Complementarity-based selection strategy for genomic selection

Saba Moeinizade  
*Iowa State University, sabamz@iastate.edu*

Megan Wellner  
*Iowa State University*

Guiping Hu  
*Iowa State University, gphu@iastate.edu*

Lizhi Wang  
*Iowa State University, lzwang@iastate.edu*

Follow this and additional works at: https://lib.dr.iastate.edu/imse_pubs

Part of the Agronomy and Crop Sciences Commons, Genomics Commons, Operational Research Commons, and the Plant Breeding and Genetics Commons

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/imse_pubs/236. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Industrial and Manufacturing Systems Engineering at Iowa State University Digital Repository. It has been accepted for inclusion in Industrial and Manufacturing Systems Engineering Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Complementarity-based selection strategy for genomic selection

Abstract
Genomic selection is a technique that breeders use to select plant or animal individuals to mate and produce new generations of species. The conventional selection method is to select individuals that are either observed or predicted to be the best based on the assumption that parents with better phenotypes will produce better offspring. A major limitation of this method is its focus on the short-term genetic gains at the cost of genetic diversity and long-term growth potential. Recently, several new genomic selection methods were proposed to maximize the long-term potential. Along this research direction, we propose a new method, the complementarity-based selection strategy (CBS), to improve the tradeoff between short-term genetic gain and long-term potential. This approach is inspired by the genetic compatibility mate-choice mechanism in animals. Our selection method selects the individual with the highest genomic estimated breeding value to emphasize short-term achievement and then pairs it with the individual that is the most complementary to the one with highest genomic estimated breeding value to emphasize the long-term growth potential. The CBS method allows favorable alleles to be accounted for within the selection and more of them to be included. We present simulation results that compare the performance of the new method against the state-of-the-art methods in the literature and show that the CBS approach has a great potential to further improve long-term response in genomic selection.

Keywords
Genomic Selection, Genetic Gain, Operations Research, Simulation

Disciplines
Agronomy and Crop Sciences | Genomics | Operational Research | Plant Breeding and Genetics

Comments
This is the peer-reviewed version of the following article: Moeinizade, Saba, Megan Wellner, Guiping Hu, and Lizhi Wang. "Complementarity-based selection strategy for genomic selection." Crop Science 60, no. 1 (2020): 149-156, which has been published in final form at DOI: 10.1002/csc2.20070 This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving. Posted with permission.
Complementarity-based Selection Strategy for Genomic Selection

Saba Moeinizade, Megan Wellner, Guiping Hu, and Lizhi Wang

Industrial and Manufacturing Systems Engineering Department
Iowa State University
Ames, IA 50011, United States

ABSTRACT
Genomic selection is a technique that breeders use to select plant or animal individuals to mate and produce new generations of species. The conventional selection method is to select individuals that are either observed or predicted to be the best based on the assumption that parents with better phenotypes will produce better offspring. A major limitation of this method is its focus on the short-term genetic gains at the cost of genetic diversity and long-term growth potential. Recently, several new genomic selection methods were proposed to maximize the long-term potential. Along this research direction, we propose a new method, the complementarity-based selection strategy, to improve the tradeoff between short-term genetic gain and long-term potential. This approach is inspired by the genetic compatibility mate-choice mechanism in animals. Our selection method selects the individual with the highest genomic estimated breeding value to emphasize short-term achievement and then pairs it with the individual that is the most complementary to the one with highest genomic estimated breeding value to emphasize the long-term growth potential. The complementarity-based selection method allows favorable alleles to be accounted for within the selection and more of them to be included. We present simulation results that compare the performance of the new method against the state-of-the-art methods in the literature and show that the CBS approach has a great potential to further improve long-term response in genomic selection.
INTRODUCTION

The continuing growth of the human population leads to the increase in global food and water demand (Rosegrant and Cline 2003, Haghiri et al. 2018). The agriculture industry has faced increasing challenges in producing more food with limited resources (Godfray et al. 2010). Plant breeding has been contributing to solving these challenges (McCouch et al. 2013). Recently, analytical decision making tools have been developed to support the plant breeding industry (Varshney et al. 2016, Shahhosseini et al. 2019). Operations Research is a discipline that applies scientific methods to provide a quantitative support for decision making (Morse, Kimball, and Gass 2003). With the growing needs for plant breeding in the agriculture industry, the role of operations research has become important with the increasing complexity in decision making.

In plant or animal breeding, one of the most important decisions is parental selection. Traditionally, selections are based on the phenotype of parents under the assumption that better parents would lead to better offspring (Lobos et al. 2017). However, phenotyping can be labor intensive, expensive, time consuming, and destructive (Akdemir, Sanchez, and Jannink 2015). Applying alternative methods for selection can help breeders make better decisions to produce the best offspring.

