3. Van Duijn, C. M., Clayton, D., Chandra, V. *et al.*, *Int. J. Epidemiol.*, suppl. 2, 1991, **20**(2), S13–S19.
4. Graves, A. B., White, E., Koepsell, T. D. *et al.*, *Ann. Neurol.*, 1990, **28**, 766–774.
5. Cook, R. H., Ward, B. E., Austin, J. H., *Neurology*, 1979, **29**, 1402–1412.
6. Goudsmit, J., White, B. J., Weitkamp, L. R. *et al.*, *J. Neurol. Sci.*, 1981, **49**, 79–89.
7. Nee, L. E., Polinsky, R. J., Eldridge, R. *et al.*, *Arch. Neurol.*, 1983, **40**, 203–208.
8. Foncin, J. F., Salmon, D., Supino-Viterbo, V. *et al.*, *Rev. Neurol.*, (*Paris*) 1985, **141**(3), 194–202.
9. Bird, T. D., Lampe, T. H., Nemens, E. J. *et al.*, *Ann. Neurol.*, 1988, **23**, 25–31.
10. Bird, T. D., Sumi, S. M., Nemens, *et al.*, *Ann. Neurol.*, 1989, **25**, 12–25.
11. Martin, J. J., Gheuens, J., Bruyland, M. *et al.*, *Neurology*, 1991, **41**, 62–68.
12. Heston, L. L., White, J. A. and Mastri, A. R., *Arch. Gen. Psychiatry*, 1987, **44**, 409–411.
13. Nochlin, D., Sumi, S. M., Bird, T. D. *et al.*, *Neurology* 1989, **39**, 910–918.
14. Zhang M. Y., Katzman, R., Salmon, D. *et al.*, *Ann. Neurol.*, 1990, **27**, 428–437.
15. Goate, A., Chartier-Harlin, M. C., Mullan, M. *et al.*, *Nature*, 1991, **349**, 704–706.
16. Schellenberg, G. D., Anderson, L., O'dahl, S. *et al.*, *Am. J. Hum. Genet.*, 1991, **49**, 511–517.
17. Mohs, R. C., Breitner, J. C. S., Silverman, J. M. *et al.*, *Arch. Gen. Psychiatry*, 1987, **44**, 405–408.
18. Breitner, J. C. S., Silverman, Mohs, R. C. *et al.*, *Neurology*, 1988, **38**, 207–212.
19. Naruse, S. *et al.*, *Lancet*, 1991, **337**, 978–979.

Principles of linkage analysis applied to genetic mapping of familial Alzheimer's disease

Ellen M. Wijsman

Division of Medical Genetics, Department of Medicine, RG-25 and Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Linkage analysis of Alzheimer's disease is technically very demanding because of the tremendous amount of missing data inherent in such a late-onset disease. However, given careful attention to the basic principles which underly the use of the analysis techniques, it should be possible to map genes responsible for the disease. Genetic heterogeneity, difficulties in determining penetrance functions, and difficulties in performing the computations are all complications in the analysis, but these issues should not be used as license to make decisions which violate assumptions behind the basic methods used.

Introduction

ALZHEIMER'S disease (AD) is the most common form of dementia in the elderly, claiming as victims as many as a quarter of individuals who live into or beyond their eighties[1]. A genetic component to this disease is strongly suggested by a number of studies[2–5], and is supported by the existence of a number of pedigrees with early onset disease in which approximately half the family members are affected in multiple generations (reviewed by Bird, this issue), as is expected for a dominant disorder. Because of this evidence, several groups have taken up the challenge trying to map, and eventually identify, the gene(s) responsible for this disease.

AD poses many challenges in the use of statistical techniques to map genes contributing to the disease. As a result, these studies are much more difficult for AD than for most Mendelian genetic diseases, and have led to controversies in the interpretations of the results of linkage studies, for example, with markers near the centromere on chromosome 21 (refs. 6–10). One explanation of the results is that there is genetic heterogeneity for the disease[7,11]. While there is now excellent evidence that there are several different mutations in the same codon in the amyloid precursor protein (APP) on chromosome 21 which cause AD[12,13], only a small handful of families so far studied has had these mutations. No sporadic cases have been found with these mutations, indicating that they are not responsible for the majority of AD in the populations so far studied, including much of the early onset familial AD[14]. As efforts continue to identify other genes responsible for this disease, similar controversial results are likely to occur; for example, suggestive reports of linkage to chromosome 19 have been

reported[15]. In order to understand apparently contradictory linkage results, and in order to plan future studies, it is useful to consider the principles behind the statistical analysis, as well as potential problems which are likely to occur in application of these principles to linkage analysis of familial AD.

