Crossover counts and likelihood in multipoint linkage analysis

E. A. Thompson,
Department of Statistics, GN-22,
University of Washington,
Seattle, WA 98195.

SUMMARY

For large numbers of genetic loci, jointly tested to determine their order along a chromosome, likelihood methods become infeasible due to the very large numbers of discrete alternative hypotheses (locus orderings) whose likelihoods must be separately evaluated. A method to order loci according to the criterion of minimising the obligatory crossover count is therefore proposed. A branch-and-bound algorithm implementing this proposal has been programmed; the properties of this algorithm are investigated. The statistical properties of the proposed method are also considered. It is shown to be consistent under wide conditions including arbitrary locus spacings, variable amounts of information per locus, and some patterns of interference. The relationship between the minimum crossover order and the maximum likelihood order is discussed. For fully informative gametes, and tight linkage, there is a virtual equivalence of the two criteria. For looser linkage there remains a close relationship.

1. Introduction

Recent developments in linkage analysis are towards the multilocus mapping of the highly polymorphic markers with codominant alleles, these then forming the framework for mapping more complex traits. With the current technology of restriction fragment polymorphisms, very large numbers of loci may be typed for each segregation, particularly in the area of plant and animal genetics where highly informative matings/crosses can be planned. The loci considered are often known to be syntenic, linked, or even tightly linked: the questions are of locus order. For the ordering problem, the analysis of three loci jointly, with the extra dimension it provides over the simple recombinant/nonrecombinant status of a pairwise observation, provides for far greater accuracy in the assessment of locus order (Thompson, 1984). This increased accuracy has two sources; at the lowest level gametes may only be pairwise informative, second they may be jointly informative but the information scored pairwise, and third they may be jointly informative and jointly scored. Although the intermediate case of joint data scored pairwise can provide good information for ordering (White et al., 1985) the third data type is the most efficient. Further, for larger numbers of loci, where the samples are of gametes and individuals informative for different subsets of the loci, joint analysis of the cosegregating loci must provide still greater benefits over the pairwise analysis of recombination rates. However, full likelihood analyses are highly computationally intensive.

There are thus two alternate simplifications for a full multipoint linkage analysis. One is first to consider data pairwise: linkage of a trait with single markers is often a useful first approach, although combination of results over different linked markers requires care. Buetow et al. (1985) have considered multidimensional scaling of maximum likelihood estimates of pairwise recombination rates to obtain a multilocus order. The other simplification is to consider data jointly but to score recombination events rather than compute a full likelihood. For the problem of ordering multiple markers, it seems that more may be lost by failure to take account of loci jointly than by failing to extract all the statistical information. That is, a simple scoring method based on joint data may prove to produce better results than a full likelihood analysis of pairwise data. While a full joint likelihood analysis must be the most statistically efficient procedure, the computations involved are large and lengthy. A scoring method can indicate the regions of the hypothesis space, in terms of both recombination values and loci orderings, which should be more fully investigated.

A heuristic scoring criterion for estimation or inference is never as satisfactory as an estimate based on the likelihood. However, the greater ease of computing such a heuristic estimate lead to its widespread use. The statistical justification for such a method must lie in the accuracy of results, and/or in the approximation of such results to those obtained via some valid method of inference, such as by maximum likelihood estimation. A useful analogy is in the use of minimum evolution as an approximation to the likelihood solution for divergence of allele frequencies in populations under random genetic drift (Cavalli-Sforza and Edwards, 1964). In that problem also, there are a large number of discrete hypotheses of primary interest -- the alternative evolutionary tree forms. The estimation of quantitative parameters, such as times of divergence, is of secondary importance. It is important to investigate the relationship between scoring estimates and inferences based upon the likelihood function, and the models and/or types of data for which the methods give similar inferences (Thompson, 1986). Although scoring methods for ordering loci date back to Sturtevant (1913), there has been no detailed investigation of their statistical properties. Such investigation is necessary, if such methods are to become widely used either in place of, or as a preliminary to, full likelihood analyses.

2. Locus ordering by branch-and-bound on a crossover count criterion

The intrinsic information for linkage analysis in data on any segregation consists of the grandmaternal/grandpaternal origins of alleles on each of the two gametes received by an offspring (figure 1). We shall therefore consider data in a partially preprocessed form of received gametes scored for each locus as grandmaternal (2), grandpaternal (1) or unknown (0). For a complex trait and/or two-generation data alone there may be "partial" knowledge, in the sense of

probabilities dependent upon the types of other offspring of the same mating or on sibs of the parents, either of which may provide partial information on parental phase. However, for three-generation data on traits with codominant alleles unknown phase will result either from homozygosity in a parent, or from two identically heterozygous parents providing a heterozygous offspring (figure 1). In the former case grandparental origin of an allele in intrinsically unknowable; i.e. is fully "unknown". The latter case occurs with low probability for polymorphic loci, and can be avoided in planned matings between lines. Thus, although in general information may be lost by scoring as unknown the grandparental origins of offspring alleles with origins not fully determined, the loss should be small in all practical cases.