A recent innovation in parental selection is called genomic selection. Conventional genomic selection (CGS) methods use genotype information to make selection decisions. Predictions for offspring are then made based on the highest genomic estimated breeding value (GEBV), instead of observed phenotypic values (Spindel et al. 2016). This method is particularly useful because genotypic information is increasingly affordable hence becoming more easy to implement than phenotyping (Akdemir, Sanchez, and Jannink 2015). Furthermore, recent improvements to
conventional genomic selection strive to save breeders time, money, and energy by selecting top performers that will create the best offspring (Spindel et al. 2016). These approaches attempt to improve crop yield, and/or optimize desired traits within plants such as height.

Among those recent approaches to improve the conventional genomic selection method, the first is weighted genomic selection (WGS). This approach highlights the fact that favorable low frequency alleles should be included in selection and therefore gives more weight to such alleles (Goddard 2009). A study showed that with WGS, long-term performance is improved partially because of the reduced number of favorable alleles lost (Lorenz et al. 2011). The second approach is the optimal haploid value (OHV), which predicts the best doubled haploid that can be produced (Daetwyler et al. 2015). Doubled haploid is the genotype formed when individuals undergo chromosome doubling (Murovec and Bohanec 2012). This method selects best haplotypes to produce elite doubled haploids, which contributes to a greater genetic gain than conventional genomic selection (Daetwyler et al. 2015). Finally, optimal population value (OPV) is created to select the candidate breeding population as a unit. This strategy looks at the haploid values of individuals but also looks at the potential of a subgroup of individual candidates to evaluate how the selected individuals are as a group (Goiffon et al. 2017). The OPV method was found to achieve greater responses than CGS, WGS, and OHV in a series of simulation experiments that spanned 10 generations (Goiffon et al. 2017).

In this paper, we design a new genomic selection method, complementarity-based selection (CBS), to further improve the response of breeding and balance the trade-off between short-term genetic gain and long-term potential. CBS is inspired by the genetic compatibility mate-choice mechanism in animals. Mate choice can be defined as any pattern of behavior that leads members of a certain sex to be more likely to mate with certain members of the opposite sex (Halliday 1983). The
genetic compatibility mate-choice mechanism indicates the importance of non-additive genetic effects from choosing a mate with alleles that complement the genome of the chooser in addition to the additive genetic effects (Andersson and Simmons 2006, Neff and Pitcher 2005). The genetic dissimilarity of potential mates can also provide fitness benefits for the offspring (Mays and Hill 2004). This concept becomes of an interest when looking at plants. Following this analogy, the proposed CBS method selects an individual with the highest GEBV to emphasize short-term achievement and then selects the most complementarity individual as its pair to emphasize long-term potential. The CBS approach allows for the inclusion of alleles that are missing in the individual with the highest GEBV by defining the complementary score. This allows favorable alleles to be accounted for within the selection and more of them to be included. In the remaining of this manuscript, “performance individual” represents the individual with the highest GEBV and “complementing individual” represents the individual that is the most complementary to the one with highest GEBV. We compare the performance of CBS approach with the previous approaches (CGS, OHV, and OPV) in simulation experiments and show that the new methodology has potential benefits to increase the efficiency of plant breeding.

MATERIALS AND METHODS
Genomic selection starts with an initial population of individuals that contains the candidates for breeding. Within that population, certain criteria determine how many and which individuals should be selected and crossed to create the next generation of plants. These individuals are determined in an attempt to create the best progeny for later generations. The main information used in genomic selection techniques is the genotype of plants and the estimated marker effects. Let \( A \in \{0, 1\}^{l \times m \times n} \) be the genomic information of locus \( l \) from chromosome \( m \) in individual \( n \) for all \( l \in \{1, 2, ..., L\}, \ m \in \{1, 2\}, \) and \( n \in \{1, 2, ..., N\} \) with 1 and 0 representing the major and minor
alleles respectively and let $e_l$ be the marker effect at locus $l$. The genomic estimated breeding value (GEBV) of individual $n$ is defined as the sum of all marker effects across the entire genome which is described mathematically in equation (1). The ultimate goal is to maximize the GEBV of the offspring and therefore their desirable traits.

