic gain by maximizing the probability of producing outstanding progenies in the final targeted generation. The proposed method can focus on achieving both short-term and long-term genetic gains and has the flexibility of adjusting based on the deadline and resource availability.

2.2 Materials and Methods

In this section, we present a uniform formula for all existing GS methods namely, conventional genomic selection (CGS), weighted genomic selection (WGS), optimal haploid value (OHV), optimal population value (OPV), and our new selection method, look-ahead selection (LAS). Equations 2.1, 2.2, and 2.3 show this uniform optimization formulation. It should be observed that the only difference among these four existing methods is in their objective functions as they aim to maximize different objectives. Equations 2.4, 2.5, 2.6 and 2.7 show the objective functions respectively for CGS, WGS, OHV, and OPV. x_n is a binary variable that shows whether individual n is selected ($x_n = 1$) or not ($x_n = 0$). Each method aims to select a subset of population (S individuals) as shown by equation 2.2.

$$\max_x \quad F^{GS} \tag{2.1}$$

$$\text{such that} \quad \sum_{n=1}^{N} x_n = S \tag{2.2}$$

$$x_n, \in \{0, 1\}, n \in \{1, ..., N\} \tag{2.3}$$

Here we define the notations used in this paper:

- N: The number of individuals in the population.

- L: The number of marker loci.
- $G \in \{0,1\}^{L\times M\times N}$: The genotypic information of individual n.
- β_l: The effect of having the major allele at locus l.
- M: The ploidy of the plants. We consider that the plants are diploid so $M = 2$.

We break down GS methods into two groups: 1. Truncation selection; and 2. population-based selection. In truncation selection approaches (CGS, WGS, and OHV), an individual is selected by ranking the candidates based on a method and then a fraction of the population with highest values are selected. In population-based selection approaches (OPV and LAS), a group of individuals that make the best combination are selected.

The objective function of the optimization problem, F^{GS} is formulated as F^{CGS}, F^{WGS}, F^{OHV}, and F^{OPV} in equations 2.4, 2.5, 2.6, and 2.7 respectively. As shown by Meuwissen et al. (2001), an individual's genomic estimated breeding value (GEBV) is the sum of all marker effects across the entire genome (2.4). This conventional GS method ranks the individuals based on their GEBVs and selects the ones with highest GEBV.

$$F^{CGS} = \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m=1}^{2} G_{l,m,n}\beta_l x_n. \tag{2.4}$$

Simulation and some empirical studies have shown that the CGS selection results in rapid genetic gains (Hayes et al., 2009; Lorenzana and Bernardo, 2009; VanRaden et al., 2009; Jannink, 2010). However, CGS focuses on one or two cycles of selection and does not guarantee long-term gain.

WGS has been proposed as a variation of CGS model that can preserve more favorable alleles than CGS. In this model (2.5), marker effects are weighted to increase the frequency of rare favorable alleles (Goddard, 2009) (Jannink, 2010).

$$F^{WGS} = \sum_{n=1}^{N}\sum_{l=1}^{L}\sum_{m=1}^{2} G_{l,m,n}\frac{\beta_l}{\sqrt{\max(w_l, 1/n)}}x_n. \tag{2.5}$$

The weight, w_l is defined as the fraction of favorable alleles to the number of individuals in the population. This model gives a higher weight to the markers that have low-frequency favorable alleles.

For OHV, OPV, and our new method, LAS, we clustered markers into haplotypes to define haplotype blocks as adjacent markers are very likely to segregate together. The following definitions will be used to take the blocks into account:

- B: the number of haplotype blocks per chromosome.
- $H(b), \forall b \in \{1, ..., B\}$: the set of marker loci that belong to haplotype block b.

Double haploids (DH) have been routinely used in breeding programs to accelerate the process. OHV has been proposed to combine the creation of double haploids with GS methods and evaluates the potential of producing elite double haploids (Daetwyler et al., 2015). Equation 2.6 shows the objective function for OHV selection. The OHV of individual n is the GEBV of the best possible DH individual derived from it. This method ranks the individuals based on their OHV and selects the best ones (Daetwyler et al., 2015).

$$F^{OHV} = 2\sum_{n=1}^{N}\sum_{b=1}^{B}\max_{m\in\{1,2\}}\sum_{l\in H(b)} G_{l,m,n}\beta_l x_n. \tag{2.6}$$

Simulation studies have shown that OHV selection results in more genetic gain and diversity when compared to conventional GS method (Daetwyler et al., 2015).

As discussed, the second group of GS methods (OPV and LAS) focus on population-based approaches. OPV selection, a population-based selection method, is an extension to OHV which evaluates the breeding merit of a set of individuals instead of evaluating the breeding value of a single individual (Goiffon et al., 2017). The OPV of breeding population S is the GEBV of the best possible progeny produced after an unlimited number of generations. Mathematically, OPV is defined as (Goiffon et al., 2017):

$$F^{OPV} = 2\sum_{b=1}^{B}\max_{n\in\{1,...,N\}}\max_{m\in\{1,2\}}\sum_{l\in H(b)} G_{l,m,n}\beta_l x_n. \tag{2.7}$$

CGS, WGS, OHV and OPV have three major limitations: 1. none of these methods are time dependent, 2. none of these methods gives an optimal strategy for mating, and 3. resource allocation is not taken into account. These three limitations serve as the major motivation for this study. In this paper, we define an innovative selection method, LAS which can address these issues by selecting individuals based on time and resource constraints, and giving an optimal mating strategy. This new method is described in detail in next section.

2.2.1 Look-ahead selection

LAS looks into the future and estimates the breeding value of progenies for the terminal generation. The main idea is to look ahead for a predefined number of generations in stochastic simulation process so that future information can be incorporated in finding an optimal selecting and pairing strategy for the current generation. The goal is to increase the probability of producing outstanding progenies in the terminal generation. Figure 2.1 displays the look-ahead stochastic simulation.

