Linkage Newsletter

Vol. 2 No. 1 July 1988

EDITORIAL

You may have been wondering what happened to the Linkage Newsletter. It is still very much alive although there has been an unplanned gap since the last issue, largely due to a lack of time on my part. But I hope that in the future, the Newsletter will appear on a more regular basis.

People interested in contributing news and comments are welcome to do so. Please submit your contributions on a diskette, either 5¼" (360KB or 1.2MB) or 3½" (DD or HD). The authors name will appear as given on the text submitted (initials or full name, with or without address).

ANNOUNCEMENT

The computer programs LODSTAT and SIMLINK (Boehnke, 1986; Boehnke and Ploughman, 1986) were developed to permit estimation of the power of a linkage study based on a set of pedigrees of known structure. The method used by these programs assumed that trait phenotype implies trait genotype and, hence, limited their usefulness to dominant, fully-penetrant traits. We have recently developed a new method (Ploughman and Boehnke, 1988) that allows estimation of the power of a linkage study for traits with more complex modes of inheritance. This method simulates genotypes at the trait locus, correctly taking into account trait information on all pedigree members. The method can be used for arbitrary pedigrees and a broad class of genetic models. These include models allowing incomplete or age-dependent penetrance, or a major locus and individual-specific environment. An updated version of the SIMLINK program employing our method is in preparation and will be made available this fall. Current users of SIMLINK and LODSTAT will be sent the updated version as soon as it is available. Other interested individuals may request the programs by writing:

Michael Boehnke, Department of Biostatistics
School of Public Health, University of Michigan
Ann Arbor, Michigan 48109

Boehnke M (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 39:513-527.

Boehnke M, Ploughman LM (1986) Estimation of the power of a proposed linkage study: computer programs for simulation. Am J Hum Genet 39:A148.

Ploughman LM, Boehnke M (1988) In preparation.

Submitted by Michael Boehnke and Lynn M. Ploughman
Department of Biostatistics, University of Michigan

COMMENTS FROM READERS

From Dr. A.W.F. Edwards, Department of Community Medicine,
University of Cambridge, Fenner's, Gresham Road,
Cambridge, CB1 2ES, England:

In QUESTIONS AND ANSWERS (Linkage Newsletter 1/2, Dec. 1987) it was stated that 'The application of results from linkage analysis to risk calculation requires the use of confidence limits for the recombination fraction that was estimated'. This is not so. Any consistent theory for the prediction of risk in terms of probabilities _must_ invoke the likelihood principle (for which see my book _Likelihood_*, p. 30). Confidence limits are not appropriate; the correct procedure is to use the complete likelihood function and derive from it the corresponding likelihood function for the risk probability. A relevant genetical example is given on pp. 61-63 of _Likelihood_; an example involving the calculation of the risk of nuclear accidents may be found in _Nature_ 324, 417-8 (1986). Of course, if a _decision_ has to be made, decision theory tells us that there is then no alternative to assuming a Bayesian prior for the recombination fraction.

*) Ref. in Linkage Newsletter 1/2; also now in paperback (1984, 1987). ■

From Dr. Martin Farrall, M.D., London, England:

It is good practice to determine the "power" of a sample of families once thay have been collected (and before any typings are undertaken) in anticipation of a linkage study. Analytic methods are suitable for simple structures (see J. Ott, "The

Analysis of Human Genetic Linkage") and simulation methods may be used for more complex pedigrees (e.g. M. Boehnke's SIMLINK program).

Another simple method involves first calculating the maximum lod score for the family (usually with the aid of LIPED or LINKAGE) by 'making' up phenotypes that are fully informative and cosegregate with no cross-overs. This maximal lod score (calculated with $\theta = 0.0$) is used to estimate the equivalent number of phase-known meioses (John Edwards has provided suitable formulae in the past). It is then trivial to calculate the lod scores under imaginary conditions, e.g. the maximal lod score if two thirds of the meioses are informative and there's 10% recombination.

