J. theor. Biol. (1986) **118**, 443–469

The Evolution of Electrophoretic Mobility of Proteins

DAN GRAUR†

Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston, Houston, Texas 77225, U.S.A and Lehrstuhl für Populationsgenetik, Institut für Biologie II, Universität Tübingen, D-7400 Tübingen 1, West Germany

(*Received* 22 *May* 1985, *and in revised form* 13 *September* 1985)

A model was constructed that predicts the electric charge of a protein and its isoelectric point from its primary and quaternary structures. By using two different patterns of mutation and purifying selection, four schemes of nucleotide substitution were simulated. In the absence of selection for a specific value of pI, proteins are expected to evolve toward a mildly basic pI. Thus, the selection for maintaining extreme values of pI must be stringent, and proteins with extreme pI's will evolve very slowly. This prediction is consistent with observations on the evolution of histones and ubiquitin. The mean charge change is expected to be about 0·005 units pI per nucleotide substitution. The amount of electrophoretically hidden variation is expected to be considerable even for large degrees of divergence at the nucleotide and amino acid levels. Electrophoretic detectability depends on the size of the protein. The longer the protein the larger the amount of variation at the amino acid level that is undetectable by isoelectric focusing. This property may be partially responsible for the imperfect correlation between molecular weight and gene diversity observed for electrophoretic data. Very basic and very acidic proteins are expected to generate less electrophoretic variablity than proteins with intermediate pI's. Unequal rates of mutation between nucleotides and asymmetrical patterns of purifying selection have almost no effect on the equilibrium pI of proteins, but affect the rates of change in pI, and increase the amount of electrophoretically hidden variation in comparison to the expectations derived from random patterns of mutation and constant selection. Comparison of detectability of protein differences among four electrophoretical techniques suggests that the best performance is obtained by the sequential electrophoresis method.

Introduction

Protein electrophoresis is used extensively to investigate genetic variation and evolutionary processes at the molecular level. Heterozygosity or gene diversity for 20 loci or more has been studied for at least 400 species (for reviews see Powell, 1975; Fuerst *et al.*, 1977; Nevo, 1978; Hamrick *et al.*, 1979; Nei & Graur, 1984; Nevo *et al.*, 1984; Graur 1985*b*). The electrophoretic mobility of a protein is mainly determined by its net electric charge, but also by its molecular weight, volume and shape. The effects of molecular weight and volume on electrophoretic mobility have been investigated both experimentally and theoretically (Weber & Osborn, 1969; Laemmli, 1970; Rodbard & Chrambach, 1970; Chrambach & Rodbard, 1971; Neville, 1971). The effects of charge changes on protein electrophoretic mobility have been

† Present address: Department of Zoology, Tel Aviv University, P.O. Box 39040, Ramat Aviv, Tel Aviv 69978, Israel.

0022-5193/86/040443+27 $03.00/0

studied experimentally (Basset *et al.*, 1978; Ramshaw *et al.*, 1979; Fuerst & Ferrell, 1980; McLellan, 1984), but the question of how much genetic variability is detectable by electrophoretic methods has not been settled.

According to the charge-state model, unit charge-change substitutions are primarily responsible for differences in the electrophoretic mobility of genetic variants, and only four amino acids (i.e. lysine, arginine, glutamic acid and aspartic acid) are assumed to contribute to the overall net charge of a protein. By assuming equal probabilities of changes between nucleotides, it has been calculated that about 25–30% of the possible base substitutions will result in a change in the electrophoretic mobility (Marshal & Brown, 1975; Nei, 1975, pp. 22–26; Wilson *et al.*, 1977). There are reasons to believe that the charge-state model does not give a good description of electrophoretic variation. First, this model assumes discrete charge changes (e.g. from + to −, from 0 to +, or from + to ++) without considering the real dynamics of charge changes imposed by the dissociation constants of charged or polar amino acids. For example, a lysine ($pK = 10{\cdot}53$) to arginine ($pK = 12{\cdot}48$) substitution is assumed to be electrophoretically undetectable because it is a change from one positive charge to another. In practice, however, this kind of substitution (e.g. hemoglobin Athens-Georgia, $\beta 40^{arg \rightarrow lys}$ (Lehmann & Kynoch, 1976, p. 274)) may be detectable, at least when moderately basic buffers are used. Second, the charge-state model does not take into account the electric contributions of other amino acids, such as histidine and cysteine.

The charge-state model assumes that many amino acid substitutions produce no detectable differences in electrophoretic mobility (Nei & Chakraborty, 1973; Ohta & Kimura, 1973; Marshal & Brown, 1975; Brown *et al.*, 1981), and that the proportion of hidden variation will strongly depend on effective population size and allele frequency (Nei & Chakraborty, 1976). Experimental data show that while hidden variation is indeed more prevalent within high frequency electromorphs (Shumaker *et al.*, 1982) as predicted by Nei & Chakraborty (1976), the total amount of hidden variation is not always as large as predicted by the charge-state model. Indeed, it has been concluded that electrophoresis reveals most amino acid variation at the majority of protein loci (Johnson 1976, 1977; Ramshaw *et al.*, 1979; Ayala, 1982; Shumaker *et al.*, 1982).

The purpose of this study is to develop a mathematical model of charge change due to point mutations, and to estimate how much of the total genetic variability in proteins can be detected by electrophoresis. The effects of amino acid substitution on electrophoretic mobility for proteins of different initial amino acid constitutions will also be examined. In this study I will mainly be concerned with the method of isoelectric focusing. The efficiency of the electrofocusing method in detecting genetic variation in comparison with other electrophoretic techniques will also be studied.

Model for Predicting the Isolectric Point of a Protein

(A) ASSUMPTIONS

The isoelectric point (pI) is the only variable which determines the mobility of a protein in a pH gradient during electrophoretic focusing. Consequently, pI values

will provide us with the means of comparing proteins and studying their evolution. I assume that the electric charge of a polypeptide at a given pH and its isoelectric point can be inferred from its primary (amino acid sequence) and quaternary (number of subunits) structures. The rationale behind ignoring the secondary and tertiary structures in calculating the expected pI is that charged amino acids as well as other polar and hydrophylic moieties which determine the pI tend to assume positions on the surface of the protein, and are, consequently, unobscured by an interior location within a molecule. Thus, by looking at the amino acid composition of a protein one is expected to detect most of the contributors to its charge. Moreover, even if some of the charged amino acids are obscured inside the bulk of the molecule, it has been shown that by denaturing proteins (e.g. with heat or urea) it is possible to reveal more electrophoretic variability than by studying their native form (Bernstein *et al.*, 1973). This means that even those few polar moieties, which are hidden inside the interior, can be exposed to the electric field by simple experimental procedures.

In this study I ignore three factors than can sometimes result in sizeable variation in pK values within each amino acid, i.e. (1) electrostatic effects resulting from the ionization of neighboring moieties, e.g. the Bohr effect, (2) medium effects due to the proximity of hydrophobic residues, and (3) hydrogen bonding effects. These three factors will be referred to as "local charge environment effects" (a term suggested by D. Hewett-Emmett, personal communication). Ignoring factors (2) and (3) should cause no major biases since they are known to have only minor effects on the pK of an amino acid. Factor (1), however, can change the pK of an amino acid considerably (e.g. Perutz, 1983). In a few cases data are available such that local charge environment effects can be taken into account, for example, by determining the effective number of masked polar amino acids that do not contribute to the charge of a polypeptide (Beslow & Gurd, 1962; Shire *et al.*, 1974), and excluding these sites from the calculations. This procedure, however, requires knowledge beside the primary and quaternary structures, and, as we shall see later, all these blunt oversimplifications introduce no serious biases in the predicted isoelectric points. pI's are also influenced by ion binding (e.g. binding to carrier ampholites); the effect being determined by experimental conditions. However, this effect can be minimized by increasing the time allowed for a protein to reach its equilibrium position.

In the following, I assume that the charge of a protein is the sum of the individual contributions of its amino acid and non-amino acid components. The electric charge of a protein is determined mainly by two strong acids: the β-carboxyl residue of aspartic acid (asp) and the γ-carboxyl of glutamic acid (glu), and two strong bases: the ε-amino residue of lysine (lys) and the guanidino of arginine (arg). In addition, lesser contributions to the overall electric charge are made by a weak acid: the phenolic hydroxyl of tyrosine (tyr), and a weak base: the imidasole residue of histidine (his). Cysteine (cys) contributes to the electric charge of a molecule only when it is not involved in a cystine (cys–cys) disulfide bond. When in the cys form, a further contribution to the overall electric charge is made by its weak acidic moiety, sulfurhydryl.

(B) MATHEMATICAL FORMULATION

In predicting the overall electric charge and subsequently the isoelectric point of a protein, the following factors are considered:

(1) The number and kinds of charged amino acids (asp, glu, lys, arg, tyr, cys and his).
(2) The number and composition of the amino and carboxyl termini.
(3) The number of free cysteins.
(4) The number of phosphorylated amino acids.
(5) The number of acetylated or otherwise neutralized charged amino acids.
(6) The number and nature of prosthetic molecules (e.g. propionic acids, iron and carboxy and amino termini of the heme moiety in hemoglobin and cytochrome c).
(7) Post-translational modifications of amino acids (e.g. glu to 5-oxopyrrolidine-2-carboxylic acid in complement C1Q β-chain, Reid *et al.* (1982)).

We note that only factors 1 and 2 can be deduced from the primary and quaternary structures. These factors will be considered separately from factors 3–7.

To predict pI values from factors 1 and 2, I used two rearranged forms of the classical Henderson–Hasselbach equation (Lehninger 1975, p. 50)

$$\mathrm{pH} - \mathrm{p}K = \log (A/D) \tag{1}$$

where pK is the negative logarithm of the dissociation constant, and A and D are the concentrations of the proton acceptors and donors, respectively.

From Table 1 we see that the number of negative charges, $[R^-]$, equals the number of proton acceptors of the first four amino acids. For each amino acid, $D = 1 - A$. The proportion of charged amino acids at a site, A_i, is, thus, $A_i = a_i/(1+a_i)$, where $a_i = 10^{(\mathrm{pH}-\mathrm{p}K_i)}$.

