Species Adaptation in a Protein Molecule

M. F. Perutz
MRC Laboratory of Molecular Biology

The allosteric properties of hemoglobins, especially their responses to ligands other than oxygen, vary widely in different classes of vertebrates. Knowing the stereochemistry of the cooperative effects in human hemoglobin, one can infer the stereochemical basis of these variations from the changes in amino acid sequence. The results indicate that the tertiary and quaternary structures of deoxy- and oxyhemoglobin have remained almost invariant during vertebrate evolution and that most of the amino acid replacements between species are functionally neutral. Adaptations leading to responses to new chemical stimuli have evolved by only a few (one to five) amino acid substitutions in key positions. Once such a response has become superfluous, it may be inactivated, not necessarily by a reversal of one of the original substitutions but by any other that happens to inhibit it.

Introduction

Species adaptation at the molecular level is a virgin field. Lewontin (1979) complained that "it has proved remarkably difficult to get compelling evidence for changes in enzymes brought about by selection, not to speak of adaptive changes. . . ." Such evidence has recently been gathered for hemoglobins whose response to different chemical stimuli varies widely in vertebrates living in different environments. How did this adaptation come about? Is it the result of changes in tertiary and quaternary structure? Has it been brought about by the gradual accumulation of minor mutations, each producing a small shift in chemical affinity, or by a few amino acid substitutions in key positions? Have similar changes in chemical affinity in different species been brought about by the same amino acid substitutions or by different ones? What selective advantage does possession of multiple hemoglobin genes offer to a species? Are most amino acid substitutions between species functionally significant, and have they evolved by Darwinian selection (Goodman 1981), or are they caused by random fixation of neutral or

1. Key words: species adaptation, proteins, hemoglobin structure.

Address for correspondence and reprints. M. F. Perutz, MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, England.

© 1983 by The University of Chicago. All rights reserved.
0737-4038/83/0101-0007$02.00
nearly neutral mutations (King and Jukes 1969)? We can now at least begin to answer these questions.

The hemoglobins of two primitive vertebrates, the lamprey and the hagfish, are monomeric when oxygenated and aggregate into oligomers on deoxygenation. They contain only one type of chain. Hemoglobins of the higher vertebrates are tetramers made up of two α- and two β-chains, each containing between 141 and 147 amino acid residues. Each chain is linked to one heme. The α-chain contains seven and the β-chain eight helical segments, designated A-H; they are interrupted by nonhelical segments marked AB, BC, and so on. A short nonhelical segment, called NA, precedes the first helix, and another, called HC, follows the last helix (fig. 1). The four chains are arranged tetrahedrally around a twofold symmetry axis which runs along a water-filled cavity. Uptake and release of oxygen are accompanied by small changes in tertiary structure of the segments surrounding the hemes and by a large change in quaternary structure involving a rotation of one αβ-dimer relative to the other by 15°, together with a relative shift of 1 Å. The rotation that occurs on dissociation of oxygen widens the cavity between the two β-chains so that the cationic groups that form its lining can bind organic phosphates. The two quaternary structures will be referred to as the oxy or relaxed (R) and the deoxy or tense (T) structures (Fermi and Perutz 1981).

All vertebrate hemoglobins react cooperatively with molecular oxygen. Their oxygen equilibria are influenced by various chemical factors, known as hetero-

---

**Fig. 1.**—Tertiary structure of globin chain, showing helical and nonhelical segments and position of heme.
tropic ligands, and by temperature. In human erythrocytes these ligands are $H^+$, $Cl^-$, $CO_2$, and D-2,3-bisphosphoglycerate (DPG); they all reduce the oxygen affinity of hemoglobin in a physiologically advantageous manner by combining preferentially with the T structure. In addition, the uptake of protons on release of oxygen, known as the alkaline Bohr effect, facilitates the transport of $CO_2$ from the tissues to the lungs in the form of $HCO_3^-$. In 0.1 M chloride the uptake of protons vanishes at pH 6; below that pH protons are released on release of oxygen. This is known as the acid or reverse Bohr effect; it does not seem to have any physiological function in mammals but is dominant even at physiological pH in certain fish and amphibia. In humans a minor part of the $CO_2$ is carried from the tissues to the lungs in the form of carbamino ion bound to hemoglobin. DPG is an allosteric regulator which lowers the oxygen affinity and thereby facilitates the release of oxygen to the tissues. The reaction of human hemoglobin with oxygen is exothermic, so that the oxygen affinity decreases with rising temperature (Weatherall 1976; Bunn et al. 1977; Imai 1982).

The effects of heterotropic ligands arise because the T structure has a low affinity for oxygen and a high one for protons, chloride, organic phosphate, and $CO_2$. In the R structure these relative affinities are reversed. The transition between the two structures gives rise to the cooperativity of oxygen binding. Hill’s coefficient $n$ provides a measure of the cooperativity. In terms of allosteric theory, the equilibrium curve is described by the oxygen dissociation constants of the two alternative structures $K_T$ and $K_R$ and by the equilibrium constant $L = [T]/[R]$ in the absence of oxygen. Heterotropic ligands strongly influence $K_T$ and $L$ but exercise only a weak influence on $K_R$ (Weatherall 1976; Imai 1982). ($K_T$ and $K_R$ are usually expressed in [mm Hg]$^{-1}$).

Structural analysis of human hemoglobin has shown that all heterotropic ligands lower the oxygen affinity because they stabilize the T structure. They do so by forming salt bridges within and between the subunits (fig. 2). The binding sites of the heterotropic ligands are as follows: DPG binds in the cavity between the two $\beta$-chains to Val NA1(1)B, His NA2(2)B, Lys EF6(82)P, and His H21(143)B (fig. 3). (In this notation the first number gives the position of the residue along the chemistry see Fermi and Perutz (1981).

![Fig. 2.—Salt bridges in human deoxyhemoglobin (Weatherall 1976). For details of the stereochemistry see Fermi and Perutz (1981).](image-url)
FIG. 3.—D-2,3-bisphosphoglycerate (DPG) binding site between the two \( \beta \)-chains in human deoxy-hemoglobin. On transition to the oxy structure the \( \alpha \)-amino groups move apart and the EF segments close up, so that the complementarity of the binding site to DPG is lost (Arnone 1972).

The helical or nonhelical segment, while the second number, in parentheses, gives its position in the amino acid sequence.) Bohr protons are bound by certain cationic groups that have their pK's raised in the \( R \rightarrow T \) transition, because they are free in the \( R \) structure but form salt bridges with carboxylates, chloride, or phosphate in the \( T \) structure. In the absence of DPG, Bohr protons bind mainly to Val \( NA1(1)\alpha \), Lys \( EF6(82)\beta \), and His \( HC3(146)\beta \), making the hydrogen bonds shown in figure 2. In the presence of DPG, Bohr protons are also bound to His \( NA2(2)\beta \) and His \( H21(143)\beta \). In vivo all these residues therefore contribute to the alkaline Bohr effect. CO, forms carboxino compounds with Val \( NA1(1)\alpha \) and \( \beta \). Chloride is bound in the internal cavity by Val \( NA1(1)\alpha \), Lys \( EF6(82)\beta \), and probably at other minor sites. In the absence of DPG, His \( H21(143)\beta \) contributes to the acid Bohr effect (Weatherall 1976; Kilmartin et al. 1980; Perutz et al. 1980).

**Fish Hemoglobins**

Many teleost (bony) fish use hemoglobin for both respiration and secretion of oxygen into the swim bladder and the eye (Wittenberg and Haedrich 1974; Wittenberg and Wittenberg 1974; Steen 1979; Ingerman 1982). Each organ possesses an elaborate vascular network, the rete, which converts a gradient of increasing lactic acid concentration into a gradient of decreasing oxygen concentration by countercurrent circulation, thus causing oxygen to be discharged. This acid-activated discharge is known as the Root effect, but it is best considered as an enhanced version of the alkaline Bohr effect which causes so drastic a drop in the oxygen affinity at low pH that several hundred atmospheres of oxygen
pressure fail to saturate the hemoglobin with oxygen. The effect arises because
the response of teleost fish hemoglobins to heterotropic ligands is markedly dif-
ferent from that of most mammalian hemoglobins. While in human hemoglobin
Hill's coefficient \( n \) remains at or just below 3.0 independent of pH, in fish hemo-
globins it drops to unity, or even below unity, near pH 6. This disappearance of
cooperativity is due to inhibition of the allosteric transition from the T to the R
structure at acid pH. Alkaline pH tends to inhibit the transition from the R to the
T structure instead, so that \( n \) rises to a maximum of only just above 2.0 near
neutral pH and in many, but not all fish, falls again at alkaline pH. Both the span
of oxygen affinities covered by the pH range 6–9 and the magnitude of the alkaline
Bohr effect (\( \Delta \log p_A / \Delta \text{pH} \)) are larger in fish than in mammalian hemoglobins.
In carp hemoglobin for instance, \( K_4 \), the association constant of the first oxygen
molecule to be taken up, drops 100-fold between pH 9 and 6.5, compared with
eightfold in human hemoglobin; over the same pH range \( K_4 \), the association con-
stant of the last oxygen, drops 10-fold in carp compared with 1.2 fold in human.
On oxygenation, human hemoglobin in deionized water releases no more than
one proton per tetramer, while carp releases 3.6. However, the enhancement of
the alkaline Bohr effect by chloride or phosphate is smaller in carp than in human
hemoglobin. While the oxygen dissociation constants of the \( \alpha \)- and \( \beta \)-subunits of
human hemoglobin differ but slightly, those of fish hemoglobins at acid pH may
differ by as much as 100-fold, which allows one pair of subunits to secrete oxygen
into the swimbladder and the eyes against hydrostatic pressures as high as several
hundred atmospheres (for details of the properties summarized here, see Noble
et al. [1970]; Binotti et al. [1971]; Gillen and Riggs [1972]; Tan et al. [1972]; Brunori
et al. [1973]; Brunori [1975]; Imai and Yonetani [1975]; Riggs [1979]; Chien and
Mayo [1980]; Morris et al. [1981]; Morris and Gibson [1982]).

What makes the allosteric properties of these fish hemoglobins with Root
effects so different from those of mammalian hemoglobins? Seeing that the amino
acid replacements between, say, human and carp hemoglobin number over 140,
this seems a forbidding problem, but systematic examination on my atomic model
of human hemoglobin led me to the conclusion that the Root effect is caused
mainly by a single amino acid replacement, that of cysteine F9B in mammals by
serine in fish. My reasoning was as follows. In the T structure of all hemoglobins
that exhibit an alkaline Bohr effect, the imidazole of the C-terminal histidine HC3B
forms a salt bridge with Asp or Glu FG1B; this raises the pH of the histidine so
that protons are bound. The C-terminal carboxyl of the same histidine forms a
salt bridge with the invariant Lys C5\( \alpha \), whether or not the hemoglobin shows a
Bohr effect. In mammals the sulfydryl of Cys F9B is in van der Waals contact
with that oxygen of the C-terminus that is not bonded to Lys C5\( \alpha \), but it forms
at most a very weak hydrogen bond with that oxygen. The OH of serine, however,
forms strong hydrogen bonds. The atomic model shows that the OH of Ser F9B
is so placed that it can donate a hydrogen bond to the free terminal oxygen atom
of His HC3B and accept a hydrogen bond from the peptide NH of His HC3B
(figs. 4, 5; table 1). These additional hydrogen bonds would stabilize the C-terminal
salt bridges in the T structure, thereby raising the allosteric constant \( L \) and the
pK of His HC3B and lowering the oxygen-binding constant \( K_T \), especially of the
\( \beta \)-chains (Perutz and Brunori 1982). Experimental tests of the theory are described
below.
Fish hemoglobins exhibit at least one other feature that is likely to shift their allosteric equilibria toward the T structure. In the R structure of mammalian hemoglobins, the terminal carboxyl of His HC3β forms an external salt bridge with Lys HC1 of the same β-chain. In teleost fish hemoglobins this external lysine is replaced by a glutamine which would not bind the carboxyl in an aqueous

Fig. 4.—C-terminal salt bridges and the role of residue F9β in the T and R structures of mammalian and fish hemoglobins. a, Human T structure; b, human R structure; c, carp T structure; d, carp R structure. The letters F, G, and H denote helical segments; FG, GH, and HC denote nonhelical segments. The same notation is used in figs. 5–6 (Perutz and Brunori 1982).

Fig. 5.—Stereo diagram showing the bonds at the C-terminus of the β-chain in the T structure of teleost fish hemoglobin. The amino group of Lys C5α donates a hydrogen bond to one of the carboxylate oxygens of His HC3β, and the OH of Ser F9β donates a hydrogen bond to the other oxygen. The imidazole donates a hydrogen bond to the carboxylate of Glu FG1. The main chain NH of His HC3 donates a hydrogen bond to the lone pair electrons of the OH of Ser F9. Note the direct link from the heme iron via His F8 and Ser F9 to the C-terminus, which would therefore feel the movement of the iron toward the porphyrin that occurs on oxygen binding (Perutz and Brunori 1982).
### Table 1
**Important Amino Acid Replacements Distinguishing Mammalian, Amphibian, and Fish Hemoglobins**

<table>
<thead>
<tr>
<th>Position of Amino Acid Residue</th>
<th>Human</th>
<th>carp</th>
<th>Trout IV</th>
<th>Trout I</th>
<th>Aquatic Frog Xenopus</th>
<th>Tadpole of Xenopus</th>
<th>Rana esculenta (Europe)</th>
<th>R. catesbeiana (America)</th>
<th>Tadpole of R. catesbeiana</th>
<th>SHARK (Heterodontus portus-jacksoni)</th>
<th>Lungfish (Lepidodactylus paradoxa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>β-chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA2</td>
<td>His</td>
<td>Glu</td>
<td>Asp</td>
<td>Glu</td>
<td>Gly</td>
<td>His</td>
<td>–</td>
<td>–</td>
<td>His</td>
<td>His</td>
<td>His</td>
</tr>
<tr>
<td>EF6</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Leu</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
</tr>
<tr>
<td>F6</td>
<td>Glu</td>
<td>Val</td>
<td>Val</td>
<td>Glu</td>
<td>Lys</td>
<td>Thr</td>
<td>Glu</td>
<td>Glu</td>
<td>Glu</td>
<td>Lys</td>
<td>His</td>
</tr>
<tr>
<td>H9</td>
<td>Cys</td>
<td>Ser</td>
<td>Ser</td>
<td>Ala</td>
<td>Ser</td>
<td>Ala</td>
<td>Ser</td>
<td>Ser</td>
<td>Ala</td>
<td>Ala</td>
<td>Ser</td>
</tr>
<tr>
<td>FG1</td>
<td>Asp</td>
<td>Glu</td>
<td>Glu</td>
<td>Asn</td>
<td>Glu</td>
<td>Glu</td>
<td>Asn</td>
<td>Gly</td>
<td>Asn</td>
<td>Gly</td>
<td>Glu</td>
</tr>
<tr>
<td>H21</td>
<td>His</td>
<td>Arg</td>
<td>Arg</td>
<td>Ser</td>
<td>Lys</td>
<td>His</td>
<td>Lys</td>
<td>Lys</td>
<td>His</td>
<td>Lys</td>
<td>Arg</td>
</tr>
<tr>
<td>HCl</td>
<td>Lys</td>
<td>Gln</td>
<td>Gln</td>
<td>Arg</td>
<td>Gly</td>
<td>Gly</td>
<td>Ala</td>
<td>Gly</td>
<td>Ser</td>
<td>Glu</td>
<td>Glu</td>
</tr>
<tr>
<td>HC3</td>
<td>His</td>
<td>His</td>
<td>His</td>
<td>Phe</td>
<td>His</td>
<td>Phe</td>
<td>His</td>
<td>Phe</td>
<td>His</td>
<td>His</td>
<td>His</td>
</tr>
<tr>
<td>α-chain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA1</td>
<td>Val</td>
<td>Ac-Ser</td>
<td>Ac-Ser</td>
<td>Ac-Ser</td>
<td>Leu</td>
<td>?</td>
<td>Gly</td>
<td>?</td>
<td>Ac-Ser</td>
<td>Ac-Ser</td>
<td>Met-Arg</td>
</tr>
<tr>
<td>C3</td>
<td>Thr</td>
<td>Gln</td>
<td>Gln</td>
<td>Gln</td>
<td>Lys</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>Gln</td>
<td>Ala</td>
<td>Gly</td>
</tr>
<tr>
<td>Root effect</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Bohr effect</td>
<td>Normal</td>
<td>Strong</td>
<td>Strong</td>
<td>Absent</td>
<td>Normal</td>
<td>?</td>
<td>Weak</td>
<td>Weak-Normal</td>
<td>Reversed</td>
<td>Weak</td>
<td>Normal</td>
</tr>
<tr>
<td>Sources</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4, 15</td>
<td>5, 16</td>
<td>6</td>
<td>7, 8</td>
<td>9</td>
<td>10, 11</td>
<td>12, 13</td>
<td>14</td>
</tr>
</tbody>
</table>

*a For meaning of the symbols, see figs. 1–6.

environment, so that less energy is needed to move His HC3β toward Asp FG1β on going from the R to the T structure (figs. 4, 5). The constellation of polar groups described here has been found in the hemoglobins of carp (Grujic-Injac et al. 1980), goldfish (Braunitzer and Rodewald 1980), trout IV (Bossa et al. 1976), and suckers (*Catostomus clarkii*) (Powers and Edmundson 1972), all of which exhibit strong Root effects.

Mammals and frogs use DPG as an allosteric effector, but teleost fish use ATP, GTP, or inositol pentaphosphate and probably also lactate (Gillen and Riggs 1977; Isaacks et al. 1977). While in mammals DPG lowers the oxygen affinity more than ATP, the opposite holds in teleost fish (Gillen and Riggs 1971). Mammalian hemoglobins whose oxygen affinity is regulated by DPG have a hydrogen donor side chain in position NA2β (His, Gln, or Asn) and they have His in position H21β (fig. 3). Teleost fish have either Glu or Asp in position NA2β and Arg at H21β. Substitution of those side chains in the atomic model of human deoxyhemoglobin produces a constellation of charged groups stereochemically complementary to strain-free ATP or GTP (fig. 6). The model suggests that, when ATP is bound, the carboxylate of either Glu or Asp NA2β, accepts a hydrogen bond from the N-6 amino group of adenine; the amino group of Val 1B, and the guanidinium group of Arg H21β, each donate a hydrogen bond to (PO₄)₃⁻, while Lys EF6β, and β, donate hydrogen bonds to either of the other two phosphates, thus neutralizing the four negative charges of ATP (Perutz and Brunori 1982).

Since this structure was first proposed, Braunitzer and his colleagues have determined the amino acid sequence of rhinoceros hemoglobin (Mazur et al. 1982). Its allosteric effector site shows only a single substitution compared with that of human hemoglobin: His Na2β → Glu, yet GTP lowers its oxygen affinity more than ATP, and ATP lowers it more than DPG, just as in teleost fish (R. Baumann, unpublished). This observation supports the hydrogen bonds between the purine and Glu NA2 proposed in figure 6; in fact, it can hardly be explained without these bonds. It also suggests that GTP is bound more tightly than ATP.

The stereochemistry of ATP and GTP in carp deoxyhemoglobin has recently been determined by nuclear magnetic resonance (Clore et al., unpublished) using

---

**Fig. 6.—Suggested ATP binding site of fish hemoglobins**
the time-dependent transferred nuclear Overhauser effect (Clore and Gronenborn 1983). This experiment showed the purine to be in the anti conformation with respect to the ribose in both trinucleotides; the ribose pucker is 3' endo; the O₅—C₅—C₄—C₃ torsion angle is trans, Pᵣ—O₁—C₂—C₃ are coplanar (trans), Pᵢ—O—P₈ and P₁—O—P₂ are as in the free nucleotide, probably all trans, so that electrostatic repulsion between the phosphates is minimized. The results confirm the structure of ATP in carp deoxyhemoglobin proposed by Perutz and Brunori (1982), except for a small change in the angle of tilt of the purine (fig. 6). To convert the structure of ATP into that of GTP, the entire nucleotide has to be turned by roughly 180° about its long axis. This leaves the γ-phosphate bound to Val 1α of subunit β₁ and allows N₁ and N₂ of the guanine to donate hydrogen bonds to Glu NA2 of subunit β₂. The formation of two hydrogen bonds between the purine and Glu NA2 in GTP, compared with the single hydrogen bond in ATP, explains why GTP lowers the oxygen affinity of carp hemoglobin twice as much as ATP (Weber and Lykkeboe 1978).

It is thought that during fast movement the pH at the gills may drop too low for efficient oxygen uptake by those hemoglobins that exhibit the Root effect; to ensure a continued oxygen supply, trout and some other fast-swimming fish have two kinds of hemoglobin, those that respond to heterotropic ligands and those that do not. In trout, the hemoglobin that exhibits the Root effect has been assigned the number IV, while the two hemoglobins that fail to respond to heterotropic ligands are called I and II (Brunori 1975). Component III makes up only 3% of the total, and its ligand-binding properties are not known. In trout I all the polar residues involved in both the alkaline and acid Bohr effects and in phosphate binding are replaced by neutral ones: Lys EF₆β is replaced by Leu, Ser F₉β by Ala, Glu FG₁β by Asn, Arg H₂₁β by Ser, and His HC₃β by Phe (table 1). The sequences of trout hemoglobins I and IV show many other replacements, but so far none of these is of a kind likely to have a significant effect on the affinity for heterotropic ligands.

The properties just described are not typical of all fish. Among elasmobranch, skate hemoglobin has no appreciable Bohr effect and chloride raises rather than lowers its oxygen affinity. Like trout I, torpedo hemoglobin fails to respond to organic phosphates (Riggs 1970, p. 209; Johansen and Weber 1976, p. 219). These properties cannot yet be interpreted for lack of amino acid sequences. Hemoglobin of the shark Heterodontus portusjacksoni, a cartilaginous fish without a swim bladder, exhibits a weak alkaline Bohr effect (Nash et al. 1976). Position F₉β is occupied by Ala; residue FG₁β is Glu, which could produce a normal alkaline Bohr effect by forming a strong salt bridge with His HC₃β if it were not for a Lys in position F₆β, one turn of π-helix away, competing with His HC₃β for that salt bridge. The lungfish Lepidosiren paradoxa also lacks a swim bladder; its hemoglobin exhibits no Root effect and has an alkaline Bohr effect similar in magnitude to that of human hemoglobin (Phelps et al. 1979). It does have Ser at F₉β and Glu at FG₁β, but as in Xenopus and in the shark, its salt bridge with His HC₃β is weakened by competition from residue F₆β, in this instance a His.

The fish discussed so far are cold blooded, but fast-swimming sharks can maintain their bodies 7—10 C and tuna (Thunnus thynnus) up to 15 C above the water temperature (Carey and Teal 1969; Carey et al. 1971); cooling of the muscles is minimized by a countercurrent exchange system between the arterial and venous circulations that transfers metabolic heat from the veins to the cold blood arriving
in the arteries from the gills. In most species the oxygen affinity of hemoglobin drops with rising temperature because the reaction of heme with oxygen is exothermic; if this were true also in tuna, heating of the cold arterial blood would cause some of its oxygen to dissociate and be transferred to the nearby veins unused. To reduce this waste, tuna have evolved a hemoglobin in which the reaction with oxygen is endothermic (Carey et al. 1971; Gibson and Carey 1982). This reversal arises because the intrinsic heat of oxygenation of the hemes is exceeded by the heat absorbed in the T → R transition. In human hemoglobin that transition absorbs about half of the -66 kcal released when four molecules of oxygen combine with one molecule of hemoglobin; in trout IV it absorbs about two-thirds. To produce a reversed temperature coefficient, the T → R transition in tuna absorbs about 100 kcal (Gibson and Carey 1982). What is the source of the extra 30 kcal? The main source of heat in the allosteric transition comes from the formation of hydrogen bonds. Taking the enthalpy of hydrogen bond formation in the dimerization of N-methylacetamide in benzene of 3.6 kcal/mol as a measure (Davies and Thomas 1956), the T structure would have to contain four additional and the R structure four fewer hydrogen bonds to achieve the desired result. Taking the enthalpy of 1.5 kcal/mol for the disruption of a CO-HN bond in water as a measure (Kauzmann 1959), these numbers would rise to the improbable value of 10. Another enthalpy sink could be generated if the sum of the van der Waals interactions in the R structure was less than in the T structure, so that heat of fusion would be absorbed in the T → R transition (Bello 1977).

Lest hemoglobin be regarded as indispensable to vertebrate life, let me draw attention to the existence of three species of large antarctic fish devoid of erythrocytes or blood pigment (Rund 1954). The amount of dissolved oxygen their very cold, colorless blood can carry is only one-eighth of that of two other antarctic fish with red blood, but that seems to be adequate for the sluggish existence of these large fish.

Experimental Tests of the Theory of the Root Effect

When X-ray analysis had indicated that a salt bridge formed by His HC3β in the T structure may be responsible for the major part of the alkaline Bohr effect in horse and human hemoglobins, Kilmartin set out to test that prediction by preparing a hemoglobin from which that C-terminal histidine had been enzymatically cleaved. X-ray analysis showed that this cleavage had no significant effect on the structure of the rest of the hemoglobin molecule. Nevertheless, it halved the alkaline Bohr effect (Kilmartin and Wootton 1970; Perutz and Ten Eyck 1971). This observation and many subsequent ones supported the dominant contribution of this one specific residue (Kilmartin et al. 1980). More recent is the discovery of an abnormal human hemoglobin with a drastically reduced alkaline Bohr effect, arising from the substitution Asp FGlβ → Asn which neutralizes the anionic arch of the salt bridge, as does the substitution Glu FGlβ → Asn that occurs between trout IV and I hemoglobins (Como et al. 1983).

If the Root effect in fish hemoglobins is due largely to the extra stabilization conferred by Ser F9β on the salt bridges of His HC3β, it should be inhibited by enzymatic cleavage of His HC3. Parkhurst, Goss, and Perutz tested that prediction by preparing des-His carp hemoglobin (Parkhurst et al. 1983). This halved the alkaline Bohr (Root) effect, lowered L, raised $K_T$, and also changed the kinetics of ligand binding. On approach to pH 6, native carp hemoglobin shows a sharp
rise in the rate of oxygen dissociation and fall in the rate of CO recombination characteristic of the Root effect. Both the rise and fall are much diminished in des-His carp hemoglobin. All these observations supported the theory.

A further test became possible when the hemoglobin of the South African frog *Xenopus* was found to have a serine in position F9β (Williams et al. 1980). According to the theory, this hemoglobin should exhibit a Root effect at low pH, and this was confirmed by Perutz and Brunori (1982). We found that it has a much lower oxygen affinity than human hemoglobin, consistent with a more stable T structure; its Hill’s coefficient tends toward unity at pH 6, and its kinetics of CO recombination at low pH are similar to those of hemoglobins exhibiting a Root effect. However, its Root effect is not as strong as that of carp hemoglobin.

**Amphibia**

To produce a strong alkaline Bohr effect and a Root effect at acid pH, *Xenopus* hemoglobin has conserved the essential Ser F9β, Glu FG1β, and His HC3β, and it has Gly rather than Lys in position HC1β to prevent anchoring of the C-terminus in the R structure (table 1). However, it has a Lys at F6β, one turn of \( \pi \)-helix away from Glu FG1β and capable of competing with His HC3β for the essential salt bridge with that Glu (figs. 4, 5). This competition may weaken its alkaline Bohr effect compared with that of carp.

*Xenopus* hemoglobin is exceptional among amphibians for its strong alkaline Bohr effect; in 0.1 M Cl\(^-\) other amphibian hemoglobins show either weak alkaline or only weak acid Bohr effects. The major hemoglobin component of the large American bullfrog *Rana catesbeiana*, when stripped and in dilute solution, shows a Bohr effect of \( \Delta \log p_{50}/\Delta \text{pH} = -0.16 \), compared with \(-0.56\) in *Xenopus* (Aggarwal and Riggs 1969). Position F9β is occupied by Ser, as in *Xenopus*, but its influence is counteracted by the substitution Gly for Glu in position FG1β (table 1; figs. 4, 5). The two major fractions of the adult hemoglobin of the European frog *R. esculenta* also exhibit only weak alkaline Bohr effects (Brunori et al. 1968). The sequence of the major component of the \( \beta \)-chain is similar to that of *R. catesbeiana*; position FG1β is occupied by Asn rather than Gly, which again inhibits the strong salt bridge with His HC3β. His H21β, which contributes to the acid Bohr effect in human hemoglobin A, is replaced by Lys in all three adult frog hemoglobins (fig. 3).

