Graph-based data selection for the construction of genomic prediction models

Steven Maenhout*,1, Bernard De Baets§, Geert Haesaert*

*Department of Biosciences and Landscape Architecture, University College Ghent, B-9000 Gent, Belgium,
§Department of Applied Mathematics, Biometrics and Process Control, Ghent University, B-9000 Gent, Belgium
Data selection for genomic prediction

Keywords: genomic prediction, $\varepsilon$-SVR, BLP, graph theory

Corresponding author:
Steven Maenhout
Department of Biosciences and Landscape Architecture
University College Ghent
Voskenslaan 270
B-9000 Gent
Belgium
Phone +32 9 248 88 60
Fax +32 9 242 42 79
Email: Steven.Maenhout@hogent.be
Abstract

Efficient genomic selection in animals or crops requires the accurate prediction of the agronomic performance of individuals from their high-density molecular marker profiles. Using a training data set that contains the genotypic and phenotypic information of a large number of individuals, each marker or marker allele is associated with an estimated effect on the trait under study. These estimated marker effects are subsequently used for making predictions on individuals for which no phenotypic records are available. As most plant and animal breeding programs are currently still phenotype-driven, the continuously expanding collection of phenotypic records can only be used to construct a genomic prediction model if a dense molecular marker fingerprint is available for each phenotyped individual. However, as the genotyping budget is generally limited, the genomic prediction model can only be constructed using a subset of the tested individuals and possibly a genome-covering subset of the molecular markers. In this paper, we demonstrate how an optimal selection of individuals can be made with respect to the quality of their available phenotypic data. We also demonstrate how the total number of molecular markers can be reduced while a maximum genome coverage is ensured. The third selection problem we tackle is specific to the construction of a genomic prediction model for a hybrid breeding program where only molecular marker fingerprints of the homozygous parents are available. We show how to identify the set of parental inbred lines of a predefined size which has produced the highest number of progeny. These three selection approaches are put into practice in a simulation study where we demonstrate how the trade-off between sample size and sample quality affects the prediction accuracy of genomic prediction models for hybrid maize.
Despite the numerous studies devoted to molecular marker-based breeding, the genetic progress of most complex traits in today’s plant and animal breeding programs still heavily relies on phenotypic selection. Most breeding companies have established dedicated databases that store the vast number of phenotypic records that are being routinely collected throughout the course of their breeding programs. These phenotypic records are, however, gradually being complemented by various types of molecular marker scores and it is to be expected that effective marker-based selection schemes will eventually allow to reduce current phenotyping efforts (Bernardo 2008; Hayes et al. 2009). The available marker and phenotypic databases already allow for the construction and validation of marker-based selection schemes. Mining the phenotypic databases of a breeding company is, however, quite different from analyzing the data that is generated by a carefully designed experiment. Genetic evaluation data is often severely unbalanced as elite individuals are usually tested many times on their way to becoming a commercial variety or sire, while less performing individuals are often disregarded after a single trial. Furthermore, the different phenotypic evaluation trials are separated in time and space and as such, subjected to different environmental conditions. Therefore, ranking the performance of individuals that were evaluated in different phenotypic trials is usually a non-trivial task.

Animal breeders are well experienced when it comes to handling unbalanced genetic evaluation data. The best linear unbiased predictor or BLUP approach (Henderson, C. R. 1975) presented a major breakthrough in this respect, especially when combined with restricted maximum likelihood or REML estimation of the needed variance components (Patterson and Thompson 1971). Somewhat later on, this linear mixed modeling approach was also adopted by plant breeders as the de facto standard for handling unbalanced phenotypic data. The more recent developments in genomic selection (Bernardo 1995; Meuwissen et al. 2001; Gianola and van Kaam 2008) and marker-trait association studies (Yu et al. 2006) are, at least partially, BLUP-based and are therefore, in theory, perfectly suited for mining the large marker and phenotypic databases that back each breeding program. In practice, however, the unbalancedness of the available genetic
evaluation data often reduces its total information content and the construction of a marker-based selection model is limited to a more balanced subset of the data.

As phenotypic data is available, genotyping costs limit the total number of individuals that can be included in the construction of a genomic prediction model. The best results will be obtained by selecting a subset of individuals for which the phenotypic evaluation data exhibits the least amount of unbalancedness. In this paper we demonstrate how this phenotypic subset selection problem can be translated into a standard graph theory problem which can be solved with exact algorithms or less time consuming heuristics.

In most plant and animal species, the number of available molecular markers is rapidly increasing, while the genotyping cost per marker is decreasing. Nevertheless, as budgets are always limited, genotyping all mapped markers for a small number of individuals might be less efficient than genotyping a restricted set of well-chosen markers on a wider set of individuals. One should therefore be able to select a subset of molecular markers that covers the entire genome as uniformly as possible. We demonstrate how this marker selection problem can also be translated into a well-known graph theory problem which has an exact solution.

The third problem we tackle by means of graph theory is more specific to hybrid breeding programs where the parental individuals are nearly or completely homozygous. This implies that we can deduce the molecular marker fingerprint of a hybrid individual from the marker scores of its parents. As the phenotypic data is collected on the hybrids, genotyping costs can be reduced by selecting a subset of parental inbreds that have produced the maximum number of genetically distinct offspring amongst themselves. Obviously, the phenotypic data on these offspring should be as balanced as possible.