$$GEBV(n) = \sum_{l=1}^{L} \sum_{m=1}^{M} A_{l,m,n} \cdot e_l$$

In equation 1, $L$ is the total number of marker loci and $M$ represents the ploidy of the plant. To select the performance individual, we use the approach that selects the individual with the largest GEBV among the individuals in the population. Next, a complementary score is defined so that the most complementary individual with respect to the performance individual can be selected as its pair. This complementary score between the performance individual, $p$ and the complementing individual, $c$ is defined as:

$$Score(p, c) = \sum_{l=1}^{L} (\sum_{m=1}^{M} A_{l,m,p} + \sum_{m=1}^{M} A_{l,m,c})^f \cdot e_l$$

Here $f \in [0,1]$ is the scaling factor. We calculate this score for any possible pair and then find individual $c$ that maximizes this score for a given individual $p$. The selected individual $c$ is assigned as the complementing individual and then paired with the performance individual, $p$. These two will be crossed with one another and the process is repeated with the other individuals from the population by choosing the next best GEBV and complementary score. To provide a more insightful description of the proposed method, we will go through an example of the CBS method.

**Example of the complementary-based selection method**

Let’s assume we have a population of 4 diploid individuals ($M = 2$) with the following genomic information and effect vector (assuming that each individual has only three SNPs for simplicity) (Figure 1):
The goal is to select two individuals to be the parents of next generation. The following describes the steps taken in the CBS method to come up with the best strategy of selection.

**Step1: Select the performance individual as the one with the highest GEBV**

Here the GEBV values for four individuals are 14, 20, 12, and 15 respectively. Hence, we select line 2 as the performance individual.

**Step2: Find the complementary score for each possible pair**

Given that line 2 is selected as the performance individual, we have three possible pairs of (Line 2, Line 1), (Line 2, Line 3), (Line 2, Line 4). Note that we disregard (Line 2, Line 2) since self-pollination is not considered in this example. According to equation 2, the scores for these three pairs can be calculated. Here, we set the scaling factor, $f = 0.7$. These scores are 27.6166, 25.1951, and 27.0334 respectively. Figure 2 is an illustration of the score calculation for one of the pairs.

**Step3: Assign the complementing individual to be the one with the highest complementary score**

We see that the highest score is for (Line 2, Line 1), therefore Line 1 is selected as the complementing individual.

In the next section, we describe the simulation platform designed to run the experiments for comparison between different methods.

**Simulation Platform**

The simulation platform is implemented in Matlab for two independent datasets. The first dataset includes 369 maize individuals. This data contains genotypes, effects of desirable alleles, and recombination frequency information. Each individual contains about 1.4 million SNPs. The genetic dataset, effects and recombination frequencies are from those in the article by Goiffon et al. (2017). From the 369 individuals, 200 were randomly selected for each initial population. The
second dataset includes 400 soybean individuals. It contains 46,361 SNPs with effects simulated uniformly between zero and ten. The dataset is from the publication of Song et al. (2015). From the 400 individuals, 200 were randomly selected for each initial population. For the simulations of both datasets, the plants were diploids, meaning that the ploidy of the plants, $M$ is two. Figure 3 shows a diagram of the simulation process. Each simulation is spanned 10 generations so the total number of generations, $T$ equals ten ($T = 10$). Figure 3 illustrates the simulation process. The resource allocation is an important step in this simulation process. Here the resource allocation refers to the allocation of budget, number of crosses and number of progenies to be made from each cross in each generation. The resource allocation decisions should be optimized systematically, given the cost of making a cross and genotyping a progeny, under the budget and deadline constraints. In our simulation, we have a pre-defined budget of $20500 where making a cross costs $5 and making a progeny costs $10. For CGS, OHV and OPV the budget is divided into 10 generations equally and for every generation $k = 10$ crosses are made each producing 20 progeny as Goiffon et al. (2017). Table 1 summarizes the resource allocation strategy used for these three methods. The budget allocation is an important problem in genomic selection since it is not known if resources should be spent evenly over time or should more investment be made in earlier generations before genetic diversity deteriorates. We use a dynamic budget allocation approach for the CBS method, where the number of crosses, $k$ is being adapted as a function of time. The idea is to make more crosses and invest more in the earlier generations. For each subsequent generation, $k$ is defined as the number of generations remaining multiplied by 2 and then subtracted by 1, i.e. $k = 2 \times (T - t) - 1$, where $T$ is total number of generations and $t$ is the current generation number. Table 2 demonstrates the resource allocation strategy used for CBS method.
One of the major parameters of the CBS approach is $f$, defined as a number between 0 and 1 that represents the scaling factor of the different potentials of the model which is used in the complementary score definition (see equation 2). The scaling factor, $f$, and the number of crosses made for a given generation, $k$, are tuned through the simulation study with different values to find the combined parameters that result in the best performance and complementing individuals. If suboptimal $f$ and $k$ parameters are used, it would lead to less-desirable offspring in the final generation. It was found that the best combined parameters are $f = 0.70$ and $k = 2 \times (T - t) - 1$. These are the parameters used for the CBS method in our simulation study.