Summing over the entire protein we obtain

$$[R^-] = \sum_{i=1}^{5} n_i \times a_i/(1+a_i) \tag{2}$$

where i stands for asp, glu, tyr, cys and COO^- termini and n_i is the number of i-th residues. In the case of cys (factor 3), I assume that all cys are free unless there is evidence in the literature to the contrary. Alternatively, we can use the rule of thumb (Lehninger, 1975, p. 945; D. Hewett-Emmett, personal communication) that intracellular proteins contain cys, while secreted proteins contains cys–cys. The number of positive charges, $[R^+]$, equals the number of proton donors of the last three amino acids in Table 1. Therefore

$$[R^+] = \sum_{j=1}^{4} n_j/(1+a_j) \tag{3}$$

where j = lys, his, arg and NH_4^+ termini and n_j is the number of j-th residues. Mean pK values for free amino acids were taken from Mahler & Cordes (1966, pp. 10–13), and are shown in Table 1. It is also possible to use pK values for amino acids in

TABLE 1

pK values and charge states of side chains of amino acids which contribute to the electric charge of molecules

Amino acid	Abbreviation	Side chain	pK'	Charge of proton	
				Donor	Acceptor
Asp	D	β-carboxyl	3·86	0	–
Glu	E	γ-carboxyl	4·25	0	–
Tyr	Y	phenolic hydroxyl	10·07	0	–
Cys	C	sulfhydryl	10·78	0	–
Lys	K	ϵ-amino	10·53	+	0
His	H	imidasole	6·00	+	0
Arg	R	guanidino	12·48	+	0

an "average" polypeptide (Stryer, 1975, p. 40, p. 80), but these are never very different from the ones I used. Ideally, one should use actual values of pK's for each protein (e.g. McLellan, 1984). Unfortunately, such data are scarce. The results, notwithstanding, are not affected significantly by the use of approximate values.

The number of negative charges contributed by the phosphoryl groups, $[P^-]$, in phosphorylated proteins (factor 4) was calculated by using the following rearranged Henderson–Hasselbach equation

$$[P^-] = n_p \times (m_1 + 2m_2 + 3m_3) \tag{4}$$

where n_p is the number of phosphoryl groups

$$m_1 = a_1/(1 + a_1 + a_1a_2 + a_1a_2a_3)$$

$$m_2 = a_1a_2/(1 + a_1 + a_1a_2 + a_1a_2a_3)$$

$$m_3 = a_1a_2a_3/(1 + a_1 + a_1a_2 + a_1a_2a_3)$$

$$a_1 = 10^{(\mathrm{pH}-\mathrm{p}K_1)}$$

$$a_2 = 10^{(\mathrm{pH}-\mathrm{p}K_2)}$$

$$a_3 = 10^{(\mathrm{pH}-\mathrm{p}K_3)}$$

pK_1 = 0·01, pK_2 = 7·21 and pK_3 = 12·32. pK_n values (n = 1, 2, 3) were taken from Mahler & Cordes (1966, p. 192).

The pI of a protein is the pH at which $[R^-] + [P^-] = [R^+]$, or, in cases we do not have knowledge on the number of phosphorylated amino acids, or the protein in question is not phosphorylated, when $[R^-] = [R^+]$.

Other contributions to the electric charge such as post-translational modifications, minor proteinous components, prosthetic groups (e.g. heme, sugar moieties) and partial acetylation, i.e. factors 5–7, were taken into account only qualitatively, and the signs [<], [>] or [?] were added to the predicted pI's to represent lower, higher or unknown effects on pI, respectively.

A computer program was written to calculate the net electric charge of a molecule at small intervals (0·01) within the pH range of 0·00–14·00, and to find iteratively the pH at which the charge is nearest to zero (the inferred pI).

ACCURACY OF THE MODEL

Let us now examine whether or not charge changes with increased pH are predicted correctly. For this purpose I compared the predicted net charge of bovine ribonuclease based on its amino acid composition with its experimental titration curve (Tanford & Hauenstein, 1956). The theoretical curve is virtually indistinguishable from the experimental one. Interestingly, this comparison revealed one potential use of the method; the possibility to check the accuracy of an amino acid sequence. For example, there was initially an ambiguity with respect to the amidation state of residue 103 of ribonuclease (asp or asn). Since this position is occupied by glu in many mammalian ribonucleases, it has been assumed that it is asp. As we see from Fig. 1, however, the theoretical curve with asp is different from the experimental curve. Indeed, amino acid 103 is asn (Smyth *et al.*, 1963). Similar fits between observed and expected titration curves were obtained for other proteins, and it is concluded that the charge of a protein at a given pH can be predicted accurately. In heme containing proteins, however, the fit was less satisfactory.

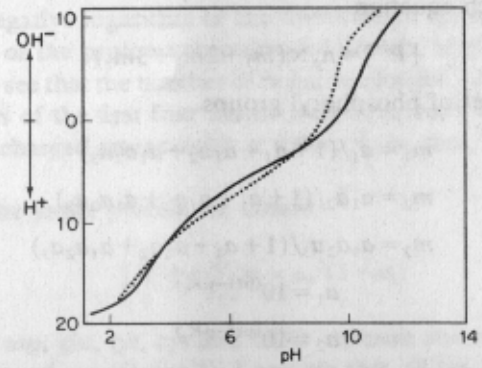

FIG. 1. Experimental and theoretical titration curves of bovine ribonuclease. Solid line: experimental results (data from Tanford & Hauenstein, 1956; Mahler & Cordes 1961, p. 52; Lehninger, 1975). Dotted line: theoretical simulation with the asp at position 103. The theoretical curve of ribonuclease is indistinguishable from the experimental result.

Next, I examined whether or not my model predicts pI values correctly. Experimental pI's for 28 proteins were taken from Mahler & Cordes (1966, p. 54), Lehninger (1970, p. 162) and Stryer (1975, p. 90). Data on the number of chains and number of disulfide bridges were taken from the articles listed in Table 2 and from Schachman (1963). Table 2 lists two predicted values of pI's (pI_A and pI_B) together with their experimentally determined pI's. pI_A values were computed by using only factors 1 and 2. For computing pI_B, factors 3–7 were also considered. The agreement between the predicted and the observed values is remarkable for both pI_A and pI_B. The

TABLE 2

Experimental and theoretically predicted values of pI

Organism	Protein	Experimental pI	Predicted pI		References
			A^1	B^2	
Chum salmon[3]	salmine A1	12.10	13.78	13.78	1
Herrings[4]	clupeines[5]	12.10	13·77	13·77	2
Caspian sturgeon[6]	sturine	11·71	13·74	13·74	3
Human	lysozyme	11·07	9·20	9·86	4
Human	cytochrome *c*	10·17	10·06	<9·96	5
Bovine	ribonuclease	9·60	9·98	?9·05[7]	6
Bovine	chymotripsinogen A	9·50	8·14	9·84[8]	7
Fruit fly[9]	alcohol dehydrogenase[10]	7·40	8·21	8·15	8
Human	myoglobin	6·99	7·65	<7·63	9
Human	hemoglobin	6·98	8·05	<8·03	10
Chicken	conalbumin (ovotransferrin)	6·95	6·34	6·38	11
Human	somatotropin	6·85	5·10	>5·10[11]	12
Bovine	β-crystallin	6·00	6·50	6·26	13
Human	prolactin	5·73	6·14	6·15	14
Human	fibrinogen	5·65	6·56	?6·60[12]	15
Sipunculids[13]	hemerythrin	5·60	6·11	<6·16[14]	16
Bovine	catalase	5·60	6·42	<6·33	17
Human	insulin	5·35	5·29	5·30	18
Rabbit	tropomyosin	5·10	4·49	<4·45[15]	19
Bovine	β-lactoglobulin	5·08	4·64	4·64[16]	20
Human	serum albumin	4·80	5·71	?5·65[17]	21
Chicken	ovalbumin	4·65	5·02	<4·82[18]	22
Pseudomonid[19]	rhodopsin	4·52	4·51	<4·51	23
Bovine	β-casein	4·50	5·08	4·51[20]	24
Chicken	α-ovomucoid	4·12	4·63	4·63[21]	25
Bovine	α_{S1}-casein	4·05	6·01	4·45[22]	26
Pig	pepsinogen A	3·70	3·75	3·71[23]	27
Pig	pepsin A	<1·00	2·98	?2·87[24]	27

References: (1) Ando & Watanabe, 1969. (2) Iwai *et al.*, 1971; Suzuki & Ando, 1972. (3) Yulikova *et al.*, 1976. (4) Canfield *et al.*, 1971; Thomsen *et al.*, 1972. (5) Matsubara & Smith, 1962, 1963. (6) Smyth *et al.*, 1963. (7) Brown & Hartley, 1966; Blow *et al.*, 1969. (8) Thatcher, 1980. (9) Romero-Herrera & Lehmann, 1971*a*, 1971*b*, 1972. (10) Lawn *et al.*, 1980; Liebhaber *et al.*, 1980; Michelson & Orkin, 1980. (11) Elleman & Williams, 1970; Jeltsch & Chambon, 1982. (12) Li & Dixon, 1971; DeNoto *et al.*, 1981. (13) Driessen *et al.*, 1980. (14) Cooke *et al.*, 1981. (15) Blombäck *et al.*, 1976; Gårlund *et al.*, 1977; Doolittle *et al.*, 1978; Watt *et al.*, 1979*a*, 1979*b*; Henschen *et al.*, 1980. (16) Klippenstein *et al.*, 1968; Ferrell & Kitto, 1971. (17) Murthy *et al.*, 1981; Schroeder *et al.*, 1982. (18) Nicol & Smith, 1960; Oyer *et al.*, 1971; Bell *et al.*, 1979, 1980; Sures *et al.*, 1980; Ullrich *et al.*, 1980. (19) Stone & Smillie, 1980; Mak *et al.*, 1980. (20) McKenzie *et al.*, 1972; Braunitzer *et al.*, 1973. (21) Walker, 1976; Saber *et al.*, 1977; Lawn *et al.*, 1981. (22) McReynolds *et al.*, 1978; Thompson & Fisher, 1978; Woo *et al.*, 1981. (23) Ovchinnikov *et al.*, 1978; Dunn *et al.*, 1981. (24) Ribadeau-Dumas *et al.*, 1972. (25) Catterall *et al.*, 1980. (26) Mercier *et al.*, 1971, 1973; Brignon *et al.*, 1977. (27) Stepanov *et al.*, 1973; Morávek & Kostka, 1974; Sepulveda *et al.* 1975.

