Exclusion mapping

J H EDWARDS

From the Genetics Laboratory, Department of Biochemistry, South Parks Road, Oxford OX1 3QU.

SUMMARY The basis for using information against linkage to some loci to support linkage to other loci is discussed and data on cystic fibrosis available up to late July 1985 used as an example.

In attempting to locate an unknown locus by comparing the cosegregation of alleles at this locus with alleles from loci of known position, every informative meiosis makes a contribution. In the simplest case of one marker locus of known position showing allelic cosegregation, then, if this is of sufficient strength to be beyond the likely vagaries of chance, the unplaced locus will be assumed to be a neighbour of the marker locus and additional support from other loci will be unnecessary. However, by considering evidence against linkage the same strength of inference can be obtained on fewer observations. Since every locus must be somewhere, it is possible to define limited probable segments of chromosome by exclusion alone. The definitive paper is by Cook *et al.*¹ They used the term exclusion mapping and distinguished it from deficiency or deletion mapping, in which a certainty of exclusion can be deduced.

The strength of evidence sufficient to justify the assumption of linkage is a matter of some complexity. If only one locus is tested, then an odds ratio of 1000:1 for the most likely recombination fraction compared to the recombination fraction of 1/2 expected from a distant or unlinked locus is a reasonable measure of certainty. Morton² defined a *z* score as the logarithm of this odds ratio, using Barnard's acronym of lod,³ but establishing a convention of lods to the base of 10. Morton used this acceptance criterion in the context of sequential analysis, and suggested a lod of 3, or an odds ratio of 1000:1 in favour, as a reasonable criterion. This is not the same as a probability of 1 in 1000 of fortuitous cosegregation, as there is no exchange rate between probability and likelihood. Probability relates to a subset of a set of possible events on a defined set of conditions, and likelihood to one event in relation to two values of a parameter. However 1000:1 or a lod of 3 is a round number giving reasonable grounds for a sound inference if

only one marker is tested; this was discussed in detail by Morton.²

If it were known that the recombination fraction were either *t* or 1/2, then the chance of obtaining *r* recombinants and *n* non-recombinants would be either

$$t^r p^n \text{ or } (1/2)^r (1/2)^n$$

where $p=1-t$

giving an odds ratio of

$$2^s : t^r p^n \text{ where } s=n+r$$

and a lod, or log of this odds ratio, of

$$r \log(t) + n \log(p) - s \log(1/2) \\ \text{or } r \log(2t) + n \log(2p).$$

If, as is usual, several loci are tested, then clearly the chance of an erroneous assignment due to chance will increase, and, for small numbers, will be proportional to the number of loci tested. For *n* loci the necessary odds ratio will be 1/1000*n* with a lod of $3 + \log(n)$, as advanced by Kidd and Ott⁴ on a somewhat more elaborate argument. An exact treatment by Thompson⁵ gives numerically similar results. This was overlooked by Botstein *et al.*⁶ and is one of several reasons why they seriously underestimated the number of observations needed to define a linkage at a given reliability.

With multiple loci failure to incorporate information against linkage is an inefficient approach: the locus must be somewhere, and every informative meiosis gives evidence for or against every position. There is no sudden stage at which linkage deserves the accolade of general approval, although the present conventions are useful provided they do not lead to the suppression of negative information. If there are many loci well spread over all chromosomes except one, and all are excluded, then the

locus must lie on that chromosome, and a location has been defined by exclusion.

The method is most easily explained by the simpler problem of one chromosome on which both the unplaced and marker loci are known to lie, as in X linkage. Information on the numbers of recombinants and non-recombinants, or cosegregants, will define the likelihood of the unplaced locus being at any defined position on this chromosome, assuming that the probability of recombination between any two points is known. In practice it is simplest to assume that recombination occurs at random, with a defined average expectation, the number of chiasmata per chromosome being proportional to the length of the mitotic chromosome. Cook *et al*¹ used a simple approach to define a distorted or meiotic length; this will be discussed later. The simplest relationship between distance and recombination fraction was defined by Haldane,⁷ who assumed a two strand cut-and-join event, with an appreciable chance of a chromosome not crossing over. Although, it is now clear that recombination occurs at a four strand stage, and that one recombination event usually inhibits a neighbouring event, this is an adequate approximation within the framework of the available data. This mapping function is based on summing the number of odd (recombinant) and even (non-recombinant) cross overs assuming both occur at random. As the number of even events, including zero, can never exceed the number of odd events the recombination fraction can never exceed a half. The average number of odd events between two loci, which defines the genetic length in Morgans, is proportional to the terms of a logarithmic series relating to twice the recombination fraction. The distance in Morgans, or units of 100 centimorgans, often termed w , is then related to the recombination fraction by the odd terms of a logarithmic series relating to twice the recombination fraction. A centimorgan (cM) is that length of a segment of chromosome which has a 1% chance of recombination.⁷

$$2w = (u + u^2/2 + u^3/3 + u^4/4 \dots \text{etc})$$

where t is the recombination fraction and $u = 2t$. Clearly the distance, w , is almost equal to t when t is small, becoming infinite when $t = 1/2$. Various modifications are possible to allow for interference. Kosambi⁸ introduced a mapping function which omits the even terms leaving those in 1, 3, 5, etc. This gives addition rules for recombination fractions which are identical to those for velocities in the theory of relativity, a recombination fraction of $1/2$ and the speed of light being absolute maxima.