Tadpole hemoglobin of *R. catesbeiana* in the absence of phosphates has only a reverse Bohr effect (Atha et al. 1979). Here Ser F9β is replaced by Ala, Glu FG1 by Asn, and Ser 1α, the other major source of the alkaline Bohr effect, is acetylated, so that the alkaline Bohr effect is completely inhibited. However, His H21β is present, consistent with its role in the acid Bohr effect of human hemoglobin. His H21β is also present in *Xenopus* tadpole hemoglobin; this has Phe in position HC3β, which would inhibit most of the alkaline Bohr effect contributed by the \( \beta \)-chains (D. Banville et al., unpublished). The amino acid sequence of the \( \alpha \)-chain is not yet known. The \( \beta \)-chains of both tadpole species carry the same residues at the phosphate binding site as does adult human hemoglobin; organic phosphates, if present, would therefore be expected to overcome the acid Bohr effect and produce a weak alkaline one instead.
In another amphibian, *Amphiuma*, the adult hemoglobin has an acid Bohr effect at physiological pH, converted in vivo to a weak alkaline one by organic phosphate (Bonaventura et al. 1977). The adult hemoglobin of the crested newt, *Triturus cristatus*, also exhibits an acid Bohr effect at physiological pH, but in this amphibian it is maintained in the hemolysate; even so, the higher concentration of organic phosphates in the red cell may convert it to a weak alkaline Bohr effect in vivo (Condo et al. 1981). Blood of the frog *Telmatobius culeus* that lives in Lake Titicaca at an altitude of 3,800 m has a higher oxygen affinity than that of any other frog (Hutchinson et al. 1976), but the oxygen equilibrium curves of its purified hemoglobin are still unknown. It will be interesting to analyze the stereochemical bases of all these unusual properties once amino acid sequences become available.

Why is the alkaline Bohr effect of the bloods of the Port Jackson shark, of the newt, of *Amphiuma*, and of the anuran tadpoles so much weaker than that of mammals? In the blood of air breathers the partial pressure of CO₂ is about 40 mm Hg, and a substantial Bohr effect is needed for its transport from the lungs to the tissues; but in the blood of water breathers the partial pressure of CO₂ is only 2-4 mm Hg because it is 28 times more soluble in water than oxygen and diffuses out of the body through the gills and the skin. Having no swim bladder and no rete to pump oxygen into the eye, the shark and the tadpoles can therefore get along with only a weak alkaline Bohr effect. The same is true of the aquatic frog *Amphiuma*, which is regarded as a permanent larva, even though the adult has lost its external gills and shows other evidence of partial metamorphosis (Rahn 1966; Lenfant and Johansen 1967; Bonaventura et al. 1977; Shelton and Boutilier 1982). Judging by the strong acid Bohr effect, the switch from tadpole to adult hemoglobin may not have occurred in this amphibian. In contrast, the adult terrestrial bullfrog *R. catesbeiana* loses only a small part of its CO₂ through the skin even when submerged and relies on ventilation through its lungs. Its blood therefore needs a larger alkaline Bohr effect than that of *Amphiuma* (Lenfant and Johansen 1967). The same appears to be true of the South American lungfish *Lepidosiren paradoxa* which lives in hot, stagnant water and surfaces to breathe (Phelps et al. 1979; Rodewald and Braunitzer 1983). *Xenopus* is anomalous. It is an aquatic frog with a low blood CO₂, capable of respiration through the skin or even of switching to anaerobic metabolism (Shelton and Boutilier 1982). At this stage I have not found any satisfying explanation for why it needs a large alkaline Bohr effect at physiological pH, or a Root effect at acid pH.

**Reptiles**

Crocodilians are able to stay underwater for as long as an hour without coming up to breathe, but they lack the high concentrations of myoglobin that provide diving mammals and birds with large stores of oxygen. Instead, crocodilians reduce oxygen consumption by shutting off the circulation to their muscles so that oxygen supply is restricted to their brain and viscera. Even so, they need to use as much as possible of the oxygen stored in their lungs and their blood. This appears to be accomplished by the unusual allosteric properties of their hemoglobin, which hardly responds to the normal heterotropic ligands (protons, CO₂, chloride, or organic phosphates) but responds strongly only to bicarbonate ion (fig. 7). Crocodilian hemoglobins bind two equivalents of bicarbonate ion per tetramer which lower the oxygen affinity as much as P6-inositol (inositol-hexa-

Model building showed that the bicarbonate binding sites lie in the cavity between the two β-chains where organic phosphates or carbamino CO₂ are bound in other species; each HCO₃⁻ is bound by Lys EF6 and Glu HC1 of one β-chain together with the N-terminal residue of its partner chain (fig. 8). In caiman hemoglobin, the N-terminal sequence is Ser-Pro-Phe, compared with Val-His-Leu in human hemoglobin (table 2). The human tripeptide is straight (fig. 3), but the

Fig. 7.—Oxygen equilibrium curves of caiman and horse hemoglobin with and without CO₂. Pco₂ = 40 torr, 25 C, I = 0.2 M, pH 7.2 for caiman (Bauer et al. 1981); pH 7.4 for horse (Kilmartin and Rossi-Bernardi 1969).

Fig. 8.—Stereo drawing of proposed bicarbonate binding site between the two β-chains of caiman deoxyhemoglobin. The central sign marks the dyad symmetry axis. Bicarbonates and their binding residues are underlined. Capital letters mark helical and interhelical segments (Perutz et al. 1981). Residues are marked in sequential, rather than structural notation. The following list gives the structural numbers with the sequential ones in parentheses: Ser NA1(1), Pro NA2(2), Phe NA3(3), Ser A1(4), Ala A2(5), His A3(6), Lys EF6(82), Glu HC1(144), Tyr HC2(145), His HC3(146), all β.
Table 2
Amino Acid Replacements for Allosteric Control in Crocodilian Hemoglobins

<table>
<thead>
<tr>
<th>POSITION</th>
<th>Human</th>
<th>Other Bony Vertebrates</th>
<th>Nile Crocodile</th>
<th>Mississippi Crocodile</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA1β</td>
<td>Val</td>
<td>Val or Gly</td>
<td>Ser</td>
<td>Ac-Ala</td>
</tr>
<tr>
<td>NA2β</td>
<td>His</td>
<td>Gln, Asn, Glu, His, Asp, Met</td>
<td>Pro</td>
<td>Ser</td>
</tr>
<tr>
<td>EF6β</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
<td>Lys</td>
</tr>
<tr>
<td>H21β</td>
<td>His</td>
<td>Lys, Arg, Ser, His</td>
<td>Ala</td>
<td>Ala</td>
</tr>
<tr>
<td>HC1β</td>
<td>Lys</td>
<td>Arg, Ala, Ser, Glu, Lys</td>
<td>Glu</td>
<td>Glu</td>
</tr>
<tr>
<td>H14α</td>
<td>Ser</td>
<td>Very variable</td>
<td>Ala</td>
<td>Ala</td>
</tr>
</tbody>
</table>

NOTE.—For full sequence, see Leclercq et al. (1981).

caiman tripeptide is forced to turn a corner at the proline. The turn brings the N-terminal scissile within reach of the bicarbonate ion, so that one of the bicarbonate oxygens can form a salt bridge with the α-NH⁺ and can also accept a hydrogen bond from the serine OH. The second bicarbonate oxygen can form a salt bridge with Lys EF6β, and the third oxygen can donate a hydrogen bond to one of the carboxylate oxygens of Glu HC1β. In the hemoglobins of the Nile crocodile and the Mississippi alligator, where the N-terminal sequence is acetyl-Ala-Ser-Phe, the α-NH of acetylated alanine could donate a hydrogen bond to O₁ of the bicarbonate at the same distance as the α-NH⁺ does in caiman, provided the chain bends at Ser 2 at the same angle as it does at Pro 2 in caiman.

Consider now the loss of allosteric inhibition by organic phosphates, carbamino CO₂, and chloride. The loss of affinity for organic phosphates is accounted for by the replacements His NA2β → Pro or Ser and His H21β → Ser (fig. 3). In human deoxyhemoglobin, carbamino CO₂ bound to Val 1α accepts a hydrogen bond from Ser H14α. In the three crocodilian hemoglobins this Ser is replaced by Ala. This replacement may also inactivate one of the strong chloride-binding sites which probably coincides with the CO₂ site. In human deoxyhemoglobin, another pair of CO₂ molecules competes with organic phosphates for binding to Val 1β; when bound, each carbamino CO₂ is stabilized by a salt bridge to Lys 82β. In caiman hemoglobin, the bend in the chain forced by Pro 2 inhibits that bridge; in the other two crocodilian hemoglobins, the blocking of Val 1β inhibits binding of CO₂ (Perutz et al. 1981).

The decrease in oxygen affinity brought about by the interaction of crocodilian hemoglobins with bicarbonate ensures that oxygen is released from the blood to the tissues at a relatively high partial pressure of oxygen. If the hemoglobin were insensitive to bicarbonate, the venous Po₂ would be only 7 torr; interaction with bicarbonate raises this to 27 torr, thus creating a large enough pressure head for the flow of oxygen from the blood to the tissues (Jelkmann and Bauer 1980a). In human hemoglobin, the same relative rise in Po₂ from 19 to 40 torr can be brought about only by the combined effects of CO₂ and DPG. It is surprising that the crocodilian hemoglobins’ simple and direct reciprocating action between oxygen and one of the end products of oxidative metabolism has not been adopted by other vertebrates.

These results show that an entirely new function can evolve in a protein by no more than three amino acid substitutions (Val NA1β → Ser, His NA2β →
Pro, and Lys HCl → Glu), requiring only four nucleotide base changes; just two more amino acid substitutions (His H21β → Ala and Ser H14α → Ala), or three base changes, are needed for inhibition of the old functions (oxygen-linked phosphate, carbamino CO₂, and Cl⁻ binding). Most of the other ~100 substitutions that distinguish crocodilian from human hemoglobins are conservative and would have little if any effect on the oxygen equilibrium. There are some significant substitutions in the heme pockets and at the subunit contacts, but none of these is unique or could play any part in the allosteric control by bicarbonate ions. All the residues which are essential for the formation of the characteristic T and R structures are either conserved or have been replaced by ones that can serve the same purpose equally well. A recent paper suggests that the hemoglobins of another class of diving reptiles, the sea turtles, have allosteric properties resembling those of the crocodilians (Isaacks et al. 1982).

Until recently investigators have searched in vain for an intermediate between the primitive dimeric hemoglobins of the lamprey and hagfish and the α,β tetramers of the higher vertebrates. Such an intermediate has now been found in the reptile *Sphenodon punctatus* that inhabits islands off the shore of New Zealand and is a survivor from the Triassic period of the ancient order of "beakhead" reptiles. It closely resembles the rhynchocephalians living 200 million yr ago. The reptile's blood and its stripped hemolysate both exhibit hyperbolic oxygen equilibrium curves, the former with a p50 = 19 mm Hg and the latter with a p50 = 1.8 mm Hg. The alkaline Bohr effect was small: Δp50/ΔpH = -0.16. The erythrocytes contain 1.2 mol ATP per mole hemoglobin tetramer, and addition of ATP restored p50 of the hydrolysate to 14 mm Hg (Wells et al. 1983).

These properties pose a riddle. Organic phosphates have never been observed to influence the oxygen equilibrium curve of any hemoglobin other than α,β tetramers that undergo an allosteric transition, yet *Sphenodon* hemoglobin looks as though it responded to ATP without such a transition, because its reaction with oxygen is devoid of cooperativity. The hemoglobin is tetrameric, but it is not clear yet whether the tetramer is made up of pairs of α- and β-like subunits. Conceivably it consists of two pairs of β-like chains, each pair capable of combining with ATP as in figure 6. The ATP would constrain the chains in their tertiary deoxy structure without being released by an allosteric transition on oxygenation. Further study of the hemoglobin may provide important clues to the evolution of cooperative oxygen binding.

**Birds**

Bird hemoglobins are functionally similar to mammalian ones but they use P₇-inositol (inositolpentaphosphate) in place of DPG as an allosteric effector. The residues at the phosphate-binding site are the same as in mammals except that position H21β is occupied by Arg rather than His, as in teleost fish. Two further replacements in the internal cavity are H13β Ala → Arg and H17β Asn → His (Takei et al. 1975; Oberthür et al. 1980; Oberthür et al. 1981; Braunitzer and Godovac 1982). These two basic residues may be too far removed from the P₇-inositol binding site to bond to its phosphates directly, but they would help to neutralize its negative charges and thus contribute indirectly to the binding energy.

One remarkable instance of adaptation has been reported in the hemoglobins of two related species of geese. The greylag goose (*Anser anser*) lives in the plains, and the oxygen affinity of its blood is normal; the bar-headed goose (*Anser indicus*)
migrates across the Himalayas at an altitude of 9,000 m, and the oxygen affinity of its blood is abnormally high (Petschow et al. 1977). In laboratory experiments the bar-headed goose proved more resistant to hypoxic stress than the Canada goose or the Pekin duck, neither of which flies at such high altitudes (Black and Tenney 1980). It was then found that the oxygen affinities of the purified hemoglobin of the two species differed only slightly; at pH 7.2 in 0.1 M NaCl at 37°C the hemoglobin of the greylag goose has a $p_{50}$ of 5.8 mm Hg and that of the bar-headed goose one of 4.5 mm Hg. Their affinities for P$_T$-inositol were identical. How was this to be reconciled with the large differences between the oxygen affinities of their bloods? Rollema and Bauer (1979) demonstrated that by lowering the oxygen affinities of both hemoglobins 10-fold, P$_T$-inositol also amplified the differences between them 10-fold, raising $p_{50}$ to 50 and 37 mm Hg, respectively: this decrease of 13 mm in $p_{50}$ is sufficient to explain the high oxygen affinity of the bar-headed goose’s blood.

The amino acid sequences of the two hemoglobins differ by only four substitutions, of which only one is unique among the bird sequences determined so far (Petschow et al. 1977; Oberthür et al. 1981). This is H2α Pro (greylag goose) → Ala (bar-headed goose). Pro H2α is invariant in nearly all other species, probably because its Cα forms an important van der Waals contact at the αβ boundary: in mammals with Met D6β and in birds with Leu D6β. Studies of abnormal human hemoglobins have shown that the loss of any interatomic contact at subunit boundaries is liable to loosen the constraints of the T structure, thereby lowering $L$ and raising $K_T$ (Fermi and Perutz 1981). The loss of this van der Waals contact in the bar-headed goose hemoglobin may therefore explain its high oxygen affinity.

Mammals

H. F. Bunn has shown that mammalian hemoglobins can be divided broadly into two groups: the great majority have an intrinsically high oxygen affinity, which is lowered in the red cell by DPG, while those of ruminants and cats (Cervidae, Bovidae, Felidae) and of one primate, the lemur, have an intrinsically low oxygen affinity that is little, or not at all, lowered by DPG (Bunn 1971; Bonaventura et al. 1974; Hamilton and Edelstein 1974; Iaketa 1974; Scott et al. 1977). Typically, hemoglobins with high intrinsic oxygen affinity have $p_{50}$ values of between 4 and 6 mm Hg, and those with low affinity have $p_{50}$ values of between 10 and 20 mm Hg (measured in stripped hemoglobin solutions in 0.05 M bis-tris, 0.1 M NaCl, pH 6.5–7.5 at 20–25°C). The low affinity is due to lower values of $K_T$ and higher values of $L$, while $K_R$ remains the same in the two groups of species (Perutz and Imai 1980).

Nearly all the amino acid replacements that distinguish the high and low oxygen affinity hemoglobins are conservative and external, but consistent differences are found in position NA2β, which is occupied by hydrophilic residues, His, Gln, or Asn in hemoglobins with high affinity and by large hydrophobic residues, Leu, Met, or Phe in those with low affinity. What part could residue NA2β play in lowering the oxygen affinity? In the human T structure, DPG forms salt bridges with ValNA1 and HisNA2; these salt bridges pull the two helices A toward the center of the molecule so that they become locked more tightly to neighboring segments of the polypeptide chain (fig. 3). In bovine, sheep, and deer hemoglobins, residue NA1 is missing and residue NA2 is Met, which almost abolishes their affinity for organic phosphates. In these hemoglobins, the hydro-
phobic side chain of the N-terminal Met probably plays the same part in locking helix A tightly in place and thereby stabilizing the T structure as organic phosphates do in human and other hemoglobins with intrinsically high oxygen affinity. In cat hemoglobin this part would be played by Phe, and, in lemur Hb, by LeuNA2β (Perutz and Imai 1980).

Several mammals living at high altitude have bloods with high oxygen affinities. They include the llama, the golden-mantled ground squirrel (Citellus lateralis), and the yellow-bellied marmot (Marmota flaviventris). For example, at 35 mm Hg CO₂ and 37 C, pSO₂ was about 20 mm Hg for the marmot, 30 mm Hg for the squirrel, and 50 mm Hg for a lowland rodent, the rat (Hall et al. 1936; Bartels et al. 1963; Bullard et al. 1966). Amino acid sequences have been reported only for llama hemoglobin (Braunitzer et al. 1977a, 1977b), which has a low oxygen affinity in the absence of phosphate, overcompensated by a low affinity for DPG, due to the replacement His NA2β → Asn (Bauer et al. 1980). The side chain of asparagine is 1.3 Å shorter than that of His or Gln. Figure 3 shows that this shortening places the amino groups of the asparagines further from the phosphates of DPG than the NH of the histidine or the amino group of a glutamine. In consequence the affinity for DPG is weakened. Asn NA2β occurs also in the hemoglobin of the elephant (Braunitzer et al. 1982), whose blood shows a high oxygen affinity for an animal that lives at low and medium, rather than at high, altitudes (pSO₂ = 22.4 mm Hg at 39 C) (Bartels et al. 1963); perhaps this helped the elephants who carried the supplies for Hannibal’s army across the Alps in 218 B.C. In fact, there exists a somewhat weak correlation between body size and oxygen affinity of the blood, of which the elephant with pSO₂ = 23 mm Hg and the mouse with pSO₂ = 46 mm Hg are two extreme examples (Scott et al. 1977).

A similar adaptive mutation is found in human fetal hemoglobin. This has a lower oxygen affinity than the adult in the absence of phosphate, but it also has a lower affinity for DPG, mainly due to the replacement His H21(143)β → Ser. As a result, the oxygen affinity of fetal blood exceeds that of the adult, which helps the transfer of oxygen from the maternal to the fetal circulation across the placenta (Bunn et al. 1977).

On going to high altitudes, humans show an increased DPG concentration in their erythrocytes, which lowers the oxygen affinity. This used to be regarded as an adaptive response, lower oxygen affinity allowing a larger fraction of the oxygen carried to be discharged in the tissues. Theoretical and experimental studies have shown that this may be true at moderate altitudes (<3,000 m), but at higher altitudes a raised oxygen affinity is more advantageous, for the following reason. Suppose we mark the arterial and the venous oxygen pressure on the oxygen equilibrium curve and connect the two points by a line. Efficiency of oxygen transport will be greatest when the line lies on the steepest slope of the sigmoid curve. At moderate altitudes a shift to lower oxygen affinity was found to move the line to a steeper part of the curve, while at higher altitudes the reverse was true (Turek et al. 1973, 1978).

The mole (Talpa europaea) lives in its burrows under hypoxic conditions, to which it is adapted by having a blood with a high oxygen affinity, a high concentration of hemoglobin per unit volume, and a low body temperature (Quilliam et al. 1971). The high oxygen affinity of its blood is reported to be due to a low affinity of its hemoglobin for DPG, but how the amino acid sequence affects this remains unclear (Kleinschmidt et al. 1981).
Species Adaptation in Enzymes

I know of only two instances of adaptation of protein molecules other than hemoglobin that have been interpreted in atomic detail. They are the enzymes ferredoxin and glyceraldehyde phosphate dehydrogenase in thermophile bacteria. In these heat-resistant enzymes an entire range of adaptive mechanisms has been encountered, from 10 amino acid substitutions in the 55-residue chain of ferredoxin to create six more external salt bridges (Perutz and Raidt 1975), to a multitude of replacements in glyceraldehyde phosphate dehydrogenase of Bacillus steator-thermophilus to produce extra salt bridges and van der Waals contacts (Walker et al. 1980). These enzymes have had many billions more generations to evolve than those of vertebrates, but even so their tertiary and quaternary structures have remained closely similar to those of mesophiles.

New enzymatic activities have evolved in the laboratory in bacteria subjected to selective pressures. For instance, growth of Klebsiella aerogenes on xylitol in place of its "natural" substrate ribitol led to a point mutation in the gene for ribitol dehydrogenase that increased that enzyme's affinity for xylitol (Hartley et al. 1976). Clarke has studied the evolution of an aliphatic amidase in Pseudomonas aeruginosa (Clarke 1980). The wild-type strain grows well on acetamide, but not on longer chain amides. Under selective pressure a single point mutation that substituted a serine for a phenylalanine in the enzyme's polypeptide chain produced a strain that grew on butyramide. A further single mutation produced a strain that grew on phenylacetamide and had lost the ability to grow on acetamide. Hall (1981) has studied the evolution of a minor β-galactosidase in a strain of Escherichia coli from which the gene for the major β-galactosidase had been deleted. He found spontaneous point mutations that altered the specificity of the enzyme, so that it hydrolyzed lactose and other sugars on which the original strain would not grow. The structures of the three enzymes used in these studies are unknown, so that we cannot unravel the stereochemical mechanisms underlying the evolution of new catalytic activities, but the fact that they did evolve by one or two point mutations is consistent with the conclusion from my hemoglobin studies that such evolution can be accomplished by very few amino acid substitutions.

There exists a large literature on enzyme polymorphism and species adaptation (Koehn et al. 1983), but none of it can as yet be interpreted in stereochemical terms. The best-studied species is Drosophila melanogaster, where the frequency of the two dominant alleles of alcohol dehydrogenase varies with latitude in several continents; one of these alleles has a consistently lower Michaelis constant for alcohols than the other (McDonald et al. 1980). The two enzymes have been found to differ by the single substitution of a lysine for a threonine at position 192 of the polypeptide chain (Thatcher 1981). Unfortunately the stereochemical meaning of the substitution remains unknown, because there is no sequence homology between the Drosophila enzyme and the horse heart enzyme whose structure has been determined (Bränden et al. 1973), but it is noteworthy that here again enzyme adaptation is the result of a single substitution. Little is known as yet about the chemistry of other polymorphic enzymes.
Discussion

We may now reconsider the questions raised in the introduction. Does adaptation involve changes in the tertiary and quaternary structure of hemoglobin? I have no firm answer, because the only structures that have been accurately determined are those of human and horse hemoglobin. Their tertiary and quaternary structures are the same within experimental error, despite 42 amino acid substitutions between them. The number of amino acid substitutions between human and fish hemoglobins is about 140 or half of the total number of amino acids; those between human and caiman number 110. Few of these lie at heme contacts or at contacts between the subunits. Comparison of the amino acid sequences of human and carp hemoglobins shows that of the residues in contact with the hemes in human hemoglobin listed by Fermi and Perutz (1981), 16 out of 20 are identical in the \( \alpha \)-chains and 14 out of 20 in the \( \beta \)-chains. At the \( \alpha,\beta \) contact, which forms the switch between the deoxy and oxy structures, 27 out of 32 residues listed by Fermi and Perutz (1981) are identical in the two species. These homologous residues would be stereochemical misfits unless the tertiary and quaternary structures superposed with a standard deviation of less than 1 Å. I am confident that X-ray analysis will bear this out. Additions and deletions have occurred in nonhelical segments: for instance, in the Port Jackson shark, which is regarded as a living fossil, the \( \alpha \)-chain contains 144 residues instead of the 142 in other fish, and the \( \beta \)-chain contains only 141 instead of the 147 in other fish (Nash et al. 1976; Fisher et al. 1977). The additional two residues in the \( \alpha \)-chain are at the amino end and can be tucked away in the internal cavity without change of structure. The deletions in the \( \beta \)-chain occur in the CD segment, which is nonhelical and external; they require only small local adjustments in tertiary structure (fig. 1).

My proposal that the tertiary and quaternary structure of the hemoglobins has been conserved throughout evolution from fish to mammals has met with disbelief among biologists and others, but it comes as no surprise to protein crystallographers, who have found that homologous proteins in distant species have closely similar structures. A few examples are glyceraldehyde phosphate dehydrogenase from *Bacillus stearothermophilus* (Biesecker et al. 1977) and lobster (Buehner et al. 1974), phosphoglycerate kinase from yeast (Bryant et al. 1974) and horse (Banks et al. 1979), bovine and fungal catalase (Murphy et al. 1981; Vainshtein et al. 1981), human and hen lysozyme (Artymiuk and Blake 1981; Mioto et al. 1972), and cytochromes c (Dickerson 1972).

The tertiary structures of vertebrate myoglobins, of the monomeric lamprey hemoglobin, and of invertebrate and leguminous plant hemoglobins all resemble those of the \( \alpha \)- and \( \beta \)-chains of vertebrate hemoglobin to the extent necessary to preserve the vitally important geometry of the heme pocket, but they vary in detail. The angles between helical segments differ by up to 30° and the points of contact between them by up to 7 Å. Many different combinations of side chains are found to produce helix interfaces that are comparably well packed, as if the tertiary structure had been conserved by a patchwork of improvisations (Lesk and Chothia 1980). In fact, the only invariant residues common to all globins are the proximal histidine F8 and phenylalanine CD1 which wedges the heme into its pocket. What maintains the tertiary structure is a set of sites that are invariably occupied by nonpolar residues (Perutz et al. 1965).
Has adaptation been brought about by the gradual accumulation of minor mutations, each producing a small shift in chemical affinity, or by a few amino acid substitutions in key positions? In the instances analyzed so far, new chemical functions appear to have evolved by only a few amino acid substitutions in key positions. Consider the evolution from cartilaginous fish to bony fish. The hemoglobin of the Port Jackson shark has only a weak alkaline Bohr effect and has Ala in position F9. It seems that the Root effect in teleost fish has evolved primarily by the substitution of this Ala by Ser. In the crocodilians, affinity for bicarbonate evolved by no more than three amino acid substitutions, requiring only four nucleotide base changes; just two more amino acid substitutions, or three base changes, were needed for inhibition of the affinity for the usual heterotropic ligands (carbamino CO2, organic phosphate, and Cl−). Five amino acid substitutions probably suffice to inhibit in trout I hemoglobin all the heterotropic interactions exhibited by trout IV hemoglobin. Conversion from an ATP to a DPG binding site needs only one amino acid substitution or one base change; modulation of the affinity for DPG in the human fetus, the llama, and the elephant also needs only a single base change.

In amphibia that do not need a Root effect, the alkaline Bohr effect is weakened by a variety of single substitutions: Ser F9 → Ala in the tadpoles of Xenopus and Rana catesbeiana, Glu FG1β → Asn or Gly in adult frogs and in one of the tadpoles, His HC3 → Phe in another tadpole. It looks as though once a function is no longer needed or has become undesirable, evolution would permit any mutation that inhibits the function to become fixed (Kimura 1983).

What selective advantage does possession of multiple hemoglobin genes provide? In fish that possess two types of hemoglobin, one responsive and the other unresponsive to heterotropic ligands, such multiplicity offers a clear advantage. One of the hemoglobins assures the supply of oxygen to the swim bladder and the eye, while the other assures the fish of a continued supply of oxygen when low pH inhibits the uptake of oxygen by the first hemoglobin (Brunori 1975). The oxygen affinities of mammalian fetal blood are higher than those of the corresponding maternal blood, thus ensuring oxygen transfer across the placenta. In primates and ruminants this need is met by fetal hemoglobins that have higher oxygen affinities than the adult ones, but other species, among them the horse, manage with the same hemoglobin in both fetus and adult (Stockell et al. 1961). Studies of dogs, rats, and rabbits suggest that in species lacking fetal hemoglobins the oxygen affinity of the fetal blood is raised relative to that of the adult by a lower concentration of 2,3-DPG. This is achieved by an increased activity of pyruvate kinase, which helps to metabolize 1,3-DPG, the precursor of 2,3-DPG (Jelkmann and Bauer 1978; Jelkmann and Bauer 1980b; Mueggler and Black 1982).

Human adults possess two α- and two β-globin genes. The two α-chain genes code for identical amino acid sequences and are both efficiently expressed, so that humans have the advantage of being effectively tetraploid for α-globin. This benefits individuals born with a deletion or malfunction of one of the α-chain genes (e.g., α-thalassemia); like sickle cell heterozygotes such individuals are healthy and also have a better chance of surviving malaria than individuals with wild-type hemoglobin genes. The two polymorphic adult β-chains differ at 10 positions and are functionally indistinguishable, but one of them, known as the δ-chain, is made in only 1/40 of the quantity of the other, so that possession of the δ gene confers no significant advantage to either β-thalassemia or sickle cell
anemia patients. It may be a "fossil" of some past stage of evolution. Many vertebrates have multiple $\alpha$ and $\beta$ genes, some coding for functionally equivalent and others for functionally different genes. It is not clear whether their possession offers any selective advantage or whether they have merely survived from selectively neutral gene duplications. H. F. Bunn has suggested to me that the presence of multiple hemoglobins might allow a higher total hemoglobin concentration to be maintained in the red cell than if there were only a single one. On the whole, the evidence supports Kimura's view that "in many cases polymorphism has no visible phenotypic effect and no obvious correlation with environmental conditions" (Kimura 1979).