Notes: (1) A = Predicted values of pI by taking into account amino acid composition only. (2) B = Predicted values of pI by taking into account disulphide bonds, radicals, prosthetic groups and post-translational modifications. (3) *Oncorhynchus keta.* (4) Pacific herring (*Clupea pallasii*) and European herring (*C. harengus*). (5) Clupeines Z and YII. (6) *Acipenster gouldenstadti.* (7) pI is affected in an unknown fashion by the binding of asn at position 34 to a carbohydrate moiety in some of the molecules (Plummer & Hirs, 1964). (8) At position 102 the amidation state of the residue was not determined

Continued overleaf

unambiguously, and it may be asp instead of asn (Meloun *et al.*, 1966). In this case the predicted pI is 8·30. (9) *Drosophila melanogaster.* (10) Adh^F. (11) Variants are known to have gln instead of glu at position 74 (Niall *et al.*, 1971) with a predicted pI value of 5·10. The variant described by Seeburg (1982) has a predicted pI of 7·80 (12) Variants are known to have glu instead of gln at position 128, asp instead of asn at positions 177, 212 and 390, and asn instead of asp at position 388 of the α-chain (Henschen *et al.*, 1980), asp instead of asn at position 202 and glu instead of gln at position 301 of the β-chain (Watt *et al.*, 1979*b*) with a predicted pI value of 6·53. The gln moiety at position 1 of the β-chain is modified post-translationally to pyrrolidine carboxylic acid (Blombäck *et al.*, 1976). Asn at position 364 of the β-chain binds carbohydrate (Henschen *et al.*, 1980). A variant present in about 15% of the molecules has a γ-chain with an extended carboxyl terminus due to alternate splicing of a single gene and having a molecular weight of about 2000 daltons greater (about 15 amino acids) than the normal γ-chain (Wolfenstein-Todel & Mosesson, 1980). (13) *Golfingia gouldii* and *Dendrostomum pyroides.* (14) The minor component sequence has glu instead of gln at position 63, asp instead of glu at position 78, asn instead of his at position 82 and ala instead of ser at position 96 (Klipenstein, 1972) with a predicted pI value of 5·80. The hemerythrin of another sipunculid, *Themiste zostericola* (Ferrell & Kitto, 1971), has a predicted pI of 6·33. (15) Ser at position 283 of the α chain is phosphorylated at a rate of approximately 10% (Mak *et al.*, 1978). (16) Variant A has asp instead of gly at position 64 and val instead of ala at position 118 (Braunitzer *et al.*, 1973) with a predicted pI value of 4·56. Variant C has his instead of gln at position 59 in the Jersey breed with a predicted pI value of 4·73. Variant *D* has gln instead of glu at position 45 (Brignon & Ribadeau-Dumas, 1973) with a predicted pI value of 4·71. (17) pI is affected in an unknown fashion by the binding of lys at position 240 to bilirubin *in vitro* and *in vivo* (Jacobsen, 1978). Variants are known to have glu instead of lys at position 396 (Méloun *et al.*, 1975) with a predicted pI value of 4·76. (18) A minor component is known to have asp instead of asn at position 311 (McReynolds *et al.*, 1978) with a predicted pI value of 4·76. (19) *Halobacterium halobium.* (20) Variant A has his instead of pro at position 67 with a predicted pI of 4·59. Variant B has his at position 67 and arg instead of ser at position 122, with a predicted pI value of 4·69. Variant C has lys instead of glu at position 37, his at position 67 and lacks a phosphate group on ser-35 (Grosclaude *et al.*, 1972) with a predicted value of 4·90. Variant A3 has gln instead of his at position 106 (Ribadeau-Dumas *et al.*, 1970) with a predicted pI value of 4·42. Variant E has lys instead of glu at position 36 (Grosclaude *et al.*, 1974) with a predicted pI value of 4·67. (21) A variant is known to have asn instead of thr at position 62, glu instead of asp at position 64, thr instead of met at position 108 and glu instead of gly at position 150 (Kato *et al.*, 1978) with a predicted pI value of 4·58. (22) Variant A has a deletion of 13 residues (Grosclaude *et al.*, 1970*a*) with a predicted pI of 4·25. Variant C has gly instead of glu at position 192 (Grosclaude *et al.*, 1970*b*) with a predicted pI value of 4·28. Variant D has a phosphorylated thr at position 53 (Grosclaude *et al.*, 1972) with a predicted pI of 4·19. (23) A variant is known to have asp instead of asn at position 307 (Sepulveda *et al.*, 1975) with a predicted pI value of 3·69. Another has ser-asp instead of asp-ser at positions 104–105 and glu instead of gln at position 113 (Stepanov *et al.*, 1973) with a predicted pI value of 3·70. (24) A variant is known to have asp instead of asn at position 262 (Sepulveda *et al.*, 1975) with a predicted pI value of 2·86. Another has ser-asp instead of asp-ser at positions 59–60 and glu instead of gln at position 68 (Stepanov *et al.*, 1973) with a predicted pI value of 2·87.

correlation coefficients between observed pI and pI_A and pI_B are 0·938 and 0·959, respectively. In addition, we see from Fig. 2 that pI_A exhibits an almost perfect linear relationship with observed pI. The same applies to pI_B. The respective regression functions (with the standard errors in parenthesis) are

$$pI = 0{\cdot}900(\pm 0{\cdot}066) \times pI_A + 0{\cdot}316(\pm 0{\cdot}502)$$

and

$$pI = 0{\cdot}885(\pm 0{\cdot}051) \times pI_B + 0{\cdot}444(\pm 0{\cdot}389)$$

where the expectations are $pI = pI_A = pI_B$. For both pI_A and pI_B, the slope and the intercept of the regression functions were not different significantly from unity and zero, respectively. pI_A explains about 88% of the variation in pI, and pI_B explains about 92%. The difference is not statistically significant. One can, therefore, predict

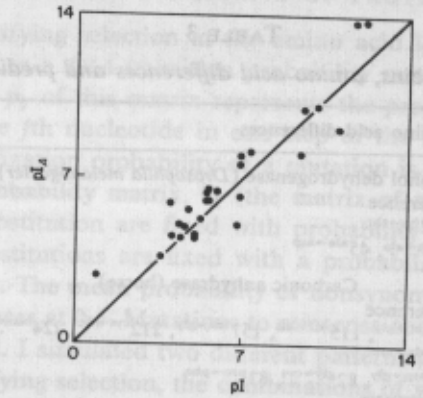

FIG. 2. Relationship between experimental values of pI and pI_A.

the pI of a protein from knowledge of its primary and quaternary structures with a rather high accuracy. Use of more detailed biophysical data, which are mostly unavailable, does not improve the accuracy significantly. The model works better with intermediate values of pI than with extreme values (Fig. 2).

(D) MINIMUM DETECTABLE DIFFERENCE IN pI

ΔpI_{min} is defined as the minimum difference in pI values between two proteins above which their electrophoretic bands are distinguishable on a gel. In this section I am interested in finding the value of ΔpI_{min} for the electrofocusing method. Obviously, ΔpI_{min} is expected to vary with different laboratory conditions (e.g. the density of the gel (Johnson, 1979)). I shall thus be interested in obtaining a range of values and not a fixed number.

The data available for answering this question are scarce. The following amino acid sequences were used: three alleles of alcohol dehydrogenase (EC 1.1.1.1) in *Drosophila melanogaster*, six alleles of equine carbonic anhydrase I (EC 4.2.1.1), six variants of apolipoprotein A-I (apoA-I), six variants of apolipoprotein E (apoE) and 13 cetacean myoglobins. The alcohol dehydrogenase alleles for which we have data on electrofocusing mobility and amino acid sequences are slow, fast and ultra-fast, abbreviated as Adh^S, Adh^F and Adh^{UF}, respectively (Thatcher, 1980; Kreitman, 1983). The carbonic anhydrase I alleles are D, D′, T, A1, A2 and B (Jabusch *et al.*, 1980). The data for the apolipoproteins were taken from Sprecher *et al.* (1984). The nomenclature for the apoE variants is very confused (Zannis *et al.*, 1982) such that apoE-2, for instance, refers to three different variants. In Table 3, I arbitrarily marked the apoE variants of known primary structure by consecutive subscripts in Arabic numerals. The myoglobin sequences were taken from Edmundson (1965), Romero-Herrera & Lehmann (1971*a*, *b*, 1972, 1974) and McLellan (1984). We do not, however, have data on electrofocusing for the myoglobins, and in the following I shall use the data obtained by McLellan (1984) at pH = 9·6, which is close to the predicted pI of myoglobins (9·04–9·37). Table 3 shows the variant

TABLE 3

Variant proteins, amino acid differences and predicted pI's

Variant	Amino acid differences	Predicted pI
	Alcohol dehydrogenase (*Drosophila melanogaster*)	
Adh^S	reference	8·75
Adh^F	$192^{lys \rightarrow thr}$	8·21
Adh^{UF}	$8^{asn \rightarrow ala}$, $45^{ala \rightarrow asp}$	7·31
	Carbonic anhydrase (horse)	
D	reference	6·62
D′	$65^{ser \rightarrow gly}$, $115^{ser \rightarrow his}$, $157^{leu \rightarrow gly}$, $212^{cys \rightarrow tyr}$, $224^{ser \rightarrow ala}$	6·66
T	$54^{ala \rightarrow asp}$	6·39
A_2	$81^{asp \rightarrow gly}$, $82^{gly \rightarrow cys}$, $83^{pro \rightarrow phe}$	6·93
A_1	$183^{ser \rightarrow arg}$	6·95
B	$183^{ser \rightarrow arg}$, $222^{gln \rightarrow arg}$	7·63
	Apolipoprotein A-I (human)	
$apoA\text{-}I_0$	reference	5·60
$apoA\text{-}I_{Milano}$	$173^{arg \rightarrow cys}$	4·65
$apoA\text{-}I_{Munster\text{-}2}$	107^{lys} deletion	4·60
$apoA\text{-}I_{Marburg}$	107^{lys} deletion	4·60
$apoA\text{-}I_{Giessen}$	$143^{pro \rightarrow arg}$	6·60
$apoA\text{-}I_{Munster\text{-}3}$	$3^{pro \rightarrow his}$, $4^{pro \rightarrow arg}$, $103^{asp \rightarrow asn}$	6·71
	Apolipoprotein E (human)	
$apoE_0$	reference	5·91
$apoE_1$	$112^{cys \rightarrow arg}$	6·36
$apoE_2$	$158^{arg \rightarrow cys}$	5·78
$apoE_3$	$145^{arg \rightarrow cys}$	5·78
$apoE_4$	$146^{lys \rightarrow gln}$	5·70
$apoE_5$	$127^{gly \rightarrow asp}$, $158^{arg \rightarrow cys}$	4·78

proteins, their amino acid differences and their predicted pI's. The myoglobin data appears in McLellan (1984). Some of these variants are easily distinguishable by electrofocusing, and some are electrophoretically indistinguishable. The indistinguishable proteins are connected by braces. All the cetacean myoglobins are indistinguishable (predicted pI values ranging from 9·25 to 9·37) except that of the finback whale, *Ballaenoptera physalus* (pI = 9·04), which forms a distinct band.

I used as the maximum estimate of ΔpI_{min} the smallest difference between the predicted pI's of any of the electrophoretically distinguishable allozymes, and as minimum estimate the largest difference in pI values between any of the six sets of indistinguishable proteins. From Table 3, we see that differences smaller than 0·1 pI units are not detected by electrofocusing, while differences above 0·2 pI units are. Thus, ΔpI_{min} ranges between 0·1 and 0·2.