A scoring method can, of course, use any score criterion, including the -log-likelihood itself: the optimal locus order is, by definition, that which achieves the minimal score. However, in order to facilitate searching one requires a score function that, for all hypothesised locus orderings, provides an easily computed non-negative increase on inclusion of data on additional loci. In order for the optimal order under the criterion to bear any relationship to the true order, we naturally require some function which we anticipate will "probably" be "small" under the true order. The criterion which we investigate is the total number of recombination events (crossovers) implied for the data by a given ordering of loci. The statistical properties of the procedure resulting from this choice of score function will be investigated below; we consider first the ordering algorithm.

Lemma: The number of implied recombinants is additive over gametes, and provides a non-negative score increase when, in any hypothesised ordering, data on an additional locus is included.

Proof:

We shall simply specify the score increases for a given gamete; these are functions of the grandparental origins for the previously ordered loci of the gamete, which are assumed separately scored for each gamete of the sample (see above). Suppose that the data for some subset of the loci under a given subset of n loci ordered without loss of generality as

$$A_1, A_2, \ldots, A_n$$

is, for some gamete,

$$d_1, d_2, \ldots, d_i, d_{i+1}, \cdots d_n \quad d_i = 0, 1 \text{ or } 2$$

and the $(n+1) sup th. $ locus is inserted after locus $A sub i $ (figure 2).

If $ d sub n+1 =0$, there is no change in score.

If $d sub n+1 != 0 $, let x be the first non-zero $d sub j $ (if any) amongst $d sub i, d sub i-1 , d sub i-2 , ...$, and y be the first non-zero $d sub j $ (if any) amongst $d sub i+1 , d sub i+2 , ... , d sub n $; if either of these does not exist let $x~(y)$ be 0. Then:

If $x=y=0$, there is no change in score.

If $x=0,~~y != 0$, the score is increased by 1 if and only if $d sub n+1 != y.$

If $x != 0,~~y=0$, the score is increased by 1 if and only if $d sub n+1 != x.$

If $x != 0,~~y != 0$, the score is increased by 2 if and only if $x=y != d sub n+1$.

Thus changes in score are non-negative; they can be zero.

The lemma is proved.

We now consider determination of optimal ordering of loci using the classic branch-and-bound algorithm for tree-structured searches (see e.g. Lawler and Wood, 1966; Knuth, 1968). (In pursuing the analogy of section 1, it is of interest that branch-and-bound has also been used in the context of minimum evolution estimates of evolutionary trees (Hendy and Penny, 1982)). The tree structure employed is of branch points corresponding to the insertion of information on additional loci, the alternative branches corresponding to alternative positions for insertion. Thus at level i, when the $i sup th.$ locus is to be inserted, every node has i branch points (figure 3). It is not necessary that the order for insertion be the same on every subtree, but of course all subtrees from a given level must correspond to insertion of the same set of loci.

The search involves evaluation of the score increase involved on each branch of the tree, and comparison with a previously determined value which is the current minimum found for a full order (at some tree tip). Since the score increase on each arc is non-negative, but need not be positive, we can terminate search of any subtree as soon as the value at the root of that subtree is not strictly smaller than the current comparison value. Thus for an efficient search, evaluating explicitly the smallest possible number of partial orderings, we require

(a) that some order with low total score is found early in the search, to give a small value for comparison, and

(b) that high score increases are acquired at low levels in the tree, to reduce the chance of having to search many subtrees to high levels.

An important aspect of any branch-and-bound algorithm is thus a specification of order of insertion of items (in our case, loci). Objective (b) is achieved by choosing for insertion at any level, that locus that has highest average immediate score increase over the alternative positions for insertion. Objective (a) is then achieved by searching the alternative subtrees at each level in increasing order of immediate score increase at that level. It is an open question how far it is worth reordering insertion of loci on different subtrees (figure 3). While the overall ordering at low levels can be very important, at high levels the procedure seldom leads to changes, and so the increased computation involved cannot provide improvement.

A branch-and-bound program has been implemented (in C under UNIX) in accordance with the above, and tested on simulated data. Some brief summary statistics are given in Table 1, but some formulae are also illuminating. Note that in a tree of l loci the total number of full and partial orderings, or the total potential number of order evaluations, is

$$\tfrac{1}{2}\,(3! + 4! + 5! + \ldots . l!) = \sum_{i=3}^{l} \prod_{j=3}^{i} j$$

while the minimum number of evaluations to achieve the first ordering is

$$3 + 4 + 5 + \ldots . + l = (l+3)(l-2)/2.$$

In many cases, the number of evaluations made is less than twice this minimum (Table 1), showing that for clear data (informative tightly linked loci) the optimal order is not only immediately found, but more important is almost immediately so recognised by the algorithm. The proportion of tree searched is closely related to the number of changes that are made before the optimum is found, although of course there is no exact relationship. This number of changes is never large; it is often zero. Six was the largest value found in any run. It is dependent on the number of gametes, as well as on the number of loci and the level of information; the overall means given in Table 1 are intended only as an indication of order of magnitude. It is also the case, qualitatively, that optima achieved only after several changes have a higher frequency of not being the true locus order.

