A Quantitative Test of the Neutral Theory Using Pooled Allozyme Data

D. O. F. Skibinski,* M. Woodwark* and R. D. Ward†

**Molecular Biology Research Group, School of Biological Sciences, University College of Swansea, Singleton Park, Swansea SA2 8PP, Wales, and †CSIRO Division of Fisheries, Hobart, Tasmania 7001, Australia*

Manuscript received June 5, 1992
Accepted for publication May 22, 1993

ABSTRACT

Neutral theory predicts a positive correlation between the amount of polymorphism within species and evolutionary rate. Previous tests of this prediction using both allozyme and DNA data have led to conflicting conclusions about the influence of selection and mutation drift. It is argued here that quantitative conclusions about the adequacy of neutral theory can be obtained by analyzing genetic data pooled from many sources. Using this approach, a large database containing information on allozyme variation in over 1500 species is used to examine the relationship between heterozygosity and genetic distance. The results provide support for the hypothesis that a major percentage of protein variation can be explained by variation in neutral mutation rate, and a minor percentage by strong selection.

IT has been suggested that allozyme analysis, which is restricted in the number of genetic loci and type of mutational change that can be investigated, lacks the power to discriminate between selection and neutrality (LEWONTIN 1991). By contrast DNA sequence analysis, which can provide almost unlimited genetic information, has led to renewed hope of resolving the "selectionist-neutralist" controversy. The power of allozyme analysis can, however, be increased by pooling data from many species. This approach has been used successfully in a variety of tests of neutral theory (for example, FUERST, CHAKRABORTY and NEI 1977; CHAKRABORTY, FUERST and NEI 1978, 1980; NEI and GRAUR 1984).

An analytical method that has been applied to both DNA sequence and pooled allozyme data is the analysis of the relationship between intraspecific variation and evolutionary rate. Neutral theory makes two predictions, first that there should be a high positive correlation between these two variables, and second that regions of the genome having identical levels of intraspecific variation should evolve at the same rate. The allozyme tests provide support for the first prediction (SKIBINSKI and WARD 1982; WARD and SKIBINSKI 1985; CHAKRABORTY 1984). The DNA tests have, by refuting the second prediction, provided evidence for selection. Unexpectedly high levels of DNA variation in regions flanking allozyme loci have been used to implicate balancing selection (HUDSON, KREITMAN and AGUADE 1987; BEGUN and AQUADRO 1991; KREITMAN and HUDSON 1991), and unexpectedly high divergence in regions low in polymorphism have been used to implicate positive directional selection (BEGUN and AQUADRO 1991; BERRY, AJIOKA and KREITMAN 1991; MARTINCAMPOS *et al.* 1992; STEPHAN and MITCHELL 1992). Evidence of a higher ratio of fixed differences to polymorphisms for replacement than synonymous mutations has also been attributed to positive selection (MCDONALD and KREITMAN 1991). However, some of the many comparisons of variation between different regions of the genome have failed to refute the null hypothesis of strict neutrality (BRADY, RICHMOND and OAKESHOTT 1990; LANGE, LANGLEY and STEPHAN 1990; BEGUN and AQUADRO 1991). In general, the results of these DNA tests suggest that variation due to amino acid replacements might be influenced by selection but that silent variation is largely neutral.

This paper describes a study of the relationship between intraspecific variation and evolutionary rate in a large body of allozyme data pooled from a wide variety of vertebrate and invertebrate sources. In tests of the first prediction, correlations between protein heterozygosity and genetic distance are found to be generally higher than in earlier studies (SKIBINSKI and WARD 1982; WARD and SKIBINSKI 1985). However, in tests of the second prediction, some proteins with similar heterozygosity are found to differ significantly in genetic distance. The results provide support for the hypothesis that a major percentage of the variation in protein heterozygosity and genetic distance can be explained by variation in neutral mutation rate but that a minor percentage of this variation is influenced by strong selection.

MATERIALS AND METHODS

The analyses were carried out using a large database of allozyme studies. Most of the data are from published sources but some unpublished data are included. The sources of most of the data are given by WARD, SKIBINSKI

and WOODWARK (1992). A full bibliography can be obtained on request from the authors. The database has 800 studies of allozyme variation in animals. Each of these studies is a comparison of two or more populations or species at a sample of allozyme loci. Analyses were carried out separately for five groups of studies: (I) the whole database (comprising groups II to V below) containing 800 studies with a total of 3728 populations or species, (II) vertebrate intraspecies containing 324 studies with a total of 1788 populations, (III) vertebrate interspecies containing 174 studies with a total of 599 species, (IV) invertebrate intraspecies containing 188 studies with a total of 910 populations, and (V) invertebrate interspecies containing 114 studies with a total of 431 species. The intraspecies studies are comparisons of two or more populations within a species, the interspecies studies are comparisons of two or more related species. (To avoid repetition, the word taxon is used below to stand for both species or populations in those circumstances where both words are equally applicable.) The data are distributed over many phyla and classes. For example, among the vertebrate studies there are 282 species of fish, 94 of reptiles, 137 of amphibians, 86 of birds and 202 of mammals. Among the invertebrate studies there are 253 species of insects, 120 species of mollusks and 111 species of crustacea as well as representatives from other less widely studied groups. Taxa included in the database have been scored for at least 15 allozyme loci in 750 of the studies and between 10 and 14 loci in the remaining 50 studies, and for at least 15 individuals per locus per population. In the analyses that follow, attention is focused on those proteins scored in 50 or more studies. Information on these proteins is given in Table 1.

PREDICTIONS OF NEUTRAL AND SELECTION THEORY

In the infinite allele model, expected heterozygosity (H) at a neutral locus in a population is given by $H = 1 - 1/(1 + 4N_e u)$ (KIMURA and CROW 1964), where u is neutral mutation rate and N_e is effective population size. The expected genetic distance (D) between this population and a sister population is given by $2ut$ (NEI 1972), where t is the time since divergence of the two populations from a common ancestor. Combining the two equations gives the relationship $D = (t/2N_e)H/(1 - H)$. Given that genetic drift dominates mutation, a test of neutral theory based on this relationship is best carried out by averaging H and D over many independent replicate loci. With DNA data, this can be done by using many estimates of site heterozygosity and distance from each of several regions of the genome differing in underlying neutral mutation rate. With allozyme data, this situation can be simulated by pooling information from many taxa such that every taxon contributes data for every one of a set of proteins. In this balanced dataset, every protein will have the same distribution of t and N_e, as in the DNA approach using two descendent taxa. Thus, the ratio t/N_e will be identical for every protein, and the form of the relationship between H and D will be a curve of positive slope. The position of a protein on this curve is a function of its neutral mutation rate.

The form of the relationship between D and H is very robust to variations in the neutral model. For example, the stepwise mutation model (OHTA and KIMURA 1973) gives rise to a curve of positive slope, which becomes increasingly asymptotic at high levels of divergence (LI 1976; CHAKRABORTY and NEI 1977; MUKHERJEE, SKIBINSKI and WARD 1987). Simulation studies (MUKHERJEE, SKIBINSKI and WARD 1987) show that the form of this curve is not affected by heterogeneity of mutation rates between replicate loci within proteins, nor by heterogeneity of divergence times between the pairs of taxa contributing to the genetic distance estimates, nor by breeding system. Nor is it affected by migration high enough to dominate drift; thus, a test can be applied at the level of population as well as species divergence. The robust behavior of the relationship between D and H is not surprising, for in a balanced dataset proteins differ in only one parameter, neutral mutation rate. In this circumstance, even if the true underlying model is unknown, a reasonable conjecture is that the relationship will be a line of positive slope lacking inflections.

Population bottlenecks cause the rapid conversion of heterozygosity into genetic distance (CHAKRABORTY and HEDRICK 1983; CHAKRABORTY and NEI 1977). This can explain observed negative correlations between D and H when the plotted points are for different populations (LIVSHITS and NEI 1990). However, it has been argued (SKIBINSKI and WARD 1983) and shown by computer simulation (MUKHERJEE, SKIBINSKI and WARD 1987) that the relationship between D and H for a set of proteins in a balanced dataset is unaffected by population expansions or contractions. Thus, a test of neutral theory is not dependent on the assumption of neutral equilibrium.

In the versions of neutral theory incorporating selection against deleterious mutations (OHTA 1976, 1977; KIMURA 1979), some scatter about the curve could only occur if proteins differ in the values of the additional parameters reflecting the distribution and intensity of selective effects. In a model with normally distributed selection coefficients, the relationship between substitution rate and heterozygosity follows the neutral expectation closely with moderate selection (selection coefficients of the order of $1/N_e$) or with weak selection (TACHIDA 1991). Thus, large differences in the strength of selection between proteins seem to be needed to cause appreciable scatter in these models.