RESULTS

We performed 1,000 independent simulations for each genomic selection method (the CGS, OHV, OPV and CBS), which is consistent with the other studies in the literature (Goiffon et al. 2017). We have not included the comparisons between the performances of CBS with WGS because the WGS method represents a small variation of the CGS method, only emphasizing the preservation of rare favorable alleles. We have tested this method for the maize dataset and the results are indeed similar to the CGS (data not shown). To compare the performance of four different methods, the cumulative distribution functions of population GEBV maximum as well as population GEBV mean are obtained for two datasets. Moreover, the genetic gain values over ten generations are presented. Major results are summarized as follow:

Cumulative distribution functions of the population GEBV maximum in the final generation

For each selection method, the cumulative distribution functions (CDF) of the population GEBV maximum in the final generation are illustrated in Figures 4 and 5. In the CDFs, the vertical values correlate to the percentage of outcomes that have a GEBV lower than the horizontal values. The 1st percentile shows one of the worst performances in 1000 simulations whereas the 99th percentile
shows one of the best performances. That being said, the best selection method will be the farthest to the right at any point on the graph. In this simulation, the goal is to obtain the largest GEBV value. The points that are farther to the right have higher GEBV values with reference to the other approaches.

In the corn dataset, the CBS method outperformed truncation selection methods, namely, CGS and OHV, for the entirety of the CDF (see Figure 4). We also observe that the CBS method is superior to OPV in most of the percentiles and within around the 10th to the 55th percentile the CBS method outperformed OPV by a high margin. In other percentiles, the two approaches had similar performances. In the soybean dataset, the CBS method outperforms all the benchmark methods for the entirety of the study (see Figure 5).

**Cumulative distribution functions of the population GEBV mean in the final generation**

The CDF of population GEBV mean looks at the average individuals (instead of the maximum performers) for the final generation in each approach. The CDFs for population GEBV mean are shown in Figures 6 and 7.

Figure 6 demonstrates that in the final population of the corn dataset, the mean individual performers of the CBS strategy outperformed all of the others with the exception of OPV in the zero to five percentile range. Figure 7 shows that the CBS method of selection outperformed CGS, OHV, and OPV by a high margin in the soybean dataset with respect to the population mean. This means that the average individuals in the CBS selection strategy outperformed the average individuals for the other selection strategies in the final generation within 1000 simulations.

**Genetic gains over ten generations**

Figures 8 and 9 show the genetic gain over ten generations for all four selection methods. The genetic gain is calculated as the difference between the mean GEBV of the current population and
that of the initial population for each generation. One can see from Figures 8 and 9 that CBS had
more genetic gain per generation than all other approaches except at generation two, when CGS
had a slightly higher genetic gain in both datasets.

DISCUSSION

We examined the effectiveness of a complementarity-based method of selection against three
state-of-the-art methods including conventional genomic selection, optimal haploid value, and
optimal population value. The CBS strategy evaluates individuals based on how well the two
individuals work together. The CBS approach allows for the inclusion of alleles that are missing
in the individual with the highest GEBV by defining the complementary score. Our model strives
to emphasize the short-term achievement of crops by selecting the performance individual with
the highest GEBV and the long-term achievement by selecting a partner that is most
complementary to the performance individual. Simulation results demonstrate that compared to
CGS, OHV, and OPV the CBS method has more genetic gain per generation in both datasets (see
Figures 8, and 9). While CBS has more genetic gain per generation in both corn and soybeans, it
outperformed by a higher margin in the middle generations in the corn dataset. The soybean dataset
has more of a gradual curve whereas the corn dataset jumps greatly in genetic gain from generation
two to five. Furthermore, the CBS method outperformed the previous methods for almost the
entirety of both studies. In the corn dataset, CBS outperformed the three benchmark methods for
the 10th to 55th percentile for population GEBV maximum (see Figure 4). From the 1st to 10th and
the 55th to 100th percentiles CBS performed similarly to OPV while beating CGS and OHV. In
contrast, in the soybean dataset CBS outperformed all three methods by a high margin for the
entirety of the study regarding population GEBV maximum in the final generation (see Figure 5).
Looking at the average individuals in the final generation, we see that the CBS has outperformed
the previous methods for both datasets with the exception of 0th to 5th percentiles in the corn dataset (see Figures 6 and 7).

