Chapter 1

Basic Building Blocks and Structure of Animal Breeding Programs

What Do We Mean By Breeding Strategies?
- Tactics designed to integrate new technologies and to improve old ones, for the purpose of maximizing performance of existing stock (Charles Smith).
- Integration of the components of a breeding program into a structured system for genetic improvement, with the aim to maximize an overall objective.

General aim for animal breeding strategies:
Obtain future generations of animals that will produce more efficiently under future production circumstances.

Basic Principle of making genetic progress in a population
Mate the “best” to the “best” and do that as quickly as possible.

Genetic Gain / Yr = genetic superiority of selected parents / generation interval
= intensity * accuracy * genetic st. dev. / generation interval

Some Questions
- How to find/identify the “best”?
- “Best” for what?
- What are the limits to use of only the “best”?
- How can we shorten the generation interval?
- What are the limits?
- How can a breeding company make a profit from this?
- “Breeding is a business” Lush, 1945
- How do technologies enter into this?

Basic Components of a Successful Breeding Program/Strategy
- Breeding Goal/Objectives
- Trait recording, Performance testing, Breeding value estimation
- Selection
- Mating
- Product Development and Dissemination
- Reproductive technology
- Molecular genetic technology
- Mating/Crossbreeding
- Improved Commercial Production

Genetic and Economic Aspects of Animal Breeding Strategies

Chapter I Introduction
- Building blocks and structure of animal breeding programs
- Methods to model/evaluate breeding programs/strategies
- Design of breeding programs/strategies
- Marker-assisted Selection

BREEDING STRATEGIES

Basic Components of a Successful Breeding Strategy
- Breeding Goal or Objectives - where should we go?
  - Which traits must be improved? - Economic traits
  - How important is each trait? - Economic values
  - Focus on improvement of Economic efficiency/profit
  - Consider (future) consumer demands
- Trait recording, Performance testing, Breeding value estimation
- Identify animals with “best” genetics - relative to breeding goal
  - Trait performance recording and testing programs
  - which traits should be recorded and on which animals?
  - field recording
  - performance test stations
  - nucleus herds
  - pedigree registration
- Genetic Evaluation
- Selection Index (Total merit index)
Components of a Successful Breeding Strategy (cont’d)

- Selection and mating
  - Use best animals to breed next generation → genetic improvement
  - How many and which animals should we select?
  - How should we mate them?
  - Should reproductive technology be used to increase # progeny/parent?
  - Balancing rate of genetic gain and inbreeding (and cost)

- Product Development and Dissemination
  - Program for marketing and distribution of superior genes into the commercial sector
    - Progeny testing, AI
    - Multipliers

- Mating/Crossbreeding
  - Optimize combinations of genetic material in commercial animals

Typical Structure of Livestock Breeding Programs

Dairy Cattle Progeny Testing Program

Basic steps in the design of breeding programs (Harris ’84)
1) Describe the production system(s)
2) Formulate the objective -simplified and comprehensive- of the system
3) Choose a breeding system and breeds
4) Estimate selection parameters and (discounted) economic values
5) Design an animal evaluation system
6) Develop selection criteria
7) Design matings for selected animals
8) Design a system for expansion - dissemination - of genetic superiority
9) Compare alternative programs

Development of Breeding Strategies Summary

- Integration of the components of a breeding program into a structured system for genetic improvement, with the aim to maximize an overall objective (genetic gain, market share).
- Evaluate opportunities for improving upon current strategies.
- Evaluate the potential of new technologies.
  - How can they best be incorporated into current strategies?
  - Can their benefits best be capitalized on in a redesigned breeding structure?

Breeding Strategies - Summary

What tools are necessary to develop optimum strategies?

- Quantitative genetics theory
  - Predicting response to selection, selection index, inbreeding, etc.
- Systems analysis
  - Predicting and optimizing response in overall objective
- Common sense
- An open mind
Chapter 2

Stochastic Methods to Model Breeding Programs

2.1 Introduction

The objective of genetic improvement of livestock is to enhance the genetic level for traits of interest in a population through genetic selection such that some overall goal is achieved or enhanced. The overall goal can usually be described in economic terms (e.g. maximize profit per animal per year) and will be discussed further in chapter 7.

There are many factors that determine the success of a breeding program. These include design and implementation issues. In this course, we will primarily focus on factors related to the design of genetic improvement programs, which include factors such as population size, numbers of animals to select, criteria for selection, etc.. Because of the number of factors involved, the number of alternative programs is numerous. However, ultimately only one program can be implemented; animal breeders don’t have the luxury of trying out different options and then deciding which one to go with. Thus, we need some other means of deciding a priori which breeding program will maximize our overall objective. This requires the ability to model breeding programs and to predict outcomes from alternative breeding programs. Furthermore, if a good understanding can be developed of the impact alternative design factors have on program outcomes, this will lead to the development and choice of better breeding programs. The development of this knowledge and associated methods and tools are the focus of this course.

