Beltsville Symposia in Agricultural Research. 2. Biosystematics in Agriculture. J. A. Romberger, R. H. Foote, L. Knutson, and P. L. Lentz, eds. Allanheld, Osmun & Co., Montclair, N. J.; and John Wiley & Sons, New York, N.Y.

13] Molecular Phylogenetics

BY LYNN H. THROCKMORTON*

ABSTRACT

Different molecular approaches to phylogenetic analysis are illustrated and evaluated. The principle of genetic indeterminism is formulated, and the restrictions it places on phylogenetic methodology are made explicit. Molecular clocks, and the uses and misuses of genetic distance measures, are commented upon, as are the so-called maximum parsimony methods. Covariation analysis is emphasized as a sound analytical tool; it is defined, and examples are provided from protein studies of *Drosophila.*

INTRODUCTION

Some years ago, when the current bandwagon got rolling, one of the common arguments for molecular approaches to phylogenetic analysis was that they permit almost direct investigation of the genetic material. DNA hybridization can measure genetic divergence, the amino acid sequences of proteins are but one step away from the genetic code itself, and so on. The implication was, of course, that the closer we are to the genetic material the closer we are to pure phylogenetic information, and that if we could read the DNA we could read off phylogeny. That is not true, and I shall begin by noting some properties of the genetic system that severely limit phylogenetic analysis. This will establish what I shall call *genetic indeterminism.* Then I shall discuss some procedures of phylogenetic method and how these apply to molecular data. This will define and illustrate *covariation analysis,* which is the core of standard phylogenetic method and widely accepted as such, although it is not generally called by that name. Finally I shall consider some of the current fads: DNA hybridization, immunological and genetic distance, common ancestor and "maximum parsimony" methods, and so on. These can be compared with standard methods, and

*Department of Biology, University of Chicago, Chicago, Illinois 60637.

with the realities of scientific method, and their advantages and disadvantages for systematics can be appraised.

My overall strategy will be first to emphasize and illustrate theoretical considerations, and second to evaluate current phylogenetic fashions against these principles. Most of my illustrations will be drawn from the virilis group of *Drosophila* (*27, 28,* and unpublished data). This is a holarctic species group with about half a dozen nearctic and half a dozen palearctic species. It probably originated in middle or late Miocene times, perhaps about fifteen million years ago, and its genealogy is quite well established (*25, 29*). The phylogeny is based on information from salivary gland chromosomes, proteins and reproductive behavior, but I shall not try to recount that evidence at this time. There are uncertainties regarding some of the most recently described species, though these do not affect the examples or the conclusions illustrated by them.

GENETIC INDETERMINISM

We have long been aware of genetic variation within natural populations, but it was only during the last few decades that this became directly measurable. Protein electrophoresis made that possible and early studies produced remarkably high estimates of the mean proportion of heterozygous loci within species gene pools. For *Drosophila* this was about 53%, and values for other insects were similar. It was somewhat higher in marine invertebrates (59%) and rather lower in vertebrates (down to 15% in birds). The average heterozygosity per individual ranged from 17% in plants through 15% in *Drosophila* and other insects to about 4% in large mammals (*21*). These estimates are only for structural gene loci. Control loci and elements of other systems may behave differently, but this is still an extraordinarily large amount of genetic variation.

Large as it is, however, recent discoveries find still more (*1, 22, 30*). The first investigations depended on electromorphs, on protein molecules that showed mobility differences during electrophoresis, and it was thought that more than two-thirds of the total genetic variation could be measured in this way. When that was actually checked within populations by using two criteria, mobility plus heat sensitivity, the average electromorph was found to be a collection of proteins with about three different heat stabilities. At the XDH locus in the virilis group, for example, 11 electromorphs became 32 thermoelectromorphs. And when three or more criteria were used to assess differences, the number of morphs increased four or five fold (*4, 5, 23*). This total now approaches theoretical expectations for the number of variants that might be hidden within a single electromorph, so it may be nearly correct. In the virilis group the number of electromorphs per locus per species ranges from 1 to 15, with a mean near 8. If these were increased four to five fold the mean number of alleles per locus per species could be as high as 30 or 40.

The point of all this is that at the level of the gene pool the system is encountered at one of its least predictable levels. Variability is high and natural selection can winnow it in many ways. And since heterozygous gene pools persist through time and many different species bud off from them (*26*), individual genic variants may be apportioned to descendent species almost haphazardly relative to the genealogy of the group. On the average, three electromorphs per locus (12 to 15 alleles?) have been segregating in the gene pools of the virilis group since its origin. Some of the consequences of this are illustrated in Fig. 13.1. First there is a two-allele polymorphism that has persisted in all but one line for about 15 million years (HP-2, lower right); then a three-allele pattern shows the origin of alleles within phylads and their subsequent haphazard distribution from then on (LAP, lower left). The upper examples (MDH, HP-1) show combinations of these patterns. If some single locus were being followed and distribution of its alleles used as evidence for phylogeny, no one of these would give the correct result.

This segregation is quite real, but operationally its effects are confounded with another aspect of genetic indeterminancy. Mutation occurs at predictable rates and individual genic variants may recur quite regularly so long as appropriate parent variants persist in the gene pool. At the upper right in Fig. 13.1 (HP-1) electromorph 6 is found almost throughout one phylad, but it is also found once in the other phylad. Is this convergence or parallelism? We cannot sharply separate the two. This instance might be mutational. The distribution of alleles 3 and 4 of HP-2 (lower right) almost certainly is not. Insofar as phylogenetic analysis is concerned it makes no difference. Whichever it is, it is noise, and any competent phylogenetic analysis must cope with it.