How many of the amino acid substitutions between the hemoglobins of different species are functionally significant? The great majority of amino acid differences between hemoglobins of different species consist of conservative or semiconservative replacements of external polar and nonpolar residues and internal nonpolar ones. Comparison of normal and abnormal human hemoglobins, or of the hemoglobins of closely related species, shows that such replacements have little, if any, influence on the functional properties of hemoglobin. For example, human and horse hemoglobins differ in sequence at 42 out of 287 positions, yet their functional properties are the same. Only a minority of the 400 or so different amino acid replacements found in abnormal hemoglobins affect any of the hemoglobin functions. It could be argued that even an amino acid replacement that produces only a very small shift in the oxygen equilibrium curve may give an animal a selective advantage that would prove decisive over thousands of generations; against this it could be held that homeostatic mechanisms allow organisms to compensate efficiently for quite large shifts in the curve. Studies of abnormal hemoglobins have shown mutations that merely change surface charges to be without significant effect on the respiratory functions of hemoglobins, but recent work has revealed that such mutations may influence the rate of assembly of the $\alpha$- and $\beta$-subunits into $\alpha_2\beta_2$ tetramers by electrostatic effects arising from their different isoelectric points ($\alpha^\alpha = 8.0; \beta^\alpha = 6.5$). At physiological pH free $\alpha$-subunits exist as monomers and dimers carrying a net positive charge, while free $\beta$-chains exist as tetramers, dimers, and monomers carrying a net negative charge. The effect of these charges on the assembly is illustrated by the following example. The sickle cell mutation Glu A4(6)$\beta$ $\rightarrow$ Val reduces the negative charge on the $\beta$-subunits, thus reducing electrostatic repulsion in the $\beta_2^\beta$-tetramer and electrostatic attraction between $\alpha^\alpha$- and $\beta^\alpha$-subunits. In consequence both the dissociation of $\beta_2^\beta$-tetramers and the association of $\beta^\alpha$- and $\alpha^\beta$-subunits is slower than the dissociation of $\beta_2^\alpha$-tetramers and the association of $\beta^\alpha$- and $\alpha^\beta$-subunits. That difference in rates appears to account for the ratio of hemoglobin A to S in sickle cell heterozygotes being about 3:2 rather than 1:1. The higher ratio of hemoglobin A contributes to the fitness of the heterozygotes. It seems important, therefore, that the net charges of the subunits in an oligomeric protein be maintained (Bunn and McDonald 1984). With that proviso, the structural evidence suggests that most of the amino acid replacements between species are neutral or nearly so, caused by random drift of selectively equivalent mutant genes, and that adaptive mechanisms generally operate by a few replacements in key positions. Here again the evidence supports Kimura's view that "adaptive mutations are much
less frequent than selectively neutral or nearly neutral substitutions caused by random drift" (Kimura 1983).

R. Lewontin has pointed out to me that a mutation adapting a species to a new environment is likely to have preceded occupation of that environment. For example, a mutation that raised the oxygen affinity of the llama's blood would have occurred before llamas discovered that they were able to graze at altitudes barred to competing species. W. Bodmer suggested that once a large change in chemical affinities produced by one mutation had enabled a species to occupy a new environment, its effect might have been refined by later adaptive mutations, each contributing minor shifts, over a long period of time. The structure of glyceraldehyde phosphate dehydrogenase from *B. stearothermophilus* (Walker et al. 1980), referred to earlier, suggests that the heat stability of this enzyme may have evolved by such a process. This might be the molecular equivalent of punctuated equilibria in the evolution of species.

**Acknowledgments**

This paper has benefited from helpful suggestions made by Prof. Christian Bauer, Dr. Walter Bodmer, Drs. Joseph and Celia Bonaventura, Dr. Sydney Brenner, Prof. Maurizio Brunori, Dr. H. Frank Bunn, Prof. Patricia H. Clarke, Dr. Thomas E. Creighton, Dr. Richard E. Koehn, Dr. Arthur M. Lesk, Dr. Richard C. Lewontin, Dr. Robin Perutz, and Dr. Graham Shelton. To all these colleagues I wish to express my gratitude.

**LITERATURE CITED**


——. 1980b. 2,3-DPG levels in relation to red cell enzyme activities in rat fetuses and hypoxic newborns. Pflügers Arch. 389:61–68.


Thatcher, D. R. 1981. The complete amino acid sequence of three alcohol dehydrogenase alleloenzymes (Adh<sup>−/−</sup>, Adh<sup>+</sup> and Adh<sup>+++</sup>) from the fruitfly *Drosophila melanogaster*. Biochem. J. **187**:875, 886.


Walter M. Fitch, reviewing editor

Received August 17, 1983; revision received August 25, 1983.
Structure and Evolution of Human and African Ape rDNA Pseudogenes

Elise Brownell
State University of New York at Stony Brook

Mark Krystal
Mount Sinai School of Medicine

Norman Arnheim
State University of New York at Stony Brook

We discuss the evolutionary significance of four aberrant 18S rDNA clones that were obtained from human, chimpanzee, and gorilla DNA libraries. We show that these clones carry representatives of a small 18S rDNA pseudogene family that arose in a common ancestor of these species. Aspects of their structure and phylogenetic distribution suggest that the 18S pseudogenes no longer interact genetically with normal ribosomal genes and therefore may not be linked to nucleolus organizer regions.

Introduction

The genes that encode 18S and 28S ribosomal RNAs (rDNA) are present in multiple tandemly arrayed copies in most eukaryotic genomes. The 18S and 28S genes of orangutan (Pongo pygmaeus), gorilla (Gorilla gorilla), chimpanzee (Pan troglodytes), and human (Homo sapiens) are found on several pairs of nonhomologous chromosomes (Henderson et al. 1972, 1974, 1976, 1979; Evans et al. 1974; Tantravahi et al. 1976). One interesting aspect of the rRNA genes in these species is that they evolve in unison, or concertedly, despite their multichromosomal distribution (Arnheim et al. 1980).

To learn more about the molecular genetic mechanisms that are responsible for this phenomenon, we isolated representatives of the functional rDNA repeats from human and chimpanzee genomic DNA libraries. In addition, we purified unusual 18S rDNA-containing recombinants from human, chimpanzee, and gorilla DNA libraries. Characterization of these aberrant clones and comparison with the normal rDNA repeats led us to ask how the unusual genes were generated and how they were evolutionarily maintained despite rDNA turnover in these species.
Material and Methods

Enzymes

Restriction endonucleases were purchased from New England Biolabs, Boehringer-Mannheim, and Bethesda Research Laboratories. We used the digestion conditions recommended by the suppliers.

Cloning Studies

We obtained the clones shown in figure 1B–E by screening partial human and chimpanzee EcoRI genomic DNA libraries (in Charon 4A) with a mouse 18S rDNA probe (Benton and Davis 1977). The human library was a generous gift from Dr. J. Slightom; the chimpanzee library was constructed using standard procedures (Blattner et al. 1977; Lawn et al. 1978). The human clones in figure 1B and C were later subcloned into the plasmid vector pBR-322 (Arnheim and Kuehn 1979). One of these subclones, a HindIII/EcoRI fragment from the human rDNA pseudogene (fig. 1C, probe 4), was used to screen a gorilla EcoRI genomic DNA library (kindly provided by Dr. Alan Scott). Two recombinants were isolated, one of which is shown in figure 1F.

Southern Transfer and Hybridization Studies

For the experiments shown in figure 2a, b, and c we used 32P-labeled probes that were nick translated (Rigby et al. 1977) to a specific activity of at least 1.5 × 106 cpm/μg with 32P dCTP. All hybridizations were carried out in 3 × SSC (1 × SSC = 0.15 M NaCl, 0.015 M Na citrate), 0.1% sodium dodecyl sulfate (SDS), and 0.06% each of polyvinylpyrrolidone (PVP), Ficoll, bovine serum albumin (BSA), with 15 mM ethylene diamine tetra acetic acid (EDTA) at 65 C for 16–18 h. For each experiment 2 × 106 cpm of probe per milliliter of hybridization buffer were used. Hybridized filters were washed three times (1 h each) at 65 C in 3 × SSC, 0.1% SDS before autoradiography.

Results and Discussion

We screened human, chimpanzee, and gorilla partial EcoRI genomic DNA libraries with rDNA and rDNA related probes. Five of the six clones that we isolated are shown in figure 1B–F. Extensive restriction enzyme mapping studies of the 18-kb human (fig. 1B) and the 11.5-kb chimpanzee (fig. 1D) clones and comparison with genomic mapping data (Arnheim et al. 1980; see fig. 1A) suggested that the clones were derived from functional ribosomal genes. Indeed, both were found to be capable of initiating transcription in a human in vitro assay system (Miesfeld and Arnheim 1982 and unpublished data).

In contrast, the remaining four recombinants were unusual in several respects that would support the view that they are pseudogenes. The human (fig. 1C), chimpanzee (fig. 1E), and two presumably nonallelic gorilla (one of which is shown in fig. 1F) “aberrant” clones had EcoRI fragments atypical in length (15 kb); they lacked the evolutionarily conserved mammalian 18S gene EcoRI site (Southern 1975; Arnheim and Southern 1977; Fuke et al. 1981); and, although they retained nontranscribed spacer and internal transcribed spacer homologies detected by restriction enzyme mapping and cross-hybridization experiments (summarized in fig. 1), they lacked 28S gene sequences. In addition, although they hybridized strongly to a normal 18S gene probe, a 250-bp fragment that contained the human
FIG. 1.—Restriction enzyme maps of human, chimpanzee, and gorilla 18S rDNA genes and pseudogenes. A, A human rDNA repeating unit. A polymorphic EcoRI site (Arnheim et al. 1977; Wellauer and Dawid 1979; Arnheim et al. 1980; Nelkin et al. 1980; Erickson et al. 1981; Higuchi et al. 1981; Miesfeld et al. 1982) is shown in parentheses. Each 45-kb repeating unit (Wellauer and Dawid 1979) is composed of nontranscribed spacer (NTS), external transcribed spacer (ETS), 18S, internal transcribed spacer (ITS), and 28S sequences. The origin and direction of transcription of the 45S precursor are shown. B, The human 18S rDNA clone. The probes used to define homologous regions between the normal genes and pseudogenes by cross-hybridization experiments are also shown. C, The human 18S pseudogene. D, The chimpanzee 18S rDNA clone. E, The chimpanzee 18S pseudogene clone. F, A gorilla 18S pseudogene clone. Sites above the solid lines denote complete mapping data; sites below the line signify that only partial mapping information is presented.
rDNA origin of transcription (Miesfeld and Arnheim 1982) (region 3, fig. 1B) reacted equally weakly with the human, chimpanzee, and gorilla 15-kb clones when compared with normal rDNA genes (human data shown in fig. 2a). This suggests that these rDNA variants may be deleted for regions around the origin of transcription. Preliminary in vitro transcription assays using subclones carrying the weakly hybridizing region of the human 15-kb clone as a template failed to show transcriptional initiation (Miesfeld and Arnheim, unpublished data). Other more extensive deletions of nontranscribed spacer segments were also characteristic of the pseudogenes (fig. 1).

To eliminate the possibility that the unusual human recombinant was a cloning artifact, we tested whether predicted restriction enzyme fragments derived from it could be detected in the human genome. Two human placental DNAs digested with SacI and hybridized with a mouse 18s probe (fig. 2b, lanes B and C) gave a strong 11-kb 18s DNA-containing fragment typical of normal human rDNA repeats and a faint 6.9-kb band that corresponded in size to the 18s-containing SacI piece from the aberrant human clone (lane A). The relative intensities of these two genomic bands also suggested that the 18s rDNA variant specific fragment might be present in low copy number in the human genome. Indeed, subsequent quantitation experiments showed that there are between four and 10 copies in each human, chimpanzee, and gorilla genome (data not shown).

Unlike normal rDNA, 28s sequences were not found 2.5 kb 3' to the 18s region of the aberrant clones. To determine whether this 3' end was derived from other regions of the normal rDNA repeating unit, we hybridized EcoRI restricted human placental DNA with a subclone from the 3' end of the human pseudogene (the 850-bp HindIII/EcoRI fragment, designated pA6; fig. 2c, lane D). Because the sizes of the bands that we detected were not attributable to any of the normal rDNA EcoRI fragments (19 kb, 18 kb, 12 kb, 7.5 kb, 6 kb; fig. 1A), we concluded that pA6 contains sequences of non-rDNA origin which are themselves present in low copy number. Sequences homologous to pA6 were also found at the 3' terminus of the chimpanzee and gorilla 18s pseudogene 15-kb fragments (fig. 1E, F). Interestingly, although each gorilla recombinant contained a large 15-kb EcoRI fragment that hybridized with pA6, they were flanked at the 3' end by different, smaller EcoRI segments. Restriction enzyme analysis (data not shown) indicated that the 15-kb EcoRI fragments were identical to each other, in contrast to their 3' flanking sequences. We inferred from this observation that the clones carried nonallelic pseudogene loci.

The presence of non-rDNA homologies at the 3' ends of these 18s pseudogenes suggests possible ways in which these pseudogenes arose and how they were evolutionarily maintained. For example, 18s rDNA pseudogene formation might have been initiated following integration of exogenous DNA sequences into the internal transcribed spacer region of a normal rDNA repeat. Ribosomal genes, in fact, may have been the targets for the integration of mobile elements during evolution (Treco et al. 1982). In many Drosophila species a fraction of the ribosomal genes have insertions in the 28s gene, and repeats bearing these insertions are almost completely transcriptionally silent in vivo (i.e., pseudogenes; see Long and Dawid [1980] and Beckingham [1982]). However, it is not clear whether these insect 28s pseudogenes constitute an evolutionarily stable subpopulation of ribosomal genes within nucleolus organizing regions. It has been proposed that the insertions may move between repeats from one generation to the next through
Fig. 2.—a, The human 15-kb clone contains only a small portion of the origin of rDNA transcription. Equimolar amounts of the normal rDNA 6-kb EcoRI fragment (lane A) and the 15-kb rDNA clone (lane B) were hybridized with a 250-bp probe containing the origin of human rDNA transcription in a Southern transfer experiment (fig. 1B, probe 3). The human 15-kb clone hybridizes at approximately 1/10 of the control level. b, The aberrant 15-kb human clone is represented in the human genome. SacI digests of the human 15-kb clone (lane A) and two human placental DNA samples (lanes B, C) were hybridized with a mouse 18S DNA probe. The normal rDNA-containing bands in lanes B and C are 11.0 kb in length. The faint bands at 6.9 kb correspond in size to the SacI fragment in lane A. c, The human and chimpanzee 15-kb clones contain non-rDNA sequences. EcoRI-digested human placental DNAs were hybridized with the following probes: Lane A—A small fragment including the terminal 600 bp of the human 28S gene immediately 3' to the EcoRI site in the gene (see fig. 1A) and additional 3' nontranscribed spacer sequences which should hybridize to the 19-kb fragment. Lane B—A lambda Charon 4A phage clone containing human ETS, 18S, ITS, and the 5' portion of the 28S gene which should hybridize to the 18-kb, 6-kb, and 7.5-kb pieces. Lane C—pA4, a 3.4-kb HindIII subclone of the human 18S pseudogene (fig. 1C) adjacent to pA6. Lane D—The 850-bp HindIII/EcoRI subclone of the human 18S pseudogene (pA6) (see fig. 1C). The sizes of the expected EcoRI fragments from normal rDNA repeats are shown to the left of lane A; the sizes of the non-rDNA fragments we observed with probes C and D are noted to the right of lane D. We conclude that fragment pA4 contains both rDNA and non-rDNA sequences; fragment pA6 contains only non-rDNA.
the normal process of unequal meiotic crossing-over (Rae 1982). This would allow some inactive 28S rDNA pseudogenes to be reactivated in subsequent generations. This hypothesis is consistent with the observation that, besides the insertion itself, no obvious structural differences can be detected between the pseudogenes and normal genes even though such differences would be expected to accumulate once an active gene were permanently silenced. Even at the nucleotide sequence level, no differences have accumulated within the origin of transcription region in the interrupted genes (Long et al. 1981). In contrast, the significant structural differences between the primate rDNA pseudogenes and their normal counterparts suggest that this pseudogene family cannot be rescued through genetic interactions with normal genes and that these pseudogenes have had a long evolutionary history in isolation from normal rDNA arrays. In addition, although other pseudogenes have been found in association with many multigene families (Jacq et al. 1977; Lacy et al. 1979; Bentley and Rabbits 1980; Jahn et al. 1980; Nishioka et al. 1980; Vanin et al. 1980; Denison and Weiner 1981; Selker et al. 1981; Steinmetz et al. 1981; Karin and Richards 1982; Lemischka and Sharp 1982; Martin et al. 1983), the persistence of 18S pseudogenes over the last 5 million years (i.e., since these species last had a common ancestor; Sarich and Wilson 1967) is unexpected in view of the evolutionary behavior of the normal rDNA clusters in these three primates. The ribosomal gene repeats of man and the apes evolve concertedly (Arnheim et al. 1980); that is, within an individual the ribosomal gene units are more similar to each other than would be expected if each repeat evolved independently of all others (see Brown et al. 1972; Hood et al. 1975; Tartof 1975; Dover et al. 1982; Arnheim 1983). By examining the molecular structure of primate rDNA repeats, Arnheim et al. (1980) found that, despite a multichromosomal distribution, different species-specific restriction sites have been fixed in the clusters of even the three most recently diverged species—human, chimpanzee, and gorilla (Sarich et al. 1967). That 18S pseudogenes escaped elimination during the normal process of rDNA turnover in these three lineages suggests that the pseudogenes are genetically isolated from normal rDNA repeats. This situation would arise, for example, if pseudogenes were relegated to junctional positions between rDNA clusters and neighboring chromosomal sequences. Being at the extreme end of the array might effectively prevent genetic interaction of the rDNA pseudogene with the remaining genes. However, a recent study of yeast rDNA organization suggests that junctional segments contain rDNA sequences which have not undergone significant structural divergence in the transcribed regions (Zamb and Petes 1982). If the human and ape pseudogenes are junctional fragments, our clones are potentially useful for analyzing the DNA sequences surrounding nucleolus organizer regions.

A second situation that would lead to genetic isolation would be one in which rDNA sequences moved to a new location outside of the array (Childs et al. 1981; Denison and Weiner 1981; Jagadeeswaran et al. 1981). Possible mechanisms include the recombinational excision of looped-out normal rDNA segments during meiotic unequal alignment of the rDNA repeating units, followed by integration at a new locus. The 3' non-rDNA sequences may define the original target site. The dispersed copy, and any descendants that arose from it by amplification, could have evolved independently of the original parental locus. Additional data will allow us to distinguish between these alternatives.
Acknowledgments

Portions of this research were funded by National Institutes of Health grants 5T32CA09176 and 5RO1GM2536503 and American Cancer Society grant CD 140. We thank P. Deininger for the Alal plasmid. The human and gorilla phage libraries were provided by Drs. J. Slightom and A. Scott, respectively. We thank Meryl Diamond for screening and preliminary characterization of the gorilla pseudogenes and Dee Sedivec for preparation of the manuscript.

LITERATURE CITED


Richard K. Koehn, reviewing editor

Received June 29, 1983; revision received August 10, 1983.
Mitochondrial DNA Differentiation during the Speciation Process in *Peromyscus*

John C. Avise, John F. Shapira, Susan W. Daniel, Charles F. Aquadro, and Robert A. Lansman

University of Georgia

We address the problem of the possible significance of biological speciation to the magnitude and pattern of divergence of asexually transmitted characters in bisexual species. The empirical data for this report consist of restriction endonuclease site variability in maternally transmitted mitochondrial DNA (mtDNA) isolated from 82 samples of *Peromyscus polionotus* and *P. leucopus* collected from major portions of the respective species' ranges. Data are analyzed together with previously published information on *P. maniculatus*, a sibling species to *polionotus*. Maps of restriction sites indicate that all of the variation observed can be reasonably attributed to base substitutions leading to loss or gain of particular recognition sites. Magnitude of mtDNA sequence divergence within *polionotus* (maximum = 2%) is roughly comparable to that observed within any of five previously identified mtDNA assemblages in *maniculatus*. Sequence divergence within *leucopus* (maximum = 4%) is somewhat greater than that within *polionotus*. Consideration of probable evolutionary links among mtDNA restriction site maps allowed estimation of matriarchal phylogenies within *polionotus* and *leucopus*. Clustering algorithms and qualitative Wagner procedures were used to generate phenograms and parsimony networks, respectively, for the between-species comparisons. Three simple graphical models are presented to illustrate some conceivable relationships of mtDNA differentiation to speciation. In theoretical case I, each of two reproductively defined species (A and B) is monophyletic in matriarchal genealogy; the common female ancestor of either species can either predate or postdate the speciation. In case II, neither species is monophyletic in matriarchal genotype. In case III, species B is monophyletic but forms a subclade within A which is thus paraphyletic with respect to B. The empirical results for mtDNA in *maniculatus* and *polionotus* appear to conform closely to case III. These theoretical and empirical considerations raise a number of questions about the general relationship of the speciation process to the evolution of uniparentally transmitted traits. Some of these considerations are presented, and it is suggested that the distribution patterns of mtDNA sequence variation within and among extant species should be of considerable relevance to the particular demographies of speciation.

1. Number 5 in the series "The Use of Restriction Endonucleases to Measure Mitochondrial DNA Sequence Relatedness in Natural Populations." Key words: mitochondrial DNA, mammalian speciation, molecular evolution, restriction digests.

Address for correspondence and reprints: John C. Avise, Department of Molecular and Population Genetics, University of Georgia, Athens, Georgia 30602.
Introduction

In recent years there has been tremendous interest in the molecular changes associated with speciation. One major motivation for this interest is that any genetic changes (such as chromosome rearrangements or iteration patterns of repetitive DNA) found to be regularly correlated with speciation might also be causally responsible for morphological divergence and/or reproductive isolation. Thus for many groups of organisms, information has been gathered about the genetic differences between closely related species. For example, in the multilocus protein-electrophoretic literature, more than 3,800 genetic distances between pairs of species have been recorded for vertebrates alone (Avise and Aquadro 1982); and the relevance of allozyme data to the speciation process has been extensively discussed (Lewontin 1974; Ayala 1975, 1976; Nei 1975). Early suggestions that changes in genome regulation (rather than replacement substitutions in structural genes) might be responsible for most organismal evolution have stimulated searches for other classes of genetic characters associated with species formation (Wallace 1963; Ohno 1969; Stebbins 1969; Britten and Davidson 1971; Wilson 1976). White (1978) makes a strong case that structural chromosomal rearrangements have commonly played major roles in initiating divergence and speciation. Recent discoveries of previously unsuspected classes of DNA (e.g., repetitive DNA and transposable elements) and mechanisms of DNA change (e.g., gene conversion and molecular drive) have led to speculation about how these processes may be related to evolution and speciation (Schopf 1981; Bonner 1982; Dover and Flavell 1982; Milkman 1982).

In this paper we address the problem of mitochondrial DNA (mtDNA) differentiation during speciation. Our concern with this issue might seem surprising. To the best of current knowledge, mtDNA appears to be strictly maternally inherited in higher animals (e.g., Dawid and Blackler 1972; Francisco et al. 1979; Lansman et al. 1983b). The apparent asexual transmission of mitochondria means that female lineages are genetically and evolutionarily isolated from one another in mtDNA composition no matter whether they belong to the same or to different species. Thus, from the point of view of mtDNA, sexual reproduction and speciation might seem to be unimportant processes (except through possible selection effects involving interactions between mtDNA and different nuclear genotypes). Furthermore, it remains a viable hypothesis that much of the extensive mtDNA polymorphism commonly assayed within and among closely related vertebrate species is selectively neutral. The majority of within-species polymorphism is attributable to base substitutions rather than to additions, deletions, or rearrangements (Aquadro and Greenberg 1983; Avise and Lansman 1983; Greenberg et al. 1983); in base sequences encoding polypeptides, silent third-position changes greatly outnumber substitutions which replace amino acids (Anderson et al. 1982; Brown et al. 1982). Why then should it be of interest to examine the differentiation of mtDNA during speciation?

It is precisely because mtDNA's are asexually transmitted, and because many of their polymorphisms may not be under strong selection, that they may provide excellent genetic markers for inferring the evolutionary histories and population dynamics of female lineages ancestral to extant species. In an example of this approach, Brown (1980) interpreted the exceptionally low level of mtDNA poly-
morphism within living humans as evidence for a severe population bottleneck in our female ancestors about 180,000–360,000 yr ago. Thus even if most assayable mtDNA divergence does not directly underlie reproductive isolation and speciation, it remains of interest because of the inferences it may allow about the genealogical and demographic histories of female populations undergoing speciation. This is the thesis we will address here.

More specifically, the purposes of this report are: (1) to estimate empirically levels and patterns of mtDNA divergence between the very closely related rodent species *Peromyscus maniculatus* and *P. polionotus* (samples of *P. leucopus* were also included as an “outgroup”), (2) to provide general conceptual scenarios of how observable mtDNA differentiation among extant species might be related to the process of species formation, and (3) to interpret the empirical data on *Peromyscus* in the context of these conceptual possibilities.

**Material and Methods**

**Taxa Examined**

Many of the approximately 60 named species of the rodent genus *Peromyscus* have been placed within taxonomic “species groups” which reflect probable evolutionary affinities. *Peromyscus leucopus* belongs to the *Leucopus* group. The *maniculatus* species group consists of about six recognized species—the widespread *maniculatus*, which occurs throughout most of North America, and geographically peripheral species in the highlands of Mexico (*melanotis*), Gulf of California (*sejuginis, sleveni*), Pacific Northwest (*sitkensis* and possibly *oreas*), and southeastern United States (*polionotus*). *Polionotus* is generally thought to have evolved from *maniculatus*-like stock originally isolated on Floridian islands as a result of changing sea levels during the Pleistocene (Blair 1950).

The geographic ranges of *polionotus* and *maniculatus* do not overlap, so the magnitude of reproductive isolation in natural sympathy is unknown. In the laboratory, reproductive isolation between *polionotus* and tested *maniculatus* exists primarily in the form of prezygotic mating barriers (Blair and Howard 1944), coupled with difficulties of implantation and embryo survival following those successful matings which do occur (Maddock and Chang 1979). (Reproductive relationships among geographic samples of *maniculatus* are poorly known, but at least some populations have been reported to exhibit mild reproductive barriers as well [Dice 1968].) Glazier (1980) suggests that *polionotus* and *maniculatus* might best be considered semispecies within a larger superspecies complex.

Two earlier reports in this series also dealt with mtDNA evolution in *Peromyscus*. The first paper (Avise et al. 1979) introduced the approach of comparing mtDNA restriction digest patterns in samples of *P. maniculatus*, *P. polionotus*, and *P. leucopus* collected from a few localities. In a later study (Lansman et al. 1983a) we greatly extended the data base for *maniculatus* by employing additional enzymes and by mapping restriction sites (“site analysis”) in mtDNA from mice collected from about 40 localities across the North American continent. Site analyses provide more refined estimates of sequence divergence than do fragment comparisons alone and in addition provide more direct information about the nature of evolutionary change in the mtDNA molecule itself. The new data for this current report consist of “site analyses” of mtDNA from a total of 82 newly collected individuals of *polionotus* and *leucopus* taken from major portions of their respective geographic ranges in eastern North America. These new data will
be analyzed together with the restriction site data previously published for *maniculatus*. Collection sites for *polionotus* and *leucopus* are listed in tables 1 and 2, respectively.

**Laboratory Procedures and Data Analyses**

Mice were returned live to the laboratory, where crude "cytoplasmic nucleic acid" fractions were prepared from livers plus kidneys and hearts of individual animals, according to published procedure (Lansman et al. 1981). These were digested with the following eight restriction enzymes: (1) HincII, (2) BglII, (3) HindIII, (4) BstEII, (5) EcoRI, (6) BamHI, (7) Xba, and (8) HpaII. Digested fragments were "end labeled" with $^{32}$P-dATP using DNA polymerase as described by Brown (1980). The labeled fragments were separated by molecular weight on agarose gels and detected by autoradiographic procedures (Lansman et al. 1981).

