GENOMICS 1, 174–181 (1987)

MAPMAKER: An Interactive Computer Package for Constructing Primary Genetic Linkage Maps of Experimental and Natural Populations

ERIC S. LANDER,*,†,‡ PHILIP GREEN,§ JEFF ABRAHAMSON,*,† AARON BARLOW,*,† MARK J. DALY,*,† STEPHEN E. LINCOLN,*,† AND LEE NEWBURG*,†

*Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142; †Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139; ‡Harvard University, Cambridge Massachusetts, 02138; and §Human Genetics Department, Collaborative Research, Inc., 2 Oak Park, Bedford, Massachusetts 07130

Received August 4, 1987; revised September 4, 1987

With the advent of RFLPs, genetic linkage maps are now being assembled for a number of organisms including both inbred experimental populations such as maize and outbred natural populations such as humans. Accurate construction of such genetic maps requires multipoint linkage analysis of particular types of pedigrees. We describe here a computer package, called MAPMAKER, designed specifically for this purpose. The program uses an efficient algorithm that allows simultaneous multipoint analysis of any number of loci. MAPMAKER also includes an interactive command language that makes it easy for a geneticist to explore linkage data. MAPMAKER has been applied to the construction of linkage maps in a number of organisms, including the human and several plants, and we outline the mapping strategies that have been used. © 1987 Academic Press, Inc.

INTRODUCTION

A primary genetic linkage map, consisting of easily scored polymorphic marker loci spaced throughout a genome, is an essential prerequisite to detailed genetic studies in any organism. Classically, it has been possible to construct such linkage maps only in intensively studied organisms, such as bacteria, yeast, or fruit flies, in which many visible mutations were available as genetic markers. Recently, however, this limitation has been removed, following the recognition that DNA polymorphisms (most conveniently visualized as *r*estriction *f*ragment *l*ength *p*olymorphisms, or RFLPs) could provide an abundant supply of codominant genetic markers (Botstein *et al.*, 1980). Projects are currently underway aimed at constructing complete RFLP linkage maps in many organisms, including human (Schumm *et al.*, 1985; White *et al.*, 1985), mouse (J. L. Guenet, personal communication), maize (Helentjaris *et al.*, 1986; D. Hoisington, personal communication), lettuce (Landry *et al.*, 1987), tomato (Helentjaris *et al.*, 1986), the mustard *Arabidopsis thaliana* (C. Chang and E. Meyerowitz, personal communication), and the fungus *Bremia lactucea* (R. Michelmore, personal communication).

Construction of a linkage map involves following the inheritance of RFLPs in appropriate pedigrees. (i) For experimental organisms in which inbred lines are available and large crosses can be conveniently arranged (e.g., maize), it is most efficient to study progeny from an F2 intercross between two inbred lines. Although more complex to analyze, intercrosses provide almost twice as much information as backcrosses because markers are segregating in both parents. (ii) For natural populations in which inbred lines are not available and matings cannot be arranged (e.g., humans or trees), the most efficient approach is to study a collection of two- or three-generation nuclear families, consisting of four grandparents (optional), two parents, and a large number of children.

In both cases, one cannot simply analyze the data by "counting recombinants," because the data are fundamentally incomplete (see, e.g., Lander and Green, 1987). In an offspring from an F2 intercross between two inbred strains, if two loci are heterozygous one cannot tell whether crossovers occurred between the loci in neither parent or in both parents. In natural populations, even thornier problems arise. These include situations in which it is not possible to infer from grandparental genotypes which alleles at various loci are in *cis* and which are in *trans* in the parents; in which it is not possible to infer which allele a child inherited from which parent because the parents have the same genotype at a locus; and in which loci are uninformative in certain families. These com-

0888-7543/87 $3.00
Copyright © 1987 by Academic Press, Inc.
All rights of reproduction in any form reserved.

plexities can make it difficult to analyze even a two-point cross by hand.

Moreover, two-point analysis is just a starting point. Because only a limited number of co-informative meioses are studied, the genetic distances based on two-point crosses may be only rough approximations to the truth. Attempting to infer gene order from such distances can lead to incorrect conclusions. To overcome this problem, one requires multipoint linkage analysis. When most loci are informative (i.e., heterozygous) in most meioses, three- and four-point crosses typically suffice for correct inference of locus order. When loci are uninformative in a significant fraction of the meioses, it may be desirable to analyze 5 or 10 markers simultaneously: this ensures that informative flanking markers are present in every meiosis in which a recombination occurred between the markers of interest. In short, computer analysis is essential.