Granting that single alleles have great limitations, one might suggest that perhaps analyses combining information from many loci will overcome the drawback. That can be true or not, depending on how the results are analyzed. If a similarity index is calculated (phenetic distance, genetic distance, etc.) it is not true. High levels of similarity, large numbers of alleles in common, can also arise in parallel. Perhaps the most obvious and most common example of this is evolutionary conservatism, where two or more species on different evolutionary pathways remain very similar through retaining a large proportion of ancestral alleles in their genotypes. Alternately, from a common heterozygous heritage they may independently generate highly similar genotypes through responding in similar ways to similar environmental challenges. Again these last two alternatives cannot be distinguished operationally, but that does not change the conclusion. So long as mechanisms exist by which high levels of similarity arise independently, degree of similarity, a blind number, cannot reliably stand as evidence of propinquity of descent.

Table 13.1 illustrates the above. The matrix to the left shows the distribution of similarities attributable to shared primitive states. In this instance

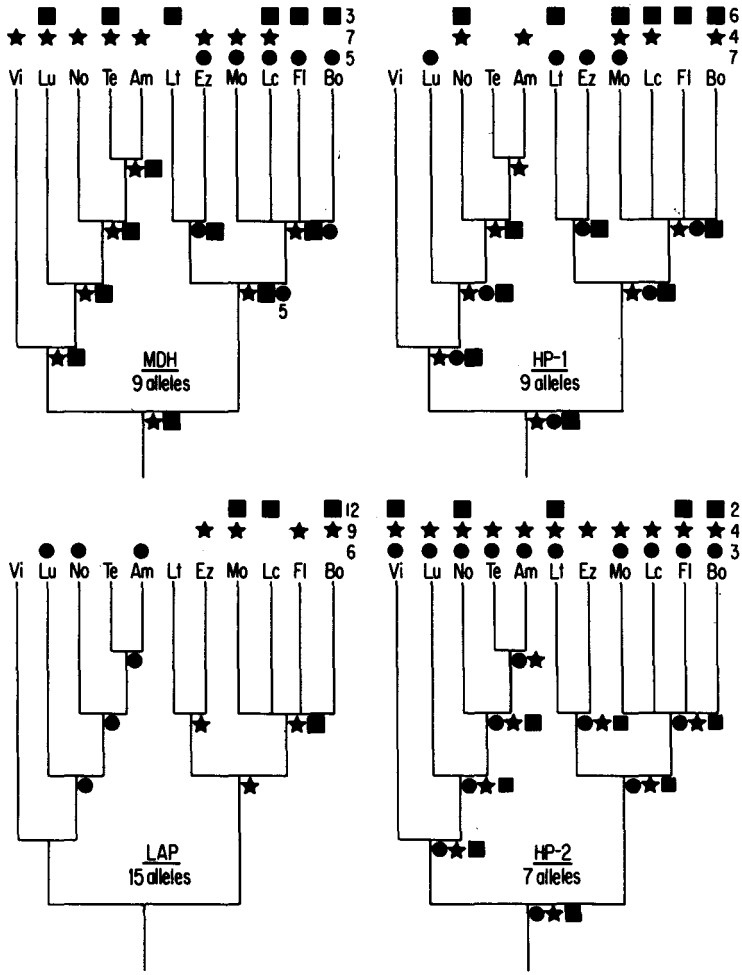

Figure 13.1. Genetic indeterminism. The segregation of different electromorphs is illustrated by symbols on the genealogy of the virilis group of *Drosophila* species. The electromorphs are numbered to the upper right. The locus and the total number of electromorphs at each locus are indicated between the phylads. MDH, malic dehydrogenase; LAP, leucine amino peptidase; HP-1, hemolymph protein 1; HP-2, hemolymph protein 2; Vi, *D. virilis*; Lu, *D. lummei*; No, *D. novamexicana*; Te, *D. a. texana*; Am, *D. a. americana*; Lt, *D. littoralis*; Ez, *D. ezoana*; Mo, *D. montana*; Lc, D. lacicola; Fl, D. flavomontana; Bo, *D. borealis*.

the primitives were detected operationally as those electromorphs that were present in at least one species of each of the two cytological phylads. Derivatives were recognized as those electromorphs restricted to the species of only one phylad. On the average, primitive states account for about 78% of the genetic similarity between species in the virilis phylad (upper left corner of the matrix) and 59% of the similarity between species in the

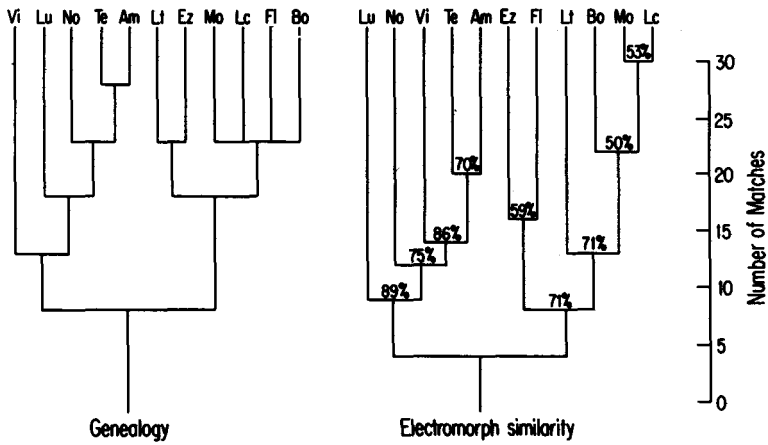

Figure 13.2. Genetic indeterminism. A contrast is shown between the phylogeny of the virilis group of *Drosophila* species (left) and the dendrogram of genetic similarity derived from protein data (right). The number of matches is shown on the right side. Numbers on crossbars are the percent contributed to that similarity by primitive electromorphs. Abbreviations are as in Figure 13.1.

montana phylad (to the lower right in the matrix). Some consequences of this for phylogeny are illustrated in Fig. 13.2. On the left I show the genealogy of the virilis group again. On the right I show a dendrogram of relationship derived by clustering species on the basis of the number of matches between them for electromorphs at 10 loci. The numbers on the crossbars indicate the percent of matches that in each case result from sharing of ancestral genes. This ranges from 50% to nearly 90%, and the overriding effect it can have on similarity, and on phylogenies blindly based on similarity indexes, is obvious.