One sample each of *polionotus* and *leucopus* was also used to construct recombinant plasmids by inserting all of the BamHI fragments of these genomes into pBR322. Much of the site mapping was accomplished by analyzing results of double-enzyme digests of the cloned sequences. The construction of clones and the mapping experiments were conducted as in Lansman et al. (1983a) (Shapira and Lansman, unpublished). The approximately 16,000 base-pair mtDNA genome of *Peromyscus* was arbitrarily assigned a total map length of 100 units, so each

<table>
<thead>
<tr>
<th>Table 1</th>
<th>mtDNA Clones Observed in Geographic Samples of <em>Peromyscus polionotus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Clone Number</td>
<td>Composite mtDNA Genotype</td>
</tr>
<tr>
<td>1</td>
<td>AAPBANEUX</td>
</tr>
<tr>
<td>2</td>
<td>APMBUEUY</td>
</tr>
<tr>
<td>3</td>
<td>AABMACEUX</td>
</tr>
<tr>
<td>4</td>
<td>AAPLAEUY</td>
</tr>
<tr>
<td>5</td>
<td>AAPLADEUY</td>
</tr>
<tr>
<td>6</td>
<td>AAPHAEUX</td>
</tr>
<tr>
<td>7</td>
<td>CAPBACEUX</td>
</tr>
<tr>
<td>8</td>
<td>EAPBACEUX</td>
</tr>
<tr>
<td>9</td>
<td>HAPBACEUX</td>
</tr>
<tr>
<td>10</td>
<td>AAPMACEUX</td>
</tr>
<tr>
<td>11</td>
<td>APMPEUEY</td>
</tr>
<tr>
<td>12</td>
<td>AAMBAEUX</td>
</tr>
<tr>
<td>13</td>
<td>oAPBHEUX</td>
</tr>
<tr>
<td>14</td>
<td>uAPBHEUX</td>
</tr>
<tr>
<td>15</td>
<td>BMMABYEUY</td>
</tr>
<tr>
<td>16</td>
<td>APMABEUY</td>
</tr>
<tr>
<td>17</td>
<td>AAPBAXEUX</td>
</tr>
<tr>
<td>18</td>
<td>AAPBAXEUX</td>
</tr>
<tr>
<td>19</td>
<td>AAPBAXEUX</td>
</tr>
<tr>
<td>20</td>
<td>ANIBAEUX</td>
</tr>
<tr>
<td>21</td>
<td>AUPCAEUX</td>
</tr>
<tr>
<td>22</td>
<td>BAPKAEVZ</td>
</tr>
</tbody>
</table>

$a$ Letters, from left to right, refer to restriction morphs for HincII, BglII, HindIII, BstEII, EcoRI, BamHI, Xba, and HpaII, respectively (see fig. 2).
Table 2  
mtDNA Clones Observed in Geographic Samples of *Peromyscus leucopus*

<table>
<thead>
<tr>
<th>Clone Number</th>
<th>Composite mtDNA Genotype*</th>
<th>Counties and States of Collection</th>
<th>Number of Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>αBAPEMAK</td>
<td>Cheshire, N.H.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>αBAOEMA</td>
<td>Cheshire, N.H.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>αBBOEMAJ</td>
<td>Union, Ga.</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>αBAOEMAK</td>
<td>Clarke, Ga.; Macon, N.C.</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>αBAPEMAJ</td>
<td>Macon, N.C.</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>αBAOENAS</td>
<td>Chester, Pa.</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>αBBOEMAE</td>
<td>Pickaway, Ohio</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>αBAODMAE</td>
<td>Giles, Va.</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>αBAOEMA</td>
<td>Giles, Va.</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>αBBOEMBS</td>
<td>Giles, Va.</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>βBAPFPFA</td>
<td>Reeves, Tex.</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>βBAQCOAG</td>
<td>Jefferson, Kans.</td>
<td>1</td>
</tr>
</tbody>
</table>

* Letters, from left to right, refer to restriction morphs for HinclI, BglII, HindIII, BstEI, EcoRI, BamHI, Xba, and HpaII, respectively (see fig. 3).

map unit corresponds to about 160 base pairs. Map position zero indicates the unique site at which the endonuclease *PstI* cleaves most *maniculatus* mtDNA.

Many of the data analyses follow procedures utilized in earlier papers of this series (Lansman et al. 1983a). Thus for each restriction enzyme and for each animal, the “raw” data consist of a restriction site map given an arbitrary letter designation for purposes of bookkeeping. The letter designations have been standardized to those of *maniculatus* presented earlier (in other words, a *BglII* map “B” for *leucopus* and/or *polionotus* would carry the identical restriction sites as a *BglII* map “B” in *maniculatus*). Considering data from all eight endonucleases, we summarize the “composite mtDNA genotype” of each animal in an eight-letter code (see note to table 1). For convenience, we say that individuals sharing a given composite genotype belong to the same mtDNA clone (with respect to the observed restriction sites). For comparisons within the species *leucopus* and *polionotus* (where mutation distances proved to be small), the eight-letter codes facilitate the linking of mtDNA clones into unrooted phylogenetic networks by a process which involves adding new clones to positions in a growing network requiring the smallest number of mutation steps (see Lansman et al. 1983a). Such networks among conspecifics were subsequently superimposed over the geographic sources of collections.

In comparisons between species and more distantly related genotypes, a useful alternative representation of raw data is a matrix of presence-absence information for each restriction site in each mtDNA clone. Our final matrix for the three species of *Peromyscus* had 10,070 elements (106 restriction sites × 95 mtDNA clones). Minimum mutation distances (the minimum number of mutation steps separating pairs of clones) were counted directly from the matrix. Estimates of nucleotide sequence divergence (*p*) between mtDNA clones were calculated by the cleavage site method of Nei and Li (1979, eq. 8). The final value of *p* was weighted according to relative numbers of cleavage sites produced by the endonucleases recognizing four- and six-base cleavage sites.
The mtDNA molecule is shown in linear form with the scale beginning at the zero PstI site. Sites in *polionotus* are shown above each line, those in *leucopus* below. Conserved sites (those present in all samples from a species) are shown by solid lines; variable sites (those absent in one or more samples) are shown by dashed lines. Stars indicate sites mapped to only one of two possible alternative positions. Checks indicate conserved sites that were also shared by all assayed members of *P. maniculatus*. The approximate positions of coding sequences at the top are for cytochrome oxidase (CO) subunits, ATPase (A), cytochrome B (B), the D-loop origin of H-strand replication (D), and 16S and 12S rRNA, all determined by comparison with *Mus musculus* mtDNA (Bibb et al. 1981; see Lansman et al. 1983a).

For a preliminary identification of phenetic groupings in the data, a matrix of minimum mutation distances between representative mtDNA genotypes was subjected to a clustering analysis by the unweighted pair-group method with arithmetic means (UPGMA [Sneath and Sokal 1973]). The original presence-absence matrix of restriction sites was also used to generate an initially unrooted phylogeny by the Wagner parsimony method (Eck and Dayhoff 1966; Kluge and Farris 1969) implemented in computer programs kindly supplied by Joe Felsenstein. This parsimony method allows character-state changes in both directions (0 → 1 and 1 → 0), assumes no knowledge of ancestral states, and seeks to minimize total changes along the network.

**Results**

**mtDNA Variability within *polionotus* and *leucopus***

The eight restriction endonucleases used to assay the 82 samples of *Peromyscus polionotus* and *leucopus* revealed a grand total of 74 restriction sites (48 in each species), distributed along the mtDNA molecule as shown in figure 1. As was also true for the earlier data on *maniculatus*, all of the observed genetic changes can reasonably be attributed to base substitutions leading to loss or gain of particular restriction sites; additions, deletions, or rearrangements affecting more than about 50–150 base pairs (0.3–1.0 map unit or less) would usually have been detectable but were not observed. Thus variable restriction sites behave independently and are gained or lost without detectable alterations in the sizes of the restriction fragments in which they occur.

Considering restriction enzymes one at a time, we were able to link mtDNA restriction maps for conspecifics of *polionotus* and *leucopus* into relatively unambiguous minimum path networks which represent the probable pathways of genotypic interconversions during evolution (figs. 2, 3). For example, *polionotus*
FIG. 2.—Top, Locations of restriction sites in the three EcoRI maps (P, A, and B) observed in *Peromyscus polionotus*. Lines connecting maps indicate probable phylogenetic links; arrows indicate direction of loss of a restriction site (but do not necessarily imply direction of evolution). Bottom, Probable phylogenetic links among the maps produced by each of six other restriction endonucleases. For BglII, all *polionotus* samples exhibited an identical three-site map (fig. 1).

EcoRI patterns "B" and "P" differ from each other by two restriction site changes but are both only one site loss from "A" (fig. 2). Similarly, by parsimony criteria, the four EcoRI maps of *leucopus* can be linked in a unique fashion (fig. 3).

The composite mtDNA genotypes of *polionotus* and *leucopus*, which summarize the combined data for all restriction enzymes, are listed in tables 1 and 2, respectively. Estimated phylogenetic networks (Lansman et al. 1983a) were generated from these composite genotypes and superimposed over the geographic sources of collections, with results shown in figures 4 and 5. For *polionotus*, a predominant mtDNA clone (no. 1) was observed in 20 mice collected from five counties ranging from extreme eastern Georgia, through central and southwest Georgia, and into southeastern Alabama. This genotype was one mutation step removed from each of seven other mtDNA clones observed in adjacent geographic regions (nos. 3, 6, 7, 8, 9, 12, 17). MtDNA clone "17," the most common genotype in central Florida, was in turn one mutation step removed from each of four additional mtDNA clones (nos. 18–21). In figure 4, for purposes of graphical clarity, the network positions of a few mtDNA clones are not pictured. Clone 15 is linked by two mutation steps to clone 16, clone 22 is three steps from clone 21, and clone 10 is three steps from clone 6.

Especially for these closely related mtDNA clones, the particular links in the network should not be interpreted as definitive statements of matriarchal relationships for two reasons. First, we have utilized only eight restriction enzymes to sample an average of about 38 sites per mtDNA genome. Because of problems of sampling error, it is very likely that particular branching patterns would be altered with additional data; thus the given network might better be thought of as an estimation of observed mtDNA character-state relationships rather than of whole organismal relationships (Avise 1983). Second, even when interpreted as observed character-state relationships, several ambiguities due to homoplasy (the
FIG. 3.—Top, Locations of restriction sites in the four EcoRI maps (C, E, D, and F) observed in *Peromyscus leucopus*. Site "e" was mapped only to one of the two alternative positions indicated by the dashed lines. Lines connecting maps indicate probable phylogenetic links; arrows indicate direction of loss of a restriction site (but do not necessarily imply direction of evolution). *Bottom*, Probable phylogenetic links among the maps produced by each of six other restriction endonucleases. For *BglII*, all *leucopus* samples exhibited an identical two-site map (fig. 1).

FIG. 4.—Geographic distribution and estimated phylogenetic network of mtDNA clones in *Peromyscus polionotus*. Numbers refer to composite genotypes as in table 1. Branches connect clones that are related by a single mutation step; in aggregate they form a network based on minimum mutation distance. The placements of clones 10, 15, and 22 (which are not shown fully in the figure) are discussed in the text. Black dots indicate individual mice sampled.
convergent loss or gain of particular restriction sites) are apparent. For example, *BamHI* map "D," which differs from "E" by absence of a site at position 60.5, was shared by clones 19 and 5, which were otherwise only distantly related in the network (fig. 4). When genotypes from pairs of enzymes are considered jointly, it can be shown that no network alteration can eliminate this kind of homoplasy. Thus the *BamHI* genotypes "D" and "E," and the *HpaII* genotypes "X" and "Y," were observed in all possible combinations (DX, DY, EX, EY) in *polionotus*. No matter which of these clonal patterns is considered ancestral, and no matter what additional data from other enzymes might reveal, it is inevitable (barring recombination) that for the character conflict exemplified by *BamHI* and *HpaII* there has been evolutionary convergence to a common restriction morph at least once (LeQuesne 1969). At least five similar instances of convergence are apparent in our samples of *polionotus*. This finding of homoplasy in the mtDNA restriction site data for *polionotus* parallels our earlier findings of extensive restriction site convergence in *maniculatus* (Lansman et al. 1983a).

For our samples of *leucopus*, no single mtDNA genotype predominated in abundance or in geographic distribution (fig. 5). Most mtDNA clones interconnect by single mutation steps in the network, but three genotypes (nos. 10, 11, and 12) were several mutation steps from all others. Two of these were from western portions of the range of *leucopus* (Kansas and Texas), and the third was from Virginia. Baker et al. (1983) have recently described two well-defined chromo-
### Table 3
Estimates of Nucleotide Sequence Divergence between Representative mtDNA Clones in *Peromyscus polionotus*, *P. leucopus*, and *P. maniculatus*

<table>
<thead>
<tr>
<th>Clone Number</th>
<th><em>P. polionotus</em></th>
<th><em>P. maniculatus</em></th>
<th><em>P. leucopus</em></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 7 17 15 10</td>
<td>35 1 32 23 30</td>
<td>2 4 6 11 12</td>
</tr>
<tr>
<td>1</td>
<td>.002 .022 .019 .011</td>
<td>.037 .026 .043 .046 .060</td>
<td>.156 .142 .150 .126 .133</td>
</tr>
<tr>
<td>7</td>
<td>.004 .022 .013</td>
<td>.039 .022 .040 .043 .056</td>
<td>.158 .136 .151 .128 .126</td>
</tr>
<tr>
<td>17</td>
<td>.21 .021 .013</td>
<td>.039 .027 .044 .048 .056</td>
<td>.158 .145 .151 .128 .134</td>
</tr>
<tr>
<td>15</td>
<td>7 8 8 8 .007</td>
<td>.052 .033 .058 .054 .072</td>
<td>.155 .141 .147 .124 .140</td>
</tr>
<tr>
<td>10</td>
<td>4 5 5 3</td>
<td>.044 .029 .047 .047 .060</td>
<td>.151 .138 .144 .122 .137</td>
</tr>
<tr>
<td>35</td>
<td>14 15 15 19</td>
<td>.029 .033 .050 .049</td>
<td>.138 .133 .132 .118 .116</td>
</tr>
<tr>
<td></td>
<td>1 10 11 13 12</td>
<td>.038 .039 .046</td>
<td>.156 .121 .151 .130 .120</td>
</tr>
<tr>
<td>32</td>
<td>16 15 17 20 18</td>
<td>14 16</td>
<td>.059 .047</td>
</tr>
<tr>
<td>23</td>
<td>17 16 18 19</td>
<td>19 15 21 .023</td>
<td>.166 .144 .159 .151 .149</td>
</tr>
<tr>
<td>30</td>
<td>19 18 18 24</td>
<td>21 19 19 19 10</td>
<td>.167 .143 .159 .151 .150</td>
</tr>
<tr>
<td>2</td>
<td>45 44 46 44 44</td>
<td>45 42 47 46 49</td>
<td>.003 .005 .021 .020</td>
</tr>
<tr>
<td>4</td>
<td>44 43 45 42 43</td>
<td>44 41 46 45 48</td>
<td>1 .008 .010 .016</td>
</tr>
<tr>
<td>6</td>
<td>43 42 44 42 43</td>
<td>43 41 45 44 46</td>
<td>2 3 .020 .026</td>
</tr>
<tr>
<td>11</td>
<td>39 38 40 38 39</td>
<td>41 40 43 42 44</td>
<td>6 5 8 .013</td>
</tr>
<tr>
<td>12</td>
<td>41 40 42 42 43</td>
<td>41 39 43 45 46</td>
<td>8 7 10 6</td>
</tr>
</tbody>
</table>

**Note.**—Entry above diagonal: mean sequence divergence estimated from analyses of proportions of shared restriction sites; below diagonal: minimum mutation distances between clones. Margin numbers label the mtDNA clones as in tables 1 and 2 of the present report, and (for *maniculatus*) table 1 of Lansman et al. (1983a). For this matrix, the product-moment correlation between the distance measures is >0.99.

mtDNA Divergence among *maniculatus*, *polionotus*, and *leucopus*

Estimates of nucleotide sequence divergence between representative mtDNA genotypes of *maniculatus*, *polionotus*, and *leucopus* are presented in table 3. Distances derived from site analyses and minimum mutation steps are highly correlated for these data (product-moment correlation > 0.99). In site comparisons between *maniculatus* and *polionotus*, $p = 0.045$ (range 0.02–0.07). In comparisons of *leucopus* versus *maniculatus* and *polionotus*, the respective estimates of sequence divergence are as follows: $\bar{p} = 0.14$ (range 0.12–0.17), and $\bar{p} = 0.14$ (range...
FIG. 6.—UPGMA dendrogram for representative mtDNA clones of *Peromyscus polionotus*, *P. leucopus*, and *P. maniculatus* generated by cluster analysis of a matrix of minimum mutation distances between mtDNA genotypes. The five "assemblages" of *maniculatus* mtDNA clones are the same as those described in our earlier study (Lansman et al. 1983a). The mtDNA clones are numbered as in tables 1 and 2 of the present study and table 1 of Lansman et al. (1983a). In this and in fig. 7, possible trifurcations in the data (e.g., clones 1, 7, 17) are shown as bifurcations to indicate one of several equally parsimonious trees that could have been drawn.

0.12–0.16). All of these between-species estimates of sequence divergence are considerably lower (roughly one-third to one-half the magnitude) than our previously published estimates for these same species based on fewer samples (Avise et al. 1979). Presumably the differences are attributable to the larger cohort of enzymes included in the present study (eight vs. six), to the fact that only three of eight endonucleases were shared by the two studies, and to the greater reliability of current estimates due to precise site mapping. In any event, these results underline the need for caution in interpreting absolute levels of mtDNA sequence divergence based on information from small numbers of restriction enzymes (Li 1981; Nei and Tajima 1981).

A UPGMA phenogram for 36 representative mtDNA clones of *maniculatus*, *polionotus*, and *leucopus* is shown in figure 6. This particular phenogram was generated from a matrix of minimum-mutation distances (the matrix of $p$ values would cluster similarly since the distance measures were highly correlated). Five *maniculatus* clusters, representing the previously identified genotypic assemblages (Lansman et al. 1983a), are again apparent. All *polionotus* clones form a tight cluster, but this cluster falls well within a larger grouping formed by all *polionotus* and *maniculatus*. *Leucopus* samples also form a well-defined cluster which is phenetically distant from the *polionotus-maniculatus* grouping. On the average, in our survey about 43 mutation steps distinguish any *leucopus* sample from any *maniculatus* or *polionotus*. 
A Wagner network generated from the presence-absence restriction site matrix of representative *Peromyscus* clones is shown in figure 7. Results agree closely with those of the phenetic analysis: all five assemblages of *maniculatus* can be identified, *polionotus* genotypes form a branch of the network, the *polionotus* branch stems from a point within the larger network linking the *maniculatus*
assemblages, and *leucopus* also forms a distinct branch of the tree. The computer output of the Wagner analysis allows identification of those restriction sites which uniquely describe the structure of the tree. Most of these (with the exception of sites present in only one or a few closely related genotypes) are shown in figure 7. Thus 16 sites were present in all assayed mtDNA clones in *leucopus* but absent in all *maniculatus* and *polionotus*. Four sites were shared by all assayed members of *polionotus* and *maniculatus* but were universally absent in *leucopus*. A single site (*HpaII 58.4*) uniquely defines *polionotus*. In cladistic terminology, since *maniculatus* and the outgroup *leucopus* lack this site, presence of *HpaII 58.4* constitutes a synapomorphy (shared-derived state) defining *polionotus* as a clade. There were no restriction sites defining all *maniculatus* assemblages as a clade distinct from *polionotus*.

Nei and Li (1979) have developed a formula to estimate time of divergence between species when data on mtDNA polymorphism within species is also available (see also Ferris et al. 1981, eq. 1). The second term of Nei and Li's equation 25 is meant to correct for mtDNA divergence before species separation (Nei, personal communication). If we assume that *P. polionotus* is most closely related to S. Cal. *P. maniculatus* (fig. 7), and that mtDNA evolves at a rate of 2% base substitutions per million years (Brown et al. 1979), the corrected date for separation of *P. polionotus* from S. Cal. *P. maniculatus* becomes about 1.5 million years. We interpret this as a very provisional estimate.

**Discussion**

Although some specific models have been developed for the evolution of reproductive isolation through nucleo-cytoplasmic incompatibility (Watson and Caspari 1960; Wright 1969), little attention seems to have been paid to the more general problem of evolution of asexually transmitted characters within otherwise sexually reproducing species (Takahata and Maruyama 1981; Birky et al. 1983). For the evolution of strictly neutral, asexually transmitted characters, of what special significance (if any) is speciation? In particular, how might the distributions of such characters among living populations have been influenced by past speciation events? Some general conceptual possibilities are pictured in figure 8.
In hypothetical case I of figure 8, each reproductively defined species constitutes a monophyletic group (clade) with respect to matriarchal genealogy. All individuals within species A can be traced to a common female parent at point $p$, and all individuals within species B can be traced to a common female parent at $q$. If it is assumed that the female-transmitted traits being assayed (such as mtDNA) diverge at the same constant rate in all lineages, A and B would also appear as coherent clusters in a phenetic analysis. However, the mtDNA sequence diversity within B, much of which would have accumulated from a time vastly predating the onset of reproductive isolation, would be much higher than that within A, all of which would have accumulated from a time well after the speciation event. Also note that sequence divergence between A and B was initiated at point $r$ and not at the time of species splitting.

In hypothetical case II of figure 8, neither species constitutes a monophyletic group with respect to matriarchal genealogy. Various groupings of individuals (such as e-j) do form apparent clades, but these are not consonant with the recognized species boundaries. Depending on the particular comparison, sequence divergence among pairs of conspecifics could range from very low (i.e., individuals $h$ vs. $i$) to very high (i.e., individuals $i$ vs. $n$ which last shared a common female parent at point $s$). In comparisons between representatives of A and B, divergence would have been initiated at points ranging from $p$ to $s$, the times at which various pairs of individuals last shared female ancestors. In case III (fig. 8), species B constitutes a monophyletic group with respect to matriarchal genealogy, but species A does not (i.e., species A is paraphyletic [Farris 1974; Wiley 1981]). Expected patterns of divergence within versus between species can again be readily interpreted from the diagram.

The primary empirical concern of this report is the pattern and magnitude of mtDNA differentiation within and among populations of the closely related species *Peromyscus maniculatus* and *P. polionotus*. From both the Wagner and UPGMA analyses of the restriction site data, the pattern of mtDNA evolution (figs. 6, 7) conforms most closely to case III of figure 8. *Peromyscus polionotus* appears to form a monophyletic matriarchal assemblage (and a phenetic grouping) within the larger clade composed of *polionotus* and *maniculatus*. Thus the various assemblages of *maniculatus* appear to be paraphyletic with respect to *polionotus* in matriarchal ancestry. Which *maniculatus* assemblage is most closely related to *polionotus* is not clearly resolved from our data. However, the magnitude of mtDNA sequence divergence within *polionotus* is roughly comparable to that observed within any of the five geographic assemblages of *maniculatus*.

If our sample of mtDNA restriction sites provides a reliable assessment of overall matriarchal relationships for these species, it appears that populations now classified as *maniculatus* are paraphyletic and can trace ancestries to female parents predating the separation of *polionotus* from *maniculatus*. This scenario is probably not inconsistent with conventional thought. Blair (1950), Hooper (1968), Bowers et al. (1973), and Greenbaum et al. (1978) have all suggested that *maniculatus* represents the ancient, geographically central evolutionary stock from which *polionotus*, *melanotis*, *sikensis*, *sleveni*, and *sejugis* were budded off via peripheral isolation. If true, the greater total sequence diversity within *maniculatus* is probably expected, as is the potential monophyletic origin of each peripheral isolate.
The primary conceptual concern of this report is how the pattern and magnitude of differentiation in asexually transmitted characters might generally be related to speciation. With the very simple models we have introduced, at this point only a few general observations can be attempted. One prediction from these scenarios is that the initiation of mtDNA sequence divergence between living species could predate a speciation, but could not postdate it (fig. 8). This possibility may have important consequences for discussions of rates of molecular evolution. For example, suppose that samples from species A and B show 10% sequence divergence in mtDNA, and from fossil or geographic evidence are known to have speciated 5 million years ago. Even barring potential problems of character-state reversals or convergence, it still does not necessarily follow that rate of divergence is 2% per million years. Depending on the particular (and in practice unknown) matriarchal genealogies involved, rate of sequence divergence could be lower than this simply because of the pattern of lineage survivorship.

This potential error is in some respects similar to previously recognized sources of error in rate estimates from paralogous versus orthologous protein comparisons (Wilson et al. 1977). In the conventional view of a paralogous comparison of proteins (such as the α and β chains of hemoglobin), any apparent distance between species would have accumulated since the time of gene duplication, which could only be as old as or older than the speciation. (This neglects recently discovered possibilities of “concerted” evolution [Coen et al. 1982].) Similarly, in comparisons of strictly maternally transmitted characters, any apparent distance between species would have accumulated since the time of last common female ancestor, which might also be older than the time of speciation. Recent empirical studies have suggested that mtDNA evolves very rapidly—perhaps five to 10 times more rapidly than single-copy nuclear DNA (Brown et al. 1979; Brown 1983). If the female ancestries of the species used for these comparisons commonly trace to times predating the speciations used in rate calibrations, the absolute rates of mtDNA evolution could be inflated for this reason (however, there are also independent mechanistic reasons for supposing that mtDNA has evolved especially rapidly [Brown 1983]).

One way that speciation could have immediate consequences for the pattern of mtDNA evolution is if each speciation is accompanied by a severe population reduction followed by long-term population expansion. With most or all mtDNA sequence diversity lost at the bottleneck, the speciation event would effectively “reset” the molecular clock; subsequently arising sequence diversity within a species could be calibrated to the time of speciation. For example, if polionotus went through a severe population reduction at time of separation from maniculatus perhaps 1–2 million years ago, rate of mtDNA sequence divergence within polionotus could be roughly 2% per million years (this is generally comparable to published rates for mtDNA divergence in other mammals [Brown et al. 1979; Ferris et al. 1981]). However, populations certainly may experience bottlenecks at times not associated with speciation. Thus the initiation of sequence divergence within a species could also predate or postdate the speciation process. Furthermore, lineage sorting can be very rapid in stable populations even in the absence of bottlenecks (Avise et al., in preparation).

Another way that past speciations will likely influence patterns of mtDNA relationships among extant species is through the necessary partial correlation between matriarchal heredity and total nuclear genome heredity. The evolutionary
demographics of female lineages have immediate effects on the patterns of evolution of mtDNA, but those females which survive and reproduce contribute to the nuclear as well as the mtDNA gene pools of future generations. The strength of the correlation of mtDNA genotype with nuclear genotype in a given situation will likely depend on mating structure, population size and subdivision, population history, and other related demographic variables. Finally, as already noted, there could be selection pressures causing partial coevolution of nuclear DNA and mtDNA due to functional interactions of their transcribed products.

Clearly, a great deal remains to be learned about the relationship of the speciation process to the evolution of uniparentally inherited traits such as mtDNA. Progress should proceed along two fronts. Empirically, many more examples are needed of the magnitude and pattern of mtDNA differentiation within and among closely related species. Perhaps examples of cases I and II (fig. 8) will be found to complement the example of case III presented here. How often will species boundaries defined by reproductive isolation be concordant with phenetic or cladistic boundaries identified in analyses of mtDNA genotype? Within what limits will conspecific individuals be found to differ in mtDNA genetic composition, and how might these values be related to the evolutionary dynamics of female lineages in particular cases? More empirical work with higher animals is also needed to verify the supposed lack of paternal mtDNA input to progeny; to examine the possibility of recombination or conversion among mtDNA’s; to understand how these possibilities, if realized, might affect the conclusions of this and similar studies; and generally to understand the molecular and cellular dynamics of mtDNA variation (Chapman et al. 1982; Lansman et al. 1983a).

On the conceptual front, more comprehensive and formal theoretical models relating the dynamics of matriarchal lineages to speciation are required. Under what demographic conditions can concordances between reproductively defined species and matriarchally defined assemblages be anticipated? How often and by how much time might the shared female ancestors of representatives of extant lineages antedate the time of origin of reproductive isolation? How might the evolutionary dynamics of female lineages lead to major genetic gaps in mtDNA sequence among populations (such as those observed in maniculatus)? Answers to these and related questions should tell a great deal about the demographics of speciation.

Acknowledgments

We are deeply grateful to Joe Felsenstein for providing computer programs for the Wagner analysis and to Chris Meacham for helping to implement these programs at the University of Georgia. Work was supported by National Science Foundation grants DEB-8022135 and BSR-8217291.

LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received February 28, 1983; revision received May 12, 1983.
Evolution of Antibiotic Resistance Genes: The DNA Sequence of a Kanamycin Resistance Gene from Staphylococcus Aureus

Gary S. Gray
Biogen S.A., Switzerland

Walter M. Fitch
University of Wisconsin—Madison

The kanamycin resistance gene from Staphylococcus aureus has been sequenced and its structure compared with similar genes isolated from Streptomyces fraiae and from two transposons, Tn5 and Tn903, originally isolated from Klebsiella pneumoniae and Salmonella typhimurium, respectively. The genes are all homologous but, since their common ancestor, have undergone extensive divergence, with more than 43% divergence between the closest pair. The phylogeny of the genes cannot be made congruent to the phylogeny of the taxa from which they were isolated without requiring rather improbable differences in rates. One is therefore led to conclude that there have been multiple occurrences of gene transfer between these species. Thus, although they are homologous, they are neither orthologous nor paralogous. It is suggested that homologous genes of this type be called xenologous.

Introduction

Staphylococcus aureus, resistant to aminocyclitol antibiotics, produces a variety of mechanistically different aminocyclitol-modifying enzymes including the aminocyclitol-3'-phosphotransferase (APH[3']-III). Similar enzymes have also been described in the gram-positive genus Streptococcus (Courvalin et al. 1980a), the Enterobacteriaceae (Smith 1978) and Pseudomonadaceae (Matsuhashi et al. 1973, aminocyclitol-producing Bacillus circulans (Courvalin et al. 1977), and actinomycetes (Davies et al. 1979; Thompson et al. 1980). The APH(3') activities from various sources all catalyze the phosphorylation of neomycin and kanamycin and can be subgrouped according to their reactivity with other aminocyclitol antibiotics (Davies and Smith 1978). Although the APH(3') genes and their gene products are functionally related, studies employing both immunochemical methods and DNA hybridization have shown only the staphylococcal and streptococcal genes to be related (Davies and Smith 1978; Smith 1978; Courvalin et al. 1980a). However, comparison of the DNA sequences of the APH(3') genes from trans-
posons Tn5 and Tn903 of gram-negative bacteria and from *Streptomyces fradiae* revealed marked similarity (Thompson and Gray 1983). It has been hypothesized that an aminocyclitol-producing bacterium was the source of the genes for the widely disseminated aminocyclitol phosphotransferases (Benveniste and Davies 1973). To extend the analyses of genes encoding aminocyclitol-modifying enzymes and to test that hypothesis, we report here the DNA sequence of the APH(3')-III gene encoded by an *S. aureus* plasmid and present a detailed comparison of the four known APH(3') gene sequences. These results confirm the close structural relationship among five phosphotransferase genes. They also show that the hypothesis about the source of the gene in resistant strains is correct.

