Eur. J. Epidemiol. 0392-2990
Suppl. 1, 1992, p. 3-9

EUROPEAN JOURNAL OF EPIDEMIOLOGY

Vol. 8, Suppl. to No. 2

APPLICATION OF CLADISTICS TO THE ANALYSIS OF GENOTYPE-PHENOTYPE RELATIONSHIPS

C.F. SING[1], M.B. HAVILAND, K.E. ZERBA and A.R. TEMPLETON

The University of Michigan Medical School - Medical Science II M4708
ANN ARBOR - MI 48109-0618 - USA.

Key words: Atherosclerosis - Cladistics - Genetics

We seek to understand the relative contribution of allelic variations of a particular gene to the determination of an individual's risk of atherosclerosis or hypertension. Work in progress is focusing on the identification and characterization of mutations in candidate genes that are known to be involved in determining the phenotypic expression of intermediate biochemical and physiological traits that are in the pathway of causation between genetic variation and variation in risk of disease. The statistical strategy described in this paper is designed to aid geneticists and molecular biologists in their search to find the DNA sequences responsible for the genetic component of variation in these traits. With this information we will have a more complete understanding of the nature of the organization of the genetic variation responsible for quantitative variation in risk of disease. It will then be possible to fully evaluate the utility of measured genetic information in predicting the risk of common diseases having a complex multifactorial etiology, such as atherosclerosis and hypertension.

INTRODUCTION

Numerous quantitative biological traits contribute to determining an individual's risk of developing a common complex disease such as atherosclerosis or hypertension (2). Dietary and life style effects interact with these traits to determine a continuous distribution of risk among individuals in the population at large. The etiological relationships among the many causal factors that underlie such continuous variation are expected to be complex.

What is the evidence that leads one to conclude that diseases like coronary artery disease (CAD) and hypertension have complex etiologies? It is clear that each of these disorders aggregates in families, but disease is not distributed among relatives in a manner consistent with a simple Mendelian model of inheritance. The phenotypic possibilities for the many biological risk factor traits that influence risk of disease are continuously distributed among relatives. There is no known combination of phenotypes in an individual for which risk is totally absent or disease an absolute certainty. All studies to date suggest that interindividual variability in every one of these biological risk factor traits is explained by the interaction of the effects of differences in many genes with exposures to variation in numerous environmental factors (7). Furthermore, variation in a particular gene may contribute to phenotypic variation in more than one of these intermediate biological risk factor traits that are in the pathway of causation between genetic variation and interindividual variability in risk (3). Thus, the complexity of the network of causation across interdependent levels of biological organization precludes those research strategies that seek to find a one-to-one mapping between a mutation in one gene and the occurrence of disease.

[1] Corresponding author.

There is no universally accepted strategy for characterizing the paths of causation between genetic variation and interindividual variability in risk of atherosclerosis or hypertension. Most researchers begin by studying the biochemical and physiological traits that have been shown by epidemiological studies to be predictors of disease. One first seeks to answer the following questions to understand the genetic architecture of each of these intermediate risk factor traits. How many genes are involved in determining interindividual variability? How many functional alleles are there for each gene? What is the relative frequency of each allele? And, what is the impact of each allele on interindividual variation in trait levels, on intraindividual variability in levels and on the pairwise correlations between traits? Once information about genetic architecture is available it becomes possible to consider a second step that addresses the role that each gene plays in predicting risk of disease. Two questions dominate. First, what is the contribution of a gene to prediction of risk as a consequence of its impact on a particular intermediate trait? Second, does knowing an individual's genetic make-up provide information about risk that is not provided by measures of available intermediate traits? Statistical associations between allelic variation in a gene and variability in risk of disease, as a consequence of its impact on a particular intermediate trait, can suggest directions for laboratory research that will enhance our understanding of disease etiology. When genetic variation improves prediction of risk beyond that provided by available measured intermediate traits, one must conclude that there are pleiotropic effects of the gene on other unmeasured intermediate traits that are also involved in the etiology of disease. Progress in understanding etiology depends on a dynamic interplay between statistical evaluation and laboratory exploration. We consider below some aspects of the statistical work.