Heterozygosity, its persistence, and its predictably unpredictable consequences over long periods of evolutionary time, cannot be avoided. Genetic indeterminism is a fact of life, and the next question is whether, or how, it can be dealt with. To my knowledge there is only one method that minimizes its effects, and that is covariation analysis.

COVARIATION ANALYSIS

Covariation and correlation are not the same thing. Attributes that covary are correlated, but things that are correlated do not necessarily covary. *Character states can be said to covary when any one of the states is not observed unless another state of the set, from at least one other independent character of the set, is present in the same unit also.* Table 13.2 shows the covariation of sets of inversions, one set for *Drosophila lummei* through *D. a. americana,* the other for *D. ezoana* through *D. borealis.* Note that *D. virilis* is not included in either set. Within the first set there is a sequential

Table 13.1 Genetic indeterminism. The distribution of similarities generated through the retention of ancestral electromorphs (left) and those resulting from segregation of derivative electromorphs (right). Abbreviations for species names are as in Fig. 13.1.

	Shared primitive states										Shared derivative states									
	Vi	Lu	No	Te	Am	Ez	Lt	Fl	Mo	Lc	Bo	Vi	Lu	No	Te	Am	Ez	Lt	Fl	Mo	Lc	Bo
Vi		9	11	12	14	7	9	10	11	8	14		2	1	2	5	0	0	0	0	0	0
Lu			8	9	9	7	9	7	12	9	8			1	1	3	0	0	0	0	0	0
No				9	12	4	8	8	11	9	10				4	5	0	0	0	0	0	0
Te					14	8	8	9	13	10	11					6	0	0	0	0	0	0
Am						7	7	9	13	10	13						0	0	0	0	0	0
Ez							6	10	11	8	8							2	7	10	6	9
Lt								9	11	10	11								2	4	4	4
Fl									13	10	12									10	6	8
Mo										16	14										14	16
Lc											12											11

Table 13.2. Covariation analysis of gene sequences. Species are indicated to the left and abbreviations are as in Fig. 13.1. The gene sequences of salivary gland chromosomes are across the top. Individual sequences are not named since only the pattern of covariation is emphasized here. Unique sequences, those found only in one species, are indicated in the column to the right.

	Character states—chromosome sequences																																	
Species																																		U
Vi																																		1
Lu	+	+	+	+																														?
No	+	+	+		+	+	+	+	+	+																								0
Te	+	+	+	+	+	+					+																							0
Am	+	+	+	+	+	+	+	+	+	+	+																							3
Ez												+	+	+	+	+	+	+	+	+														4
Lt												+	+	+	+	+					+	+												6
Fl												+	+	+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	+	4
Mo												+	+	+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	+	33
Lc												+	+	+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	+	13
Bo												+	+	+	+	+	+	+					+	+	+	+	+	+	+	+	+	+	+	7

Table 13.3. Covariation analysis of electromorphs. Species are indicated to the left and abbreviations are as in Fig. 13.1. Different electromorphs are shown horizontally. The upper section shows the ancestral electromorphs and uniques. The lower section shows derivative electromorphs and should be compared with the pattern seen in Table 13.2.

Character states—ancestral electromorphs

Species																																						U
Vi	+	+	+	+	+	+	+				+	+			+	+	+	+	+								+	+	+	+									3
Lu		+	+	+		+		+	+		+	+			+												+		+	+	+								0
No	+	+	+	+							+	+			+		+					+	+	+	+		+	+	+	+									2
Te	+	+		+		+			+		+	+	+	+	+		+	+	+	+	+								+	+	+								0
Am	+	+	+	+		+				+	+	+	+	+	+	+	+	+	+	+			+	+					+	+	+								4
Ez					+	+		+			+	+	+		+				+		+	+	+				+		+	+	+								1
Lt	+	+		+	+	+		+	+						+											+	+	+	+	+	+	+							4
Fl				+	+	+	+	+		+		+			+				+	+	+	+				+		+	+	+	+								5
Mo		+	+	+	+	+	+	+	+	+	+	+	+	+	+					+	+		+	+	+	+	+		+	+	+								6
Lc		+		+	+	+		+	+	+	+			+	+						+		+	+	+			+	+	+								3
Bo	+	+	+		+	+	+	+	+	+				+	+	+	+	+	+			+		+		+		+	+	+	+							5

Character states—derivative electromorphs

Species																																						
Vi	+	+	+	+	+																																		
Lu	+	+				+																																
No			+			+	+	+	+		+																											
Te	+			+			+	+	+	+	+																											
Am	+	+	+	+	+	+	+	+	+	+																												
Ez											+	+	+	+	+	+	+	+	+			+	+															
Lt											+	+									+	+																
Fl											+	+	+	+			+	+					+	+	+	+												
Mo											+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Lc											+	+	+	+	+	+			+	+			+	+			+									+	+	+
Bo											+	+	+	+	+	+	+	+	+	+			+	+			+	+	+									

subdivision, with three taxa, *D. novamexicana, D. a. americana,* and *D. a. texana,* sharing inversions not yet seen in *D. lummei.* The second set is subdivided dichotomously, with *D. ezoana* and *D. littoralis* partitioned off from the other four species. This happens to be the cytological evidence for the phylogeny of the virilis group, and it should be pointed out that inversion phylogenies are nondirectional. In themselves they contain no information as to the position of the "root" of the tree. Non-directionality is a peculiar property of inversion data and not a general property of

covariation. In this instance information from chromosome fusions, and also other evidence, indicates that the tree is "rooted" in such a way that *D. virilis* is included phylogenetically with *D. lummei, D. novamexicana,* and *D. americana.*

A more typical set of data is given in Table 13.3. This shows some of the proteins from the virilis group. Species are again listed to the left and alleles are shown horizontally. In the top section ancestral alleles and the uniques are given. At the bottom the derivative sets are shown. The pattern seen for the ancestral alleles gives some indication of the "noise" in the system and of the haphazardness of distributions generated through evolutionary time as a consequence of heterozygosity. The sets of derivative and primitive electromorphs were, of course, intermingled in the original data set. Ninety-five electromorphs are present; 30 (31%) are primitive, 32 (34%) are derivative, and 33 (35%) are unique. Since species boundaries are known, the uniques are easy to detect. Of the remainder about half are "information" and half are "noise," and identifying which is which can be difficult. It follows that I am emphasizing covariation analysis as an effective phylogenetic method, not as an easy one.