This should give a 'feel' for the "power" of the study which although rough and ready, is very easy to do. ■

A similar view has been expressed by Dr. Aravinda Chakravarti, Pittsburgh, who suggests first calculating the maximum lod score. If that is only equal to 1, for example, there might be little need to go into more detailed analyses.

SOFTWARE NOTES

My Linkage Utility programs (version 2.1, list available on request) have been modernized and updated to run under Turbo Pascal version 4. Two bugs have been eliminated. One was in the CHIPROB program which calculates the p-value for a chi-square variable with given number of degrees of freedom; when chi-square was larger than 169, due to a programming error, the p-value returned may have been larger than 1. The other bug was in the ASSOC program which partitions the usual chi-square for the association between the phenotypes at two codominant loci into two portions, one due to allelic association and the other due to all other interactions; the algorithm used in the previous program version did not always produce reliable results, particularly for small numbers of observations. Users of the ASSOC program are encouraged to ask for the new version (please send a disk, 5¼" or 3½").

In the LINKAGE program package, several improvements have been made in the transition from version 4.6 to 4.7, but the small number of remaining bugs can still make life difficult for the unwary. Here is a list of the problems I know about.

- When you invoke the programs through LCP and want to calculate genetic risks, the risk locus must not be the last locus, otherwise the SETUP program aborts with an error message. You may, however, use any of the calculating programs (eg, MLINK) without problem.
- In MLINK, risks for homozygous normal and affected are

reversed on screen but correct in the output file (FINAL.OUT).

- In ILINK, the female/male distance ratio cannot be kept fixed unless it is equal to 1.
- When more than 9 loci are present, the PREPLINK program will write the 2-digit locus numbers without intervening spaces in the line for locus order. The analysis programs (MLINK etc.) are then unable to read the locus numbers properly and (may) report error messages difficult to interpret. The cure is to manually insert spaces in the output file produced by PREPLINK (the present version sent out by me does things right).
- The MAKEPED program cannot handle ID numbers larger than 700. It is generally preferable to use names rather than ID numbers in which case this problem is avoided.
- MAKEPED does not object when in a pedigree two unrelated sets of individuals exist (it does detect a single unrelated individual). MLINK, on the other hand, will use only one set of related individuals and completely disregards the others. Often, presumably, it is due to an error that sets of unrelated individuals occur in the same pedigree. Note that the LIPED program takes any number of unrelated sets of relatives in a single pedigree into account but it does not make the user aware of the presence of unrelated individuals.
- MAKEPED may not properly brake loops. The following example shows two bugs in the program.

```
        [1]-----(2)
             |
             |
          -------
         |        |
(5)-----[3]      [4]-----(6)
     |                |
     |                |
    [7]--------------(8)
             |
             |
            [9]
```

When the individuals in the pedigree above are given in the order of their ID numbers and the MAKEPED program is asked to break the loop at individual no. 7, it produces the following pedigree file:

```
1 1 0 0 3 0 0 1 1  1 1 2  Ped: 1  Per: 1
1 2 0 0 3 0 0 2 0  1 2 2  Ped: 1  Per: 2
1 3 1 2 7 4 4 1 0  2 1 2  Ped: 1  Per: 3
1 4 1 2 8 0 0 1 0  1 2 2  Ped: 1  Per: 4
1 5 0 0 7 0 0 2 0  1 1 2  Ped: 1  Per: 5
1 6 0 0 8 0 0 2 0  1 1 1  Ped: 1  Per: 6
1 7 3 5 0 0 0 1 2  2 1 2  Ped: 1  Per: 7
110 0 0 9 0 0 1 2  2 1 2  Ped: 1  Per: 7
1 8 4 6 9 0 0 2 0  1 1 2  Ped: 1  Per: 8
```

```
1 9 7 8 0 0 0 1 0  1 1 2  Ped: 1  Per: 9
```