Evolutionary Change of Electrophoretic Properties of Proteins

(A) COMPUTER SIMULATION OF THE EVOLUTION OF NUCLEOTIDE AND PROTEIN SEQUENCES

For simulating the evolutionary change of protein sequences I used Gojobori's (1983) stochastic model, considering different patterns of point mutation at the

nucleotide level and purifying selection at the amino acid level. The mutational changes occur according to a 4×4 transition probability matrix, **P** (the matrix of mutations). The element p_{ij} of this matrix represents the probability that the i-th nucleotide mutates to the jth nucleotide in one step of the simulation (one unit evolutionary time). The fixation probability of a mutation is defined by a 20×20 amino acid exchange probability matrix, **Q** (the matrix of selection). Mutations which result in silent substitution are fixed with probability 1. Mutations which result in amino acid substitutions are fixed with a probability, q_{ij}, given by the elements of the **Q** matrix. The mean probability of nonsynonymous substitutions, $\bar{q}_{ij}$, has been fixed in all cases at 0·5. Mutations to nonsense codons are given a zero probability of being fixed. I simulated two different patterns of mutation and two different patterns of purifying selection, the combinations of which resulted in the following four nucleotide substitution schemes:

Scheme I: Unequal mutation rates and varying purifying selection. I used the inferred estimated pattern of mutation derived from pseudogenes in Table 2 of Li *et al.* (1984). Mutations to nonsynonymous codons were fixed with probabilities defined by an amino acid substitution matrix whose individual elements are negatively correlated with Grantham's (1974) chemical distances (Gojobori *et al.*, 1982; Graur, 1985*a*, *c*). The matrices for the unequal mutation rates and varying purifying selection are given in Tables 4 and 5, respectively.

TABLE 4

P *matrix (estimated pattern of relative mutation frequencies) used in simulation schemes I and II*

	A	T	C	G
A	—	0·046	0·042	0·088
T	0·046	—	0·088	0·042
C	0·068	0·208	—	0·048
G	0·208	0·068	0·048	—

Note: The different probabilities of a nucleotide mutating were taken into account in constructing the seed sequences.

Scheme II: Unequal mutation rates and constant purifying selection. The pattern of mutation is the same as in scheme I. The constant pattern of purifying selection means that all nonsynonymous substitutions are given equal probabilities, i.e. 0·5.

Scheme III: Equal mutation rates and varying purifying selection. All elements of the matrix of mutation rates are given equal probabilities of occurrence. The pattern of purifying selection is the same as in scheme I.

Scheme IV: Equal mutation rates and constant purifying selection. The pattern of mutation is the same as that of scheme III. The pattern of purifying selection is the same as that of scheme II.

I consider scheme I to be the most realistic one, and in all subsequent analyses this scheme will be used primarily.

TABLE 5

Q matrix (probabilities of fixation of amino acid substitutions used in simulation schemes I and III)

	Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp
Ser	0·488	0·326	0·656	0·730	0·540	0·424	0·740	0·340	0·279	0·330	0·479	0·586	0·684	0·786	0·437	0·697	0·628	0·372	0·177
Arg		0·526	0·521	0·670	0·479	0·553	0·419	0·549	0·549	0·642	0·163	0·865	0·800	0·600	0·879	0·553	0·749	0·577	0·530
Leu			0·544	0·572	0·553	0·851	0·358	0·977	0·898	0·833	0·079	0·540	0·474	0·288	0·502	0·200	0·358	0·930	0·716
Pro				0·823	0·874	0·684	0·805	0·588	0·470	0·488	0·214	0·642	0·647	0·577	0·521	0·498	0·567	0·595	0·316
Thr					0·730	0·679	0·726	0·586	0·521	0·572	0·307	0·781	0·805	0·698	0·637	0·605	0·698	0·623	0·405
Ala						0·702	0·721	0·563	0·474	0·479	0·093	0·600	0·577	0·484	0·507	0·414	0·502	0·609	0·312
Val							0·493	0·865	0·767	0·744	0·107	0·609	0·533	0·381	0·549	0·293	0·437	0·902	0·591
Gly								0·372	0·288	0·316	0·260	0·544	0·595	0·628	0·409	0·563	0·544	0·409	0·144
Ile									0·902	0·847	0·079	0·563	0·493	0·307	0·526	0·219	0·377	0·953	0·716
Phe										0·898	0·047	0·535	0·460	0·265	0·526	0·177	0·349	0·870	0·814
Tyr											0·098	0·614	0·540	0·335	0·605	0·256	0·433	0·833	0·828
Cys												0·191	0·284	0·353	0·060	0·284	0·209	0·088	0·000
His													0·888	0·684	0·851	0·623	0·814	0·595	0·465
Gln														0·786	0·753	0·716	0·865	0·530	0·395
Asn															0·563	0·893	0·805	0·340	0·191
Lys																0·530	0·740	0·558	0·488
Asp																	0·791	0·256	0·158
Glu																		0·414	0·293
Met																			0·688

Seed sequences were generated as follows. First, I computed the expected frequencies of nucleotides in functional genes in equilibrium (Gojobori *et al.*, 1982). According to these frequencies I generated 10000 sequences for each of two lengths of proteins: 20 and 100 amino acids, and calculated their predicted pI's. Subsequently, five sequences were chosen at random for each of the following categories of protein acidity: (1) very acidic (VA, pI = 2·5–3·0), (2) acidic (A, 3·5–4·0), (3) moderately acidic (MA, 4·5–6·0), (4) neutral (N, 6·5–7·5), (5) moderately basic (MB, 7·5–8·5), (6) basic (B, 10·5–11·0) and (7) very basic (VB, 12·5–12·5) proteins.

Base changes were introduced one at a time, randomly distributed spatially. Fixations were counted, and the simulation was terminated when a predetermined number of fixations has occurred. Depending on the purpose, I could fix either nucleotide or amino acid substitutions. In this way 20 derived sequences were generated from each of the five seed sequences for each category of acidity and degree of divergence. This process was repeated five times. Derived nucleotide sequences were translated into amino acid sequences by using the "universal" genetic code, and their predicted pI was computed.

(B) EVOLUTIONARY CHANGE OF pI

We first note that in the present simulation no selection for pI is imposed. From Fig. 3 we see that under this condition proteins evolve toward a mildly basic value

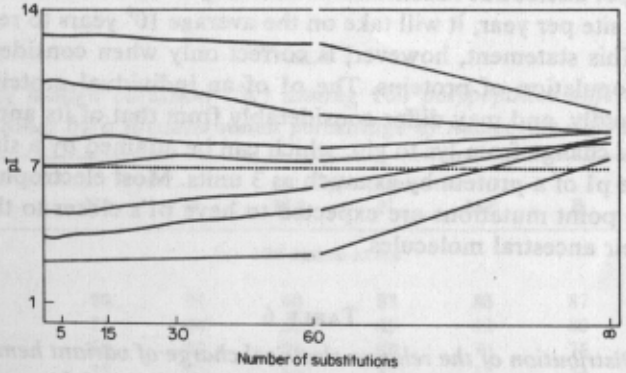

FIG. 3. Change in pI with nucleotide substitutions for proteins of length 100 amino acids with different initial pI's.

of pI. This means that stringent selection is required to maintain pI values above or below the range of 7–9. This is compatible with the observation that proteins which maintain extreme pI's for purposes of function (e.g. the very basic histones which bind deoxyribonucleic acids) evolve very slowly. In histones, mutations will almost always lower the pI, and, thus, most mutations will be selected against, resulting in a low substitution rate (for other reasons see Graur, 1985*c*). This is also true with proteins having many charged molecules (e.g. ubiquitin, which maintains both very acidic and very basic independent domains although its overall charge

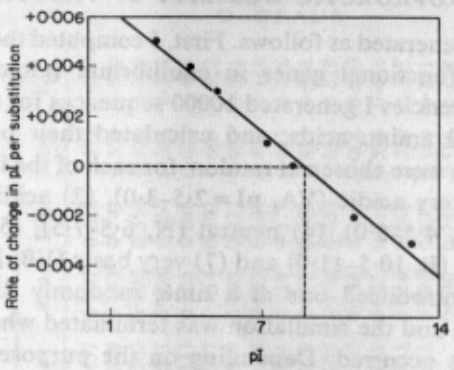

FIG. 4. Rate of change in pI per nucleotide substitution for proteins of length 100 with different pI's At equilibrium (rate of change = 0) the pI is 8·6.

is neutral). As seen from Fig. 4, the rate of change in pI is different for proteins with different initial pI's. VA and VB proteins are expected to experience a much more rapid change in pI during evolutionary times than proteins closer to the equilibrium pI. From Fig. 4, we estimate that the mean pI value at equilibrium (when the rate of change in pI equals 0) is 8·7.

From Figs 3 and 4, we see that the evolutionary change in pI is very slow. For example, a VA protein of length 100 will experience a positive change of about 0·005 units pI per nucleotide substitution. Assuming a rate of nucleotide substitution of 2×10^{-9} per site per year, it will take on the average 10^9 years to record a change of 1 unit pI. This statement, however, is correct only when considering the mean change of a population of proteins. The pI of an individual protein may change much more rapidly, and may differ considerably from that of its ancestral protein. For example, a change from lys to glu, which can be attained by a single mutation, can change the pI of a protein by as much as 3 units. Most electrophoretic variants resulting from point mutations are expected to have pI's closer to the equilibrium range than their ancestral molecules.

TABLE 6

Distribution of the relative electrical charge of variant hemoglobins in comparison with standard hemoglobin

Chain	0	+	++	−	−−
hemoglobin α	11	37	3	23	2
hemoglobin β	48	54	5	42	6
hemoglobin γ	3	7	3	0	1
hemoglobin δ	0	4	1	3	0
Total	62	102	12	68	9

0 = no charge change; +, − = a charge change of less than 2 units pI; ++, −− = a charge change of more than 2 units pI; +, ++ = positive charge changes; −, −− = negative charge changes

That this is the case in actuality can be demonstrated with human hemoglobin variants (Lehmann & Kynoch, 1976). The most common amino acid sequence for human hemoglobin is assumed to be the ancestor, since it is identical to that of chimpanzee. All variants are assumed to be newer mutants. Hemoglobin has a pI of 6·98, lower than the equilibrium value. Hence, I predict that mutations resulting in a more basic protein will outnumber those resulting in a more acidic protein. The data are shown in Table 6. In accordance with the theoretical expectations, we find a significant excess of positive changes ($\chi^2 = 7{\cdot}18$, $0{\cdot}05 < P < 0{\cdot}01$). (The number of mutants resulting in no electric charge change is most probably a gross underestimate due to biases in assessment which, in cases of no clinical symptoms, heavily favor electrophoretically detectable mutants.)

(C) AMOUNT OF ELECTROPHORETICALLY HIDDEN VARIATION

In this section I computed the fraction of derived sequences that are different from the seed sequence by a fixed number of nucleotides, and are indistinguishable from each other by their pI's, i.e. the percentage of sequence variation that is undetectable by the electrofocusing method. I call this value the amount of hidden variation. This definition differs from that of Nei & Chakraborty (1976) and Chakraborty & Nei (1976). We see from Table 7 that, as expected, the amount of hidden variation decreases as the degree of nucleotide divergence increases. Nevertheless, the amount of hidden variation remains considerable even for large degrees of divergence.

TABLE 7

Amount of hidden variation (%) among 100 polypeptides differing from each other by a predetermined percentage of nucleotide difference

% divergence	Type of protein: VA	A	MA	N	MB	B	VB
	(a) 100 amino acids						
2	89	91	90	87	88	87	87
6	71	84	85	85	83	80	83
10	68	82	79	80	74	76	82
30	38	63	66	61	55	54	63
60	21	50	48	50	30	43	55
	(b) 20 amino acids						
10	68	73	81	70	68	69	80
30	48	45	67	54	48	43	67
60	33	30	49	38	37	25	46

Note: Simulation scheme I.