In a sense it is misleading to refer to the ancestral electromorphs as noise. This is a matter of perspective. Insofar as intragroup relations are concerned, they are noise. But within the genus these are the information that sets the group apart from all other groups. In an early study of sibling species from all parts of the genus, it was found that less than six percent of electromorphs are

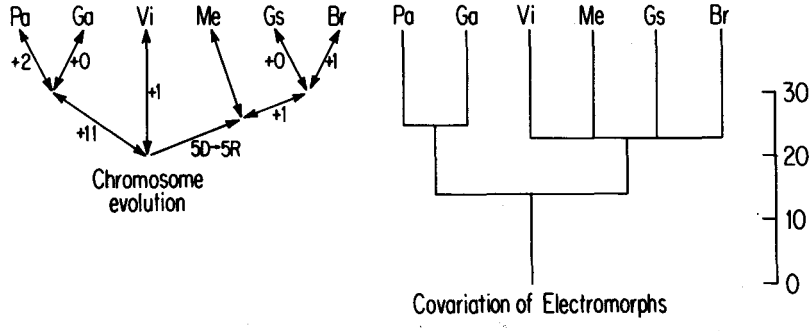

Figure 13.3. Covariation analysis of the mesophragmatica group of *Drosophila* species. To the left is given the diagram of chromosome relationships. Most arrows are bidirectional. In one case, 5D to 5R, the dot chromosome was increased by heterochromatin to become a small rod. This may have been a directional change since the dot is primitive for the genus. Numbers along the right side indicate the size of the covariation set that establishes each group. Chromosome data are recalculated from Brncic *et al.* (*3*) and protein data are recalculated from Nair *et al.* (*18*). Pa, *D. pavani*; Ga, *D. gaucha*; Vi, *D. viracochi*; Me, *D. mesophragmatica*; Gs, *D. gasici*; Br, *D. brncici*; U, uniques.

Table 13.4. Covariation analysis of the mesophragmatica group. Species are indicated to the left and character states across the top. Chromosome data are recalculated from reference *13* and protein data are recalculated from reference *14*.

Chromosomes																Ancestral electromorphs														
Species															U															U
Pa	+	+	+	+	+	+	+	+	+	+	+	+			10	+		+	+	+	+	+	+	+	+	+		+	+	5
Ga	+	+	+	+	+	+	+	+	+	+	+	+			0	+	+		+		+	+	+		+	+				2
Vi	+														1					+	+		+	+	+	+				7
Me													+		4			+	+		+	+			+		+	+	+	3
Gs													+	+	6	+	+				+	+					+		+	2
Br													+	+	0	+		+			+	+	+	+	+	+	+			0

Derivative electromorphs																				
Species																				
Pa	+	+	+	+	+	+	+	+	+	+	+									
Ga	+	+	+	+	+	+	+	+	+	+	+									
Vi												+	+	+			+			+
Me												+		+		+		+	+	+
Gs													+		+	+	+	+	+	+
Br													+		+	+	+			+

shared between species of different species groups (*14*). If we use 94% (100%-6%) as a guide, approximately 28 of the 30 ancestral electromorphs are distinctive relative to those of other species groups in the genus; which gives some idea of the power of this approach for partitioning taxa.

Over the years I have gotten the impression that some systematists envy *Drosophila* their salivary gland chromosomes. Or at any rate they envy the systematists who make use of them. Protein electrophoresis can, to some degree, give all taxonomists a tool as powerful as salivary gland chromosomes. Each has its advantages and disadvantages, as is true for almost all methods. The covariation patterns of the chromosomes (Table 13.2) and of the derivative states of the electromorphs (Table 13.3, bottom) can be contrasted. That of the chromosomes is a much tighter pattern, with fewer gaps in it, indicating that these chromosomes were less sensitive, on the average, to the indeterminism of heterozygosity than were the electromorphs. Thus these inversions are somewhat better evidence of genealogy than are the proteins. On the other hand, covariation patterns of proteins are directional while those of the inversions are not, so what is gained in one way can be lost in another.

Not all chromosomes give such definitive patterns as those of the virilis

group. In Table 13.4 and Fig. 13.3 an example is given from the mesophragmatica group (*3, 18*). These six species are split into two clusters, a sibling species pair, *D. gaucha* and *D. pavani,* and four other species separated from it by at least 11 inversions. In one case (5D to 5R, Fig. 13.3) the dot chromosome has been converted to a small rod by the addition of heterochromatin. Since the dot-like chromosome 5 is primitive for the genus, this is probably a directional change, but it and the few remaining inversions fixed between species cannot help much to delineate further the genealogy of these species. Their numbers are so low that they could easily have segregated in this pattern from heterozygous ancestral gene pools with no relationship to phylogeny.

The protein data from the mesophragmatica group parallels the chromosome data and resolves some of the phylogenetic problems (Fig. 13.3, right). The "root" of the tree appears to be between the sibling species pair and the other four species. *D. viracochi* belongs to the phylad on the right and the origin of the rod-like fifth chromosome probably occurred during the early evolution of that phylad. We still cannot tell how to apportion the eleven paracentric inversions that separate the two clusters, and we do not know the details of the relationships of the four species in the viracochi phylad. This is the parsimonious answer that can be derived from these data, and, while we might like to have more, we certainly cannot complain of what we do have.