Individual 9 is now said to have mother "8", which is fine, and father "9" which is wrong - it should be "10", the 'double' of individual no. 7 without parents in the pedigree. In addition, the ID number of individual 10 is written without a space after the pedigree number which leads to the analysis program to read a pedigree number of 110. In contrast, the PEDPOINT program, version 3.5, processes the pedigree correctly:

```
   1   5   3   4   0   0   0 1 2   2 1 2   id=   7
   1   4   0   0   5   0   0 2 0   1 1 2   id=   5
   1  10   9   8   0   0   0 1 0   1 1 2   id=   9
   1   9   0   0  10   0   0 1 2   2 1 2   id=   7
   1   8   6   7  10   0   0 2 0   1 1 2   id=   8
   1   7   0   0   8   0   0 2 0   1 1 1   id=   6
   1   6   1   2   8   0   0 1 0   1 2 2   id=   4
   1   3   1   2   5   6   6 1 0   2 1 2   id=   3
   1   2   0   0   3   0   0 2 0   1 2 2   id=   2
   1   1   0   0   3   0   0 1 0   1 1 2   id=   1
```

USEFUL HINTS FOR LINKAGE ANALYSTS

The ILINK program in the LINKAGE package contains a constant, NBIT, which before compilation has to be set to the "number of bits of machine precision", that is, to the length of the mantissa of real (floating point) numbers. As distributed, NBIT has been set equal to 23 which in several compilers is the mantissa length (excluding the sign bit) for single precision variables. However, in Turbo Pascal, the mantissa is 39 bits long without 8087 support and 52 bits long with 8087 support. Thus, it will be beneficial to adjust NBIT accordingly. In the ILINK versions presently sent out by me, NBIT is a variable whose value is determined by a small routine, 'precision', inside the program (within the 'initilink' overlay procedure). That routine reads as follows:

```
procedure precision(VAR nbit:integer);
{ Determines number of bits of machine precision, ie, length of
mantissa.  Based on BYTE magazine, Feb. 1985, page 231 }
CONST
  one = 1.0;
  zero = 0.0;
  minusone = -1.0;
VAR
  radix,precision,width,wide : real;
  x,y,z : real;
BEGIN {precision}
  wide := one;
  REPEAT
```

```
    wide := wide + wide;
    x := wide + one;
    y := x - wide;
    z := y - one;
  UNTIL (minusone + ABS(z)) >= zero;
  y := one;
  REPEAT
    radix := wide + y;
    y := y + y;
    radix := radix - wide;
  UNTIL radix <> zero;
  precision := zero;
  width := one;
  REPEAT
    precision := precision + one;
    width := width * radix;
    y := width + one;
  UNTIL (y-width) <> one;
  nbit := round(precision);
END; {precision}
```

A bivariate table of lod scores for the male and female recombination fraction, *m* and *f*, is easy to obtain in LIPED. The structure of the LINKAGE programs, however, makes this more difficult as these programs work in terms of the male recombination fraction, *m*, and the ratio, *r*, of the female to male map distance (Haldane measure). The following conversion formula yields the female recombination value, *f*, from given *m* and *r*:

$$f = \frac{1}{2}[1-(1-2m)^r].$$

Conversely, for given male and female recombinations, *m* and *f*, the corresponding ratio of female to male map distance is obtained as

$$r = \log(1-2f)/\log(1-2m),$$

where log is the logarithm to any base. For example, the following table shows the *r* values required for each given pair of male and female recombination fractions:

	Female recombination fraction, *f*				
m	0.01	0.10	0.20	0.30	0.499
0.499	0.0032	0.036	0.082	0.147	1
0.30	0.022	0.244	0.557	1	6.782
0.20	0.040	0.437	1	1.794	12.17
0.10	0.091	1	2.289	4.106	27.85
0.01	1	11.04	25.3	45.4	307.6

QUESTIONS AND ANSWERS

Q: How many loci can be analyzed at once with ILINK on a 256K PC? How do you increase the number to 5 or more? (JS)

A: Parameters such as the max. number of loci, individuals, pedigrees etc., have to be set as constants at the beginning of the program, source code before compilation. The values of these constants depend on each other in the sense that one is able to run a larger number of loci with only a few pedigrees than with a large number of families/individuals. One will simply have to try out what works and what does not. Two hurdles will have to be taken: (1) Some constellations of constant declarations will make it impossible to even compile the program, usually because the arrays would occupy more than 64K of space. (2) After successful compilation, a set of data may not run because there is not enough memory (RAM) in the computer; this latter problem can often be overcome by reducing the number of indivuals per run. ■

