A computer programme to calculate risk in X linked disorders using multiple marker loci

JOHN F CLAYTON

From the Clinical Section, MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Edinburgh.

SUMMARY A computer programme package has been written which calculates the genetic risk (recurrence risk) for X linked disorders, incorporating data from multiple probes. The programme can allow new mutation and the incorporation of other predictors of the disease genotype of the subjects in the family.

The increasing number of polymorphic DNA probes specific to the human X chromosome,¹ and the steadily increasing data on the linkage relationships between them and X linked disease loci,^{2–5} promise that multiprobe diagnosis will soon become a feasible proposition in many of these disorders. The use of closely linked probes known to lie on either side of the disease locus (bridging probes) would, in multiply informative families, reduce the rate of misdiagnosis to the order of 1% or less in situations where linkage phase is known with certainty. However, combining the results of even loosely linked probes with other predictors of carrier status, for example, creatine kinase levels in Duchenne muscular dystrophy, can produce estimates of risk sufficiently low (or high) to be likely to influence an at risk female's course of action. Unfortunately, except in the simplest cases, the calculation of such combined risks is tedious in the extreme. If other variables are to be incorporated, such as new mutation, the complexity will be such that errors will be almost inevitable. There is, therefore, a need for automation of the calculation process.

One method, using linkage analysis programmes such as LIPED,⁶ has been described⁷ but such programmes do not allow the possibility of new mutation or the easy incorporation of other predictors of genotype. In order to perform the calculation allowing for these factors, multiple runs of the programme are necessary, the various likelihoods produced being combined by hand to evaluate the risk. In most situations, this method will be almost as tedious and error prone as hand calculation. A further disadvantage is that information from only two marker loci can be considered by such a

method. In extensive families, it is likely that different markers will be informative (that is, key subjects are heterozygous for the marker) in different parts of the pedigree and that useful information may be neglected if a programme cannot cope with more than two markers.

In this paper, a pair of programmes is presented which allows the speedy calculation of risk of carrier status of X linked diseases, incorporating family structure, linkage data of multiple probes, other predictors of carrier status, predictors of hemizygote status, and the possibility of new mutation.

The programme: 'moll' and 'tarde'

There are two programmes used in this package for risk calculation in X linked disorders, called 'moll' and 'tarde'. Only the briefest account of the mechanisms of these programmes will be given here, the emphasis being to demonstrate their use. The programmes are flexible in that they allow the user to define the type of function required. For example, the programme must perform different functions when dealing with diseases with high or negligibly small mutation rates. These different functions are declared by the user by taking various 'options' when running the programmes. Thus, the description of the programmes is essentially a description of the available options.

'MOLL'

The first programme of the pair is 'moll', which takes input very similar to that taken by LIPED⁶ and PEDIG.⁸ Its function is to assign values to each and every unassigned allele in the input data and to calculate the prior probability of each permutation of the assigned values. The output from 'moll' is the

count of recombinant and non-recombinant meiotic events between each pair of loci in a standardised format, given the currently assumed permutation of alleles and the probe/disease ordering. This count is followed by a modified prior probability, which is calculated from population frequencies of alleles, mutation rates (if allowed), and factors of 0.5 (one for each child of a woman who is heterozygous at any locus). In effect, this programme produces an equation of risk which is a function of the various recombination fractions.

It is assumed that the order of the DNA probes along the chromosome is known, but that the position of the disease locus is not. Thus, the significance of the family has to be assessed not only for each possible allele/haplotype permutation, but also for each possible disease/probe ordering. The output of this programme forms the input to the second programme, 'tarde'.

