Lecture Notes on
Construction of Genetic Linkage Maps

Piet Stam

Laboratory for Plant Breeding
Wageningen University
2. Construction of Genetic Linkage Maps

2.1. Introduction

A linkage map is a representation, in the form of a table or a graphic, of the position of genes (or markers) within a linkage group. The map positions are inferred from estimates of recombination frequencies between genes. The earliest linkage maps contained, in most cases, only a few morphological markers, mostly representing genes for which mutant phenotypes were available. Usually only two or three marker genes were segregating simultaneously, so that linkage maps had to be assembled from a number of different experimental populations. Nevertheless, even before the advent of molecular markers, so called multi-marker lines were constructed for a number of plant species (barley, maize, Arabidopsis). These multi-marker lines carried mutant alleles at a number of loci, so that these loci would segregate simultaneously in the offspring (F$_2$) when crossed with a wild type (= non-mutant).

The problem with multi-mutant lines is that these are very often ‘crippled’ phenotypes, since most mutants have a deleterious effect (e.g. chlorophyll deficiency, dwarfism, malformation of flowers, deterioration of development, etc.) so that such lines are hard to maintain, and, additionally, the phenotype of one mutant can often only be clearly recognized in absence of another mutant. Figure 2.1. schematically represents the stepwise assemblage of a linkage map based on a number of different crosses.

![Figure 2.1](image)

With advent of molecular markers the problems inherent to multimarker lines disappeared. At the DNA level the amount of variation is so vast that in a particular cross between genotypes over thousand markers may be segregating at the same time. Although the degree of DNA polymorphism between genotypes of a crop species varies from species to species, the number of DNA markers that can be obtained to discriminate between genotypes is in
principle unlimited. Figure 2.2. shows, as an example, a recently published linkage marker map of tomato chromosome 1.

Figure 2.2 An example of a marker linkage map. Chromosome 1 of tomato. The map contains 139 markers. Left of chromosome bar: map distances in centimorgans. Right of chromosome bar: marker names. Markers at the same horizontal position, as well as markers in a closed box map to the same position (perfect co-segregation). From Haanstra et al., 1999.
2.2. **Assessment of linkage groups**

When a large number of markers (1000, say) are segregating simultaneously in a mapping population and these markers are to be placed on a linkage map, the first step is to group the markers into linkage groups. (Ideally the number of linkage groups equals the haploid chromosome number.) Linkage groups are established by considering all estimates of recombination frequencies. If two markers are significantly linked (by LOD value) they belong to the same linkage group. A computerized search through all pairs of markers, using a certain threshold value of the LOD score, will then produce a grouping of the markers. Within a group the markers ‘stick together’. Figure 2.3 shows the assignment of markers to linkage groups. Notice that a conservative threshold of the LOD score may lead to more linkage groups than the haploid chromosome number. When constructing a linkage map from scratch (with a set of markers that have not previously been assigned to linkage groups or chromosomes) it is always wise to stay ‘at the safe side’ by using a conservative (= high) LOD threshold. This will prevent that groups of markers on different chromosomes are incorrectly assigned to a single linkage group.
2.3. Ordering of markers within a linkage group

Finding the correct order of markers within a linkage group and calculating the map distances between them is the major job of software packages for the construction of linkage maps. For most users of mapping software this process unrolls ‘behind the screen’ and does not bother them. However, to get a taste of what is going on below the surface, we pay some attention to the computational aspects of map construction.

The most simple ordering problem arises when we have recombination estimates for each pair of a set of three markers (see Figure 2.4). Let

\[
\hat{r}_{AB} = 0.15; \hat{r}_{AC} = 0.10 \quad \text{and} \quad \hat{r}_{BC} = 0.07. 
\]

It then takes little more than a glance to see that A - C - B is the 'best' order.

Remark: We will return later to the question: "What is 'best' in this context?", "Is there an objective criterion by which we can qualify a given order of markers as 'superior' over another order?". For the moment we will assume that for any given order a 'fitting criterion' can be calculated.