**Material and Methods**

The aminocyclitol-3'-phosphotransferase gene from *Staphylococcus aureus* plasmid pSH2 was cloned in *Escherichia coli* (Courvalin and Fiandt 1980) and subsequently subcloned after partial digestion with HindIII. The resulting recombinant plasmid, pAT48, confers kanamycin resistance to *E. coli* and possessed 3.5 kbp of inserted DNA which contained two internal Hind111 sites. An attempt to subclone the large insert-specific HindIII fragment from pAT48 by selection of transformants for kanamycin resistance was unsuccessful, suggesting that a DNA sequence surrounding at least one of the two Hind111 sites was essential for transcriptional or translational activity. Therefore, DNA sequencing was initiated at the ends of this large HindIII fragment. The DNA sequence around one of the HindIII sites proved to be similar to that found for other APH (3') genes (Thompson and Gray 1983); subsequently, adjacent DNA sequences were determined by the method of Maxam and Gilbert (1980) according to the strategy in figure 1.

**Results**

Figure 2 shows the primary structure of the APH(3') gene of *Staphylococcus aureus* and the predicted protein sequence for the open reading frame of 789 base pairs (fig. 2, residues 293–1082). Within the coding region, there were 62 translation termination codons distributed among the other five possible reading frames. The translation initiation site of the staphylococcal APH(3') gene was assigned to the methionine codon at residue 293; initiation at this codon would yield a protein of MW = 30,724 (MW = molecular weight, in daltons). This estimate agrees with the value found for purified APH(3') protein (MW = 29,000 ± 300) (Smith 1978) and that determined for the insert-specific protein (MW = 31,000) obtained after coupled in vitro transcription and translation of pAT48 DNA. Inspection of the DNA sequences immediately preceding the proposed initiation codon reveals a ribosome binding site (GGAAAGG, residues 277–282) and transcription initiation sequences (residues 230–260 or 210–240) similar to those found for other staphylococcal genes (Horinouchi and Weisblum 1982a, 1982b). The region following the translational stop codon (residue 1033) contains an inverted repeated (IR) sequence of 19 bp which presumably functions in RNA transcription termination (Rosenberg and Court 1979); similar sequences occur after the staphylococcal chloramphenicol (Horinouchi and Weisblum 1982a) and erythromycin resistance genes (Horinouchi and Weisblum 1982b) and many gram-negative sequences (Rosenberg and Court 1979). The two HindIII sites located in the IR sequence (residues 1109–1122) mark the end of the large HindIII fragment; their location suggests that the failure to obtain kanamycin resistant clones when sub-
The proposed amino acid sequence for the *S. aureus* APH(3')-III gene and those predicted for the resistance genes of transposons Tn5 and Tn903 and the neomycin-producing *Streptomyces fradiae* are compared in figure 3. The amino acid sequences exhibit regions of extensive similarity, with the carboxyl terminal segments being the most similar. A conserved cysteine residue at position 145 is flanked by two regions (residues 115–175) containing a high frequency of acidic amino acids which may function in the binding of the basic aminocyclitol antibiotics to the modifying enzyme.

That the sequences are very divergent is shown by counting the number of positions with different numbers of nucleotides in them. They all may be different (1:1:1:1), or may have only one pair (2:1:1), two pairs (2:2), a triplet (3:1), or a quartet (4) of nucleotides alike. These cases number 26, 270, 103, 246, and 213, respectively, and include the termination codon. A gapped nucleotide was counted as identical to the most frequent nucleotide in this count so that the differences reflect nucleotide substitutions exclusive of insertions or deletions and are therefore conservative with respect to total divergence. There are 21 distinguishable gaps comprising 39 positions. Less than 25% of the positions have no nucleotide differences, while more than 42% have three or more different nucleotides.
Fig. 2—DNA sequence of the aminocyclitol-3'-phosphotransferase gene of *Staphylococcus aureus*. The sequence presented represents the sense strand of DNA commencing at the 5' end with the HinfI site. The nucleotide residue corresponding to the proposed initiation codon is located at residue 293. Pregene sequences include a Shine and Dalgarno (1974) sequence (S.D.; residue 280) and two possible promoter sequences (−10, −35; residue 210–240 and 230–260). The predicted amino acid sequence is shown below the DNA sequence and the location of translational stop codons indicated (***). Arrows indicate the block of inverted repeated symmetric sequences immediately following the predicted coding region.
The phylogeny was investigated using a parsimony analysis (Fitch 1971), for which only the 103 2:2 cases are informative and, for any one nucleotide position, may be assigned to one of three ordered sets, MNMN = 48, MMNN = 31, and MNNM = 24, where M and N represent the two kinds of nucleotides in the same sequence order as in figure 3 and where 48 + 31 + 24 = 103 accounts for all the

Fig. 3.—Comparison of the amino acid sequences of aminocyclitol 3'-phosphotransferases of *Staphylococcus aureus*, *Streptomyces fradiae* (Thompson and Gray 1983), and gram-negative transposons Tn5 (Beck et al. 1982) and Tn903 (Oka et al. 1981). Sequence locations containing identical amino acid residues are enclosed with solid lines, and sites of conservative amino acid replacements are enclosed within dotted lines. Conservative means within the groups [D,E], [K,R], [S,T], [F,Y], and [I,L,V,M]. Gaps have been inserted to improve the sequence alignment. The one-letter amino acid codes are A = ala, C = cys, D = asp, E = glu, F = phe, G = gly, H = his, I = ileu, K = lys, L = leu, M = met, N = asn, P = pro, Q = gin, R = arg, S = ser, T = thr, V = val, W = trp, and Y = tyr.
cases. There are only three possible unrooted tree structures, and a tree requiring only one nucleotide substitution for one of these three sets will require two substitutions in each of the other two trees. Thus the three possible unrooted topologies for these crucial positions will require $48 + 2(24 + 31) = 158$, $31 + 2(48 + 24) = 175$, and $24 + 2(48 + 31) = 182$ nucleotide substitutions. To each of these must be added 864 additional substitutions for the uniquely derived nucleotides. These trees are shown in their respective order at the top of figure 4.

Of the 1,022 nucleotide substitutions in the left tree, 242 do not change the encoded amino acid. The number of nucleotide substitutions in the first, second, and third codon positions is 313, 260, and 449, respectively. The ratio of transversions to transitions was 666/356 overall and not greatly different by codon position.

**Discussion**

The alignment in figure 3 shows sufficient amino acid identities and similarities that one can hardly doubt that the sequences are homologous (i.e., have a common ancestor). In this analysis, codon gaps were constrained to be in common where possible and to include only whole codons. The *Streptomyces fradiae* DNA se-
quence contained an additional 12 nucleotides at the 5’ end because of the possibility of translational initiation at either of the methionine codons located here.

While parsimony clearly favors tree I (fig. 4), the total amount of change is so great as to make any definite conclusion uncertain. One is no better off using a method based on overall amounts of change (e.g., Fitch and Margoliash 1967) since any pair of sequences is about as divergent as any other pair (table 1, range 397–454 base differences), although again tree I is marginally preferred over tree II. If the nucleotide sequences are ignored and one resorts to the amino acid sequences, tree III is preferred on the basis of minimum base differences (see table 1) when an average linkage method such as that of Fitch and Margoliash (1967) is used.

Since two of the sequences are encoded by transposable elements, it is probably not meaningful to ask if tree I agrees with the presumed phylogeny of the four taxa that are host to these sequences. It is instructive, however, to inquire about the relation between the two since any discrepancy between them would be accountable by, and expected on the basis of, a plasmid-mediated acquisition by its host of another taxon’s defense against aminocyclitol antibiotics. For this purpose, we assume that the first maker of the antibiotic had to have the enzymatic defense mechanism to protect itself from its own weaponry. We shall assume for the moment that the *Streptomyces fradiae* lineage serves that role since it is the only taxon present that makes aminocyclitols. This origination is represented by the small dotted lines uppermost on trees IV and V in figure 4 where topologies I and II have been superimposed (dashed lines) on the presumed phylogeny of the four taxa (solid lines).

The implication of tree IV is that all three of the other lineages independently acquired their gene copies from the *fradiae* line. This can be reduced to two acquisitions in tree V by having the Klebsiella(Tn5)-Salmonella(Tn903) lineage acquire the gene (transposon) prior to the divergence of those two genera. The principal argument in favor of this version is that by splitting the 134 substitutions into 75 and 59 on the upper left and right dashed lines of descent, one obtains a solution with a magnificently uniform rate of descent in all four lines.

However, since the two topologies (species and gene) are identical, no independent gene acquisitions need to be postulated, as can be seen by stretching the upper bifurcation point on the *fradiae* line upward to the apex representing the common ancestor of all four taxa until the branch points of both trees are superimposed. The principal argument against this is the correspondingly large distortion of uniformity of rate of change with time as more than 40% of all the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Base Differences for APH(3’) Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TN5</td>
</tr>
<tr>
<td>TN5 . . . . . .</td>
<td>...</td>
</tr>
<tr>
<td>Sfr . . . . . .</td>
<td>397</td>
</tr>
<tr>
<td>903 . . . . . .</td>
<td>421</td>
</tr>
<tr>
<td>Sau . . . . . .</td>
<td>422</td>
</tr>
</tbody>
</table>

NOTE.—The portion of the table above the diagonal shows the minimum base differences as if only the amino acid sequences in fig. 3 were known; the portion below the diagonal shows the actual base differences observed when the nucleotide sequences are examined in the same coding alignment.
required changes must be incorporated in the relatively small amount of evolutionary time (crudely, 20%) since the common ancestor of *Klebsiella* and *Salmonella* compared with the sum of the other evolutionary times on the tree. Of course, being resident on a transposon may somehow contribute the necessary additional increase in rate. But there is no reason to stretch the point to avoid independent gene acquisitions since they cannot be avoided. Courvalin (personal communication) has sequenced this gene isolated from a transposon from *Streptococcus faecalis*, and it is identical to that of *Staphylococcus aureus* reported here. These two species are much too distantly related for at least one of these two genes not to have been recently acquired by transfection. A similar situation would appear to have occurred in the histones of sea urchins (Busslinger et al. 1982), except that the responsible vector is unknown in their case.

The trees in figure 4 make it clear once again that “homology is not enough.” That is to say, any attempt to use homologous sequences to infer the phylogeny of the taxa is in danger of error if the sequences are not the right type of homology (Fitch 1970). If the sequences diverged because the taxa containing the gene diverged, this is the subcategory of homology called orthology, and a good species phylogeny requires orthologous sequences. If the sequences diverged following a gene duplication, this is the subcategory of homology called paralogy, and mixing paralogous sequences such as the alpha and beta hemoglobins (one or the other for each taxon) would be a good method for getting a bad species phylogeny. Clearly transfection (and symbiosis and parasitism as well) represents a way that cells and organisms have acquired foreign genes in the past, and, since they are neither orthologous nor paralogous but are clearly homologous (sensu strictu), perhaps they should be called xenologous (xeno = foreign).

Finally, it should be noted that it would be premature to assume that these genes originated in the *Streptomyces* line. Since *Bacillus circulans* also produces an aminocyclitol antibiotic, one might as easily believe, on the limited evidence available, that these genes first appeared in the *Bacillus* lineage which, phylogenetically, would form a branch off the *Staphylococcus* line in figure 4 (trees VI and VII). Remarkably, shifting the origin to the latter lineage would have little effect on the preceding discussion. The double trees of figure 4 VI and VII differ from IV and V only in that shift. Hypotheses VI and VII do, however, raise a new question, namely, Why don’t staphylococcal species make aminocyclitols if their ancestors did? One is still left with a narrow choice between tree I and tree II on the grounds of parsimony versus uniform rates, neither of which is a given principle of nature (tree III is the worst on both grounds). Nor can one dismiss the possibility that different lineages acquired their defenses from different offensive lineages such as other aminocyclitol producing actinomycetes. The homology of these sequences is clear, as is their transfection, and one cannot help being tantalized by the prospect and potential of further sequence information to resolve the question of their origins.

*Note added in proof.*—C. J. Herbert, I. G. Giles, and M. Akhtar (FEBS Letters 160:67–71, 1983) have now sequenced the APH gene from *Bacillus circulans*. It is homologous to the genes in this paper and has a range of distances to these genes similar to that which these genes have among themselves. The *B. circulans* gene is marginally closer to the *Staphylococcus aureus* gene than to the others. This seems to imply that either the *Streptomyces* or the *Bacillus* lineage (probably the latter) acquired this defense by transfection relatively recently in its history.
and raises the question whether its offensive weaponry might not also have been obtained by transfection.

Acknowledgments

We thank Julian Davies for helpful discussions and a critical review of this manuscript and Sylvia Robertson and Ron Niece for technical assistance in the computing. This investigation was supported by Biogen S.A. and National Science Foundation grant DEB-78-14197.

LITERATURE CITED


ROY J. BRITTON, reviewing editor

Revision received September 2, 1983.
Geographic variation in the genetic structure of natural enteric populations of *Escherichia coli* was assessed at both the single-locus and dilocus levels from allozyme genotypes at 12 enzyme loci in 178 cell lines isolated from human hosts in Sweden, Iowa, and Tonga. Although there was significant heterogeneity in allele frequencies at six of the 12 loci, geographic variation accounted for only 2.0% of the total genetic diversity ($H_T = 0.518$). Ohta's $D$-statistics were used to partition the total variance of dilocus linkage disequilibrium into within-population and between-population components. The observed total variance in disequilibrium (0.0339), averaged over 66 locus-pairs, was significantly greater than would be expected (0.0103) if alleles were randomly associated in an unstructured total population; and both within-locality and between-locality components made substantial contributions to the total variance. Half the locus-pairs exhibited the specific dual relationship among components expected when random factors are generating disequilibrium, but 20% of the locus-pairs showed the opposite relationship, reflecting systematic allele associations. The magnitude of dilocus disequilibrium apparently is unrelated to the chromosomal distance between loci. This and other evidence indicates that substitutive recombination rates in natural populations are sufficiently low to permit indirect periodic selection to play a prominent role in generating multilocus genetic structure.

Introduction

Nonrandom associations of alleles in haploid genotypes (linkage disequilibria) are often attributed to the direct action of epistatic natural selection on the loci involved (Bodmer 1979; Klein and Figueroa 1981; and earlier examples in Hedrick et al. 1978). However, it is well known that linkage disequilibrium can also arise without epistasis as a result of random genetic drift within single populations (Hill and Robertson 1968; Ohta and Kimura 1969a, 1969b; Hill 1975, 1976) and sub-

1. Key words: allozymes, geographic variation, multilocus genetic structure, random genetic drift, periodic selection.

Address for correspondence and reprints: Thomas S. Whittam, Department of Biological Sciences, Rutgers University, Piscataway, New Jersey 08854.
divided populations (Nei and Li 1973; Li and Nei 1974; Feldman and Christiansen 1975; Ohta 1982a, 1982b) and by founder effects (Avery and Hill 1979). Furthermore, natural selection acting on variation at one locus can indirectly affect associations between neutral alleles either at closely linked loci through "hitchhiking" (Thomson 1977; Hedrick 1980, 1982; Asmussen and Clegg 1981) or, in the case of clonally structured populations, at all loci in the genome as a result of the "dynasty effect" of periodic selection (Koch 1974; Kubitschek 1974; Levin 1981). With several potential sources of disequilibrium, the evolutionary interpretation of the adaptive nature of particular multilocus associations requires a cautious approach.

To study the contributions of epistatic natural selection and population subdivision to observed disequilibrium, population geneticists have recently developed methods for partitioning multilocus associations into within- and between-population components (Brown and Feldman 1981; Ohta 1982a, 1982b). One objective of this approach is to detect systematic associations among alleles in isolated populations of a species—a condition that, with certain reservations (Hedrick et al. 1978), may be taken as circumstantial evidence of the direct action of natural selection on the loci involved (Lewontin 1974, p. 315). For systematic associations, there is a relatively large within-population component and a relatively small between-population component, because disequilibrium is in the same direction in each population. In contrast, a large between-population component of disequilibrium is most readily attributable to nonselective effects of population subdivision or founder effects (Brown and Feldman 1981; Ohta 1982a, 1982b).

Recent studies of allozymic variation in *Escherichia coli* have revealed that the multilocus genotypes observed in natural populations represent a highly nonrandom subset of a vast array of possible genetic combinations (Selander and Levin 1980; Caugant et al. 1981, 1983; Ochman et al. 1983; Whittam et al. 1983). In an earlier analysis of multilocus structure in *E. coli* (Whittam et al. 1983), we demonstrated that many alleles are in linkage disequilibrium in the total population and that the species is subdivided into three overlapping groups of strains characterized by distinctive combinations of alleles. However, in the absence of direct information on rates of migration, recombination, and extinction in natural populations, it is difficult to assess the contributions of natural selection and nonselective forces to the generation and maintenance of the multilocus genetic population structure of this species.

In an effort to determine the evolutionary sources of linkage disequilibrium in *E. coli*, we have analyzed polymorphism at 12 enzyme loci in samples from three localities within a framework provided by the following questions: (1) What is the degree of population subdivision and genetic differentiation between widely separated geographic populations? (2) To what extent can the effects of population subdivision and genetic drift account for the multilocus structure? (3) Is the strength of the statistical association between alleles within populations related to the proximity of loci on the chromosome and, hence, to the potential rate of recombination between them? We begin with a standard analysis of allele frequencies, then apply Ohta's (1982a, 1982b) method of partitioning the total variance of dilocus linkage disequilibrium into several components, and, finally, compare the strength of observed allele associations with the chromosomal map distances between pairs of loci.
Material and Methods

The analysis is based on a total of 178 isolates of *Escherichia coli*, including 30 from Iowa, the United States; 100 from Göteborg, Sweden; and 48 from Tongatapu, Tonga. Each isolate was obtained from a fecal sample of a separate individual human host. The isolates from the United States and Sweden were selected from a large collection of samples used in earlier studies of genotypic diversity in natural populations of *E. coli* (Selander and Levin 1980; Caugant et al. 1983; Ochman et al. 1983; Whittam et al. 1983).

Protein extracts from single-cell isolates were prepared, subjected to starch-gel electrophoresis (as described by Selander et al. 1971; Caugant et al. 1981), and selectively stained for the 12 enzymes listed in tables 1 and 2. For each enzyme, we distinguished from three to 20 allozymes, which were equated with alleles. Isolates lacking activity for a particular enzyme were designated as “null” at that locus. Thus, the combination of alleles observed at 12 structural gene loci characterizes the multilocus genotype of each isolate.

The sampling employed in this study differed from that used in an earlier study of multilocus structure in *E. coli* (Whittam et al. 1983). In the earlier study, both allele frequencies and measures of disequilibrium were calculated over the set of unique genotypes (multilocus electrophoretic types), and the relative abundance of each type in the total sample of isolates was not considered. This procedure was necessary because data were combined from several surveys in which the relative frequencies of genotypes in the total sample of isolates were strongly influenced by sampling effort (i.e., the numbers of isolates obtained from individual hosts) and, therefore, probably did not reflect their actual abundance in nature. In the present study, however, we have standardized the sampling effort by using only one isolate per host; calculations of allele frequencies and disequilibrium coefficients incorporate estimates of the relative frequencies of the various multilocus genotypes in the natural populations sampled.

Results

Genetic Diversity and Geographic Variation

In the sample of 178 isolates, 103 allelic states (allozyme alleles and null alleles) were distinguished over the 12 enzyme loci assayed. All loci were polymorphic in the three populations, except for GOT and G6PD, which were monomorphic in Iowa (table 1). The number of alleles in the total population averaged 8.6 per locus, ranging from 3 at GOT to 21 at 6PGD. Table 2 presents the total allelic diversity (Nei 1975) for each locus. The average total diversity \(H_T = 0.518\) is nearly identical to that obtained earlier for 279 electrophoretic types isolated from both human and other animal sources (Whittam et al. 1983).

Table 2 also presents \(F_{ST}\), the standardized variance in allele frequency among localities, for each locus. Six loci (MDH, 6PGD, G6PD, GOT, AK, and ACO) have relatively small \(F_{ST}\) values, indicating essentially no variation among localities. At each of the remaining six loci, which have relatively large \(F_{ST}\) values, one or more alleles vary significantly among localities (heterogeneity \(\chi^2\) test; table 1). The average \(F_{ST}\) over loci is 0.020, which means that 98% of the total genetic diversity can be found within a geographic population and only 2% can be accounted for by variation in allele frequencies (Nei 1977). Although the amount of
### Table 1

Allele Frequencies at 12 Enzyme Loci in Three Geographic Populations

<table>
<thead>
<tr>
<th>Locus and Allele</th>
<th>Locality</th>
<th>Locality</th>
<th>Locality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Iowa (n=30)</td>
<td>Sweden (n=100)</td>
<td>Tonga (n=48)</td>
</tr>
<tr>
<td>ADH:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.10</td>
<td>.25</td>
<td>.17</td>
</tr>
<tr>
<td>1</td>
<td>.27</td>
<td>.25</td>
<td>.33</td>
</tr>
<tr>
<td>2</td>
<td>.33</td>
<td>.33</td>
<td>.46</td>
</tr>
<tr>
<td>3</td>
<td>.63</td>
<td>.47</td>
<td>.46</td>
</tr>
<tr>
<td>6</td>
<td>.33</td>
<td>.17</td>
<td>.46</td>
</tr>
<tr>
<td>7</td>
<td>.23</td>
<td>.13</td>
<td>.15</td>
</tr>
<tr>
<td>9</td>
<td>.23</td>
<td>.13</td>
<td>.15</td>
</tr>
<tr>
<td>10</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>11</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>12</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>13</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>14</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>15</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>16</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>17</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>18</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>19</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>20</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>21</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>22</td>
<td>.03</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>G6PD:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.03</td>
<td>.03</td>
<td>.03</td>
</tr>
<tr>
<td>2</td>
<td>.04</td>
<td>.04</td>
<td>.04</td>
</tr>
<tr>
<td>3</td>
<td>.06</td>
<td>.06</td>
<td>.06</td>
</tr>
<tr>
<td>4</td>
<td>.08</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>5</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>6</td>
<td>.12</td>
<td>.12</td>
<td>.12</td>
</tr>
<tr>
<td>8</td>
<td>.16</td>
<td>.16</td>
<td>.16</td>
</tr>
<tr>
<td>9</td>
<td>.18</td>
<td>.18</td>
<td>.18</td>
</tr>
<tr>
<td>10</td>
<td>.20</td>
<td>.20</td>
<td>.20</td>
</tr>
<tr>
<td>11</td>
<td>.22</td>
<td>.22</td>
<td>.22</td>
</tr>
<tr>
<td>12</td>
<td>.24</td>
<td>.24</td>
<td>.24</td>
</tr>
<tr>
<td>14</td>
<td>.28</td>
<td>.28</td>
<td>.28</td>
</tr>
<tr>
<td>15</td>
<td>.30</td>
<td>.30</td>
<td>.30</td>
</tr>
<tr>
<td>16</td>
<td>.32</td>
<td>.32</td>
<td>.32</td>
</tr>
<tr>
<td>17</td>
<td>.34</td>
<td>.34</td>
<td>.34</td>
</tr>
<tr>
<td>18</td>
<td>.36</td>
<td>.36</td>
<td>.36</td>
</tr>
<tr>
<td>19</td>
<td>.38</td>
<td>.38</td>
<td>.38</td>
</tr>
<tr>
<td>20</td>
<td>.40</td>
<td>.40</td>
<td>.40</td>
</tr>
<tr>
<td>21</td>
<td>.42</td>
<td>.42</td>
<td>.42</td>
</tr>
<tr>
<td>22</td>
<td>.44</td>
<td>.44</td>
<td>.44</td>
</tr>
</tbody>
</table>

**Note.** See Table 2 for names of enzyme loci.

* Allele designations are cognate with those of earlier studies (Ochman et al. 1983; Whittam et al. 1983). Zero denotes null phenotype, that is, no detectable enzyme activity at the locus. Leading alleles are italicized. \( n \) = number of isolates.

* Significant heterogeneity in frequency among localities is indicated by \( \chi^2_{10} > 5.99, P < .05 \).

** Significant heterogeneity in frequency among localities is indicated by \( \chi^2_{10} > 9.21, P < .01 \).
alleles frequency variation among localities is small relative to the total genetic diversity, the absolute degree of differentiation among populations is substantial (minimum genetic distance, $D_u = 0.028$), being equivalent to that reported for several animal species (Nei 1975, p. 152).

Variance Components of Linkage Disequilibrium

In a theoretical analysis of linkage disequilibrium in a finite population with a subdivided structure, Ohta (1982a, 1982b) devised a method of partitioning the variance of di-locus disequilibrium coefficients in a manner analogous to Wright's $F$-statistics for allele frequencies (e.g., Wright 1978, pp. 86–89). Definitions of Ohta's five coefficients of disequilibrium for a subdivided population are presented in table 3. For a pair of loci, these $D$-statistics partition the total variance of disequilibrium into two sets of within- and between-population components. In our application of this method, the sampling properties of these statistics were ignored and the empirical values were used as direct estimates of the population parameters.

From the definition in table 3, the total variance of disequilibrium, $D_{ij}$, was calculated for each pair of loci. For locus-pair $A$ and $B$, $D_{ij}$ is the average squared deviation over all allelic pairs of the observed frequency of chromosome $A,B$ in the $k$th population from the expected frequency calculated as the product of average allele frequencies. The frequency distribution of the total variance for 66 pairs of loci is presented in figure 1. The average total variance is 0.0339. For comparative purposes, we randomized the allele combinations in 178 hypothetical isolates 50 times by computer, each time calculating 66 pairwise sets of $D$-statistics (table 4). Of the 3,300 randomly generated values of $D^2$, 98% fell below 0.0200,
Table 2
Total Genetic Diversity, $H_T$, and Relative Differentiation, $F_{ST}$, at 12 Enzyme Loci Assayed in Three Geographic Populations of Escherichia coli

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>E.C. Symbol</th>
<th>Number</th>
<th>$H_T$</th>
<th>$F_{ST}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol dehydrogenase</td>
<td>ADH</td>
<td>1.1.1.1</td>
<td>.672</td>
<td>.055</td>
</tr>
<tr>
<td>Malate dehydrogenase</td>
<td>MDH</td>
<td>1.1.1.37</td>
<td>.304</td>
<td>.000</td>
</tr>
<tr>
<td>Isocitrate dehydrogenase 6-phosphogluconate</td>
<td>IDH</td>
<td>1.1.1.41</td>
<td>.446</td>
<td>.071</td>
</tr>
<tr>
<td>Glutamic oxaloacetic transaminase</td>
<td>GOT</td>
<td>2.6.1.1</td>
<td>.044</td>
<td>.003</td>
</tr>
<tr>
<td>Adenylate kinase</td>
<td>AK</td>
<td>2.7.4.3</td>
<td>.451</td>
<td>.009</td>
</tr>
<tr>
<td>β-galactosidase</td>
<td>βGA</td>
<td>3.2.1.23</td>
<td>.888</td>
<td>.031</td>
</tr>
<tr>
<td>Phenylalanyl-leucine peptidase</td>
<td>PEP2</td>
<td>3.4.11.2</td>
<td>.611</td>
<td>.016</td>
</tr>
<tr>
<td>Aconitase</td>
<td>ACO</td>
<td>4.2.1.3</td>
<td>.665</td>
<td>.002</td>
</tr>
<tr>
<td>Phosphoglucose isomerase</td>
<td>PGI</td>
<td>5.3.1.1</td>
<td>.596</td>
<td>.021</td>
</tr>
<tr>
<td>Mannose phosphate isomerase</td>
<td>MPI</td>
<td>5.3.1.8</td>
<td>.715</td>
<td>.010</td>
</tr>
</tbody>
</table>

* Corrected for sampling variance (Wright 1978, pp. 86–89).

Table 3
Variance Components of Linkage Disequilibrium

<table>
<thead>
<tr>
<th>Level</th>
<th>Deviation</th>
<th>Variance Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>$g_{ij} - \bar{g}i\bar{g}j$</td>
<td>$D_{ij}$</td>
</tr>
<tr>
<td>Within</td>
<td>$g_{ij} - x_{ij}y_{ij}$</td>
<td>$D_{1j}$</td>
</tr>
<tr>
<td></td>
<td>$g_{ij} - \bar{g}i\bar{g}j$</td>
<td>$D_{3j}$</td>
</tr>
<tr>
<td>Between</td>
<td>$x_{ij}y_{ij} - \bar{g}i\bar{g}j$</td>
<td>$D_{ij}$</td>
</tr>
<tr>
<td></td>
<td>$\bar{g}i - \bar{g}j\bar{g}j$</td>
<td>$D_{3j}$</td>
</tr>
</tbody>
</table>

*NOTE.—Coefficients of linkage disequilibrium defined by Ohta (1982a) for a pair of loci, A and B, in which $g_{ij}$ denotes the frequency of chromosome $A_iB_j$, and $x_{ij}$ and $y_{ij}$ denote allele frequencies in the $k$th locality. Bar notation denotes a weighted average of frequencies over localities. The variances are calculated by summing the squared deviations and calculating the weighted average, where necessary, over localities. The $I$, $T$, and $S$ subscripts are interpreted in the same manner as Wright’s $F$-statistics. Note that $D_{ij} = D_{1j}^2 + D_{3j}^2$, but that $D_{ij}^2 \neq D_{1j}^2 + D_{3j}^2$. The standardized linkage disequilibrium, $s_{ij}$, is calculated by dividing $D_{ij}$ by the weighted average over localities of $\sum_{ij} x_{ij}(1 - x_{ij})y_{ij}(1 - y_{ij})$. 
Linkage Disequilibrium in *E. coli*

![Graph of frequency distribution](image)

**Fig. 1.**—Observed frequency distribution of total variance of linkage disequilibrium, $D^2_t$ for 66 pairs of enzyme loci.