The most satisfactory and general approach to linkage analysis is the *method of maximum likelihood* (Haldane and Smith, 1947; Morton, 1955; Ott, 1985). For each possible map (consisting of an order for the loci and recombination fractions between them), one can compute the probability that the map would have given rise to the observed data; this probability is called the *likelihood* of the map. The "best" map is the one with the highest likelihood. (When it is possible to count recombinants, the resulting map is in fact the maximum likelihood solution; thus the method of maximum likelihood is a generalization of counting recombinants.) The ratio of the likelihoods between two maps provides a simple measure of how much better one fits the data than the other. The method is widely favored because it can be applied even if the modes of inheritance and amounts of data vary among loci.

Elston and Stewart (1971) provided the first general algorithm for computing the likelihood of any given map; by searching over many possible maps, one could find the map with maximum likelihood. The widely used programs LIPED (Ott, 1976) (for two-point analysis) and LINKAGE (Lathrop and Lalouel, 1984) (for multipoint analysis) implement this approach for very general pedigrees and arbitrary traits and are the workhorses of linkage analysis. The Elston–Stewart algorithm, however, is not well-suited to the sort of multilocus analysis involving a large number of loci required for constructing primary linkage maps of genomes: the computation time needed to calculate such likelihoods grows *exponentially* with the number of loci. Consequently, it has been written that multilocus linkage analysis is "prohibitively time-consuming even on a supercomputer" (Morton *et al.*, 1986) and that "some shorter and easier method is urgently needed" (Smith, 1986).

Several approaches have been proposed recently for overcoming this exponential bottleneck, in order to aid in the construction of linkage maps. Lathrop *et al.* (1986) have described a modification of the Elston–Stewart algorithm which involves the fact that one can sometimes factor the likelihood calculation into two or more parts—such as whenever parental phases are completely known at a locus (as may occur in three-generation, but not two-generation, pedigrees). This approach has been developed in a special-purpose version of the program LINKAGE, resulting in a substantial increase in speed (J.-M. Lalouel, personal communication).

In addition, Lander and Green (1987) have described a different algorithm for computing likelihoods, one whose computation time has been mathematically proven to scale *linearly* rather than exponentially with the number of loci.

We describe here a new computer package, MAPMAKER, specifically designed for the construction of primary genetic linkage maps from RFLP data either from F2 intercrosses in experimental populations or from two- and three-generation nuclear families in natural populations. MAPMAKER provides an interactive, user-friendly environment designed to let a geneticist easily explore his or her data. The package uses the Lander–Green algorithm to calculate the "best" map for any given order of loci. The favorable scaling properties of the algorithm make it practical to study a large number of loci simultaneously. In addition, the package includes an interactive command language which allows one to compare different genetic orders and to map new loci to genetic intervals.

We also describe systematic strategies that can be used for constructing detailed genetic linkage maps, which become feasible with the ability to perform multilocus analysis rapidly.

OVERVIEW OF MAPMAKER

The MAPMAKER program is written in the C programming language, with versions for both the UNIX and the VAX/VMS operating systems. The program and a short user's manual are available to academic researchers without charge by writing to the authors.

Interactive Shell

In the hope of making linkage analysis more directly accessible to the working geneticist without extensive computer experience, we have designed MAPMAKER to be interactive: one uses a simple vocabulary of commands that instructs MAPMAKER to perform various types of analysis on subsets of loci. One can instruct MAPMAKER to record a verbatim

transcript of the interactive session by typing **"photo output,"** where *output* is replaced by the name of the file in which the transcript should be placed.

Data

At the outset of a MAPMAKER session, one loads a file containing either of two types of information, called "F2 data" or "CEPH-type data." (i) F2 data refers to data from F2 intercrosses or backcrosses between homozygous inbred lines. The data may contain co-dominant, dominant, or recessive markers, as well as missing data. (ii) CEPH-type data refers to data on segregation of co-dominant markers such as RFLPs in two- or three-generation families in a natural population. The name CEPH is taken from the Centre d'Etude du Polymorphisme Humain, an international collaboration that has collected cell lines from 40 such human families. The genotype of each individual is listed in the form "a/b," where a and b are names assigned to the alleles, or "." to denote missing data. (There is no limit on the number of alleles, nor is there any need to recode alleles for increased efficiency.)

