Arch. Tierz., Dummerstorf **50** (2007) 3, 294-308

Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere (FBN), Dummerstorf, Germany

CHRISTINE BAES and NORBERT REINSCH

Computing the condensed conditional gametic QTL relationship matrix and its inverse

Abstract

The inverse of the conditional gametic relationship matrix ($\mathbf{G}^{-1}$) for a marked quantitative trait locus (MQTL) is required for estimation of gametic effects in best linear unbiased prediction (BLUP) of breeding values if marker data are available. Calculation of the "condensed" gametic relationship matrix $\mathbf{G}^{*}$ - a version of $\mathbf{G}$ where linear dependencies have been removed - and its inverse $\mathbf{G}^{*-1}$ is described using a series of simplified equations following a known algorithm. The software program COBRA (**co**variance **b**etween **r**el**a**tives for a marked QTL) is introduced, and techniques for storing and computing the condensed gametic relationship matrix $\mathbf{G}^{*}$ and the non-zero elements of its inverse are discussed. The program operates with both simple pedigrees and those augmented by transmission probabilities derived from marker data. Using sparse matrix storage techniques, $\mathbf{G}^{*}$ and its inverse can be efficiently stored in computer memory. COBRA is written in FORTRAN 90/95 and runs on a variety of computers. Pedigree data and information for a single MQTL in the German Holstein population are used to test the efficiency of the program.

Key Words: marker assisted selection, best linear unbiased prediction, gametic relationship matrix

Zusammenfassung

Titel der Arbeit: **Berechnung der eingedampften gametischen QTL Verwandtschaftsmatrix und ihrer Inversen**

Die Inverse $\mathbf{G}^{-1}$ der Gametischen Verwandtschaftsmatrix für einen QTL (quantitative trait locus) mit gekoppelten Markern wird für eine markerunterstützte Zuchtwertschätzung benötigt. Es wird beschrieben wie die eingedampfte gametische Verwandtschaftsmatrix $\mathbf{G}^{*}$ - eine Variante von $\mathbf{G}$ ohne lineare Abhängigkeiten - in einer Abfolge einfacher Rechenschritte nach einem bekannten Algorithmus berechnet werden kann. Das EDV-Programm COBRA (**co**variance **b**etween **r**el**a**tives for a marked QTL) wird vorgestellt, um anschließend Techniken für die Speicherung und die Berechnung von $\mathbf{G}^{*}$ und die Nicht-Nullelemente ihrer Inversen zu besprechen. COBRA kann sowohl einfache Pedigreedaten als auch Pedigreedaten mit Markerinformation verarbeiten, wobei Speichertechniken für dünnbesetzte Matrizen eingesetzt werden. COBRA ist in FORTRAN 90/95 geschrieben und kann auf einer Vielzahl verschiedener Rechner laufen. Pedigreedaten und Informationen für einen einzelnen MQTL in der deutschen Holstein-Population werden benutzt, um die Effizienz des Programms zu prüfen.

Schlüsselwörter: Marker-gestützte Selektion, bester linearer unverzerrter Prediktor, gametische Verwandtschaftsmatrix

1. Introduction

The joint utilization of marker and phenotype information in current genetic prediction models is evolving rapidly. FERNANDO and GROSSMAN (1989) incorporated marked quantitative trait loci (MQTL) information into the existing best linear unbiased prediction (BLUP) breeding value estimation model by splitting the 'genetic' portion of the model into the additive effects of the unique QTL gametes and the polygenic effects. In order to include MQTL information in the estimation model, the

inverse of the conditional covariance matrix of QTL allele effects $\mathbf{G}$ is required ($\mathbf{G}^{-1}$).The efficient computation of matrices like $\mathbf{G}$ and $\mathbf{G}^{-1}$ for large pedigrees is crucial for further successful incorporation of marker data in genetic evaluation models.

A numerically efficient algorithm for the calculation of $\mathbf{G}$ and its inverse for an MQTL was developed by ABDEL-AZIM and FREEMAN (2001) based on the work of FERNANDO and GROSSMAN (1989), VAN ARENDONK et al. (1994) and WANG et al. (1995). TUCHSCHERER et al. (2004) showed that the calculation of $\mathbf{G}$ depends on the mode of gamete identification (gametes identified by markers *vs.* gametes identified by parental origin), although the final MA BLUP breeding value of each animal is identical irrespective of the gamete identification method employed. They suggested that gamete identification by parental origin may have practical advantages compared to that by markers; for example, fewer values are required to denote marker related transmission probabilities. More importantly, the generalisation developed by TUCHSCHERER et al. (2004) showed that under certain circumstances identical rows and columns in $\mathbf{G}$ may occur if parents pass identical copies of their gametes to their offspring (*i.e.* $\mathbf{G}$ may be rank deficient); in such cases, the inverse matrix $\mathbf{G}^{-1}$ is not defined. By excluding duplicate gamete information, the number of gametic effects in $\mathbf{G}$ can be reduced to a smaller set of unique effects in a 'condensed' gametic relationship matrix $\mathbf{G}^{*}$; $\mathbf{G}^{*}$ is always of full rank and $\mathbf{G}^{*-1}$ is defined. Additionally, the smaller $\mathbf{G}^{*}$ matrix requires less memory and may be calculated faster than a larger one.

In the first section of this article, the calculation of $\mathbf{G}^{*}$ and its inverse is described to illustrate the practical application of the generalised algorithm presented by TUCHSCHERER et al (2004). Secondly, the software program COBRA (**co**variance **b**etween **r**el**a**tives at a marked QTL) is introduced, and techniques for determining, storing and computing the condensed gametic relationship matrix $\mathbf{G}^{*}$ and the non-zero elements of its inverse are discussed. Finally, pedigree data and information for a single marker in the German Holstein population are used to test the efficiency of the program.

2. Calculating $\mathbf{G}^{*}$ and its inverse

The matrix $\mathbf{G}$ was developed by SMITH (1984) and SMITH and ALLAIRE (1985) to calculate the probability of any two gametes being identical by descent in an inbred population (cited by SCHAEFFER et al., 1989). It is symmetric and contains one row and one column for each gamete (*i*), which are calculated from the rows and columns belonging to the gametes' predecessor gametes (called the gametes' parents here forth). The diagonal elements of $\mathbf{G}$ are equal to one ($\mathbf{G}^{*}_{(i,i)} = 1.0$ for gamete *i*), as the probability of a gamete being identical to itself is always equal to one. The probability that the parental gamete received from the sire of the animal, Pg_a and the parental gamete from the dam Mg_a are identical by descent is the inbreeding coefficient f_a of individual *a*, $\mathbf{G}^{*}_{(Pg_a, Mg_a)} = f_a$).