One last example of covariation can be noted (Table 13.5). This does not come from *Drosophila,* but it is of a kind familiar to almost everyone. I picked, almost at random, a 30-residue segment from amino acid sequence data published by Dayhoff (*6*). It happens to be from some vertebrate polypeptide hormones. The different covarying sets are indicated by underlining and spacing. The point of the example is to show the second major way covariation patterns are seen in molecular data. Processing is the same as for electromorph data. Here the amino acid residue at a particular site in a protein is the character state. There the particular electromorph at a specific locus is the character state. In both cases direction of evolution is resolved by reference to outside groups, with states encountered outside the group regarded as operational primitives of no value for determining within-group relationships. Since, on the average, fewer than one in ten of the electromorphs within a group are found outside a group, I have not emphasized that aspect of the analysis for electromorphs. For other types of data, anatomical as well as molecular, this is a very important consideration.

Before continuing to the fads and fashions that we encounter today in molecular phylogenetics, it is necessary to comment on the scientific, or epistemological, justification for covariation analysis and for regarding the product of such an analysis as a nonarbitrary and parsimonious indication of the true genealogy of the species involved.

The argument is brief, but all the more potent for being concise. First

Table 13.5. Covariation of amino acid residues. A 30-residue section of some vertebrate hormones is shown. Horizontal lines separate the different covarying sets. 1, cow prolactin; 2, sheep prolactin; 3, human lactogenic hormone; 4, human growth hormone; 5, cow growth hormone; 6, sheep growth hormone. Data are from Dayhoff *(6)*.

Species	Amino acid residues
1	L L F A H L E F R M S N S E M D R V I D H V S V F D Y K S N
2	L L F A H L E F R M S N S E M D R V I D H V S V F D Y K S N
3	L L F A H L E F S L A Q A D T E H Q A D H V P M Y R R E I Q
4	L L F A H L E F S L A Q A D T E N R L D H I P M Y R R E F E
5	L L F A H L E F S L A Q A D T E N R L A Q M S V F G H R A K
6	L L F A H L E F S L A Q A D T E N R L A Q M S V F G H R A K

recall the definition of covarying character states. *Character states are said to covary when any one of the states is not observed unless another state of the set, from at least one other independent character of the set, is present in the same unit also.* An analysis based on covariation is phylogenetic. A highly improbable distribution (covariation) of character states is observed. Such an association of character states cannot be excused on the grounds of chance and some explanation is required for it. Only a limited number of alternatives is available. Covariation can be a consequence of pleiotropy, or it can be due to independent adaptation to common environmental factors (convergence), or it can be due to common descent uncomplicated by other elements. If no positive evidence for pleiotropy or convergence can be adduced, common descent is accepted as the parsimonious inference. Alternatives involve *additional assumptions* (and so are less parsimonious). Special creation is one example of an additional assumption, if you do not mind calling it that, and independent origin through mutation is another. But whatever the case, so long as there is no *concrete* evidence for other factors, the inference of common descent is justified and so is the conclusion of phylogenetic accuracy that derives from it.

It would be wrong to decide that this implies a method of absolute accuracy. Those seem to be rare in this universe and covariation analysis is no paragon. All attributes that are accessible to covariation analysis, whether anatomical or molecular, are influenced by genetic indeterminancy. The only way the consequences of this can be minimized is through the use of large sets of characters. Choice between conflicting sets of covarying states, and these do appear frequently, is based on the nature and relative sizes of the sets concerned. Naturally, the larger and "tighter" sets are the weightier sets. If alternative sets are nearly the same size, no choice is possible and the details of relationship must remain unresolved.

In brief, then, covariation analysis is not necessarily easy, but it can be effective in producing a nonarbitrary and parsimonious phylogeny. As we all know, parsimony does not guarantee truth. What it does guarantee is the best answer from the existing data. It is neither right nor wrong. It is just very correct!

FADS AND FASHIONS

In criticizing methods it is necessary to recognize at the outset that almost every method is good for something, although it may not have the usefulness its proponents claim for it. Some methods produce useful preliminary results and are fine for exploration. Others produce definitive results, but they may be too cumbersome or expensive for general use. There are four chief sources of data for molecular phylogenetics. I can describe none of them in detail, but can only direct attention to each method, indicating generally its approach, and occasionally noting technical or practical difficulties. Then I consider the manner in which these data have been, or can be, processed and the ways in which this restricts the systematic uses to which they can be put. Here the emphasis must be not so much on the steps in the procedure as on the premises that underlie it. It is these premises that validate an analysis, that give it whatever scientific merit it can have, hence they are of the utmost importance.

Obtaining the data. Electrophoresis, amino acid sequence analysis, immunology and DNA hybridization have been most frequently exploited to provide molecular data for phylogenetic analysis, and each is optimistically reviewed elsewhere (*6, 11, 19, 24, 31*). Electrophoresis makes use of samples of body fluids, homogenized tissues, homogenized whole organisms, and so on. For my own work a single fly is homogenized per sample, which gives an indication of the amount of material required. Some of the results I gave earlier were taken from a study of the virilis group involving more than 60,000 samples (flies) run in acrylamide gel over a period of about five years. Once in the gel the samples are subjected to a charge differential for varying lengths of time, then they are removed to a suitable reaction mixture to visualize the proteins. The gels can then be photographed to record the data. This method assays only products of structural gene loci, and of these it deals only with those that produce soluble products, products that are detectable by available assay systems, and so on. It is generally thought that they represent a random sample of structural gene loci. At least there is no evidence that they comprise a sample biased in some evolutionarily significant way, and they are evaluated accordingly. Without special additional procedures this method detects electromorphs, not necessarily alleles, so similarities calculated from electrophoretic data are as much phenetic distances as they are genetic distances. I have already given examples of results from electrophoresis, and since one dendrogram is

much like another, I shall not give additional specific examples for the other methods.

The second important method is that of amino acid sequence analysis (*6, 11*). Probably the best known examples of this are the cytochromes *c* and the hemoglobins (*9*). This method requires a purified protein as its starting point, and there must be suitable amounts of protein from each of the organisms to be investigated—sometimes not an easy requirement to meet. The cytochrome *c* of a gastrotrich will probably be quite hard to acquire, for example. Much of the procedure for sequencing proteins has been automated, but it is still not simple and relatively few laboratories are equipped to do it. Once the sequence has been determined, it is possible to infer the DNA sequence that produced it, to compare DNA sequences, calculate mutation distance, etc. An example of amino acid sequence data is given in Table 13.5, and it can be regarded as true genetic data, not just approximately genetic data, as was the case for the electromorphs.