To start, S individuals (in Figure 2.1, $S = 8$) are picked as breeding parents from the initial population (generation t). These breeding parents are paired sequentially to make crosses and produce the next generation (generation $t+1$). From this generation forward, a lot of random crosses are made to produce progenies for upcoming generations. This process will continue until getting a large number of progenies in the terminal generation (generation T). Now we can evaluate the selected breeding parents as a group by looking at the breeding values of progenies. The key point is that LAS has the ability of estimating the breeding value of progenies without necessarily going through all generations. We have formulated the transition probabilities that allow the simulation jump from generation $t+1$ to the targeted genetration (generation T). These transition probabilities are defined as *Look-ahead inheritance distribution* in 2.2.1.

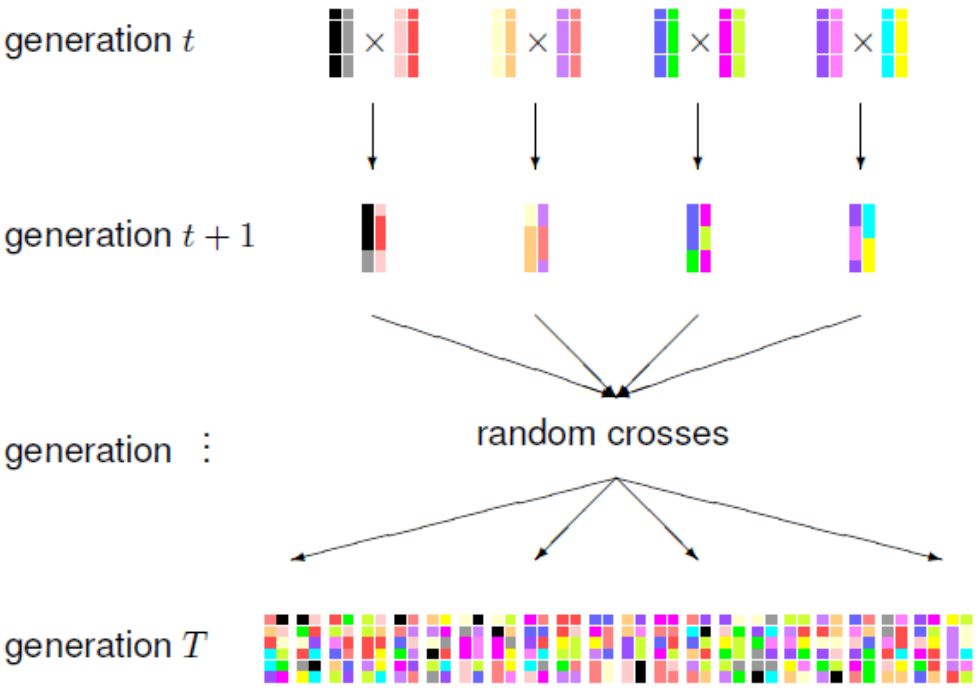
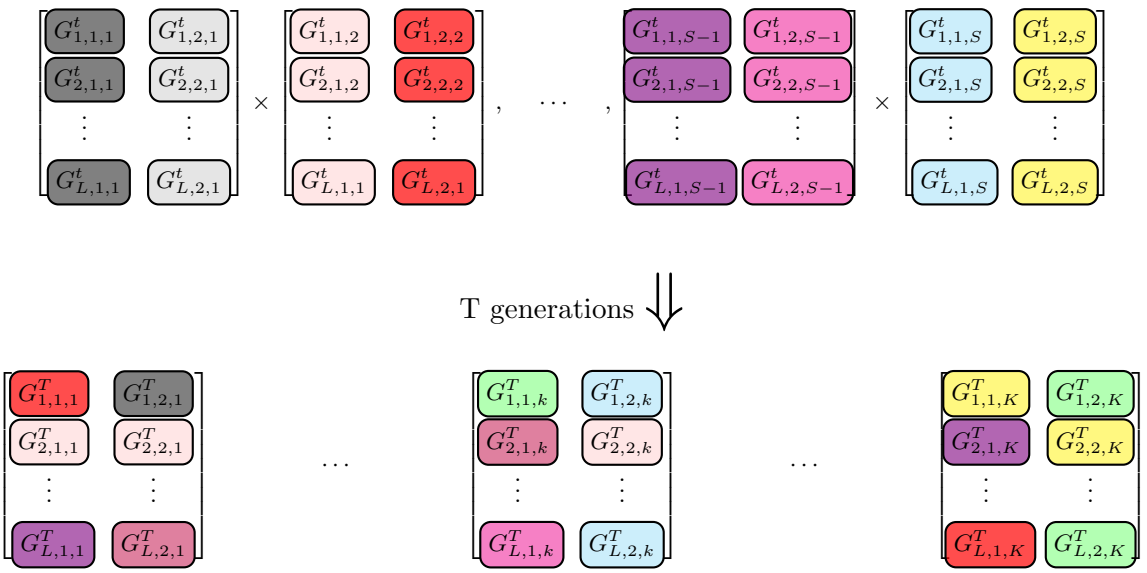


Figure 2.1 Look-ahead stochastic simulation.