In a pedigree with *n* animals, the matrix $\mathbf{G}$ has the dimension 2*n* x 2*n* because every animal is assumed to have two unique gametes. However, if an exact copy of a parental gamete is passed to its offspring, the effects of that gamete are included twice

in $\mathbf{G}$; the computation of $\mathbf{G}^{-1}$ fails due to the linear dependencies in $\mathbf{G}$ (see section 4. in TUCHSCHERER et al., 2004). If copied alleles are excluded and only unique gametes are assigned rows and columns in $\mathbf{G}^{*}$, the linear dependencies caused by identical rows and columns no longer exist. This means that animals with two unique gametes contribute two rows and two columns to $\mathbf{G}^{*}$, animals with one unique gamete and one copy of a gamete contribute only one row and column, and for animals with two copied gametes no rows or columns are added. The determination of unique and copied gametes depends on the pedigree and transmission probability information and is outlined below.

2.1 Assumptions

Consider one or several marker loci closely linked to a quantitative trait locus (MQTL), with linkage equilibrium between the markers and the MQTL. A recombination rate of zero may occur between an MQTL and one (or more) marker(s). Markers may be single, flanking or multiple for the MQTL, and the number of markers and the distance between them is not limited. The first allele is assumed to be the paternal gamete *(Pg)* and the second is considered to be the maternal gamete *(Mg)*. The paternal transmission probability $(T_{(Pg)})$ is the probability that the sire of an animal passed his paternal (first) gamete to his offspring. Likewise, the maternal transmission probability $(T_{(Mg)})$ is the probability that the dam of an animal passed her paternal (first) gamete to her offspring. The probability of the maternal (second) gamete of the sire or dam being passed on to the offspring can be calculated by subtracting the paternal or maternal transmission probability (respectively) from one. Thus two values (paternal and maternal) actually represent all four possible probabilities.

2.2 Method

The calculation of the gametic relationship matrix using recursive algorithms requires an ordered pedigree in which parental animals occur before their progeny, and transmission probabilities for all non-founder individuals conditional on marker information for the paths sire → progeny and dam → progeny. The calculation of transmission probabilities is described by MAYER et al. (2007, submitted). For missing and non-informative markers the transmission probability is 0.5; the resulting contributions to the relationship matrix are identical to those in the classical numerator relationship matrix.

If an animal has a transmission probability of one for a certain gamete, that animal will receive the paternal gamete from its respective parent with 100% certainty; the gamete received by the animal is an exact copy of the parental gamete and is not unique. Conversely, if a transmission probability of zero occurs for a given gamete, the animal will receive the maternal gamete from its respective parent with 100% certainty.

It is possible to set up a gametic pedigree following the calculation of transmission probabilities; the number of unique gametes can be determined and the size of $\mathbf{G}^{*}$ can be calculated. All unique paternal and maternal gametes are assigned an integer identification number in ascending order. Predecessor gametes of paternal gametes, $Pg_{(Pg)}$ and $Mg_{(Pg)}$, and predecessor gametes of maternal gametes, $Pg_{(Mg)}$ and $Mg_{(Mg)}$ must also be considered.

In contrast to the gametic identification numbering method employed in the calculation of $\mathbf{G}$, non-unique gametes do not receive unique identification numbers in

$\mathbf{G}^*$. If a gamete is passed from a parent to its offspring with 100% certainty, it is included in the gametic pedigree with the same identification number as the gamete from which it originated.

2.3 Calculation of the matrix $\mathbf{G}^*$

The calculation of $\mathbf{G}^*$ is described by equation (9) in TUCHSCHERER et al. (2004).
If the information in the pedigree is sorted with parents precede their progeny, it is possible to build $\mathbf{G}^*$ recursively starting with the top left corner and working towards the right. The transmission probability assigned to the gamete depends on whether it is the paternal or maternal gamete of the animal in question.
The matrix $\mathbf{G}^*$ is symmetric, therefore only the upper triangular matrix needs to be calculated; the lower triangular matrix is identical to its upper counterpart. Let (*i,j*) represent the row and column indices for an element in $\mathbf{G}^*$ and assume that $i>j$. The elements of each column are calculated from row one until the last row before the diagonal (j-1), using the information from the predecessor gametes and the transmission probability of the gamete being calculated.
For columns of paternal gametes with known predecessors (*i.e. j* is a paternal gamete),

$$\mathbf{G}^*_{(i,j)} = T_{(Pg)_j} \times \mathbf{G}^*_{(i,Pg_{(Pg)j})} + (1 - T_{(Pg)_j}) \times \mathbf{G}^*_{(i,Mg_{(Pg)j})} \qquad (1)$$

where $T_{(Pg)j}$ is the transmission probability for the paternal gamete *j*, *i* denotes the row to be calculated, $Pg_{(Pg)j}$ is the paternal predecessor gamete of the paternal gamete *j* and $Mg_{(Pg)j}$ is the maternal predecessor gamete of the paternal gamete *j*. The calculation of columns of maternal gametes with known parents (*i.e. j* is a maternal gamete) is accomplished similarly,

$$\mathbf{G}^*_{(i,j)} = T_{(Mg)_j} \times \mathbf{G}^*_{(i,Pg_{(Mg)j})} + (1 - T_{(Mg)_j}) \times \mathbf{G}^*_{(i,Mg_{(Mg)j})} \qquad (2)$$

where $T_{(Mg)j}$ is the transmission probability for the maternal gamete *j*, *i* is the row to be calculated, $Pg_{(Mg)j}$ is the paternal predecessor gamete of the gamete *j* and $Mg_{(Mg)j}$ is the maternal predecessor gamete of the gamete *j*. If the predecessors of *i* and *j* are unknown, $\mathbf{G}^*_{(i,j)} = 0$. In this way, $\mathbf{G}^*$ can be calculated for all unique gametes and no linear dependencies occur.