Each of the options available within 'moll' is designated by a single letter. The most important of these will be described here.

r. This option indicates that the output is to be used to calculate the classical genetic risk to a subject, such as would be required in carrier identification. The programme requires the user to define the subject to be counselled and checks each permutation of alleles to see if that subject is a carrier under the current assumption. If she is, the line on output (which describes the current permutation) will be flagged. If not, no flag will be set.

p. This option allows the production of data suitable for the calculation of a probability distribution on risk. It is envisaged that this option would be used where antenatal diagnosis was being undertaken. Under this option, the user must define the fetus and the at risk parent. The p and r options cannot be run simultaneously.

m. This option allows the programme to permit a single new mutation in its description of the family. The user must specify male to child and female to child mutation rates separately. If this option is not taken, the processing of any allele/haplotype assumption which requires a mutation (for example, a normal mother having an affected son) will be abandoned without producing any output. This option can be taken with any of the other options.

l. This option requires the incorporation of another predictor of carrier status, such as creatine kinase (CK) data. The relative likelihood of the predictor value of each female is given in the input data.

h. This option is similar to option 1 in that it incorporates other predictors, but this time for males and females. This might be used when the disease status of hemizygotes can be statistically

inferred, as in 'age of onset' calculations for late onset disorders.

a. This option allows the user to change the number of alleles at each locus. The default number of alleles is two.

This programme can cope with families of virtually arbitrary structure. Up to 150 subjects can be specified in the input data, with data on up to five probes. If more probes are informative, it is only necessary to alter a single line of the programme to increase this number. Each probe is allowed to have a maximum of nine alleles, but the disease locus can have only two, the disease gene being defined as allele 2 at the disease locus. Again, alteration of a single line of the programme allows the number of alleles to be increased to 99.

'TARDE'

This programme takes the output data from 'moll' and translates it into a risk for the subject. It solves the equation of risk by substituting the values of the recombination fractions. Three options are currently available.

f. This option specifies that the user is confident of the position of the disease locus and that he is willing to specify values for each recombination fraction and a probe/disease order. Therefore, a single figure of risk is output.

t. Under this option, the user specifies a position for the disease locus, but not recombination fractions. Risks are calculated and presented in tabular form for all sets of probe/disease recombination fractions in 5% increments. Probe-probe recombination fractions are calculated from the probe/disease fractions. It is not feasible to produce three dimensional tables so this option is restricted to a maximum of two probes.

p. This option can be used on data produced by 'moll' under its p option to produce a probability distribution on risk. The classical risk (being the mean) is also produced. The user must define the order and the recombination fractions, as in option f.

Both these programmes are written in the 'C' programming language⁹ and have been tested on the UNIX operating system. When compiled, 'moll' occupies 25 k of space and 'tarde' occupies 15 k.

Examples of use

The programmes have been used to calculate risks for a number of hypothetical pedigrees of varying complexity. In all of these the results have been checked by hand. Three examples are presented here.

The first pedigree (fig 1) has information from

FIG 1 A small family with an X linked disorder. Results from two probes known to bridge the disease gene are available and are indicated in the brackets (probe 1, probe 2). Each probe recognises two alleles (1 and 2), the mother being heterozygous at both loci. Since there is a finite new mutation rate, the mother's genotype at the disease locus cannot be inferred, except to say that she cannot be homozygous for the disease gene.

two probes. In the calculation of risk of the proband being a carrier (table 1), the r and m options of 'moll' and the t option of 'tarde' were taken.

In the second pedigree (fig 2), the problem is to provide data for prenatal diagnosis for the proband. Here there are data from two probes and also from a predictor of carrier status. An edited form of the probability distributions of risk for the four possible male fetuses of the proband are shown in table 2. In this the risk is taken to be the proportion of offspring which can be expected to be affected. Where the two probes agree that the fetus is at low risk, such risks will be of the order of 1%, possibly extending up to 5% for useful probes. If they

disagree, the risk will be around 50% if the probes have equal recombination fractions or to one side if these are unequal. Thus, the risk will tend to fall into one of the four regions shown in the table. In this calculation, the p, m, and l options of 'moll' and the p option of 'tarde' were taken.