In considering the minimum number of gametes to infer the order of equally spaced loci note that the probability of an observed recombinant between two adjacent loci is $ r (1-h) sup 2 $, where r is the recombination fraction and h the probability that a locus is *not* informative, assumed the same for all loci. Thus the approximate number of gametes to be sampled before there is an observed recombinant in every one of the $(l -1)$ gaps between adjacent loci is

$$\frac{1}{r(1-q)^2} \sum_{j=1}^{l-1} \frac{1}{j}\,.$$

Loci cannot be ordered until there is at least a recombinant distinguishing the members of every adjacent pair (Thompson, 1984). The above formulae is an approximation, since loss of information at a locus simultaneously affects two adjacent gaps, but it provides a useful lower bound on the number of gametes it is reasonable to consider. Another consideration is of the smallest scores that are likely to give reliable orderings. Table 1 shows that high scoring optima are almost invariably true, although it is also the case that untrue optima tend to have higher than average

score for that set of locus characteristics. Where there are a sufficient number of gametes for an expected score of 50 or greater one can place high reliance on the optimum order; those few incorrect optima with values over 50 all corresponded to situations in which the average was less than 50. For equally spaced and equally informative loci we can consider the smallest total score that has a given probability of some score attributable to every gap; again a necessary condition for order information to be present. For fully informative loci, this is simply an occupancy problem (Feller, 1968 Pp.102-5); the probability that each of the $(l -1)$ gaps is "occupied" by at least one of the items in a total score of s required recombinants is

$$p_0(s, l-1) = \sum_{j=0}^{l-1} (-1)^j ((l-1)!/j!(l-1-j)!)(1 - \frac{j}{l-1})^s$$

which can be approximated (for l large) by $ e sup {- lambda }$, where $ lambda = (l -1)e sup {-s/(l -1)}$ (Feller, 1968).

The runs summarised on Table 1 correspond to a "clear" and a "less clear" case for each of 5, 8 and 10 loci. Other runs, for larger numbers of loci and for unequally spaced loci, were also made. For clear-cut cases (recombination rates less than 0.1, and no missing data) up to 15 loci can be handled with ease, but for other cases time constraints precluded the gathering of samples of useful size. For unequally spaced loci, no clear pattern emerged. Cpu times (given for a VAX780 but almost identical to those for a MASSCOMP) are almost proportional to the number of gametes times the number of (full and partial) orders evaluated. For a given number of loci this is expected, since each position evaluation must be made for every gamete. More surprising was the fact of virtual constancy of the time per gamete per order evaluation across the range from 5 to 10 loci and across the range of levels of information. Although times varied slightly, the mean for every group was between 0.00009 and 0.00011 seconds/gamete/order-evaluated. Thus it would be efficient to first scan gametes, discarding those with no obligatory recombinants, either because of tight linkage or because of high levels of missing data. This was not done in the runs summarised in Table 1. Times are also dependent on the proportion of tree searched, but this dependence is not great.

While fuller analyses of performance are required, these preliminary results confirm the both the efficiency of recombination scores in determining a true locus ordering, and the efficiency of branch-and-bound in determining the order with optimal score. Comparisons with other methods, such as a full joint linkage analysis, are not possible, since there is no program available which will deal readily with more than four loci simultaneously.

3. Consistency of the Minimum Crossover Scoring Method

In a statistical analysis of any estimation procedure, consistency is an important consideration -- as the number of informative gametes becomes large, is the correct order inferred? Note that the evolutionary tree analogy would here suggest that consistency might fail. Felsenstein (1978) showed that disparate evolutionary rates lead to inconsistency of the "minimum evolution" tree, although when overall evolutionary rates are small consistency is more likely to obtain. Here, by contrast, we shall show consistency not only for tightly linked and/or equally spaced loci, but far more widely. Consistency depends only on the expected value of the score function under the alternative discrete hypotheses. If this is necessarily minimal for the true order, then asymptotically the true order will be (with probability tending to one) that of minimum score. The scoring criterion is then consistent, giving asymptotically correct results.

Consider, in particular, the case of four loci, and assume that gametes are fully informative for these four loci. Assume that the true order is $ABCD$, with inter-locus recombination rates r, s, and t (figure 4). Assume also absence of interference, so that these three parameters determine the probabilities of all recombination events. Then there are eight possible gamete types. The score and the probabilities of these under the true order are given in Table 2. The score which each assigns to each of the other eleven locus orderings may be likewise computed. Table 3 gives again the score for the true order, $ABCD$, and the score **differences** between all eleven other orders and the true order. The probabilities of the eight events, and the score differences which

they provide, give immediately the expected score differences, which are also given in Table 3. We see that these are non-negative for all alternative hypotheses (locus orderings) regardless of the locus spacing (i.e. for any r, s, and t less than 0.5). Under our assumptions the scoring algorithm is consistent for four loci.