2.4 The inbreeding coefficient

The inbreeding coefficient at the MQTL, f_a, is the probability that the gametes at the MQTL in individual *a* are identical by descent and can be found in the element (*Pg*,*Mg*) of $\mathbf{G}^*$ for each individual. In $\mathbf{G}$, this value is always located directly above the diagonal of the maternal gamete (or below the diagonal of the paternal gamete) for every individual. Although $\mathbf{G}^*$ also contains inbreeding coefficients for each individual, they are not necessarily located in the same position as in $\mathbf{G}$ (*i.e.* directly above the maternal gamete of the individual) because the matrix is ordered differently due to the exclusion of non-unique gametes.
It is important to realize that the f_a of an animal is a function of its respective transmission probabilities and the relationships between its parental gametes. Although the f_a are the only elements of $\mathbf{G}^*$ required to calculate its inverse, it is apparent that the f_a of an individual depends on certain pre-existing elements of $\mathbf{G}^*$. Technically,

only very specific elements of $\mathbf{G}^*$ are needed for the calculation of its inverse. TIER (1990) described how the minimum subset of matrix $\mathbf{G}$ required for the calculation of $\mathbf{G}^{-1}$ can be determined. However, the computation method presented here is sufficient for medium-sized pedigrees with a half sib structure.

2.5 Calculation of the matrix $\mathbf{G}^{*-1}$

The detailed calculation of $\mathbf{G}^{*-1}$ is described by equation (10) in TUCHSCHERER et al. (2004). The matrix $\mathbf{G}^{*-1}$ can be calculated using only the pedigree information (from which the gametic pedigree is derived), the inbreeding coefficient of each individual and the paternal and maternal transmission probabilities. The size of $\mathbf{G}^{*-1}$ is the same as that of $\mathbf{G}^*$, and each column and row are once again assigned to one gamete. Only the elements of the upper triangular matrix are calculated; the lower triangular matrix is obtained by symmetry. Whereas the calculation of $\mathbf{G}^*$ consisted of adding one column (and therefore one diagonal) for every unique gamete to an existing matrix, $\mathbf{G}^{*-1}$ is calculated by adding one column and one diagonal for every gamete as well as adding values to specific existing elements in $\mathbf{G}^{*-1}$.
It is possible to build $\mathbf{G}^{*-1}$ for one individual at a time by passing through the pedigree from the first individual to the last, where one of four possible situations may occur:

1 - the individual has no unique gametes (no columns or rows are added):
- no changes are made to the existing matrix

2 - only one gamete is unique (only one column is added, paternal or maternal):
- one new diagonal exists,
- up to two new values exist in the new column and
- up to three values in the existing matrix are altered

3 - both gametes are unique (two columns added):
- two new diagonals exist,
- up to four new values exist in the new columns and
- up to six values in the existing matrix are altered

Let *Pg* and/or *Mg* represent the new column(s) in $\mathbf{G}^{*-1}$, and therefore also the unique paternal or maternal gamete (respectively) of an individual. If situation 1 occurs, no calculations need to be made and no columns are added because the contributions of such individuals have already been accounted for.
If an individual has one unique paternal gamete, the following calculations are made using the paternal transmission probability $T_{(Pg)}$ of the individual and the inbreeding coefficient of the individuals' sire. The elements of $\mathbf{G}^{*-1}$ where values are to be added are located in the columns of the paternal gametes' predecessor gametes $Pg_{(Pg)}$, $Mg_{(Pg)}$, and the elements where values are to be subtracted are located in the column of the gamete *Pg* with the exception of the diagonal, which is positive. For the new diagonal element:

$$\mathbf{G}^{*-1}_{(Pg,Pg)} = 1/d_i \qquad (3a)$$

for the new elements in the added column:

$$\mathbf{G}^{*-1}_{(Pg_{(Pg)},Pg)} = -(T_{(Pg)} / d_i) \quad (3b)$$

$$\mathbf{G}^{*-1}_{(Mg_{(Pg)},Pg)} = -(1-T_{(Pg)}) / d_i \quad (3c)$$

and for the elements in the existing matrix:

$$\mathbf{G}^{*-1}_{(Pg_{(Pg)},Pg_{(Pg)})} = (T_{(Pg)}{}^2 / d_i) \quad (3d)$$

$$\mathbf{G}^{*-1}_{(Mg_{(Pg)},Mg_{(Pg)})} = (1-T_{(Pg)})^2 / d_i \quad (3e)$$

$$\mathbf{G}^{*-1}_{(Pg_{(Pg)},Mg_{(Pg)})} = (T_{(Pg)})(1-T_{(Pg)}) / d_i \quad (3f)$$

where $d_i = 2T_{(Pg)}(1-T_{(Pg)})(1-f_{sire})$ if the sire is known, and d_i=1 if the sire is unknown (see equation 10 in TUCHSCHERER et al. (2004) and appendix A for a proof of d_i).

If an individual has one unique maternal gamete, the same calculations are made, however, the maternal transmission probability $T_{(Mg)}$ of the individual and the inbreeding coefficient of the individuals' dam are used. The elements where values are to be added are located in the columns of the maternal gametes' predecessor gametes *Pg(Mg)*, *Mg(Mg)*, and the elements where values are to be subtracted are located in the column of the gamete *Mg* with the exception of the diagonal, which is positive. For diagonal elements:

$$\mathbf{G}^{*-1}_{(Mg,Mg)} = 1 / d_i \quad (4a)$$

for the elements in the added columns:

$$\mathbf{G}^{*-1}_{(Pg_{(Mg)},Mg)} = -(T_{(Mg)} / d_i) \quad (4b)$$

$$\mathbf{G}^{*-1}_{(Mg_{(Mg)},Mg)} = -(1-T_{(Mg)}) / d_i \quad (4c)$$

and for the elements in the existing matrix:

$$\mathbf{G}^{*-1}_{(Pg_{(Mg)},Pg_{(Mg)})} = (T_{(Mg)}{}^2 / d_i) \quad (4d)$$

$$\mathbf{G}^{*-1}_{(Mg_{(Mg)},Mg_{(Mg)})} = (1-T_{(Mg)})^2 / d_i \quad (4e)$$

$$\mathbf{G}^{*-1}_{(Pg_{(Mg)},Mg_{(Mg)})} = (T_{(Mg)})(1-T_{(Mg)}) / d_i \quad (4f)$$

where $d_i = 2T_{(Mg)}(1-T_{(Mg)})(1-f_{dam})$ if the dam is known, and d_i=1 if the dam is unknown.