Definition 2.2.1. For a given vector of recombination frequencies, $r \in [0, 0.5]^{L-1}$, and a given set of individuals, S, the *Look-ahead inheritance distribution* is defined as transition probabilities in equation 2.8 and 2.9:

$$P(G^T_{1,m',k} = G^t_{1,m,i}) = \frac{1}{2S}, \forall i \in \{1, 2, ..., S\}, \forall k \in \{1, 2, ..., K\}, \forall m, m' \in \{1, 2\}. \quad (2.8)$$

Where $G^T \in \{0, e_l\}^{L\times 2\times K}$ is the genotypic information of random progenies produced in terminal generation, and S is the number of breeding parents selected from the initial population. Equation 2.8 explores the transition probability for the the first locus and states that the first allele of a progeny has an equal probability of inheriting information from the initial population. The following matrices show the genotypic notation for the breeding parents and the progenies that were produced T generations later. The color codes are a representation of recombination.



$$T_{i,j,l} = P(G^T_{l+1,m',k} = G^t_{l+1,m'',j} | G^T_{l,m',k} = G^t_{l,m,i}), \forall i, j \in \{1, 2, ..., S\}, \tag{2.9}$$

$$\forall l \in \{1, 2, ..., L-1\}, \forall k \in \{1, 2, ..., K\}, \forall m, m^{'}, m^{''} \in \{1, 2\}$$

Equation 2.9 explores the transition probability for all the loci rather than the first one. This equation describes the transition matrix of inherited genetic information and is defined mathematically in equation 2.10. This transition matrix is an extension to the simple case of having one pair of breeding parents, described by Han et al. (2017).

$$T_{i,j,l} = \begin{cases} (1-r_l)^2(1-R_l), & \text{if } j \in J_1 \\ r_l(1-r_l)(1-R_l), & \text{if } j \in J_2 \\ \frac{1}{2}\, r_l(1-R_l), & \text{if } j \in J_3 \\ \frac{\frac{1}{4}\, R_l}{\frac{S}{2}-1}, & \text{Otherwise} \end{cases}, \tag{2.10}$$

$$\forall l \in \{1, 2, ..., L-1\}, \forall i, j \in \{1, 2, ..., S\}$$

Where:

$$J_1 = i \tag{2.11}$$

$$J_2 = 4\lceil i/2 \rceil - i - 1 \tag{2.12}$$

$$J_3 = 8\lceil i/4 \rceil - i - 3 \quad \text{or} \quad i + 2\sqrt{2}(\sin(\frac{i\pi}{2} - \frac{\pi}{4})) \tag{2.13}$$

Here, R_l is the look-ahead recombination frequency defined in the APPENDIX as (.0.1):

$$R_l = \frac{(S/2-1)(1-(1-r_l)^t)}{S/2} \tag{2.14}$$

To provide a more insightful description of the look-ahead transition matrix, we elaborate on four different cases of transition as follow:

Case 1: No recombination happens (J_1).

$$\begin{bmatrix} G^t_{1,1,1} \\ G^t_{2,1,1} \\ \vdots \\ G^t_{L,m,i} \end{bmatrix} = \begin{bmatrix} G^T_{1,m',k} \\ G^T_{2,m',k} \\ \vdots \\ G^T_{L,m',k} \end{bmatrix}$$

Case 2: Recombination happens within an individual (J_2).

$$\begin{bmatrix} G^t_{1,1,1} \\ G^t_{2,2,1} \\ \vdots \\ G^t_{L,m,i} \end{bmatrix} = \begin{bmatrix} G^T_{1,m',k} \\ G^T_{2,m',k} \\ \vdots \\ G^T_{L,m',k} \end{bmatrix}$$

Case 3: Recombination happens within the paired individual (J_3).

$$\begin{bmatrix} G^t_{1,1,1} \\ G^t_{2,1,2} \\ \vdots \\ G^t_{L,m,i} \end{bmatrix} = \begin{bmatrix} G^T_{1,m',k} \\ G^T_{2,m',k} \\ \vdots \\ G^T_{L,m',k} \end{bmatrix}, \text{ or } \begin{bmatrix} G^t_{1,1,1} \\ G^t_{2,2,2} \\ \vdots \\ G^t_{L,m,i} \end{bmatrix} = \begin{bmatrix} G^T_{1,m',k} \\ G^T_{2,m',k} \\ \vdots \\ G^T_{L,m',k} \end{bmatrix}$$

Case 4: This case considers all possible remaining recombination (J_4).

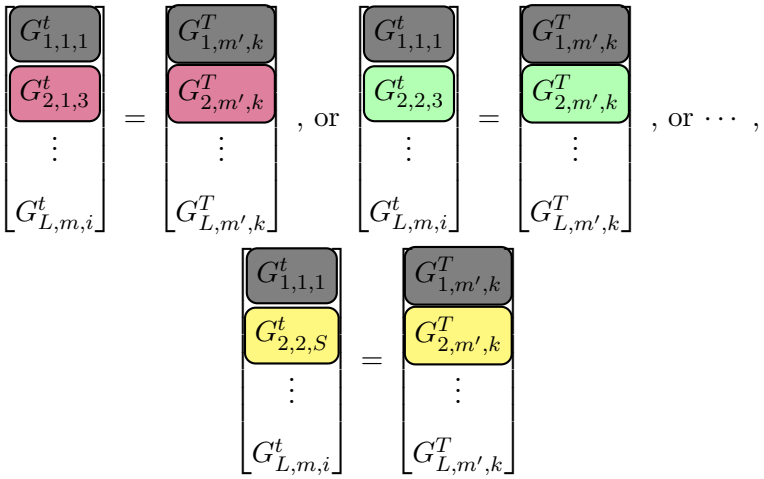


Figure 2.2 shows examples of these transitions through generating a single chromosome in the targeted generation.