Immunology is the third important method, the most common procedure being that of microcomplement fixation (*20*). This method also requires samples of purified protein, although usually not in the amounts needed for sequence analysis. An antiserum is produced against this protein and reacted against serum from the same (homologous) and different (heterologous) species. The factor by which the serum concentration from heterologous samples must be raised to give a reaction equal to that of homologous serum provides an index of dissimilarity. The higher the required concentration of heterologous serum the greater the dissimilarity. The log of the index of dissimilarity is approximately linear with respect to the proportion of different amino acids between the sequences of the proteins used in the reaction if the proteins are not too different, at least for a few proteins that have been tested so far. To obtain approximate linearity, however, it is necessary to make full reciprocal comparisons between species. This is a warning that the procedure is less trustworthy than might be wished. From a study of bird transferrins (*12*), for example, I calculated the mean percent difference between reciprocal comparisons, and for more than half (52.5%) this difference was 20 percent or greater. Thus, microcomplement fixation measures an approximate genetic distance. The distance is roughly comparable to that obtained for single proteins by amino acid sequence analysis. It is not an accurate measurement, however.

The fourth method is DNA hybridization (*2, 10, 13, 15, 16, 17, 19*). For this the DNAs are hybridized and the thermal stability of the hybrid DNA is measured. At the temperature at which one-half the hybrid DNA dissociates, a 1.5% mismatch between strands lowers the thermal stability by 1°C. This is, accordingly, a convenient measure of differences between DNAs, the lower the thermal stability of the hybrid DNA the greater is the distance between the parent DNAs (*16*). Several years ago DNA phylogenies, mostly based on a less accurate technique, were much in evidence. Then serious complications became apparent (*19*). It is now known that the DNA of an

organism is complex; exactly how complex is not yet certain. There are at least two fractions of DNA, one unique and presumably coding for structural gene loci, the other repetitive and acting as spacer DNA, etc. Different species may have different relative quantities of these types of DNA, and the different ways these different kinds of DNA might evolve are scarcely understood at all (*10*). In consequence, DNA comparisons between species do not necessarily involve homologous DNAs, or involve homologous DNAs only in very uncertain proportions, even when care is exercised to use just what are hoped to be the unique sequences. If it were *known* that only unique sequences were involved, the genetic distance calculated from the DNA might be roughly comparable to the genetic distances calculated from amino acid sequence data. It would estimate genetic distance between total genomes, however, rather than over just one structural gene as in the case of genetic distances between molecules of, say, hemoglobin. Since the DNAs involved in such comparisons cannot be sharply defined yet, the significance of the difference values is uncertain.

Analyzing the data. Phylogeny means different things to different people. I use phylogeny in the sense of genealogy, to refer to the sequence by which different groups arise. A construction that cannot accurately represent genealogy cannot be regarded as a phylogeny, and a method that does not accurately reproduce genealogy is not a phylogenetic method, as I use the term here. Of course, a phylogeny does not need to be completely detailed to be accurate. Accuracy is required only for the details that are shown. There are three fundamentally different ways of approaching phylogenetic analysis. One of these is covariation analysis, which I have already considered briefly. A second is the "clock" approach, and the third is one based on some concept or another of "minimal" evolution.

The concept of a molecular clock has been with us now for almost a generation (*6, 8, 19*). It grew out of the observation that the degree of amino acid sequence difference between certain proteins is nearly linearly related to the times of divergence between them. Different proteins have evolved at different rates, and hence it should be possible to use slowly evolving proteins to clock ancient events and rapidly evolving ones to clock recent ones. Unfortunately, there are serious reservations regarding just how such a clock can be used.

First, it is now realized that constancy of rate is a consequence of the long periods of time involved. A protein can change at quite different rates over short periods of time and its average rate can still be quite constant. Evidence for inconstant rates comes from a variety of sources, a convenient one for me being the data on protein evolution in the virilis group. The two phylads of this group have been evolving for the same length of time, but one phylad added 20 new electromorphs while the other added 48. Of the 10 loci investigated, three evolved faster in the virilis phylad, six evolved faster in the montana phylad, and one did not change in either phylad. The net

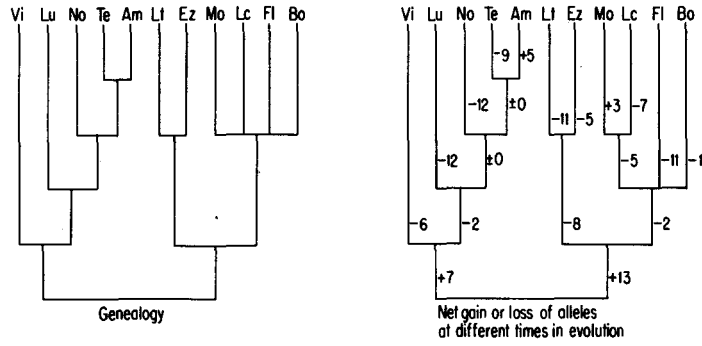

Figure 13.4. The irregular molecular clock. The genealogy of the virilis group of *Drosophila* species is given to the left. On the right a slightly modified version is shown with the net gain or loss of alleles during each interval shown beside the internodes. In total the virilis phylad gained 20 electromorphs while the montana phylad gained 48.

result is that the montana phylad evolved 2.4 times faster than the virilis phylad. The general pattern for gain and loss of alleles is shown in Fig. 13.4. The net gain or loss of electromorphs is given for each internode of the dendrogram and ranges from +13 to -16, with two internodes showing no net gain or loss. The rate of molecular evolution is not constant here. The ecologically conservative phylad changed slowly; the one encountering evolutionary opportunity in new environments changed rapidly. Hence, there is no molecular clock that can be used for timing short intervals like the branches of most phylogenetic trees. It is, in fact, a very poor clock since, as we all know, an inaccurate clock can be worse than none at all. It may be correct sometimes, but one never knows when those times are.