If both *Pg* and *Mg* are unique, all 12 of the above calculations (*i.e.* Equations 3 a-f and 4 a-f) are used. For base animals, two columns are added; the parental gametes are unknown, and only a one is added in the diagonal.

The matrix $\mathbf{G}^{*-1}$ is calculated for one individual at a time and therefore 'layer by layer' using the equations given above. Once all individual layers have been calculated, they are added to form the final matrix $\mathbf{G}^{*-1}$.

3. Computing $\mathbf{G}^*$ and its inverse

The computation of matrices like $\mathbf{G}$ and $\mathbf{G}^{-1}$ require very large amounts of computer memory. ABDEL-AZIM and FREEMAN (2001) described computational techniques for calculating and storing the minimum possible elements of $\mathbf{G}$ and its inverse using computational methods described by TIER (1990) and linked list storage techniques. Gametes were identified by markers. To the authors' knowledge, computational techniques for efficiently calculating $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$ have not yet been published.

The following section introduces and describes the software program COBRA, which is a FORTRAN (90/95) program designed to determine, store and compute the non-zero elements of $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$ using the TUCHSCHERER et al. (2004) algorithm with gametes identified by parental origin. It is shown that $\mathbf{G}^*$ and its inverse can be efficiently stored in computer memory or saved to file using sparse matrix storage techniques.

COBRA operates in three main steps. In the first step, pedigree information is read in, checked for errors and written to three index matrices. The gametic index matrix is calculated and the user is informed of the number of animals and 'unique' gametes in the pedigree. In step two, $\mathbf{G}^*$ is computed from the index information and the inbreeding coefficients calculated are saved. The user is informed of the number of non-zero elements in $\mathbf{G}^*$ and the fill density of $\mathbf{G}^*$ (in percent). Finally, the index information from step one and the inbreeding coefficients from step two are used to calculate the $\mathbf{G}^{*-1}$ matrix. The user is informed of the number of non-zero elements in $\mathbf{G}^{*-1}$ and the fill density of $\mathbf{G}^{*-1}$ (in percent). Additionally, the number of individuals with no unique gametes, one unique paternal gamete, one unique maternal gamete and two unique gametes is given and the program is complete. COBRA includes the option of saving information on gamete occurrence, inbreeding coefficients with a reference list ordering gametes to animals and the non-zero elements of $\mathbf{G}^{*-1}$ to file.

3.1 Preparation of index matrices

A text file containing the identification numbers of sire and dam, followed by the transmission probabilities of paternal and maternal gametes is the input pedigree. Each line in the input pedigree file signifies one individual and the pedigree must be ordered such that parents precede their progeny. The column j of all index matrices corresponds to the animal identification number (1 to n). The input pedigree file is first read to determine the number of animals (n) and to test for simple errors in the data (duplicate entries, wrong order, transmission probabilities > 1). The file is rewound, memory is allocated for the first two index matrices (index matrix N1 has dimension 2 x n, index matrix N2 has dimension 3 x n) and the pedigree is read in.

Index matrix N1 contains the parental identification numbers of animals 1 to n, with the identification number of the sire in the first row and that of the dam in the second. Unknown parents are assigned zeros. At this point, only the first two rows of index matrix N2 are read in; they contain the transmission probabilities of the paternal and maternal gametes of individuals 1 to n. If the sire or dam is unknown, the transmission probability for the paternal or maternal gamete is considered to be 0.5. Linkage equilibrium is assumed. The third row of N2 remains empty at this point, but is filled with the inbreeding coefficients of animals 1 to n once step two is complete.

The third index matrix N3 (dimension 6 x n) is now calculated from the information in N1 and N2. This matrix contains the paternal and maternal gamete identification

numbers of individual *n* in rows 1 and 4, respectively. The identification numbers for the paternal pair of predecessor gametes are saved in rows 2 and 3, and those of the maternal pair of predecessor gametes in rows 5 and 6. All paternal and maternal gametes are assigned an integer identification number in ascending order if the transmission probability of the gamete is not equal to 1.0 or 0.0, starting with the first animal in the pedigree. If the origin of a gamete is unknown, zeros are assigned as parental paternal and maternal gamete identification numbers.

3.2 Writing $\mathbf{G}^*$

The $\mathbf{G}^*$ matrix is calculated using the generalization of the tabular method of WANG et al. (1995) presented by TUCHSCHERER et al. (2004). This formulation adds the covariance information for each gamete column-wise to the upper triangle of $\mathbf{G}^*$, starting with column one. Columns are calculated from row one to the last row before the diagonal using the equations shown in section two, with the diagonal elements always equal to one.

Only the non-zero elements of the upper half of $\mathbf{G}^*$ are saved in sparse IA, JA, A format (KNUTH, 1997), which reduces the memory requirements of the program. The position of non-zero elements in $\mathbf{G}^*$ is initially unknown and is calculated from information in the index matrices N2 and N3 for each particular gamete using the methods described in section two. As $\mathbf{G}^*$ is calculated, hashing is used to store and retrieve the required non-zero elements of $\mathbf{G}^*$. The organisation of the hash table is arranged on a “first come first serve” basis; when an element is to be inserted, the locations of its probe sequence are examined sequentially until an empty spot in the vector A is found. The new element is saved to that location and its coordinates are saved in the parallel IA JA vectors.

Diagonals of gametes without offspring are included, however their columns are not calculated. Should certain elements of such columns be required for other calculations, only those specific elements are computed. This ‘economised’ method of calculating $\mathbf{G}^*$ saves a large amount of memory and increases the speed of the program considerably, especially if the input pedigree contains many individuals without offspring.

Once $\mathbf{G}^*$ has been calculated, the inbreeding coefficient of each animal is retrieved from the IA, JA, A vectors and written into row three of the index matrix N2. The IA, JA and A vectors are cleared. As an option a list of animals, their respective paternal and maternal gametes and their inbreeding coefficients can be generated. The IA, JA, A vectors with the $\mathbf{G}^*$ matrix coordinates and values may also be sorted and saved to file before clearing if required. This may be useful when the data are to be processed by other software.