The bottom-up approach to defining genetic architecture of a quantitative trait

Sing and Moll (8) discuss the top-down and bottom-up strategies for unraveling the genetic architecture of a quantitative intermediate risk factor trait. In the top-down approach, a biometrical analysis is carried out to evaluate the impact of genetic variability on phenotypic differences among individuals. For example, complex segregation analysis is often used to determine if there is statistical evidence for allelic variation at a single locus with large phenotypic effects. Families segregating for large phenotypic differences are candidates for linkage studies to identify the region of the genome where the responsible gene may be located. Once there is evidence for linkage to a marked region, individuals from these families are used in molecular studies to characterize the DNA sequences responsible for the major phenotypic effects. This approach is limited to searching for those few loci that have a major impact on interindividual variability. The bottom-up approach evaluates the impact of allelic variation in candidate genes on phenotypic variability. A gene is a candidate if its product is known to be involved in the metabolism of the trait(s) of interest. We limit our discussion here to the bottom-up approach to understanding the genetic architecture of intermediate traits.

In the bottom-up approach, each candidate gene is expected to have many allelic forms (6). Most DNA sequence variations that define these gene differences are not expected to influence variation of an intermediate trait (4, 6). Hence, research on a candidate gene involves sorting out functional allelic differences from those that have no effect on the expression of the intermediate trait. When these genetic variants have been identified, the individuals who carry them can be studied in greater depth. Such studies will increase our understanding of the etiological pathways between the candidate gene and risk of disease. We have been developing a statistical method to identify the subset of haplotypes that carry functional mutations that are responsible for phenotypic differences in quantitative risk factor traits in a population. We present here an overview of this strategy. Much work remains to be done to offer a complete theory. Hence, the following is a summary of work in progress.

The use of cladistics to define genotypic variability

An overview. It is tempting to approach the problem of defining mutations in a candidate gene responsible for quantitative phenotypic effects simply by comparing the DNA sequences of individuals with extremely high values with homologous sequences from individuals with extremely low values of the trait of interest. Five limitations prevent meaningful inferences from this strategy. 1) Individuals selected may not have detectable differences in the candidate gene region because environmental factors or mutations in regions other than that sequenced may cause the observed differences in trait levels. Comparing individuals with high and low values provides no information to guide sequencing work. 2) There may be many DNA sequence differences that are not involved in determining the observed extreme phenotypic differences. Comparing individuals with high and low values does not provide a strategy for distinguishing those mutations that are responsible for the observed phenotypic differences from those that do not influence phenotypic expression. 3) There may be more than one sequence difference that will give the same observed phenotypic difference. 4) Multiple mutations in the gene may combine to produce an observed effect. Thus, the effect of any particular mutation will be small and difficult to distinguish as

being involved in determining the observed phenotypic differences. 5) Most of the mutations responsible for quantitative variation in the population at large will not result in extreme values. Sampling only individuals with extreme values will give a biased view of genetic architecture.

To address the above issues one must sample DNA variation in multiple regions of the candidate gene using a representative sample of the population at large. Restriction fragment length polymorphisms (RFLPs) provide a means to mark specific gene variants. Differences in single markers define alleles. Multiple markers of the same gene define haplotypes. Haplotypes associated with phenotypic effects are those which should be sequenced to identify the responsible mutations.

We have proposed (10) a statistical strategy that employs cladistics to identify those haplotypes which are likely to carry mutations that determine quantitative phenotypic effects. A cladistics analysis involves two basic steps. First, a cladogram is constructed using the observed haplotypes defined by simultaneously considering multiple polymorphic restriction sites of the candidate gene. Second, the cladogram is used to completely define a powerful and efficient analysis of associations between haplotype variation and phenotypic variability.