The third pedigree (fig 3) is one of the most complex ones which have been tested and the results checked. In this pedigree, information can be obtained from three probes and a predictor of carrier status. The assumed probe/disease order is

FIG 2 As fig 1, but with relative likelihoods for each female of a predictor of carrier status. This is the ratio of the proportion of normal females producing the laboratory result and the proportion of carrier females who produce it. The assumed value for each female is shown in square brackets.

TABLE 1 The calculated risks for the proband in fig 1 for a series of different recombination fractions. This is the tabular (or t) option of the programme 'tarde'. The new mutation rate was set at 0.00007 and the disease gene frequency (p) at 0.00014, that is, the heterozygote frequency (2p) is four times the mutation rate.¹⁰ The population frequency of allele 1 of each probe was set at 0.80.

	Probe 1 recombination fraction (rf)									
	0.050	0.100	0.150	0.200	0.250	0.300	0.350	0.400	0.450	0.500
Probe 2										
rf										
0.05	0.007	0.015	0.023	0.031	0.040	0.047	0.054	0.059	0.062	0.063
0.10	0.015	0.031	0.047	0.062	0.078	0.091	0.103	0.112	0.118	0.120
0.15	0.023	0.047	0.070	0.092	0.114	0.132	0.148	0.160	0.167	0.170
0.20	0.031	0.062	0.092	0.121	0.147	0.169	0.188	0.202	0.210	0.213
0.25	0.040	0.078	0.114	0.147	0.176	0.202	0.222	0.238	0.247	0.250
0.30	0.047	0.091	0.132	0.169	0.202	0.229	0.251	0.267	0.277	0.280
0.35	0.054	0.103	0.148	0.188	0.222	0.251	0.274	0.290	0.300	0.303
0.40	0.059	0.112	0.160	0.202	0.238	0.267	0.290	0.307	0.317	0.320
0.45	0.062	0.118	0.167	0.210	0.247	0.277	0.300	0.317	0.327	0.330
0.50	0.063	0.120	0.170	0.213	0.250	0.280	0.303	0.320	0.330	0.333

TABLE 2 The classical genetic risks and probability distributions of risk to the four possible offspring of the proband in fig 2. The table shows the probability that the risk to the offspring lies in the range given at the top of the table. The classical genetic risk is the sum of products of these probabilities and the risks. The probes were assumed to bridge the disease locus and each to have a recombination fraction of 0.10 with respect to the disease locus and 0.18 with respect to each other. The other parameters used in this calculation are defined in the legend to table 1.

Alleles of fetus	Mean	0%	1-5%	20-80%	95-99%
1,1	0.262	0.138	0.484	0.240	0.138
1,2	0.829	0.109	0.003	0.099	0.789
2,1	0.062	0.109	0.789	0.099	0.003
2,2	0.599	0.138	0.138	0.240	0.484

FIG 3 A complex pedigree with information for three probes, again shown in brackets (probe 1, probe 2, probe 3), this order being their order along the chromosome, the disease locus lying between probes 1 and 2. Each probe recognises two alleles. There is also a carrier status predictor, as in fig 2.

probe 1, disease, probe 2, and probe 3, with recombination fractions of 0.1, 0.05, and 0.05 between adjacent loci. Without the predictor data, the risk to the proband of being a carrier is 0.00021. With the predictor data, it is 0.0075. With all data except that of probe 3, the risk is 0.040. Without probe 2 it is 0.157 and without probe 1 it is 0.141. It is therefore of value in this pedigree to consider all three probe loci in the calculation of risk. These calculations used the *r*, *l*, and *m* options of 'moll' and the *f* option of 'tarde'. The other parameters used in this example are defined in the legend to table 1.

Discussion

In this paper, a programme package has been presented which calculates classical genetic risk and probability distribution on risk for X linked dis-

orders, incorporating multipoint DNA probe data. It can allow for the possibility of new mutation and other predictors of genotype status. The package has been tested with a number of pedigrees. Whenever there was a discrepancy, the error was found to be in the hand worked version or in the input data to the programmes. It appears therefore that the package functions correctly.