Q: How do -2 Ln(L) and location score in LINKMAP relate to likelihood? (JS)

A: It is somewhat confusing that various different transformations of the likelihood are in use, and it is important to distinguish them carefully. The common denominator is the likelihood, L, which, formally being a probability, is a number between 0 and 1. Below, log denotes the logarithm to base 10, and Ln denotes the natural log (base e).

The MLINK program, when analyzing 2 loci, reports the lod score defined in the standard manner, $Z = \log[L(\theta)/L(\frac{1}{2})]$. With more than 2 loci, it reports a quantity called LOG LIKE DIFFERENCE which is equivalent to the former LOCATION SCORE in LINKMAP and is obtained by varying one of the recombination fractions, θ, in a given interval while holding the recombination fractions in the other intervals constant. Specifically, the LOG LIKE DIFFERENCE is the difference, D, obtained by subtracting $-2\ \mathrm{Ln}[L(\theta)]$ from $-2\ \mathrm{LN}[L(\frac{1}{2})]$. Note that this is not a lod score which for this situation may be defined as $Z = \log[L(\theta)] - \log[L(\frac{1}{2})]$ so that it represents one of the possible extensions to the multipoint case of the standard lod score. The two quantities, D and Z, are related by the formula $Z = D/4.6$. ■

Q: How do you account for interference in the calculations when using the LINKAGE programs? (JS)

A: Interference can only be allowed for when 3 loci are analyzed. There are two possible options, (1) one either invokes a mapping function that is one of the first procedures in the programs, or (2) interference is defined through 3 recombination fractions, that is, the 2 θs in the two intervals and the θ between the two flanking loci. One chooses between these options through a switch (1 or 2) on one of the last lines in the data-file input file (see documentation) where a switch value of 0 specifies absence of interference. With more than 3 loci, absence of interference is assumed. It is, of course, always possible to convert resulting θ estimates to map distances using a mapping function such as Kosambi's that has interference built in, but the programs will carry out internal calculations (probabilities of haplotypes, joint recombination events) under absence of interference whenever the number of loci is larger than 3. ■

COMPARISONS AMONG SOME OF THE FASTER MS-DOS COMPUTERS

To compare execution speed between several of the newer microcomputers regarding linkage analyses, the example pedigree in Lathrop and Lalouel, Am J Hum Genet 42/3, 1988, p.502, was used for test runs. No special approximations or other measures to reduce execution time were taken. In individual 32, the third marker phenotype was changed form 11 to 12 (incompatibility). The MLINK program was used with the {R+} switch which allows range checking in arrays but adds 15-20% execution time (compiled with 8087 support). 4 loci total were used. All machines, below, had an 80287 coprocessor installed except for the Model 70 which had an 80387. The execution times given in the following table reflect the time required by MLINK per likelihood calculation (the UNKNOWN program was run prior to the test runs).

	Execution	Relative
Microcomputer	time	speed
IBM PS/2 Model 70, 20 MHz	42 sec.	3.4
PC Designs GV386, 16 MHz	79 sec.	1.8
Compaq 386, 16 MHz	73 sec.	1.9
Compaq 286, 8 Mhz	141 sec.	1

On the IBM PS/2, the calculations were also carried out with MLINK compiled by Turbo Pascal version 4. They took 12 seconds only. However, under Turbo version 4, the programs sometimes yield wrong results whose causes are still unknown.

COMPUTER PROGRAMS

Several descriptions of computer programs were received in response to the questionnaire in the last Newsletter. Due to lack of time, a list of available programs will appear in the next issue of the Newsletter only.

Jurg Ott
Columbia University, Box 58
722 West 168 Street
New York, NY 10032

Tel. (212) 960-2504
FAX (212) 795-5886
EARN/Bitnet: OTT@NYSPI