Not all these assumptions are in fact required. In particular, the assumption of fully informative gametes is for convenience only and can be relaxed (see below). Also, complete absence of interference is not required. We need only sufficient restriction on the event probabilities to ensure strictly positive expected score differences for all untrue orders. From Tables 2 and 3, the necessary and sufficient conditions for consistency for four informative loci are

$$q_D + q_{AB} > q_C + q_{BD}, \quad q_A + q_{BC} > q_B + q_{BD}, \quad q_A + q_D > q_B + q_{C,} \tag{1a}$$

$$q_D + q_{AB} > q_B + q_{BC}, \quad q_A + q_{AB} > q_C + q_{BC}, \quad 2q_{AB} > 2q_{BD}, \tag{1b}$$

$$q_A + q_{AB} + q_{BC} > q_C + 2q_{BD}, \quad q_D + q_{AB} + q_{BC} > q_B + 2q_{BD}, \tag{1c}$$

and

$$2q_{AB} + q_C > q_D + q_{BC} + q_{BD}, \quad 2q_{AB} + q_A > q_B + q_{BC} + q_{BD}. \tag{1d}$$

Sufficient conditions are thus

$$q_{AB} > q_{BC} > q_{BD} \quad \text{and} \quad min(q_A, q_D) > max(q_B, q_C). \tag{2}$$

It is therefore sufficient for consistency that, amongst events when the loci segregate three-and-one, single crossover events have higher probability than double crossovers, and that separately for the two-and-two segregating gametes, a single crossover should have higher probability than doubles, and doubles than triples. If loci are equally spaced, these conditions must always be satisfied for any level of interference. Positive interference in fact will enhance the chance of consistency, in that it decreases the probability of double and triple crossovers relative to single ones. Note, however, that the final sufficient conditions are not satisfied by loci with spacing of different orders of magnitude; there is in general no reason why $ q sub A $ should exceed the double crossover rate $ q sub C $. (In the absence of interference it will do so only if $ r(s-1)(1-t)~>~(1-r)st $.) Thus the necessary and sufficient conditions (1) imply consistency in the absence of interference, and for equally spaced loci regardless of interference, but neither of these cases subsumes the other. Equation (1) will also provide for consistency in many cases intermediate between these two special cases.

In the absence of interference, there are, of course, many alternative ways to write the expected scores of Table 3 in terms of pairwise recombination fractions. One way that will be illuminating for the general case below is shown in Table 4. Each order has an expected score which is the sum of the expected score for the loci A, B and C in the 3-locus order obtained by deletion of D, plus an "increment" which depends only on the immediate neighbour(s) of D in the 4-locus order. Tedious but simple algebra will confirm that the scores of Table 3 can indeed be written in the form shown in Table 4. We use this idea of an expected score for n loci, plus an increment, to extend the above now to arbitrary numbers of loci, assuming absence of interference.

Framework; Suppose we have loci whose true order is

$$A_1, A_{2,} A_{3,} \cdots, A_n, A_{n+1}, \ldots.$$

and suppose that the first n have been placed in some hypothetical order, giving rise to some expected score relative to the true order. Consider now the addition of locus $ A sub n+1 $, and suppose that for a given hypothesised order it is placed between loci $ A sub i $ and $ A sub j $ in the previous hypothesised order, where 1 <= i < j <=n. Let $ r sub i,j $ denote the pairwise recombination rate between loci $ i $ and $ j $, and suppose that $ 0 < r sub i,j < 1/2 $. Consider a single fully informative gamete, and, for this gamete, the expected score increase on addition of locus $A sub n+1 $ into the count of crossovers, for any placement of this additional locus. Scores are measured relative to the true ordering of the same loci, but we wish to consider the increment in score due to addition of the $(n+1) sup th. $ locus. We consider therefore, for a

hypothesised order $ H sub n+1 $ loci, the expression

$$(S(H_{n+1}) - S(T_{n+1})) - (S(H_n) - S(T_n)) \tag{3}$$

where S is the total expected score function, T denotes a true order, and $H sub n $ is the n-locus order formed by deletion of $A sub n+1 $ from the (n+1)-locus order $H sub n+1 $. Now for purposes of evaluation we may rewrite (3) as

$$(S(H_{n+1}) - S(H_n)) - (S(T_{n+1}) - S(T_n)). \tag{4}$$

In the absence of interference, the values $S(T)$ are easily evaluated, being simply the expected number of recombinations between adjacent loci, or

$$S(T_n) = \sum_{i=2}^{n} r_{i-1,i},$$

so that (4) reduces to

$$(S(H_{n+1}) - S(H_n)) - r_{n,n+1}. \tag{5}$$

Lemma; If $ A sub n+1 $ is located between $A sub i $ and $ A sub j $ with $ 1 <= i < j <= n $ then