As the number of markers increases the number of possible orders increases rapidly. For \( n \) markers the number of orders equals \((\frac{1}{2})n!\). Table 2.1 gives an impression of the magnitude of this value.
Table 2.1

<table>
<thead>
<tr>
<th>n</th>
<th>$(\frac{1}{2})n!$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$1.8 \times 10^6$</td>
</tr>
<tr>
<td>20</td>
<td>$1.2 \times 10^{18}$</td>
</tr>
<tr>
<td>50</td>
<td>$1.5 \times 10^{24}$</td>
</tr>
<tr>
<td>100</td>
<td>$4.6 \times 10^{57}$</td>
</tr>
</tbody>
</table>

Suppose that 20 markers (a very modest number) are to be ordered. Suppose further that a modern high speed computer can calculate the fitting criterion of a million orders per second. It would take this computer 38.57 millennia to consider all possible orders. It is beyond imagination how long it would take to do the job for 100 markers. Obviously, an exhaustive search for the best fitting order is not feasible. For this reason most mapping software packages apply shortcuts and 'tricks' to arrive at the best order or an order that is 'almost best'. One of the 'cheapest' algorithms to achieve this, known as a greedy algorithm, is the stepwise build-up of a map by adding one marker at a time. This works as follows. One starts with two markers, A and B, say. Then the third marker (C) is to be added. The orders CAB, ACB and ABC are compared and the best fitting one is chosen, say ACB. When adding the next marker (D) the orders DACB, ADBC, ACDB and ACBD are compared, and, again, the best one is chosen. This continues until all markers have been placed in the sequence. This procedure requires that only for

$$\sum_{i=1}^{n} \frac{n(n+1)}{2}$$

different orders the fitting criterion has to be calculated.

Remark: The JoinMap package, which uses this 'greedy' algorithm, has several refinements to this general scheme. For example, the order in which markers are added to the sequence is not random, but depends on the 'amount of information' a marker contains. In addition, after a marker has been added, a 'local reshuffling' can be applied in order to prevent that the previous sequence will not be changed any more and the algorithm is trapped in a local optimum from which it cannot escape.

2.3.1. What fitting criterion?

As outlined above, for any given order a number has to be calculated which indicates the 'goodness-of-fit' of that order. Different software packages use different criteria. For the sake of completeness we mention a few of the most widely used ones.

(a) Full likelihood.
This requires that the complete likelihood of the observed genotypes is calculated, and at the same time the recombination frequencies between adjacent markers be estimated. Computationally this is 'expensive', resulting in a slow performance.

(b) Sum of adjacent distances (SAD).
This method uses the estimates of the recombination frequencies between adjacent markers.
The order with the smallest value of SAD is regarded as ‘best’. In the example given earlier the SAD values of the three possible orders are

\[
\begin{align*}
\text{ABC:} & \quad \hat{r}_{AB} + \hat{r}_{BC} = 0.15 + 0.07 = 0.22 \\
\text{ACB:} & \quad \hat{r}_{AC} + \hat{r}_{BC} = 0.10 + 0.07 = 0.17 \\
\text{CAB:} & \quad \hat{r}_{AC} + \hat{r}_{AB} = 0.10 + 0.15 = 0.25.
\end{align*}
\]

By this criterion ACB is the best order, which is also intuitively obvious.

A problem with SAD as goodness-of-fit is that it ignores that the different estimates may vary widely in precision. This may be due to different population sizes that were used to obtain the estimates or due to incomplete information (caused by dominance, for example) for some markers.

(c) Sum of adjacent LOD scores (SAL).

This method calculates the sum of the LOD scores for each adjacent marker pair in the sequence. The order with the highest value of SAL is regarded as best. This method circumvents, to a certain extent, the problems we mentioned for method (b) (= SAD). LOD score is, in a sense, the ‘amount of linkage information’. Thus, the SAL criterion considers all pieces of information in a sequence and adds these. From a number of alternative sequences the most informative (with respect to linkage) is considered ‘best’.

(d) Number of observable recombination events.

Some of the population types used for linkage analysis, like BC (backcross) and DH (doubled haploids) allow direct counting of the number of recombination events (in the meioseses that gave rise to the gametes). This number depends on the assumed order of markers. As an example consider the following data in a three-point backcross \((AaBbCc \times aabbcc)\).