3.3 Writing $\mathbf{G}^{*-1}$

The inverse $\mathbf{G}^{*-1}$ of the condensed gametic relationship matrix can now be calculated using the TUCHSCHERER et al. (2004) generalization of the WANG et al. (1995) algorithm. Using index matrices N1, N2 and N3, values are either added to or subtracted from the required elements of $\mathbf{G}^{*-1}$ as described in section two. A simplified flowchart of the COBRA procedure for calculating $\mathbf{G}^{*-1}$ is included in Appendix B. As in the calculation of $\mathbf{G}^*$, only the non-zero elements of the upper half of the $\mathbf{G}^{*-1}$

matrix are calculated and saved in sparse IA, JA, A format. Hashing methods are used to store and retrieve the required elements. A list including animal identification number, row and column indices and values of non-zero elements is generated and sorted for the further calculation of BLUP breeding values.

4. Practical validation

Pedigree data from 12008 German Holstein bulls and bull dams (7174 males and 4834 females) was used to test the efficiency of the program. Pedigree and marker information originated from the second phase of the genome analysis project of the German Cattle Breeders Federation (Arbeitsgemeinschaft Deutscher Rinderzüchter, ADR) and is currently used for the MA BLUP evaluation of the trait somatic cell score. The pedigree was obtained from the United Information Systems Animal Production (Vereinigte Informationssysteme Tierhaltung w.V., VIT) in Verden, Germany. Information on a single highly polymorph marker with 15 alleles on chromosome 18 was included for 6520 typed animals (6050 males, 470 females). Allele frequencies ranged from 0.01% to 32.83%.

Transmission probabilities were calculated by extracting standard partial pedigrees (individual, sire, dam, sire's sire, sire's dam, dam's sire, dam's dam) from original pedigree data and applying SimWalk2 software as described by MAYER et al. (2007, submitted) and TUCHSCHERER et al. (2007, in preparation). The partial pedigrees were grouped by available genotype information. The individual and its male ancestors (sire and dam's sire) were genotyped in 33.6% of all partial pedigrees, followed closely by partial pedigrees in which the individual and its sire were genotyped (32.8%). Only 1.2% of the partial pedigrees contained full marker information for all individuals.

At the marker, 4584 of the typed males were not identical by descent, while the remaining 1466 typed males had an average inbreeding coefficient of 0.0332. Typed females included 373 animals which were not identical by descent at the marker, while the remaining 97 animals had an average inbreeding coefficient of 0.0342. A summary of results, along with maximum and minimum inbreeding coefficients, is shown in Tab. 1.

Table 1
Average, minimum and maximum inbreeding coefficients at the marker for typed males and females given observed marker genotype (1,466 male and 97 female typed animals and 12,008 animals including non-typed animals) (Durchschnittliche, minimale und maximale Inzuchtkoeffizienten typisierter Tiere am Marker, abgeleitet aus den beobachteten Markergenotypen (für 1.466 männliche und 97 weibliche typisierte Tiere, sowie alle 12.008 Tiere einschließlich der nicht typisierten))

		Typed animals[1]		All animals
		Male (1466)	Female (97)	(12008)
Including Marker Information	Average IBC[2]	0.0332	0.0342	0.00433
	Minimum	0.0000834	0.000635	0.000
	Maximum	0.455	0.377	0.455

[1] Only typed animals with an inbreeding coefficient > 0 are included.
[2] IBC = Inbreeding coefficient.

The pedigree was analysed twice; once including marker information and once with transmission probabilities set to 0.5 (*i.e.* not including marker information). Tab. 1 shows selected computational results from the COBRA program using the same pedigree with and without marker information.

It is apparent that the percentage fill in $\mathbf{G}^*$ is not as high as that in $\mathbf{G}$ (see '% Fill', Tab. 2). Although $\mathbf{G}^{*-1}$ seems to be slightly more full than $\mathbf{G}^{-1}$ (0.0278% compared to 0.0232%), it should be noted that only 51,553 non-zero elements are included in $\mathbf{G}^{*-1}$ compared to 66,820 non-zero elements in $\mathbf{G}^{-1}$, equating to a difference of more than 15,000 elements (see 'Remaining non-zero elements', Tab. 2). Furthermore, the condensed matrix has approximately 4,700 fewer columns and rows than the uncondensed counterpart; 19,238 unique gametic effects were included in $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$ compared to 24,016 effects in $\mathbf{G}$ and $\mathbf{G}^{-1}$ (see 'Diagonals', Tab. 2).

Table 2
Fill density (in percent), storage requirements, number of non-zero elements and execution time requirements for building $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$ using data from a 12008-animal pedigree with (first row) and without (second row) marker information (Besetzungsgrad (in Prozent), Anzahl der Nicht-Nullelemente, Ausführungszeit und Speicherauslastung für die Aufstellung von $\mathbf{G}^*$ und $\mathbf{G}^{*-1}$ mit einem Pedigree bestehend aus 12.008 Tieren mit (oben) und ohne (unten) Markerinformation)

		% Fill	Elements saved for gametes without offspring	Remaining Non-zero Elements	Execution time[1] (seconds)	Diagonals
Including marker information	$\mathbf{G}^*$	2.47	131,739,817	4,567,164	23.869	19,238
	$\mathbf{G}^{*-1}$	0.0278		51,553	0.0359	
Not including marker information	$\mathbf{G}$	3.29	212,134,432	9,499,265	46.900	24,016
	$\mathbf{G}^{-1}$	0.0232		66,820	0.0452	

[1]A Dell Precision 630 Workstation with double processor (2x3.6 GHz Xeon, 2x72 GB SCSI hard drive, 8 GB RAM) running Suse-Linux 9.3 (64 bit Version) was used in the evaluation. The program was compiled with Intel-Fortran Version 8.1

The pedigree contained approximately 9,000 animals without offspring, which improved the efficiency of the program greatly since columns of $\mathbf{G}$ and $\mathbf{G}^*$ relating to these gametes were not calculated (see 'Elements saved for gametes without offspring', Tab. 2), probably because the non-unique gametes often belong to younger animals which do not yet have offspring.