Principles of linkage analysis

The Lod score method of linkage analysis, first elaborated by Morton[16], is by far the most frequently used method for gene mapping in humans. This method is based on likelihood ratio testing set in the context of sequential analysis, and assumes that the disease being mapped is a single-locus disorder with a well-defined mode of inheritance. The disease need not be fully penetrant, but the penetrance is assumed known. The usual test statistic is expressed as a base 10 logarithm, or Lod score, of the likelihood ratio of the data given a specific recombination fraction vs. a recombination fraction of 0.5 (free recombination). Lod scores are additive across independent pedigrees, and accumulate over time as the study proceeds. The recombination fraction at which the Lod score is maximized is the maximum likelihood estimate of the distance between the disease and marker. In analyses performed with a single pair of loci, the interpretation of the meaning of significant evidence for linkage (Lod score >3) or against linkage (Lod score <-2) is straightforward when assumptions behind the method are met. An odds ratio of 1000:1 (Lod score $=3$) gives a testwise significance level of approximately 2–5% when a single marker locus has been tested because of the low prior probability of linkage of a random pair of loci[17]. For a small number of randomly chosen markers, the use of a Lod score of 3.0 as a critical point continues to yield similar testwise significance levels[18]. However, use of large numbers of markers or selection of multiple linked markers because of initial 2-point results requires a correction for the multiple tests performed[19,20].

The Lod score method is ideal for use in human genetics for a number of reasons. Because it is based on maximum likelihood methods, it is highly efficient, and can use all the data from pedigrees of arbitrary structure with missing data. If ascertainment and diagnostic procedures are appropriately followed, this method requires an average of about 50% of the sample size required by fixed sample techniques to obtain significant evidence in favor of linkage[16]. In addition, it is highly advantageous to be able to analyze data as it accumulates since pedigree collection is slow. Significant results obtained in a subset of the data may then be confirmed in additional independent data. Also, the cumulative results may be changed if diagnoses change, as is likely for AD where previously unaffected individuals become affected over time. As long as pairwise results are available for individual families at a common set of recombination fractions, new results can be pooled with the old results for an overall estimate of current results.

Parameter choices

The outcome of a linkage analysis is not only a function of the actual data collected, but also of investigator-specified parameters and decisions made about what data to include in the analysis. This is particularly true in linkage analyses of AD where large amounts of data are usually missing, and the maximum likelihood approach uses the parameter values supplied to calculate all possible outcomes of the missing data which are compatible with the observed data. Poor choice of parameters can obscure evidence for linkage[21] or can inflate evidence against linkage by increasing the apparent number of recombinant events scored in a set of pedigrees. In rare cases, inflated evidence in favor of linkage may even result[22] although in most cases misspecification of parameter values has the effect of reducing power to detect linkage[23]. Inflated evidence against linkage can be very damaging since once evidence against the existence of the gene in a region is strong, the region will probably not be reconsidered until the gene has not been mapped elsewhere in the genome[24]. For complex diseases such as AD, it is imperative that the precise assumptions and values of parameters should be made known when an analysis is presented since the choice of such parameters implies certain assumptions. If different laboratories obtain different results in an analysis, it is then at least possible to determine whether this might have been caused by different conditions of analysis.

The values of several parameters are needed for the computation of the likelihood of the observed data in the pedigrees. These include the gene frequency of the disease allele, penetrance functions of the disease for the genotypes at the AD locus, and marker allele frequencies. The marker gene frequencies, including haplotype frequencies for multi-locus analyses, are used to account for the large number of missing marker genotypes for deceased pedigree members inherent in the study of very late onset diseases. Penetrance functions are used to compute the probability that an individual with a particular phenotype (affected or unaffected) has a particular disease-locus genotype, and account for the possibility that an apparently unaffected individual may be a gene carrier, or that an affected individual is a sporadic (non-genetic) case. The gene frequency for the disease locus provides the probability that an unaffected spouse might be a gene carrier. For an uncommon disease such as AD, this may be

considerable because of the late age of onset of the disease, and unaffected individuals who marry into a pedigree should not be assumed to be non-carriers for the disease allele. This latter situation can result in data which have a likelihood of 0 (Lod score of minus infinity) at 0% recombination if, e.g., two affected siblings are discordant for their marker genotypes. Similarly, a person from an early-onset family who becomes affected late in life should not be discarded from the analysis, or assumed to be a non-carrier for the AD gene. Appropriate allowance for sporadic cases by specifying a non-zero penetrance for the non-carrier genotype takes into account the relative probability of such an event as a sporadic vs. a genetic case.