$$S(H_{n+1}) - S(H_n) = 2\, r_{j,n+1}(1 - r_{i,j}) \tag{6}$$

and if, in the hypothesised ordering $A sub n+1 $ is a terminal locus with neighbour $A sub j $ with $ 1 <= j <= n $ then

$$S(H_{n+1}) - S(H_n) = r_{j,n+1} \tag{7}$$

Note; Note first that equations (5), (6) and (7) hold in the transition from three to four loci (Table 4). (It also holds trivially in the transition from two to three, but that example is not illuminating.) Our proof will use this framework of four loci, since it is shown that, in the absence of interference, we need consider only the two immediate neighbours of $ A sub n+1 $ in the new order, to obtain (6) and (7), while locus $ A sub n $ enters to provide the term $r sub n,n+1 $ for the comparison of true orders (5).

Proof;

Let $<A sub i >$ denote now the totality of all events in which $ A sub i $ segregates separately from $ A sub j$ and $ A sub n+1 $, and similarly for the other events with respect to these three loci (Table 5). Then the probabilities of these events combined over all other loci under the true order are as given in Table 5.

Now in comparing $H sub n+1 $ with $ H sub n $ the only score difference is an increase of 2 for all those events in the set $<A sub i , A sub j >$. Thus the expected score difference is as stated.

For the case where $A sub n+1 $ is terminal in $H sub n+1 $, we have a score increase of 1 for precisely those events in which $A sub j $ segregates separately from $A sub n+1 $; the probability of all such events is simply $r sub j,n+1 $.

The lemma is proved.

We have shown that for any fully informative gamete, for any number of loci, the true order minimises the expected score for that gamete. But since this argument is on a per gamete basis, we can now apply it to the informative loci for a given gamete. For any given gamete, the order that minimises the expected score is any order in which the informative loci for that gamete are correctly ordered. Hence consistency of the minimum scoring method is not dependent upon having fully informative gametes, but only on having loci that have strictly positive probability of being informative, independently of other loci.

4. Likelihood and crossover counts

The -log-likelihood itself is a consistent score function; but two consistent score functions need not necessarily provide identical results in analysis of an actual data set. A different question is thus of possible differences of inferred order provided by the likelihood and by the minimum cross-over criteria.

Consider first the maximum likelihood estimates of recombination fractions for the case of four loci, under any specified hypothesised order, on the basis of fully informative gametes, assuming absence of interference. Then the estimated recombination fractions r, s and t, between adjacent loci consist of combinations of three out of a total of six possible frequencies. Table 6 shows these maximum likelihood estimates, for the twelve possible locus orderings. The $ x sub A $ etc. denote the sample frequencies of the events <A> etc., whose probabilities $ q sub A $ etc. are given in Table 2. The maximised log-likelihood is (Thompson (1984))

$$N\ (\ h(r) + h(s) + h(t))$$

where

$$h(r)\ =\ r \log r + (1-r)\ log(1-r),$$

r, s and t are the maximum likelihood estimates, and N the total number of (fully informative) gametes in the sample. Thus in computing the log-likelihood difference between orders for which two of the estimates are in common (e.g. ABCD and ABDC, DCAB or CDAB; see Table 6), the larger log-likelihood will be for the order with the larger h-value for the third estimate. Since h is monotone decreasing in r (for r< 1/2), the order with higher log-likelihood will be that for which the third recombination estimate is smallest. But the estimates are precisely the score counts of recombinants; the lower scoring order will have higher log-likelihood. For three loci, this ensures that the maximum likelihood order is always that which requires fewest total recombinants (Thompson, 1984; see also Bishop 1985). For four loci there are, however, other cases, where comparisons must be made between orders having only one, or even no, estimated recombinations in common. Where there is one common estimate (e.g. ABCD with ADCB, CBAD, DBAC, or ACDB; see Table 6), the log-likelihood order will depend on the ordering of $ h(r)˜+˜h(s) $, where r and s are the estimates **not** held in common, while the score will depend on the ordering of $(r˜+˜s)$. While $h(r)˜+˜h(s)$ is close to monotone in $ (r+s)$ it is not exactly so (figure 5); thus there can exist data values for which the minimum score and maximum likelihood orders are not identical. For the remaining four orders (Table 6) there are no pairwise estimates in common. The log-likelihood for the order depends on $ h(r)+h(s)+h(t) $ while the score depends on $ r+s+t $; again the ordering by maximised log-likelihood need not coincide with that by minimum crossover score. On the other hand, the probability of a data set in which the maximum likelihood and minimum score orders do not coincide is very small. The orders which "compete" with the true order, are mainly those for which there are two estimates in common (see Table 3).