2.2 Quantitative Genetic Model

Because most traits of interest in livestock are multifactorial in nature, i.e. affected by a potentially large number of individual genes along with environmental factors, quantitative genetic theory has become the primary basis for the development of methods to develop, model, and evaluate alternative breeding programs. The basis of this theory is the infinitesimal genetic model (Falconer and Mackay, 1996). The purpose of this section is to briefly review this theory as a basis for developing methods to model breeding programs.

The quantitative genetic model for the phenotype of animal $i$ is:

\[ y_i = \mu + g_i + e_i \]  

(2.1)

where $\mu$ is an overall mean (or sum of fixed effects), $g_i$ is the animal’s genetic value, and $e_i$ it’s random environmental effect. For the purposes of the majority of this course, we will assume we are dealing with additive traits such that $g_i$ refers to the additive genetic or breeding value.
Variables $g_i$ and $e_i$ are assumed normally distributed with means zero and standard deviations $\sigma_g$ and $\sigma_e$. Strictly, these assumptions hold for $g_i$ only for an unselected (base) population and both the mean and variance will change as a result of selection, as will be described later on in the course.

With the exception of the sex chromosomes, which we will ignore for the moment, all animals carry two copies of every gene. One copy is inherited by random sampling from the two copies carried by the male parent (sire) and the other copy is inherited by random sampling from the two copies carried by the female parent (dam). It follows that the additive genetic value of an offspring, $g_o$, can be partitioned into three sources, and modeled as follows:

$$g_o = \frac{1}{2} g_s + \frac{1}{2} g_d + g_m$$

(2.2)

where $g_s$ and $g_d$ are the additive genetic values of the sire and dam and $g_m$ is the Mendelian sampling contribution. The Mendelian sampling contribution reflects the random selection of copies of parental genes. Since genes are inherited at random from the parents, the average values of $g_m$ over a large number of progeny is expected to be zero.

Mathematically, it is said that the expectation of $g_m$, $E(g_m)$, is zero. But for any particular individual, $g_m$ has a real value which varies between individuals. The range of values of $g_m$ is determined by its variance, which in the absence of inbreeding, is expected to be

$$E(\sigma_{g_m}^2) = \frac{1}{2} \sigma_{g_0}^2$$

(2.3)

where $\sigma_{g_0}^2$ is the initial genetic variance in the population prior to any selection. The reason for noting the requirement that there be no prior selection in the population will become clear later in the course.

With inbreeding, the expected variance of Mendelian sampling terms is reduced by a factor $[1 - \frac{1}{2}(F_s + F_d)]$, where $F_s$ and $F_d$ are the inbreeding coefficients of the sire and dam. Thus:

$$E(\sigma_{g_m}^2) = \frac{1}{2} \left[ 1 - \frac{1}{2}(F_s + F_d) \right] \sigma_{g_0}^2$$

(2.4)

### 2.3 Stochastic Models for Evaluation of Breeding Programs

The simple quantitative genetic models described in the previous paragraph can be used to simulate a breeding program and evaluate its outcomes. Simulations in animal breeding can be divided into three types:

1) stochastic simulation (or sometimes called Monte Carlo simulation)
2) deterministic simulation
3) combination of stochastic and deterministic simulation.

With stochastic simulations in animal breeding, which will be described here, a population of animals is simulated by generating records for each animal in the population by random sampling from pre-defined distributions which are determined by the rules of inheritance and
origins of environmental effects imposed on the model. A model for stochastic simulation of a breeding program is schematically represented in Figure 2.1. The steps involved are described in further detail in what follows.

**Figure 2.1** General schematic of a stochastic simulation of a breeding program with *t* time periods and *m* replicates.