Besides solving the above-mentioned selection problems by means of graph theory algorithms, we demonstrate their use in a simulation study that allows to determine the optimum trade-off between the number of individuals and the size of the genotyped molecular marker fingerprint for predicting the phenotypic performance of hybrid maize by means of $\varepsilon$-insensitive support vector machine regression ($\varepsilon$-SVR) (MAENHOUT et al. 2007, 2008, 2010) and Best Linear Prediction
SELECTING INDIVIDUALS FROM UNBALANCED PHENOTYPIC EVALUATION DATA

In most plant or animal breeding programs, all phenotypic measurements that were recorded during genetic evaluation trials are stored for future reference. The all-encompassing experimental design that generated this data is likely to be very unbalanced. The most extreme case of an unbalanced design is a disconnected design. Table 1 gives an example of a disconnected sire evaluation design taken from Kennedy and Rus (1993). The breeding values of four sires are evaluated by measuring the performance of their offspring in three different herds. Sires having offspring in different herds provide vertical connections between herds while herds containing offspring of different sires provide horizontal connections. In a perfectly balanced design, each sire would have the same number of offspring tested in each herd. In the presented scenario however, sires $s_1$ and $s_2$ are disconnected from sires $s_3$ and $s_4$ as there is no possible path between these groups. This means that if we analyze the phenotypic data from this design with an ordinary least squares model, contrasts involving sires that belong to the disconnected groups would be inestimable. However, if we fit a linear mixed model to this data in which we assume herds as fixed and sires as random effects, contrasts involving sire BLUPs belonging to these disconnected groups are perfectly estimable.

Ignoring connectivity issues by treating sire effects as random variables is, however, not without consequence. This approach implicitly assumes that all evaluated individuals originate from the same population and as such have the same breeding value expectation. This assumption is generally not valid in animal breeding programs as the better sires are usually evaluated in the better herds (Foulley et al. 1990). A similar stratification can be observed in genetic evaluation trials performed by plant breeders where late and therefore higher yielding individuals are generally tested in geographical regions with longer growing seasons. As a consequence, BLUP-based genomic selection routines will be less efficient, while marker-trait association studies will
Table 1

Example of a disconnected sire × herd design taken from Kennedy and Trus (1993). The cell numbers indicate how many offspring of the sire pertaining to that particular column were evaluated in the herd pertaining to that particular row.

<table>
<thead>
<tr>
<th></th>
<th>s₁</th>
<th>s₂</th>
<th>s₃</th>
<th>s₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>h₁</td>
<td>3</td>
<td>_</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>h₂</td>
<td>3</td>
<td>_</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>_</td>
<td>_</td>
<td>_</td>
<td>_</td>
</tr>
<tr>
<td>h₃</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>_</td>
</tr>
</tbody>
</table>

Suffer from increased false positive rates and reduced power. A very unbalanced but nevertheless connected design will also reduce the effectiveness of marker-based selection approaches as the prediction error variance of the estimated breeding values increases substantially. Furthermore, the estimated breeding values will be regressed towards the mean and will not account for the true genetic trend.

We can assume that the available data set contains unbalanced phenotypic measurements on t individuals, where t is generally a very large number. The available phenotypic data allows the breeder to try out one or more of the more recent BLUP-based genomic selection approaches without setting up dedicated trials. Given his financial limit for genotyping, he wants to select exactly p individuals from this data set. The selection of p individuals should be optimal in the sense that the precision of the BLUPs of the p breeding values that are obtained from a linear mixed model analysis of the full set of phenotypic records, is superior to the precision of any other set of BLUPs with cardinality p. This optimality criterion requires a measure of precision of a subset of BLUPs obtained from a linear mixed model analysis. To introduce this criterion, we will make the general assumption that the applied linear mixed model takes the form

\[ y = X\beta + Zu + e, \]

where y is a column vector containing n phenotypic measurements on the t individuals. \( \beta \) is a
vector of fixed nuisance effects like trial, herd and replication effects and $u$ is vector containing random genetic effects for each of the $t$ individuals. For ease of explanation, we will assume that $u$ only contains $t$ breeding values, but the presented approach can easily be generalized to cases where $u$ is made up from general (GCA) and specific combining abilities (SCA) and possibly the different levels of various genotype-by-environment (G×E) interaction factors. Vector $e$ contains $n$ random residuals. Matrices $X$ and $Z$ link the appropriate phenotypic records to the effects in $\beta$ and $u$ respectively. Furthermore we assume that we can represent the variance of $u$ and $e$ as

$$\text{Var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}.$$ 

$G$ can contain an assumed covariance structure for the $t$ individuals, typically a scaled numerator relationship matrix calculated from available pedigree or marker data. It is, however, important to realize that fitting a covariance between breeding values allows the BLUPs from individuals which have little phenotypic information themselves, to borrow strength from phenotypic records on closely related individuals. As a result, the $p$ individuals with highest BLUP precision will most likely be close relatives which is detrimental for the generalizing capabilities of the marker-based selection model. If we want the selection of $p$ individuals to rely completely on the amount of information and the structure (balancedness) of their phenotypic records, $G$ should be a scaled identity matrix. Once the $p$ individuals have been selected, a pedigree or marker-based covariance structure can be incorporated in $G$ for the construction of the actual marker-based prediction model. The covariance structure of the residuals in matrix $R$ can contain heterogeneous variances for the different production environments, or in case that data originates from actual field trials, spatial information like row or column correlations. The BLUPs in vector $u$ are obtained by solving the mixed model equations (HENDerson, C.R. 1984)

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}.$$
The inverse of the coefficient matrix allows to obtain the prediction error variance (PEV) matrix of vector $\hat{u}$ as

$$
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}Z \\
Z'R^{-1}X & Z'R^{-1}Z + G^{-1}
\end{bmatrix}^{-1} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},
$$

where

$$
\text{PEV}(\hat{u}) = \text{Var}(\hat{u} - u) = C_{22}.
$$

A logical choice of measure to express the precision of a selection of $p$ BLUPs from the $t$ candidates in vector $\hat{u}$ would be some function of the $p \times p$ principal submatrix $C_{22}^p$, obtained by removing the rows and columns of $C_{22}$ that pertain to individuals that are not in that particular selection.

