STATISTICAL GENETICS '98
Complex Segregation Analyses: Uses and Limitations

Gail Pairitz Jarvik

Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine, Seattle

Complex segregation analysis (CSA) is a general method for evaluating the transmission of a trait within pedigrees. It proceeds by testing models of varying degrees of generality, both to determine whether a Mendelian locus is likely to exert a large effect on the phenotype of interest and to estimate the magnitude of genetic sources of variation in the trait. This information is valuable both as a prelude to linkage analysis, which generally assumes Mendelian transmission, and in providing a model on which to base parametric linkage methods. CSA is distinguished from (simple) segregation analysis, in that the latter evaluates whether the proportion of affected and unaffected offspring in families is consistent with Mendelian expectations. In contrast, CSA can be applied to any pedigree structure and works with both qualitative and quantitative traits. In the past, complex pedigrees might have been broken into their constituent nuclear families, and quantitative traits would have been dichotomized, but CSA obviates the need for either of these steps, either of which can lead to an unacceptable loss of transmission information. Additionally, CSA can consider more complicated patterns of transmission and environmental perturbations.

Simulation studies have consistently found CSA to be both reliable and robust. The individual contributions and summary papers for past Genetic Analysis Workshops suggest that the model underlying simulated data can be reasonably recovered by a variety of standard CSA methods, despite unmodeled effects that might be expected in a complex disorder (Blangero 1995). In addition, a simulation study (Hodge 1995) showed that genetic effects in a data set simulated for an oligogenic trait could neither be determined by CSA nor mapped by linkage, indicating that, like positive evidence for a Mendelian locus, negative results from CSA may be of great utility. In practice, there is usually considerably less information in the analysis of real diseases than would

be available in simulation studies, but, judged on the basis of its record, CSA is clearly a powerful tool in the elucidation of the genetic basis of both qualitative and quantitative traits.

Application of CSA

I will describe the CSA methods with record brevity, facilitated by the absence of any equations; more-detailed descriptions should be sought elsewhere (Pairitz et al. 1988; Khoury et al. 1993). In CSA, a variety of models are considered. Parameters commonly estimated by CSA include the transmission probabilities; the allele frequencies; either the genotype means, for quantitative traits, or the penetrances for each genotype, for qualitative traits; the variance within genotypes; and any residual genetic correlations not explained by the Mendelian locus. A general model will often contain the most parameters. This model is then compared with a Mendelian transmission model, an environmental transmission model, and a polygenic model. Under a Mendelian model, the transmission probabilities—that is, the probabilities that the AA, Aa, and aa genotypes will pass on an A allele—do not significantly differ from the Mendelian expectations of 1, .5, and 0, respectively, whereas in the general model these transmission probabilities can take any value. Under the environmental model, these probabilities are all equal—the phenotypic mode that a child is in is unrelated to the mode that the parent is in. Both the Mendelian and environmental models can contain multiple small genetic and environmental effects; however, these are the only effects in a polygenic model, which has no large deviation in the trait caused by either a major locus or the environment. Should a Mendelian model be favored in a data set, dominant and recessive Mendelian submodels can then be evaluated.

Important complicating factors include the assumption of a normal distribution of a quantitative trait, incorporation of adjustments both for covariates and for possible covariate interactions, and the adjustment for the ascertainment scheme. Although a simple application of CSA is to investigate the possible effects of a single two-allele locus with or without residual genetic correlations, CSA methods can also be extended to incorporate more-complex modes of inheritance. Finally, the

choice of the phenotype to be modeled is not always simple, particularly for a disease. Consistent, rational diagnostic criteria are required.

Models are generally compared by a likelihood ratio test. The likelihood of each model, which is proportional to the probability of the data, given the model and the family structure, is approximated or computed. If one model is a submodel of the other, then the likelihoods of the two models can be compared to test whether the fit of the restricted model is significantly worse than that of the unrestricted model, in which more parameters are estimated. Although CSA is generally done by approximation methods as implemented by the computer packages PAP (Hasstedt and Cartwright 1981; Hasstedt 1993) and SAGE (1994), newer Monte Carlo–Markov chain methods, such as that implemented in the MORGAN program (Thompson 1994), provide estimates from the true likelihood surface.

Uses of CSA

To date, LOD-method linkage analysis (i.e., "model based" linkage analysis, as described by Elston [1998], in this issue of the *Journal*) has been the most successful method in the mapping of disease loci. This method depends on specification of a reasonable approximation of the mode of inheritance. This approximation can be derived from the parameters estimated by CSA in appropriate samples. Although power may be affected, two-point LOD-method linkage analysis is reasonably robust to model misspecification, as long as the mode of inheritance is correct. Misspecification may include over- or underestimation of the disease-allele frequency, variation from the true age-dependent penetrance function, or errors in the phenocopy rate. Indeed, misspecification may even include the failure to model the presence of a second, epistatic genetic locus (Vieland et al. 1992). Model misspecification tends to result in the overestimation of the recombination fraction between the disease locus and the marker. Since parametric multipoint linkage methods measure the recombination fractions from several markers, they do not allow such overestimation, and they are therefore less robust to model misspecification (Risch and Giuffra 1992).