<table>
<thead>
<tr>
<th>genotype</th>
<th>parental gamete</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{AaBbCc}</td>
<td>\text{ABC}</td>
<td>28</td>
</tr>
<tr>
<td>\text{aabbcc}</td>
<td>\text{abc}</td>
<td>28</td>
</tr>
<tr>
<td>\text{AabbCc}</td>
<td>\text{Abc}</td>
<td>7</td>
</tr>
<tr>
<td>\text{aaBbCc}</td>
<td>\text{aBC}</td>
<td>7</td>
</tr>
<tr>
<td>\text{AaBbcc}</td>
<td>\text{ABc}</td>
<td>3</td>
</tr>
<tr>
<td>\text{aabbCc}</td>
<td>\text{abC}</td>
<td>3</td>
</tr>
<tr>
<td>\text{AabbCc}</td>
<td>\text{AbC}</td>
<td>12</td>
</tr>
<tr>
<td>\text{aaBbcc}</td>
<td>\text{aBc}</td>
<td>12</td>
</tr>
</tbody>
</table>

For counting the total number of recombination events we consider the parental gametes and write these down for each of the three possible orders.

(1) Order A-B-C

<table>
<thead>
<tr>
<th>gamete</th>
<th>number</th>
<th>number of rec. events</th>
</tr>
</thead>
<tbody>
<tr>
<td>\text{ABC}</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>\text{abc}</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>\text{Abc}</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>\text{aBC}</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>\text{ABc}</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>\text{abC}</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>\text{AbC}</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>\text{aBc}</td>
<td>12</td>
<td>2</td>
</tr>
</tbody>
</table>

Total number of recombination events =

\[(28 \times 0) + (28 \times 0) + (7 \times 1) + (7 \times 1) + (3 \times 1) + (3 \times 1) + (12 \times 2) + (12 \times 2) = 68\]
(2) Order A-C-B

<table>
<thead>
<tr>
<th>gamete</th>
<th>number</th>
<th>number of rec. events</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACB</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>acb</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>ACb</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>acB</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Acb</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>aCB</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>AcB</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>aCb</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Total number of recombination events = 50

(3) Order B-A-C

<table>
<thead>
<tr>
<th>gamete</th>
<th>number</th>
<th>number of rec. events</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>bac</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Bac</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>bAC</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>BAc</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>baC</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>BaC</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>bAc</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

Total number of recombination events = 58

We see that the order A-C-B corresponds to the smallest number of recombination events. For that reason we decide that A-C-B is the best order.

Notice that this criterion is quite similar to SAD. (Verify that the SAD criterion also yields A-C-B as the best order by calculating \( \hat{r}_{AB}, \hat{r}_{AC}, \text{ and } \hat{r}_{BC}. \))

We have not yet explained why an order with less recombination events is ‘better’. The reason is that the probability of a recombination event in any interval (of finite length) is always less than the probability of no recombination event (remember: \( r \leq 0.5 \)).

(e) **Weighted Least Squares (WLS)**

The WLS criterion uses the estimates of all (pairwise) recombination frequencies of a sequence of markers simultaneously, i.e. not only those between adjacent markers. The

---

**Figure 2.5.** A ‘spring model’ illustrating the weighted least squares method for calculation of map distances from a series of pairwise recombination estimates. Each spring represents an estimated recombination frequency. Blocks (loci) are interconnected by springs. The ‘stiffness’ or elasticity of a spring corresponds to the weights assigned to the individual estimates. The method calculates the position of the blocks in the equilibrium status of the system.
mathematical details of the method are beyond the scope of this course, but to get a taste of what the method does, the illustration of Figure 2.5 is useful. (See Stam, 1994 for a detailed account of WLS.) The blocks in Figure 2.5 are the markers. The blocks are connected by springs. The length of a spring corresponds to the estimated recombination frequency. The elasticity of a spring corresponds to the precision of the estimate (the ‘weight’ of the estimate). The WLS method calculates the equilibrium state of the system of connected blocks and the total ‘internal tension’ of the system (the sum of ‘pulling’ & ‘pushing’ forces exerted on all blocks).

The WLS method is computationally ‘expensive’ as compared to SAD, SAL or counting. However, it is much faster than maximum likelihood. One of the advantages of WLS is that it can combine recombination estimates that were obtained in different experiments. (WLS was originally designed to assemble maps from a series of experiments in which partially overlapping sets of markers are segregating; see Jensen & Jorgenson, 1975, Hereditas) The well known JoinMap package uses the WLS method to find the best fitting order and simultaneously calculates the map positions corresponding to that order.