As a good design is strongly associated with the precision of pairwise contrasts (Bueno Filho and S.G. Gilmour 2003), we use the lowest precision of all possible pairwise contrast vectors between the $p$ selected individuals as optimization criterion. A pairwise contrast vector $q_{ij}$ for the individuals $i$ and $j$ is a vector where $q_{ij} = 1$ and $q_{ji} = -1$, while all other elements of $q_{ij}$ are zeros. Laloë (1993) and Laloë et al. (1996) propose to express the precision of a linear contrast vector $q$ by means of the generalized Coefficient of Determination which is defined as

$$
\text{CD}(q) = \frac{q' (G - C_{22}) q}{q' G q},
$$

where CD$(q)$ always lies within the unit interval. They indicate that CD$(q)$ can be obtained as a weighted average of the $t - 1$ non-zero eigenvalues $\mu_i$ of the generalized eigenvalue problem

$$
((G - C_{22}) - \mu_i G) v_i = 0,
$$

as

$$
\text{CD}(q) = \frac{\sum_{i=2}^{t} a_i^2 \mu_i}{\sum_{i=2}^{t} a_i^2},
$$

where $a_i^2$ is the weight for eigenvalue $i$ and the first eigenvalue $\mu_1$ always equals zero as a conse-
quence of the well-known summation constraint $1'G^{-1}1$ (see e.g. FOULLEY et al. (1990)). Each linear contrast vector $q$ can be expressed as a linear combination of the $t-1$ non-zero eigenvectors $v_i$ as

$$q = \sum_{i=2}^{t} a_i v_i.$$ 

In fact, all linear contrast vectors that are estimable in the least squares sense, are linear combinations of the eigenvectors $v_i$ of Eq. (2) that are associated to non-zero eigenvalues $\mu_i$, while those contrasts that are not estimable in a least squares sense, are linear combinations of eigenvectors for which at least one associated eigenvalue is zero. This implies that the CD of a pairwise contrast vector involving two individuals that were evaluated in two disconnected groups does not necessarily become zero as several eigenvalues $\mu_i$ in Eq. (3) might be non-zero. This might bias the selection procedure to favor a disconnected set of individuals with a high information content (i.e. a high level of replication) instead of a connected set of individuals with low information content. To avoid this situation, the CD of pairwise contrast vectors between disconnected individuals should be forced to zero. In case Eq. (1) represents a simplified animal model where $G = I \sigma^2_g$ and $R = I \sigma^2_e$, disconnected pairs of individuals can be easily identified by examining the block diagonal structure of the PEV matrix $C_{22}$ as explained in Appendix A1. In Appendix A2 we show how disconnected pairs of individuals can be identified by means of the transitive closure of the adjacency matrix of the $t$ individuals.

Now that we have the corrected CD for each of the $\frac{p(p-1)}{2}$ pairwise contrast vectors, we can represent the $t$ individuals as vertices (also called nodes) of a weighted complete graph where the edge between individual $i$ and individual $j$ carries the weight $CD(q_{ij})$, expressing the precision of the pairwise contrast as a number between zero and one. We need to select exactly $p$ vertices such that the minimum edge weight in the selected subgraph is maximized. This problem is equivalent to the ‘discrete $p$-dispersion problem’ from the field of graph theory. This problem setting is encountered when locating facilities that should not be clustered in one location, like nuclear plants or franchises belonging to the same fast-food chain. This problem is NP-hard even when the distance matrix satisfies the triangle inequality. ERKUT (1990) describes two exact algorithms based on a
A branch and bound strategy and compares 10 different heuristics (ERKUT et al. 1994). An interesting solution lies in the connection between the discrete $p$-dispersion problem and the maximum clique problem. A clique in a graph is a set of pairwise adjacent vertices, or in other words, a complete subgraph. The corresponding optimization problem, the maximum clique problem, is to find the largest clique in a graph. This problem is also NP-hard (CARRAGHAN and P.M. PARDALOS 1990). The idea is to decompose the discrete $p$-dispersion problem in a number of maximum clique problems by assigning different values to the minimum required contrast precision $CD_{\text{min}}$. Initially, $CD_{\text{min}}$ is low (e.g. $CD_{\text{min}} = 0.1$) and we define a graph $G'(V, E')$ where the edges of the original graph $G$ are removed when their edge weight is smaller than $CD_{\text{min}}$. This implies that there will be no edges between disconnected pairs of individuals in the derived graph as these edge weights have been set to zero by the CD correction procedure. Solving the maximum clique problem in $G'(V, E')$ allows to identify a complete subgraph for which all edge weights are guaranteed to be greater than $CD_{\text{min}}$. The number of vertices in this complete subgraph is generally smaller than $t$ but greater than $p$. By repeating this procedure with increasing values of $CD_{\text{min}}$ one can make a trade-off between sample size and sample quality as is demonstrated in Figure 1 for a representative sample of size $t = 4236$ individuals for which genetic evaluation data was recorded as part of the grain maize breeding program of the private company RAGT R2n. Each dot represents the largest possible selection of individuals where $CD_{\text{min}}$ ranges from 0 to 0.97. The data used in this example is connected as there is no sudden drop in the number of individuals when $CD_{\text{min}}$ is raised from 0.0 to 0.1. In general, the surface below the curve represents a measure of data quality. If one is only interested in obtaining the optimal selection of exactly $p$ individuals from a set of $t$ candidates, one can implement the described maximum clique-based procedure in a binary search.