One of the interesting observations from the results is that CBS has outperformed the previous methods by a higher margin in the soybean dataset. A possible explanation for this is that the two species have a different genetic architecture. Another possible reason is that the soybean dataset contained 46,361 SNPs while the corn dataset contained 1.4 million which is a significant difference. It is worth mentioning that in OHV and OPV methods, the recombination events are assumed to be possible between haplotypes but not within them. This assumption helps reduce the computational effort of the algorithms. However, the CBS approach is computationally faster and doesn’t require this assumption. Hence, the CBS method does not group adjacent markers into haplotype blocks. Introducing haplotype blocks in a selection method adds more parameters to tune, making them less robust. As a result, for CBS method, the recalibration of parameters was not necessary for running the soybean simulation after doing the corn simulation which means that the CBS dataset is more robust.

CONCLUSION

As demand for crops and food throughout the world increases, meeting that demand is becoming more and more crucial. Genomic prediction which focuses on the accuracy of phenotype prediction has transformed breeding. However, selection and mating steps have not received equal attention and typically remain the same as in traditional phenotyping selection. Optimizing these two steps is the focus of this study. With better selection and mating strategies, bigger yields and better crops could be produced. In this paper, the new method for genomic selection that we propose, the complementarity-based selection method, helps improve yields and crop quality. The reason for
the improved performance compared to other methods is that the CBS approach allows for the
inclusion of alleles that are missing in the individual with the highest GEBV by searching for their
complementary partners. This allows favorable alleles to be accounted for within the selection and
more of them to be included. By defining a complementary score, our model strives to emphasize
the short-term achievement by selecting the performance individual with the highest GEBV and
also the long-term achievement by selecting a complementing individual.

The research in this paper is subject to a few limitations which suggest future research directions.
Firstly, further research could be done to better define complementary score. Our model designed
a simple and effective score but there could be a better way to define complementarity between
two individuals. Secondly, the CBS method takes selection on an individual basis focusing on the
complementarity of two individuals as parents. If this could be done with a group selection
approach which also focus on the complementarity between all selected individuals, perhaps the
results could be improved. Lastly, the resource allocation strategies can be improved further by
spreading out the budget systematically among different generations given the cost of making a
cross and genotyping a progeny under deadline constraints.

ACKNOWLEDGMENTS

This work is partially supported by project 1011702 from the USDA National Institute of Food
and Agriculture, project 1929681 from National Science Foundation and Plant Sciences Institute
at Iowa State University. The authors are also grateful to William Beavis and Vishnu
Ramasubramanian at Iowa State University for assistance in the collection of the soybean data set
used in our simulations.
REFERENCES


FIGURES AND TABLES

Figure 1. Illustration of genomic information and marker effects for four individuals

<table>
<thead>
<tr>
<th>Line 1</th>
<th>Line 2</th>
<th>Line 3</th>
<th>Line 4</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 2. Illustration of complementary score calculation for Line 2, and Line 4

\[
\text{Score(Line2, Line4)} = 10.8 + 16.2 + 0 = 27
\]

Figure 3. The Simulation process used for comparing the four methods (CGS, OHV, OPV, and CBS)
Figure 4. Cumulative distribution functions of the Genomic Estimated Breeding Value (GEBV) population maximum for corn dataset
Figure 5. Cumulative distribution functions of the Genomic Estimated Breeding Value (GEBV) population maximum for soybean dataset
Figure 6. Cumulative distribution functions of the Genomic Estimated Breeding Value (GEBV) population mean for corn dataset
Figure 7. Cumulative distribution functions of the Genomic Estimated Breeding Value (GEBV) population mean for soybean dataset.
Figure 8. Genetic gain per generation of the mean Genomic Estimated Breeding Value (GEBV) for corn dataset
Figure 9. Genetic gain per generation of the mean Genomic Estimated Breeding Value (GEBV) for soybean dataset.

Table 1. The resource allocation methods strategy for CGS, OHV, and OPV

<table>
<thead>
<tr>
<th>Generation number</th>
<th>Budget allocated</th>
<th>Number of crosses</th>
<th>Number of progenies produced per cross</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$2050</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>$2050</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>$2050</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>$2050</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>$2050</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Generation number</td>
<td>Budget allocated</td>
<td>Number of crosses</td>
<td>Number of progenies produced per cross</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------</td>
<td>------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>$3895</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>$3485</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>$3075</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>$2665</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>$2255</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>$1845</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>$1435</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>$1025</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>$615</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>$205</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

Total Budget=$20500

Table 2. The resource allocation strategy for CBS method