2.3. Genetic Mapping Functions

A genetic mapping function (mf) describes the mathematical relation between recombination frequency (r) and map distance (x). The map distance between two markers is the average number of recombination events per meiosis between those markers. The exact relation \( f \) between \( r \) and \( x \) \((r = f(x))\) depends on the degree on interference between recombination events, in adjacent intervals. In the absence of interference recombination events in adjacent intervals (and as a result, in any pair of intervals) are independent. Denoting recombination and no recombination by 1 and 0, respectively, independence means

\[
P_{00} = (1 - r_x)(1 - r_y), \\
P_{10} = r_x(1 - r_y), \\
P_{01} = (1 - r_x)r_y, \text{ and} \\
P_{11} = r_xr_y.
\]

See Figure 2.6

![Figure 2.6](image_url)
Notice that under the assumption of no interference

\[ r_{x+y} = r_x(1-r_y) + (1-r_x)r_y = r_x + r_y - 2r_xr_y. \]

The above is a mathematical formulation of the fact that recombination frequencies are non-additive. (The above relation is often written as

\[ (1-2r_{x+y}) = (1-2r_x)(1-2r_y) \]

Map distances are, of course, additive; when the mean number of recombination events in two intervals are \( x \) and \( y \), respectively, then the mean number of events in the joint interval equals \( x+y \).

It can be shown that in the absence of interference the mapping function reads

\[ r = \frac{1}{2}(1-e^{-2x}), \]

which is known as Haldane’s m.f., in honour of the famous geneticist J.B.S. Haldane.

Interference means that recombination events in adjacent intervals interfere. The occurrence of an event in a given interval may reduce or enhance the occurrence of an event in its neighbourhood. Positive interference refers to the ‘suppression’ of recombination events in the neighbourhood of a given one. Negative interference refers to the opposite: enhancement of clusters of recombination events. In most organisms positive interference has been observed, to various degrees of intensity. Positive interference results in less double recombinants (over adjacent intervals) than expected on the basis of independence of recombination events. Referring to Figure 2.6 it results in

\[ P_{11} < r_xr_y. \]

Another consequence of positive interference is

\[ r_{x+y} > r_x + r_y - 2r_xr_y. \]

The observation that in many organisms less double recombinants occur than expected on the assumption of independence has led to the conclusion that positive interference is a widely occurring phenomenon.

The exact mathematical relation between \( r \) and \( x \) in the case of interference depends on the degree/strength of the interference. A mapping function that (empirically) appears to be fairly well describing the strength of interference in many organisms is Kosambi’s m.f. It reads

\[ r = \frac{1}{2} \tanh \left( 2x \right). \]
It can be shown that for adjacent interval the addition rule reads

\[ r_{x+y} = \frac{r_x + r_y}{1 + 4r_x r_y} \]

when Kosambi’s m.f. holds.

Figure 2.7 shows the graph of both Haldane’s and Kosambi’s m.f. In addition to these two m.f.’s a whole family of genetic mapping functions has been designed, each corresponding to a given degree of interference. However, only Haldane’s and Kosambi’s m.f. are the most widely used ones. Most software packages for linkage mapping allow optional application of Haldane’s or Kosambi’s m.f. to ‘translate’ recombination frequencies in map distances. From Figure 2.7 it is easily seen that Haldane’s m.f. results in longer maps than Kosambi’s m.f.

\[ \text{Fig. 2.7. Haldane’s (H) and Kosambi’s (K) mapping functions. Note that Kosambi’s m.f. leads to a shorter map distance for a given recombination frequency.} \]

**Remember:** The unit of map distance is one centimorgan (1 cM). A distance of 100 cM means that over that distance an average of 1 recombination event (per chromatid) occurs.

### 2.4. Errors in genotype scoring

Errors in genotype scoring, or genotype classification, will lead to incorrect estimates of recombination frequencies. We illustrate this effect by means of a simple numerical example (see Figure 2.8). Suppose that in a backcross the loci A/a and B/b are segregating; suppose further that the true recombination frequency between these loci is 10% \((r = 0.10)\). Suppose further that in 10% of the cases the genotype at the B/b-locus is misclassified as indicated in the right hand column of Fig. 2.8.
We see that based on the incorrect observations the estimate of $r$ equals 0.18, almost twice as large as the true value! This illustrates that misclassification of genotypes leads to an overestimation of recombination frequencies. This is not only the case for a BC or DH population, but holds for all types of mapping populations.