For tight linkage, there is an even closer relationship between likelihood and crossover counts. We consider a very large number of gametes, sufficient to ensure a crossover between ever pair of loci, but such that the *number* of crossovers in the total length remains bounded, since the total map length is small. Now let the number of crossovers between the $ (i-1) sup th $ and $ i sup th , $ loci under any assumed (not necessarily true) order $ pi $ be $ x sub i (pi)$, so that the maximum likelihood estimates are $ r sub i ˜=˜x sub i (pi)/N.$ Then the maximised log-likelihood for order $ pi $ is

$$LL(\pi) =\ \ N \sum_{1}^{l-1} h(x_i(\pi)/N)$$

which reduces to

$$\sum_{0}^{l-1} x_i(\pi) log(x_i(\pi)) + (logN - 1)(-\sum_{1}^{l-1} x_i(\pi)) + O(N^{-1})$$

or

$$LL(\pi)/log(N) \ \rightarrow \ - \sum_{1}^{l-1} x_i(\pi) = - \, S(\pi).$$

Thus as linkage becomes very tight, but samples sufficiently large to ensure identifiability of orders, the maximised log-likelihood for any order is determined by, and is a fixed multiple of, the minimum crossover score. Thus not only do both methods infer the same order (which in these circumstances has a high probability of being the correct order), but the two measures are equivalent for *any* order.

The above discussion requires fully informative gametes. For gametes that are equally informative about all loci, similar results appear to hold, although the formulae cannot be so neatly expressed. On the other hand, the above results are untrue where the information about loci is highly disparate. The minimum crossover rate order (counts adjusted to rates to adjust for different numbers of informative gametes) and maximum likelihood orders can then differ even for three loci (Thompson, 1984).

5. Discussion

As the number of loci involved in such analyses increases, it will become more important to have effective sorting and scoring methods to screen the very large numbers of discrete hypotheses (locus orderings) involved. We have seen that a branch-and-bound method search performs impressively on multi-locus simulated data. The advantage of such a deterministic method, over stochastic methods such as simulated annealing (for example), is that the optimum determined is necessarily the global minimum of the objective function. Scoring methods have also the general advantage of being more visibly dependent upon particular items of data: the precise gametes or segregations contributing to anomalous conclusions can be identified, and perhaps reassessed.

An alternative approach would be to use a tree search method over locus orderings, but to use minus the log-likelihood directly as the objective function. This is feasible only if the log-likelihood for an order can be very rapidly evaluated, but the computation of this likelihood involves maximisation over $(l - 1)$ recombination rates. For fully informative gametes, the estimates can be explicitly determined. For gametes which are highly informative, so that no two adjacent loci are unscored for any gamete, the three locus EM-approach (Thompson, 1984) allows conditional estimation of recombinants, and hence rapid determination of maximum likelihood estimates. For less informative gametes and/or loci the extended EM-approach of Lander et al. (1986) may provide rapid evaluation of the likelihood of any locus order. However, any iterative approach to parameter estimation must involve substantially more computation than the immediate evaluation of a minimum crossover score. Note also that the recombinant score is ideally suited to a tree-structured search, with the effect of addition of any locus in any position in any partial order being easily evaluated, and bounded below only by zero. If likelihoods are to be computed, the addition of information will again bound the change in -log-likelihood by zero, but since estimates of all recombination rates may change (particularly where loci are not highly informative) the -log-likelihood increase is not immediate. Nor will zero in general be a tight lower bound, and this may adversely affect the performance of a branch-and-bound procedure. It will be important to determine whether a non-zero lower bound can be predetermined, in order to allow the search to be curtailed at lower levels in the tree. Finally, the maximum likelihood estimation methods referred to here assume absence of interference. The minimum-crossover score method makes no such explicit assumption, and the four-locus results (section 3) show that this method should be robust to a variety of patterns of interference and "improved" by many patterns of positive interference. Overall, there seems to be more to be lost than gained by combination of the scoring method with simultaneous likelihood evaluations, although subsequent likelihood analysis of low-scoring orders may be a useful procedure.

For fully informative gametes, there is a close analytical relationship, although not complete equivalence, between the likelihoods for alternative locus orderings and scores based on the implied numbers of recombinants. However, when loci are differentially informative the low-

scoring order may differ from the maximum likelihood order (Thompson, 1984). Since it assigns "equal weight" to all observed recombinants, the crossover count criterion is also then prone to error. A method that counted recombinants with appropriate weighting might resolve this problem, but would require increased computation, and considerations of aspects akin to those involved in analysis of the likelihood function. Note that where a locus is highly uninformative, likelihood estimation also encounters difficulties. The variance of the natural sufficient statistics (the unobserved true numbers of recombinants underlying the data) are high for those intervals adjacent to a low-information locus (Thompson, 1984). Thus any estimation procedure relying on estimation of these underlying recombinants will be slow to converge, even where there is good information for order.

6. References

Bishop, D.T. (1985). The information content of phase-known matings for ordering genetic loci. *Genetic Epidemiology* **2,** 349-361.

Buetow K., Chakravarti A., Murray J. and Ferrel R. (1985). Multipoint mapping using seriation. *Am. J. Hum. Genet.* **37** (supplement), A190.