1. Generate a base population of parents.

```
\[ \downarrow \]
```

2. Generate progeny of defined family structure.

```
\[ \downarrow \]
```

3. Perform genetic evaluation to obtain selection criteria.

```
\[ \downarrow \]
```

4. Rank animals on selection criteria.

```
\[ \downarrow \]
```

5. Select animals, following defined rules.

```
\[ \downarrow \]
```

6. Mate parents and generate individual progeny. If *time < t*

```
\[ \downarrow \text{if } time = t \]
```

7. Output or store results. If *replicate < m* Go to next replicate.

```
\[ \downarrow \text{if } replicate = m \]
```

8. Output mean and variances of results and/or stop program.

### 2.3.1 Generating Base Population Parents

A base population is generated according to the rules of inheritance and structure of the population defined by the program control variables. For example, if the phenotype of a single trait, explained by the simple additive inheritance model plus a random environment effect, is

\[ y_i = \mu + g_i + e_i \]
and there are $n_m$ males and $n_f$ females in the base assumed to be randomly selected, unrelated, and non-inbred, then the effects for an animal in the base population could be defined by the following programming steps:

**Do for each animal i:**

1. $r = \text{random number from normal distribution with mean 0 and variance 1}$

2. $g_i = r \ast \sigma_{g_o}$ where $\sigma_{g_o}$ is the additive genetic st. deviation in the base population.

3. $r = \text{new random number from normal distribution with mean 0 and variance 1}$

4. $e_i = r \ast \sigma_e$ where $\sigma_e$ is the standard deviation of environmental effects.

5. $p_i = \mu + g_i + e_i$, where $\mu$ is the pre-defined population mean, i.e. a constant.

6. Store $p_i$, $g_i$; and $e_i$

This can be repeated for all animals in the base population. In order to enable the construction of a pedigree file, animals should be given a unique identification number. The simulation can be extended to include other genetic effects, such as dominance or systematic environmental effects such as age, herd or year. Virtually all programming languages have a random number generator or an associated library of subroutines containing a routine for random number generation. For those programming in PASCAL, C or FORTRAN, many useful subroutines are described in Press et al. (1986).

### 2.3.2 Generating Progeny

Once parents are generated, mating pairs are allocated and progeny generated. Recalling from equation 2.1 that the phenotype of progeny $k$ of male parent $i$ and female parent $j$ is

$$y_{ijk} = \mu + \frac{1}{2} g_{s_i} + \frac{1}{2} g_{d_j} + g_{m_{ij}} + e_{ijk}$$

(2.5)

where $g_{s_i}$ and $g_{d_j}$ are the known additive genetic values of the sire and dam, $g_{m_{ij}}$ is the Mendelian sampling contribution for individual $k$ and $e_{ijk}$ is the environmental effect. The contributions of $g_{m_{ij}}$ and $e_{ijk}$ are obtained for each progeny in turn by sampling from a random normal distribution with mean 0 and variance 1 and multiplying the random number by $\sigma_{g_o}$ or $\sigma_e$, where

$$\sigma^2_{g_o} = \frac{1}{2} \sigma^2_e$$

in the absence of inbreeding, or

$$\sigma^2_{g_o} = \frac{1}{2}(1 - \frac{1}{2} F_{s_i} - \frac{1}{2} F_{d_j}) \sigma^2_e$$
in the presence of inbreeding, where $F_s$ and $F_d$ are the inbreeding coefficients of the two parents. Fixed effects can then be added to $p_{ijk}$ according to the structure specified by the design.

### 2.3.3 Deriving the Selection Criterion

The selection criterion, such as the phenotypic record, a selection index, or BLUP evaluation, would be estimated for each simulated animal as if in real life. A subroutine of the program would be written to perform the evaluations. The nature of the selection criterion will determine the amount of data to be stored. For example, a selection index involving only collateral relatives would not require the parental records to have been stored, whereas animal model BLUP evaluation would require all animals and relationships back to the base population to be stored. In contrast to selection indexes, BLUP evaluation will be expensive for computing time because of the iterative nature.

Selection index or BLUP requires defined variances of traits for single trait evaluation and variance/covariance matrixes for multiple traits. Usually these would be set to the base population values, though false values may be given deliberately if estimation of sensitivity to parameter for BLUP is under investigation. If relationships back to the base generation are included, BLUP automatically allows for change in genetic variance due to selection (see Chapter 5).

With selection indexes, the appropriate variance/covariance among traits and relatives at each generation are required. A decision will therefore have to be taken as to whether to use constant parameters over time or to allow them to change. When the same set of parameters is used over time it seems logical to use the parameters from the base population, which were also used in simulating the data. In real life, the base population parameters can only be estimated and it might therefore be interesting to investigate the consequences of using other than the true parameters. Population parameters will change over time as a result of selection. These changes can be allowed for in constructing the selection index. In that case a method is needed to obtain the parameters at each point in time. The parameters could be estimated from the phenotypic and the true additive genetic values ($g_{ijk}$, $g_s$, $g_d$). This, however, would not be possible in real life and hence would not give realistic results. Alternatively, parameters could be estimated using phenotypic records or changes in parameters could be predicted from the selection strategy. Interpretation of the results will obviously depend on the assumptions made.