Let us now check these results against experimental data. Unfortunately, there are no electrofocusing data in the literature that would confirm or contradict my results. However, one can get an idea on whether the theoretical expectations are

reasonable or not from the experiments of Fuerst & Ferrell (1980) on fixed pH electrophoresis. In their Fig. 1, we see that out of the 24 mammalian hemoglobins only 6 classes are readily distinguishable. This means that the amount of hidden variation is 75%. The degree of amino acid divergence between the species is about 7% (10–12% nucleotide differences). From Table 7 we see that the expected amount of hidden variation is around 74%. There is good agreement between the simulation and the experimental results.

As mentioned above, it has been calculated that under equal mutation rates and no selection about 25–30% of all possible base substitutions will result in a detectable change in electrophoretic mobility. With my model I was able to estimate the percentage of nucleotide and amino acid substitutions that will result in detectable electromorphs under specific conditions of mutation and purifying selection. For each seed sequence I generated mutants differing from the seed sequence by either one nucleotide (column a in Table 8) or one amino acid (column b). I then computed the percentage of mutants of which the pI value is different from that of the seed sequence. The results are presented in Table 8.

The most striking observation from Table 8 is that detectability depends on the length of the protein, such that in short polypeptides most nucleotide substitutions will be detected, while in long polypeptides most nucleotide substitutions will remain undetected. For proteins of 100 amino acids, one is expected to detect 3–12% of all single nucleotide mutants depending on the initial pI of the ancestral molecule. For short proteins of 20 amino acids we are expected to detect 21–71% of the variants. Since polypeptides used in electrophoretic studies are, mostly, much longer (about 400 amino acids on the average, (Nei, 1975)), I conclude that 25–30% detectability is an overestimate.

Koehn & Eanes (1977, 1978) and Nei *et al.* (1978) found a positive correlation between the subunit molecular weight of a protein and heterozygosity. Under the mutation-drift hypothesis these quantities are indeed expected to be correlated since a large molecule will generally sustain a higher mutation rate than a small one. The correlation found by Nei *et al.* (1978), however, was not very high, although it was statistically significant. These authors derived the expected correlation between the two quantities, and showed that, because of the nonlinear relationship between mutation rate and expected heterozygosity, the expected correlation is usually much less than 1. Nevertheless, even when this factor was taken into account there remained a discrepancy between expected and observed correlations. The authors speculated that the discrepancy is caused by the incomplete correlation between mutation rate and molecular weight.

The present study shows that the expected substitution rate for electrophoretically detectable alleles is indeed affected by the length of the protein (approximately equivalent to molecular weight), and, consequently, the expected heterozygosity will not be linear with the mutation rate. The rationale is as follows. Average heterozygosity (H) is given by

$$H = 4N_e\nu_n/1+4N_e\nu_n$$

(Kimura & Crow, 1964) where N_e is the effective population size and ν_n is the

TABLE 8

Percentage of detectable electromorphs per nucleotide substitution (a) and per amino acid substitution (b)

Type of protein	Detectable variants (a)	Detectable variants (b)	Type of protein	Detectable variants (a)	Detectable variants (b)
		Substitution Scheme I			
length 100			length 20		
VA	12·4 (500)	17·0 (100)	VA	71·0 (100)	81·0 (100)
A	3·2 (500)	8·8 (400)	A	32·5 (200)	58·5 (200)
MA	3·6 (500)	6·1 (700)	MA	21·0 (100)	37·0 (200)
N	4·6 (500)	8·4 (320)	N	34·0 (100)	42·0 (100)
MB	5·8 (500)	14·4 (180)	MB	34·0 (100)	52·0 (100)
B	7·0 (500)	14·2 (240)	B	42·0 (100)	61·0 (100)
VB	8·6 (500)	11·1 (180)	VB	42·0 (100)	66·0 (100)
		Substitution Scheme II			
length 100			length 20		
VA	13·8 (500)	21·0 (100)	VA	63·8 (80)	79·0 (100)
A	4·0 (500)	7·8 (600)	A	48·0 (200)	62·5 (160)
MA	4·0 (500)	6·5 (600)	MA	11·0 (100)	36·3 (160)
N	6·0 (500)	9·7 (300)	N	35·0 (100)	53·8 (80)
MB	6·8 (500)	10·0 (240)	MB	34·0 (100)	46·3 (80)
B	7·0 (500)	17·1 (140)	B	40·0 (100)	55·0 (80)
VB	6·8 (500)	16·4 (140)	VB	28·0 (100)	70·0 (80)
		Substitution Scheme III			
length 100			length 20		
VA	13·8 (480)	18·0 (100)	VA	65·0 (80)	66·0 (100)
A	5·4 (500)	8·0 (400)	A	54·5 (200)	68·8 (160)
MA	4·0 (500)	6·4 (700)	MA	21·0 (100)	32·1 (240)
N	3·4 (500)	8·6 (500)	N	26·0 (100)	40·0 (80)
MB	8·2 (500)	14·0 (100)	MB	30·0 (100)	42·5 (80)
B	7·2 (500)	13·5 (200)	B	33·0 (100)	61·3 (80)
VB	6·4 (500)	13·5 (200)	VB	42·0 (100)	57·5 (80)
		Substitution Scheme IV			
length 100			length 20		
VA	15·0 (500)	20·0 (100)	VA	66·3 (80)	80·0 (100)
A	4·2 (500)	9·0 (400)	A	46·5 (200)	70·7 (160)
MA	5·0 (500)	7·3 (400)	MA	17·0 (100)	30·0 (260)
N	7·2 (500)	13·0 (300)	N	39·0 (100)	51·3 (80)
MB	7·2 (500)	11·8 (300)	MB	35·0 (100)	49·0 (100)
B	8·2 (500)	12·5 (200)	B	54·0 (100)	73·8 (80)
VB	9·8 (500)	10·5 (200)	VB	43·0 (100)	56·3 (80)

Note: The numbers in parentheses are the sample sizes.

effective substitution rate per n amino acids. Only when $4N_e\nu_n$ is much smaller than 1, is H expected to be approximately linear with ν_n. This, however, should not cause difficulties since for most organisms $4N_e\nu_n$ is indeed less than 1. ν_n is composed of three factors: (1) a constant, the intrinsic substitution rate per amino acid, ν, (2) the number of amino acids, n, and (3) the electrophoretic detectability,

f_n. In other words

$$\nu_n = \nu \times n \times f_n$$

Our results show that f_n decreases as n increases. If the decrease is linear with n, f_n can be expressed as a function of n as follows

$$f_n = k_1 - k_2 n$$

where k_1 and k_2 are constants. We can easily see that in this case ν_n is no longer linear with n, and we will expect a decrease in the correlation coefficient between H and molecular weight. If, on the other hand, the relationship between f_n and n is not linear, and can be expressed, for example, as

$$f_n = k/n$$

then ν_n becomes independent of n, and we will expect no correlation at all between H and molecular weight. The situation in reality is most probably in between these two extreme cases, but no matter what the exact situation is, the dependence of f_n on n means that even for $4N_e\nu_n$ values smaller than 1, H will not be linear with ν_n, and consequently, the correlation between H and molecular weight will, in all cases, be less than 1.

(D) DISTRIBUTION OF VARIANTS THAT ARE DETECTABLE BY THE ELECTROFOCUSING METHOD AND GENE DIVERSITY

Previously I have determined the minimum detectable difference in pI. Let us now consider how much protein variation is detectable with this resolving power. For this purpose, I first consider the distributions of pI's for mutant proteins for various initial pI's. The distributions are shown in Figs. 5(a) and 5(b). In these distributions pI is shown with an interval of 0·1. The number of variant classes is larger in reality than that shown. It is noted, however, that the number of classes is quite limited. The distribution of variants does not follow the strict premises of the stepwise model (Ohta & Kimura, 1973). Rather, in accordance with experimental observations (Ramshaw *et al.*, 1979; Fuerst & Ferrell, 1980; McCommas, 1983), the distributions depicted in Figs 5(a) and 5(b) indicate that the infinite-allele model (Kimura & Crow, 1964) is more appropriate for electrophoretic data.

Flake & Lennington (1977) and Brown *et al.* (1981) have suggested that the average difference in mobility can be used as a measure of taxonomic distance. According to our results, however, the difference in pI between the electromorphs correlates well with the distance in terms of amino acid substitutions only for VA and VB proteins. For the majority of proteins, however, I obtain low correlation coefficients between the average absolute difference in pI of the ancestral and derived sequences, and the number of amino acid substitutions (taxonomic divergence). This is understandable, since on the one hand, there can be two proteins that are similar in amino acid sequence, and yet have very different pI's, and on the other hand, there can be proteins which are different in terms of amino acid sequence, and yet have approximately the same pI. There are other compelling reasons to