2.2.2 Optimization of Look-ahead selection

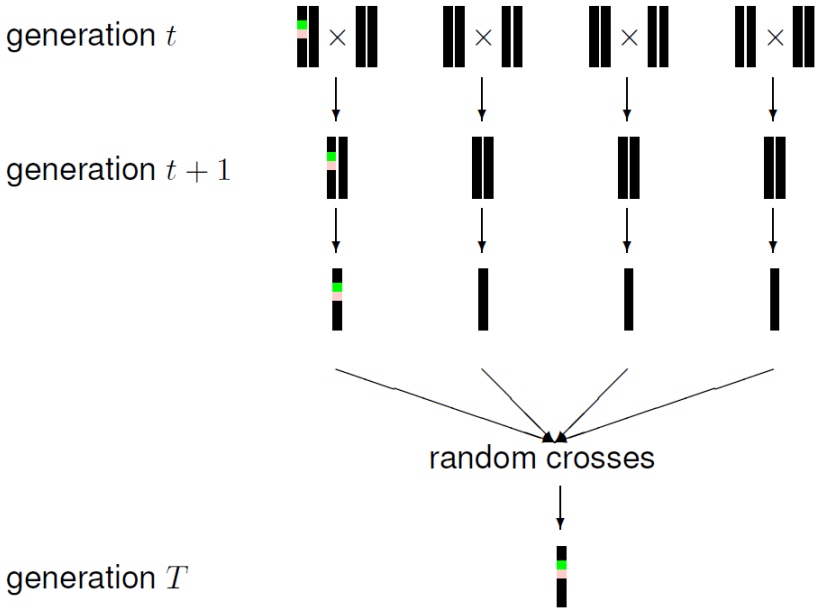
A decision-making model is formulated to find the optimal set of the breeding population in each generation. The decision variable, x_n, is a binary variable which becomes 1 when individual n is selected.

$$\max_x \quad F(S, g) \tag{2.15}$$

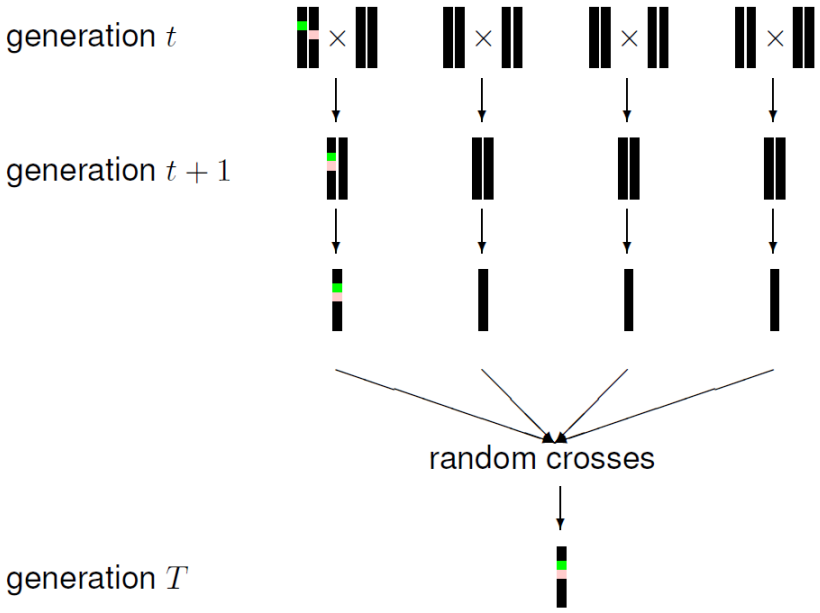
$$st. \quad \sum_{n=1}^{N} x_n = S \tag{2.16}$$

$$x_n \in \{0, 1\}, n \in \{1, ..., N\} \tag{2.17}$$

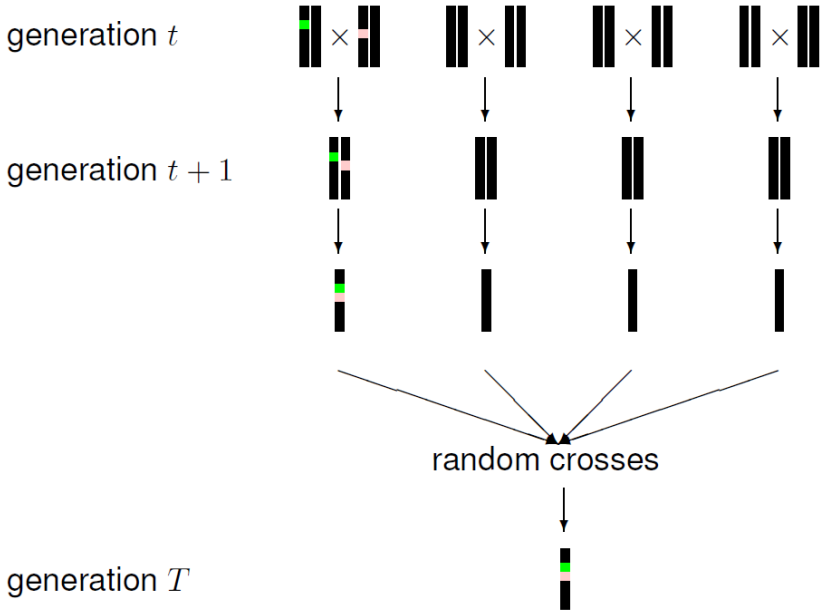
The objective function, F, is the probability of producing outstanding progenies in the terminal generation. The goal is to increase this probability, which is a function of selected breeding parents, S , and a threshold value. The threshold value, g, is a parameter that will help define an outstanding progeny. We say a progeny is outstanding if it has a GEBV greater than the threshold value. Constraint (2.16) indicates that S number of individuals will be selected in each generation as breeding parents.



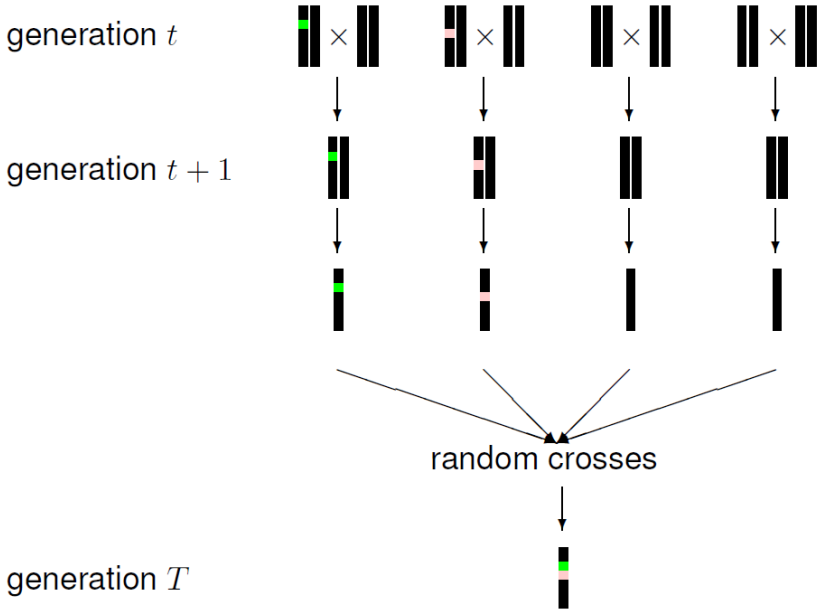
(a) Case 1: No recombination.