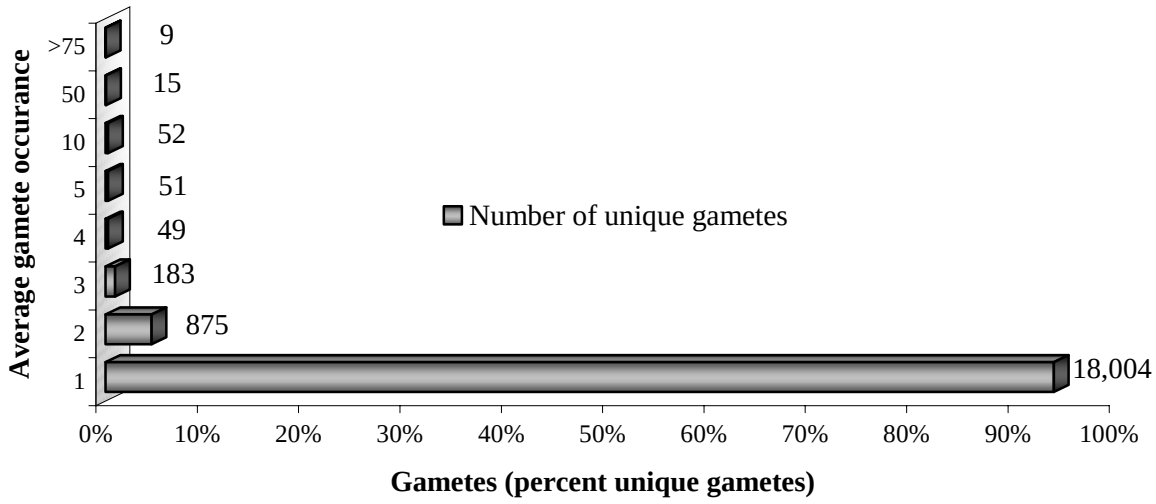

Fig. 1: Average gamete occurrence (animals per gamete) in a 12,008 animal pedigree containing 19,238 unique gametes and 6,520 genotyped animals (Durchschnittliche Häufigkeit des Auftretens von Gameten (Anzahl Tiere je Gamet) in einem Pedigree mit 12.008 Tieren, 19.238 eindeutigen Gameten und 6.520 typisierten Tieren)

Non-unique gametes have a potentially large effect on the population, especially if their copies occur frequently. Gametes in the current pedigree, which includes both typed and non-typed individuals, occurred an average of 1.25 times, with a maximum occurrence of 396. Fig. 1 shows that 18,004 gametes (93.59%) occur only once, while only 9 gametes (0.05%) have over 75 copies in bulls or bull dams.

5. Discussion

This article describes how the conditional gametic relationship matrix and its inverse can be condensed using the TUCHSCHERER et al. (2004) algorithm. It describes how ordered pedigree information and transmission probabilities can be used to set up a gametic pedigree for the calculation of $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$. The computer program COBRA was introduced and its structure was explained. Finally, pedigree data from 12,008 German Holstein animals was analysed and used to test the efficiency of the program.

The concept of reducing the size of $\mathbf{G}$ was proposed by MEUWISSEN and GODDARD (1996) for multiple markers: parents and offspring sharing the same marker haplotype were both assigned the same gametic QTL effect by assuming that the probability of double recombination was equal to zero for informative animals (transmission probabilities were rounded to 0, 0.5 or 1.0). In reality, marker information (and therefore the transmission probability for each marker) is variable and may have any value between zero and one; the values approach zero or one for more informative markers (LIU and MATHUR, 2005). The condensing algorithm presented by TUCHSCHERER et al. (2004) leads to the same result as the reducing algorithm when transmission probabilities of one and zero occur and no recombination takes place. However, the condensing algorithm uses the original transmission probabilities instead of rounded ones. If transmission probabilities very close to one or zero occur, a predefined threshold (e.g. ε=0.03) can be entered in COBRA, causing all transmission probabilities below 0.03 to be treated as zero and all those above 0.97 to be treated as one. This further condenses $\mathbf{G}^*$ and $\mathbf{G}^{*-1}$ and provides similar results to those of MEUWISSEN and GODDARD (1996).

Economising on the calculation of elements in rows and columns related to gametes without offspring can reduce the number of non-zero elements in $\mathbf{G}^*$ significantly (*i.e.* minimal computation of non-parental gametic effects). The fill density and the number of non-zero elements with (2.47%, 4,567,164 elements) and without (21.31%, 39,430,374) economising on non-parental gametic effects ($\mathbf{G}_a^*$) for the pedigree described in section 4 underlines the significance of this method with regard to matrix size .

The distribution pattern of non-zero elements in $\mathbf{G}_a^*$ is presented in Fig. 2, where the upper triangular matrix shows the distribution pattern of non-zero elements in $\mathbf{G}_a^*$. Each non-zero element in the matrix is represented by a single dot; the matrix may appear darker than in reality due to low print resolution.