From a selectionist viewpoint, few restrictions are placed on the relationship between H and D. Strong directional selection can cause rapid divergence of populations at low or high heterozygosity; balancing selection can cause high heterozygosity in the presence of rapid divergence or evolutionary stasis. Therefore, although a positive correlation between D and H is perfectly consistent with selection, there is, in contrast

TABLE 1

Proteins used in the analysis with numbers of studies scored for groups II (vertebrate intraspecies), III (vertebrate interspecies), IV (invertebrate intraspecies) and V (invertebrate interspecies)

Protein	Code	EC numbers	Number of studies in group[a]			
			II	III	IV	V
Alcohol dehydrogenase	1	1.1.1.1	126	77	31	21
Malate dehydrogenase	2	1.1.1.37	316	172	169	105
α-glycerophosphate dehydrogenase	3	1.1.1.8	222	132	97	55
Isocitrate dehydrogenase	4	1.1.1.42	293	149	131	86
Sorbitol dehydrogenase	5	1.1.1.14	127	76	45	23
6-phosphogluconate dehydrogenase	6	1.1.1.44	249	138	122	66
Lactate dehydrogenase	7	1.1.1.27	319	171	55	36
Malic enzyme	8	1.1.1.40	173	88	127	77
Glucose-6-phosphate dehydrogenase	9	1.1.1.49	85	45	51	30
Superoxide dismutase	10	1.15.1.1	273	141	108	70
Aspartate aminotransferase	11	2.6.1.1	277	154	137	91
Phosphoglucomutase	12	5.4.2.2	299	162	156	95
Esterase (nonspecific)	13	*.*.*.*	221	119	137	85
Phosphoglucose isomerase	14	5.3.1.9	288	143	165	102
Xanthine dehydrogenase	15	1.2.1.37	70	40	56	38
Glutamate dehydrogenase	16	1.4.1.2/3	94	51	27	11
Peptidases	17	3.4.11.*	188	95	72	47
Fumarase	18	4.2.1.2	96	51	44	32
Leucine aminopeptidase	19	3.4.11.1/2	75	60	95	62
Transferrin	20	*.*.*.*	58	32		
General protein	21	*.*.*.*	175	101	49	41
Haemoglobin	22	*.*.*.*	78	45	2	1
Albumin	23	*.*.*.*	83	51		
Adenylate kinase	25	2.7.4.3	110	37	60	40
Creatine kinase	26	2.7.3.2	134	59	7	5
Adenosine deaminase	27	3.5.4.4	89	40	16	8
Mannose phosphate isomerase	28	5.3.1.8	154	70	86	56
Acid phosphatase	29	3.1.3.2	93	41	72	50
Triose phosphate isomerase	30	5.3.1.1	18	6	20	18
Diaphorase	32	1.6.2.2	25	8	19	10
Amylase	33	3.2.1.1	13	6	18	17
Alkaline phosphatase	35	3.1.3.1	20	7	45	31
Octanol dehydrogenase	37	1.1.1.*	17	11	27	24
Glutamate pyruvate transaminase	40	2.6.1.2	24	12	18	11
Esterase-D	41	3.1.1.1	29	8	9	5
Aldolase	43	4.1.2.13	38	19	44	33
Glyceraldehyde-3-phosphate dehydrogenase	45	1.2.1.12	83	31	38	35
Aconitase	46	4.2.1.3	60	28	26	10
Hexokinase	47	2.7.1.1	26	9	106	59
Aldehyde oxidase	48	1.2.3.1	5		50	35
Nucleoside phosphorylase	50	2.4.2.1	56	21	8	6
Pyruvate kinase	69	2.7.1.40	18	6	28	11

[a] The number for group I is the total for groups II to V.

to neutral theory, no strong prediction of one pattern of variation and exclusion of others. Such models thus provide no test; and selection theory, unlike neutral theory, can receive no support.

In the SAS-CFF model (see GILLESPIE 1991) short-term environmental fluctuations maintain variation and cause substitutions through an allelic exchange process. Periodic long-term environmental fluctuations also cause episodes of substitutions. The model has five parameters and thus can be fit to almost any data (GILLESPIE 1989). The correlation between heterozygosity and evolutionary rate is negative in the exchange process but positive in the episodic process. Thus, the observation of a positive correlation would discriminate in favor of the episodic process, as noted by GILLESPIE (1991).

HYPOTHESES TESTED

The first prediction of neutral theory, that heterozygosity and evolutionary rate should be highly correlated, is tested by carrying out regression of protein genetic distance on protein heterozygosity. Regression analysis is used because residual deviations are

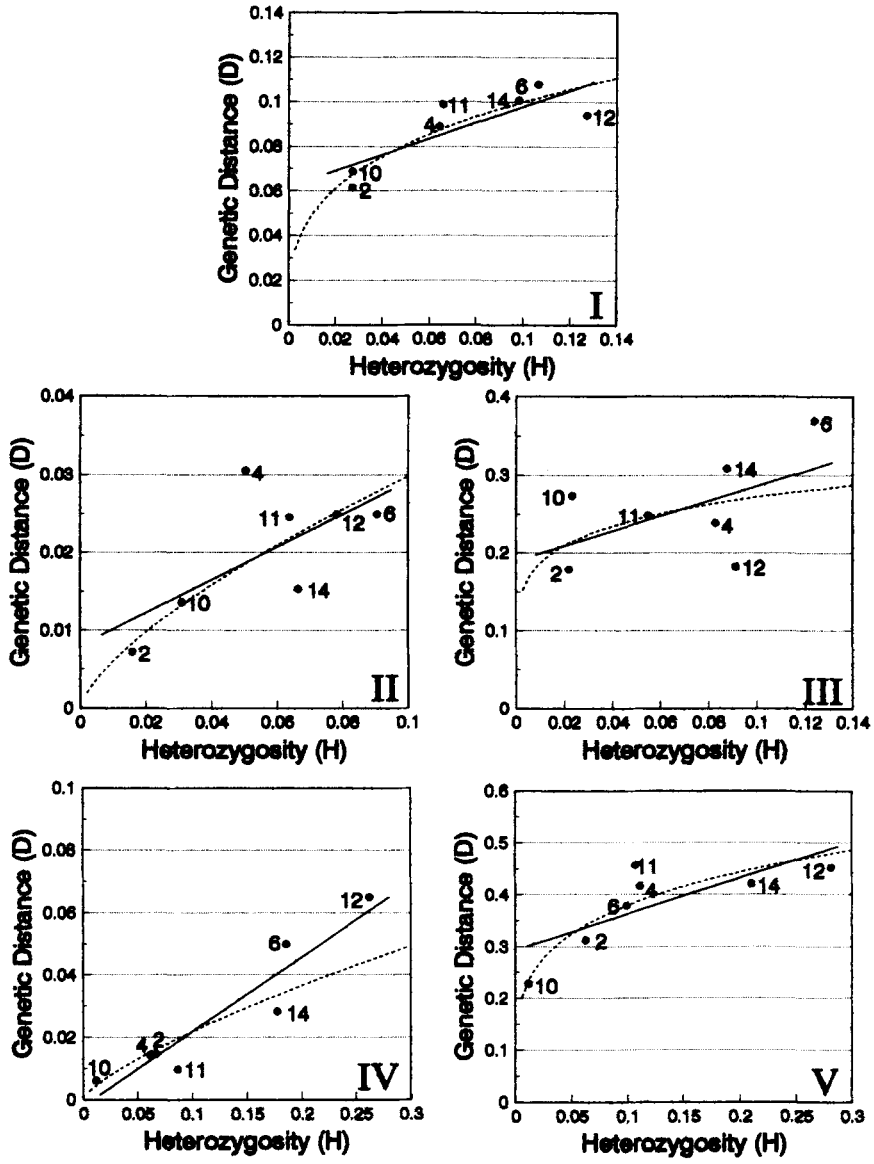

FIGURE 1.—Mean distance (D) plotted against mean heterozygosity (H) for seven proteins for the first method of analysis with random selection of loci for multilocus proteins. 2: malate dehydrogenase; 4: isocitrate dehydrogenase; 6: 6-phosphogluconate dehydrogenase; 10: superoxide dismutase; 11: aspartate aminotransferase; 12: phosphoglucomutase; and 14: phosphoglucose isomerase. (I) Whole database; (II) vertebrate intraspecies; (III) vertebrate interspecies; (IV) invertebrate intraspecies; (V) invertebrate interspecies. Dotted line = fitted curvilinear function; straight line = fitted linear regression.

more easily interpreted than deviations from a principal axis. The estimated correlation can be explained quantitatively by neutral theory (KIMURA 1989, 1991a) and provides a measure of the adequacy of neutral theory. The coefficient of determination quantifies the total variation in protein heterozygosity and distance that can be accounted for by variation in neutral mutation rate between proteins. Approaches to testing that conclude with the rejection, or failure to reject, a specific null hypothesis derived from neutral theory might be less informative than this quantitative approach. In the first and second methods of analysis (see below) a balanced dataset is used with seven different proteins. The third method of analysis uses a larger sample of proteins with a split half design to simulate a balanced dataset.

The second prediction of neutral theory is approached by testing the null hypothesis that two proteins that have closely similar heterozygosity have identical genetic distance, using a balanced dataset that contains only these two proteins. Eight pairs of proteins are considered separately (the fourth method of analysis). Another approach, performed as part of the first and second methods of analysis, is to test for significant residual deviations from the fitted regression lines.