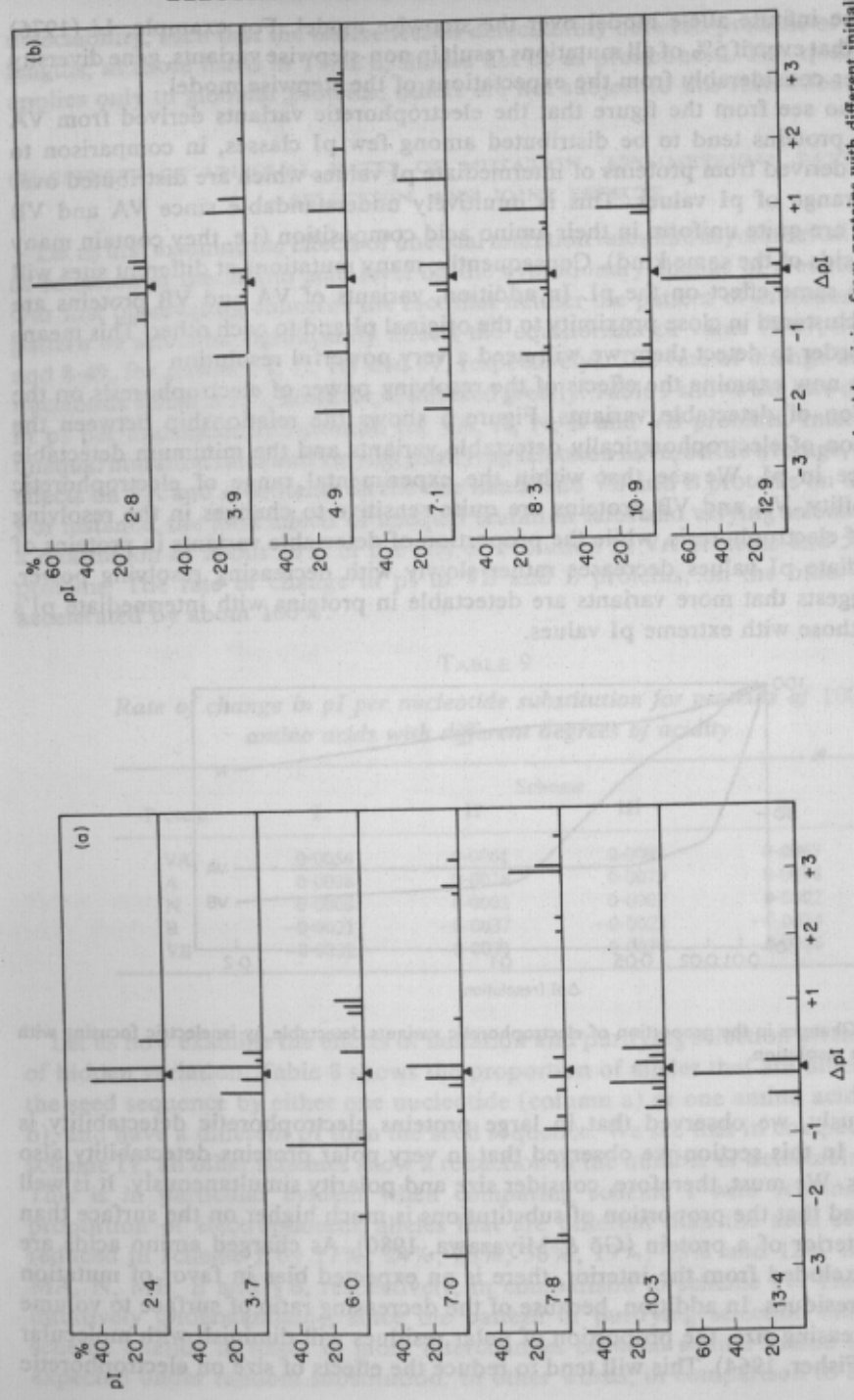

FIG. 5. Distribution of isoelectric variants which are different than their ancestral sequence by one nucleotide arising from proteins with different initial pI's. (a) proteins of length 20 amino acids. (b) proteins of length 100 amino acids.

prefer the infinite allele model over the stepwise model. For example, Li (1976) showed that even if 5% of all mutations result in non-stepwise variants, gene diversity will differ considerably from the expectations of the stepwise model.

We also see from the figure that the electrophoretic variants derived from VA and VB proteins tend to be distributed among few pI classes, in comparison to variants derived from proteins of intermediate pI values which are distributed over a wide range of pI values. This is intuitively understandable since VA and VB proteins are quite uniform in their amino acid composition (i.e. they contain many amino acids of the same kind). Consequently, many mutations at different sites will have the same effect on the pI. In addition, variants of VA and VB proteins are usually clustered in close proximity to the original pI and to each other. This means that in order to detect them we will need a very powerful resolution

Let us now examine the effects of the resolving power of electrophoresis on the proportion of detectable variants. Figure 6 shows the relationship between the proportion of electrophoretically detectable variants and the minimum detectable difference in pI. We see that within the experimental range of electrophoretic detectability, VA and VB proteins are quite sensitive to changes in the resolving power of electrophoresis, while the proportion of detectable variants in proteins of intermediate pI values decreases rather slowly with decreasing resolving power. This suggests that more variants are detectable in proteins with intermediate pI's than in those with extreme pI values.

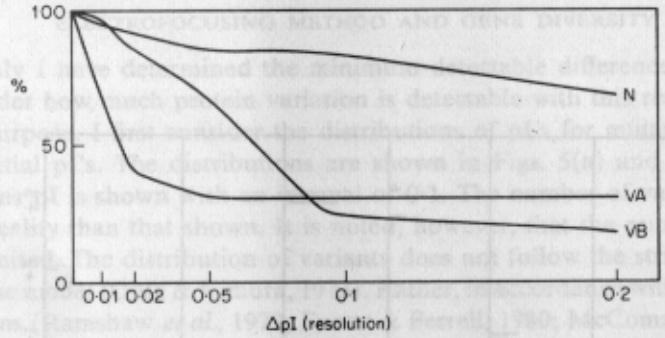

FIG. 6. Changes in the proportion of electrophoretic variants detectable by isoelectric focusing with decrease in resolution.

Previously, we observed that in large proteins electrophoretic detectability is reduced. In this section we observed that in very polar proteins detectability also decreases. We must, therefore, consider size and polarity simultaneously. It is well established that the proportion of substitutions is much higher on the surface than in the interior of a protein (Gō & Miyazawa, 1980). As charged amino acids are mostly excluded from the interior, there is an expected bias in favor of mutation in polar residues. In addition, because of the decreasing ratio of surface to volume with increasing size, the proportion of polar residues will diminish with molecular weight (Fisher, 1964). This will tend to reduce the effects of size on electrophoretic

detectability, such that the differences in detectability between proteins of different lengths, as those listed in Table 8, should not be as pronounced. This qualification applies only to globular proteins; others are not subject to this restriction.

(E) EFFECTS OF UNEQUAL RATES OF MUTATION, ASYMMETRICAL PATTERN OF SELECTION, AND JOINT EFFECTS

Let us first examine the effects of unequal mutation rates and asymmetrical patterns of selection at the amino acid level on the evolutionary change of electric charge. Our first observation concerns the fact that neither the pattern of mutation nor the pattern of selection significantly affects the equilibrium pI value (8·67, 8·42, 8·92 and 8·49, for schemes I, II, III and IV, respectively). The rate of change in pI with nucleotide substitution, however, is affected greatly. Table 9 shows the rate of change in pI per nucleotide substitution for VA, A, N, B and VB proteins. Interestingly, unequal mutation rates and varying purifying selection have, on the average, opposite effects on VA and A proteins on the one hand, and VB and B proteins on the other. For instance, the joint effects of unequal mutation rates and varying selection cause a retardation of about 20% in the rate of evolution in VA proteins and 50% in A proteins. The rate of change in pI in VB and B proteins, on the other hand, is accelerated by about 100%.

TABLE 9

Rate of change in pI per nucleotide substitution for proteins of 100 amino acids with different degrees of acidity

Protein	Scheme I	II	III	IV
VA	0·0054	0·0061	0·0081	0·0067
A	0·0038	0·0028	0·0075	0·0076
N	0·0008	0·0003	0·0002	0·0002
B	−0·0021	−0·0037	−0·0023	−0·0034
VB	−0·0032	−0·0031	−0·0030	−0·0016

Let us now examine the effects of mutation and purifying selection on the amount of hidden variation. Table 8 shows the proportion of alleles that are different from the seed sequence by either one nucleotide (column a) or one amino acid (column b), and have a different pI than the seed sequence. We see that in comparison with scheme IV, all other schemes show a reduction in the number of detectable variants. This is in particular evident when comparing scheme I with scheme IV. The proportion of electrophoretic alleles that are different than the seed sequence is reduced in scheme I by 17%, 24%, 28%, 36%, 19%, 15% and 12% for VA, A, MA, N, MB, B and VB, respectively, in comparison to scheme IV. This result is intuitively understandable, since the pattern of purifying selection employed in scheme I tends to result in more interchanges between similar amino acids than expected under random substitution. In other words, in comparison to a situation

where no directional purifying selection operates (scheme IV), the more realistic simulation (scheme I) results in many of the variants having similar electric charges as their ancestral molecule.

The amount of detectable variation also decreases in scheme I in comparison to the other schemes due to the fact that many of the variants are more similar to each other than randomly expected. This will reduce the robustness to changing resolutions. Table 10 shows a typical example of the reduction in the number of detectable variants from a perfect resolution ($\Delta pI_{min} = 0{\cdot}0$) to a resolution of 0·2 units pI, under the four substitution schemes. We can see that under the realistic scheme (scheme I) the number of alleles is very sensitive to the resolution power of the method. The same applies to schemes II and III. Only when both mutation and purifying selection are random (scheme IV), are the number of alleles less sensitive to the resolving power.

TABLE 10

Percent detectability with reduction in resolution power of a protein of length 100 *subjected to different patterns of mutation and selection*

	Resolution (ΔpI_{min})				
Scheme	0·01	0·02	0·05	0·10	0·20
I	71	47	28	23	18
II	74	44	28	24	19
III	65	46	30	24	16
IV	90	74	43	28	13

Comparison Between Electrophoretic Methods

The purpose of this section is to calculate the amount of variation at the nucleotide level one can expect to detect by using different electrophoretic methods. The methods considered are:

(1) Electrophoresis at a randomly chosen constant pH.
(2) Electrophoresis at the best performing (optimal) constant pH.
(3) Parallel (sequential) electrophoresis at five different constant pH's (i.e. 3, 5, 7, 9 and 11)
(4) Isoelectric focusing.

I generated populations of proteins differing from each other by a fixed number of nucleotide substitutions from protein sequences of different initial pI's, and then, calculated the charge of each of the proteins in each population at five different values of pH, and their isoelectric point. For each population of proteins and for each method, I computed the value of the single locus heterozygosity (h) as in Nei (1975, p. 131, 1978).

The results are given in Table 11. As expected the lowest values of h, and, hence, the poorest performance, were obtained by using method 1. The results are better when using method 2, namely by choosing the best performing single pH. Since,

TABLE 11

Comparison of the efficiency of four electrophoretic methods in detecting genetic variability. The parameter compared is h (see text)