As each locus is loaded, it is assigned a number that can be used to refer to it. One can also refer to the locus by a name, if one prefers.

Making a Map

The most basic operation in MAPMAKER is to construct the maximum likelihood map for a particular set of loci in a particular order. One could type the command

"sequence 9 3 1 7 8".

This command tells MAPMAKER that the set of five loci numbered 9 3 1 7 8, in this fixed order, should be used in all subsequent analysis (until a new sequence is specified). If one next typed

"map"

MAPMAKER would compute the maximum likelihood map for the five loci numbered 9 3 1 7 8, in this presumed genetic order. For example (using a human data set), MAPMAKER output is

MAP:

9—3	**36.6 cM**	**25.9%**
3—1	**8.4 cM**	**7.7%**
1—7	**11.9 cM**	**10.6%**
7—8	**20.4 cM**	**16.7%**

log-likelihood = −280.66 (11 iterations)

For each interval, the maximum likelihood estimates of the recombination fraction and recombination distance are given. (Haldane's mapping function is used to compute the latter from the former, although alternatives can be used instead.)

The likelihood is $10^{-280.66}$, meaning that this is the probability that the given map would exactly give rise to the observed data. Note that the likelihood is necessarily very small, because it is the probability that *each* meiosis under study would come out exactly the same if the experiment were repeated. Thus, likelihoods are useful only for comparative purposes. For example, if an alternative map had a 1000-fold lower chance of giving rise to the data, one might choose to reject it.

If one next typed

"sex on"

"map"

one would obtain the maximum likelihood map allowing for sex-specific recombination fractions. In the example above, the output was

SEX-SPECIFIC:

	MALE-MAP:		FEMALE-MAP:	
9—3	**41.0 cM**	**28.0%**	**32.8 cM**	**24.0%**
3—1	**8.7 cM**	**8.0%**	**8.0 cM**	**7.4%**
1—7	**2.2 cM**	**2.2%**	**47.8 cM**	**30.8%**
7—8	**11.0 cM**	**9.9%**	**20.6 cM**	**16.9%**

log-likelihood = −277.52 (14 iterations)

Considering many sequences. A sequence can refer to more than one order of the markers. If one types

"sequence {9 3} 1 7 8"

"map"

MAPMAKER will output two maximum likelihood maps corresponding to the two orders obtained by permuting 9 and 3. Similarly,

"sequence {9 3} 1 {7 8}"

will refer to the four orders obtained by permuting both pairs, and

"sequence {9 3 1 7 8}"

will cause MAPMAKER to consider each of the 60 possible orders obtained by permuting the five loci. Instead of asking for all 60 maps to be printed, one might prefer to type

"compare".

MAPMAKER will then compute the maximum likelihood map for each of the 60 orders, will sort the orders by likelihood, and will print out a summary table of the best 20 orders:

order 1: 9 3 1 7 8 Log-likelihood: −280.66
order 2: 8 9 3 1 7 Log-likelihood: −284.37

order	**3:**	**9**	**1**	**3**	**7**	**8**	**Log-likelihood:**	**−284.59**
order	**4:**	**9**	**3**	**1**	**8**	**7**	**Log-likelihood:**	**−285.49**
order	**5:**	**9**	**7**	**1**	**3**	**8**	**Log-likelihood:**	**−286.35**
order	**6:**	**9**	**1**	**7**	**3**	**8**	**Log-likelihood:**	**−286.41**
order	**7:**	**9**	**3**	**7**	**1**	**8**	**Log-likelihood:**	**−286.50**
order	**8:**	**9**	**7**	**3**	**1**	**8**	**Log-likelihood:**	**−286.56**
order	**9:**	**9**	**8**	**7**	**1**	**3**	**Log-likelihood:**	**−287.41**
order	**10:**	**8**	**9**	**7**	**1**	**3**	**Log-likelihood:**	**−287.51**
order	**11:**	**8**	**9**	**1**	**3**	**7**	**Log-likelihood:**	**−287.76**
order	**12:**	**8**	**9**	**1**	**7**	**3**	**Log-likelihood:**	**−288.03**
order	**13:**	**8**	**9**	**3**	**7**	**1**	**Log-likelihood:**	**−288.58**
order	**14:**	**8**	**9**	**7**	**3**	**1**	**Log-likelihood:**	**−288.64**
order	**15:**	**9**	**1**	**3**	**8**	**7**	**Log-likelihood:**	**−288.66**
order	**16:**	**9**	**8**	**3**	**1**	**7**	**Log-likelihood:**	**−288.90**
order	**17:**	**9**	**8**	**7**	**3**	**1**	**Log-likelihood:**	**−289.81**
order	**18:**	**9**	**8**	**1**	**3**	**7**	**Log-likelihood:**	**−290.25**
order	**19:**	**9**	**8**	**1**	**7**	**3**	**Log-likelihood:**	**−290.48**
order	**20:**	**9**	**7**	**8**	**1**	**3**	**Log-likelihood:**	**−291.09**