(b) Case 2: Recombination within an individual.



(c) Case 3: Recombination within the paired individual.



(d) Case 4: All possible remaining recombination.

Figure 2.2 Four different cases of transitions.

We can find F with a two-step simulation approach:

Step 1 : Produce K progenies according to Look-ahead inheritance distribution after t generations.

Step 2 : Get the proportion of outstanding progenies by dividing the number of outstanding progenies to K and use this proportion as an estimate of F.

To solve the model, we design a four-step heuristic algorithm:

Step 1 : Select S individuals randomly.

Step 2 : Find F.

Step 3 : Propose pairwise swaps between a selected individual and every other unselected one, evaluate F for all and keep the one with highest F.

Step 4 : Repeat step 3 until no improvements can be achieved.

2.3 Simulation

We compare four different methods of CGS, OHV, OPV, and LAS through simulation implemented in MATLAB. The genetic data and recombination rates are based on Goiffon et al. (2017). Genetic data contains 369 maize inbred lines with approximately 1.4 million SNPs. To facilitate the comparisons, the genetic data was scaled such that the maximum potential of the initial breeding population is 100. Similar to Goiffon et al. (2017), we assumed that marker effects were known.

In this paper, the plant breeding process starts with the initial population and iteratively goes through: 1. selection 2. reproduction. This continues until getting the final population in T generations. Figure 2.3 describes the in silico breeding process (Goiffon et al., 2017). Four different methods of CGS, OHV, OPV and LAS are used in the selection step for comparison.

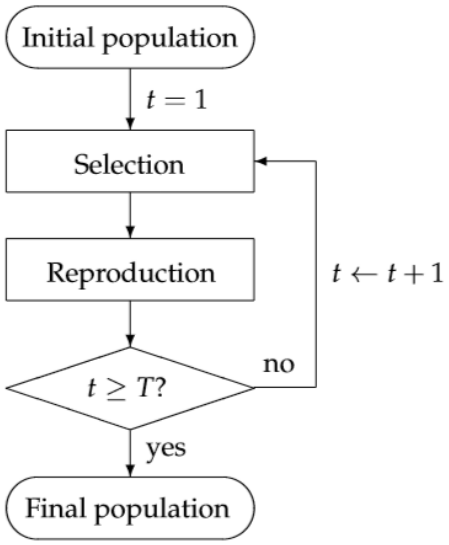


Figure 2.3 The simulation diagram.

Two hundred, individuals are randomly selected in each selection step. To make the comparisons consistent we used the same set of initial population for all GS methods. A breeding population of $S = 20$ individuals are selected in each generation to make 10 crosses. The number of progenies produced for each cross is proportional to the genetic diversity of the breeding parents. This results in having different number of progenies for each cross by producing more progenies for the breeding parents that have a higher genetic diversity.

The number of haplotype blocks, B, and the discarded percentage of individuals, F, are two parameters that can effect the performance of selection methods. When B is small, the selection method will focus on genetic gain in short-term while when it is a large number the selection method will focus on long-term gain. F shows the percentage of individuals with the lowest GEBV that will be removed before optimizing the selection strategy. When F is large, the process focuses on short-term gain while when it is a small number the focus would be on long-term gain. The best values for B and F were determined in an experiment by Goiffon et al. (2017) through testing different combinations of parameters. We adopted the same optimized parameter setting which is $B = 12$, and $F = 70\%$ for OHV and,

$B = 1$, $F = 40\%$ for OPV. In Look-ahead selection, we do not remove any lowest GEBV individuals from the population, but we define haplotype blocks. Here B is set to be 1000. Additionally, LAS has one more parameter which is the number of progenies (K) produced by the look-ahead method. Here we set $K = 10000$. The number of progenies produced by the look-ahead method needs to be large enough to be able to capture different inheritance possibilities and needs to be small enough due to time constraints. In the next section, we will discuss the results from the simulation.

2.4 Results

One thousand independent simulations were performed for four selection approaches. From each approach, the cumulative distribution functions (CDFs) of the population maximum in final generation were generated and compared (Figure 2.4).

It should be noted that the best CDF curve should be the farthest right as the vertical value of each point on the curve gives the percentage of random outcomes which have a lower GEBV than the corresponding horizontal value. To provide a more insightful assessment of different selection methods, we identified markers on each curve with a 10 percent interval. Comparing all methods for the same percentile makes it clear that LAS has the higher phenotype without any exceptions. As can be seen from the CDF curves, LAS outperforms truncation selection methods (GEBV and OHV), and also outperforms the only population-based method (OPV) at every percentile.

Furthermore, simulation results show that population-based methods preserve more genetic diversity. Figure 2.5 displays the genetic diversity of four GS methods in 10 generations where genetic diversity has been defined as the difference between the maximum potential and the minimum potential of the current generation. We see that LAS loses diversity faster in short term, but then has a consistent rate until losing more diversity in final generations. Overall, population-based methods, OPV and LAS seem to be a better approach in preserving genetic diversity.