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Comparison of Observed Variance Components of Linkage Disequilibrium with Those Generated by Computer Randomization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component</td>
<td>Observed</td>
</tr>
<tr>
<td>Total:</td>
<td>$D^2_t$</td>
</tr>
<tr>
<td>Within:</td>
<td>$D^2_{is}$</td>
</tr>
<tr>
<td></td>
<td>$D^2_{ir}$</td>
</tr>
<tr>
<td>Between:</td>
<td>$D^2_{sr}$</td>
</tr>
<tr>
<td></td>
<td>$D^2_{sr'}$</td>
</tr>
</tbody>
</table>

**NOTE:**—Entries are average variances $\pm$ SEs for all 66 possible pairs of 12 loci.

and the average value was 0.0103. Thus, the observed average total variance of disequilibrium is much greater than would be expected if the total population lacked genetic structure.

We next partitioned the total variance, $D^2_t$, into within- and between-population components (table 3) to assess the contribution of geographic subdivision to the total disequilibrium for each locus-pair. The average squared disequilibrium within populations, over all locus-pairs, $D^2_{is} = 0.017$, is about four times greater than the average generated by the randomization procedure (table 4).

For a pair of loci, the between-population coefficient, $D^2_{sr}$, is the deviation of expected chromosome frequencies at a locality relative to the total population. The observed chromosome frequencies, $g_{i,k}$, do not enter into the calculation of $D^2_{sr}$ (table 3), which is, therefore, sensitive only to variations in allele frequencies among subdivisions. The average $D^2_{sr} (= 0.0156)$ over all locus-pairs is approximately 2.5 times the randomly generated values (table 4), reflecting the observed heterogeneity in allele frequencies among localities. (Both $D^2_{is}$ and $D^2_{ir}$ contribute to the total variance, although they are not additive.) $D^2_{is}$ is greater than $D^2_{sr}$ for more pairs of loci than expected under the random model (47% vs. 38%), which
suggests that a larger part of the total variance of disequilibrium results from deviations within localities than from variation in allele frequencies among localities.

The second set of coefficients is \( D'_{is} \), the variance of disequilibrium in the total population, and \( D'_{3T} \), the average deviation of frequencies of different dilocus genotypes in geographic populations from their average frequencies in the total population (table 3). (For a pair of loci, these components are additive, such that their sum equals \( D'_{is} \).) Over all locus-pairs, both the observed average \( D'_{is} (= 0.0212) \) and \( D'_{3T} (= 0.0128) \) exceed the randomly generated averages (table 3) and account for 62% and 38% of the total variance of disequilibrium, respectively. This result agrees with that of the analysis above in that a larger part of the total disequilibrium is accounted for by the within-locality component. However, with random genetic structure, \( D'_{is} \) is expected to exceed \( D'_{3T} \), because the variance of disequilibrium in the total population should be very close to zero, whereas genotypic frequencies within each geographic sample may deviate relative to those in the total population. And, in fact, this relationship was found for 99% of the comparisons generated by computer randomization. But the opposite relationship was found for 20% of the observed comparisons, indicating significant disequilibrium in all populations but relatively little variance in disequilibrium among localities relative to the total population.

Ohta (1982a) has suggested that the dual relationships among the four components of total disequilibrium for a pair of loci may be useful in evaluating the roles of various evolutionary factors producing nonrandom associations of alleles. From both analytical and numerical considerations, she concluded that, at equilibrium, the dual relationships \( D'_{is} < D'_{3T} \) and \( D'_{is} > D'_{3T} \) should hold in general for two loci when migration among subdivisions is limited. Under the conditions of the model, allele frequencies diverge and inflate \( D'_{3T} \) as a result of genetic drift, which occurs both within subdivisions—because of finite population size—and between subdivisions—by the extinction and replacement of local populations. In this case, disequilibrium within subdivisions is nonsystematic, and \( D'_{is} > D'_{3T} \). In contrast, the opposite dual relationship, \( D'_{is} > D'_{3T} \) and \( D'_{is} < D'_{3T} \), obtained when disequilibrium is systematic. A third dual relationship, which is intermediate between these extremes, is expected when \( D'_{is} > D'_{3T} \) but the disequilibrium is nonsystematic (\( D'_{is} > D'_{3T} \)).

As shown in table 5, 35 of the 66 pairwise locus comparisons (53%) exhibit the dual relationship among the components of total disequilibrium predicted by Ohta for nonsystematic disequilibrium. For these locus-pairs, much of the disequilibrium results from variation in allele frequencies among localities, as reflected in the relatively large average value of \( D'_{3T} (= 0.0177) \). Moreover, the observed average chromosome frequencies for these locus-pairs show little deviation from the expected chromosome frequencies calculated from the products of allele frequencies in the total population: \( D'_{3T} \) averaged over these pairs of loci constitutes only about 15% of the average total disequilibrium \( D'_{is} = 0.024 \). For the remaining pairs of loci, the within-population disequilibrium, as measured by \( D'_{is} \), exceeds the between-population component due to allele frequency variation, \( D'_{3T} \). For 18 of the 66 total pairs of loci (27%), linkage disequilibrium varies in direction among localities, and thus \( D'_{is} > D'_{3T} \). These 18 locus-pairs exhibit a greater total disequilibrium than do pairs of the first group of loci, and a larger part of the total (38%) is due to variance of disequilibrium in the total population,
Table 5
Dual Relationships and Average Values of Disequilibrium Coefficients for 66 Pairs of Loci

<table>
<thead>
<tr>
<th>Dual Relationship</th>
<th>Number of Locus-Pairs</th>
<th>$D^2_{IS}$</th>
<th>$D^2_{ST}$</th>
<th>$D^2_{IS}$</th>
<th>$D^2_{ST}$</th>
<th>$D^2_{II}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. $D^2_{IS} &lt; D^2_{ST}$</td>
<td>35</td>
<td>.0052</td>
<td>.0177</td>
<td>.0204</td>
<td>.0036</td>
<td>.0240</td>
</tr>
<tr>
<td>II. $D^2_{IS} &gt; D^2_{ST}$</td>
<td>18</td>
<td>.0238</td>
<td>.0135</td>
<td>.0237</td>
<td>.0145</td>
<td>.0382</td>
</tr>
<tr>
<td>III. $D^2_{IS} &gt; D^2_{ST}$</td>
<td>13</td>
<td>.0400</td>
<td>.0127</td>
<td>.0198</td>
<td>.0350</td>
<td>.0548</td>
</tr>
</tbody>
</table>

Finally, 20% of the locus-pairs (13 of 66) display the dual relationship of systematic disequilibrium, which means that allele associations are in the same directions in each population. These 13 pairs have the greatest average within-population component of disequilibrium and also the greatest average total variance in disequilibrium. For these pairs, 64% of the total variance is due to variance of disequilibrium in the total population.

Table 6 presents chromosome frequencies for specific pairs of alleles that are positively associated in all three geographic populations. Fifteen of the 20 cases involve one or two common alleles at the four loci ADH, IDH, ACO, and PGI.

Interpretation of these relationships in terms of the effects of various non-selective and selective forces that can generate and maintain linkage disequilibrium depends critically on the rates of recombination in natural populations between the enzyme loci assayed. Recombination has not been measured in natural populations of *Escherichia coli*, but one can indirectly assess its effect on linkage disequilibrium by taking the distance between loci on the chromosome as an index of potential rate of recombination between them (Langley 1977).

Linkage Disequilibrium and Chromosomal Map Distance

The 100-min chromosomal map of *E. coli* (Bachmann et al. 1976; Bachmann and Low 1980) specifies the locations of 11 of the 12 loci listed in table 1 (the exception is ACO), from which one can estimate the shortest map distances, measured in minutes, from 0 to 50, between all pairs of loci. (One minute corresponds to a molecular distance of approximately 1% of the bacterial genome or about $10^4$ nucleotides; Glass [1982].) We assume that the rate of recombination and, hence, the rate of decay of linkage disequilibrium between any two loci increase with the map distance between them. This assumption is warranted to the extent that genetic transfer between cells occurs via plasmid-mediated mobilization of the bacterial chromosome in such a way that only neighboring enzyme loci are likely to be cotransferred. Thus, one would expect a negative correlation between map distance and the magnitude of disequilibrium for pairs of loci, provided that (1) recombination via plasmid-mediated conjugation occurs with significant frequency between cell lines in natural populations; (2) substitutive recombination rates increase with map distance; and (3) the linkage disequilibrium is not stably maintained by epistatic selection.

Using the frequency of the leading allele (see table 1) versus that of the pooled remaining alleles at each locus, we tested for disequilibrium between pairs of loci.
Table 6

Chromosome Frequencies and Coefficients of Linkage Disequilibrium for Pairs of Alleles That Are Positively Associated within Geographic Populations

<table>
<thead>
<tr>
<th>Allele Combinations</th>
<th>Iowa</th>
<th>Sweden</th>
<th>Tonga</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDH^6PGD^6</td>
<td>0.47 (.109)</td>
<td>0.48 (.077)</td>
<td>0.48 (.090)</td>
</tr>
<tr>
<td>ACO^6</td>
<td>0.43 (.101)</td>
<td>0.43 (.060)</td>
<td>0.29 (.055)</td>
</tr>
<tr>
<td>MPI^6</td>
<td>0.50 (.117)</td>
<td>0.47 (.075)</td>
<td>0.29 (.094)</td>
</tr>
<tr>
<td>MDH^6MPI^6</td>
<td>0.23 (.148)</td>
<td>0.13 (.095)</td>
<td>0.04 (.020)</td>
</tr>
<tr>
<td>BGA^7 PGI^7</td>
<td>0.10 (.033)</td>
<td>0.26 (.103)</td>
<td>0.04 (.037)</td>
</tr>
<tr>
<td>PGI^7</td>
<td>0.17 (.100)</td>
<td>0.16 (.094)</td>
<td>0.17 (.080)</td>
</tr>
<tr>
<td>AK^4 PEP^2</td>
<td>0.53 (.156)</td>
<td>0.45 (.119)</td>
<td>0.58 (.037)</td>
</tr>
<tr>
<td>PGI^4</td>
<td>0.33 (.111)</td>
<td>0.35 (.097)</td>
<td>0.52 (.119)</td>
</tr>
<tr>
<td>PGI^4</td>
<td>0.33 (.111)</td>
<td>0.29 (.133)</td>
<td>0.17 (.111)</td>
</tr>
<tr>
<td>PEP^2^2 PGI^7</td>
<td>0.20 (.067)</td>
<td>0.11 (.056)</td>
<td>0.02 (.014)</td>
</tr>
<tr>
<td>PEP^2^2 PGI^7</td>
<td>0.13 (.044)</td>
<td>0.21 (.088)</td>
<td>0.08 (.028)</td>
</tr>
<tr>
<td>PEP^2^2 PGI^7</td>
<td>0.23 (.044)</td>
<td>0.26 (.061)</td>
<td>0.42 (.048)</td>
</tr>
<tr>
<td>ACO^6</td>
<td>0.37 (.121)</td>
<td>0.30 (.096)</td>
<td>0.38 (.095)</td>
</tr>
<tr>
<td>IDH^6 PGI^7</td>
<td>0.33 (.222)</td>
<td>0.17 (.104)</td>
<td>0.38 (.136)</td>
</tr>
<tr>
<td>IDH^6</td>
<td>0.27 (.122)</td>
<td>0.15 (.082)</td>
<td>0.31 (.131)</td>
</tr>
<tr>
<td>PGI^7</td>
<td>0.67 (.222)</td>
<td>0.46 (.068)</td>
<td>0.23 (.069)</td>
</tr>
<tr>
<td>ACO^6</td>
<td>0.37 (.078)</td>
<td>0.42 (.068)</td>
<td>0.17 (.027)</td>
</tr>
<tr>
<td>PGI^7 ACO^6</td>
<td>0.21 (.122)</td>
<td>0.28 (.124)</td>
<td>0.33 (.127)</td>
</tr>
<tr>
<td>PGI^7 ACO^6</td>
<td>0.37 (.078)</td>
<td>0.32 (.104)</td>
<td>0.17 (.069)</td>
</tr>
<tr>
<td>ADH^6</td>
<td>0.65 (.211)</td>
<td>0.31 (.079)</td>
<td>0.29 (.139)</td>
</tr>
</tbody>
</table>

NOTE.—Coefficients of linkage disequilibrium in parentheses. \( D = x_{14} - x_{17} x_{47} \). 

by calculating Lewontin’s \( D' = \frac{D}{D_{	ext{max}}} \) coefficient (Lewontin 1964) and its standard error (Cole 1949) for all of 55 pairwise comparisons of the 11 mapped loci in each of the three samples. There are 27 pairs of loci in which the leading alleles were significantly associated at one or more localities. Figure 2 shows map distance plotted against a measure of linkage disequilibrium for these 27 locus-pairs. (In this case, we used the standardized within-population component of linkage disequilibrium \( s_{ij} \); see table 3) because it, like Lewontin’s \( D' \) coefficient, is relatively insensitive to differences in allele frequencies between loci.) Pairs of loci that are in systematic disequilibrium and those that show nonsystematic disequilibrium are indicated by different symbols.

Several conclusions can be drawn from the analysis summarized in figure 2. First, over all 27 pairs of loci, there is no apparent relationship between the chromosomal map distance between loci and the standardized variance of within-population linkage disequilibrium. Second, \( s_{ij} \) is on average larger for the nine pairs of loci that exhibit systematic disequilibrium than for other pairs of loci, and, again, the magnitude of disequilibrium is unrelated to map distance. For these nine pairs, relatively large values of \( s_{ij} \) are not surprising, because a large \( D_{ij}^2 \) is one way to fulfill the dual relationship of coefficients that is characteristic of systematic disequilibrium (see above). Third, for the 18 pairs of loci exhibiting significant allele associations that vary in direction among localities (i.e., non-systematic disequilibrium), there is a tendency for \( s_{ij} \) to decrease with increasing
map distance. The product-moment correlation coefficient between $s_{ij}$ and map distance for these pairs is approximately $-0.38$, but this correlation is in large part accounted for by one point in the upper left-hand corner of figure 2 ($r = -0.19$ with this point excluded). Probability values are not given for these correlation coefficients, because each sample represents a pairwise comparison of loci, and, therefore, the samples are not independent.

In sum, associations between loci that are close together and would, therefore, be expected to have a greater likelihood of cotransfer during conjugation are no stronger than associations between loci that are widely separated on the chromosome.

Discussion

The objective of this study was to elucidate the geographic aspects of the genetic structure of natural populations of *Escherichia coli*, at both the single-locus and dilocus levels. A key result is that enteric populations of *E. coli* in human hosts in widely separated regions of the world exhibit relatively little genetic differentiation. More than 95% of the total genetic diversity at single loci can be found at a locality. In absolute terms, the amount of geographic differentiation is equivalent to about twice that found among the three major races of man (Nei 1975, p. 152). Geographic differentiation in *E. coli* is largely a result of the occurrence of uncommon alleles unique to each locality (table 1). The observed pattern of apportionment of genetic diversity within and between populations suggests that rates of migration are sufficiently high to prevent strong divergence of allele frequencies at single loci among populations, a result in agreement with earlier conjectures of Levin (1981), Whittam et al. (1983), and Achtman et al. (1983).

Extensive geographic structuring of *E. coli* is revealed when the frequencies of dilocus genotypes are considered. Ohta’s method of partitioning the total variance of disequilibrium permits one to quantify the contributions of the within- and between-population components to the multilocus structure and to categorize each locus-pair by the relationships among these components (see table 5). Brown and Feldman (1981) have independently developed a set of indices for measuring multilocus associations in a structured population, in which the within- and be-
tween-population components contribute to the total variance in the number of heterozygous loci in two randomly chosen haploid genotypes (gametes). Table 7 presents these indices for the total *E. coli* sample and for the locus-pairs exhibiting the three dual relationships indicated in table 5. (Brown and Feldman [1981] should be consulted for definitions of their components.) In the total *E. coli* sample, only 57% of the variance in the number of heterozygous comparisons, \( \sigma_k^2 \), is accounted for by the sum of the single-locus effects, which indicates that the variance is almost doubled by the association among alleles at different loci. The mean disequilibrium, \( MD \), is a prominent component (40%) of the total variance. For the 35 locus-pairs exhibiting the first dual relationship, 97% of \( \sigma_k^2 \) is accounted for by the sum of single-locus components: for these pairs of loci, there is virtually no correlation in allelic state among loci over haploid genotypes. For the second dual relationship, the 18 pairs of loci show a genetic contribution of the two-locus effects, with 18% of the \( \sigma_k^2 \) being accounted for by the average linkage disequilibrium over loci, \( MD \). However, the disequilibrium is not systematic, as indicated by the relatively large variance in disequilibrium, \( VD \), which accounts for 12% of the average variance of the number of heterozygous comparisons, \( \sigma_k^2 \). The 13 locus-pairs that exhibited systematic associations, as defined by the third dual relationship of Ohta’s components, also show the characteristics of systematic association defined by Brown and Feldman, that is, high \( MD \), positive \( AI \), low \( VD \), and low \( CI \).

The effects of various evolutionary factors on the multilocus genotypic distributions in natural populations of *E. coli* depend on the rates of genetic recom-

<p>| Table 7 |
|------------------|------------------|------------------|------------------|------------------|
| <strong>Brown and Feldman’s Components of Variance in Number of Heterozygous Comparisons between Haploid Genotypes for the Total Sample and for Pairs of Loci Exhibiting the Dual Relationships Indicated in Table 5</strong> |</p>
<table>
<thead>
<tr>
<th><strong>COMPONENT</strong></th>
<th><strong>DUAL RELATIONSHIP</strong></th>
<th><strong>I</strong></th>
<th><strong>II</strong></th>
<th><strong>III</strong></th>
<th><strong>TOTAL SAMPLE</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-locus effect:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( MH ), mean genetic diversity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.260</td>
<td>1.934</td>
<td>2.143</td>
<td>2.260</td>
<td></td>
</tr>
<tr>
<td>( VH ), variance in diversity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.036</td>
<td>.032</td>
<td>.033</td>
<td>.036</td>
<td></td>
</tr>
<tr>
<td>( WH ), Wahlund’s allele frequency variance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.049</td>
<td>-.052</td>
<td>-.051</td>
<td>-.049</td>
<td></td>
</tr>
<tr>
<td>Two-locus effect:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( WC ), covariance of allele frequencies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.007</td>
<td>.026</td>
<td>.011</td>
<td>.044</td>
<td></td>
</tr>
<tr>
<td>( MD ), mean linkage disequilibrium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.061</td>
<td>.458</td>
<td>1.039</td>
<td>1.557</td>
<td></td>
</tr>
<tr>
<td>( AI ), interaction of ( WC \times MD )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.012</td>
<td>.100</td>
<td>.045</td>
<td>.066</td>
<td></td>
</tr>
<tr>
<td>( VD ), variance of disequilibrium</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.128</td>
<td>.354</td>
<td>.179</td>
<td>.662</td>
<td></td>
</tr>
<tr>
<td>( CI ), covariation of interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.063</td>
<td>.022</td>
<td>.008</td>
<td>.093</td>
<td></td>
</tr>
<tr>
<td>( \sigma_k^2 ), average variance of ( K )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.524</td>
<td>2.868</td>
<td>3.414</td>
<td>4.638</td>
<td></td>
</tr>
<tr>
<td>( \tau \sigma_k^2 ), variance of ( K ) in mixed pool</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.327</td>
<td>2.498</td>
<td>3.220</td>
<td>3.915</td>
<td></td>
</tr>
</tbody>
</table>

**NOTE.**—The number of loci and the number of locus-pairs used in the calculations are given in parentheses.

\( a \) The effect of Wahlund’s allele frequency variance, \( WH \), is negative for a single locus when \( 3H_s - H_T > 1 \), where \( H_s \) is the average genetic diversity within subpopulations and \( H_T \) is the total genetic diversity at the locus.
Linkage Disequilibrium in *E. coli*

Combination occurring between cell lines. Available evidence indicates that these rates are very low. Our demonstration that the strength of association between alleles is unrelated to the molecular distance between the various loci on the chromosome suggests that plasmid-mediated conjugation involving the cotransfer of enzyme genes and subsequent integration of chromosomal DNA is a rare event in nature. This inference is supported by the observation that plasmids integrate into the bacterial chromosome with low frequency in experimental laboratory populations (e.g., Schwesinger 1977) and is compatible with the hypothesis that the primary role of conjugation is the dissemination of "sex factors" and other plasmids (Achtman and Skurray 1977). However, recombination at rates sufficiently high to cause decay in linkage disequilibrium may still occur via phage-mediated transduction, in which the cotransfer of the enzyme loci assayed is unlikely because only a short segment of bacterial chromosome is involved in any transfer event (Hayes 1968).

Two lines of evidence suggest that transduction occurs infrequently in nature. First, estimated rates of phage-mediated gene transfer occurring in chemostat populations, and, by inference, in natural populations as well, appear to be very low, being on the order of magnitude of the mutation rate (Levin 1981). Second, isolates of *E. coli* from natural populations apparently are resistant to many of the common phages used in the laboratory. In laboratory tests of the resistance of 99 naturally occurring isolates to nine phages (T₂, T₄, T₅, T₆, T₇, λV, φ80, P1, and Mu), Bruce K. Levin (personal communication) found that, on average, individual isolates were resistant to about six of them.

If rates of recombination are very low in natural populations of *E. coli*, linkage disequilibrium generated by both selective and nonselective mechanisms will decay at a very slow rate. Nonselective factors—such as mutation, genetic drift, and the admixture of differentiated populations—could generate transient disequilibrium that would persist for many generations. In fact, half the pairs of loci studied here show the dual relationship predicted by Ohta (1982a) to hold for disequilibrium arising by mutation and genetic drift in a subdivided population. Genetic drift is possible in *E. coli* despite the enormous number of individual cells in the total population of the species, because the frequent extinction of lines can profoundly reduce the effective population size (Maruyama and Kimura 1980).

Although the interaction of genetic drift and population subdivision may account for much of the total disequilibrium, it does not explain the presence of systematic associations of alleles at 13 (20%) of the pairs of loci. The systematic disequilibria involving ADH, IDH, ACO, and PGI demonstrated in the present study were detected earlier in an analysis of more than 1,700 isolates of *E. coli* and *Shigella* species, representing 302 multilocus genotypes (Whittam et al. 1983). (That larger sample included the isolates from Sweden and Iowa but not those from Tonga.) Largely on the basis of nonrandom associations of alleles at these four loci, it was possible to distinguish three clusters of strains within the species *E. coli* as a whole. The present analysis suggests that representatives of these three groups of strains are worldwide in distribution.

Admixture of populations may account for the systematic pattern of disequilibrium, as well as for the nonsystematic disequilibrium. If populations of *E. coli* had diverged in geographic isolation, perhaps on different continents, in the past through nonselective processes and different selection pressures, then, as
the mobility of human populations increased in recent times, cell lineages from such differentiated populations may have been carried over wide geographic areas, eventually achieving a worldwide distribution. In the virtual absence of recombination, genetic combinations representing the original differentiation would persist at many localities, thereby yielding systematic linkage disequilibria. Nonsystematic associations of alleles would subsequently arise within localities as a result of random line extinction, immigration of new clones, and mutation.

Another possibility is that rates of recombination are sufficiently low for systematic allele associations to have been produced by periodic selection (Koch 1974; Kubitschek 1974; Levin 1981), a process in which certain combinations of neutral alleles are driven (hitchhike) to high frequencies through the indirect effects of selection on other loci. The selected loci may be located on the chromosome or on plasmids. Moreover, neutral allozymes could be hitchhiking with favorable adaptive gene complexes that have arisen through additive or substitutive recombination events. Clones harboring adaptive mutations or gene complexes may subsequently have spread throughout the geographic range of *E. coli*, thus producing the observed systematic disequilibria. This process would not require geographic isolation of populations of either *E. coli* or its human or other animal hosts.

A hitchhiking effect on neutral alleles by indirect selection has been demonstrated theoretically (e.g., Maynard Smith and Haigh 1974; Thomson 1977; Hedrick 1980) and has also been observed in chemostat populations of *E. coli* (Novick and Szilard 1950; Atwood et al. 1951; Nestman and Hill 1973; see also the recent work on asexual populations of yeast by Paquin and Adams [1983]). In addition, indirect selection may generate linkage disequilibrium between neutral alleles at linked loci (Thomson 1977). And by reducing effective population size through the extinction of less fit clones, indirect selection also promotes nonsystematic associations of alleles through genetic drift (Maruyama and Kimura 1980).

Epistatic natural selection can, of course, always be invoked as a general explanation for the occurrence of linkage disequilibrium. In the case of *E. coli*, epistasis could be increasing favorable allozyme combinations and maintaining stable disequilibria in all populations. Systematic disequilibria would be observed for those genetic combinations that are favored at all localities, and nonsystematic disequilibrium would represent combinations that are locally adaptive. We are, however, reluctant to invoke direct epistatic selection on enzyme variants as a general explanation of the observed disequilibrium, because it seems unlikely that selection can distinguish among combinations of different allozymes of enzymes having generally unrelated physiological functions (Selander and Levin 1980). As shown in table 6, pairs of alleles at nine different loci are positively associated in all three geographic populations sampled in this study. Several of these enzymes are, in fact, functionally interrelated, occurring in the same metabolic pathway (e.g., ACO and IDH in the TCA cycle); and it is possible that physiological differences between allozymes of these enzymes are expressed as selective differences between genotypes in certain genetic and/or environmental backgrounds. Indeed, selective differences between some allozymes of 6PGD have been demonstrated on certain genetic backgrounds in laboratory populations of *E. coli* (Dykhuizen and Hartl 1980). However, the same experimental approach has shown that most allozymes of 6PGD and PGI are neutral or nearly so (Dykhuizen 1982;
Dykhuizen and Hartl 1983). Until more experimental evidence of this type is available, the role of epistasis in promoting specific nonrandom combinations of allozymes in natural populations of *E. coli* will remain unclear.

**Acknowledgments**

We thank A. H. D. Brown, P. W. Hedrick, and M. Nei for critical comments on the manuscript. The strains from Tonga were supplied by B. R. Levin. P. Pattison assisted in preparation of the manuscript. This research was supported by grants from the National Institutes of Health and the National Science Foundation.

**LITERATURE CITED**


Whittam, Ochman, and Selander


Linkage Disequilibrium

in *E. coli*. 83


MASATOSHI NEI, reviewing editor

Received June 13, 1983; revision received July 25, 1983.
Based on the neutral theory of molecular evolution and polymorphism, and particularly assuming "the model of infinite alleles," a method is proposed which enables us to estimate the fraction of selectively neutral alleles (denoted by $P_{\text{neut}}$) among newly arisen mutations. It makes use of data on the distribution of rare variant alleles in large samples together with information on the average heterozygosity. The formula proposed is $P_{\text{neut}} = \frac{\bar{H}/(1 - \bar{H})}{\log(2\bar{n}q)/\bar{n}}$, where $\bar{n}_x(x < q)$ is the average number of rare alleles per locus whose frequency, $x$, is less than $q$; $\bar{n}$ is the average sample size used to count rare alleles; $\bar{H}$ is the average heterozygosity per locus; and $q$ is a small preassigned number such as $q = 0.01$. The method was applied to observations on enzyme and other protein loci in plaice, humans (European and Amerindian), Japanese monkeys, and fruit flies. Estimates obtained for them range from 0.064 to 0.21 with the mean and standard error $P_{\text{neut}} = 0.14 \pm 0.06$. It was pointed out that these estimates are consistent with the corresponding estimate $P_{\text{neut}}(\text{Hb}) = 0.14$ obtained independently based on the neutral theory and using data on the evolutionary rate of nucleotide substitutions in globin pseudogenes together with those in the normal globins.

**Introduction**

During the past decade and a half, much attention has been paid to protein polymorphisms (and more recently, DNA polymorphism), and various statistical methods have been developed to analyze the data (see Kimura [1983], pp. 271–281, for review). There has also been much discussion, particularly in the form of the neutralist-selectionist controversy (Crow 1972, 1981; Calder 1973; Lewontin 1974; Harris 1976; Ruffie 1976; Selander 1976), regarding the mechanism by which molecular polymorphisms are maintained.

However, very little attention has been paid to rare variant alleles whose frequencies in the population are too low for them to be regarded as members of polymorphic systems. This is understandable, because such alleles do not make...
any significant contribution to the genetic variability of the species, and also
because they cannot be detected unless the sample size is unusually large.

In this paper I intend to show that such rare variant alleles can, nevertheless,
supply valuable information on the mechanism by which polymorphism at the
molecular level is maintained. In particular, I shall demonstrate that the neutral
theory of protein polymorphism (Kimura 1968a, 1968b; Kimura and Ohta 1971;
Kimura [1983] for review) can supply a theoretical basis to connect observations
on rare variant alleles with those of polymorphic alleles. Furthermore, I shall
endeavor to show that the result of data analysis fits well the larger picture of
molecular evolution as seen from the standpoint of the neutral theory.