Length	Acidity	% Divergence	Random pH	Optimal pH	Parallel electrophoresis	Isoelectric focusing
100	VA	2	0·223±0·027	0·247±0·030	0·248±0·031	0·179±0·054
		6	0·383±0·048	0·465±0·048	0·508±0·042	0·518±0·055
		10	0·632±0·033	0·727±0·037	0·781±0·045	0·682±0·067
		30	0·792±0·017	0·863±0·020	0·904±0·016	0·889±0·005
	A	2	0·079±0·031	0·094±0·029	0·112±0·034	0·076±0·019
		6	0·186±0·051	0·199±0·040	0·232±0·056	0·201±0·033
		10	0·319±0·032	0·393±0·064	0·446±0·047	0·330±0·088
		30	0·611±0·029	0·684±0·031	0·732±0·040	0·652±0·051
	MA	2	0·067±0·028	0·091±0·048	0·109±0·052	0·093±0·041
		6	0·126±0·034	0·147±0·042	0·184±0·028	0·182±0·049
		10	0·301±0·044	0·314±0·040	0·420±0·038	0·298±0·051
		30	0·492±0·072	0·553±0·068	0·632±0·080	0·577±0·042
	N	2	0·081±0·044	0·111±0·044	0·111±0·044	0·149±0·022
		6	0·186±0·029	0·214±0·051	0·248±0·054	0·180±0·062
		10	0·336±0·033	0·356±0·029	0·404±0·049	0·337±0·062
		30	0·714±0·031	0·748±0·041	0·840±0·016	0·684±0·022
	MB	2	0·101±0·057	0·145±0·059	0·145±0·059	0·147±0·034
		6	0·290±0·067	0·308±0·062	0·371±0·069	0·250±0·031
		10	0·427±0·037	0·459±0·035	0·544±0·032	0·431±0·034
		30	0·724±0·021	0·771±0·023	0·851±0·026	0·711±0·023
	B	2	0·123±0·033	0·130±0·035	0·130±0·035	0·160±0·071
		6	0·345±0·066	0·367±0·077	0·386±0·069	0·280±0·061
		10	0·394±0·029	0·416±0·039	0·437±0·034	0·373±0·050
		30	0·667±0·032	0·712±0·024	0·770±0·020	0·771±0·027
	VB	2	0·125±0·039	0·145±0·053	0·163±0·052	0·181±0·039
		6	0·407±0·075	0·487±0·082	0·502±0·088	0·487±0·063
		10	0·494±0·046	0·526±0·049	0·571±0·053	0·524±0·055
		30	0·800±0·027	0·845±0·031	0·893±0·028	0·778±0·021
20	VA	10	0·597±0·014	0·714±0·016	0·768±0·010	0·736±0·036
		30	0·796±0·021	0·878±0·024	0·914±0·015	0·856±0·016
	A	10	0·464±0·053	0·533±0·057	0·567±0·056	0·546±0·056
		30	0·827±0·010	0·878±0·014	0·930±0·011	0·872±0·009
	MA	10	0·334±0·078	0·366±0·067	0·409±0·090	0·353±0·032
		30	0·693±0·023	0·741±0·034	0·806±0·034	0·635±0·092
	N	10	0·296±0·075	0·383±0·095	0·404±0·095	0·532±0·036
		30	0·630±0·048	0·714±0·042	0·762±0·041	0·775±0·018
	MB	10	0·379±0·036	0·413±0·061	0·491±0·019	0·539±0·022
		30	0·677±0·019	0·718±0·021	0·806±0·022	0·790±0·028
	B	10	0·545±0·044	0·589±0·052	0·640±0·036	0·611±0·041
		30	0·800±0·019	0·833±0·022	0·892±0·017	0·868±0·014
	VB	10	0·621±0·025	0·737±0·030	0·753±0·037	0·580±0·016
		30	0·776±0·007	0·864±0·016	0·895±0·019	0·742±0·030

however, this procedure requires running gels at different pH's, we might as well consider the data together (method 3). Sequential electrophoresis (method 3) proved to be the best method. It outperformed isoelectric focusing (method 4) in 33 out of the 42 cases listed in Table 11. Notwithstanding, if the degree of divergence among the proteins in a population is low, method 4 performs about as well as method 3.

There are, unfortunately, almost no studies that can be used to examine the agreement between these predictions and empirical observations. The only example that I am aware of is that of Basset *et al.* (1978) which showed that for 70 hemoglobin variants, 31 could not be separated by electrophoresis in a single constant pH, two could not be separated by isoelectric focusing, and all variants were resolved by using parallel electrophoresis, which is how these variants came to light in the first place. This limited information supports the results of the simulation.

I wish to express my gratitude to Drs Masatoshi Nei, David Hewett-Emmett and Klaus Wöhrmann for their criticism and support. This work was aided by research grants NIH GM-20293 and NSF BSF-8315115 to Dr M. Nei and a fellowship from the Alexander von Humbold Foundation to the author.

REFERENCES

ANDO, T. & WATANABE, S. (1969). *Int. J. Protein Res.* **1**, 221.
AYALA, F. J. (1982). *Proc. natn. Acad. Sci. U.S.A.* **79**, 550.
BASSET, P., BEUZARD, Y., GAREL, M. C. & ROSA, J. (1978). *Blood* **51**, 971.
BELL, G. I., SWAIN, W. F., PICTET, R. L., CORDELL, B., GOODMAN, H. M. & RUTTER, W. J. (1979). *Nature* **282**, 525.
BELL, G. I., PICTET, R. L., RUTTER, W. J., CORDELL, B., TISCHER, E. & GOODMAN, H. M. (1980). *Nature* **284**, 26.
BERNSTEIN, S. C., THROCKMORTON, L. H. & HUBBY, J. L. (1973). *Proc. natn. Acad. Sci. U.S.A.* **70**, 3928.
BLOMBÄCK, B., HESSEL, B. & HOGG, D. (1976). *Thromb. Res.* **8**, 639.
BLOW, D. M., BIRKTOFT, J. J. & HARTLEY, B. S. (1969). *Nature* **221**, 337.
BRAUNITZER, G., CHEN, R., SCHRANK, B. & STANGL, A. (1973). *Hoppe-Seyler's Z. physiol. Chem.* **354**, 867.
BRESLOW, E. & GURD, F. R. N. (1962). *J. biol. Chem.* **237**, 371.
BRIGNON, G. & RIBADEAU-DUMAS, B. (1973). *FEBS Lett.* **33**, 73.
BRIGNON, G., RIBADEAU-DUMAS, B., MERCIER, J. C., PELISSIER, J. P. & DAS, B. C. (1977). *FEBS Lett.* **76**, 274.
BROWN, A. D. H., MARSHALL, D. R. & WEIR, B. S. (1981). In: *Genetic Studies in* Drosophila *Populations.* (Gibson, J. B. & Oakeshott, J.G. eds). pp. 15–43. Australia: ANUP.
BROWN, J. R. & HARTLEY, B. S. (1966). *Biochem. J.* **101**, 214.
CANFIELD, R. E., KAMMERMAN, S., SOBEL, J. H. & MORGAN, F. J. (1971). *Nature New Biol.* **232**, 16.
CATTERALL, J. F., STEIN, J. P., KRISTO, P., MEANS, A. R. & O'MALLEY, B. W. (1980). *J. Cell Biol.* **87**, 480.
CHAKRABORTY, R. & NEI, M. (1976). *Genetics* **84**, 385.
CHRAMBACH, A. & RODBARD, D. (1971). *Science* **172**, 440.
COOKE, N. E., COIT, D., SHINE, J., BAXTER, J. D. & MARTIAL, J. A. (1981). *J. biol. Chem.* **256**, 4007.
DENOTO, F. M., MOORE, D. D. & GOODMAN, H. M. (1981). *Nucleic Acid Res.* **9**, 3719.
DOOLITTLE, R. F., TAKAGI, T., WATT, K., BOUMA, H., COTTRELL, B. A., CASSMAN, K. G., GOLDBAUM, D. M., DOOLITTLE, L. R. & FRIEZNER, S. J. (1978). In: *Regulatory Proteolitic Enzymes and their Inhibitors.* (Magnusson, S., Ottesen, M., Foltmann, B., Danø, K. & Neurath, H. eds). pp. 1–118. Oxford: Pergamon Press.
DRIESSEN, H. P. C., HERBRINK, P., BLOEMENDAL, H. & DE JONG, W. W. (1980). *Expl Eye Res.* **31**, 243.
DUNN, R., MCCOY, J., SIMSEK, M., MAJUMDAR, A., CHANG, S. H., RAJBHANDARY, U. L. & KHORANA, H. G. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78**, 6744.
EDMUNDSON, A. B. (1965). *Nature* **205**, 883.
ELLEMAN, T. C. & WILLIAMS, J. (1970). *Biochem. J.* **116**, 515.
FERRELL, R. E. & KITTO, G. B. (1971). *FEBS Lett.* **12**, 322.
FISHER, H. F. (1964). *Proc. natn. Acad. Sci. U.S.A.* **51**, 1285.
FLAKE, R. H. & LENNINGTON, R. K. (1977). *Biol. Zbl.* **96**, 451.
FUERST, P. A. & FERRELL, R. E. (1980). *Genetics* **94**, 185.
FUERST, P. A., CHAKRABORTY, R. & NEI, M. (1977). *Genetics* **86**, 455.

GÅRLUND, B., HESSEL, B., MAGUERIE, G., MURANO, G. & BLOMBÄCK, B. (1977). *Eur. J. Biochem.* **77**, 595.
GŌ, M. & MIYAZAWA, S. (1980). *Int. J. Peptide Protein Res.* **15**, 211.
GOJOBORI, T. (1983) *Genetics* **105**, 1011.
GOJOBORI, T., LI, W. H. & GRAUR, D. (1982). *J. mol. Evol.* **18**, 360.
GRANTHAM, R. (1974) *Science* **185**, 862.
GRAUR, D. (1985*a*). *J. mol. Evol.* **21**, 221.
GRAUR, D. (1985*b*). *Evolution* **39**, 190.
GRAUR, D. (1985*c*). *J. mol. Evol.* **22**, 53.
GROSCLAUDE, F., MAHÉ, M. F., MERCIER, J. C. & RIBADEAU-DUMAS, B. (1970*a*). *FEBS Lett.* **11**, 109.
GROSCLAUDE, F., MERCIER, J. C. & RIBADEAU-DUMAS, B. (1970*b*). *Compt. Rend. Acad. Sci.* (D) **268**, 3133.
GROSCLAUDE, F., MAHÉ, M. F., MERCIER, J. C. & RIBADEAU-DUMAS, B. (1972). *Eur. J. Biochem.* **26**, 328.
GROSCLAUDE, F., MAHÉ, M. F. & VOGLINO, G. F. (1974). *FEBS Lett.* **45**, 3.
HAMRICK, J. L., LINHART, Y. B. & MITTON, J. B. (1979). *Ann. Rev. Ecol. Syst.* **10**, 173.
HENSCHEN, A., LOTTSPEICH, F., SOUTHAN, C. & TÖPFER-PETERSEN, E. (1980). In: *Protides of the Biological Fluids.* (Peeters, H. ed.). pp. 51–56. Oxford: Pergamon Press.
IWAI, K., NAKAHARA, C. & ANDO, T. (1971). *J. Biochem.* **69**, 493.
JABUSCH, J. R., BRAY, R. P. & DEUTSCH, H. F. (1980). *J. biol. Chem.* **255**, 9196.
JACOBSEN, C. (1978). *Biochem J.* **171**, 453.
JELTSCH, J. M. & CHAMBON, P. (1982). *Eur. J. Biochem.* **122**, 291.
JOHNSON, G. B. (1976). *Genetics* **83**, 149.
JOHNSON, G. B. (1977). *Biochem. Genet.* **15**, 665.
JOHNSON, G. B. (1979). *Biochem. Genet.* **17**, 499.
KATO, I., KOHR, W. J. & LASKOWSKI, M. (1978). In: *Regulatory Proteolitic Enzymes and their Inhibitors.* (Magnusson, S., Ottesen, M., Foltmann, B., Danø, K. & Neurath, H., eds). pp. 197–206. Oxford: Pergamon Press.
KIMURA, M. & CROW, J. F. (1964). *Genetics* **49**, 725.
KLIPPENSTEIN, G. L. (1972). *Biochemistry* **11**, 372.
KLIPPENSTEIN, G. L., HOLLEMAN, J. W. & KLOTZ, I. M. (1968). *Biochemistry* **7**, 3868.
KOEHN, R. K. & EANES, W. F. (1977). *Theor. Pop. Biol.* **11**, 330.
KOEHN, R. K. & EANES, W. F. (1978). *Evol. Biol.* **11**, 39.
KREITMAN, M. (1983). *Nature* **304**, 412.
LAEMMLI, U. K. (1970). *Nature* **227**, 680.
LAWN, R. M., EFSTRATIADIS, A., O'CONNEL, C. & MANIATIS, T. (1980). *Cell* **21**, 647.
LAWN, R. M., ADELMAN, J., BOCK, S. C., FRANKE, A. E., HOUCK, C. M., NAJARIAN, R. C., SEBURG, P. H. & WION, K. L. (1981). *Nucleic Acids Res.* **9**, 6103.
LEHMANN, H. & KYNOCH, P. A. M. (1976). *Human Haemoglobin Variants and their Characteristics.* Amsterdam: North-Holland.
LEHNINGER, A. L. (1975). *Biochemistry*, 2nd edn. New York: Worth.
LI, C. H. & DIXON, J. S. (1971). *Arch. Biochem. Biophys.* **146**, 233.
LI, W. H. (1976). *Genetics* **83**, 423.
LI, W. H., WU, C. I. & LUO, C. C. (1984). *J. mol. Evol.* **21**, 58.
LIEBHABER, S. A., GOOSSENS, M. J. & KAN, Y. W. (1980). *Proc. natn. Acad. Sci. U.S.A.* **77**, 7054.
MAHLER, H. R. & CORDES, E. H. (1966). *Biological Chemistry.* New York: Harper & Row.
MAK, A. S., SMILLIE, L. B. & BÁRÁNY, M. (1978). *Proc. Natn. Acad. Sci. U.S.A.* **75**, 3588.
MAK, A. S., SMILLIE, L. B. & STEWART, G. R. (1980). *J. Biol. Chem.* **255**, 3647.
MARSHALL, D. R. & BROWN, A. H. D. (1975). *J. mol. Evol.* **6**, 149.
MATSUBARA, H. & SMITH, E. L. (1962). *J. biol. Chem.* **237**, PC3575.
MATSUBARA, H. & SMITH, E. L. (1963). *J. biol. Chem.* **238**, 2732.
MCCOMMAS, S. A. (1983). *Genetics* **103**, 741.
MCKENZIE, H. A., RALSTON, G. B. & SHAW, D. C. (1972). *Biochemistry* **11**, 4539.
MCLELLAN, T. (1984). *Biochem. Genet.* **22**, 181.
MCREYNOLDS, L., O'MALLEY, B. W., NISBET, A. D., FOTHERGILL, J. E., GIVOL, D., FIELDS, S., ROBERTSON, M. & BROWNLEE, G. G. (1978). *Nature* **273**, 723.
MELOUN, B., KLUH, I., KOSTKA, V., MORÁVEK, L., PRUSÍK, Z., VANĚČĚK, J., KEIL, B. & SŌRM, F. (1966). *Biochim. biophys. Acta* **130**, 543,
MELOUN, B., MORÁVEK, L. & KOSTKA, V. (1975). *FEBS Lett.* **58**, 134.
MERCIER, J. C., GROSCLAUDE, F. & RIBADEAU-DUMAS, B. (1971). *Eur. J. Biochem.* **23**, 41.