The presented approach for solving the discrete $p$-dispersion problem requires an efficient algorithm to obtain the maximum clique from a graph. Several exact algorithms and heuristics have been published, but comparing these is often difficult as the dimensions and densities of the provided example graphs as well as computational platforms tend to differ between papers. The exact algorithm of CARRAGHAN and P.M. PARDALOS (1990) is, however, considered as the basis for
FIGURE 1.—Graphical representation of the trade-off between the selection size and the selection quality for a sample of the RAGT grain maize breeding pool. For each examined level of $CD_{\text{min}}$, ranging from 0.0 to 0.97, the dot represents the maximum cardinality selection of individuals for which the minimum precision of a pairwise contrast is at least $CD_{\text{min}}$. 
most later algorithms. Although the efficiency of this algorithm has been superseded by that of more recent developments (ÖSTERGÄRD, P.R.J. 2002; TOMITA and SEKI 2003), its easy implementation often makes it the method of choice. If the available run-time is limited, a time-constrained heuristic like the Reactive Local Search approach presented by BATTITI and PROTASI (2001) might be more appropriate. BOMZE et al. (1999) give an overview of several other heuristic approaches found in literature, in particular greedy construction and stochastic local search including simulated annealing, genetic algorithms and tabu search.

SELECTING MARKERS FROM A DENSE MOLECULAR MARKER FINGERPRINT

The construction of a genomic prediction model requires genotypic information on each of the \( p \) selected individuals. Generally it is assumed that a good prediction accuracy can only be achieved by maximizing the genome coverage, which implies genotyping a large number of molecular markers. This approach seems particularly attractive as genotyping costs are decreasing rapidly. However, as shown by MAENHOUT et al. (2010), the relation between the number of genotyped markers and the obtained prediction accuracy seems to be subject to the law of diminishing marginal returns. This means that it might be more efficient to construct the genomic prediction model using a larger number of individuals in combination with a smaller molecular marker fingerprint. There is obviously an upper limit to the sparsity of the applied molecular marker fingerprint and its genome coverage should be as uniform as possible such that the probability of detecting a marker-trait association is maximized.

We start by solving this selection problem on a single chromosome for which \( t \) candidate molecular markers have been mapped. We want to select exactly \( q \) of these markers such that the chromosome coverage is optimal compared to all other possible selections of \( q \) markers. Maximizing the chromosome coverage could mean several things, including maximizing the average intermarker distance and maximizing the minimum marker distance. We prefer the latter definition as it implies
a one-dimensional version of the discrete $p$-dispersion problem. In this restricted setting, a reduc-
tion to a series of maximum clique problems is not necessary as RAVI et al. (1991) have published
an algorithm that obtains the optimal solution in an overall running time of $O(\min(t^2, qt \log(t)))$.

The extension to $c > 1$ chromosomes is again dependent on the interpretation of uniform
genome coverage. For example, we can use the above-mentioned algorithm to select $\frac{q\ell_i}{\sum_{i=1}^c \ell_i}$ markers
on each chromosome $i$ with length $\ell_i$. As these fractions will generally not result in integers, the
remainder after division could be attributed to each of the different chromosomes in decreasing
order of their minimum intermarker distance after the addition of one marker. A more intuitive
interpretation of a uniform genome coverage entails a selection of markers such that the minimum
intermarker distance over all chromosomes is maximized. This can be achieved by linking all
chromosomes head to tail as if all markers were located on a single chromosome. To be able to
use the above-mentioned algorithm, the distance between the last marker of the first chromosome
and the first marker of the second chromosome of each linked chromosome pair should be set to
infinity.

SELECTING PARENTAL INBRED LINES

In hybrid breeding programs, the molecular marker fingerprint of a single-cross hybrid can be
easily deduced from the fingerprints of its two homozygous parents. This allows to reduce the
total genotyping cost of the genomic prediction model considerably. If we assume we have a bud-
get for fingerprinting exactly $k$ parental inbred lines, we can maximize the number of genotyped
single-cross hybrids by selecting the set of lines which have produced the maximum number of
single-cross hybrids amongst themselves. We approach this selection problem by representing the
total set of parental inbred lines as the vertices of an unweighted pedigree graph where an edge be-
tween two vertices represents an offspring individual (i.e. a single-cross hybrid) for which genetic
evaluation data is available. Figure 2 shows such a graph representation of the sample used in Fig-
ure 1 containing 487 inbred lines and 4236 hybrids. We need to select a $k$-vertex subgraph which
has the maximum number of edges. In graph theory parlance, this problem is called the ‘dens-
FIGURE 2.—Graph representation of a sample of the RAGT grain maize breeding pool. The blue vertices represent inbred lines and the grey edges are single-cross hybrids.

The presented selection procedure does not consider the quality of the genetic evaluation data that is available for the hybrids. As a result, the optimal selection with respect to the maximization of the number of training examples might turn out to be a very poor selection with respect to the quality of the phenotypic data. To enforce these data quality constraints, the described inbred line selection procedure should be performed after a preselection of the hybrids based on the precision of pairwise contrasts. If we select \( k \) inbred lines where \( k \) ranges from the total number of candidate parents to 3 for each level of \( CD_{\text{min}} \) ranging from 0.1 to 0.97 we get a 3-dimensional representation of the data quality as shown in Figure 3. Similarly to Figure 1, each dot on the surface represents the size of the optimal selection of hybrids under the constraints of a genotyping budget for \( k \) parental
FIGURE 3.—Graphical representation of the trade-off between the selection size and the selection quality when only $k$ parental inbred lines are being genotyped. For each examined level of $C_{D_{\text{min}}}$ ranging from 0.0 to 0.97 the number of genotyped inbred lines $k$ is reduced from 487 to 3. Each dot in the plotted surface represents the maximum cardinality selection of hybrid individuals for which the minimum precision of a pairwise contrast is at least $C_{D_{\text{min}}}$ and the number of parents is exactly $k$.

inbred lines and a minimum pairwise contrast precision of $C_{D_{\text{min}}}$. We can see that for high levels of $C_{D_{\text{min}}}$ and high levels of $k$, the cardinality of the resulting selection becomes 0, indicating that there are no hybrids that comply with both constraints. As soon as the constraint on $C_{D_{\text{min}}}$ is relaxed, the selection cardinality increases gradually as more parental inbred lines are being genotyped.