Unfortunately, it is difficult to accurately define the model parameters for AD, in particular the penetrance functions, largely because of the late onset of the disease. Especially in the later onset families, the probability is high that there is death in gene carriers from another cause before symptoms of AD become apparent, resulting in underestimation of the mean age of onset. This bias in the estimate of the mean age of onset can be as high as several years[25]. It is even harder to estimate the family-specific variance in age of onset because of the small sample sizes. A variance estimate which is too small is more likely to cause inflation of the magnitude of the Lod score than one which is too large. Therefore, the solution of estimating the variance from a group of families should produce estimates which provide a more robust analysis because different families tend to have strongly clustered ages of onset, while differing among families.

Because the results of linkage analyses for AD are usually heavily dependent on the data available in unaffected individuals, it can be desirable to determine whether the results of a linkage analysis (both positive and negative) are based largely on the more secure data (the affected individuals) or the data which depend more heavily on the model parameters. This can be assessed by using the disease phenotypes of only the affected individuals in the analysis. These are the individuals for which the inference about the disease genotype is strongest and least sensitive to these other assumptions. The analysis can be done by discarding the phenotypes of the unaffected individuals, by assuming a constant, low penetrance for the disease (such as 1% probability of disease given carrier genotype), or by using methods such as the affected pedigree member (APM) method[26]. The APM method has low power, and the results are often sensitive to how gene frequencies are used to weight the evidence in favor of linkage, but the method has the advantage that it does not depend on the assumption that a single locus is segregating in the disease pedigrees.

Because the massive amounts of missing data intrinsic in linkage analyses of AD tend to cause severe reductions in power to detect linkage, when linkage results are suggestively positive, it is tempting to try to maximize the Lod scores by varying the values of several parameters in addition to the recombination fraction. Examples of this include the penetrance functions, gene frequency, use of a multipoint analysis, or even the markers used in the analysis. Unfortunately, once additional parameters are allowed to vary, the critical Lod score must increase above 3 to achieve a testwise significance level which is equivalent to that obtained when only the recombination fraction is allowed to vary in the analysis[23]. The significance level attached to Lod scores which reach 3 only in the context of multipoint analysis with several linked markers remains to be determined, especially when the multipoint analysis is performed because of suggestive results obtained in a previous two-point analysis[20]. A requirement for a higher Lod score in order to retain an appropriate significance level may paradoxically produce the effect that in an attempt to increase power by improving multiple parameter estimates, the power may actually decrease. It is therefore best to decide ahead of time what parameter values to use in the analysis.

Decisions about data inclusion

It is not only necessary to clearly specify the parameters used in an analysis, but also specify all aspects of the diagnostic procedures and reasons for including pedigrees and individuals. Examples for AD might be autopsy documentation or age of onset. This is important information when there is evidence that there might be heterogeneity and that clinical or other criteria may differentiate between groups of pedigrees. Linkage analysis can be performed in the presence of genetic heterogeneity. The analysis can be performed on subgroups of a total data set, or on the whole data set and then analyzed for the presence of heterogeneity[23]. However, the reason(s) for subdividing data sets in order to detect this heterogeneity must be made prior to any analyses with the data because one important assumption made in using this approach is that selection of disease phenotype and marker genotypes is made independently of each other. Once the results of an analysis are known to individuals making decisions about whether or not to include a pedigree in the analysis, it is highly unlikely that the decision can be made blind to the effect of including or excluding the pedigree. The cumulative nature of the Lod score means that any experimenter bias which favors one result or the other will tend to cause spurious cumulative Lod scores.

This principle of making decisions about which data to include before the analysis is performed also applies

to selection of individuals within pedigrees and to selection of pedigrees for analysis. It is possible to collect for mapping purposes pedigrees which are heavily loaded with affected individuals. There should be no bias in estimation of the recombination fraction as long as all such collected individuals are included in the analysis once marker data are available on members of the pedigrees. However, it is important not to let the results obtained for individual pedigrees decide whether or not to include the pedigrees in the analysis (unless the results are statistically significant for the individual pedigrees). The cumulative nature of the Lod score implies that any bias towards, say, labelling families as of the 'linked' or 'unlinked' variety based on non-significant Lod scores will tend to cause false evidence in favor of or against the hypothesis of linkage.