MERCIER, J. C., GROSCLAUDE, F. & RIBADEAU-DUMAS, B. (1973). *Eur. J. Biochem.* **40,** 323.
MICHELSON, A. M. & ORKIN, S. H. (1980). *Cell* **22,** 371.
MORÁVEK, L. & KOSTKA, V. (1974). *FEBS Lett.* **43,** 207.
MURTHY, M. R. N., REID, T. J., SICIGNANO, A., TANAKA, N. & ROSSMANN, M. G. (1981). *J. mol. Biol.* **152,** 465.
NEI, M. (1975). *Molecular Population Genetics and Evolution.* Amsterdam: North-Holland.
NEI, M. (1978). *Genetics* **89,** 583.
NEI, M. & CHAKRABORTY, R. (1973). *J. mol. Evol.* **2,** 323.
NEI, M. & CHAKRABORTY, R. (1976). *J. mol. Evol.* **8,** 381.
NEI, M. & GRAUR, D. (1984). *Evol. Biol.* **17,** 73.
NEI, M., FUERST, P. A. & CHAKRABORTY, R. (1978). *Proc. natn. Acad. Sci. U.S.A.* **75,** 3359.
NEVILLE, D. M. (1971). *J. biol. Chem.* **246,** 6328.
NEVO, E. (1978). *Theor. Pop. Biol.* **13,** 121.
NEVO, E., BEILES, A. & BEN-SHLOMO, R. (1984). In: *Evolutionary Dynamics of Genetic Diversity.* (Mani, G. S. ed.). pp. 13–213. Berlin: Springer-Verlag.
NIALL, H. D., HOGAN, M. L., SAUER, R., ROSENBLUM, I. Y. & GREENWOOD, F. C. (1971). *Proc. natn. Acad. Sci. U.S.A.* **68,** 866.
NICOL, D. S. H. W. & SMITH, L. F. (1960). *Nature* **187,** 483.
OHTA, T. & KIMURA, M. (1973). *Genet. Res.* **22,** 201.
OVCHINNIKOV, Y. A., ABDULAEV, N. G., FEIGINA, M. Y., KISELEV, A. V., LOBANOV, N. A. & NASIMOV, I. V. (1978). *Bioorg. Khim.* **4,** 1573.
OYER, P. E., CHO, S., PETERSON, J. D. & STEINER, D. F. (1971). *J. biol. Chem.* **246,** 1375.
PERUTZ, M. F. (1983). *Mol. Biol. Evol.* **1,** 1.
PLUMMER, R. L. & HIRS, C. H. W. (1964). *J. biol. Chem.* **239,** 2530.
POWELL, J. R. (1975). *Evol. Biol.* **8,** 79.
RAMSHAW, J. A. M., COYNE, J. A. & LEWONTIN, R. C. (1979). *Genetics* **93,** 1019.
REID, K. B. M., GAGNON, J. & FRAMPTON, J. (1982). *Biochem. J.* **203,** 559.
RIBADEAU-DUMAS, B., GROSCLAUDE, F. & MERCIER, J. C. (1970). *Compt. Rend. Acad. Sci.* (D) **270,** 2369.
RIBADEAU-DUMAS, B., BRIGNON, G., GROSCLAUDE, F. & MERCIER, J. C. (1972). *Eur. J. Biochem.* **25,** 505.
RODBARD, D. & CHRAMBACH, A. (1970). *Proc. natn. Acad. Sci. U.S.A.* **65,** 970.
ROMERO HERRERA, A. E. & LEHMANN, H. (1971*a*). *Nature New Biol.* **232,** 149.
ROMERO HERRERA, A. E. & LEHMANN, H. (1971*b*). *Biochim. biophys. Acta* **251,** 482.
ROMERO HERRERA, A. E. & LEHMANN, H. (1972). *Biochim. biophys. Acta* **278,** 62.
ROMERA HERRERA, A. E. & LEHMANN, H. (1974). *Biochim. biophys. Acta* **336,** 318.
SABER, M. A., STÖCKBAUER, P., MORÁVEK, L. & MELOUN, B. (1977). *Collect. Czech. Chem. Commun.* **42,** 564.
SCHACHMAN, H. K. (1963). *Cold Spring Harbor Symp. Quant. Biol.* **28,** 409.
SCHROEDER, W. A., SHELTON, J. R., SHELTON, J. B., ROBBERSON, B., APELL, G., FANG, R. S. & BONAVENTURA, J. (1982). *Arch. Biochem. Biophys.* **214,** 397.
SEEBURG, P. H. (1982). *DNA* **1,** 239.
SEPULVEDA, P., MARCINISZYN, J., LIU, D. & TANG, J. (1975). *J. biol. Chem.* **250,** 5082.
SHIRE, S. J., HANANIA, G. I. H. & GURD, F. R. N. (1974). *Biochemistry* 13, 2967.
SHUMAKER, K. M., ALLARD, R. W. & KAHLER, A. L. (1982). *J. Hered.* **73,** 86.
SMYTH, D. G., STEIN, W. H. & MOORE, S. (1963). *J. biol. Chem.* **238,** 227.
SPRECHER, D. L., TAAM, L. & BROWER, H. B. (1984). *Clin. Chem.* **30,** 2084.
STEPANOV, V. M., BARATOVA, L. A., PUGACHEVA, I. B., BELYANOVA, L. P., REVINA, L. P. & TIMOKLINA, E. A. (1973). *Biochem. biophys. Res. Commun.* **54,** 1164.
STONE, D. & SMILLIE, L. B. (1980). *J. biol. Chem.* **253,** 1137.
STRYER, L. (1975). *Biochemistry.* San Francisco: W. H. Freeman.
SURES, I., GOEDDEL, D. V., GRAY, A. & ULLRICH, A. (1980). *Science* **208,** 57.
SUZUKI, K. & ANDO, T. (1972). *J. Biochem.* **72,** 1419.
TANFORD, C. & HAUENSTEIN, J. D. (1956). *J. Am. chem. Soc.* **78,** 5287.
THATCHER, D. R. (1980). *Biochem. J.* **187,** 875.
THOMPSON, E. O. P. & FISHER, W. K. (1978). *Austral. J. biol. Sci.* **31,** 443.
THOMSEN, J., LUND, E. H., KRISTIANSEN, K., BRUNFELDT, K. & MALMQUIST, J. (1972). *FEBS Lett.* **22,** 34.
ULLRICH, A., DULL, T. J., GRAY, A., BROSIUS, J. & SURES, I. (1980). *Science* **209,** 612.
WALKER, J. E. (1976). *FEBS Lett.* **66,** 173.

WATT, K. W. K., COTTRELL, B. A., STRONG, D. D. & DOOLITTLE, R. F. (1979*a*). *Biochemistry* **18,** 5410.
WATT, K. W. K., TAKAGI, T. & DOOLITTLE, R. F. (1979*b*). *Biochemistry* **18,** 68.
WEBER, K. & OSBORN, M. (1969). *J. biol. Chem.* **244,** 4406.
WILSON, A. C., CARLSON, S. S. & WHITE, T. J. (1977). *Ann. Rev. Biochem.* **46,** 573.
WOLFENSTEIN-TODEL, C. & MOSESSON, M. W. (1980). *Proc. natn. Acad. Sci. U.S.A.* **77,** 5069.
WOO, S. L. C., BEATTIE, W. G., CATTERALL, J. F., DUGAICZYK, A., STADEN, R., BROWNLEE, G. G. & O'MALLEY, B. W. (1981). *Biochemistry* **20,** 6437.
YULIKOVA, E. P., EVSEENKO, L. K., BARATOVA, L. A., BELYANOVA, L. P., RYBIN, V. K. & SILAEV, A. B. (1976). *Bioorg. Khim.* **2,** 1613.
ZANNIS, V. I., BRESLOW, J. L., UTERMANN, G., MAHLEY, R. W., WEISGRABER, K. H., HAVEL, R. J., GOLDSTEIN, J. L., BROWN, M. S., SCHONFELD, G., HAZZARD, W. R. & BLUM, C. (1982). *J. Lipid Res.* **23,** 911.