The existence of a Mendelian trait is an underlying assumption of linkage analyses used to map trait loci. An interesting proposal for proceeding with linkage analysis even in the absence of an accurate model of transmission comes in a pair of recent papers. Hodge et al. (1997) show that, because of the robustness of two-point LOD-method linkage analysis to model misspecification, arbitrary dominant and recessive models may often be applied to the transmission of a qualitative trait. This approach, which can lead to a variable loss of statistical power (Greenberg et al. 1998), assumes the pres-

ence of one or more genetic loci underlying the trait, and the criteria for the assessment of linkage may need to be modified when the probability of an underlying major locus is unknown. Should there be evidence of Mendelian transmission, however, the approach advocated by Greenberg et al. may be very useful when multiple loci contribute to a qualitative trait and when the model estimated by CSA does not match the transmission of some locus of interest. However, such an application does not obviate the need for CSA in general. CSA can support, although it cannot prove, the Mendelian segregation of a trait. If Mendelian segregation is not supported, analyses of candidate loci or random markers for linkage to the trait of interest would likely be unproductive, at least in the same data set. However, because CSA can both falsely reject and falsely support a Mendelian locus, it would be prudent to require consistent evidence of a Mendelian locus from several different CSA studies before linkage studies are undertaken.

CSA can also be used to further define the genetic features of a trait. It can be used to evaluate etiologic heterogeneity in a trait, either by doing CSA in defined subsets or by contrasting the likelihoods under competing models for each family. It should be noted that there are clinical applications of transmission information from CSA that are not addressed here. CSA can be used either to look for residual genetic effects after adjustment for known genetic loci or to evaluate the independence of two traits (Pairitz et al. 1988). Probable genotypes can be assigned to pedigree members, and the effects of these genotypes on other traits can then be estimated (Jarvik et al. 1994).

Limitations of CSA

The major limitation of CSA is that a large amount of a very specific type of data is generally needed. Ascertainment of an appropriate sample is also necessary. The most appropriate samples are either population based, in which case no ascertainment correction is needed, or they are selected through a phenotype-based ascertainment scheme, for which ascertainment corrections can be implemented. Adjustment for ascertainment in CSA still needs development (Wijsman and Amos 1997). For instance, there are no appropriate methods to adjust for the type of ascertainment that is generally used to collect families for linkage analysis of rarer traits. Because linkage analysis generally relies on the collection of a highly selected group of densely affected families, and because these families are rare, the model estimates for use in LOD-method linkage analysis often must be drawn from CSA on samples other than the sample of densely affected families collected for linkage analysis.

The amount of necessary data is expected to be proportionate to the number of parameters estimated, lim-

iting the ability to evaluate more-complex models. Additionally, there is no reliable method to determine the sample size required for a desired level of power to detect a Mendelian locus by CSA. Sample sizes generally range in the mid hundreds of individuals, for quantitative traits, and can be in the thousands, for rare, dichotomous traits. Larger-sized families are expected to reduce the needed sample size, as a result of the increase in transmission information. Fortunately, experience suggests that nonpaternity, at least when it occurs at an average frequency, has little effect on the model chosen under CSA, so genetic testing is usually not necessary.

Another practical limitation is the inability to distinguish between the effect of a single locus that underlies a trait and the effects of two or more independently acting loci with similar transmission patterns. For example, all dominant hereditary prostate cancer loci would be detected as if they were a single locus with a disease-allele frequency equaling that of the sum of several disease alleles. The resulting overestimate of allele frequency would result both in an overestimate of the statistical power to detect a locus, as well as in model misspecification that could reduce the power to detect each contributing locus, using linkage analysis.

Analytic limitations such as the issues of local maxima for the parameter maximum-likelihood estimates, the estimation of parameters whose expected values are on the 0 and 1 boundaries, and convergence problems are not considered here. Analytic methods for CSA continue to improve (Snow and Wijsman 1998), but programs for doing CSA remain unwieldy.

Potential Errors

In the implementation of CSA, one common error that can lead to a false conclusion of Mendelian segregation is the failure to adjust adequately for the nonrandom ascertainment of pedigrees. Should pedigrees be selected for multiple affected individuals, the disorder may appear to have a Mendelian pattern of inheritance.