In the first step of cladogram construction, we assume the DNA region being marked is short and that recombination and gene conversion in this region are rare. Thus, each RFLP arises as a consequence of a mutation that has occurred at some point in the evolutionary history of the population. All copies of the RFLP in the contemporary population being studied may be traced through the descendants of the individual in whom the mutation first occurred. We also assume that the farther back in time that a RFLP mutation occurred, the greater the chance that a second RFLP mutation has occurred in the intervening time period in the haplotype carrying the original mutation. It follows from this basic evolutionary argument that haplotypes in the present population differing by many RFLPs will have diverged from a common ancestor longer ago than haplotypes that differ by only a few RFLPs. Conversely, we assume that the greater the similarity between two haplotypes, the closer their time of divergence in the evolution of the genetic material that is being marked. Hence, the number of RFLP differences between two haplotypes may be taken as a measure of genetic distance that roughly estimates the time of their evolutionary divergence. A cladogram is a graphical representation of these evolutionary relationships among haplotypes. It contains all of the mutational genetic distances among the haplotypes in the sample. However, the cladogram is much more than a chart of genetic distances; it is more like a map that provides detailed information on the specific mutational changes that relate one haplotype to any other haplotype in the sample. That is, the cladogram is analogous to a road map rather than a distance chart that is commonly found in an atlas.

We employ the method of maximum parsimony to construct cladograms. The objective is to link together the observed haplotypes into a network that involves the fewest mutational steps. This method is easy to implement and produces equivalent results to more complex methods under the reasonable assumption of a short evolutionary time period involved in accumulation of mutational differences between haplotypes. An illustration of the method of parsimony is given in Figure 1 where four RFLPs have been used to measure haplotype variation. Here 1 denotes the presence of a restriction enzyme cleavage (cut) site and 0 denotes the absence of a cut site. We consider a sample of haplotypes that includes four of the possible 16 haplotypes that may be distinguished by four restriction sites. They are 1111, 1100,1101 and 1001. There are 16 possible ways to link these four haplotypes into a cladogram. Three of these possibilities are also shown in Figure 1. The network that requires three mutations to explain the evolution of the four haplotypes is considered to be the most parsimonious cladogram (9). No haplotype is identified as the ancestral form; a root is not necessary for the application of the cladogram as a tool to identify those haplotypes that carry mutations with phenotypic effects.

The most parsimonious cladogram is then used in step two as a tool for the study of associations between haplotype variation and phenotypic variability. We assume that if an undetected functional mutation causing a phenotypic effect occurred at some point in the evolutionary history of the population, it will be embedded within the same historical structure represented by the cladogram. The cladogram is used to define a nested analysis of the phenotypic differences among groups of individuals that differ in degree to which their marker haplotypes differ. At each level in the hierarchy, phenotypic differences between groups of haplotypes with varying degrees of evolutionary divergence are compared. If there is a significant phenotypic difference between groups, one infers that it is attributable to a functional mutation in the gene region being marked. A nested analysis defined by the structure of the cladogram is a powerful tool which focuses the analysis of phenotypic variability on the subset of haplotype contrasts that provides the greatest amount of information about haplotypes that carry mutations having a phenotypic effect. The success of the search for this mutation is enhanced by comparison of the DNA sequences of two haplotypes that have the fewest number of RFLP differences yet are associated with a significant phenotypic difference.

Below we give a previously published example of a cladogram built using RFLPs and then turn to how it has been applied to the analysis of a sample of experimental *Drosophila* data (10). We close with a discussion of the issues that are involved in the extension of cladistics to the analysis of human quantitative data.