Both linear and nonlinear methods of curve fitting are used. The former is conservative from the neutralist viewpoint because the relationship between distance and heterozygosity is unlikely to be a straight line, except at low levels of divergence. The latter is conservative from the selectionist viewpoint because the best possible curve is being fit to the data. Regression through the origin is used as well as conventional regression about the mean. The former provides an estimate of the percentage of the sum of squared deviations of distance from the abscissa that can be explained by regression. It is appropriate in circumstances in which the fitted regression line follows the theoretical expectation, and it is only the scatter of

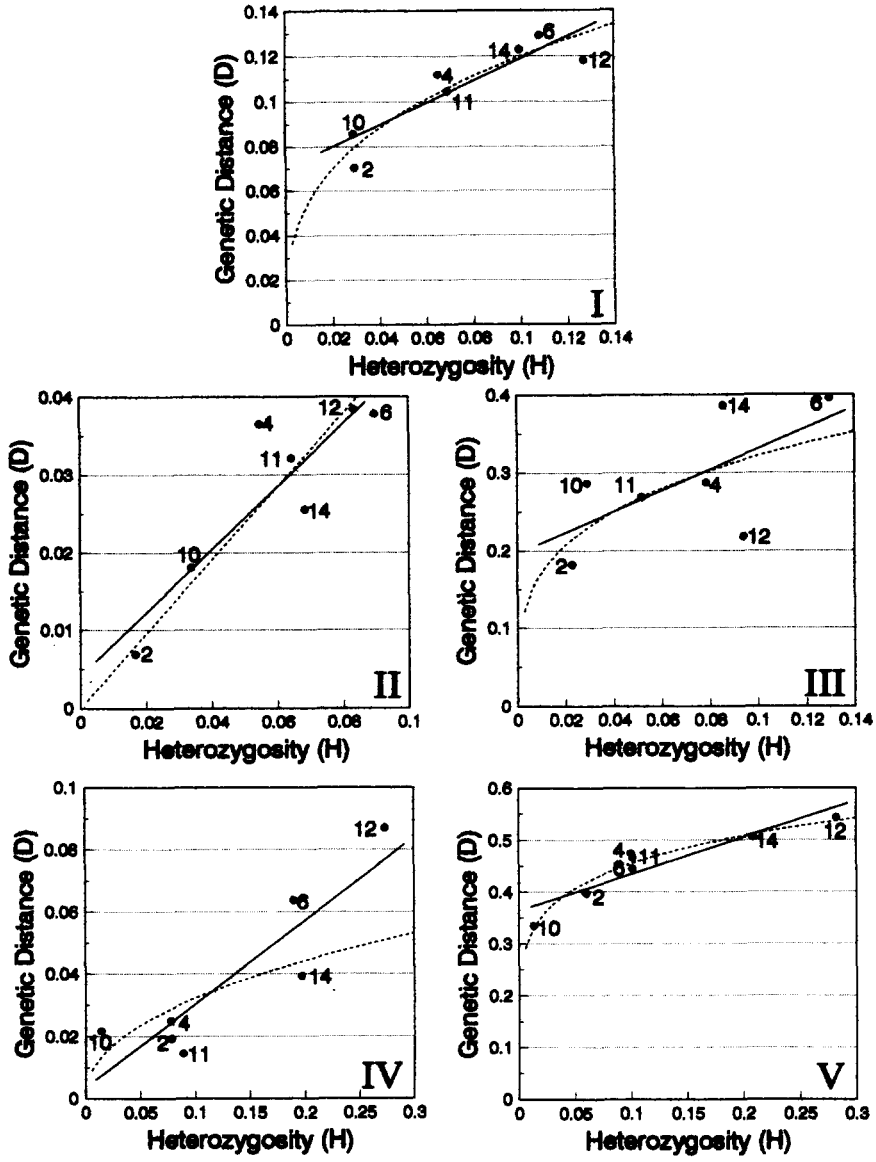

FIGURE 2.—Mean distance (D) plotted against mean heterozygosity (H) for seven proteins for the second method of analysis using all loci for multilocus proteins. Key as in Figure 1.

points about the fitted curve that is seen as inconsistent with the theory being tested.

ANALYSIS AND RESULTS

The balanced dataset (the first and second methods of analysis): The amount of data in a balanced dataset is at a maximum when the product (number of proteins) × (number of studies) is at a maximum. This occurs with the seven proteins malate dehydrogenase (code 2), isocitrate dehydrogenase (4), 6-phosphogluconate dehydrogenase (6), superoxide dismutase (10), aspartate aminotransferase (11), phosphoglucomutase (12) and phosphoglucose isomerase (14). These have been scored in 291 of the 800 studies, of which 144 are in group II, 70 in group III, 42 in group IV and 35 in group V.

As the number of taxa in a study increases, the number of comparisons between taxa increases exponentially. The number of comparisons was weighted by study size by comparing taxon one with taxon two, taxon two with taxon three, and so on within a study so that the number of taxon pairs (m) is one less than the number of taxa. Calculated in this way, the total number of taxon pairs for the 291 studies is 1086, of which 705 are in group II, 145 in group III, 127 in group IV and 109 in group V.

Heterozygosity according to Hardy-Weinberg expectation was used as the measure of intraspecific variation, and genetic distance (NEI 1972) was used as the measure of evolutionary rate. Four basic statistics were calculated. These are the genetic distance between two taxa at an allozyme locus (d), the mean heterozygosity of the two taxa at this locus (h), and the mean of d and of h for a protein across pairs of taxa and studies (D and H, respectively). The precise method of computation of these statistics for proteins in the database is described in the Appendix.

In the first method of analysis, in situations where

TABLE 2

Results of regression analysis of genetic distance on heterozygosity for seven proteins

Group	R_{lin}(H and d)	R^2_{adj}(%)		R^2_{adj} ratio(%)	
		Proteins	Heterozygosity(h)	H_{lin}/proteins	H_{poly}/proteins
First method of analysis					
Whole dataset (I)	0.070	0.63	2.51	75.0	100
	(0.041–0.101)	(0.29–1.47)	(1.71–3.80)	(30.1–95.3)	(55.1–100)
	[0.047–0.092]				
Vertebrate (II) intraspecies	0.054	0.52	4.39	53.3	65.9
	(0.017–0.091)	(0.26–1.81)	(3.28–6.51)	(1.8–97.7)	(13.7–100)
	[0.026–0.082]				
Vertebrate (III) interspecies	0.095	2.56	5.02	31.3	47.7
	(0.006–0.196)	(1.34–6.04)	(3.11–9.52)	(0–81.6)	(0–97.6)
	[0.033–0.155]				
Invertebrate (IV) intraspecies	0.189	3.37	8.53	100	100
	(0.082–0.274)	(0.97–8.41)	(5.48–17.39)	(38.7–100)	(40.2–100)
	[0.125–0.252]				
Invertebrate (V) interspecies	0.167	4.14	4.35	64.2	95.0
	(0.068–0.244)	(1.58–8.65)	(1.91–8.92)	(9.6–94.0)	(38.6–100)
	[0.097–0.235]				
Second method of analysis					
Whole dataset (I)	0.133	2.15	7.30	81.8	93.7
	(0.094–0.178)	(1.36–3.94)	(5.10–10.80)	(45.5–93.3)	(59.5–99.2)
	[0.111–0.155]				
Vertebrate (II) intraspecies	0.147	2.58	13.92	82.6	87.6
	(0.084–0.212)	(1.85–6.07)	(10.83–20.29)	(22.1–95.7)	(36.5–98.6)
	[0.119–0.174]				
Vertebrate (III) interspecies	0.182	7.45	11.91	43.2	42.5
	(0.060–0.301)	(4.38–15.02)	(7.61–20.86)	(3.2–79.4)	(9.4–83.8)
	[0.122–0.240]				
Invertebrate (IV) intraspecies	0.247	7.04	18.76	85.5	94.2
	(0.071–0.427)	(1.75–20.96)	(7.61–38.83)	(14.3–100)	(39.7–100)
	[0.186–0.307]				
Invertebrate (V) interspecies	0.283	8.79	9.12	89.5	100
	(0.128–0.418)	(3.48–21.24)	(3.93–22.54)	(29.2–98.9)	(46.1–100)
	[0.216–0.347]				

95% confidence limits based on z statistics are given in square brackets, those based on bootstrapping are given in round brackets. The first method uses random locus selection, the second method uses all loci.

two or more loci are scored for a protein within a study, just one of the loci was selected at random as the representative of that protein in the study. In the second method of analysis, all loci were used for analysis. The first method avoids bias that might arise because of differences between proteins, within taxa, in the mean number of loci scored. The second method has the advantage of using all available loci within the balanced dataset.

Analysis was carried out separately for the whole dataset (group I) and for each of groups II to V. Multiple linear regression analysis was used, with d the dependent variable. The independent variables were added to the regression equation in a stepwise manner in the order H first, the protein codes second, and h third. The nominal protein codes were used to assign dummy variables for each protein following the method described by NIE *et al.* (1975) and were entered into the equation as a group. The analysis was repeated with polynomial regression by adding H^2 to the regression equation after adding H. The R^2 (coefficient of multiple determination) values obtained on entering variables in the regression equation were converted to adjusted R^2 (R^2_{adj}) using the equation given by SOKAL and ROHLF (1981).