In the example given, the best map for the genetic order 9 3 1 7 8 is $10^{3.71} = 5128$ times more likely to have given rise to the data than the best map for any of the other 59 alternative genetic orders. This would be strong support for this genetic order over the alternatives.

The **sequence** command provides other options. If one types

"sequence {1 2 3 [4 5] 6}"

then MAPMAKER will try all permutations which do not interpose any loci between the loci 4 and 5 and which maintain their order. This is useful if 4 and 5 are already known to be extremely close. If one types

"sequence ⟨1 2 3 4⟩ ⟨5 6 7 8⟩"

MAPMAKER will consider the four orders obtained by inverting one or both of the lists of four loci. This is useful if one has two mapped linkage groups whose relationship to one another is not yet determined.

Any command given to MAPMAKER will be performed on all the locus orders implied by the current sequence.

Placing a new locus. If one has previously determined a genetic order with a high degree of certainty, new loci can be added to the map by determining the interval into which they fall. For example, suppose that we had determined the correct order for four of the loci discussed above: 9 3 7 8. We wish now to determine the position of locus 1. To position the new locus, we should compare the likelihoods for six different maps: the best maps obtained for each of the five positions into which the new locus can be placed in the order 9 3 7 8, as well as the best map that can be made if the new locus is forced to lie at 50% recombination distance (i.e., unlinked). If we typed

"try 1"

MAPMAKER would compute the required maps and print out their relative log-likelihoods:

```
RELATIVE LIKELIHOODS:
                1
           -----------
        |   -24.90
  9     |
        |    -3.93
  3     |
        |     0.00
  7     |
        |    -5.84
  8     |
        |   -18.19
  inf   |
        |   -34.38
           -----------
```

The table indicates the relative likelihoods for the best maps that can be made with 1 placed in the indicated positions. The most favorable position for 1 is between 3 and 7. If 1 is instead placed between 9 and 3, then the best map that can be made has a likelihood which is smaller by a factor of $10^{3.93} = 8511$. Finally, if 1 is placed at 50% recombination (i.e., unlinked), the best map that can be made has a likelihood that is $10^{34.38}$ smaller. These results provide strong support for the location of locus 1. (Of course, in this case, we already knew the location of locus 1 from the exhaustive comparison of all 60 possible orders.)

It is important to note that each time MAPMAKER tests a locus in an interval it reestimates *all* the recombination distances. This contrasts with approaches in which the recombination distances for all but the new locus are held fixed. Such approaches do not use the full information in the data and may allow errors to propagate; they are sometimes adopted to reduce computation time. An advantage of the algorithm used in MAPMAKER is that it does not take much more time to reestimate all recombination fractions than it would just to reestimate one of them.

Finally, if one has already narrowed down the possible positions for the locus, one can also instruct MAPMAKER to restrict the comparison to a particular subset of the intervals.

Testing linkage between two groups. One can test for linkage between two groups of loci (as opposed to two individual loci) by using the command **"linked?"**. One would first enter a single sequence listing both linkage groups and then type

"linked?".