### 2.3.4 Selecting and Mating Animals for Breeding

In order to produce the next generation of offspring, one needs to define the method of selecting the animals to be used as parents and the procedure used in mating the selected parents. In the previous step, the selection criterion has been estimated for all candidates for selection. Truncation selection is commonly used for selection, in which the animals with the highest value for the selection criteria are selected. This requires that males and females are separately ranked in order of merit for the selection criteria. Efficient ranking routines are available in most
language libraries. Apart from the method of selection, the user has to specify the number of animals to be selected and the category of animals, which are eligible for selection. One might, for example, restrict the selection to animals of one particular age class only or have no restriction other than that animals need to be old enough to be able to reproduce. In the latter case, selection will be across age groups and it is important to specify up to what age animals are eligible for selection.

In the absence of restrictions on selection, selection is simply a process of designating the required number of top ranking animals as parents. With complete assortative mating, the top ranked male is allocated to the \( n \) top ranked females, the second ranked male to the next \( n \) females and so on; where \( n \) is the number of females per male. With random mating, each selected female is allocated a random deviate, and the females are then ranked on the random deviate and mating proceeds as above.

An advantage of stochastic simulation is that restrictions can be imposed on selection and mating. Common examples would be restrictions defining the maximum number of full and half sibs that can be selected as parents, and restrictions that full and half sibs may not be mated together. The imposition of restrictions may make some animals ineligible for mating so that more animals must be available for mating than indicated by the defined proportions to be selected.

2.3.5 Inbreeding Coefficients

Traditional methods of estimating inbreeding coefficients of individual animals by tracing path coefficients, or directly from a complete relationship matrix rapidly become time consuming and expensive of storage space as population sizes and number of generation's increase. With this method it was often impractical to estimate inbreeding coefficients in stochastic simulations. Several algorithms have been developed, however, for efficiently deriving inbreeding coefficients from a pedigree file (e.g. Tier, 1990). Use of these algorithms reduces computer time 10-100 fold compared to traditional methods. An additional trick is to recognize that all full sibs have the same inbreeding coefficient so that only one member of the family needs to have the coefficient estimated. Even so, calculation of inbreeding coefficients can still be expensive of computing time when simulating several thousand animals in each of several generations.

2.3.6 Completing the Cycle

Once mating pairs are allocated, progeny can be produced and the cycles repeated until the desired number of time periods has been achieved. At this point, summary statistics can be printed or stored, and the next replicate started. The number of replicates required will depend principally on the required accuracy of estimates of response and variance of response, which are largely dependent on the size of the population and the number of generations simulated. Large populations have low variance of response and therefore require fewer replicates for a given level of accuracy.
Stochastic simulations are often used to validate deterministic simulations. In this case it is desirable to have very accurate estimates of output parameters to estimate biases in the deterministic program. Typically, with smaller populations, several hundreds to 1000 replicates are run. But when using stochastic simulations to evaluate alternative breeding programs, very small differences between alternatives are rarely of practical interest so that often fewer, say 100, replicates can suffice. In practice the number of replicates required can be determined once a few initial runs have indicated the variance to be expected between runs for a particular size and type of population.

2.3.7 Multiple Trait Simulations

Multiple trait simulations are a little more difficult, but can be achieved by deriving the $n$ uncorrelated principal components of the genetic and environmental variance covariance matrix among the $n$ traits, generating random deviates for each principal component in turn and then back-transforming these to obtain random deviates for the original traits. Alternatively, an approach using Cholesky decomposition of the original variance covariance matrixes can be used which has advantages in terms of computing ease and time. The Cholesky decomposition approach is explained in Appendix C and some examples of simulating correlated traits and records for related individuals are given by Van Vleck (1993). These same methods can deal with simulations involving other covariances among random variables, such as $g \times e$ covariance and additive x dominance genetic covariances.

2.3.8 Finite locus models

In the previous, the genetic component was modeled as a normally distributed variable, using the infinitesimal genetic model. This model assumes that the trait is affected by a large number of unlinked loci, each of small effect. Stochastic models also allow the modeling of a more realistic genetic architecture of the trait by simulating individual loci and their placement on chromosomes within the genome. These so-called finite locus models require specification of the number of loci, the number and length of chromosomes that these loci are located on, and their position (in centi-Morgans, cM) on these chromosomes. Then, the following parameters must be specified for each locus:

1) Locus position - loci could be positioned on chromosomes at random by sampling from a uniform distribution, or evenly distributed across the genome.
2) Number of alleles.
3) Allele frequencies in the base population – these could be set to be equal or sampled from some distribution
4) Genotypic effects associated with each genotype - these can, for example, for a locus with two alleles B, b, be based on the standard single locus genetic model with genotypic values of $+a_l$, $d_l$, and $-a_l$ for genotypes BB, Bb, and bb at locus $l$ (Falconer and MacKay, 1996). Genotypic values assigned to each locus could be sampled from an assumed distribution of gene effects, such as from a gamma distribution (e.g. Hayes and Goddard, 2003), in an attempt to reflect reality. In addition, epistatic effects could be allowed for by assigning genotypic effects to combinations of genotypes at multiple loci.
For the base population, alleles at locus $l$ for individual $i$ can then be assigned by drawing two random numbers $u$ from a uniform $(0,1)$ distribution. For example, for a locus $l$ with allele frequency $f_j^l$ for alleles $B_j (j=1, \ldots, m_l)$, allele $j$ is assigned if $\sum_{k=1}^{j-1} f_k^l < u < \sum_{k=1}^{j} f_k^l$. This random sampling of alleles assumes the base population is in Hardy-Weinberg and gametic phase equilibrium (Falconer and MacKay, 1996).

The genetic value of individual $i$ then is the sum of genetic effects at each of the $q$ loci:

$$g_i = \sum_{l=1}^{q} g_i^l,$$

where $g_i^l$ is the genotypic value at locus $l$ for individual $i$, which is based on the simulated genotype of for locus $i$ and the genotypic value that is associated with this genotype.

If all loci are unlinked, progeny genotypes at each locus can be simulated by randomly drawing one of the two alleles of the sire and one of the two alleles of the dam. If loci are linked, recombination must be allowed for. Consider the two haplotypes for a parent in Figure 2.1.

![Figure 2.1. Simulation of Mendelian inheritance with linked loci with recombination in intervals 23 and 56.](image)

To create a progeny from this parent, the first step is to simulate the production of two gametic chromosomes through meiosis. This can be simulated as follows:

1) Starting with the first interval, 12, the probability of recombination ($r_{12}$) or not ($1-r_{12}$) is drawn from a uniform normal distribution. If $u[0,1] < r_{12}$, then a recombination takes place and we end up with the following two recombinant haplotypes: $Q_1 q_2, q_3 Q_4 q_5 Q_6 q_7 Q_8$ and
\[ q_1, Q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_8, \] since all alleles downstream from the cross-over are switched. If \( u[0,1] > r_{12} \) then the parental chromosomes stay intact.

2) Proceed to the next interval and draw presence or absence of a recombination event in that interval: if \( u[0,1] < r_{23} \) then there is a recombination event and we end up with the following two recombinant haplotypes (assuming there also was recombination in interval 12): \( Q_1, q_2, Q_3, Q_4, Q_5, Q_6, Q_7, Q_8 \) and \( q_1, Q_2, q_3, q_4, q_5, q_6, q_7, q_8 \). If there is no recombination event, then the haplotypes generated in step 1 remain intact.

3) Proceed through all intervals consecutively as described above.

Once a pair of recombinant gametes has been created, a random one of the two is sampled to generate the progeny. A similar procedure is used to generate the other parental chromosome.

Note that this method assumes that recombination events in adjacent intervals are independent (no interference – Haldane mapping function). If there is interference, probabilities of recombination in interval \( i \) must be adapted, depending on presence or absence of a recombination event in interval \( i-1 \).

### 2.4 Advantages and Disadvantages of Stochastic Models

Stochastic simulation depends on relatively simple rules determining inheritance from one generation to the next, along with description of the criteria on which all animals will be selected for breeding. Thus, for a given degree of complexity of the breeding program, stochastic simulations are often relatively easy to write compared to the deterministic models that will be described later. In addition, stochastic models allow alternative genetic models to be evaluated, while deterministic models are primarily restricted to the infinitesimal genetic model. However, see Chapter 12 for deterministic models with individual genes along with an infinitesimal polygenic component.

With stochastic simulation, the result of any one run reflects random sampling events so that to obtain the mean expected response, many replicate runs must be made; but this also allows the variance of the response to be estimated. Because each animal in the population is individually identified, stochastic programs can take up a large amount of storage space and involve a very large number of mathematical operations for every run. This, combined with the need to replicate, means that stochastic programs take much longer, often very much longer, to run than deterministic programs.

Stochastic simulation also does not provide much insight into the impact of various factors on response to selection and does not lend itself easily to optimization of breeding programs. Hence, in the remainder of this course, the main focus will be on deterministic models, to facilitate an understanding of the factors that affect the outcomes of breeding programs. With the tremendous increases in computing power, however, stochastic models have become more and more attractive and used for the evaluation and analysis of breeding programs in both research and practice.