Confidence intervals for the multiple regression statistics were obtained by bootstrapping across studies. For example, for the whole dataset (group I), 291 studies were taken at random, one at a time with replacement, from the original dataset of 291 studies to generate the first bootstrap sample. The multiple regression analysis was then carried out on this bootstrap sample. The entire procedure was repeated 1000 times and the 95% confidence intervals determined for regression statistics over the 1000 separate bootstrap analyses. Confidence intervals were obtained by bootstrapping in the manner described, for both the first and second methods of analysis. For comparison, confidence intervals were also obtained from the analysis of z statistics (SOKAL and ROHLF 1981) based on the total number of taxon pairs. Bootstrapping over studies rather than taxon pairs avoids problems

TABLE 3

Analysis of deviations from regression for seven proteins

Group	Linear regression			Polynomial regression		
	$100\text{-}R^2_{adj}$ ratio	F	Deviant protein codes	$100\text{-}R^2_{adj}$ ratio	F	Deviant protein codes
First method of analysis						
Whole dataset (I)	25.0 (4.7–69.9)	(1.34–11.88)	12	0 (0–44.9)	(0.41–8.18)	
Vertebrate (II) intraspecies	46.7 (2.3–98.2)	(1.07–16.57)		34.1 (0–86.3)	(0.80–16.22)	
Vertebrate (III) interspecies	68.7 (8.4–100)	(2.09–12.43)	6,12	52.3 (2.4–100)	(1.19–13.66)	12
Invertebrate (IV) intraspecies	0 (0–61.3)	(0.32–5.28)		0 (0–59.8)	(0.16–5.79)	
Invertebrate (V) interspecies	35.8 (6.0–90.4)	(1.25–10.94)	10,11	15.0 (0–61.4)	(0.46–8.16)	
Second method of analysis						
Whole dataset (I)	18.2 (6.7–54.5)	(3.34–24.29)	2	6.3 (0.8–40.5)	(1.28–22.07)	
Vertebrate (II) intraspecies	17.4 (4.3–77.9)	(2.29–36.55)		12.4 (1.4–63.5)	(1.58–39.20)	
Vertebrate (III) interspecies	56.8 (20.6–96.8)	(4.60–29.44)	12,14	57.5 (16.2–90.6)	(4.10–32.25)	12,14
Invertebrate (IV) intraspecies	14.5 (0–85.7)	(0.86–12.67)		5.8 (0–60.3)	(0.34–11.69)	
Invertebrate (V) interspecies	10.5 (1.1–70.8)	(1.16–17.08)		0 (0–53.9)	(0.60–13.81)	

Values for R^2_{adj} ratio are obtained from Table 1. F is the ratio of mean square for deviation of protein means from regression divided by the mean square for the residual variation in d. 95% confidence limits are given in brackets. See text for proteins corresponding to given codes.

of independence arising from uncertain knowledge of the true phylogeny of the taxa within studies, but gives larger confidence intervals.

Mean genetic distance (D) is plotted against mean heterozygosity (H) for the seven proteins for groups I to V in Figure 1 (for the first method of analysis) and Figure 2 (for the second method of analysis). Regression statistics for the two methods of analysis are given in Table 2. The linear correlation (R_{lin}) between H and d is positive and significantly different from zero, as judged by the confidence intervals, for all groups for both methods of analysis. The limits obtained by bootstrapping (round brackets) are wider than those obtained using z statistics (square brackets) where standard error is a function of the number of taxon pairs. There are significant differences in D between proteins though the R^2_{adj} values are generally small in magnitude ranging from 0.5–8.8%. The R^2_{adj} ratio given in Table 2 represents the variation in d explained by regression on H expressed as a percentage of the variation in d explained by regression on proteins. It is thus the percentage of the variation in D that can be accounted for by regression on H.

Differences in the values of the percentages are reflected in the scatter of points plotted in Figures 1 and 2. For example, this can be seen by comparing the percentages and plots for the vertebrate interspecies data (III) and the invertebrate intraspecies data (IV). The general picture to emerge is that between 50% and 100% of the variation in D can be accounted for by variation in H. The percentages are generally higher in the second method of analysis, and in the invertebrate and intraspecies datasets. The confidence intervals are narrower in the second method of analysis, which is the expected result of using data for all loci for each protein. The percentages are mostly somewhat higher with linear regression, which is the more conservative test from the neutralist viewpoint, than with polynomial regression.

Figures 1 and 2 provide evidence of a difference between the datasets when a curvilinear function is fitted. In line with neutral theory (MUKHERJEE, SKIBINSKI and WARD 1987), the relationship between D and H is close to a linear one for the intraspecies datasets, but more closely asymptotic for the interspecies datasets.

It can be seen from Table 2 that additional variation in d can be explained by adding h to the regression equation after proteins have been added. The partial regression coefficients (not tabulated) are all positive, therefore within proteins, those loci having high heterozygosity will have accumulated more genetic distance on average in the recent past than those loci with low heterozygosity.

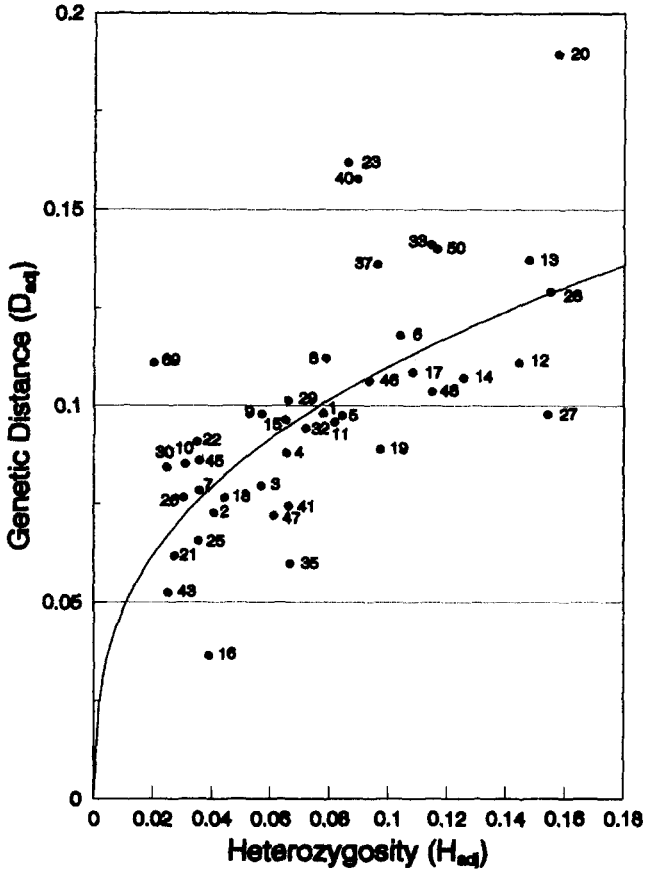

FIGURE 3.—Genetic distance (D_{adj}) plotted against heterozygosity (H_{adj}) for 42 proteins for the whole dataset (I). The curvilinear function $\hat{D}_{adj} = 0.259\ \hat{H}_{adj}^{0.358}$ derived from regression analysis is fitted to the plotted points. See text for explanation of protein codes.

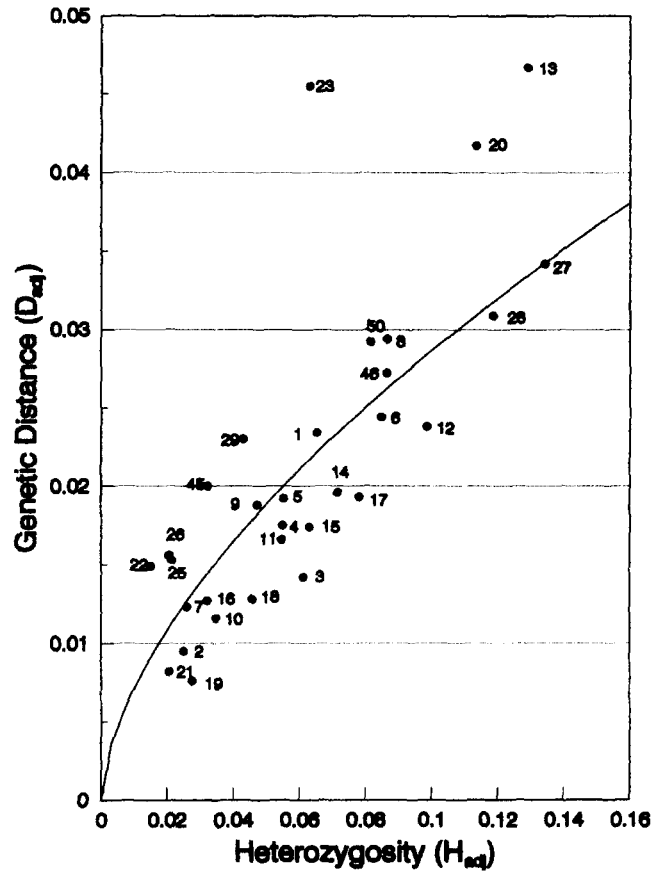

FIGURE 4.—Genetic distance (D_{adj}) plotted against heterozygosity (H_{adj}) for 31 proteins for the vertebrate intraspecies dataset (II). The curvilinear function $\hat{D}_{adj} = 0.122\ \hat{H}_{adj}^{0.606}$ is fitted to the plotted points. See text for explanation of protein codes.

The R^2_{adj} ratios and their confidence intervals, transformed by subtraction from 100, are given in Table 3, and represent the variation in D that cannot be explained by regression on H. Of the 20 transformed ratios in the table, 12 are significant and 8 nonsignificant. To confirm this result, bootstrapping was carried out on the F-ratio obtained by dividing the mean square for deviations of protein means from regression by the mean square for the residual variation in d. The confidence intervals for the F-ratio are given in Table 3. It can be seen that only the transformed ratios which have confidence intervals that overlap zero are associated with an F-ratio in which the lower confidence interval is less than one. Table 3 also gives the individual proteins whose 97.5% confidence intervals for D fall on one side of the fitted regression line. These deviant proteins can be seen to be outliers in relation to the fitted linear regression lines in Figures 1 and 2. For example in group III, proteins 12 (phosphoglucomutase) and 14 (phosphoglucose isomerase) have similar H values but different D values and clearly both cannot be made to fit the line closely. Hence phosphoglucomutase deviates significantly in the first method of analysis and both proteins deviate significantly in the second method of analysis.