**SIMULATION STUDY**

The construction of a hybrid maize prediction model based on $\varepsilon$-insensitive support vector machine regression ($\varepsilon$-SVR) (Maenhout et al. 2008, 2010) or Best Linear Prediction (BLP) (Bernardo 1995) requires a combination of genotypic and phenotypic data on a predefined number of inbred lines and their hybrid offspring respectively. As phenotypic data is available from past genetic evaluation trials, the number of training examples that is used for the construction of this prediction model is constrained by the total genotyping cost. If we reduce the size of the fingerprint, more
inbred lines can be genotyped and more training examples become available which should result in a better prediction accuracy of the model. However, reducing the size of the molecular marker fingerprint comes at the price of a reduced genome coverage and an increased number of selected hybrids results in a reduced precision of BLUP contrasts due to connectivity issues (e.g. Figure 1). Therefore, it is to be expected that within the constraints of a fixed genotyping budget, maximum prediction accuracy can be achieved by finding the optimal balance between the fingerprint size and the number of training examples. The location of this optimum is obviously highly dependent on the information content of the available phenotypic data and the applied linkage map, but can be estimated by means of the afore mentioned graph theory algorithms for each specific data set.

**Simulation setup:** To demonstrate the approach, we use the phenotypic data that was generated as part of the grain maize breeding program of the private breeding company RAGT R2n and their proprietary SSR linkage map. We assume to have a limited budget for genotyping 101 SSR markers on 200 inbred lines or 20200 markers in total. We will also assume that we can limit the number of candidate inbred lines to 400 by restricting the prediction model to a specific heterotic group combination, a specific environment (i.e. maturity rating, irrigation and fertilizer treatments, . . . ) and the set of inbred lines that are available at the moment of genotyping. This intensive preselection of candidate lines is mainly needed for keeping the simulations tractable. In a more realistic setting, calculations are only performed once so the set of initial candidate lines can be larger. Table 2 gives a schematic overview of the different steps that are performed at each iteration of the simulation routine.

Again we make use of the pedigree graph representation where inbred lines are represented as vertices and each single-cross hybrid is represented as an edge between two vertices as shown in Figure 2. In this graph, the degree of a vertex (i.e. the number of edges incident to the vertex) therefore equals the number of distinct single-cross hybrids of which the inbred line is a parent. Figure 4 shows the empirical distribution of these degrees on a log scale for the entire RAGT grain maize breeding pool. The observed long-tailed behavior of the empirical distribution is not unexpected as most inbred lines only have a limited number of children, while inbred lines with higher progeny
TABLE 2
Description of each step that is performed during a single iteration of the simulation routine. The goal is to find the optimal trade-off between the number of genotyped inbred lines and the size of their molecular fingerprint, when the total genotyping budget is fixed.

<table>
<thead>
<tr>
<th>step</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sample 400 vertices from the pedigree graph by means of the ‘Forest Fire’ algorithm: ⇒ indirect sampling of hybrids ⇒ indirect sampling of multi-environment trials</td>
</tr>
<tr>
<td>2</td>
<td>partition sampled inbred lines in $c$ heterotic groups by means of the Dsatur vertex coloring algorithm</td>
</tr>
<tr>
<td>3</td>
<td>simulate 8 breeding cycles on each of the $c$ heterotic groups</td>
</tr>
<tr>
<td>4</td>
<td>simulate phenotypic records on the sampled hybrids</td>
</tr>
<tr>
<td>5</td>
<td>reduce the number of sampled hybrids by gradually increasing $CD_{min}$</td>
</tr>
<tr>
<td>6</td>
<td>reduce the number of genotyped inbred lines by means of the greedy densest $k$-subgraph algorithm</td>
</tr>
<tr>
<td>7</td>
<td>select $q$ SSR markers with maximal genome coverage</td>
</tr>
<tr>
<td>8</td>
<td>determine the prediction accuracy of $\varepsilon$-SVR and BLP using the reduced set of training examples</td>
</tr>
</tbody>
</table>

numbers (i.e. the tester lines) are rare. In an attempt to parametrize the underlying distribution from which the observed vertex degrees were drawn, several candidate distributions among which the Poisson, geometric, discrete log-normal and discrete power-law distributions were fitted by means of likelihood maximization. The best fit was observed for the discrete power-law distribution with a left threshold value of 6 which is indicated as a straight line on Figure 4. The fit of this distribution is, however, insufficient as indicated by the significantly large Kolmogorov-Smirnov $D$-statistic where significance is determined by means of the parametric bootstrap procedure described by CLAUSET et al. (2009).

As no conclusive evidence on the underlying distribution of the observed vertex degrees was found, we prefer to sample inbred lines from the full RAGT graph directly. However, taking a representative sample from a large graph is not a trivial task. The sample quality of various published graph sampling algorithms seems to be highly dependent on the properties of the graph. To decide which sampling routine is optimal for the RAGT data, we first need to decide on a measure of...
FIGURE 4.—Log-scaled degree distribution of the graph created from part of the RAGT R2n grain maize breeding program. In this undirected, unweighted graph, parental inbred lines are represented as vertices and single-cross hybrids as edges. Each dot represents a unique log-scaled vertex degree (horizontal axis) and the log of its frequency in the graph (vertical axis). The red line represents the fitted power law distribution by means of likelihood maximization. The threshold value of 6 was determined by minimizing the Kolmogorov-Smirnov statistic as described by Clauset et al. (2009).
sample quality. We compare the empirical cumulative distribution (ECD) of the vertex degrees in
the full graph with those ECDs of 100 samples containing 400 vertices. From these ECDs, we
calculate the average Kolmogorov-Smirnov $D$-statistic for each examined sampling routine. For
the RAGT data, the ‘Forest Fire’ vertex sampling approach resulted in the smallest average $D$-
statistic compared to the alternative methods described by Lescovec and Faloutsos (2006).
This sampling routine starts by selecting a vertex $v_0$ uniformly at random from the graph. Vertex
$v_0$ now spreads ‘the fire’ to a random selection of its neighbors which are then in turn allowed
to infect a random selection of their own neighbors. This process is continued until exactly 400
vertices are selected. If the fire dies out before the sample is complete, a new starting vertex is
selected uniformly at random. The number of neighbors that is infected at each selected vertex, is
obtained as a random draw from a geometric distribution where the parameter $p$ was set to 0.62,
as this value resulted in the best average sample quality. All hybrids for which both parents were
sampled (i.e. the edges of the subgraph) have associated phenotypic records and as such indirectly
sample a set of multi-environment trials (METs). All hybrids that were not indirectly selected by
the inbred line sample, but do have phenotypic records in the sample of METs, are included in the
selection as data connecting check varieties. Despite the fact that the RAGT data already provides
phenotypic records for the selected hybrids and check varieties, we replace these by simulated
measurements as we want to be able to assess the actual prediction accuracy of $\varepsilon$-SVR and BLP
under various levels of data quality.