Because of the difficulty of diagnosing the disease, the same caution holds for the diagnosis of individuals in the pedigrees after marker data have been collected. For AD, such diagnoses are often made after analysis of a pedigree has begun. In order to avoid introducing the sort of bias mentioned above, once marker data have been collected on these family members, the diagnostician must remain blind to the marker data if changes in the diagnoses are to be made in the same families. In addition, once a decision has been made to include a family or individual in the analysis, the marker data must not be used to determine whether or not to reevaluate any of the decisions to keep or exclude individuals which were made prior to the collection of the marker data.

Computational difficulties

The number of genotypic unknowns caused by missing data in AD can cause severe computational problems. In a pedigree with only 12 untyped individuals and a single marker with 4 alleles and 10 potential genotypes, there may be as many as 10^{12} possible realizations of possible combinations of marker genotypes to use in computing the weights. While clever algorithms[27] reduce considerably the number which actually need to be computed, it is still important to minimize the number of alleles necessary in the analysis by recoding genotypes of married-in spouses when possible, and to analyze pedigrees in groups defined by the alleles shared among members of the group. If these cautions are taken, 2-point analyses are feasible on a workstation, even for highly polymorphic markers. The analysis time is dependent on both the pedigrees analyzed and the computer used to do the analysis, but is unlikely to require more than a few days of computer time, even if not all pedigrees in the analysis have a simple structure.

Multipoint analyses of AD, on the other hand, are not currently feasible for routine analyses. Because the computations increase exponentially with the number of unknowns, the addition of even a few additional marker loci increases substantially the computer time required for the computations. For example, a recent multipoint analysis of 22 of our pedigrees for three chromosome 21 markers (4 alleles each) required approximately 3 months of VAX 8820 CPU time.

The computational difficulty of these analyses for AD increases rapidly with the number of alleles at the marker loci. This is because with the large number of missing founders in the pedigrees, it is not possible to reduce the number of allele labels in a pedigree to only three or four as is possible in fully typed pedigrees[28]. This is unfortunate since at least in principle, highly polymorphic markers can be used to somewhat compensate for the lower power of AD pedigrees. For example, the power of our 50 late onset pedigrees to detect linkage at 5% recombination increases from less than 40% for a marker with PIC=0.5 (ref. 29) to 70% for a marker with PIC=0.7. Note, however, that this requires a fairly large sample size of over 400 pedigree members, approximately 25% of which are available for marker typing.

Conclusions

Linkage analysis of AD pushes our techniques of analysis to the limit. The principles behind the analysis are no different than for early onset diseases with known modes of inheritance. However, because of the difficulty of carrying out the computations, the low power of individual data sets, and the probable heterogeneity in the disease, it is more important than usual to make decisions which obey the rules behind sequential sampling procedures, and to avoid procedures which simultaneously estimate multiple parameters from the data. Positive reports of linkage are more difficult to confirm for this disease than many others, even when the results are true, because of the problem of low power. It therefore becomes tempting to try to find conditions which will maximize the Lod score, such as penetrance probabilities or diagnostic criteria such as age-of-onset, even though such maximization over multiple parameters is known to cause an excess of false positive results[23]. Therefore, controversy is more likely to result from reports of linkage for AD than for analyses of early onset dominant diseases. Unnecessary and pointless controversy can be avoided by making decisions about analysis conditions before marker data are collected, by performing all diagnoses and decisions about who to include in the analysis blind to marker data, and by reporting in detail all assumptions made in the analysis.

If these simple guidelines are followed, then discussions of how best to analyze the data can more fruitfully be directed towards the question of how best to maximize power to detect linkage with these pedigrees, and towards the difficult question of how to actually get the relevant computations done.