The results of the first and second methods of analysis support the hypothesis that a substantial percentage of the variation in genetic distance of seven proteins in a balanced dataset result from variation in neutral mutation rate. However, a smaller percentage cannot be accounted for in this way.

The entire database (the third method of analysis): To take full advantage of the size of the database, the relationship between D and H has also been studied using 42 proteins that have been scored over the whole database in 50 or more studies. To simulate the approach used with a balanced dataset, the values of H and D were adjusted for differences between studies in the pattern of representation of different proteins (see Appendix).

Adjusted genetic distance (D_{adj}) is plotted against adjusted heterozygosity (H_{adj}) for the five groups in Figures 3-7. For groups II to V, fewer than 42 proteins meet the criterion of having been scored in 50 or more studies. The figures demonstrate a clear positive correlation between heterozygosity and genetic distance. Greater genetic distance is accumulated in the interspecies than intraspecies datasets. A regression

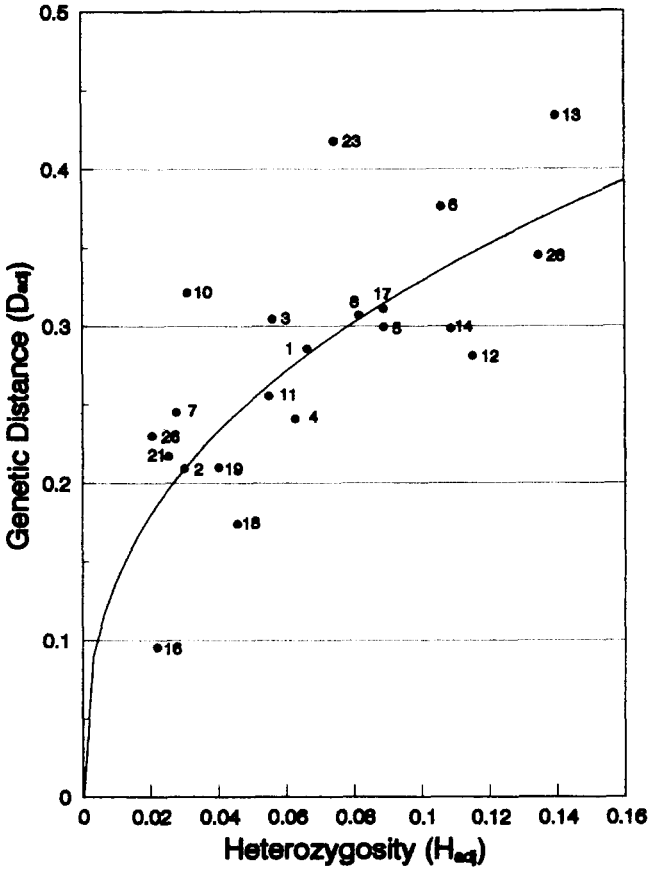

FIGURE 5.—Genetic distance (D_{adj}) plotted against heterozygosity (H_{adj}) for 21 proteins for the vertebrate interspecies dataset (III). The curvilinear function $\hat{D}_{adj} = 0.797\ \hat{H}_{adj}^{0.377}$ is fitted to the plotted points. See text for explanation of protein codes.

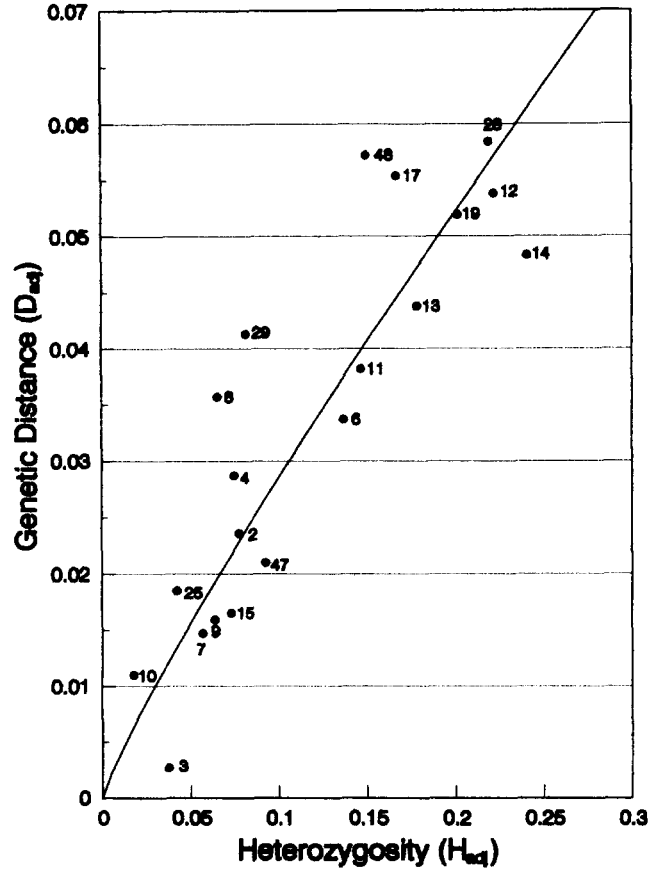

FIGURE 6.—Genetic distance (D_{adj}) plotted against heterozygosity (H_{adj}) for 20 proteins for the invertebrate intraspecies dataset (IV). The curvilinear function $\hat{D}_{adj} = 0.215\ \hat{H}_{adj}^{0.871}$ is fitted to the plotted points. See text for explanation of protein codes.

line was fitted to the plotted points using the curvilinear function $D_{adj} = a\ H_{adj}{}^{b}$. In line with the expectations of neutral theory, the relationship is close to a linear one for the intraspecies datasets but is asymptotic for the interspecies datasets.

Statistics from the regression analysis are given in Table 4. The R^2_{adj} values for the fitted curvilinear function (column 1) are similar in value to those obtained in the first and second methods of analysis. The fourth column in Table 4 gives the R^2_{adj} values for the model with regression through the origin. The percentage of the sum of squared D_{adj} values about the origin explained by regression equals $100 \times (1 - (\text{residual sums of squares of } D_{adj})/(\sum D^2_{adj}))$. Values of R^2_{adj} of over 95% are obtained, indicating a very good fit to this model.

An analysis of the deviations from regression in the third method of analysis was undertaken in the following way. The 800 studies in the database were sorted into 400 matched pairs. The variables that were used for matching the studies were the group (II to V), heterozygosity and genetic identity averaged over all proteins and taxon comparisons in the study, and the number of taxa in the study. The two studies in each matched pair were then assigned arbitrarily, one to each of two sets (A and B). The regression analysis was then repeated independently on each of sets A and B using the same proteins as in the original analysis. The residual deviation values, one for each set, for each protein were calculated and the correlation of residuals ($R(res)$) between sets across proteins calculated for each of groups I to V. If proteins have consistent deviations above or below the fitted regression line, there will be a positive correlation and the values of $R^2_{adj}(res)$ will reflect the amount of residual variation attributable to deviations from neutrality. In the split-half design used here, the expected coefficient of determination for this residual variation for the total sample of 800 studies ($R^2_{adj}(total)$) can be estimated from the equation $R^2_{adj}(total) = (2R^2_{adj}(res))/(1 + R^2_{adj}(res))$ (ROSCOE 1969). The values obtained are given in columns two and five of Table 4. For the whole dataset and the vertebrate interspecies dataset, the values indicate that a substantial proportion of the residual variation does reflect true deviations from regression and is not simply a result of the sampling variance of loci within proteins, which in the neutral model would be the drift variance. For the other groups, the values are smaller and, having confidence intervals that overlap zero, are nonsignificant. The

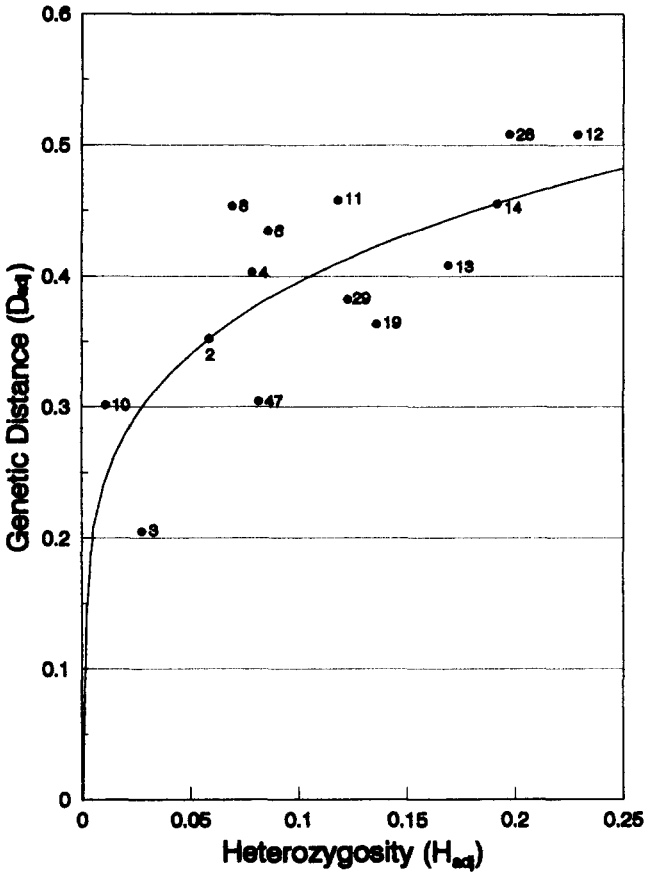

FIGURE 7.—Genetic distance (D_{adj}) plotted against heterozygosity (H_{adj}) for 14 proteins for the invertebrate interspecies dataset (V). The curvilinear function $\hat{D}_{adj} = 0.660\,\hat{H}_{adj})^{0.217}$ is fitted to the plotted points. See text for explanation of protein codes.

values of R^2_{adj} for the fitted curves given in columns 1 and 4 can be recalculated as a percentage of the total variation after eliminating the variation unexplained by the correlation of residuals. The equation used is $R^2_{adj}(ratio) = [R^2_{adj}/[R^2_{\text{adj}}(total)\ [1 - R^2_{adj}] + R^2_{adj}]$. The resulting values, given in columns three and six of Table 4, are analogous to the R^2_{adj} ratios in Table 2. The results for the curvilinear function suggest that at least 60% of the variation is explained by regression and that this percentage is closer to 90% for the invertebrate groups. The values for regression through the origin are close to 100% for all groups.