Cavalli-Sforza L.L. and Edwards A.W.F. (1964). Analysis of human evolution. In *Genetics Today* S.J. Geerts ed. (Proc. XI Int. Cong. Human Genetics; The Hague, 1963). New York; Pergamon Press.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications, Volume 1.* (3 rd. ed.) Wiley; New York.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology,* **27,** 401-410.

Hendy, M. D., and D. Penny. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences,* **59,** 277-290.

Knuth, D.E. (1968). *The Art of Computer Programming; Fundamental Algorithms,* Addison-Wesley; Reading, Mass.

Lander, E.S., Green, P., Barlow, A. and Newberg, L. (1986). A new and efficient approach to multipoint linkage analysis. *Am. J. Hum. Genet.,* **39** *(Supplement),* A161.

Lawler, E.L. and Wood, D.W. (1966). Branch and bound methods: a survey, *Operations Research,* **14,** 699-719.

Sturtevant, A.H. (1913). The linear association of six sex-linked factors in *Drosophila,* as shown by their mode of association. *J Exp Zool* **14,** 43−59.

Thompson, E.A. (1984). Information gain in joint linkage analysis. *IMA J of Math Appl in Med and Biol* **1,** 31−50.

Thompson, E.A. (1986). Likelihood and parsimony; comparison of criteria and solutions. *Cladistics* **2,** 43-52.

White, R., Leppert, M., Bishop, D.T., Berkowitz, J., Brown, C., Callahan, P., Holm, T. and Jerominski, L. (1985). Construction of linkage maps with DNA markers for human chromosomes. *Nature,* **313,** 101-105.

Figure captions.

Figure 1: Scoring of gametes on the basis of three generation data on a pedigree. The maternal and paternal gametes in an individual are each scored. In scoring at a given locus, an alleles from the a male grandparent is scored 1, from a female grandparent is scored 2, and one not identifiable is scored 0.

Figure 2: Increase in crossover score for a gamete on inclusion of information on an additional locus positioned at any specified point within the current partial order.

Figure 3: A part of a branch-and-bound tree structure; at each level information on an additional locus is incorporated.

Figure 4: The case of four loci; the situation providing the results of Tables 2 and 3.

Figure 5: Comparison of the functions $h(r)+h(s)$ and $r+s$; the first determines the log-likelihood and the second the crossover score, in the case where just two estimated recombination rates differ between the orders to be compared.

Keywords: Multipoint linkage, branch and bound, crossover counts, likelihoods, consistency.

Table 1; Performance of C-program for branch-and-bound locus ordering.

Characteristic	nloc=5		nloc=8		nloc=10	
Number of evaluations;						
potential maximum	75		23,115		2,018,765	
minimum number	12		33		52	
Proportion min:total possible	0.16		0.00143		0.000026	
Score for complete occupancy*						
with 90% probability	14.6		29.4		40.0	
with 99% probability	24.0		45.8		61.2	
Cpu time/gamete/order (secs)	$1x10 sup -4$ in all cases					
Typical cpu time for correct inference (secs)	0.3		2.4		4.6	
Data regime	$r=0.1$ $h=0.0$	$r=0.2$ $h=0.4$	$r=0.1$ $h=0.0$	$r=0.2$ $h=0.2$	$r=0.1$ $h=0.0$	$r=0.1$ $h=0.4$
Expected gametes for existent recombinants*	20.8	28.9	25.9	20.2	28.3	78.6
Mean score per gamete	0.37	0.47	0.76	1.23	0.86	0.92
Mean proportion of tree searched;						
if nchange=0	0.160	0.234	0.00165	0.00526	0.000043	0.000084
if nchange=1	0.280	0.275	0.00294	0.00870	0.000093	0.000217
if nchange=2	***	0.40**	***	0.01661	0.000113**	0.000337
if nchange>2	***	***	***	0.015**	0.000162**	0.000404**
Mean number of changes during search	0.25	0.5	0.2	0.9	0.4	1.7
Percentage of true orders amongst optimal scores						
over 150	***	100	100	100	100	100
50 to 149	100	98	100	97	100	88
30 to 49	100	70	88	50	92	50
15 to 29	93	40**	***	36**	***	20**

*: See equations and/or explanation in text

**: small samples

***: insufficient information

Table 2; Gamete events and their probabilities and scores under the true hypothesis ordering $ABCD$ figure 4.

gamete	event	score	description (cosegregating loci)	probability (under figure 4)	general probability
2222/1111	0	0	all together	(1-r)(1-s)(1-t)	$q sub 0 $
2111/1222	<A>	1	A segr. alone	r(1-s)(1-t)	$q sub A$
2122/1211	<B>	2	B segr. alone	rs(1-t)	$q sub B$
2212/1121	<C>	2	C alone	(1-r)st	$ q sub C$
2221/1112	<D>	1	D alone	(1-r)(1-s)t	$q sub D$
2211/1122	<AB>	1	AB/CD split	(1-r)s(1-t)	$q sub AB$
2112/1221	<BC>	2	AD/BC split	r(1-s)t	$q sub BC$
1212/2121	<BD>	3	AC/BD split	rst	$q sub BD$

Table 3; Crossover counts and expectations for four loci.
Scores are given relative to the order ABCD, which is assumed true for the purposes of computing the expectations.