1. Breitner, J. C., Silverman, J. M., Mohs, R. C. *et al.*, *Neurology*, 1988, **38**, 207–212.
2. Heyman, A., Wilkinson, W. E., Hurwitz, B. J. *et al.*, *Ann. Neurol.*, 1983, **14**, 507–515.
3. Mohs, R. C., Breitner, J. C., Silverman, J. M. and Davis, K. L., *Arch. Gen. Psychiatry*, 1987, **44**, 405–408.
4. Graves, A. B., White, E., Koepsell, T. D. et al., *Ann. Neurol.*, 1990, **28**, 766–774.
5. Van Duijn, C. M., Clayton, D., Chandra, V. *et al.*, *Int. J. Epidemiol.*, 1991, **20**(2) (suppl 2), s13–s19.
6. St. George-Hyslop, P. H., Tanzi, R. E., Polinsky, R. J. *et al.*, *Science*, 1987, **235**, 885–890.
7. Schellenberg, G. D., Bird, T. D., Wijsman, E. M. *et al.*, *Science*, 1988, **421**, 1507–1510.
8. Pericak-Vance, M. A., Yamaoka, L. H., Haynes, C. S. *et al.*, *Exp. Neurol.*, 1988, **102**, 271–279.
9. Goate, A., Haynes, A. R., Owen, M. J. *et al.*, *Lancet*, 1989, **i**, 352–355.
10. Schellenberg, G. D., Pericak-Vance, M. A., Wijsman, E. M. *et al.*, *Am. J. Hum. Genet.*, 1991, **48**, 563–583.
11. St. George-Hyslop, P. H., Haines, J. L., Farrer, L. A. *et al.*, *Nature*, 1990, **347**, 194–197.
12. Goate, A., Chartier-Harlin, C. M., Mullan, M. *et al.*, *Nature*, 1991, **349**, 704–706.
13. Naruse, S., Igarashi, S., Aoki, K. *et al.*, *Lancet*, 1991, **337**, 978–979.
14. Schellenberg, G. D., Anderson, L.-J., O'Dahl, S. *et al.*, *Am. J. Hum. Genet.*, 1991, **49**, 511–517.
15. Pericak-Vance, M. A., Bebout, J. L., Gaskell, P. C. *et al.*, *Am. J. Hum. Genet.*, 1991, **48**, 1034–1050.
16. Morton, N. E., *Am. J. Hum. Genet.*, 1955, **7**, 277–318.
17. Haldane, J. B. S. and Smith, C. A. B., *Ann. Eugenics*, 1947, **14**, 10–31.
18. Elston, R. C. and Lange, K., *Ann. Hum. Genet.*, 1975, **38**, 341–350.
19. Thompson, E. A., *Genet. Epidemiol.*, 1984, **1**, 314–331.
20. Edwards, J. H., *Ann. Hum. Genet.*, 1990, **54**, 253–275.
21. Greenberg, D. A. and Hodge, S. E., *Genet. Epidemiol.*, 1989, **6**, 259–264.
22. Green, P., *Comment. Genet. Epidemiol.*, 1990, **7**, 25–27.
23. Ott, J., *Analysis of Human Genetic Linkage*, Revised edition. The Johns Hopkins University Press, Baltimore, 1991.
24. Risch, N., *Genet. Epidemiol.*, 1990, **7**, 41–45.
25. Farrer, L. A., Myers, R. H., Cupples, L. A. *et al.*, *Neurology*, 1990, **40**, 395–403.
26. Weeks, D. E. and Lange, K., *Am. J. Hum. Genet.*, 1988, **42**, 315–326.
27. Elston, R. C. and Stewart, J., *Hum. Hered.*, 1971, **21**, 523–542.
28. Braverman, M. S., *Comput. Biomed. Res.*, 1985, **18**, 24–36.
29. Botstein, D., White, R. L., Skolnick, M. and Davis, R. W., *Am. J. Hum. Genet.*, 1980, **32**, 314–331.

ACKNOWLEDGEMENTS. This work was supported by NIH AG06781, NIH AG07584, and NIH AG05136.

The molecular genetics of familial Alzheimer's disease

Gerard D. Schellenberg

Division of Neurology, School of Medicine, University of Washington, Seattle, Washington, 98195, USA

Familial Alzheimer's disease is genetically heterogeneous. In some families, early onset of disease is caused by mutations in the APP gene. In other families, a second gene on chromosome 21, a gene on chromosome 19, or a gene located somewhere else in the genome causes Alzheimer's disease.

THE role of inheritance in the occurrence of Alzheimer's disease (AD) has come under intense scrutiny during the past decade. Epidemiologic survey studies where family history was studied have repeatedly shown that families of AD probands have more cases of AD in close relatives than families of controls[1]. Twin studies not only show that many identical twins are often concordant for AD[2] but also that the families of concordant twins have more AD cases than the families of twins discordant for AD[3]. These types of studies are highly suggestive that the inheritance of defective genes is important in the pathogenesis of AD. However, population-based studies cannot clearly determine what mode of inheritance is responsible for the observed family clustering.

Some authors have suggested that all AD is autosomal dominant[4]. This hypothesis is difficult to prove by family studies since the onset of AD typically occurs late in life and relatives of probands often die of other causes before they reach the age of appropriate risk (age censoring). Thus even if all AD was the result of autosomal dominant inheritance, in most cases of randomly ascertained AD probands, secondary cases in close relatives would not be observed due to age censoring[4]. Alternatively, single cases in a family could