There is little doubt that there is a need for a well tested risk calculating programme. However, the major problem arises in the testing. With simple pedigrees, the results can be checked by hand. On the other hand, pedigrees of even moderate complexity can involve a thousand or more possible permutations of alleles in their full description. Each permutation can involve 20 or more separate calculations. It is not possible to check such a quantity of arithmetic by hand. Therefore, the risks calculated by this package for complex families with information from three or more probes should be viewed with a degree of suspicion, until the workings have been corroborated by an independent programme, such as 'LINKAGE',¹¹ in its multipoint form.

The assumption made in the present options of 'tarde', that probe/disease ordering can be known with certainty, warrants some discussion. It is certainly possible to define the order of DNA probes with high polymorphic information contents to a high degree of accuracy using standard families.¹² It is unlikely, however, that such accuracy will ever be achieved with disease loci. In general, there will be few families who are multiply informative and in whom linkage haplotypes are known with certainty. In those where they are, the probability of the various orders will be heavily dependent on those meiotic events which show recombination with respect to the DNA markers. Such recombinations will themselves be rare if the loci are closely linked. It is predictable, therefore, that a major contribution to the risk in most clinical situations will be the probability that the probe/disease order has been wrongly inferred. The risks produced by the *f* and *t* options will therefore tend to be underestimates of the true risk. The organisation of these programmes, with one programme producing an equation of risk for each possible order and a second programme solving them, makes it an easy matter to allow a valid risk to be calculated even when order is not certain. However, this option (in 'tarde') requires order specific probability distributions of probe/disease separations to be available. As yet none are, nor are there programmes to produce them, although it is intended to develop this package in order to do this. When such distributions do become available, the new option

will be implemented to produce indefinite order risks and the standard errors on risk.

This programme package and its documentation will be made available to interested parties.

References

- ¹ Wieacker P, Davies KE, Cooke HJ, *et al.* Towards a complete linkage map of the human X chromosome. *Am J Hum Genet* 1984;**36**:265–76.
- ² Murray JM, Davies KE, Harper PS, Meredith L, Mueller CR, Williamson R. Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature* 1982;**300**:69–71.
- ³ Kingston HM, Thomas NS, Pearson PL, Sarfarazi PL, Harper PS. Genetic linkage between Becker muscular dystrophy and a polymorphic DNA sequence on the short arm of the X chromosome. *J Med Genet* 1983;**20**:255–8.
- ⁴ Harper K, Winter RM, Pembrey ME, Hartley D, Davies KE, Tuddenham EGD. A clinically useful DNA probe closely linked to haemophilia A. *Lancet* 1984;**ii**:6–8.
- ⁵ Bhattacharya SS, Wright AF, Clayton JF, *et al.* Close genetic linkage between X linked retinitis pigmentosa and a restriction fragment length polymorphism identified by recombinant DNA probe L1.28. *Nature* 1984;**309**:253–5.
- ⁶ Ott J. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 1974;**26**:588–97.
- ⁷ Winter RM. The estimation of recurrence risks in monogenic disorders using flanking marker loci. *J Med Genet* 1985;**22**:12–5.
- ⁸ Heuch I, Li FHF. PEDIG—a computer program for calculation of genotype probabilities using phenotype information. *Clin Genet* 1972;**3**:501–4.
- ⁹ Kernighan BW, Ritchie DM. *The C programming language*. New Jersey: Prentice-Hall, 1978.
- ¹⁰ Emery AEH. Duchenne muscular dystrophy: genetic aspects, carrier detection and antenatal diagnosis. *Br Med Bull* 1980;**36**: 117–22.
- ¹¹ Lathrop GM, Lalouel JM. Easy calculation of Lod scores and genetic risks on small computers. *Am J Hum Genet* 1984;**36**: 460–5.
- ¹² White R, Leppert M, Bishop DT, *et al.* Construction of linkage maps with DNA markers for human chromosomes. *Nature* 1985;**313**:101–5.

Correspondence and requests for reprints to Dr J F Clayton, Clinical Section, MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Edinburgh EH4 2XU.