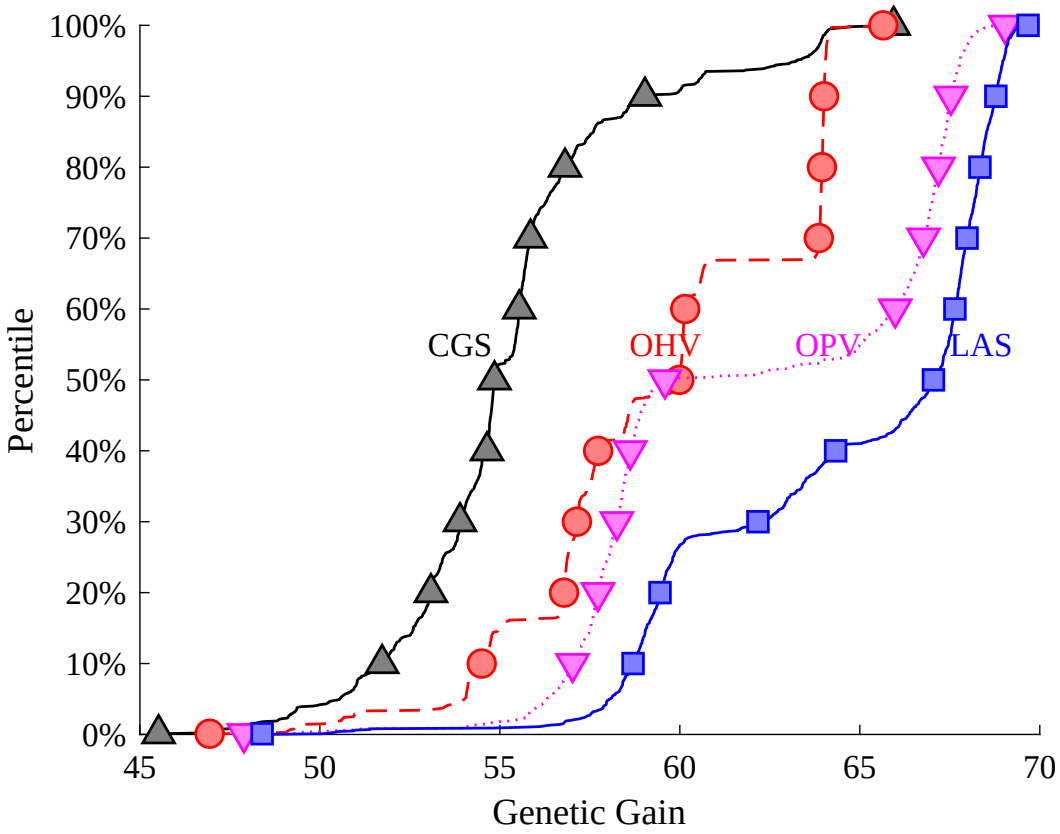


Figure 2.4 CDFs of population maximum for four selection methods.

Figure 2.6 shows the genetic gain in each generation. We define the genetic gain as the difference between the mean GEBV of the current generation and the initial mean GEBV. The interesting thing is that in the first generation, LAS rises faster than other three approaches and then increases with a consistent slope until generation 8. When reaching to the deadline LAS rises fast again. CGS performs well in the first three generations and then the curve flatters. Similarly, OHV and OPV have a higher slope for the first two generations and then increase with a lower speed until the final generation. This validates LAS method is able to incorporate the deadline into the selection while other methods are not. In addition, the look-ahead selection is capable of making a trade-off between achieving short-term genetic gain and preserving long-term growth potential.

We examined the effectiveness of a look-ahead method against three state-of-the-art

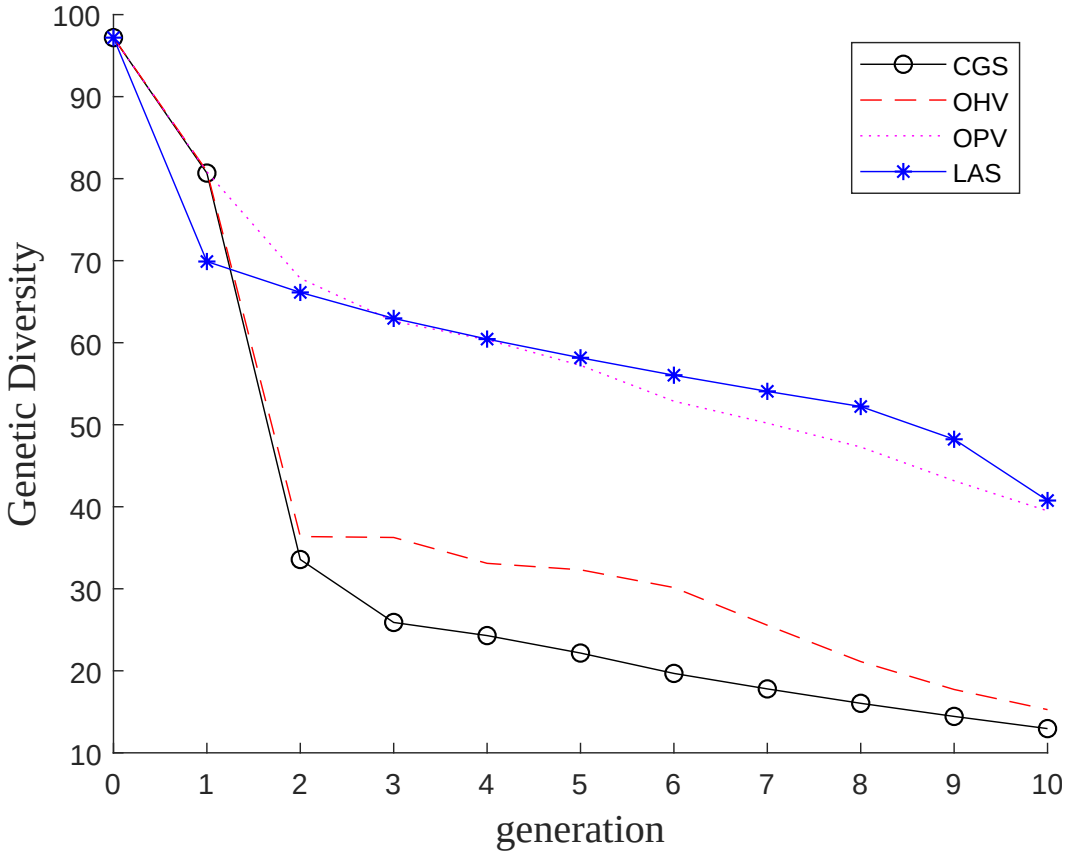


Figure 2.5 Genetic diversity in 10 generations for four GS methods.