Another function, which includes only terms in 1, 5, 9, . . . etc was advanced by Carter and Falconer.⁹ As such algebraic functions are only the roughest of guides and lack any cytogenetic justification, there are advantages in using the simplest.

Haldane's mapping functions can also be expressed as

$$w = 1/2(1 - 2 \ln(t))$$

and

$$t = 1/2(1 - e^{-2w}).$$

This is shown, with Kosambi's function, in the table.

Given these assumed relationships between distance and recombination, we can consider the implications of observing r recombinants and n non-recombinants on a chromosome with a length of 100 cM with the marker locus at its centre.

If we start at one end, the distance from the marker is 50 cM, giving a recombination fraction of 0.32 and a likelihood of $0.32^r 0.68^n$ while in the middle the likelihood will be $0.00^r 1.00^n$.

In likelihoods the non-observance of an impossible event has a unit likelihood, for it is bound to happen, so that, if no recombinants were observed, both the latter terms would be unity.

On this simple model we would have various likelihood distributions, and, as the likelihoods are relative to each other, we can define the probability distribution by computing the likelihoods and then multiplying them by some factor to preserve the unit area of the distribution. Figs 1, 2, and 3 show the three types of distribution possible for no recombinants, for some recombinants but far more non-recombinants, and for equal numbers of recombinants and non-recombinants.

If there were two chromosomes only one of which has a marker locus, then the probability distribution of the unmarked chromosome will be rectangular, but the area of this rectangle will be influenced by the information from the other chromosome. If $r = n$

TABLE Mapping functions of Haldane and Kosambi. Distances in centimorgans against percentage recombination.

Theta	Haldane	Kosambi
5.00	5.27	5.02
10.00	11.16	10.14
15.00	17.83	15.48
20.00	25.54	21.18
25.00	34.66	27.47
30.00	45.81	34.66
35.00	60.20	43.37
40.00	80.47	54.93
45.00	115.13	73.61

FIG 1 Likelihood distribution for $r:n$ of 0:10. Chromosome length 100 cM. Marker locus at 50 cM.

FIG 2 Likelihood distribution for $r:n$ of 5:5. Chromosome length 100 cM. Marker locus at 50 cM.

FIG 3 Likelihood distribution for $r:n$ of 2:18. Chromosome length 100 cM. Marker locus at 50 cM.

then this will favour the unmarked chromosome. If r is substantially less than n , this will favour the marked chromosome. If r is far less than n , and n is large, for example 2 and 18 respectively, then the likelihood distribution will so favour the marked chromosome that an assignment to this chromosome would be reasonable. The ratio of the areas defined by these likelihood ratios would be 525:1 and with a lod of 2.7, while the ratio of the peak to the mean would be

$$2 \log (0.2) + 18 \log (1.8) \text{ or } 3.2$$

with an odds ratio of 1584. The peak odds ratio will necessarily exceed the average lods ratio.

If we had a second locus on the first chromosome at position 75 units and had data on linkage of r_1 and n_1 for the first and r_2 and n_2 for the second, then the joint likelihood would be the product of the likelihoods defined by each, as in the example in fig 4.

The 'exclusion' zone defined by two nearby loci will often exclude intervening segments if they are analysed together, but not if they are analysed separately.

In practice matters are more complicated. Recombinant events are not at random and differ by sex. Mapping functions are not simple, and almost all those in common use are seriously flawed by the algebraic assumptions being inconsistent with the known mechanism of meiosis. However, these functions provide a rough guide and allow an easy visual display. It is only rough and, in view of the absence of any precise data on chromosome length, or of any preferential distribution of chiasmata by sex, or of the intensity of interference, there is little prospect of much refinement on present data. Nor is there much need of this, for such displays are only intended as rough guides.

Where information is not available in the form of recombinants and non-recombinants, equivalent observations may be used.¹⁰ These are defined as the number of recombinants and informative meioses which would give a likelihood curve with a peak of the same maximum at the same recombination fraction. Where the peak is undefined but positive it may be defined approximately by interpolation, assuming a parabolic distribution of lod score against recombination fraction, an approximation which becomes exact as the amount of data becomes infinite. Where the peak is at 0.5, as it will be with most loci tested, since most will be on different chromosomes, then it is convenient to define the lod at the value of 0.1. In practice a lod of -1.0 at a recombination fraction of 0.1 is equivalent

FIG 4 Chromosome length 100 cM. Marker loci at 50 cM and 75 cM with r:n of 2:8 and 3:7 respectively. The dots show the likelihoods, the solid areas are after standardisation to a unit area.

to about five informative meioses, or to about 2.5 recombinants out of five meioses. The non-integral values are immaterial.