The simulation of these phenotypic records for the sampled hybrids starts by partitioning the
selected inbred lines into heterotic groups. This partitioning should ensure that the two parents
of each single-cross hybrid always belong to distinct heterotic groups, while the total number of
groups needs to be minimized. The graph theory equivalent of this problem is called the ‘vertex
coloring problem’ which, as all previously described graph theory problems, belongs to the com-
plexity class of NP-hard problems. The minimum number of colors (i.e. heterotic groups) is called
the chromatic number of the graph. The vertex coloring problem has been extensively studied in
graph theory literature (Jensen and Toft 1995) and several efficient heuristics are available. The
greedy desaturation algorithm or Dsatur published by Brézal (1979) is often used as a benchmark method to assess the efficiency and precision of newly developed vertex coloring algorithms. Its good performance on a variety of graphs and easy implementation makes it the method of choice for designating inbred lines to heterotic groups at each iteration of the simulation routine.

Once the chromatic number $c$ has been determined for the sampled set of inbred lines, an entire breeding program is simulated starting from $c$ open-pollinated varieties and resulting in $c$ unrelated heterotic groups. The simulation of this breeding program mimics the maize breeding program of the university of Hohenheim as described by Stich et al. (2007) and modified by Maenhout et al. (2009). In short, the simulation routine uses the proprietary linkage map of the breeding company RAGT R2n containing 101 microsatellites and adds an additional 303 evenly distributed, simulated SSRs. It also generates 250 QTL loci of the selection trait (e.g. yield) which are randomly positioned on the genetic map. The number of alleles for each SSR or QTL is drawn from a Poisson distribution with an expected value of 7. Each simulation starts by generating an initial base population in Hardy-Weinberg equilibrium. Allele frequencies for each locus are drawn from a Dirichlet distribution and used to calculate the allele frequencies in each of the $c$ subpopulations assuming an $F_{st}$ value of 0.14. We perform 8 breeding cycles where each cycle consists of 6 generations of inbreeding and subsequent phenotypic selection based on line per se or testcross performance as described by Stich et al. (2007). The result is a set of 400 highly selected inbred lines partitioned in $c$ unrelated heterotic groups. Within each of these groups, the simulated inbred lines are randomly assigned to the sampled inbred lines and a genotypic value is generated for each interheterotic hybrid by summing the effects of the 250 QTL alleles of both parents and adding a normally distributed SCA value. The size of the SCA variance component depends on the heritability of the trait under consideration, but is assumed to be only $\frac{1}{8}$ of the total non-additive variance (SCA + G×E and residual error) as this was the average of observed ratios for the traits grain yield, grain moisture contents and days until flowering in the actual RAGT data. The genotypic values of the check varieties are generated from a single normal distribution where the variance is the sum of the additive variance and SCA variance of the sampled hybrids. The
simulated genotypic values of hybrids and check varieties are used to generate phenotypic records according to the sampled MET data structure, assuming a single replication in each location of a MET. This implies that G×E effects are confounded with the residual error and only a single effect is drawn from a normal distribution where the variance is \( \frac{7}{8} \) of the total non-additive variance. The main environmental effect of each location is also drawn from a normal distribution for which the variance is twice the additive variance of the hybrids.

The simulated phenotypic records that are associated with the sampled data structure allow to estimate the genotypic value of each hybrid by means of a linear mixed model analysis. We fit individuals (hybrids and check varieties) as random and locations as fixed effects as this approach should result in an \( \varepsilon \)-SVR model with a superior prediction accuracy (Maenhout et al. 2010). To avoid selections of closely related hybrids, the variance-covariance matrix of the genotypic effects is fitted as a scaled identity matrix. The resulting PEV matrix of the random genotypic effects is used to iteratively select a smaller subset of the sampled hybrids by gradually increasing the minimum required precision of each pairwise contrast in the selection. Initially, the required \( CD_{\text{min}} \) value is set to 0 which implies that all hybrids are selected. The next examined level of precision requires \( CD_{\text{min}} > 0 \) which effectively excludes selections containing disconnected individuals. More stringent levels of precision are enforced by requiring \( CD_{\text{min}} > q_p \) where \( q_p \) is the \( p \)th quantile of the observed distribution of CD values in the complete sample and \( p \) ranges from 0 to 0.875 in steps of 0.125. Defining \( CD_{\text{min}} \) values as quantiles allows to compare the obtained prediction accuracies over the different samples of the simulation routine.