selection methods including conventional genomic selection, optimal haploid value and optimal population value. Results of simulation show that LAS outperforms all other methods with no exception. We did not compare LAS with WGS since WGS has the similar growth rate to conventional GS method. In conclusion, LAS is able to achieve short-term genetic gain, preserve long-term genetic diversity and is sensitive to the deadline.

2.5 Conclusions

As global food demand increases, plant breeding has been critical in improving production yield. Genomic selection has been instrumental in efficiency improvement in plant breeding. In this study, we introduce a new selection method, LAS, which has the potential

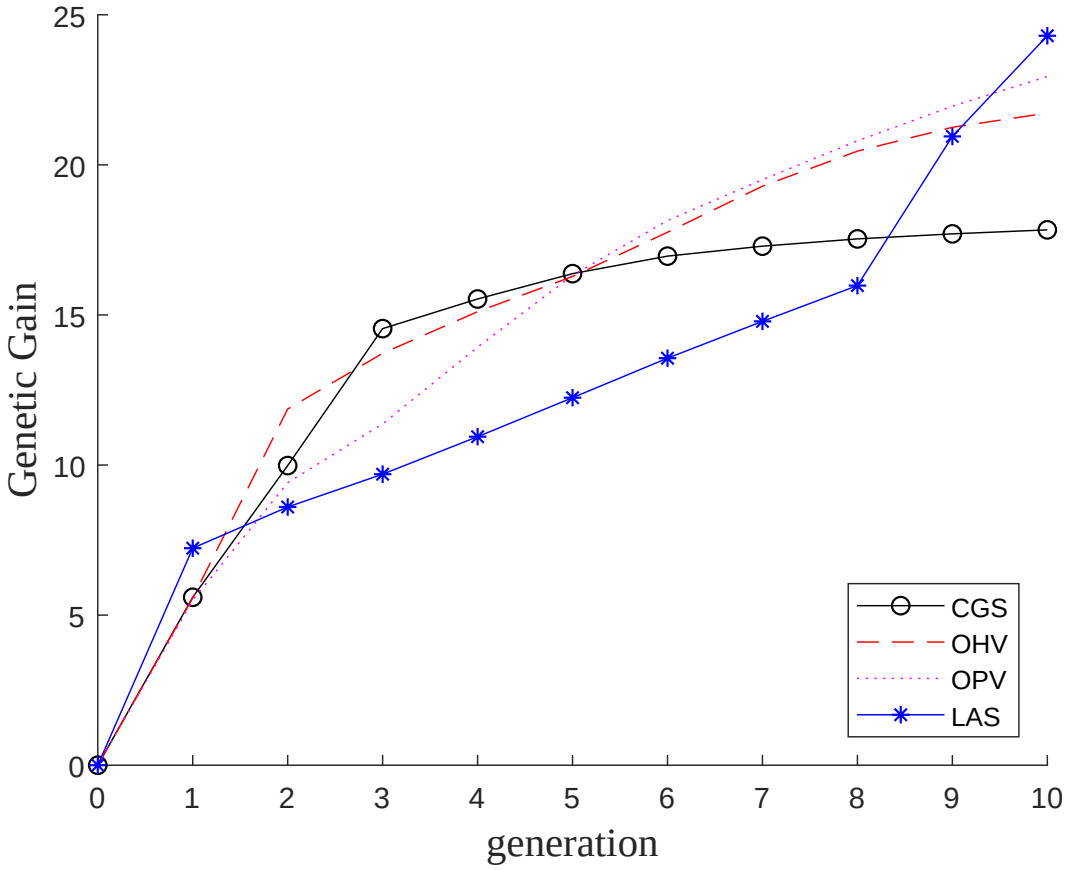


Figure 2.6 Genetic gain in 10 generations for four GS methods.

to improve the breeding efficiency given the limited resources and target delivery date.

The new selection method, look-ahead selection evaluates the genetic merit of a set of selection candidates. We showed that LAS outperforms other methods in a series of simulation experiments by using empirical data from an inbred maize population. LAS has three major contributions: The first one is time managing. This method is sensitive to the deadline and is able to make a trade-off between short-term genetic gain and long-term genetic diversity. The second contribution is optimizing the pairing strategy. This method selects an order dependent set of individuals as the breeding parents to find the best possible pairing strategy. Finally, the third contribution is allocating the recourses such that the number of progenies produced from each cross is proportional to the genetic

diversity of the breeding parents. This can preserve more genetic diversity in the breeding process. The research in this paper was subject to a few limitations which suggest future research directions. Further research can focus on resource allocation in genomic selection problems and utilize reinforcement learning for optimizing different parameters discussed in this research.

CHAPTER 3. GENERAL CONCLUSIONS AND FUTURE WORK

In order to feed the world's growing population, an interdisciplinary effort is needed. In this research, operations research tools are applied to the problem of genomic selection by integrating stochastic simulation and optimization. This paper considers three components of a breeding process that have been ignored in previous approaches. The first component is **time dependency**. The look-ahead selection method decides on the selection and mating strategy with considering the deadline. This results in making a trade-off between short-term genetic gain and long-term growth potential. The second component is **optimizing mating strategy**. The LAS method selects a set of individuals which are order dependent. This results in finding the best pair for each individual and optimizing the pairing strategy. The third component is **resource allocation**. While previous approaches make same number of progenies from each cross, LAS can vary the number of progenies based on the genetic diversity of their selected parents to produce more progenies for individuals which have more genetic diversity.