Failure of the correction method used in the calculation of H_{adj} and D_{adj} should be considered as a possible source of error in the measured residual deviations. Moreover, the matching procedure used in the split-half design is expected to correct only for the effects of the variables used when assigning the matched pairs. These factors will cause some uncertainty in the values of the ratios given in columns three and six of Table 4.

The third method of analysis allows the conclusions of the first and second methods of analysis to be extended to a larger sample of proteins.

Paired protein comparisons (the fourth method of analysis): In Figures 3–7, several proteins appear to be outliers of the fitted curve. For example, for the whole dataset, transferrin (20), albumin (23) and pyruvate kinase (69) have unusually high values of genetic distance given their heterozygosity values, whereas glutamate dehydrogenase (16) has an unusually low heterozygosity. To assess the significance of these outlying proteins, the approach used in the fourth method of analysis has been to select a pair of enzymes with similar mean heterozygosity and compare their genetic distance values in a $2 \times n$ balanced dataset, where n is the number of studies across the whole database where both proteins are scored. For example, in Figure 3 it can be seen that malate dehydrogenase (2) and glutamate dehydrogenase (16) are closely similar in heterozygosity but differ greatly in genetic distance. These two proteins have both been scored in 177 of the 800 studies in the database. Values of H and D were calculated for each protein in the 2×177 balanced dataset using the analytical approach of the second method of analysis. The values of D and H for these two proteins are plotted in Figure 8, the points linked by a straight line.

Bootstrapping across studies has been used to test for a significant difference in distance between the two proteins. For each of 5000 bootstrap samples, 177 studies were selected at random with replacement and values of H and D for the two proteins recalculated. The difference is significant if the proportion of bootstrap samples in which the D value for the higher distance protein (2, malate dehydrogenase) exceeds that for the lower distance protein (16, glutamate dehydrogenase), is greater than 0.975. The test has been modified by dividing the bootstrap samples into three categories according to whether the bootstrapped H value for the protein having the highest D value in the raw data (2, malate dehydrogenase) is greater than, equal to or less than the bootstrapped H value for the second protein. The results are given in Table 5. Malate dehydrogenase (2) has a slightly higher value of H in the raw data than glutamate dehydrogenase (16). Thus, for a critical test emphasis should be placed on the category containing 2718 bootstrap samples (Table 5, column 5) in which the value of H for glutamate dehydrogenase (16) is greater than for malate dehydrogenase (2). Of these, 2704 (a proportion of 0.994) showed malate dehydrogenase to have the higher D value. It can be concluded that the two proteins differ significantly in genetic distance though they have similar heterozygosity.

The analysis was repeated with seven other pairs of proteins. Five of these involve one outlying protein with respect to the fitted line in Figure 3 (20, 23 or 69) and a more frequently scored protein falling close to the line. Although transferrin (20) appears to be an outlying protein in Figure 3, use of a balanced

TABLE 4

Results of regression analysis of genetic distance on heterozygosity for third method of analysis

Group	Curvilinear function R^2_{adj} %			Regression through origin R^2_{adj} %			Number of proteins
	Fitted curve	Residuals: split design	Ratio %	Fitted curve	Residuals: split design	Ratio %	
Whole dataset (I)	43.6 (18.4–63.1)	48.5 (15.8–70.0)	61.4	95.2 (90.5–97.1)	49.2 (16.6–70.4)	97.6	42
Vertebrate (II) intraspecies	60.1 (31.5–76.8)	0 (0–25.7)	60.1	92.7 (84.0–95.9)	0 (0–26.8)	100	31
Vertebrate (III) interspecies	50.2 (12.6–73.6)	60.0 (11.8–81.7)	62.6	96.4 (89.8–98.2)	59.9 (11.6–81.6)	97.8	21
Invertebrate (IV) intraspecies	75.3 (44.1–87.8)	12.9 (0–57.2)	96.0	94.7 (84.9–97.3)	10.0 (0–55.3)	99.4	20
Invertebrate (V) interspecies	53.9 (5.6–79.1)	18.9 (0–67.7)	86.1	98.1 (92.2–99.1)	18.5 (0–67.5)	99.7	14

95% confidence limits derived using the z distribution are given in brackets.

dataset as in Figure 8 suggests that the high D value might simply be a reflection of high H, in line with the predictions of neutral theory. Of the three comparisons involving transferrin (20), only that with esterase (13) has enough bootstrap samples in the critical category to allow a test. This fails to provide evidence of a significant difference. By contrast, significant differences are found in the comparisons of albumin (23) with 6-phosphogluconate dehydrogenase (6) and pyruvate kinase (69) with malate dehydrogenase (2). One comparison was made involving an outlying protein with low D (16 in the pair 2 + 16) and as reported above a significant difference was found. The remaining two comparisons involve pairs in which both proteins are close to the fitted line. For sorbitol dehydrogenase (5) and malic enzyme (8), the proportion has a high value in the critical test, but is not significant. For malate dehydrogenase (2) and lactate dehydrogenase (7), the difference in D is significant.

The fourth method of analysis provides evidence of significant differences in genetic distance between proteins with similar heterozygosity, even if, as with malate dehydrogenase (2) and lactate dehydrogenase (7), the differences in distance are quite small. These differences between proteins cannot be attributed to differences in neutral mutation rate.

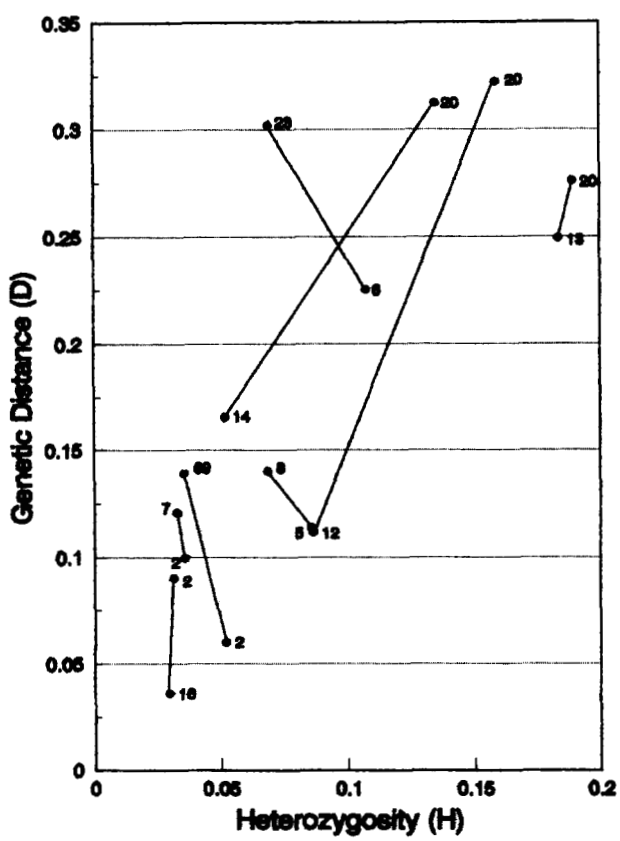

FIGURE 8.—Mean distance (D) plotted against mean heterozygosity (H) for pairs of proteins compared in balanced datasets. Members of a pair are joined by straight lines. The proteins are 2: malate dehydrogenase; 5: sorbitol dehydrogenase; 7: lactate dehydrogenase; 8: malic enzyme; 12: phosphoglucomutase; 13: esterase; 14: phosphoglucose isomerase; 20: transferrin; 23: albumin; 69: pyruvate kinase.