Failure to test whether the transmission probabilities are Mendelian—that is, failure to contrast the general and the Mendelian model—can also lead to the false conclusion that a Mendelian locus exists, particularly in the case of skewed data (Demenais and Bonney 1989). An entertaining paper reporting that misapplied CSA might conclude that there is an autosomal recessive gene for medical-school attendance did, in fact, find that the fit of the Mendelian model was significantly worse than that of the general model (McGuffin and Huckle 1990); in other words, the Mendelian transmission probabilities could be rejected, and one could not conclude that there is a Mendelian locus for medical-school attendance. Examination of the transmission probabilities is now rou-

tinely done, but it may have been neglected in older papers in the literature.

Practical Examples of CSA

Traits for which CSA results have contributed to the identification of genes include breast cancer and prostate cancer. CSAs have supported an infrequent autosomal dominant breast cancer locus (Williams and Anderson 1984; Newman et al. 1988). Newman et al. estimated that the breast cancer–susceptibility allele has a frequency of .0006 and that carriers have an 82% lifetime risk of breast cancer whereas noncarriers have an 8% lifetime risk. The strength of this result supported attempts to map such a gene and provided parameter estimates for the linkage analyses. These efforts were successful, resulting in the mapping and subsequent cloning of the autosomal dominant breast cancer–susceptibility loci, BRCA1 (Hall et al. 1990) and BRCA2 (Wooster et al. 1994). The existence of more than one locus should not be surprising, given the preceding discussion of CSA limitations.

A similar story has emerged for prostate cancer. CSAs by several groups support an autosomal dominant prostate cancer–susceptibility locus. The early report by Carter et al. (1992) suggested that an autosomal dominant prostate cancer–susceptibility allele occurs with a frequency of .003 and confers an age-dependent risk maximizing at a penetrance of 88% by age 85 years in males. By means of models based at least in part on the autosomal dominant CSA results of Carter et al., two autosomal dominant prostate cancer–susceptibility loci have been mapped: HPC-1 (Smith et al. 1996) and PCaP (Berthon et al. 1998). As was the case for breast cancer, CSA did not distinguish the presence of more than one dominant locus for prostate cancer. Indeed, the two loci described do not appear to be segregating in many high-risk prostate cancer families, suggesting that additional loci are yet to be found (McIndoe et al. 1997; Berthon et al. 1998; Eeles et al. 1998). CSA can be used as a tool to investigate disease heterogeneity when the inheritance pattern may differ in subsets of a sample. Moll et al. (1989) exploited this in their investigation of the transmission of apolipoprotein A1. By contrasting the likelihoods of the data for each family, under alternative models, Moll et al. were able to estimate the fraction of families segregating a Mendelian locus. Similarly, by comparing individual family likelihoods for alternative models, Jarvik et al. (1994) concluded that the transmission of apolipoprotein B level is genetically heterogeneous.

In the case of a complex disease for which subsets can be defined, CSA can be performed on the subsets, and the final models can be compared by a heterogeneity test. An example of such an analysis is the division of

families with Alzheimer disease, according to whether they do or do not have an apolipoprotein E, $\epsilon 4$ allele. When models for each subset were contrasted, it was concluded that the $\epsilon 4$ group and the non-$\epsilon 4$ group had different patterns of transmission of Alzheimer disease (Jarvik et al. 1996), supporting a role for this allele in Alzheimer disease.

Using the probability of each genotype for each individual, one can also assign putative genotypes for a Mendelian locus detected by CSA. For an apolipoprotein B–elevating locus detected by CSA, assigned genotypes have been shown to predict the presence of familial combined hyperlipidemia but not of hyperapobetalipoproteinemia. The results support both the separate etiology of these two overlapping disorders and the role of the putative apolipoprotein B–elevating locus in the former (Jarvik et al. 1993, 1994).

Conclusions

To quote Elston and Stewart (1971, p. 523), "the purpose of analysing pedigree data is to establish the presence or absence of a genetic mechanism for the manifestation of a particular trait of set of traits; to elucidate such a mechanism, if it is present; and to classify individuals for their genotypes." A disease with complex inheritance, one that is not transmitted in a simple Mendelian fashion, often occurs through the action of multiple Mendelian loci, polygenic loci, and environmental components. As simply inherited traits are elucidated and mapped, we are left to the job of understanding the complex traits, which are generally of more significance in human disease. When applied to such a complex disease or trait, CSA allows us to further the goals stated by Elston and Stewart.