For each level of \( CD_{\text{min}} \), the number of genotyped inbred lines is reduced from 400 to 50 in steps of 50, while at the same time the number of markers in the molecular marker fingerprint is increased from 50 to 404. For each combination of \( CD_{\text{min}} \) and number of genotyped inbred lines, the BLUPs of the selected hybrids are used to construct an \( \varepsilon \)-SVR and a BLP-based prediction model. In fact, the prediction accuracy of both methods is verified by randomly assigning the BLUPs to one of five groups. For each of these groups, a separate \( \varepsilon \)-SVR and BLP prediction model is constructed using all BLUPs in the remaining four groups as training data. The resulting prediction model is then
FIGURE 5.—Accuracy of the genotypic value BLUPs of the hybrids selected using the described graph-based procedures. The three examined heritability levels $h^2 = 0.25$, $h^2 = 0.5$ and $h^2 = 0.75$ are represented by the bottom, middle and top wireframe surfaces respectively. Each point on a surface is the squared Pearson correlation between the BLUPs and the actual (simulated) genotypic values of the selected hybrids under the constraints of a minimum required contrast precision $CD_{\text{min}}$, expressed as a percentile of the sampled $CD$ values, and the number of genotyped inbred lines, averaged over 100 iterations of the simulation routine.

used to make predictions on the hybrids in the selected group (i.e. the validation data). Combining the predictions of all five models allows to obtain a measure of prediction accuracy by correlating them against the simulated genotypic values.

**Simulation results:** We expect that enforcing a minimum required pairwise contrast precision $CD_{\text{min}} > 0$, results in a selection of BLUPs that has greater accuracy compared to the full set of hybrids. In Figure 5 this BLUP accuracy is plotted against $CD_{\text{min}}$ and the maximum number of inbred lines for each of the three examined heritability levels. Each point on these wireframe
surfaces represents the squared Pearson correlation between the BLUPs and the actual, simulated genotypic values of the selected hybrids at that particular level of \( CD_{\text{min}} \) and number of parental inbred lines, averaged over 100 iterations of the simulation routine. We can see that an increase in \( CD_{\text{min}} \) results in an almost linear increase in BLUP precision for each heritability level. This effect is especially pronounced for the lowest heritability level \( h^2 = 0.25 \). As expected, the BLUP precision is not influenced by the number of parental inbred lines.

Figure 6 presents the prediction accuracy of both \( \varepsilon \)-SVR and BLP for increasing values of the minimum required contrast precision \( CD_{\text{min}} \) and a decreasing number of genotyped inbred lines. The height of each point in the wireframes represents the average prediction accuracy, expressed as a squared Pearson correlation, over 100 iterations of the simulation routine. For each of the examined heritability levels, \( \varepsilon \)-SVR generally performs better than BLP. The negative effect of disconnected hybrids in the selection of training examples is visualized as the sharp increase in prediction accuracy when the minimum required contrast precision is slightly constrained from \( CD_{\text{min}} = 0 \) to \( CD_{\text{min}} > 0 \). This effect is more pronounced for BLP than for \( \varepsilon \)-SVR. Increasing \( CD_{\text{min}} \) any further, generally decreases the prediction accuracy, especially for traits with lower heritability. This observation implies that, at least for the RAGT data set, a larger number of training examples of lower data quality is to be preferred over a smaller selection of hybrids for which more and better connected phenotypic information is available, as long as disconnected individuals are excluded.

BLP and \( \varepsilon \)-SVR do not take an unanimous stand on the optimal number of genotyped inbred lines. For BLP, the optimum seems to lie somewhere around 100 inbred lines for \( h^2 = 0.25 \) and 150 for \( h^2 = 0.50 \) and \( h^2 = 0.75 \), the equivalent of fingerprint sizes of 202 and 134 SSR markers respectively. This optimum is, however, less pronounced for the higher heritability levels. For \( \varepsilon \)-SVR, the optimal number of inbred lines is 150, 200 and 350 for \( h^2 = 0.25 \), \( h^2 = 0.5 \) and \( h^2 = 0.75 \) respectively. At the highest heritability level, \( \varepsilon \)-SVR seems to prefer training sets of maximum size, at the cost of a very small molecular marker fingerprint size. The observed behavior of both BLP and \( \varepsilon \)-SVR is consistent with the results of a previous study (MAENHOUT et al. 2010) where it was
FIGURE 6.—Average prediction accuracy of $\varepsilon$-SVR and BLP prediction models over 100 iterations of the simulation routine for varying levels of the minimum required contrast precision $CD_{min}$, expressed as a percentile of the sampled CD values ranging from 0 to 0.875, and the number of genotyped inbred lines. The height of each point in the wireframe represents the prediction accuracy obtained by $\varepsilon$-SVR and BLP when training on the optimal selection of hybrids under the constraints imposed by the levels of the two independent variables. Prediction accuracy is expressed as the average squared Pearson correlation between the simulated and the predicted genotypic values of the hybrids. The interval at the bottom of each wireframe provides the minimum and maximum standard error of the mean. The scales of the vertical axes are only comparable within the same heritability level.
shown that BLP is less sensitive to a reduction of the number of training examples compared to $\varepsilon$-SVR, as long as the molecular marker fingerprint is dense. $\varepsilon$-SVR on the other hand, although requiring a training set of considerable size, handles smaller or less informative molecular marker fingerprints better than BLP.