If in the above example not only locus $B/b$ but also locus $A/a$ had been misclassified at the same rate, the overestimation of $r$ would even have been larger. When the error rate for both loci equals $x$, the resulting estimate of $r$ equals

$$\hat{r} = r + 2x(1-x)(1-2r).$$

Figure 2.8 The effect of errors in genotype scoring, illustrated with a numerical example. Gametes are supposed to be produced by a double heterozygote. Left: correct counts; $r = 0.10$. Right: counts resulting from 10% misclassification; $r = 0.18$. Recombinant gametes are boxed.

Figure 2.9. The effect of errors in genotype scoring for different error rates ($x$). Note the considerable overestimation of recombination frequency at even moderate error rates.
This relation has been plotted in Figure 2.9 for various values of the error rate \( x \). Notice that the overestimation (= bias) relative to the true value of \( r \) increases as \( r \) decreases.

The value of the ratio

\[
\frac{2x(1-x)(1-2r)}{r} = \text{bias, true value}
\]

the ‘inflating factor’, has been plotted in Figure 2.10 for several values of \( x \).

![Figure 2.10 The ‘inflation factor’ (relative bias) caused by errors in genotype scoring for various error rates (x). An inflation factor of 4.0 means a fivefold overestimation.](image)

The consequences of an error rate as small as 2% \( (x = 0.02) \) are enormous, especially for the construction of high density maps where the average distance between markers is below 5 cM, say. From Figure 2.10 we see, for example, that an error rate of 2% will inflate a true value of \( r = 0.02 \) by almost 200%! The resulting map length will be inflated by a similar percentage.

Figure 2.11 shows two calculated maps, in which the second (E) is based on the same data set as the first (C), but 3% of the genotype readings has been replaced by the wrong one \( (A \rightarrow a \text{ and } a \rightarrow A) \).

Figure 2.10 clearly indicates that the inflation increases as the map density increases.
The above considerations emphasise the necessity to avoid errors in ‘genotyping’ as far as possible. Errors can, of course, never be avoided completely; practices such as standard double checking of gel readings and automation of gel reading may greatly increase the overall correctness of genotype scores.

Figure 2.11. Two linkage maps constructed with simulated data. Three percent of the correct genotype scores were replaced by an incorrect score, to produce the data set with errors. Note the inflation of map length caused by the incorrect scores.
Several mapping software packages have a module that enables the user to identify ‘suspect’ genotype scores. Although these error detection routines are helpful, they are not the ultimate solution, because a ‘suspect’ genotype may be a correct one even when it is very unlikely (‘suspectness’ in error detection algorithms are based on likelihood).

2.5. Final remarks

The outcome of a mapping experiment depends on the composition of the sample population. The larger the mapping population, the more confidence we have in the estimates of recombination frequencies and map distances. For most purposes populations of size in the range 80-400 are used. (Remember that the population type also influences the standard errors of the estimates.) It is good to realise that, for example, a replicate experiment with 100 RILs will result in a (slightly) different map. Figure 2.12 shows the simulated result of replicate sampling of an F2 population and the best map corresponding to each sample. Although the variation between these maps with respect to marker order is nil, the resulting total map length as well as the inter-marker distances are quite variable. This demonstrates that the ‘ultimate true map’ does not exist. A calculated map is the best statistical approximation, given the sample population.

Figure 2.12 Graphical representation of ten linkage maps. For each of these maps a data set was generated by stochastic simulation. The true map (length 100 cM) on the left corresponds to the recombination rates between adjacent markers used in the simulation. Due to the variation in the sample composition, the resulting maps are different. Notice that for a given marker pair the estimated map distance may vary considerable (see lines connecting marker pairs between maps).
The awareness that calculated map positions are estimates, not certainties, has created the need to visualise, in one way or another, along the picture of a map the uncertainty of the marker positions. Some software packages have such an option, although these are mostly (computer) time consuming.