Acknowledgments

References

Berthon P, Valeri A, Cohen Akenine A, Drelon E, Paiss T, Wöhr G, et al (1998) Predisposing gene for early-onset prostate cancer, localized on chromosome 1q42.2-43. Am J Hum Genet 62:1416–1424

Blangero J (1995) Genetic analysis of a common oligogenic trait with quantitative correlates: summary of GAW9 results. Genet Epidemiol 12:689–706

Carter B, Beaty TH, Steinberg GD, Childs B, Walsh PC (1992) Mendelian inheritance of familial prostate cancer. Proc Natl Acad Sci USA 89:3367–3371

Demenais F, Bonney GE (1989) Equivalence of the mixed and regressive model for genetic analysis. Genet Epidemiol 6:597–617

Eeles R, Durocher F, Edwards S, Teare D, Badzioch M, Hamoudi R, Gill S, et al (1998) Linkage analysis of chromosome 1q markers in 136 prostate cancer families. Am J Hum Genet 62:653–658

Elston RC (1998) Methods of linkage analysis—and the assumptions underlying them. Am J Hum Genet 63:931–934 (in this issue)

Elston R, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Greenberg DA, Abreu P, Hodge SE (1998) The power to detect linkage in complex disease by means of simple LOD-score analyses. Am J Hum Genet 63:870–879

Hall J, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC (1990) Linkage of early onset familial breast cancer to chromosome 17q21. Science 250:1684–1689

Hasstedt S (1993) Variance components/major locus likelihood approximation for quantitative, polychotomous, and multivariate data. Genet Epidemiol 10:145–158

Hasstedt S, Cartwright P (1981) PAP: pedigree analysis package. Department of Medical Biophysics and Computing, University of Utah, Salt Lake City

Hodge S (1995) An oligogenic disease displaying weak marker associations: a summary of contributions to problem 1 of GAW9. Genet Epidemiol 12:545–554

Hodge SE, Abreu PC, Greenberg, DA (1997) Magnitude of type I error when single locus linkage analysis is maximized over models: a simulation study. Am J Hum Genet 60:217–227

Jarvik G, Beaty TH, Gallagher PR, Coates PM, Cortner JA (1993) Genotype at a major locus with large effects on apolipoprotein B levels predicts familial combined hyperlipidemia. Genet Epidemiol 10:257–270

Jarvik G, Brunzell JD, Austin MA, Krauss RM, Motulsky AG, Wijsman EM (1994) Genetic predictors of FCHL in four large pedigrees: influence of ApoB level major locus predicted genotype and LDL subclass phenotype. Arterioscler Thromb 14:1687–1694

Jarvik GP, Larson EB, Goddard K, Kukull WA, Schellenberg GD, Wijsman EM (1996) Influence of apolipoprotein E genotype on the transmission of Alzheimer disease in a community-based sample. Am J Hum Genet 58:191–200

Khoury M, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York

McGuffin P, Huckle P (1990) Simulation of Mendelism revisited: the recessive gene for attending medical school. Am J Hum Genet 46:994–999

McIndoe RA, Stanford JL, Gibbs M, Jarvik GP, Brandzel S, Neal CL, Li S, et al (1997) Linkage analysis of 49 high-risk families does not support a common familial prostate cancer–susceptibility gene at 1q24-25. Am J Hum Genet 61:347–353

Moll PP, Michels VV, Weidman WH, Kottke BA (1989) Genetic determination of plasma apolipoprotein AI in population-based sample. Am J Hum Genet 44:124–139

Newman B, Austin M, Lee M, King MC (1988) Inheritance

of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc Natl Acad Sci USA 85:3044–3048

Pairitz G, Davignon J, Mailloux H, Sing CF (1988) Sources of interindividual variation in the quantitative levels of apo-lipoprotein B in pedigrees ascertained through a lipid clinic. Am J Hum Genet 43:311–321

Risch N, Giuffra L (1992) Model misspecification and multipoint linkage analysis. Hum Hered 42:77–92

SAGE (1994) Statistical analysis for genetic epidemiology. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Smith J, Freije D, Carpten JD, Gronberg H, Xu J, Isaacs SD, Brownstein MJ, et al (1996) Major susceptibility locus for prostate cancer on chromosome 1 suggested by a genome-wide search. Science 274:1371–1374

Snow G, Wijsman EM (1998) Pedigree analysis package (PAP) vs MORGAN: model selection and hypothesis testing on a large pedigree. Genet Epidemiol 15:355–369

Thompson E (1994) Monte Carlo programs for pedigree analysis: 1990–1993. Department of Statistics, University of Washington, Seattle

Vieland V, Hodge SE, Greenberg DA (1992) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 9:45–59

Wijsman EM, Amos CI (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 14:719–735

Williams W, Anderson DE (1984) Genetic epidemiology of breast cancer: segregation analysis of 200 Danish pedigrees. Genet Epidemiol 1:7–20

Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, et al (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 265:2088–2090