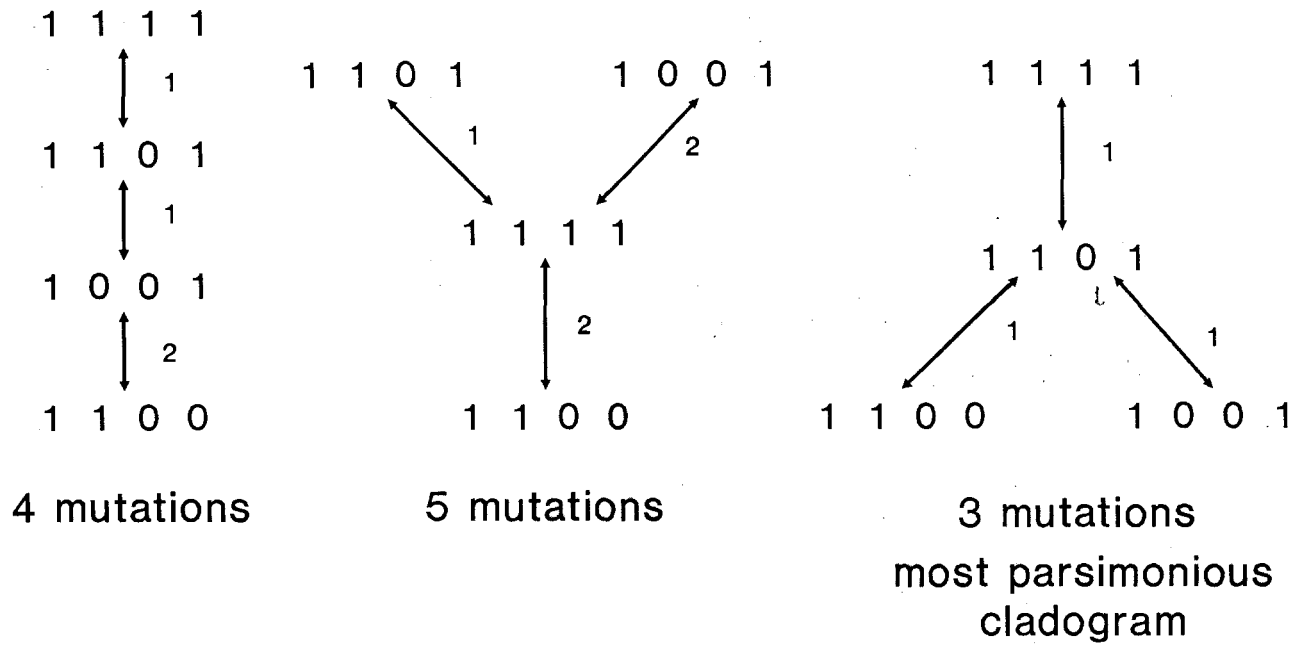

Figure 1. - 3 of 16 Possible Cladograms Involving 4 of 16 Possible Haplotypes.

An application of cladistics to Drosophila Adh data

To simplify the illustration of our approach, we consider markers in the alcohol dehydrogenase (Adh) gene region and Adh activity data collected from homozygous, inbred lines of *Drosophila melanogaster* by Aquadro et al. (1). Adh activity level is continuously distributed among 25 inbred lines. Details of the analyses to find mutations responsible for genetic variation in Adh activity among these lines are given in Templeton et al. (10). In this application of cladistics, each inbred line is homozygous for a particular haplotype. The consequences of heterozygosity encountered in natural human populations present a special problem of estimation of haplotype frequencies. Work in progress discussed below addresses this added complexity.

Figure 2 presents the most parsimonious cladogram of 25 haplotypes defined by 14 markers located in a 13-kb region of the Adh gene. Each haplotype is represented by a number. This network of haplotypes represents the minimum number of evolutionary steps that interrelate the 14 closely linked markers. For example, haplotype labeled as 25 differs from haplotype 24 at one marker site and from haplotype 22 at two marker sites. Haplotypes that are most dissimilar for marker types are positioned at the greatest distance. The O's denote intermediate haplotypes that were not present in the sample but are needed to interconnect the observed haplotypes.

A one-way analysis of variance, which ignores information about the relatedness of the 25 haplotypes contained in Figure 2, detects significant variation in the average level of Adh activity among the 25 haplotype classes ($Pr < 0.001$). It does not, however, give information about which haplotypes are most likely to carry the mutations that determine significant phenotypic effects. One might carry out multiple contrasts between pairs of haplotype classes to localize the effects to particular haplotypes (5). There are 300 possible pairwise contrasts but only 24 degrees of freedom (independent comparisons) available for comparing haplotypes. The difficulty in choosing an appropriate algorithm for performing the contrasts is discussed by Templeton et al.(10). These considerations provide the motivation for using a nested analysis defined by the structure of the cladogram.

The cladogram given in Figure 2 completely defines the nesting of haplotypes by degree of genetic relatedness. Haplotypes that share marker types are grouped together into clades (branches). Haplotypes that are most dissimilar for marker types are positioned farthest apart. Nesting of haplotypes into clades according to the similarity of their marker types gives a network that represents the evolutionary relationships among haplotypes. These relationships may be used to select comparisons among groups of haplotypes that will be most informative about the presence of a mutation with a phenotypic effect. A nested analysis of variance is a much more powerful alternative to the one-way analysis of variance because it takes into account this information about relatedness of haplotype classes. This analytical approach has the advantage of providing a means to

detect the existence of DNA sequence changes in haplotypes in different parts of the cladogram that must represent different mutational events because of the separation of the haplotypes in evolutionary time. The nesting of statistical comparisons according to the cladogram also allows full use of the available degrees of freedom in making all relevant contrasts of haplotypes. We now discuss the algorithm for grouping of the haplotypes into clades.