A programme has been written to display exclusion areas following the approach of Cook *et al.*¹ The only data required are the length of the human chromosomes, the position of all markers in percentages of length of chromosome from the tip of the short arm, and the linkage data, which may be in various forms including recombinants and non-recombinants, equivalent observations, or various formats for lod scores from which the equivalent observations can be derived. Some published series of lods include one or more negative values separated by a positive value: these are almost certainly

misprints or other errors and should not be used. Data in which the absolute value of any lod is less than 1/2 are unlikely to be suited to translation to equivalent observations, and in any case contain little information. Such data are commented on at input and then ignored. They are worth including in any tabulation since they can be added to future data.

As an example, the data on cystic fibrosis (fig 5) available from the Human Gene Mapping Conference in Helsinki¹¹ and from previous publications^{12,13} are presented in the form shown at the end of that meeting. Strong support was provided by a very similar exclusion map for paraoxonase for which linkage data with cystic fibrosis were presented by Eiberg *et al.*¹⁴

While it would be simple to use distorted or meiotic maps, either following the simple algorithms of Cook *et al.*¹ or the algebraic distortions suggested by various authors, including Edwards *et al.*¹⁵ and Morton *et al.*¹⁶ this has not been done. The various summaries based on the Human Gene Mapping conferences provide annotated mitotic chromosomes, and any benefit from using a meiotic map on present data is likely to be offset by the lack of any standard map.

One source of misinformation is the independence of the information from neighbouring loci which are both informative. As yet this has been a serious problem, since most linkage data are based on loci with only two common alleles, so that most loci are usually uninformative for any given meiosis. It is, of course, essential not to use individual loci which are known to be closely linked as though they were independent. For example, in the HLA region HLA-B and Bf should not both be entered, since in most cases the information provided by the Bf locus would also be provided by the HLA-B locus. Similarly, it should be appreciated that the coarseness of the map is related to recombination being an all or nothing event, so that in any series showing a hopeful bulge in likelihood, testing the same series by more probes in the same region will lead to diminishing returns, or, if precautions are not taken to restrict analysis to previously uninformative families, to spurious confirmation.

The question is sometimes asked "How much of the genome has been excluded?". This is not a question which has an answer, for nothing is excluded: the likelihoods change, and a decision has to be made when to accept the evidence as sufficient to justify the expense of direct molecular analysis, or the casualty rate inseparable from the use of linkage in fetal diagnosis. The only certain method of exclusion is by the study of deficiencies, in the same way as the only certain method of inclusion is by the

FIG 5 Exclusion map of cystic fibrosis locus up to July 1985. The number to the right of the chromosome number is the probability of assignment to that chromosome. The black area is the likely area of occupancy of the locus.

study of duplications or trisomies or by in situ methods.

A copy of the programme, written in Turbo Pascal, accompanied by various files of published or publicly available data, is available on receipt of a 5" disk formatted for an IBM pc or similar.

References

- Cook PJJ, Noades JE, Lomas CG, Buckton KE, Robson EB. Exclusion mapping illustrated by the MNSs blood group. *Ann Hum Genet* 1980;44:61-73.
- Morton NE. Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7:277-318.
- Barnard GA. Statistical inference. *J Roy Statist Soc B* 1955;11:115-49.
- Kidd KK, Ott J. Power and sample size in linkage studies. *Human Gene Mapping* 1984;7:510-1.
- Thompson EA. Interpretation of lod scores with a set of marker loci. *Genet Epidemiol* 1984;1:357-62.
- Botstein D, White RL, Skolnick MH, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980;32:314-31.
- Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet* 1919;8:299-309.
- Kosambi DD. The estimation of map distances from recombination values. *Ann Eugen* 1944;12:172-5.
- Carter T, Falconer DS. Stocks for detecting linkage in the mouse and the theory of their design. *J Genet* 1951;50:307-25.
- Edwards JH. The analysis of X linkage. *Ann Hum Genet* 1971;34:229-50.
- Eighth International Workshop on Human Gene Mapping. *Cytogenet Cell Genet* 1985;40.
- Steinberg AG, Shwachman H, Allen FH, Dooley RR. Linkage studies in cystic fibrosis. *Am J Hum Genet* 1955;7:162-76.
- Goodchild MC, Edwards JH, Glen KP *et al*. A search for linkage in cystic fibrosis. *J Med Genet* 1976;13:417-9.
- Eiberg H, Mohr J, Schmiegelow K, Williamson R. Linkage relationship of paraoxonase and other markers: indication of PON-cystic fibrosis synteny. *Clin Genet* 1985;28:265-71.
- Edwards JH, Parekh J, Kirtton V, Hultén M. A meiotic linkage map of the human male. *Cytogenet Cell Genet* 1978;22:698-701.
- Morton NE, Lindsten J, Iselius L, Yee S. Data and theory for a revised chiasma map of man. *Hum Genet* 1982;62:266-70.

Correspondence and requests for reprints to Professor J H Edwards, Genetics Laboratory, Department of Biochemistry, South Parks Road, Oxford OX1 3QU.