MAPMAKER will ask the user to indicate the interval to the test for linkage (i.e., the interval between the two linkage groups). The program will then compute (i) the likelihood for the best map if the recombi-

nation fraction for the designated interval is held at 50% and (ii) the likelihood for the best map if it is allowed to vary. Such a test may allow one definitively to detect linkage between two linkage groups, even when no pair of markers in the two linkage groups is sufficiently informative to allow linkage to be definitively detected between them. This can be quite useful in CEPH-type data.

Two-Point Analysis

Before performing multipoint analysis as described above, it is best to begin with two-point analysis. MAPMAKER provides a number of commands to facilitate two-point analysis. The command

"sequence *all"

"two-point"

will cause MAPMAKER to compute two-point recombination fractions and maximum LOD scores between all pairs of loci in the data file. (The maximum LOD score for a pair of loci is the traditional measure of two-point linkage (Morton, 1955). It is defined as the $\log_{10}$ of the ratio of the likelihoods when the loci are taken to be at their maximum likelihood recombination fraction and when the loci are taken to be unlinked.)

All of the two-point distances and maximum LOD scores are stored internally for analysis. To simply print out the table of distances and maximum LOD scores, one would type

"lodtable".

To see just a list of the maximum LOD scores over 3.0, one would type

"biglods 3.0".

One can also instruct MAPMAKER to infer linkage groups from the two-point data. If one typed

"group 0.30 3.0"

MAPMAKER would determine the linkage groups that would be inferred from the two-point data if loci are considered as linked whenever the recombination fraction between them is less than 0.30 and the maximum LOD score exceeds 3.0. In an example with F2 data involving 40 loci in a plant genome, MAPMAKER responded to the **"group"** command with

Suspected linkage groups:

1 3 5 8 13 14 16 18 19 21 22 26 29 32 33 34 37 38

2 4 11 15 20 25 36

6 10 23 28 39 40

7 17 27

9

12 24 30 35

31

The five linkage groups with multiple loci correspond to the five known chromosomes of the plant, while the two singletons are loci sufficiently far out on a chromosome that they do not yield a maximum LOD score of at least 3.0 for linkage.

Finally, after a two-point analysis, one can examine all three-point crosses between nearby loci by typing

"three-point 0.20 3.0".

For a set of three loci, a, b, c, in the current sequence, if the pairs a,b and b,c are separated by less than 20% recombination and have a maximum LOD score of at least 3.0, MAPMAKER will compare the relative likelihoods for the best maps that can be made for each of the three possible orders: abc, acb, and cab. A portion of the output, pertaining to the three loci numbered 12, 24 and 30, was

30 12 24 > 12 24 30 > 12 30 24

0.00 2.64 5.23

[best map: 19.4 16.9 cM]

The output means that the order 30 12 24 is favored over the two alternatives: the best map for this order has a likelihood that is $10^{2.64}$ higher than the best map for the order 12 24 30 and $10^{5.23}$ higher than the best map for the order 12 30 24. The best map for the preferred order is then given.

After the **"three-point"** command completes this task, MAPMAKER then determines all n-point orders compatible with the three-point data. In particular, the program considers as excluded any three-point order whose likelihood falls below the likelihood for an alternative order by a given threshold, specified by the user. It then finds all n-point orders which contain no excluded three-point order.

In addition, one can ask for the closest locus to a given locus and for all loci within a specified distance of a locus (by using the commands called **"closest"** and **"near"**).

Other Functions

MAPMAKER provides a number of other commands, including ones that compute the likelihood at any desired point on the likelihood surface, that apply a permutation test to determine the significance of

sex-specific (as opposed to sex-nonspecific) recombination fractions, and that allow the user to name and record sequences of linkage groups for future use. It is our intention to add additional functions as needs arise. Finally, MAPMAKER provides an on-line **"help"** facility.

Algorithmic and Statistical Considerations

MAPMAKER uses the algorithm described by Lander and Green (1987) to find the maximum likelihood genetic linkage map, with the genetic reconstruction being performed by the Markov reconstruction approach detailed there. The procedure is declared to have converged when the log-likelihood increase from one iteration to the next is below a given tolerance t. Since it is mathematically proven that the likelihood can never decrease (Lander and Green, 1987), the only issue is to choose an adequately fine tolerance. We find that $t = 0.01$ is useful for initial exploration of new data, but we use $t = 0.001$ in later stages of analysis. (Convergence is essentially complete by this stage, in all cases we have examined.) The threshold for declaring convergence can be changed at any time with the command **"tolerance 0.001"**, for example. The user also can specify an initial point at which to begin the iterative search, but we find that the choice makes little difference and, in practice, we use the default values.