Arch. Tierz. **50** (2007) 3

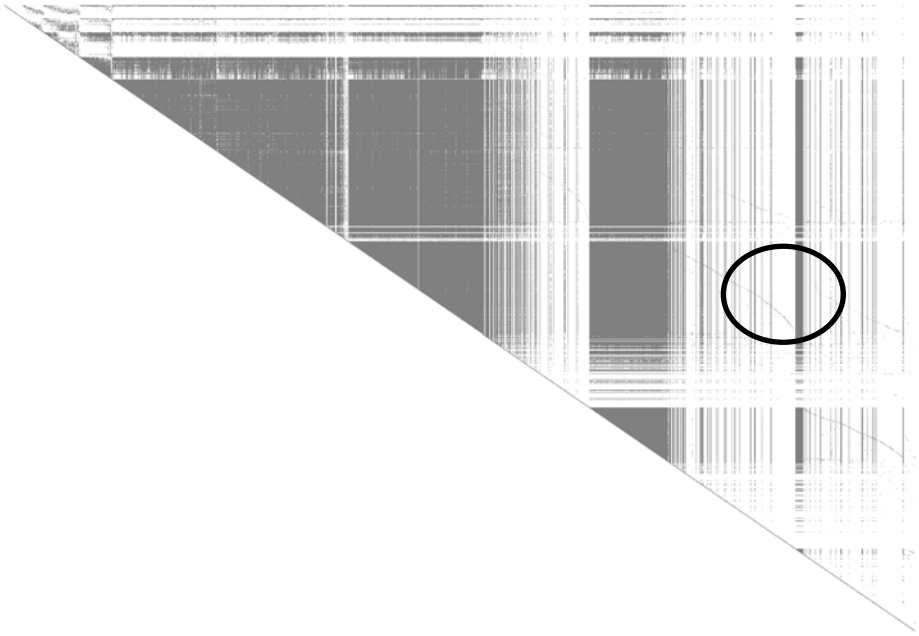

Fig. 2: Distribution pattern of non-zero elements in $\mathbf{G}_{\mathbf{a}}^{*}$ (upper triangular matrix); non-zero elements are only calculated for animals with progeny. Empty columns contain inbreeding coefficients of animals without progeny, visible as a diagonal trace of non-zero elements in the circle. (Verteilungs-Muster der Nicht-Nullelemente in $\mathbf{G}_{\mathbf{a}}^{*}$ (obere-Dreiecksmatrix), d.h. ohne die Berechnung von Nicht-Nullelementen für Tiere ohne Nachkommen, leere Spalten enthalten Inzuchtkoeffizienten von Tieren ohne Nachkommen, sichtbar als diagonale Spur von Nicht-Nullelementen innerhalb des Kreises)

The circle in Fig. 2 draws attention to a faint arrangement of 'irregular' dots in the otherwise regular checked-pattern visible in $\mathbf{G}_{a}^{*}$. These irregular elements are computed when certain constituents of economised columns are required for other calculations, such as those required for inbreeding coefficients of animals without unique gametes.

In order to successfully incorporate MQTL information in the BLUP breeding value estimation model, the marker genotype information of as many animals as possible is required. Fig. 1 shows that the vast majority of gametes occurred only once in the pedigree (93.59%), however transmission probabilities for marker genotypes were only included for 6,520 of 12,008 animals. The average gamete occurrence would likely increase if the missing genotype information, especially for bull dams, was included.

The efficient calculation of $\mathbf{G}$ and $\mathbf{G}^{-1}$ is imperative if marker data are to be practically used in genetic evaluation models. The computational techniques proposed by ABDEL-AZIM and FREEMAN (2001) for constructing the inverse are efficient, however the condensing algorithm proposed by TUCHSCHERER et al. (2004) is not included. The relatively uncomplicated programming methods for the calculation of $\mathbf{G}^{*}$ and $\mathbf{G}^{*-1}$ described in this article prove adequate for medium sized pedigrees with simple structure. However, computation time required may be too high for the calculation of larger pedigrees (>100,000 animals) or pedigrees with a more complex

constitution. Computation time may therefore be improved in later versions of the COBRA program.

Acknowledgements

Financial support from the FUGATO MAS.Net project is gratefully acknowledged.

References

ABDEL-AZIM, G.; FREEMAN, A.:
A rapid method for computing the inverse of the gametic covariance matrix between relatives for a marked Quantitative Trait Locus, Genet. Sel. Evol. **33** (2001), 153-173

FERNANDO, R.L.; GROSSMAN, M.:
Marker-assisted selection using best linear unbiased prediction, Genet. Sel. Evol. **21** (1989), 467-477

KNUTH, D.:
The Art of Computer Programming, Volume 3: Sorting and Searching. Addison-Wesley (1997), 513–558

LIU, Y.; MATHUR, P.K.:
Simplification of marker-assisted genetic evaluation and accounting for non-additive interaction effects. Arch. Tierz., Dummerstorf **48** (2005), 460-474

MEUWISSEN, T.H.E.; GODDARD, M.E.:
The use of marker-haplotypes in animal breeding schemes, Genet. Sel. Evol. **28** (1996), 161-176

SCHAEFFER, L.R.; KENNEDY, B.W.; GIBSON, J.P.:
The Inverse of the Gametic Relationship Matrix. J. Dairy Sci. **72** (1989), 1266-1272

SMITH, S.P.:
Dominance relationship matrix and inverse for an inbred population. Mimeo, Dep. Dairy Sci., Ohio State Univ. (1984)

SMITH, S.P.; ALLAIRE, F.L.:
Efficient Selection rules to increase non-linear merit: applica tion in mate selection. Genet. Sel. Evol. **17** (1985), 387-395

TIER, B.:
Computing inbreeding coefficients quickly. Genet. Sel. Evol. **22** (1990), 419-430

TUCHSCHERER, A.; MAYER, M.; REINSCH, N.:
Identification of gametes and treatment of linear dependencies in the gametic QTL-relationship matrix and its inverse. Genet.Sel. Evol. **36** (2004), 621-642

VAN ARENDONK, J.A.M.; TIER, B.; KINGHORN, B.P.:
Use of multiple genetic markers in prediction of breeding values. Genetics **137** (1994), 319-329

WANG, T.; FERNANDO, R.L.; VAN DER BEEK, S.; GROSSMAN, M.; VAN ARENDONK, J.A.M.:
Covariance between relatives for a marked quantitative trait locus. Genet. Sel. Evol. **27** (1995), 251-274

Received: 2006-06-21

Accepted: 2007-03-22

Authors' address
CHRISTINE BAES
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere (FBN)
Forschungsbereich Genetik und Biometrie
Wilhelm-Stahl-Allee 2, 18196 DUMMERSTORF, GERMANY

E-Mail: baes@fbn-dummerstorf.de

Prof. Dr. NORBERT REINSCH*
Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere (FBN)
Forschungsbereich Genetik und Biometrie
Wilhelm-Stahl-Allee 2, 18196 DUMMERSTORF, GERMANY