DISCUSSION

This article presents three selection problems that are relevant to the budget-constrained construction of a genomic prediction model from available genetic evaluation data. The first problem considers the selection of exactly $p$ individuals from a set of $t$ candidates that will be genotyped to serve as training examples for the construction of the prediction model. This selection should be optimal in the sense that a linear mixed model analysis of the associated phenotypic records should result in a set of $p$ BLUPs of genotypic values that have the highest precision of all possible selections. By defining the precision of a selection as the minimum generalized coefficient of determination of a pairwise contrast, this selection problem can be translated to the ‘discrete $p$-dispersion problem’ from the field of graph theory. The reduction of this problem to a set of maximum clique problems allows to visualize the trade-off between selection size and selection quality. The greedy nature of a breeding program does unfortunately bias the presented selection approach towards high-performing individuals. These are generally tested more thoroughly than their low-performing colleagues. As the latter generally only have a few associated phenotypic records, the pairwise contrasts involving these individuals have a low precision which in turn makes their selection by the described procedure very unlikely. As a consequence, the resulting genomic prediction model is likely to overestimate the capabilities of the low-performing individuals. To avoid this bias, the selection procedure should optimize two objectives simultaneously: (1) maximizing the minimum precision of all pairwise contrasts in the selection and (2) maximizing the genetic variance in the selection. Even if one would succeed in finding an acceptable trade-off between these conflicting objectives, the estimates of the genotypic value of low-performing individuals will always suffer from large standard errors which makes them unreliable training examples.
The second problem we discuss deals with the selection of exactly \( q \) molecular markers from a set of \( t \) candidates for which the relative positions on a genetic map are known. To guarantee that the selection has an optimal genome coverage, we maximize the minimum intermarker distance. We show that this problem can be translated to a one-dimensional discrete \( p \)-dispersion problem for which an exact algorithm is available.

The third problem is specific to hybrid breeding programs and entails the selection of exactly \( k \) parental inbred lines such that the number of single-cross hybrids in the selection is maximized. If we represent the inbred lines as vertices of a graph and each single-cross hybrid as an edge between its parental vertices, this problem can be translated to the ‘densest \( k \)-subgraph problem’ which we solve by using a greedy heuristic.

The presented solutions to the three selection problems are put into practice in a simulation study where the goal is to find the optimal number of training examples for the construction of \( \varepsilon \)-SVR and BLP prediction models with maximal prediction accuracy under a fixed genotyping budget. At each iteration of the simulation routine, inbred lines, hybrids and their associated phenotypic data structure are sampled from actual genetic evaluation data. The number of training examples is gradually reduced by putting constraints on the data quality and the number of genotyped inbred lines. The results indicate that selections of training examples containing disconnected individuals are detrimental for the prediction accuracy of both \( \varepsilon \)-SVR and BLP. More stringent data quality constraints are however not necessary. \( \varepsilon \)-SVR performs best if the number of parental inbred lines (i.e. the number of training examples) is maximized at the cost of a reduced genome coverage. BLP on the other hand performs best when trained on a smaller set of training examples for which a dense fingerprint is available.

Despite the fact that these conclusions are most likely specific to maize breeding programs and possibly even specific to the heterotic groups and breeding methods used by the data-providing breeding company, the presented graph-based data selection algorithms should prove themselves to be useful for the construction of genomic prediction models in other plant and animal species as well. Evidently, more species-specific case studies are required to ascertain this claim.
Acknowledgments: The authors would like to thank the people from RAGT R2n for their unre-served and open-minded scientific contribution to this research.

REFERENCES


BERNARDO, R., 2008 Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Sci. 48: 1649–1664.


APPENDIX: IDENTIFYING DISCONNECTED PAIRS OF INDIVIDUALS

A1: by examination of the PEV matrix  If we analyze the available genetic evaluation data with a linear mixed model according to Eq. (1) where the variance structure is simplified to

\[
\text{Var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} \sigma^2_g & 0 \\ 0 & \sigma^2_e \end{bmatrix},
\]

we can express the prediction error variance matrix as (HENDERSON, C.R. 1984)

\[
\text{PEV}(\hat{u}) = (Z'MZ + I\lambda)^{-1}\sigma^2_e,
\]

where \( \lambda = \frac{\sigma^2_e}{\sigma^2_g} \) and \( M \) is the orthogonal projector on the column space of matrix \( X \) as \( M = I - X(X'X)^{-1}X' \). The matrix product \( Z'MZ \) is in fact the information matrix of the genotypic effects if we would consider both environments and genotypic effects as fixed and analyze the data as a block design in a linear least squares setting. CHAKRABARTI, M.C. (1964) proves that if such a block design is fully connected (i.e. all elementary contrasts are estimable in a least squares sense), the rank of this information matrix equals \( t - 1 \), where \( t \) is the number of fitted genotypic effects. Furthermore, HEILIGERS (1991) proves that in case the information matrix has a lower rank \( t - p \), where \( p \geq 2 \), the design is disconnected and the symmetric matrix \( Z'MZ \) can always be put in a block diagonal form with \( p \) distinct blocks around the principal diagonal, by simply permuting the appropriate rows and columns. Each of these blocks represents a set of fully connected individuals that are disconnected from all other individuals that are not represented in that particular block. If we assume that we are dealing with a disconnected design and that the columns of \( Z \) (i.e. the individuals) are ordered in such a way that \( Z'MZ' \) is in block diagonal form, it should be fairly obvious that also \( \text{PEV}(\hat{u}) \) is block diagonal as inversion preserves this matrix property. As most linear mixed model packages provide the PEV matrix, the identification of disconnected pairs of individuals can be performed by recovering this block diagonal structure by appropriate row and column permutations.
**A2: by computing the transitive closure**  If \( \text{Var}(u) \) is not a diagonal matrix, the block diagonal structure of \( Z'MZ \) is not preserved in the PEV matrix. One could of course examine the structure of \( Z'MZ \) instead, but this matrix is generally not available. It might therefore be easier to identify disconnected pairs of individuals by determining the transitive closure of their adjacency matrix. This is a symmetric, Boolean \( t \times t \) matrix where the element on row \( i \) and column \( j \) is set to 1 if individuals \( i \) and \( j \) have been evaluated in a common environment and 0 otherwise. For the example in Figure 1, this adjacency matrix looks like

\[
A = \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{bmatrix}
\]

The transitive closure of this matrix is again a symmetric, block-diagonalizable, Boolean matrix which can be interpreted in a similar way as \( Z'MZ \). **Warshall** (1962) describes a concise and efficient algorithm for computing the transitive closure of an adjacency matrix which has a worst case complexity of \( O(t^3) \). More advanced algorithms are described by **Naessens et al.** (2002) and **De Meyer et al.** (2004).