Details of the nesting algorithm are given in Templeton et al.(10). Each haplotype is a 0-step clade. Beginning at the terminal ends of the branches of the cladogram, 1-step clades are formed. The terminal haplotypes in Figure 2 are 4, 5, 7, 9, 10, 11, 12, 14, 16, 18, 19, 21, 23 and 25. The 1-step clades are defined by linking these haplotypes to haplotypes that have an allelic difference at one marker only. Haplotype 22 joins with 23 to form a 1-step clade. Other 1-step clades include 16 with 15, 25 with 24, 20 with 21 and 18 and 19 with 17. The 1-step clades are then combined to form 2-step clades. To do this each 1-step clade is considered to be a terminal clade and connected to other 1-step clades that differ for only one marker allele. Finally, 3-step clades are formed by joining 2-step clades that differ by one marker.

We next discuss the analysis of variability of Adh activity among branches of the cladogram to detect the presence of functional mutations. Table 1 presents the analysis of variance that is defined by the nesting of the 25 Adh haplotypes into 1-, 2- and 3-step clades. The first comparison is between the average level of Adh activity of the two 3-step clades (denoted X and Y in Figure 2). A highly significant difference suggests that the 3-step clade, X, that includes haplotypes 1-14 (average = 7.83 units of activity as nanomoles NAD reduced/min./fly) contains a functional mutation that is not shared by the 3-step clade, Y, that includes haplotypes 15-25 (average = 3.52 units). An asterisk is placed in Figure 2 between haplotypes 1 and 15 to denote that the two 3-step clades may differ for a functional mutation that influences Adh activity. It is not possible from this analysis to establish if all the haplotypes in clade X (or Y) are homogeneous or if reversions and/or other mutations with phenotypic effects have occurred. The nested analysis of variance within the X and Y clades, and the subsequent sequence analysis of haplotypes, can establish the extent of genetic heterogeneity within these 3-step clades.

The analysis of differences among 2-, 1- and 0-step clades presented in Table 1 focus on comparisons among haplotypes within the 3-step clades. There are three pairwise comparisons (3 df) of 2-step clades within the two 3-step clades. The nested analysis of variance suggests that none of these comparisons of average Adh activity is statistically significant.