The uses of this kind of clock are very limited, since they must be compatible with uses of an *average*. The restriction is very simple. One average can be used to estimate another, with more or less success depending on circumstances. But one cannot take an average and from it discover the particular individual observations that contributed to it. One cannot, for example, use the height of the average adult U.S. male and assert that *that* will be the height of the next man one meets. But that is effectively what is done when genetic distance is used to specify the individual internodes on a dendrogram. For the length of an internode (= time) is established in proportion to genetic distance. Genetic distance and time are presumed to be equivalent and the rate of evolution is treated as constant over the entire dendrogram. The data from the virilis group illustrate an instance where that is not the case; which thus introduces a constraint that eliminates all genetic distance measures as effective means of constructing phylogenies. Genetic distances can be used in phylogenetic analysis only if molecular change occurs with clocklike regularity, and present evidence shows that that is not true.

One might suppose that there is a situation where genetic distance can be used, as an average, to time divergence between groups. If one had, say, two groups of several species each, one could make pairwise measurements of genetic distance between them, and from these observations calculate a pretty fair average genetic distance between them. Then, if one could somehow calibrate this genetic distance and determine how much time, on the average, was related to each distance unit, one might have an estimate of the time his groups diverged. One still would not know how much evolution occurred in the line leading to one group, and how much occurred in the other, but that might not be important. But there is another problem, and that is calibration. This genetic distance must be measured between groups for which times of divergence are known from fossil, or other evidence. And those are hard to come by. Calibration must usually be carried out on one group of organisms and the results applied to others. This creates an epistemological problem. What guarantees that the rate observed in one group actually applies in others? After all, if I want to find out what is in my pocket I do not look in yours, and if I measure the rate of evolution in frogs or bacteria I cannot have too much assurance that it applies for *Drosophila.* In short, if the method cannot be properly calibrated it cannot be used, and if it can be calibrated, it is not needed, since the required times of divergence are already available from other more reliable sources.

It follows from this that phylogenetic methods depending on molecular clocks, as do all varieties of genetic distance, whether they be calculated from DNA, electrophoretic data, amino acid sequences, or immunological data of one kind or another, are inadequate. At least they cannot produce non-arbitrary and parsimonious phylogenies. They can, however, be very useful for exploratory work, and they should not be discounted for systematic research of that kind.

The minimal distance methods (*7, 9*), including "common ancestor" methods (*6*), present a somewhat different scientific problem. Those who have worked much with arranging dendrograms know that many different trees may often be created by slight changes in clustering methods, sequence of analyzing the data, and similar trivial variations. How does one choose among the trees that can be produced? In fact, there is no nonarbitrary choice, and that in itself is sufficient to eliminate these methods as definitive phylogenetic methods. If the choice of dendrogram is arbitrary, there can be no necessary relation between it and the genealogy of the species involved. And there is an additional problem. A favorite method is to try, by any one of many ingenious means, to discover the "shortest" possible tree that is compatible with the data. The argument for this is that it reflects "minimal evolution" of some kind and is therefore the "best" tree. The difficulty with this approach should be obvious to everyone. Formal evolution theory does not assert, and never has, that evolution proceeds from one particular species to another by the shortest possible route. What it does say

Table 13.6. Relationship between kind of data and method of analysis. In the body of the table a "yes" indicates that data from a given source is suitable for analysis by a given method. Only two kinds of data are suitable for covariation analysis, and covariation analysis is the only epistemologically sound method of analysis by which genealogies can be inferred.

	Method of analysis		
Source of data	Genetic distance	Minimum distance[a]	Covariation
Electrophoresis	Yes	Yes	Yes
Amino Acid Sequence	Yes	Yes	Yes
Immunology	Yes	Yes	No
DNA Hybridization	Yes	Yes	No

[a]Minimum mutational distance, maximum parsimony, etc.

is that change from one generation to another will be small (but not necessarily the smallest possible change), and that one character state will evolve into another gradually. That is a far cry from saying, for some haphazard assortment of extant (or fossil) species, that the lot of them must have evolved from each other, or from annectant ancestors, by the shortest possible route. To regard such a procedure as parsimonious completely misconsceives the intent and use of parsimony in science.

Parsimony is a means of choosing between answers on the basis of the inferences employed in arriving at those answers from the available data. One asks which of the analyses involves the fewest unnecessary inferences, the fewest presumptions not directly dictated by the data themselves. It is most definitely not a method that involves looking at two "answers" and deciding which is the least "extreme." Minimum distance methods force data to conform to a specific preset bias. Data are not permitted to "disclose the structure of the universe," as is the intent of scientific method. Instead they must produce an answer as nearly as possible in conformity with a predetermined model. Far from being parsimonious, minimum distance methods are thoroughly procrustean, and they succeed only in giving inherently arbitrary choices quite undeserved dignity and standing.

CONCLUSIONS

The gene pool carries a great deal of variation, and while its distribution from generation to generation can be predicted with some facility, its course through longer periods of evolutionary time is quite unpredictable. This indeterminancy, reflecting persistent heterozygosity, haphazard retention of ancestral alleles among descendent species, convergent mutation, and parallel generation of similar genotypes, places severe restrictions on meth-

ods of phylogenetic analysis. And the closer the material is to the gene itself, the more sensitive it is to these perturbing influences. Thus, the distribution of alleles at one locus must be quite uncertain, while the joint distribution of many alleles at many loci (covariation) can, if properly analyzed, accurately evidence genealogy.

Genetic distance methods depend on the reliability of unreliable molecular clocks, and so are not adequate for phylogenetic analysis. The minimum-distance methods force data to conform to preset bias, hence are procrustean, and so are wholly nonparsimonious and also unsuited to produce phylogeny. Only covariation analysis (Table 13.6), which searches for covarying character sets and hence can in principle discover nonarbitrary partitions within the data, can produce sound evidence from which parsimonious phylogenetic inferences can be drawn. Present immunological and genetic distance methods cannot produce data suitable for covariation analysis. Only electrophoretic data and amino acid sequence data are suitable, and definitive molecular phylogenies must come from them. At best, other so-called phylogenetic methods are suitable only for exploration, to provide tentative evidence of probable lines of descent. Such preliminary work can precede more discriminating study by some form of covariation analysis, but it cannot replace it.