**Basic Theory**

Let us assume a random mating, diploid population of effective size \( N_e \).
Consider a particular locus and assume the infinite allele model (Kimura and Crow
1964), that is, assume that whenever mutation occurs it leads to an allele not
already existing. Let \( v \) be the mutation rate per locus per generation. I denote by
\( \Phi(x) \) the distribution of allelic frequencies under the assumption of an equilibrium
in which mutational production of new alleles is balanced by random extinction
of existing alleles. This distribution means that \( \Phi(x)dx \) represents the expected
number of alleles whose frequencies lie in the range between \( x \) and \( x + dx \). When
all of the mutations are selectively neutral, it was shown by Kimura and Crow
(1964) that

\[
\Phi(x) = 4N_e v (1 - x)^{M v - 1} x^{-1}.
\]

(1)

Note that this distribution refers only to those alleles actually contained in the
population (\( x > 0 \)); although I assume there are an infinite number of possible
alleles, only a limited number of them are present at any given moment in the
population, and I do not include countless missing alleles. In the following, I shall
use the letter \( M \) to stand for \( 4N_e v \), so that the right-hand side of equation (1)
becomes \( M (1 - x)^{M v - 1} x^{-1} \).

The average value of the sum of squares of allelic frequencies or the average
homozygosity is

\[
\bar{H}_o = \int_0^1 x^2 \Phi(x)dx = 1/(M + 1),
\]

(2)

and, therefore, the average heterozygosity is

\[
\bar{H}_e = 1 - \bar{H}_o = M/(M + 1).
\]

(3)

This means that, if I know the value of \( \bar{H}_e \) from observation, I can estimate the
parameter \( M \) by the formula

\[
M = \bar{H}_e / (1 - \bar{H}_e).
\]

(4)

As emphasized by Nei (1975), a reliable estimate of the average heterozygosity
of any species can only be obtained by averaging heterozygosities over a number
of loci. Therefore, it is necessary to consider the possibility that mutation rates
for neutral alleles differ among loci. To take such a variation into account, Nei
et al. (1976) proposed an infinite allele model assuming that \( M \) or \( 4N_e v \) among loci
follows a gamma distribution with the mean $\bar{M}$ and the variance $V_M$. In this model, the relationship between the average heterozygosity and $\bar{M}$ is more complicated, but Nei (1975) derived a useful approximation formula,

$$\bar{H}_e = \frac{\bar{M}}{1 + \bar{M}} - \frac{V_M}{(1 + \bar{M})^3},$$

(5)

which is valid unless $\alpha = \bar{M}^2/V_M$ is small and $\bar{M}$ is large. According to Nei et al. (1976), an appropriate value of $\alpha$ is about 1. Also, for a wide range of organisms, $\bar{M}$ seldom exceeds 0.3 (see, e.g., Nevo 1978), so this formula should have wide applicability. Note that under these circumstances $V_M$ is much smaller than $\bar{M}$, and therefore variation of $M$ among loci has a relatively small effect on $\bar{H}_e$, as easily seen by comparing equations (3) and (5).

One important point which I should note in estimating $M$ from $\bar{H}_e$ using equation (4) is that $M$ is determined largely by polymorphic alleles; rare alleles contribute very little to $\bar{H}_e$ and therefore to $M$.

Next, let us examine the occurrence of rare alleles whose frequencies are less than a certain small value $q$. Bearing in mind the standard practice of defining a polymorphic locus as one in which the most frequent allele does not exceed 99%, I find it appropriate to take $q = 0.01$.

It can be shown mathematically (see, e.g., Kimura 1983, p. 227) that, in the neighborhood of $x = 0$, the population behavior of alleles in general, including those having mild selective advantage or disadvantage, is essentially the same as that of selectively neutral mutants. Thus the average number of alleles per locus whose frequencies are less than $q$ is

$$\bar{n}_q(x < q) = \int_0^q \Phi(x) dx \approx M \log_2(2nq),$$

(6)

where $n$ is the sample size. This formula is valid if $4N_e|s|q$ is small, where $|s|$ is the absolute value of the selection coefficient of a mutant allele. In this formula $M$ stands for $4N_e\nu$; however, $\nu$ here represents the mutation rate for practically all types of alleles, as pointed out by Nei (1977), and not just for selectively neutral alleles. In fact, he proposed the use of equation (6) for estimating the mutation rate for protein loci. I shall denote $M$ in this equation by $M_q$ in order to distinguish it from $M$ obtained by equation (4). If rare variants are scored at more than one locus with a large sample for each locus (the mean size being $\bar{n}$ per locus), and if $N_e$ is known, I can estimate the total mutation rate per locus by

$$\nu_{(E)} = M_q/(4N_e),$$

(7)

where the subscript $E$ refers to electrophoretically detectable alleles, and

$$M_q = \bar{n}_q(x < q)/\log_2(2\bar{n}q).$$

(8)

However, if I denote by $\nu_{(0\infty)}$ the mutation rate per locus for selectively neutral (and electrophoretically detectable) alleles, then $M = 4N_e\nu_{(0\infty)}$, so that

$$\nu_{(0\infty)} = M/(4N_e).$$

(9)
Thus, I can estimate the fraction of neutral alleles among all the mutations that can be detected electrophoretically by

$$P_{\text{neut}} = \frac{v_{\text{neut}}}{v_{\text{tot}}} = \frac{M}{M_n}. \quad (10)$$

What is important here is that, even when the actual value of $N_e$ is not known, we can estimate the fraction of neutral alleles at the time of their occurrence by the ratio $M/M_n$, provided that the neutral theory is correct. This equation may be rewritten as

$$P_{\text{neut}} = \frac{\bar{H}_e}{1 - \bar{H}_e} \cdot \log(2\bar{q})$$

where $\bar{H}_e$ is the mean heterozygosity per locus estimated by averaging over a number of polymorphic as well as monomorphic loci, $\bar{n}_o(x < q)$ is the number of rare variant alleles per locus, and $\bar{n}$ is the average sample size over loci used to count the rare alleles whose frequencies are less than $q$. As mentioned already, an appropriate value for $q$ is 0.01, although other values, such as $q = 0.005$, may be used.

Equations (6)–(10a) contain several assumptions and approximations. In particular, equation (6) is derived by assuming that the distribution of rare alleles in the sample is sufficiently close to that in the population. More accurate (but more complicated) formulas on the subject have been derived by Chakraborty (1981). Also, the use of the average sample size $\bar{n}$ in equation (8) requires that variation of the sample size among loci is relatively small. In the plaice data, $n$ varies around the mean = 1,956 with the standard deviation 508, so that use of the mean ($\bar{n} = 1,956$) alone will not cause much error. It is hoped that these approximations are acceptable for the moment, and that, in the future, more extensive data will be analyzed with better statistical methods.

**Data Analysis**

As the first example of the application of the theory above to estimate the fraction of selectively neutral mutations, I shall use the data from Ward and Beardmore (1977) on protein variation in the plaice, *Pleuronectes platessa*, a marine flatfish. They screened electrophoretically detectable variation at 46 protein loci (39 enzyme and seven nonenzyme proteins), taking very large samples from the Bristol Channel population. This probably represents the most comprehensive investigation of protein variation in fishes. The mean heterozygosity per individual per locus turned out to be $0.102 \pm 0.026$. The sample sizes differ from locus to locus: they are more than 2,000 in 8 loci, between 2,000 and 500 in 9 loci, between 500 and 100 in 16 loci, but less than 100 in the remaining 13 loci. Using equation (4), I get $M = 0.114$ for $\bar{H}_e = 0.102$. If I use Nei's formula 5, then $\bar{M} = 0.128$ for $\alpha = 1$ and $\bar{H}_e = 0.102$, so the effect of variation of mutation rate among loci is rather small. Therefore, in the following, I shall use equation (4) to simplify our calculation.

In order to estimate $M_n$ from observations on rare variants, the sample size must be large. Therefore, I have chosen from the data of Ward and Beardmore (1977, tables 2 and 3) 11 loci for which the sample size per locus is larger than 1,000. The average sample size per locus for them turned out to be $\bar{n} = 1,956 \pm$
Of these 11 loci, 8 are polymorphic and 3 are monomorphic. The average heterozygosity of these 11 loci is 0.147, which is not very different from the average heterozygosity of 46 loci, that is, 0.102. Among these loci, 30 alleles are found whose frequencies in the sample are less than 0.01. Thus, \( \bar{n}_{a}(x < 0.01) = 30/11 = 2.73 \) per locus. Then, applying equation (8), where I assume \( \bar{q} = 0.01 \), I get
\[
M_q = \frac{2.73}{\log_2(2 \times 19.56)} = 0.744.
\]
This leads to \( v_{Ne} = 0.744/(4N_e) \). Although the real value of \( N_e \) is not known, if it is \( 10^4 \), we have \( v_{Ne} = 1.86 \times 10^{-7} \). Finally, substituting \( M = 0.114 \) and \( M_q = 0.744 \) in equation (10), I get
\[
P_{neut} = 0.15.
\]
This means that one mutation out of 6.5 on the average is selectively neutral while the remaining 5.5 are too deleterious to contribute to protein polymorphism.

As the second example, I shall use the data on human populations of Harris et al. (1974), who reported the incidence of rare alleles determining electrophoretic variants at 43 enzyme loci in Europeans. From their table 1, I have chosen 26 loci for which the sample size is larger than 1,000. The average sample size for them is \( \bar{n} = 4,058.04 \). The average number of rare alleles per locus has turned out to be 49/26 or 1.88. Since Harris et al. defined rare alleles as those alleles whose individual frequency in the sample was less than 0.005, this corresponds to \( q = 0.005 \) of equation (8). Then, substituting \( \bar{n}_{a}(x < q) = 1.88, \bar{n} = 4,058.04 \), and \( q = 0.005 \) in this equation, I get \( M_q = 0.508 \). This gives an estimate for \( 4N_{e}v_{Ne} \), where \( v_{Ne} \) is the total mutation rate per enzyme locus for electrophoretically detectable alleles. However, from equation (4), I obtain \( M = 0.0718 \) by assuming \( H_e = 0.067 \), which is an approximate value for the average heterozygosity per locus due to common polymorphic alleles (Harris and Hopkinson 1972). Then from equation (10) the fraction of mutations that are selectively neutral among all electrophoretically detectable mutations is
\[
P_{neut} = M/M_q = 0.14.
\]
This value is very close to the corresponding estimate obtained for the plaice.

Extensive studies of rare variants in human populations have also been done by Neel and his associates on Amerindians, and valuable data have been obtained. I use the data presented in table 1 of Neel (1978), which lists the occurrence of rare variants at 28 loci in 21 Amerindian tribes. His definition of rare variant alleles corresponds to \( q = 0.01 \) in my terminology, and from his table I obtain
\[
\bar{n}_{a}(x < q) = 1.29 \quad \text{and} \quad \bar{n} = 6,442.07,
\]
giving \( M_q = 0.266 \). An interesting feature of his data is that some of the variants represent what he calls “private polymorphisms,” that is, they are concentrated in a single or several related tribes where their frequencies are well above the minimum for a polymorphism. For example, an allele called YAN-2 at the albumin (Alb) locus is present in more than 6% of the members of the Yanomama tribe but absent in other tribes. We can calculate the value of \( M \) using the average heterozygosity at 23 loci over 12 tribes as listed in table 5 of Neel (1978), where I find \( H_e = 0.054 \). Thus, I obtain \( M = 0.057 \). Therefore, the fraction of neutral mutations, as estimated by \( M/M_q \), turns out to be \( P_{neut} = 0.21 \), which is not very different from the corresponding value obtained for European populations (i.e., \( P_{neut} = 0.14 \)).

The occurrence of rare variants is also reported in the Japanese macaque (Macaca fuscata fuscata) studied by Nozawa and his associates (see, e.g., Nozawa et al. 1982). They surveyed 32 independent protein loci and obtained 1.3% as the average heterozygosity, which is a rather low value even for mammals. Their extensive studies so far yield the following data (Nozawa, personal communication, 1981). The average number of rare variants per locus is
\[
\bar{n}_{a}(x < 0.01) = 23/32 = 0.719,
\]
and the average sample size is \( \bar{n} = 1,609.9 \). Thus I get
\[
M_q =
0.207 from equation (8). The observed average heterozygosity per locus is \( \bar{H}_e = 0.013 \pm 0.0014 \), from which I get \( M = 0.0132 \). Using these values, I obtain \( P_{\text{neut}} = M/M_q = 0.064 \). This means that, roughly speaking, one mutation out of 16, on the average, is selectively neutral in Japanese monkeys. This is less than half as large as the corresponding value obtained for the plaice.

The census number of the total population of the Japanese macaque is estimated to be 20,000–70,000. It is also estimated that the effective population size is about one-third of its census number (cited from Nozawa et al. 1975). Following Nei (1977), if I assume \( N_e = 2 \times 10^4 \), I obtain \( v_{(E)} = M_q/(4N_e) = 2.6 \times 10^{-6} \). However, the mutation rate for neutral alleles is \( v_{(E)} = M/(4N_e) = 1.65 \times 10^{-7} \) per generation.

As the final example, I shall analyze the data from Drosophila willistoni group studied by Ayala and his associates (1974). The sample size per locus per species in this case is not as large as in the previous examples. Of the five species studied, only the D. willistoni data are extensive enough for the average sample size per locus to be larger than 500 (in terms of the gene number, i.e., \( 2n \)), so I shall concentrate on this species. From their table 1, which lists allelic frequencies at 31 loci, I have chosen alleles whose frequencies are less than 1% (\( q = 0.01 \)). There are 85 such alleles, so that \( \bar{H}_e(x < q) = 85/31 = 2.74 \). The average sample size per locus is \( 2\bar{n} = 568.06 \). From these values, I obtain \( M_q = 1.60 \). The average heterozygosity per locus \( (\bar{H}_e) \) as listed in table 6 of Ayala et al. (1974) is 0.177, from which I obtain \( M = 0.215 \). Therefore, the estimate for the fraction of neutral mutations among all electrophoretic mutations at the time of occurrence is \( P_{\text{neut}} = M/M_q = 0.13 \). This estimate is not very different from the corresponding estimates obtained for human populations, as well as for the plaice.

The results of analyses of the five examples above are summarized in table 1. The average of five \( P_{\text{neut}} \) values turns out to be 0.14 ± 0.06.

**Discussion**

From the standpoint of the neutral theory, the rare variant alleles are simply those alleles whose frequencies within a species happen to be in a low-frequency range \( (0, q) \), whereas polymorphic alleles are those whose frequencies happen to be in the higher-frequency range \( (q, 1 - q) \), where I arbitrarily take \( q = 0.01 \). Both represent a phase of molecular evolution.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
</table>
| **Proportion of Selectively Neutral Mutations**
| **at the Time of Occurrence among Electrophoretically Detectable Mutations \( (P_{\text{neut}}) \)**
| **Estimated from Five Data Sets**
<table>
<thead>
<tr>
<th>( \bar{H}_e )</th>
<th>( M )</th>
<th>( M_q )</th>
<th>( P_{\text{neut}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plaice ( (Pleuronectes platessa) )</td>
<td>.102</td>
<td>.114</td>
<td>.744</td>
</tr>
<tr>
<td>Human: European</td>
<td>.067</td>
<td>.072</td>
<td>.509</td>
</tr>
<tr>
<td>Amerindian</td>
<td>.054</td>
<td>.057</td>
<td>.266</td>
</tr>
<tr>
<td>Japanese macaque</td>
<td>.013</td>
<td>.013</td>
<td>.207</td>
</tr>
<tr>
<td>Fruit fly ( (Drosophila willistoni) )</td>
<td>.177</td>
<td>.215</td>
<td>1.60</td>
</tr>
</tbody>
</table>

Rare Variant Alleles 89
However, in contrast to polymorphic alleles, which are predominantly neutral, the rare variant alleles may include slightly deleterious and sometimes even definitely deleterious alleles in addition to selectively neutral ones. This means that rare variant alleles reflect the total mutation rate much more faithfully than polymorphic alleles. In this connection, Harris et al.'s (1974) observation is relevant. They found that "polymorphic" and "monomorphic" loci do not differ in the average heterozygosity for rare alleles if the placental alkaline phosphatase, an unusually variable locus, is excluded. This is easy to understand if we note that the intrinsic mutation rates ($v_{neut}$) at these two classes of loci may essentially be the same.

In the analysis above, I have estimated the fraction ($P_{neut}$) of selectively neutral mutations at the time of occurrence among mutations that can be detected by electrophoretic method, using data from the plaice, humans, the Japanese macaque, and the fruit fly (see table 1). It is remarkable that this fraction is relatively uniform among widely separated species with highly different average heterozygosities.

The present analysis is consistent with Ohta's (1975) finding on the excess of rare alleles: using Drosophila and human data, she noticed that the observed and theoretical distributions of allelic frequencies agree quite well under the neutral theory with respect to polymorphic alleles but that there is a marked excess of rare alleles in the observed distribution. Ohta (1976) went further and showed, using data on D. willistoni, that the excess of rare alleles is more pronounced in the substrate-specific enzymes than the substrate-nonspecific enzymes (see Ohta's [1976] table 3). Actually, if we apply the present method for each of these classes of enzymes separately (16 specific and 15 nonspecific enzymes in D. willistoni), we obtain $P_{neut} = 0.070$ for the substrate-specific group and $P_{neut} = 0.204$ for the substrate-nonspecific group. From these two $P_{neut}$ values, it is evident that the probability of a mutational change being selectively neutral is much smaller for the substrate-specific enzymes than for the nonspecific enzymes.

A similar calculation can be done using human data (Harris et al. 1974) following the classification of group I (substrate-specific) and group II (substrate-nonspecific) enzymes proposed by Gillespie and Langley (1974). For 13 loci of the group I enzymes, it turns out that $P_{neut} = 0.11$, and for 10 loci of the group II enzymes, $P_{neut} = 0.43$.

Previously, Gillespie and Langley (1974) showed that the average heterozygosity ($H_e$) per locus is much lower for substrate-specific than for the nonspecific enzymes not only for Drosophila but also for the human and the mouse. These observations are compatible with the neutral theory if it is assumed that selective constraint (negative selection) is stronger for substrate-specific than nonspecific enzymes. This means that the probability of an amino acid change being not harmful, that is, selectively neutral, is smaller for the substrate-specific than for nonspecific enzymes, even if the total mutation rate per locus ($v_{TCE}$) is the same for these two types of loci.

Finally, I would like to show that the present analysis on $P_{neut}$ is consistent with the results obtained from recent studies on the evolutionary rate of globin pseudogenes. As shown by Miyata and Yasunaga (1981) and Li et al. (1981), the evolutionary rate of nucleotide substitutions is very high for pseudogenes. This is easily understandable from the neutral theory, because pseudogenes can be regarded as "dead genes" which have been liberated from the constraint of neg-
ative selection, so that all the mutations in them become selectively neutral. Thus, pseudogenes accumulate mutational changes at the maximum speed as predicted by the neutral theory. This can be explained in more quantitative terms as follows: if I denote the fraction of neutral mutations by \( f_0 \) (which is determined by the degree of selective constraint), the rate of evolution in terms of mutant substitutions is

\[
k = v_0 = f_0 v_r,
\]

where \( v_0 \) is the neutral mutation rate and \( v_r \) is the total mutation rate. Note that, under the neutral theory, the rate of evolution is equal to the mutation rate for neutral alleles (Kimura 1968a). As predicted by Kimura (1977), the maximum evolutionary rate is attained when \( f_0 = 1 \), and it is likely that pseudogenes indeed represent such a case.

If I adopt the estimates given in table 3 of Li et al. (1981), the average rate for the three globin pseudogenes, mouse \( \psi \alpha 3 \), human \( \psi \alpha 1 \), and rabbit \( \psi \beta 2 \), is \( 4.6 \times 10^{-9} \) substitutions per nucleotide site per year. However, the rates of nucleotide substitutions at the first, second, and third positions of the codons in the normal globin genes are \( 0.71 \times 10^{-9}, 0.62 \times 10^{-9}, \) and \( 2.64 \times 10^{-9} \), respectively (Li et al. 1981). In order to estimate the mutation rates \( v_{1IE} \) and \( v_{0IE} \) from these observed values, one needs to know what fraction of nucleotide changes at each of the three positions of the codon cause electrophoretically detectable amino acid changes. For this purpose, assume that electrophoretic mobility of a protein is determined solely by its net charge and that, among 20 amino acids, aspartic and glutamic acids are acidic and negatively charged, lysine and arginine are basic and positively charged, while the rest are electrically neutral. Then, from the standard code table, we find that the probability of a random nucleotide change causing an electrophoretically detectable amino acid change is about 0.28 for the first position, one-third for the second position, and only one-twelfth for the third position of the codon.

I also note that nucleotide changes always cause amino acid changes at the second position and predominantly so at the first position. However, at the third position, nucleotide changes cause amino acid changes in only some one-third of the cases, the rest being synonymous. Furthermore, in globins, the synonymous component of nucleotide substitutions has an evolutionary rate at least two or three times as high as the amino acid altering nucleotide substitutions (Jukes 1980; Kimura 1981), suggesting that the probability of a random nucleotide change being selectively neutral is much higher for the synonymous than for amino acid altering changes. There is also the phenomenon of nonrandom usage of synonymous codons (Grantham 1980; Ikemura 1981), and this, too, complicates the problem. For these reasons, I exclude the data from the codon's third position in the following calculation.

Then, using Li et al.'s (1981) estimates, I can compute the fraction of neutral mutations with respect to electrophoretically detectable changes in hemoglobin by the ratio \( (0.71 \times 0.28 + 0.62/3)/(4.6 \times 0.28 + 4.6/3) \), which gives \( P_{neu}(Hb) = 0.14 \). Although I do not know the evolutionary rates in terms of amino acid substitutions of the various enzymes and other proteins used to estimate \( P_{neu} \) in table 1, it is likely that their average evolutionary rate is not very different from the evolutionary rate of hemoglobin which is near the median of the evolutionary rates of proteins (Kimura 1974). Considering the many uncertainties involved in
the process of estimating the fraction of neutral mutations, the agreement between the two independent estimates above, that is, \( P_{\text{neut}} = 0.14 \pm 0.06 \) for enzyme and other protein loci in the four organisms and \( P_{\text{neut}}(\text{Hb}) = 0.14 \) for hemoglobin in mammals, is impressive. I believe that this consistency strongly supports the neutral theory. I also think that a detailed study of rare variant alleles is just as important for understanding the mechanism of the maintenance of genetic variability as that of polymorphic alleles. It is hoped that more data on rare variants will be obtained for wild species whose ecologies, particularly the population sizes, are well known.

**Acknowledgment**

This work is supported in part by a grant-in-aid from the Japanese Ministry of Education, Science and Culture.

**LITERATURE CITED**


Rare Variant Alleles


Received June 22, 1983; revision received August 2, 1983.
Statistical analyses of DNA sequences of globin genes (β^A, β^C, and γ) from goat and sheep (including new sequence information for the second intron of sheep β^A and γ, kindly provided by A. Davis and A. W. Nienhuis) indicate that the rates of nonsynonymous substitution in these genes have been greatly accelerated following the gene duplication separating γ and the ancestor of β^A and β^C and the gene duplication separating β^A and β^C. In both cases the acceleration was apparently due to relaxation of purifying selection (functional constraints) rather than advantageous mutations because acceleration occurred only in less important parts of the β globin chain. The rates of nonsynonymous substitution in these genes are estimated to be about 2.3 × 10^{-9} per site per year, which is three times higher than that for the divergence between human β and mouse β major globin genes. Our analyses further suggest that the rate of synonymous substitution in functional genes and the rate of substitution in pseudogenes are approximately equal and are between 2.8 × 10^{-9} and 5.0 × 10^{-9} and that the rate of substitution in introns is about 3.0 × 10^{-9}. The divergence time between β^A and β^C and that between γ and the β^A-β^C pair are about 12 and 30 million years, respectively. The proportion of transition mutations is estimated to be 64%, two times higher than expected under random mutation but considerably lower than the 96% estimated for animal mitochondrial DNA.

Introduction

In their studies of the evolution of hemoglobins and myoglobins, Goodman (1976, 1981) and his associates (Czelusniak et al. 1982) have concluded that extremely high rates of amino acid substitution occurred following the gene duplication separating myoglobin and hemoglobin and the gene duplication separating α and β hemoglobins, and that the high rates were due to advantageous mutations, improving the function of myoglobin and hemoglobin. Wilson et al. (1977) and Kimura (1981), however, have challenged this view and contended that the alleged
high rates are overestimates because the actual dates of gene duplication appear to be much older than those assumed by Goodman et al. Kimura (1981) has also contended that, while rapid evolution may occur following gene duplication, it is more likely to arise from relaxation of purifying selection (functional constraints) than from advantageous mutations. Some other cases of accelerated evolution have also been proposed, but, like the above two cases, they are all of ancient duplication and the dates of duplication are uncertain (see the review by Wilson et al. [1977]). Thus it remains to be determined whether the rate of amino acid replacement can be accelerated following gene duplication and, if acceleration indeed occurs, whether it is due to relaxation of purifying selection or advantageous mutations.

To resolve the two issues above we studied the evolution of the adult (β^A), preadult (β^C), and fetal (γ) globin genes of goat and sheep. The DNA sequence data have indicated that γ and the ancestor of β^A and β^C were derived from a relatively recent duplication event and that β^A and β^C were derived from a block duplication event that also produced the two pseudogenes, ψβ^X and ψβ^Z (Cleary et al. 1981; Schon et al. 1981). Since the two duplication events are quite recent, it should not be difficult to infer whether the rates of nonsynonymous substitution in β^A, β^C, and γ have been substantially accelerated and also to infer the main cause of acceleration if this has indeed occurred in any of these genes.

We also studied the proportion of transition and transversion mutations. Gojobori et al.'s (1982) study of the substitution pattern in pseudogenes suggests that in nuclear DNA the proportion of transition mutations is about 56%, which is almost two times higher than the 33% expected under random mutation. Is there really such a strong bias in the pattern of point mutation? Since the goat and sheep sequences are closely related, they are useful for resolving this question.

**Sequence Data**

The nucleotide sequences used in this study are goat β^A (Gβ^A), β^C (Gβ^C), γ (Gγ), pseudogene β^X (ψβ^X), and pseudogene β^Z (ψβ^Z) (Cleary et al. 1981; Schon et al. 1981); the partial sequences of sheep β^A (Sβ^A) and γ (Sγ) published in Kretschmer et al. (1981) and the sequences for the second intron of sheep β^A and γ to be presented below; human β (Hβ) (Lawn et al. 1980); and mouse β major (Mβ^Ma) (Konkel et al. 1978). The coding regions of these sequences can easily be aligned (see, e.g., Schon et al. 1981). The introns of Hβ and Mβ^Ma are not considered in this study because it is difficult to align them with those of the other genes. No sequence data are available for the first intron of sheep β^A and γ. Thus, the first introns used are those from the five goat sequences; we have excluded five nucleotides from each of the 5' and 3' ends, for these regions are highly conservative (Rogers and Wall 1981). The second introns used are described below. We have also examined the 5' region starting from the CCAAT box to the nucleotide before the initiation codon of the goat sequences; we used the alignments given by Cleary et al. (1981) and Schon et al. (1981).

The second introns of sheep β^A and γ have been sequenced by A. Davis and A. W. Nienhuis. They have kindly provided these unpublished sequences for our analysis and for inclusion in the present paper. We have aligned these sequences with the second introns of Gβ^A and Gγ (fig. 1a and b). This alignment can easily be superimposed on the alignment of the second introns of goat β^A, β^C, γ, ψβ^X, and ψβ^Z given by Schon et al. (1981). The length of the second intron is 906 base
pairs (bp) in Gβ^, 903 bp in Sβ^, 827 bp in Gγ, and 832 bp in Sγ. The difference in size can largely be accounted for by two insertions or deletions. First, approximately 180 bp from the 5' end there appears a 247 bp sequence that is common to Gγ and Sγ but absent in Gβ^ and Sβ^. This sequence is flanked at both ends by a perfect direct repeat of 13 bp and has been suggested to be a transposable element (Schon et al. 1981). Second, approximately 45 bp from the 3' end there appears a 318 bp sequence that is common to Gβ^ and Sβ^ but absent in Gγ and Sγ. This sequence is flanked at both ends by a perfect direct repeat of 7 bp and also has been suggested to be a transposable element. These two subsequences and their flanking repeats are not included in the present analysis. The parts of the sequences used in our analysis are the segment from position 6 to 166, the segment from position 428 to 797, and the segment from position 1131 to 1164 (fig. 1).

**Number of Nucleotide Substitutions between Sequences**

Table 1 shows the proportion of different nucleotides and the estimated number of nucleotide substitutions per site between genes in the intron regions and in the exon regions. The proportion of different nucleotides at synonymous and nonsynonymous sites are estimated by the method of Miyata and Yasunaga (1980). This method tends to give an overestimate of the number of synonymous differences and an underestimate of the number of nonsynonymous differences, if one or both of the sequences compared are pseudogenes and are not closely related, because the method was intended to be applied to functional genes and assumes that synonymous substitutions occur much more often than nonsynonymous substitutions, an assumption that is obviously not true for pseudogenes. Therefore, the estimates for the cases involving one pseudogene and one functional gene should be taken with caution. In the case of ψβ^ versus ψβ^, we have counted the numbers of synonymous and nonsynonymous substitutions and found them to be in agreement with the numbers obtained by Miyata and Yasunaga's method; this is reasonable because the degree of sequence divergence in this case is fairly small. In both the exon regions and the intron regions, the mean and standard error of the number (d) of substitutions per nucleotide site are computed by the method of Jukes and Cantor (1969) and Kimura and Ohta (1972). In table 1 we have not considered introns 1 and 2 separately because in all comparisons the d values for the two introns are not significantly different from each other.