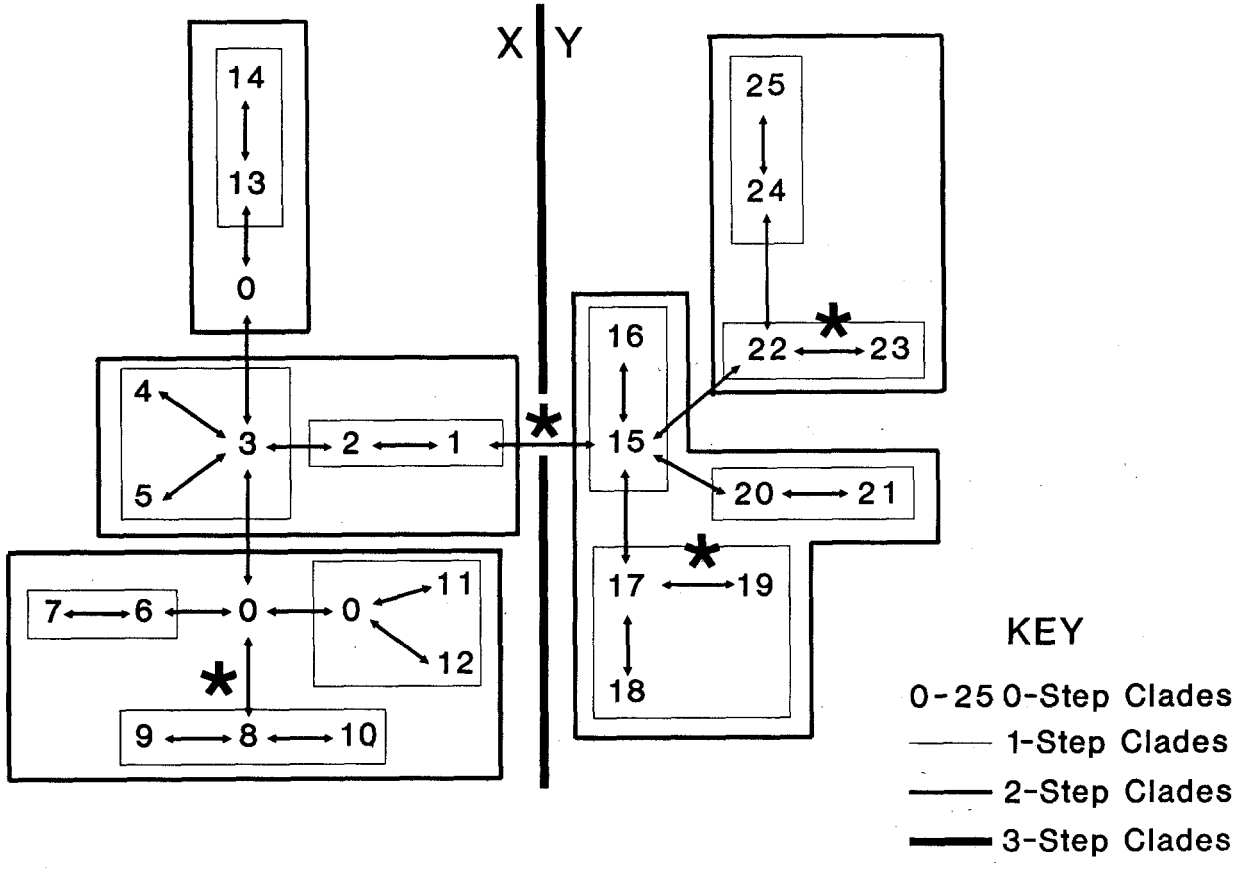

Figure 2. - The Cladogram of Adh Haplotypes.

TABLE 1. - Nested analysis of Adh activity variability.

Source	df	F-stat	P value
3-Step Clades	1	136.33	0.001
2-Step Clades in 3-Step Clades	3	0.78	NS
1-Step Clades in 2-Step Clades	6	5.71	0.01
0-Step Clades in 1-Step Clades	14	3.62	0.01
	24		

However, the nested analysis of differences among the 1- and 0-step clades, also presented in Table 1, detected significant effects of three additional mutations within the 3-step clades. The positions of these mutations are denoted in Figure 2 by asterisks. There are 6 pairwise comparisons of 1-step clades within the five 2-step clades. The significant result reported in Table 1 is due to the comparison of average Adh activity of individuals with haplotypes 8-10 with the average of those having the haplotypes 6, 7, 11 and12 included in the same 2-step clade. There are 14 pairwise comparisons of 0-step clades (haplotypes) within 1-step clades. Only two of these comparisons (22 vs. 23 and 17 vs. 19) are involved in determining statistically significant differences suggesting the presence of two additional functional mutations in the Adh gene region.

The results of the nested analysis of variance guide one in making decisions about which individuals should be sequenced to define the mutations that cause the phenotypic effects detected. For example, the marker information used to build the cladogram argues that individuals with haplotype 23 are more similar in their DNA sequence to individuals with haplotype 22 than they are to individuals with any other haplotype considered. Hence, the probability of finding the sequence responsible for the observed significant phenotypic difference in Adh activity between those individuals with haplotype 22 and those with haplotype 23 is expected to be maximized by comparing the DNA sequence of the Adh gene for individuals with haplotype 23 with the sequence obtained from individuals with haplotype 22. Likewise, individuals with haplotype 17 and those with haplotype 19 are candidates for DNA comparisons to detect a second Adh mutation that has a significant effect on the level of Adh activity.