ACKNOWLEDGMENTS

This research was supported in part by grants GM 11216 and GM 23007 from the National Institutes of Health.

LITERATURE CITED

1. Bernstein, S. C., L. H. Throckmorton, and J. L. Hubby. 1973. *Still more genetic variability in natural populations.* U.S. Natl. Acad. Sci., Proc. 70: 3928-3931.

2. Britten, R. J., and D. E. Kohne. 1968. *Repeated sequences in DNA.* Science 161: 529-540.

3. Brncic, D., P. S. Nair, and M. R. Wheeler. 1971. *Cytotaxonomic relationships within the mesophragmatica species group of Drosophila.* Univ. of Texas Publ. 7103: 1-16.

4. Cochrane, B. J. 1976. *Heat stability variants of esterase-6 in Drosophila melanogaster.* Nature 263: 131-132.

5. Coyne, J. A. 1976. *Lack of genic similarity between sibling species of Drosophila as revealed by varied techniques.* Genetics 84: 593-607.

6. Dayhoff, M. O. 1972. *Atlas of protein sequence and structure.* Natl. Biomed. Res. Found., Washington, D.C.

7. Farris, J. S. 1972. *Estimating phylogenetic trees from distance matrices.* Amer. Nat. 106: 645-668.

8. Fitch, W. M. 1976. *Molecular evolutionary clocks.* Pages 160-178 *in* F. J. Ayala, ed. *Molecular Evolution.* Sinauer, Sunderland, Mass.

9. Fitch, W. M., and E. Margoliash. 1967. *Construction of phylogenetic trees.* Science 155: 279-284.

10. Galau, G. A., M. E. Chamberlin, B. R. Hough, R. J. Britten, and E. H. Davidson. 1976. *Evolution of repetitive and nonrepetitive DNA.* Pages 200-224 *in* F. J. Ayala, ed. *Molecular Evolution.* Sinauer, Sunderland, Mass.

11. Goodman, M. 1976. *Protein sequences in phylogeny.* Pages 141-159 *in* F. J. Ayala, ed. *Molecular Evolution.* Sinauer, Sunderland, Mass.

12. Ho, D. Y.-K., E. M. Prager, A. C. Wilson, D. T. Osuga, and R. E. Feeney. 1976. *Penguin evolution: protein comparisons demonstrate phylogenetic relationships to flying aquatic birds.* J. Mol. Evol. 8: 271-282.

13. Hoyer, B. H., B. J. McCarthy, and E. T. Bolton. 1964. *A molecular approach to the systematics of higher organisms.* Science 144: 959-967.

14. Hubby, J. L., and L. H. Throckmorton. 1968. *Protein differences in Drosophila. IV. A study of sibling species.* Amer. Nat. 102: 193-205.

15. Kohne, D. E. 1970. *Evolution of higher-organism DNA.* Quart. Rev. Biophys. 3: 327-375.

16. Kohne, D. E., J. A. Chiscon, and B. H. Hoyer. 1972. *Evolution of primate DNA Sequences.* J. Hum. Evol. 1: 627-644.

17. Laird, C. D., and B. J. McCarthy. 1968. *Magnitude of interspecific nucleotide sequence variability in Drosophila.* Genetics 60: 303-322.

18. Nair, P. S., D. Brncic and K. Kojima. 1971. *Isozyme variations and evolutionary relationships in the mesophragmatica species group of Drosophila.* Univ. of Texas Publ. 7103: 17-28.

19. Nei, M. 1975. *Molecular population genetics and evolution.* North-Holland, Amsterdam, and Elsevier, New York. 288 pp.

20. Sarich, V. M., and A. C. Wilson. 1966. *Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation.* Science 154: 1563-1566.

21. Selander, R. K. 1976. *Genic variation in natural populations.* Pages 21-45 *in* F. J. Ayala, ed. *Molecular Evolution.* Sinauer, Sunderland, Mass.

22. Singh, R. S., J. L. Hubby, and L. H. Throckmorton. 1975. *The study of genic variation by electrophoretic and heat denaturation techniques at the octanol dehydrogenase locus in members of the Drosophila virilis group.* Genetics 80: 637-650.

23. Singh, R. S., R. C. Lewontin, and A. A. Felton. 1976. *Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in Drosophila pseudoobscura.* Genetics 84: 609-629.

24. Soule, M. 1976. *Allozyme variation: its determinants in space and time.* Pages 60-77 *in* F. J. Ayala, ed. *Molecular Evolution.* Sinauer, Sunderland, Mass.

25. Stone, W. S., W. C. Guest, and F. D. Wilson. 1960. *The evolutionary implications of the cytological polymorphism and phylogeny of the virilis group of Drosophila.* U. S. Natl. Acad. Sci., Proc. 46: 350-361.

26. Throckmorton, L. H. 1962. *The problem of phylogeny in the genus Drosophila.* Univ. of Texas Publ. 6205: 207-343.

27. Throckmorton, L. H. 1972. *Genetic change and speciation in the virilis group of Drosophila.* Presented at 14th Internatl. Congr. of Entomol., Canberra.

28. Throckmorton, L. H. 1972. *Abstracts,* 14th Internatl. Congr. of Entomol. (Canberra) Page 16 (Abstr.)

29. Throckmorton, L. H. 1975. *The phylogeny, ecology, and geography of Drosophila.* Pages 421-469 *in* R. C. King, ed. *Handbook of Genetics 3.* Plenum, New York.

30. Throckmorton, L. H. 1977. *Drosophila systematics and biochemical evolution.* Ann. Rev. Ecol. Syst. 8: 235-254.

31. Wilson, A. C., S. S. Carlson, and T. J. White. 1977. *Biochemical evolution.* Ann. Rev. Biochem. 46: 573-639.