MAPMAKER simultaneously reestimates all the recombination fractions for each map. This is preferable to using previous estimates, since it avoids the propagation of slight errors. If desired, however, one can instruct MAPMAKER to treat certain distances as fixed.

USE OF MAPMAKER

MAPMAKER has been extensively tested in the course of collaborative projects with a number of researchers involved in RFLP mapping. Using the F2 data option, MAPMAKER has been used to analyze genetic maps of the entire genome of maize, *Zea mays* (with D. Hoisington, University of Missouri, Columbia), of the mustard, *Arabidopsis thaliana* (with C. Chang and E. Meyerowitz, Cal. Tech.), and of lettuce, *Lactuca sativa* (with R. Michelmore, University of California, Davis). Using the CEPH-type data option, MAPMAKER has been used to analyze a genetic map of 63 RFLP loci on human chromosome 7 (with H. Donis-Keller and colleagues, Collaborative Research, Inc.) and a partial linkage map of the genome of the lettuce mildew, *Bremia lactucea* (with S. Hulbert and R. Michelmore, University of California, Davis). The human chromosome 7 project (Barker *et al.*, 1987) provided a particularly important test, because one of us (P. Green) independently constructed a linkage map by using a separate program, CRI-MAP, which he has written; both programs produced the same preferred orders and maps.

The results of these projects are or will be reported elsewhere by the various groups. We summarize here some general considerations about the use of MAPMAKER in construction of genetic linkage maps.

(i) Analysis of F2 Data

In a single experimental cross, informative data are available on all individuals at all loci under study (apart from missing information due to uninterpretable lanes on Southern blots, typically 5–10%). It is thus fairly easy to infer locus order.

We begin by finding all the apparent linkage groups (by using **"two-point"** followed by **"group"**). We then perform all three-point crosses within a linkage group (by using **"three-point"**). We instruct MAPMAKER to determine all orders of loci which are compatible with the three-point data, treating as excluded any three-point order whose likelihood falls 1000-fold below an alternative. Finally, multilocus crosses are used to determine the correct order from among the possibilities compatible with the three-point data. After linkage groups are constructed, they can be recorded in a file for future use. As new loci are added to the data set, they are placed relative to the previously constructed linkage groups (by using **"try"** or related functions to compare the results of positioning the marker in each interval) and new maps derived.

Computation times required for each step are minimal. Interactive analysis of the entire linkage map of a genome can usually be completed in less than a day.

The computer program generally used for analyzing F2 data (Suiter *et al.*, 1983) performs only two-point analysis. In a number of cases we have examined, rigorous multipoint linkage analysis has revealed errors in genetic order when maps were constructed by hand using only such two-point information. We would suggest that the construction of such linkage maps in F2 populations be performed via multipoint analysis.

(ii) Analysis of CEPH-Type Data

Data from natural populations are more difficult to analyze. There are two issues: the first fundamental and the second computational.

The fundamental issue is that there may not be enough data to order loci definitively, due to the fact that each locus is informative in only a fraction of the families. To order the loci, one must have several meioses in which both loci are informative, in which a recombination has occurred between them, and in

which a flanking marker is also informative. Since different loci are informative in different meioses, a number of flanking loci may be required in order to make full use of all the available data. Accordingly, three-point and four-point crosses are often not sufficient for resolving the order of the loci.

Briefly, we have used the following strategy. Based on two-point analysis, we infer apparent linkage groups. For each linkage group, we use the two-point analysis to select a handful of relatively informative loci that appear to be separated by gaps of 10–20 cM. We compare the likelihoods for the best maps with each possible order for the loci (by use of "**compare**"). If one or two locus orders have much higher likelihoods than the rest (say, by 10,000:1 likelihood ratio), we accept these orders as a framework for further analysis. The positions of each of the remaining loci, relative to this framework, are tested (by use of "**try**"). If a locus clearly maps to an interval (we frequently require a likelihood ratio of 100:1 for one interval over all others), it is added to the framework. We repeat the process as new loci are added to the framework, since these often allow further loci to be placed uniquely. In the end, some nearby loci cannot be ordered with respect to one another. As a test of the final order, we do the following: for each set of three consecutive loci, we test all six permutations of these loci while keeping the order of the remaining loci fixed. Finally, we attempt to detect linkage between two apparent linkage groups (by use of "**linked?**"). It is worth noting that one of us (P. Green) has implemented a somewhat different strategy, based on a breadth-first search, in conjunction with his program CRI-MAP and this approach has led to the same conclusions.