*Corresponding author, E-Mail: reinsch@fbn-dummerstorf.de

Arch. Tierz. **50** (2007) 3

APPENDIX A

From page 634 in TUCHSCHERER ET AL. (2004):

$$d_i^* = (1 - \mathbf{Q}_i^k \hat{\mathbf{G}}_i \mathbf{Q}_i^{k'})$$

For paternal gamete $i=Pg$,

$$= 1 - \begin{bmatrix} T_{(Pg)} & (1-T_{(Pg)}) & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{G}_{s(i)} & \mathbf{G}_{s(i)d(i)} \\ \mathbf{G}_{d(i)s(i)} & \mathbf{G}_{d(i)} \end{bmatrix} \begin{bmatrix} T_{(Pg)} & 0 \\ (1-T_{(Pg)}) & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$= 1 - \begin{bmatrix} T_{(Pg)} + (1-T_{(Pg)}) f_{sire} & T_{(Pg)} f_{sire} + (1-T_{(Pg)}) \end{bmatrix}$$

$$= 1 - \left[T_{(Pg)} + (1-T_{(Pg)}) f_{sire} T_{(Pg)} + (T_{(Pg)} f_{sire}) + (1-T_{(Pg)})(1-T_{(Pg)}) \right]$$

$$= 1 - \left[{T_{(Pg)}}^2 + (1-T_{(Pg)}) f_{sire} T_{(Pg)} + T_{(Pg)} f_{sire} (1-T_{(Pg)}) + (1-T_{(Pg)}) \right]$$

$$= 1 - {T_{(Pg)}}^2 - (1-T_{(Pg)})^2 + 2T_{(Pg)}(1-T_{(Pg)}) f_{sire}$$

$$= 1 - {T_{(Pg)}}^2 - (1-T_{(Pg)})^2 - 2T_{(Pg)}(1-T_{(Pg)}) f_{sire} + 2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire}) - 2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire})$$

$$= 2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire})$$

because ${T_{(Pg)}}^2 + (1-T_{(Pg)})^2 + 2T_{(Pg)}(1-T_{(Pg)}) f_{sire} + 2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire}) = 1.$

Therefore for animal a,

$$\mathbf{D}_a = \begin{bmatrix} 2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire}) & 0 \\ 0 & 2T_{(Mg)}(1-T_{(Mg)})(1- f_{dam}) \end{bmatrix}$$

and

$$\mathbf{D}_a^{-1} = \begin{bmatrix} \dfrac{1}{2T_{(Pg)}(1-T_{(Pg)})(1- f_{sire})} & 0 \\ 0 & \dfrac{1}{2T_{(Mg)}(1-T_{(Mg)})(1- f_{dam})} \end{bmatrix}$$

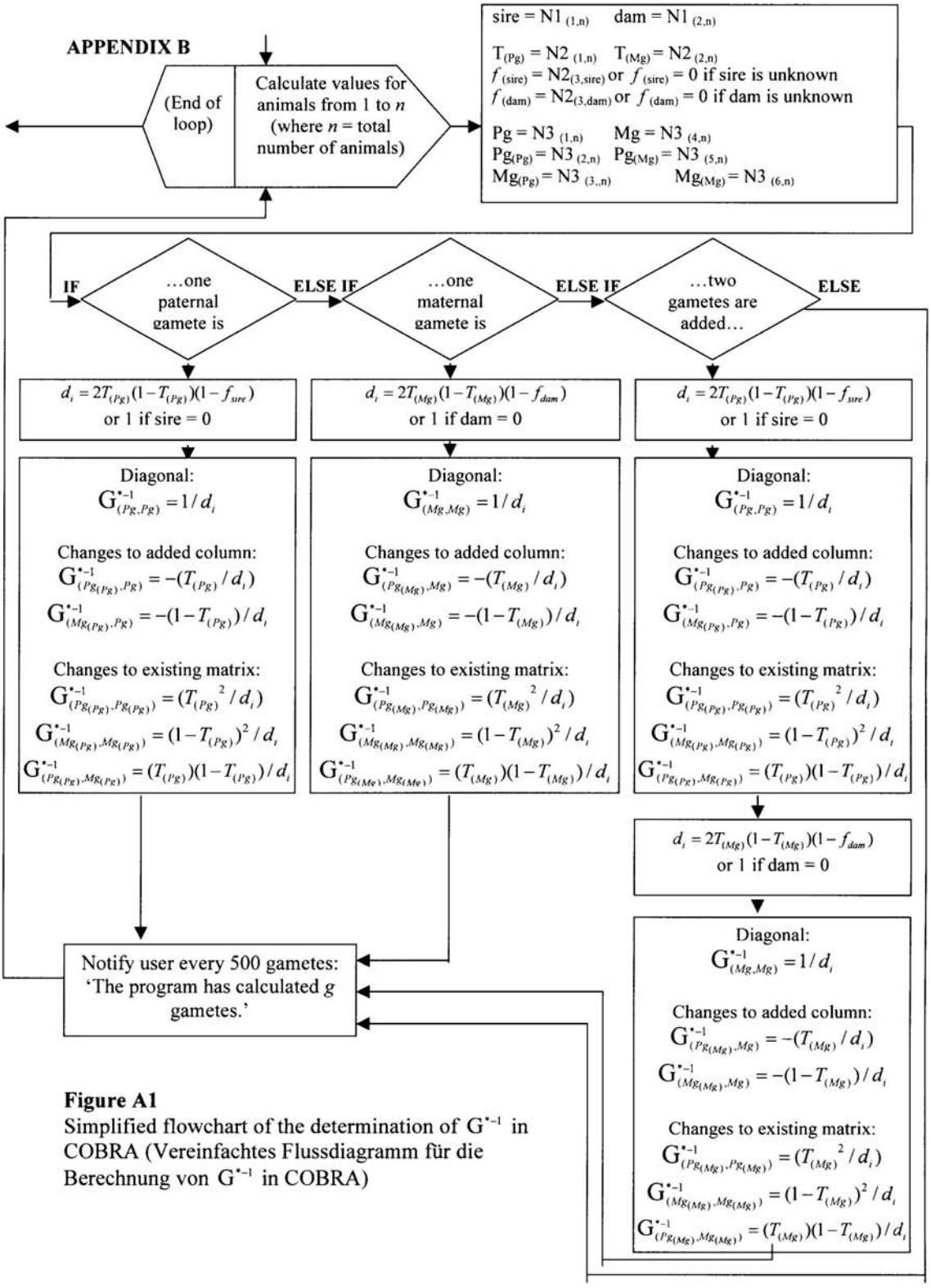

Figure A1
Simplified flowchart of the determination of G^{*-1} in COBRA (Vereinfachtes Flussdiagramm für die Berechnung von G^{*-1} in COBRA)