DISCUSSION

The results of the analyses described in this study demonstrate that most of the variation in mean genetic distance of proteins can be explained statistically by regression on mean protein heterozygosity. With regression about the mean, the percentage is between 50% and 100%. The correlations obtained are higher than comparable values obtained in previous studies that employed a smaller body of data (WARD and SKIBINSKI 1985). The results provide quantitative support for neutral theory, which predicts a high correlation between the variables, caused by differences between proteins in neutral mutation rate. The results

TABLE 5

Results of genetic distance comparisons of paired proteins

Protein codes		Heterozygosity category			Number of studies
HI	*LO*	$Het_{HI} > Het_{LO}$	$Het_{HI} = Het_{LO}$	$Het_{HI} < Het_{LO}$	
20	13	0.649 (3173)	0.625 (8)	0.426 (1819)	70
20	12	1.000 (4998)	(0)	1.000 (2)	86
20	14	0.990 (4998)	(0)	1.000 (2)	75
23	6	1.000 (11)	(0)	0.999 (4989)	114
8	5	0.945 (787)	1.000 (9)	0.928 (4204)	146
7	2	1.000 (1064)	1.000 (38)	0.997 (3898)	567
2	16	0.999 (2245)	1.000 (37)	0.994 (2718)	177
69	2	1.000 (2056)	1.000 (16)	0.999 (2928)	61

Columns three, four and five give the proportion of bootstrap replicates in which the mean distance (*D*) for the higher distance protein (*HI*) is greater than for the lower distance protein (*LO*) for replicates in which average heterozygosity (*H*) for protein designated *HI* is either greater (column 3), equal to (column 4) or less than (column 5) the average heterozygosity for the protein designated *LO*. The number of bootstrap replicates in each category is given in brackets. See text for proteins corresponding to given codes.

might also be consistent with models incorporating moderate selection; however, variation in neutral mutation rate would not be displaced as the dominant cause of the variation in protein heterozygosity and distance. If it is assumed that the fitted regression lines represent genetic distance accumulated by neutral evolution, so that regression through the origin is relevant, the percentage is over 95%. In the future, it may be possible to check this assumption by plotting comparable data for heterozygosity and genetic distance of silent DNA variation on the same graphs as the allozyme data.

The results are not in agreement with the view that, although silent variation in the genome is selectively neutral, protein variation is strongly influenced by selection. Using an argument based on the index of dispersion, GILLESPIE (1991) concluded that the upper limit for the percentage of neutral amino acid replacement substitutions in three mammalian orders is closer to 10% than 90%. The allozyme data used in the present study cover a wider taxonomic range, involve a special subset of replacement variation and, particularly at low divergence, reflect allele frequency shifts rather than allelic substitutions. Thus, the conclusions from the two approaches are not necessarily in conflict. Other lines of evidence appear to favor selection over neutral theory. For example, detailed studies of individual polymorphisms frequently reveal functional differences between allozyme genotypes (POWERS 1990). Yet little is known about how representative these studies are of allozyme variation in general nor about the fitness consequences of functional differences (LEWONTIN 1974). Thus, their relevance to the quantitative conclusions of the present study is difficult to gauge. Much evidence in favor of selection, including that from the DNA tests analyzing the relationship between polymorphism and evolutionary rate, comes from the refutation of null hypotheses derived from neutral theory. The intensity of selection required to allow rejection of null hypotheses may however be quite small and consistent with the quantitative dominance of neutral mutation as a cause of the variation in mean protein heterozygosity and distance.

This study also reports significant residual variation in mean protein distance about the fitted regression lines and significant differences in distance between proteins with similar heterozygosity. This additional variation refutes the null hypothesis derived from strictly neutral theory and might also be incompatible with models incorporating moderate selection. If a fitted-line follows the neutral expectation, positive directional selection for advantageous mutations could shift proteins above the line (as with pyruvate kinase), while balancing selection or mutation selection balance could shift proteins below the line (as with glutamate dehydrogenase). However, factors other than selection might also contribute to the residual variation, for example, biased gene conversion.

Theories of strong selection permit many parameters. Thus, both the positive correlation and the deviations from regression are easily consistent with an entirely selectionist interpretation. In the SAS-CFF model, heterozygosity would be a function of the mean and variance of allelic fitness determined by short-term environmental fluctuations, genetic distance a function of the rate of longer term environmental fluctuations in the episodic process. Polymorphism and evolutionary rate are said to be uncoupled (GILLESPIE 1984, 1989), and it is not clear whether a causal link in the form of a single parameter, analogous to mutation rate in neutral theory, should exist to explain the observed dominant positive correlation. A possible compromise between neutral and selection theory is that neutral polymorphism possesses latent adaptive potential that is realized by selection when environments change (GILLESPIE 1984; KIMURA 1991b). However, if a single parameter relating het-

erozygosity to the probability of a selective shift is introduced, neutral mutation rate could remain as the dominant cause of variation in protein distance.

In neutral theory, differences in neutral mutation rate between proteins are related to differences in constraint. With reference to strong selection, GILLESPIE (1991) uses the term "environmental challenge" to refer to the analogous propensity of proteins to respond to or experience a changing environment. In this context, it is important to note that in a balanced dataset all proteins share the same overall external environment. Thus, the search for deeper understanding of adaptive causes of differences in heterozygosity and distance should not be directed outward toward a study of the ecology and natural history of individual species but inward toward a study of the structure, biochemical environment or other properties of proteins, common to all species, that might influence responsiveness to environmental change. The observation that subunit size and number are more highly correlated with protein heterozygosity than is protein function (see WARD, SKIBINSKI and WOODWARK 1992) points toward the importance of structural features.

A relatively small percentage of the total variation in locus distance (d) is accounted for by regression on H. Some of the remaining variation, attributable in neutral theory to genetic drift and the stochastic nature of the mutational process, can be explained statistically by locus heterozygosity (h). This suggests that loci having high heterozygosity will have accumulated more genetic distance on average in the recent past than loci with low heterozygosity, a result consistent with those of an earlier study which suggested that as species diverge, the most highly heterozygous loci tend to be those that accumulate genetic distance most rapidly (SKIBINSKI and WARD 1981). This correlation of h and d within proteins is an expected historical consequence not only of genetic drift but probably also of strong selection, given that polymorphism at a locus is a prerequisite for evolutionary change.

It is of interest to compare the approach used in the present study with those used in DNA tests that consider the relationship between polymorphism and evolutionary rate. In DNA tests, contrasting levels of variation in, for example, coding and noncoding regions, are assumed in the null hypotheses tested to reflect differences in underlying neutral mutation rate. However, evidence for differential mutation rates and complex selective pressures in synonymous codons within genes (*e.g.*, RILEY 1989), insertion deletion polymorphism in regions flanking genes (*e.g.*, AQUADRO *et al.* 1986) and concerted evolution in noncoding regions (DOVER 1988) throws some doubt on the validity of estimates of DNA site variation in tests of neutral theory. There should be less doubt for estimates of allozyme variation, which arises from a less heterogeneous mutational process, that of base substitution in coding regions. Another point for comparison concerns the degree of independence of linked replicate sites in, for example, a given coding region. Within a species these may have been influenced as a group by historical factors such as population bottlenecks, migrational events or hitchhiking. This complicates the execution and interpretation of DNA tests of neutral theory. By contrast, the functionally homologous replicate loci for a specific protein within a balanced allozyme dataset are isolated in different taxa and will thus experience different and independent historical influences. Over a large dataset, the net influence should be similar for different proteins. Finally, theoretical studies (OHTA and TACHIDA 1990; OHTA 1992) demonstrate that variation in population structure and effective size, in the pattern of subpopulation extinction and recolonization, and in spatial patterns of selection, can influence moderately selected and neutral variation differently. If replacement variation is assumed to be moderately selected, and silent variation neutral, the additional parameters can generate those differences observed between these two types of variation in DNA tests and usually attributed to the contrasting influences of strong selection and neutral evolution. If allozyme variation is moderately selected, differences between the results of the present study and those of the DNA tests can be reconciled without invoking strong selection.

Because it uses pooled information covering much of the animal kingdom, an advantage of the allozyme database approach to the analysis of the relationship between intraspecific variation and evolutionary rate, compared with the taxonomically more limited DNA approaches, is that the global relevance of theories can be tested. To some, the pooling of allozyme studies from diverse sources is distasteful, and alien to the philosophy that an understanding of patterns of variation requires careful study of the ecology and natural history of individual species. However, what is relevant in the testing of neutral theory is estimation of only a few parameters. These are common to all species, and thus the detection of their global influence is not necessarily obscured by pooling. The same might be true with the testing of a unifying selection theory of global relevance that incorporates few parameters.

This work was supported by grants from the Natural Environment Research Council of the UK. We gratefully acknowledge the help of CHRISTINE BEYNON in the construction of the database. We are indebted to the following who kindly provided us with unpublished data for inclusion in the database: T. M. BERT, B. CROUAU-ROY, P. DRATCH, K. L. HAMRICK, J. D. HERNANDEZ-MARTICH, N. KOMALAMSIRA, E. PILLA and D. W. SUGG. Previous contributors to the database have been acknowledged in WARD, SKIBINSKI and

WOODWARK (1992) and WOODWARK, SKIBINSKI and WARD (1992). We wish to thank DERWYN EVANS of the Computer Centre, Swansea, for his patient help with the bootstrapping, and JOHN GILLESPIE and an anonymous reviewer for helpful critical comments of an earlier version of this paper.

LITERATURE CITED

AQUADRO, C. F., S. F. DEESE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

BEGUN, D. J., and C. F. AQUADRO, 1991 Molecular population genetics of the distal portion of the *X* chromosome in Drosophila: evidence for genetic hitchhiking of the yellow-achaete region. Genetics **129:** 1147–1158.

BERRY, A. J., J. W. AJIOKA and M. KREITMAN, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

BRADY, J. P., R. C. RICHMOND and J. G. OAKESHOTT, 1990 Cloning of the Esterase-5 locus from *Drosophila pseudoobscura* and comparison with its homologue in *D. melanogaster*. Mol. Biol. Evol. **7:** 525–546.