The computation time needed to construct a map with CEPH-type data depends on the degree of ambiguity in the data. The ambiguity may be of two sorts: (i) *parental phases unknown,* which occurs if the grandparents are missing or if the parent and his two grandparents are heterozygotes of the same genotype; and (ii) *child phases unknown,* which occurs if the two parents and their child are heterozygotes of the same genotype. If neither situation arises, the data are said to be *phase-known.*

Computation time increases as one moves (i) from analyzing completely phase-known data; (ii) to data for which child phases are known, but parental phases may be unknown; and (iii) to data for which child phases are unknown. Within case (iii), the computation time rises with the number of children for whom phase is unknown. It becomes substantial only when the number of such children is about 10, which occurs only rarely. The bulk of the computing time is thus spent on a small and identifiable subset of the data.

Accordingly, MAPMAKER allows the user to vary the portion of the data included in any particular analysis. For each family, the program can automatically (i) omit those loci in which either parental or child phases are unknown, (ii) omit those loci in which child phases are unknown for more than a specified number of children, or (iii) include all loci. In this way, one can perform extremely rapid initial analyses, often sufficient to eliminate many possibilities, or else slower, more complete analyses.

It is impossible to provide an absolute measure of computation time, since this varies with the type of computer, the load on the computer, the size of the data set, and the data for the particular loci analyzed. To obtain a rough measure of the required computation times, we computed multipoint maps for a few thousand sets of ordered loci using RFLP data from some 23 CEPH families on an HP9000 minicomputer. For various examples involving 10 loci, the typical times needed to compute the maximum likelihood map varied in the range of (i) about 1–2 s, for phase-known data; (ii) about 3–10 s for data with parental phases unknown, but child phases known; and (iii) about 20 s to 6 min for data with child phases unknown. (When we omitted loci at which child phases were unknown for more than six children, computation times fell to under 1 min.) Computations were about 40% faster on a VAX 8350, a small model in the VAX line.

In the construction of linkage maps in organisms with long generation times, it may be more convenient to employ two-generation nuclear families rather than three-generation families. In this connection, it is worth noting that the lack of parental phase information imposes no serious computation limit. This contrasts sharply with the Elston–Stewart algorithm, for which multilocus analysis of a two-generation family would unavoidably lead to exponential explosion of computing time as loci are added. Although grandparental data are unnecessary from the point of view of computation efficiency, they do contribute some additional information, roughly equivalent to one to two additional children.

CONCLUSION

Genetic linkage maps consisting of RFLPs will likely be assembled over the next decade for many organisms of interest, both in experimental and in natural populations. The degree of DNA sequence polymorphism seems adequate in most cases to make feasible the isolation and study of the RFLPs.

Accurate construction of these linkage maps will be best performed via multipoint linkage analysis, using the method of maximum likelihood. Such rigorous analysis is crucial in the initial stages of map con-

struction and frequent reanalysis is important as additional loci are added to refine the map. The MAPMAKER package may offer a useful and convenient analytical tool to assist in this purpose, because it combines a computationally efficient algorithm with an extensive, interactive command language designed for studying genetic order.

We should emphasize that MAPMAKER is *not* a general-purpose linkage analysis program, such as LIPED and LINKAGE, which allow analysis for arbitrary traits in arbitrary pedigrees. Its scope is limited to the construction of primary genetic linkage maps using two types of information: (i) codominant, dominant, or recessive traits in F2-type pedigrees; and (ii) codominant traits in CEPH-type pedigrees. Investigators engaged in such studies may find it of value.

Until recently, it has been believed that extensive multipoint linkage analysis was computationally impractical (Morton *et al.*, 1986; Smith, 1986). With the development of new algorithms and new programs, such as a fast special-purpose version of LINKAGE for three-generation families and the MAPMAKER package, this limitation has been removed for the types of pedigrees used for the construction of primary linkage maps.