We conclude from the nested analysis of variance, defined by the cladogram given in Figure 2 and presented in Table 1, that there are four mutations in the Adh gene region that have phenotypic effects on the level of Adh activity. Assuming that the evolutionary relationships among haplotypes defined by the marker haplotypes are true, we conclude that these four mutations in the Adh gene, arose at different times in the evolutionary history of the Adh region being marked. To characterize these mutations, the strategy would be to sequence the DNA of individuals that share the greatest number of marker types used to develop the cladogram but are included in different clades having significantly different average levels of Adh activity. That is, **rather than selecting individuals with high and low levels of Adh activity for DNA sequencing**, the cladistics strategy suggests that the sequences of individuals with haplotype 22 should be compared with the sequences of individuals with haplotype 23; those with haplotype 17 compared with those with haplotype 19; those in the 3-step clade including haplotype 1 compared to those in the 3-step clade including haplotype 15 and those with haplotype 8, 9 or 10 compared with those with haplotype 6, 7, 11 or 12. Comparisons of sequences will establish the DNA differences that are candidates for explaining the observed phenotypic effects. It will also be possible to establish whether a recurrent mutation is responsible for the observed phenotypic heterogeneity among haplotypes or each mutational event involves a unique DNA change.

Work in progress to facilitate an application of cladistics to human data

Obviously, it will be more difficult to apply this strategy to human observational data. However, the approach has its greatest potential in our search for the genetic basis of quantitative traits that have a complex multifactorial etiology. We are currently addressing a number of theoretical issues involved in the application of cladistics to the search for mutations responsible for quantitative phenotypic effects in humans. They include 1) modifications of the algorithm for estimating the number and frequency of haplotypes suggested by Templeton et al. (11), 2) evaluation of the limits of parsimony as a method for building cladograms and 3) developing a strategy for handling the situation where there is more than one parsimonious cladogram that is consistent with the observed haplotype data (12).

Acknowledgments

This work was supported in part by NIH grants HL24489, HL30248 and HL39107. We appreciate the interest in this

work and the constructive suggestions given by Steve Humphries, Anna Kessling, Pat Moll, Jim Neel, Tim Rebbeck, Sharon Reilly, Jack Schull and Ken Weiss during the preparation of this paper.

REFERENCES

1. *Aquadro C.F., Desse S.F., Bland M.M., Langley C.H.* and *Laurie-Ahlberg C.C.* (1986): Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster* - Genetics *114:* 1165-1190.

2. *Davignon J., Dufour R.* and *Cantin M.* (1983): Atherosclerosis and Hypertension - In: Hypertension, Second Edition, Genest J. et al. (eds). McGraw-Hill - Chapter 53, pp: 810-852.

3. *Kaprio J., Ferrell R.E., Kottke B.A., Turner S.T.* and *Sing C.F.* (1991): Effect of polymorphisms in apolipoproteins E, A-IV and H on quantitative traits related to risk for cardiovascular disease - Arteriosclerosis - *11:* 1330-1348.

4. *Kimura M.* (1983): The Neutral Theory of Molecular Evolution. Cambridge University Press, New York.

5. *Neter J., Wasserman W.* and *Kutner M.H.* (1985): Applied Linear Statistical Models, Edition 2 - Richard D. Irwin, Inc. Homewood, Ill.

6. *Roychoudhury A.K.* and *Nei M.* (1988): Human Polymorphic Genes - World Distribution - Oxford University Press, New York.

7. *Sing C.F., Boerwinkle E., Moll P.P.* and *Templeton A.R.* (1988): Characterization of genes affecting quantitative traits in humans. In: Proceedings of the 2nd International Conference on Quantitative Genetics, B.S. Weir, E.J. Eisen M.M. Goodman and G. Namkoong (eds.). Sinauer - Sunderland - MA. pp: 250-269.

8. *Sing C.F.* and *Moll P.P.* (1990): Strategies for unravelling the genetic basis of coronary artery disease. In: From Phenotype to Gene in Common Disorders, K. Berg, N. Retterstol, S. Refsum (eds.). Munksgaard A/S International Publishers, Copenhagen, Denmark.

9. *Sober E.* (1983): Parsimony in Systematics: Philosophical Issues. Ann. Rev. Ecol. Syst. *14:*335-357.

10. *Templeton A.R., Boerwinkle E.* and *Sing C.F.* (1987): A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila* - Genetics *117:*343-351.

11. *Templeton A.R., Sing C.F., Kessling A.* and *Humphries S.* (1988): A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. - Genetics *120:*1145-1154.

12. *Templeton A.R., Crandall K.A.* and *Sing C.F.* (1992): A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA Sequence data. III. Cladogram estimation. Genetics - In press.