Recently, OPV was proposed as the first population-based selection method and now LAS pushes the frontier of population-based approaches. In this study, we see that population-based methods can preserve more genetic diversity than truncation selection methods.

LAS opens a potentially fruitful direction of genomic selection to future research. In this regard, two follow-up studies are recommended:

1) Investigate the performance of selection methods for different number of generations: In the present study, the effectiveness of four selection methods are compared in 10 generations. We believe that LAS can make a trade-off between short-term and long-term goals

for different number of generations as it is the only time dependent method. Future work can focus on comparing CGS, OHV, OPV, and LAS for different time horizons through simulation.

2) Applying reinforcement learning for allocating resources: Genomic selection is implemented in a breeding process to increase the response, but little is known how to allocate the resources optimally under a budget. Reinforcement learning (RL) is a type of machine learning that allows agents to automatically determine the ideal action within a specific context to maximize its performance. The agent can learn the optimal action by getting a reward feedback. Markov decision processes (MDP) are an intuitive and fundamental formalism for RL and other learning problems in stochastic domains. RL methods can be applied to GS for optimizing the resource allocation. Future research can focus on modelling a GS problem in the framework of MDP and using RL methods to allocate resources. To do this, states, actions and transition probabilities should be defined in the context of GS. In other words, the RL model can optimize the crossing, and pairing strategies as well as the number of progenies to be produced from each cross. This new research area would help breeders to utilize the resources according to their budget in an efficient manner.

BIBLIOGRAPHY

Bernardo, R. and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082–1090.

Brumlop, S. and Finckh, M. R. (2011). *Applications and potentials of marker assisted selection (MAS) in plant breeding*, volume 298.

Daetwyler, H. D., Hayden, M. J., Spangenberg, G. C., and Hayes, B. J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics*, 200(4):1341–1348.

Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, 136(2):245–257.

Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P. (2017). Optimal Population Value Selection: A Population-Based Selection Strategy for Improving Response in Genomic Selection. *Genetics*.

Han, Y., Cameron, J. N., Wang, L., and Beavis, W. D. (2017). The Predicted Cross Value for Genetic Introgression of Multiple Alleles. *Genetics*, 205(4):1409 LP – 1423.

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92(2):433–443.

Heffner, E. L., Lorenz, A. J., Jannink, J. L., and Sorrells, M. E. (2010). Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Science*, 50(5):1681–1690.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution*, 42(1):1–11.

Lande, R. and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., and Jannink, J. L. (2011). *Genomic Selection in Plant Breeding. Knowledge and Prospects.*, volume 110. Elsevier Inc., 1 edition.

Lorenzana, R. E. and Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, 120(1):151–161.

Mcdowell, R., Beavis, W., Professor, C.-M., and Lübberstedt, T. (2016). Genomic selection with deep neural networks.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Rincent, R., Charcosset, A., and Moreau, L. (2017). Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theoretical and Applied Genetics*, 130(11):2231–2247.

VanRaden, P., Van Tassell, C., Wiggans, G., Sonstegard, T., Schnabel, R., Taylor, J., and Schenkel, F. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science*, 92(1):16–24.

APPENDIX. PROOF FOR CHAPTER 2

Definition .0.1. For a given vector of recombination frequencies, $r \in [0, 0.5]^{L-1}$, and a given set of individuals, the look-ahead recombination frequency, $R \in [0, 0.5]^{L-1}$ is defined as:

$$R = \frac{(S/2-1)(1-(1-r)^t)}{S/2} \tag{.1}$$

Proof. Define P_i as the probability that two consecutive alleles would stay together after i generations:

$$\begin{aligned}
P_0 &= 1 \\
P_1 &= P_0(1-r_l) + \frac{r_l}{S/2} \\
P_2 &= P_1(1-r_l) + \frac{r_l}{S/2} \\
&\vdots \\
P_t &= P_{t-1}(1-r_l) + \frac{r_l}{S/2}
\end{aligned}$$

Where r_l is the l^th recombination frequency for $l \in \{1, 2, ..., L\}$ and S is number of breeding parents. The last equation can be expanded as follow:

$$\begin{aligned}
P_t &= P_{t-1}(1-r_l) + \frac{r_l}{S/2} \\
&= P_{t-2}(1-r_l)^2 + (1-r_l)\frac{r_l}{S/2} + \frac{r_l}{S/2} \\
&= P_{t-3}(1-r_l)^3 + (1-r_l)^2\frac{r_l}{S/2} + (1-r_l)\frac{r_l}{S/2} + \frac{r_l}{S/2} \\
&\vdots \\
&= (1-r_l)^t + (1-r_l)^{t-1}\frac{r_l}{S/2} + \cdots + (1-r_l)\frac{r_l}{S/2} + \frac{r_l}{S/2} \\
\Rightarrow P_t &= (1-r_l)^t + \frac{r_l}{S/2}\left(\sum_{i=0}^{t-1}(1-r_l)^i\right) \\
&= \frac{1 + (S/2-1)(1-r_l)^t}{S/2}
\end{aligned}$$

We get the last equation by using the finite geometric series formula. From this we obtain P_t, the probability that two consecutive alleles stay together after t generations. Next, we compute $P_t^{'}$, the probability that two consecutive alleles would recombine after t generations:

$$\begin{aligned}
P_t^{'} &= 1 - P_t \\
&= 1 - \frac{1 + (S/2-1)(1-r_l)^t}{S/2} \\
&= \frac{(S/2-1)(1-(1-r_l)^t)}{S/2}
\end{aligned}$$

□