ACKNOWLEDGMENTS

We thank David Page and David Botstein for comments on the manuscript. This work was supported in part by grants from the National Science Foundation and System Development Foundation (to E.S.L.). The work of several authors (J.A., A.B., M.D., S.L., and L.N.) was carried out under the auspices of the Undergraduate Research Opportunities Program at MIT.

REFERENCES

1. BARKER, D., GREEN, P., KNOWLTON, R., SCHUMM, J., LANDER, E., OLIPHANT, A., WILLARD, H., AKOTS, G., BROWN, V., GRAVIUS, T., HELMS, C., NELSON, C., PARKER, C., RISING, M., REDIKER, K., WATT, D., WEIFFENBACH, B., AND DONIS-KELLER, H. (1987). A linkage map of human chromosome 7 with 63 DNA markers. *Proc. Natl. Acad. Sci. USA*, in press.
2. BERNATZKY, R., AND TANKSLEY, S. D. (1986). Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics* **112:** 887–898.
3. BOTSTEIN, D., WHITE, R. L., SKOLNICK, M. H., AND DAVIS, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Amer. J. Hum. Genet.* **32:** 314–331.
4. ELSTON, R. C., AND STEWART, J. (1971). A general model for the analysis of pedigree data. *Hum. Hered.* **21:** 523–542.
5. HALDANE, J. B. S., AND SMITH, C. A. B. (1947). A new estimate for the linkage between the genes for colour-blindness and haemophilia in man. *Ann. Eugen.* **14:** 10–31.
6. HELENTJARIS, T., KING, G., WRIGHT, S., SCHAEFFER, A., AND NIENHUIS, J. (1986). Construction of linkage maps in maize and tomato based on isozymes and random cDNA sequences. *Theor. Appl. Genet.* **72:** 761–769.
7. LANDER, E. S., AND GREEN, P. (1987). Construction of multilocus linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84:** 2363–2367.
8. LANDRY, B. S., KESESELI, R. V., FARRARA, B., AND MICHELMORE, R. W. (1987). A genetic linkage map of lettuce (*Lactuca sativa* L.) with restriction fragment length polymorphism, isozyme, disease resistance and morphological markers. *Genetics* **116:** 331–337.
9. LATHROP, G. M., AND LALOUEL, J. M. (1984). Easy calculation of lod scores and genetic risks on small computers. *Amer. J. Hum. Genet.* **36:** 460–465.
10. LATHROP, G. M., LALOUEL, J. M., AND WHITE, R. L. (1986). Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis. *Genet. Epidemiol.* **3:** 39–52.
11. MORTON, N. (1955). Sequential tests for the detection of linkage. *Amer. J. Hum. Genet.* **7:** 277–318.
12. MORTON, N. E., MACLEAN, C. J., LEW, R., AND YEE, S. (1986). Multipoint linkage analysis. *Amer. J. Hum. Genet.* **38:** 868–883.
13. OTT, J. (1976). A computer program for linkage analysis of general human pedigrees. *Amer. J. Hum. Genet.* **28:** 528–529.
14. OTT, J. (1985)."Analysis of Human Genetic Linkage," Johns Hopkins Press, Baltimore, MD.
15. SCHUMM, J., KNOWLTON, R., BRAMAN, J., BARKER, D., VOVIS, G., AKOTS, G., BROWN, V., GRAVIUS, T., HELMS, C., REDIKER, K., THURSTON, J., BOTSTEIN, D., AND DONIS-KELLER, H. (1985). Identification of more than 500 RFLPs by random screening: 8th Human Gene Mapping Workshop. *Cytogenet. Cell Genet.* **40:** 576.
16. SMITH, C. A. B. (1986). The development of human linkage analysis. *Ann. Hum. Genet.* **50:** 293–311.
17. SUITER, K. A., WENDEL, J. F., AND CASE, J. S. (1983). Linkage-1: A Pascal computer program for the detection and analysis of genetic linkage. *J. Hered.* **74:** 203–204.
18. WHITE, R., LEPPERT, M., BISHOP, D. T., BARKER, D., BERKOWITZ, J., BROWN, C., CALLAHAN, P., HOLM, R., AND SEROMINSKI, L. (1985). Construction of linkage maps with DNA markers for human chromosomes. *Nature (London)* **313:** 101–105.