Event	0	<D>	<C>	<AB>	<B>	<BC>	<A>	<BD>	expectation.
Null score	0	1	2	1	2	2	1	3	
Order:									
ABDC	0	+1	-1	0	0	+1	0	-1	t(1-2s)
DCAB	0	0	0	0	-1	+1	+1	-1	r(1-2s)
CDAB	0	+1	-1	0	-1	0	+1	0	(1-2s)(r+t-2rt)
ADCB	0	+1	0	+1	-1	-1	0	0	(1-2r)(s+t-2st)
CBAD	0	0	-1	+1	0	-1	+1	0	(1-2t)(r+s-2rs)
ACBD	0	0	0	+2	0	0	0	-2	2s(1-r-t)
DBAC	0	0	-1	+1	0	+1	+1	-2	r(1-2s)+s(1-2t)
ACDB	0	+1	0	+1	-1	+1	0	-2	t(1-2s)+s(1-2r)
ADBC	0	+1	-1	+2	0	-1	0	-1	2s(1-r)(1-2t)+t(1-2r)
BCAD	0	0	0	+2	-1	-1	+1	-1	2s(1-t)(1-2r)+r(1-2t)
BDAC	0	+1	-1	+2	-1	0	+1	-2	2s(1-r-t)+(1-2s)(r+t-2rt)

Table 4; Increments of expected score in the transition from three to four loci. The expected scores are given in Table 3, and for every 4-locus ordering it may be checked that the 4-locus expected score is the sum of the 3-locus score and the given increment. The results thus illuminate the lemma, the final column giving expressions in the more general notation of the multilocus case.

3-locus order		ABC	ACB	CAB	
3-locus expected score		0	s(1-2r)	r(1-2s)	
locus D inserted between;	score increment	new order	new order	new order	comment
end and C	0	ABCD	N/A	DCAB	$r sub CD $-t=0
end and B	(s+t-2st)-t	N/A	ACDB	CABD	$ r sub BD $-t
end and A	((r+s+t)-2(rs+rt+st) +4st)-t	DABC	DACB	N/A	$r sub AD $-t
A and B	2(s+t-2st)(1-r)-t	ADBC	N/A	BDAC	$ r sub BD $=(s+t-2st)
A and C	2t(1-(r+s-2rs))-t	N/A	ADCB	CDAB	$ r sub AC $=(r+s-2rs)
B and C	2t(1-s)-t	ABDC	ACDB	N/A	$r sub BC $=s

Table 5; Values of scoring when additional locus is inserted in the ordering. All events, orders etc. are specified only with respect to the three loci $ (A sub i , A sub j , A sub n+1)$: for details see text.

Event	Probability under true order $(A sub i , A sub j , A sub n+1)$	score under $H sub n $	score under $H sub n+1 $
$A sub n+1 $ interior;			
0	$(1-r sub i,j)(1-r sub j,n+1)$	w*	w
$<A sub i >$	$ r sub i,j (1-r sub j,n+1)$	w	w+1-1=w
$<A sub j >$	$ r sub i,j r sub j,n+1 $	w	w-1+1=w
$<A sub i , A sub j >$	$(1-r sub i,j)r sub j,n+1 $	w	w+1+1=w+2
$A sub n+1 $ terminal;			
0	$(1-r sub j,n+1)$	w	w
$<A sub j >$	$r sub j,n+1 $	w	w+1

*: w here denotes the score for the given event in the order $H sub n $. It differs, of course, between event groups, and between events within a group such as $<A sub i >$, but the comparison between $H sub n $ and $H sub n+1 $ is the same for all members of the group.

Table 6; Patterns of recombination estimates for 4-locus orderings.		
Interval	1	2
Estimate of	r	s
Order:		
ABCD	$r sub 0 = x sub A + x sub B + x sub BC + x sub BD$	$r sub 2 = x sub B + x sub C + x sub AB + x
ABDC	$ r sub 0$	$t sub 1 = x sub B + x sub D + x sub AB +
DCAB	$t sub 0$	$r sub 1 = x sub A + x sub C + x sub AB +
CDAB	$t sub 0$	$ t sub 2 = x sub A + x sub D + x sub AB +
ADCB	$t sub 2$	$t sub 0$
CBAD	$r sub 2$	$r sub 0$
ACBD	$r sub 1$	$r sub 2$
DBAC	$t sub 1$	$r sub 0$
ACDB	$ r sub 1$	$ t sub 0$
ADBC	$ t sub 2$	$ t sub 1$
BCAD	$r sub 2$	$r sub 1$
BDAC	$t sub 1$	$t sub 2$