CHAKRABORTY, R., 1984 Relationship between heterozygosity and genetic distance in the three major races of man. Am. J. Phys. Anthropol. **65:** 249–258.

CHAKRABORTY, R., P. A. FUERST and M. NEI, 1978 Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. Genetics **88:** 367–390.

CHAKRABORTY, R., P. A. FUERST and M. NEI, 1980 Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics **91:** 1039–1063

CHAKRABORTY, R., and P. W. HEDRICK, 1983 Heterozygosity and genetic distance of proteins. Nature **304:** 755.

CHAKRABORTY, R., and M. NEI, 1977 Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution **31:** 347–356.

DOVER, G. A., 1988 Three into two won't go. Nature **331:** 121.

FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations I. Distribution of single locus heterozygosity. Genetics **86:** 455–483.

GILLESPIE, J. H., 1984 The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA **81:** 8009–8013.

GILLESPIE, J. H., 1989 Could natural selection account for molecular evolution and polymorphism? Genome **31:** 311–315.

GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution.* Oxford University Press, New York.

HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

KIMURA, M., 1979 Model of effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA **76:** 3440–3444.

KIMURA, M., 1989 The neutral theory of molecular evolution and the world view of neutralists. Genome **31:** 24–31.

KIMURA, M., 1991a The neutral theory of molecular evolution: a review of recent evidence. Jpn. J. Genet. **66:** 367–386.

KIMURA, M., 1991b Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc. Natl. Acad. Sci. USA **88:** 5969–5973.

KIMURA, M., and J. R. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LANGE, B. W., C. H. LANGLEY and W. STEPHAN, 1990 Molecular evolution of Drosophila metallothionein genes. Genetics **126:** 921–932.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

LEWONTIN, R. C., 1991 Electrophoresis in the development of evolutionary genetics: milestone or millstone? Genetics **128:** 657–662.

LI, W. H., 1976 Electrophoretic identity of proteins in a finite population and genetic distance between taxa. Genet. Res. **28:** 119–127.

LIVSHITS, G., and M. NEI, 1990 Relationships between intrapopulational and interpopulational genetic diversity in man. Ann. Hum. Biol. **17:** 501–513.

MARTINCAMPOS, J. M., J. M. COMERON, N. MIYASHITA and M. AGUADE, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

MUKHERJEE, M., D. O. F. SKIBINSKI and R. D. WARD, 1987 A simulation study of the neutral evolution of heterozygosity and genetic distance. Heredity **58:** 413–423.

NEI, M., 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

NEI, M., and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73–118.

NIE, N. H., C. H. HULL, J. G. JENKINS, K. STEINBRENNER and D. BENT, 1975 *Statistical Package for the Social Sciences*, Ed 2. SPSS, Inc., New York.

OHTA, T., 1976 Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theor. Popul. Biol. **10:** 254–275.

OHTA, T., 1977 Extension of the neutral mutation drift hypothesis, pp. 148–167 in *Molecular Evolution and Polymorphism*, edited by M. KIMURA. National Institute of Genetics, Mishima, Japan.

OHTA, T., 1992 Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. Genetics **130:** 917–923.

OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

OHTA, T., and H. TACHIDA, 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219–229.

POWERS D. A., 1990 The adaptive significance of allelic isozyme variation in natural populations, pp. 324–339 in *Electrophoretic and Isoelectric Focusing Techniques in Fisheries Management*, edited by D. H. WHITMORE. CRC Press, Boca Raton, Fla.

RILEY, M. A., 1989 Nucleotide sequence of the *Xdh* region in *Drosophila pseudoobscura* and an analysis of the evolution of synonymous codons. Mol. Evol. Biol. **6:** 33–52.

ROSCOE, J. T., 1969 *Fundamental Research Statistics for the Behavioral Sciences.* Holt, Rinehart and Winston, New York.

SKIBINSKI, D. O. F., and R. D. WARD, 1981 Relationship between allozyme heterozygosity and rates of divergence. Genet. Res. **38:** 71–92.

SKIBINSKI, D. O. F., and R. D. WARD, 1982 Correlations between heterozygosity and evolutionary rate of proteins. Nature **298:** 490–492.

SKIBINSKI, D. O. F., and R. D. WARD, 1983 Heterozygosity and genetic distance of proteins—reply to Chakraborty and Hedrick. Nature **304:** 755–756.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry*. Freeman, San Francisco.

STEPHAN, W., and S. J. MITCHELL, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics **132:** 1039–1045.

TACHIDA, H., 1991 A study on a nearly neutral mutation model in finite populations. Genetics **128:** 183–192.

WARD, R. D., and D. O. F. SKIBINSKI, 1985 Observed relationships between protein heterozygosity and protein genetic distance and comparisons with neutral expectations. Genet. Res. **45:** 315–340.

WARD, R. D., D. O. F. SKIBINSKI and M. WOODWARK, 1992 Protein heterozygosity, protein structure, and taxonomic differentiation. Evol. Biol. **26:** 73–160.

WOODWARK, M., D. O. F. SKIBINSKI and R. D. WARD, 1992 A study of inter-locus allozyme heterozygosity correlations: implications for neutral theory. Heredity **69:** 190–198.

Communicating editor: C. C. LAURIE

APPENDIX

Calculation of *d* and *h*: Genetic distance (d) between two taxa at an allozyme locus in study n is defined as:

$$d = 1 - \Sigma x_j y_j / (\Sigma x_j^2 \cdot \Sigma y_j^2)^{1/2},$$

where x_j and y_j are the allele frequencies for the jth allele for taxa A and B, respectively. The mean expected heterozygosity (h) for the locus in the two taxa is defined as:

$$h = [(1 - \Sigma x_j^2) + (1 - \Sigma y_j^2)]/2.$$

In the first method of analysis, in situations where two or more loci are scored for a protein within a study, just one of the loci was selected at random as the representative of that protein in the study. This results in a balanced dataset in which each of the taxon pairs in the group of studies is represented by paired values of d and h for each of the seven different protein loci.

In the second method of analysis, within each study, the locus values of d and h were averaged for each protein over all loci scored for that protein and over all possible taxon pairs. This results in seven paired values of mean heterozygosity and distance for each study in the group. The symbols d and h are retained for these within study protein means.

In the first method of analysis, to correct for differences in overall heterozygosity and distance between taxon pairs, each d and h value was recalculated as:

$$d_{(\text{corrected})} = d - d_{(\text{taxon pair mean})} + d_{(\text{grand mean})}$$

$$h_{(\text{corrected})} = h - h_{(\text{taxon pair mean})} + h_{(\text{grand mean})}$$

where (taxon pair mean) refers to the mean value of d or h over the seven proteins for a taxon pair and (grand mean) refers to the mean of d or h over all taxon pairs and proteins in the group.

In the second method of analysis an analogous correction was applied, substituting study for taxon pair. Furthermore, to weight studies according to size as in the first method of analysis, the seven paired values of corrected h and d for a study were replicated in the group by a number of times equal to the number of independent taxon pairs in the study.

Calculation of *D* and *H*: In the first method of analysis, mean distance (D) and mean heterozygosity (H) for a given protein such as phosphoglucomutase are defined as:

$$D = \left(\sum_{n=1}^{N} \sum_{m=1}^{M_n} d_{nm} \right) \Big/ T,$$

and

$$H = \left(\sum_{n=1}^{N} \sum_{m=1}^{M_n} h_{nm} \right) \Big/ T,$$

where d_{nm} and h_{nm} are the values of d and h in taxon pair m in study n, and M_n is the number of taxon pairs in study n. N is the total number of studies, and T the total number of taxon pairs.

In the second method of analysis H and D are defined in an analogous manner except that the averaging is carried out over the replicated study means rather than taxon pair means. At low to moderate levels of levels of divergence, D is similar in value to the standard genetic distance of NEI (1972).

Calculation of adjusted *D* (D_{adj}) and adjusted *H* (H_{adj}): In the third method of analysis, independent taxon comparisons were used as in the first method of analysis, but for each taxon pair, h and d for a protein were calculated as means over all loci scored for that protein as in the second method of analysis. H and D, the means of h and d over taxon pairs and studies were calculated as in the first method of analysis.

Genetic distance and heterozygosity averaged over all loci scored in a pair of taxa are defined as:

$$d_p = \sum_{k=1}^{K} d_k/K,$$

and

$$h_p = \sum_{k=1}^{K} h_k/K,$$

where K is the number of loci scored in the taxon pair including loci for any of the 42 widely studied proteins and for any other proteins scored. A pair of d and h values for a given protein within a given taxon pair can be linked with the values of d_p and h_p for the taxon pair. By substituting d_p and h_p for d and h in the above equations for D and H, the analogous quantities D_p and H_p can be calculated.

Adjusted values of D and H are then calculated using the equations:

$$D_{\text{adj}} = D(d_{(\text{grand mean})}/D_p).$$

and

$$H_{\text{adj}} = H(h_{(\text{grand mean})}/H_p),$$

where $d_{(\text{grand mean})}$ and $h_{(\text{grand mean})}$ are as defined previously the means of h and d over all taxon pairs and all proteins. D_p is a measure of the amount of genetic distance, averaged over all loci, accumulated in those taxon pairs in which a particular protein is scored. Thus, proteins scored in a subset of taxon pairs with high divergence times will have a D_{adj} lower than D. An analogous adjustment is achieved in the calculation of H_{adj}.