**The Block Duplication Hypothesis**

Cleary et al. (1981) proposed that the two pseudogenes ψβ^ and ψβ^ were derived from a common defective sequence because they share several identical deleterious mutations, for example, a single nucleotide insertion within or immediately following codon 11. Under this hypothesis, d should be similar for all gene regions since, if pseudogenes are subject to no functional constraint, all regions would evolve at the same rate. Indeed, table 2 shows that the d values for different regions are not statistically different from one another. (As will be explained below, the higher d value for synonymous sites than for nonsynonymous sites may be partly due to nonrandom mutation.) In particular, the 5' region, which is known to be extremely conservative in functional genes, has evolved as fast as the other regions. This observation strongly supports the hypothesis that the two pseudogenes were derived from duplication of a nonfunctional gene.
FIG. 1.—a and b, Alignment of the nucleotide sequences of the second introns of goat β⁺(GBA), sheep β⁺(SBA), goat γ(GY), and sheep γ(SY). The sequences for the introns of sheep β⁺ and γ were obtained by A. Davis and A. W. Nienhuis, who have kindly provided these unpublished sequences. Boxed regions are repeated sequences. 

Cleary et al. (1981) proposed further that β⁺ and βᶜ were derived from the same duplication event that produced the two pseudogenes because the degree of sequence divergence between β⁺ and βᶜ is similar to that between ψβ⁺ and ψβᶜ and because β⁺ and βᶜ are located downstream from ψβ⁺ and ψβᶜ, respectively (fig. 2a). The observation that in the introns the d value is considerably larger for
the $\beta^A-\beta^C$ pair than for the $\psi\beta^X-\psi\beta^Z$ pair would suggest, in the absence of gene conversion, that the divergence time for the former pair is significantly longer than that for the latter pair. However, the observation that at synonymous sites the $d$ value is considerably smaller for the $\beta^A-\beta^C$ pair than for the $\psi\beta^X-\psi\beta^Z$ pair would suggest that the reverse is true. One simple way to reconcile these two observations is to assume that the two pairs were duplicated at the same time and the differences above in the $d$ value were due to random errors. Actually, the block duplication hypothesis is supported by the observations that, among the goat sequences, $\beta^A$ is most closely related to $\beta^C$ and $\psi\beta^X$ is most closely related to $\psi\beta^Z$ (table 1), and that the $d$ values for the four pairs $\beta^A-\psi\beta^X$, $\beta^A-\psi\beta^Z$, $\beta^C-\psi\beta^X$, and $\beta^C-\psi\beta^Z$ are very similar (table 2). These observations can also be explained by assuming that the $\psi\beta^X-\psi\beta^Z$ pair and the $\beta^A-\beta^C$ pair were produced by two
Table 1
Proportions of Different Nucleotides between Genes and Estimated Numbers of Nucleotide Substitutions per Site in the Introns and the Coding Regions

<table>
<thead>
<tr>
<th>Gene Pair</th>
<th>Introns</th>
<th>Non-synonymous</th>
<th>Synonymous</th>
<th>Introns</th>
<th>Non-synonymous</th>
<th>Synonymous</th>
<th>Non-synonymous/Synonymous</th>
</tr>
</thead>
<tbody>
<tr>
<td>GβA-SβB</td>
<td>18/549</td>
<td>3/239</td>
<td>3/76</td>
<td>3.4 ± 0.8</td>
<td>1.3 ± 0.7</td>
<td>4.1 ± 2.4</td>
<td>0.31</td>
</tr>
<tr>
<td>Gγ-Sγ</td>
<td>17/537</td>
<td>5/259</td>
<td>4/86</td>
<td>3.2 ± 0.8</td>
<td>2.0 ± 0.9</td>
<td>4.8 ± 2.4</td>
<td>0.42</td>
</tr>
<tr>
<td>GβC-GβA</td>
<td>69/667</td>
<td>21/324</td>
<td>8/99</td>
<td>11.1 ± 1.4</td>
<td>6.8 ± 1.5</td>
<td>8.5 ± 3.1</td>
<td>0.80</td>
</tr>
<tr>
<td>GβC-SβA</td>
<td>54/550</td>
<td>16/232</td>
<td>10/74</td>
<td>10.5 ± 1.5</td>
<td>7.2 ± 1.8</td>
<td>14.9 ± 4.9</td>
<td>0.48</td>
</tr>
<tr>
<td>ψβX-ψβY</td>
<td>52/679</td>
<td>18/250</td>
<td>9/77</td>
<td>8.1 ± 1.1</td>
<td>7.6 ± 1.8</td>
<td>12.7 ± 4.3</td>
<td>0.60</td>
</tr>
<tr>
<td>Gγ-GβA</td>
<td>77/655</td>
<td>34/327</td>
<td>11/102</td>
<td>12.8 ± 1.5</td>
<td>11.2 ± 2.0</td>
<td>11.6 ± 3.6</td>
<td>0.97</td>
</tr>
<tr>
<td>Gγ-SβA</td>
<td>64/528</td>
<td>32/327</td>
<td>11/78</td>
<td>13.0 ± 1.7</td>
<td>14.9 ± 2.7</td>
<td>15.7 ± 4.9</td>
<td>0.95</td>
</tr>
<tr>
<td>Gγ-GβB</td>
<td>73/660</td>
<td>35/319</td>
<td>15/101</td>
<td>12.0 ± 1.7</td>
<td>11.7 ± 2.0</td>
<td>17.1 ± 4.5</td>
<td>0.68</td>
</tr>
<tr>
<td>Sγ-GβA</td>
<td>60/539</td>
<td>32/258</td>
<td>12/84</td>
<td>12.1 ± 1.6</td>
<td>13.6 ± 2.5</td>
<td>15.8 ± 4.7</td>
<td>0.86</td>
</tr>
<tr>
<td>Sγ-SβA</td>
<td>57/540</td>
<td>30/206</td>
<td>13/70</td>
<td>11.4 ± 1.5</td>
<td>16.2 ± 3.1</td>
<td>21.3 ± 6.2</td>
<td>0.76</td>
</tr>
<tr>
<td>Sγ-GβC</td>
<td>52/543</td>
<td>35/351</td>
<td>15/82</td>
<td>10.2 ± 1.4</td>
<td>15.3 ± 2.7</td>
<td>21.6 ± 5.7</td>
<td>0.71</td>
</tr>
<tr>
<td>ψβX-GβA</td>
<td>183/648</td>
<td>65/328</td>
<td>28/101</td>
<td>35.4 ± 2.8</td>
<td>22.9 ± 3.0</td>
<td>35.0 ± 7.1</td>
<td>0.65</td>
</tr>
<tr>
<td>ψβX-SβA</td>
<td>158/535</td>
<td>53/339</td>
<td>24/76</td>
<td>37.5 ± 3.3</td>
<td>26.1 ± 3.8</td>
<td>41.8 ± 9.3</td>
<td>0.62</td>
</tr>
<tr>
<td>ψβX-GβC</td>
<td>186/654</td>
<td>61/320</td>
<td>26/100</td>
<td>35.8 ± 2.8</td>
<td>21.8 ± 2.9</td>
<td>32.7 ± 6.8</td>
<td>0.67</td>
</tr>
<tr>
<td>ψβX-Gγ</td>
<td>176/641</td>
<td>66/325</td>
<td>32/104</td>
<td>34.2 ± 2.8</td>
<td>27.4 ± 3.4</td>
<td>40.6 ± 7.8</td>
<td>0.58</td>
</tr>
<tr>
<td>ψβX-Sγ</td>
<td>153/528</td>
<td>58/258</td>
<td>32/84</td>
<td>36.6 ± 3.2</td>
<td>26.5 ± 3.7</td>
<td>54.1 ± 10.9</td>
<td>0.49</td>
</tr>
<tr>
<td>ψβZ-GβA</td>
<td>179/651</td>
<td>40/244</td>
<td>18/74</td>
<td>34.3 ± 2.8</td>
<td>18.7 ± 3.1</td>
<td>28.6 ± 7.3</td>
<td>0.65</td>
</tr>
<tr>
<td>ψβZ-SβA</td>
<td>145/534</td>
<td>28/156</td>
<td>11/48</td>
<td>33.7 ± 3.0</td>
<td>21.0 ± 4.1</td>
<td>25.6 ± 8.3</td>
<td>0.82</td>
</tr>
<tr>
<td>ψβZ-GβC</td>
<td>184/557</td>
<td>42/237</td>
<td>20/72</td>
<td>35.1 ± 2.8</td>
<td>20.5 ± 3.3</td>
<td>33.3 ± 8.1</td>
<td>0.62</td>
</tr>
<tr>
<td>ψβZ-Gγ</td>
<td>171/644</td>
<td>46/245</td>
<td>23/76</td>
<td>32.8 ± 2.7</td>
<td>21.6 ± 3.3</td>
<td>38.8 ± 8.8</td>
<td>0.56</td>
</tr>
<tr>
<td>ψβZ-Sγ</td>
<td>141/527</td>
<td>38/177</td>
<td>21/57</td>
<td>33.1 ± 3.0</td>
<td>25.2 ± 4.3</td>
<td>51.2 ± 12.7</td>
<td>0.49</td>
</tr>
<tr>
<td>GβA-Hβ</td>
<td>37/328</td>
<td>29/104</td>
<td>...</td>
<td>12.2 ± 2.0</td>
<td>35.2 ± 7.1</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Gγ-Hβ</td>
<td>43/326</td>
<td>30/106</td>
<td>...</td>
<td>14.5 ± 2.3</td>
<td>37.2 ± 7.3</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>GβA-Mβni</td>
<td>51/330</td>
<td>40/102</td>
<td>...</td>
<td>17.3 ± 2.5</td>
<td>55.9 ± 10.2</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>Gγ-Mβni</td>
<td>54/326</td>
<td>40/103</td>
<td>...</td>
<td>18.7 ± 2.6</td>
<td>54.8 ± 10.2</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Hβ-Mβni</td>
<td>39/332</td>
<td>39/106</td>
<td>...</td>
<td>12.8 ± 2.1</td>
<td>50.7 ± 9.2</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

a The numerator denotes the number of synonymous or nonsynonymous differences and the denominator the number of synonymous or nonsynonymous sites. In all cases these numbers have been rounded to the nearest integers; these numbers are often not integers because a site can be partly synonymous and partly nonsynonymous.

b The numbers are estimated by Jukes and Cantor’s (1969) method (see text).

Table 2
Estimated Numbers (x 100) of Substitutions per Site between Goat Sequences in Various Gene Regions

<table>
<thead>
<tr>
<th>Region</th>
<th>β^A-β^C</th>
<th>ψβ^A-ψβ^C</th>
<th>β^A-ψβ^B</th>
<th>β^A-ψβ^X</th>
<th>β^C-ψβ^B</th>
<th>β^C-ψβ^X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introns</td>
<td>11.1 ± 1.4</td>
<td>8.1 ± 1.1</td>
<td>36.2 ± 2.9</td>
<td>34.3 ± 2.8</td>
<td>35.8 ± 2.8</td>
<td>35.1 ± 2.8</td>
</tr>
<tr>
<td>Synonymous</td>
<td>8.5 ± 3.1</td>
<td>12.7 ± 4.3</td>
<td>35.0 ± 7.1</td>
<td>28.6 ± 7.3</td>
<td>32.7 ± 6.8</td>
<td>33.3 ± 8.1</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>6.8 ± 1.5</td>
<td>7.6 ± 1.8</td>
<td>22.9 ± 3.0</td>
<td>18.7 ± 3.1</td>
<td>21.8 ± 2.9</td>
<td>20.5 ± 3.3</td>
</tr>
<tr>
<td>5' region</td>
<td>0 ± 0</td>
<td>11.1 ± 3.2</td>
<td>25.4 ± 5.3</td>
<td>25.1 ± 5.3</td>
<td>25.4 ± 5.3</td>
<td>25.1 ± 5.3</td>
</tr>
</tbody>
</table>

Note.—In the case of 5' region, the number of nucleotides compared is about 120. In the other cases the numbers are given in Table 1.
Separate duplication events. Under this assumption, however, it is difficult to explain the linkage relationship of these four genes (fig. 2a). In summary, the results of our data analysis are consistent with the block duplication hypothesis.

(W. M. Fitch [personal communication] has pointed out that under the block duplication hypothesis the γ gene, whose chromosomal location is not known yet, should be located 3' downstream from β^A. This is because β^A and β^C are more closely related to γ than to ψβ^Z and ψβ^X [fig. 2b] so that neither ψβ^Z nor ψβ^X can lie in between γ and β^A, if all of these duplicated genes and pseudogenes were produced by unequal crossing-overs [see also Fitch 1977]. This prediction of the chromosomal location of γ may provide another test for the block duplication hypothesis.)

**Transition and Transversion Substitutions**

Mutations involving single nucleotide changes can be classified into transitions and transversions. Transitions include mutations between A and G and those between T and C, and transversions include mutations between A and T, A and C, G and T, and G and C. Table 3 shows the proportions of transition differences
in three different gene regions: introns, synonymous sites, and nonsynonymous sites. As most parts of introns are apparently subject to no strong functional constraint (purifying selection), the proportion of transition differences between closely related introns would reflect approximately the proportion of spontaneous transition mutations. The average for the first four comparisons between introns shown in table 3 is 64%. This is not far from the value (56%) estimated from pseudogene sequences (Gojobori et al. 1982), because both estimates presumably have large standard errors. Thus, in nuclear DNA the proportion of transition mutations is almost two times as high as the 33% expected under random mutation. The proportion is, however, considerably lower than the 96% estimated for mammalian mitochondrial DNA (Brown et al. 1982; Aquadro and Greenberg 1983). It seems that the patterns of mutation for the two kinds of DNA are quite different.

In their study of primate mitochondrial DNA's, Brown et al. (1982) found that the proportion of transition differences decreases as the degree of sequence divergence increases. The proportions of transition differences for introns shown in table 3 also decrease with increasing degree of sequence divergence, though at a rate considerably slower than that observed by Brown et al.

Before the observed proportions of transition differences at synonymous and nonsynonymous sites are discussed, it is useful to get a rough idea about the expected proportions among synonymous mutations and among nonsynonymous mutations. For this purpose we shall use the mutation pattern estimated by Gojobori et al. (1982) and consider a random sequence, that is, a sequence with equal codon frequencies. Under these conditions the proportion of transition mutations is approximately 68% among synonymous mutations and 51% among nonsynonymous mutations (Li et al., unpublished). The former proportion is substantially higher than the average value, 56%, for all mutations, because in many codons (e.g., the two codons TTT and TTC coding for phenylalanine) a synonymous mutation can arise only from a transition mutation. None of the observed proportions for synonymous differences shown in table 3 deviates significantly from 68%, though two of them are considerably higher than expected. Similarly, none of the observed proportions for nonsynonymous differences deviates significantly from 51%, though the proportion for the first comparison is only 25%. Thus, in these goat and sheep globin genes the proportions of transition substitutions at

Table 3
Percentage of Transition Differences between Sequences

<table>
<thead>
<tr>
<th></th>
<th>GβA vs.</th>
<th>GβA, SβA,</th>
<th>Gγ vs.</th>
<th>ψβX vs.</th>
<th>GβC vs.</th>
<th>Gγ vs.</th>
<th>ψβX vs.</th>
<th>ψβX vs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GβA</td>
<td>GβA, SβA</td>
<td>Gγ</td>
<td>ψβX</td>
<td>GβC</td>
<td>Gγ</td>
<td>ψβX</td>
<td>ψβX</td>
</tr>
<tr>
<td>Intron</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>63 (22/35)</td>
<td>67 (35/52)</td>
<td>61 (42/69)</td>
<td>64 (96/150)</td>
<td>56 (98/176)</td>
<td>53 (83/158)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymous differences</td>
<td>71 (5/7)</td>
<td>78 (7/9)</td>
<td>88 (7/8)</td>
<td>72 (18/25)</td>
<td>64 (23/36)</td>
<td>67 (20/30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous differences</td>
<td>25 (2/8)</td>
<td>67 (12/18)</td>
<td>60 (12/20)</td>
<td>46 (35/76)</td>
<td>52 (33/64)</td>
<td>57 (37/65)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—For the numbers in parentheses, the numerator denotes the number of transition differences and the denominator the total number of nucleotide differences.

* Estimated number of nucleotide substitutions per site (taken from table 1).
both synonymous and nonsynonymous sites were apparently not significantly disturbed by natural selection.

The fact that the proportion of transitions is higher among synonymous mutations than among nonsynonymous mutations has an interesting consequence. It is in effect equivalent to asserting that the rate of synonymous mutation per synonymous site is higher than the rate of nonsynonymous mutation per nonsynonymous site. To see this point let us consider the codon TTC for phenylalanine. By convention, one-third of the third position of this codon is counted as a synonymous site (and two-thirds as a nonsynonymous site) because only one of the three possible single nucleotide changes at this position is synonymous. Suppose that the frequency of the transition C → T, which is synonymous, is equal to the sum of the frequencies of the two transversions C → A and C → G. Then, dividing the first frequency by one-third of a site and the sum of the last two frequencies by two-thirds of a site, one can easily show that at this position the rate of synonymous mutation per synonymous site is two times that of nonsynonymous mutation per nonsynonymous site. Note, however, that, in the case in which all of the three possible single nucleotide changes at the third position of a codon are synonymous, the rate of synonymous mutation per synonymous site at this position is simply equal to the rate of mutation per nucleotide site. For a sequence with equal codon frequencies, the ratio of the rate of synonymous mutation per synonymous site to that of nonsynonymous mutation per nonsynonymous site is 1.18, if mutation follows the pattern estimated by Gojobori et al. (1982). Under the same assumptions the rate of synonymous mutation is 1.14 times the rate of mutation per nucleotide site in a random sequence. Thus, even under the assumption that all mutations are neutral, the rate of synonymous substitution is expected to be higher than the rate of nonsynonymous substitution and higher than the rate of nucleotide substitution in a random sequence. For this reason the observation that the rate of synonymous substitution is generally higher than the rate of nonsynonymous substitution might be partly the result of nonrandom mutation.

Rate of Nucleotide Substitution
Pseudogenes, Synonymous Sites, and Introns

Miyata and Hayashida (1981) have suggested that the rate of synonymous substitution per synonymous site in functional genes is slightly lower than the rate of nucleotide substitution per site in pseudogenes. This can be tested by comparing the number of substitutions per synonymous site between $\beta^A$ and $\beta^C$ with the number of substitutions per nucleotide site between $\psi\beta^x$ and $\psi\beta^z$, because under the block duplication hypothesis the divergence time between $\beta^A$ and $\beta^C$ should be the same as that between $\psi\beta^x$ and $\psi\beta^z$. The number of substitutions per site between the latter pair (entire sequences) is 0.086 with a standard error of 0.01. This is almost the same as the number of synonymous substitutions per synonymous site between G$\beta^c$ and G$\beta^A$ but is considerably lower than the number of synonymous substitutions per synonymous site between G$\beta^c$ and S$\beta^A$ (table 1). Thus, our results do not support Miyata and Hayashida's suggestion. As noted above, the rate of synonymous mutation per synonymous site is higher than the rate of mutation per site in pseudogenes. This difference in mutation rate may be enough to compensate the effect of purifying selection on synonymous mutations. It is possible that $\psi\beta^x$ and $\psi\beta^z$ have evolved more slowly than expected or that
β^A and β^C happen to have evolved very fast. This possibility is suggested by the fact that in the intron regions the number (d) of substitutions per site between ψβ^X and ψβ^Z is substantially lower than those between Gβ^C and Gβ^A and between Gβ^C and Sβ^A (table 1). Thus, it remains to be determined whether the two rates are equal or not. For simplicity, however, we shall assume that they are equal to \( r \) substitutions per site per year; all rates are in these units with the type of site appropriate to the type of substitution (e.g., synonymous, nonsynonymous).

Under the assumption above we can estimate the \( r \) value as follows. The number of substitutions per site between ψβ^X and ψβ^Z is 0.086 and the numbers of synonymous substitutions per synonymous site between Gβ^C and Gβ^A and between Gβ^C and Sβ^A are 0.085 and 0.149. The \( d \) value thus ranges from 0.085 to 0.149, with an average of 0.107. It is known that β^C is absent in cow (Huisman 1974). Thus, it probably arose (or was lost) since the goat-cow split, which is unlikely to have occurred earlier than the mid-Miocene, about 15 million years (Myr) ago (Romer 1966). We may therefore take \( T_2 = 15 \) Myr as an upper bound for the date of the block duplication event (fig. 2b). The \( r \) value can then be estimated to be between 2.8 \( \times \) 10^{-9} and 5.0 \( \times \) 10^{-9}. (The latter may not be an upper bound if \( T_2 \) is considerably shorter than 15 Myr.) The estimated range of \( r \) is in good agreement with Li et al.’s (1981) estimate, (4.6 \( \pm \) 3.0) \( \times \) 10^{-9}, of the rate for pseudogenes. Conversely, if we assume that ψβ^X and ψβ^Z and the synonymous sites in β^A and β^C have evolved at the rate of 4.6 \( \times \) 10^{-9}, we obtain \( T_1 = 0.107/(2 \times 4.6 \times 10^{-9}) = 12 \) Myr.

The average number of synonymous substitutions per synonymous site for the comparisons between Gβ^A and Sβ^A and between Gγ and Sγ is 0.045. The divergence time (\( T_1 \)) between goat and sheep has been estimated to be between 6 and 8 Myr (L. van Valen, quoted in Langley and Fitch [1974]; Novacek 1982). If we take \( T_1 = 7 \) Myr, \( r = 3.2 \times 10^{-9} \). This is within the range of \( r \) estimated above. If we assume \( r = 4.6 \times 10^{-9} \), we obtain \( T_1 = 5 \) Myr. This is close to the estimate, 6 Myr, given by Novacek (1982).

The average number of synonymous substitutions per synonymous site for the comparisons between Gγ and Gβ^A, Gγ and Sβ^A, Gγ and Gβ^C, Sγ and Gβ^A, Sγ and Sβ^A, and Sγ and Gβ^C is 0.172. The γ gene is also present in the white-tailed deer (Kitchen and Brett 1974), which separated from goat probably in the late Oligocene, that is, about 26 Myr ago (Romer 1966). Thus, the divergence time (\( T_2 \)) between γ and β^A-β^C should be at least 26 Myr and \( r \) should be smaller than 3.3 \( \times \) 10^{-9}. This latter estimate is again within the range of \( r \) given above. If we assume \( r = 2.8 \times 10^{-9} \), the lower bound given above, we arrive at \( T_2 = 31 \) Myr for the divergence between γ and β^A-β^C.

The average number of synonymous substitutions per synonymous site between the two pseudogenes and the goat and sheep functional genes is 0.382. If we assume \( r = 4.6 \times 10^{-9} \), the divergence time (\( T_3 \)) between ψβ^X-ψβ^Z and β^A-β^C-γ is estimated to be 42 Myr. This agrees well with the estimates, 41–46 Myr, given by Cleary et al. (1981). Using the goat sequences, human β, mouse β major, and rabbit β, and considering the numbers of substitutions at the three positions of codons, W.-H. Li and T. Gojobori (unpublished) estimated that the ancestor of the two pseudogenes became nonfunctional about 34 Myr ago, that is, \( T_n = 34 \) Myr (fig. 2). This estimate is lower than that (about 41 Myr) given by Cleary et al. (1981).
Next, let us consider the rate of substitution in introns. In eight out of the 10 comparisons between the functional genes of goat and sheep (table 1), the number of nucleotide substitutions per site in introns is smaller than the number of synonymous substitutions per synonymous site, though in only two cases is the difference statistically significant. It therefore seems that the rate of substitution in introns is somewhat lower than the rate of synonymous substitution. This is possible because, as mentioned above, the rate of synonymous mutation per synonymous site is about 1.14 times higher than the rate of mutation per nucleotide site in introns. The conclusion is, however, quite tentative because the difference in rate disappears if we eliminate the sheep sequences from comparison. Moreover, Miyata et al. (1980) found that the rate of substitution in the second intron of human, rabbit, and mouse β globin genes is somewhat higher than the rate of synonymous substitution in these genes, though the rate in the first intron is substantially lower than the rate of synonymous substitution. More data are required to draw a definite conclusion. If we assume $T_1 = 7$ Myr, $T_2 = 12$ Myr, and $T_3 = 30$ Myr, and use the $d$ values given in table 1, we obtain an average rate of $3.0 \times 10^{-9}$ for introns.

Nonsynonymous Sites

Our estimates of the rates of nonsynonymous substitution for the divergence between GβA and SβA and the divergence between Gγ and Sγ are $0.9 \times 10^{-9}$ and $1.4 \times 10^{-9}$, respectively (table 4). These two rates are not statistically different from that for the divergence between Hβ and Mβmaj (table 4).

The rate of nonsynonymous substitution for the divergence between GβC and the GβA-SβA pair is approximately $2.3 \times 10^{-9}$ (table 4) if we assume that the date ($T_1$) of the block duplication event is 15 Myr ago and $2.9 \times 10^{-9}$ if we assume $T_2 = 12$ Myr. These estimates are three to four times higher than the rate of nonsynonymous substitution for the divergence between Hβ and Mβmaj. It therefore appears that the rate of nonsynonymous substitution in βA and βC has been greatly accelerated since their divergence. This conclusion is supported by two observations. First, the numbers of nonsynonymous substitutions per nonsynonymous site for the comparisons between GβC and GβA and between GβC and SβA are

<table>
<thead>
<tr>
<th>Genes Compared</th>
<th>Divergence Time (Myr)</th>
<th>Rate ($\times 10^{-9}$)</th>
<th>Acceleration Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hβ vs. Mβmaj</td>
<td>80</td>
<td>0.8 ± 0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>GβA vs. SβA</td>
<td>7</td>
<td>0.9 ± 0.5</td>
<td>1.1</td>
</tr>
<tr>
<td>Gγ vs. Sγ</td>
<td>7</td>
<td>1.4 ± 0.6</td>
<td>1.8</td>
</tr>
<tr>
<td>GβC vs. GβA, SβA</td>
<td>15</td>
<td>2.3 ± 0.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Gγ, Sγ vs. GβA, SβA</td>
<td>30</td>
<td>2.3 ± 0.3</td>
<td>2.9</td>
</tr>
</tbody>
</table>

$^a$ Hβ = human β; Mβmaj = mouse β major; GβA and SβA = goat and sheep βA; GβC = goat βC; Gγ and Sγ = goat and sheep γ.

$^b$ In all estimates the standard error includes only the error in the estimation of the number of substitutions; it does not include the error in the estimation of divergence time.

$^c$ The acceleration factor is computed using the rate of nonsynonymous substitution for the divergence between Hβ and Mβmaj as a standard.
almost as large as the number of substitutions per nucleotide site between $\psi \beta^x$ and $\psi \beta^z$ (table 1). Since pseudogenes are expected to evolve very rapidly (Kimura 1980; Li et al. 1981; Miyata and Yasanuga 1981), we can conclude that $\beta^A$ and $\beta^C$ have diverged rapidly. Second, the ratios of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site are 0.80 and 0.48 for the comparisons between G$\beta^C$ and G$\beta^A$ and between G$\beta^C$ and S$\beta^A$ (table 1). The average ratio is 0.64, considerably higher than generally observed (Miyata et al. 1980); for example, this ratio is only 0.25 for the comparison between H$\beta$ and M$\beta^{maj}$ (table 1). This again indicates that $\beta^A$ and $\beta^C$ have diverged rapidly, for the rate of synonymous substitution is known to be high and relatively constant among genes (Kimura 1977; Miyata et al. 1980). Since the number of nonsynonymous substitutions per nonsynonymous site between $\beta^A$ and $\gamma$ is similar to that between $\beta^C$ and $\gamma$, $\beta^A$ and $\beta^C$ have evolved at similar rates.

The rate of nonsynonymous substitution for the comparison between G$\gamma$-S$\gamma$ and G$\beta^A$-S$\beta^A$-G$\beta^C$ is $2.3 \times 10^{-9}$ if we assume that the divergence time ($T$) is 30 Myr (table 4). This rate is about three times higher than the rate of nonsynonymous substitution for the divergence between H$\beta$ and M$\beta^{maj}$ (table 4). This indicates that the rate of nonsynonymous substitution in $\gamma$ has also been greatly accelerated after its separation from the ancestor of $\beta^A$ and $\beta^C$. (The rate of nonsynonymous substitution in $\gamma$ is about the same as that in $\beta^A$, for the numbers of nonsynonymous substitutions between G$\beta^A$ and H$\beta$ and between G$\gamma$ and H$\beta$ are very similar [table 1].) The conclusion is also supported by the high ratios of the number of nonsynonymous substitutions per nonsynonymous site to the number of synonymous substitutions per synonymous site (table 1) and by the observation that in all of the comparisons involved the number of nonsynonymous substitutions per nonsynonymous site is comparable to the number of substitutions per nucleotide site in introns (table 1). Apparently we have another case of rapid evolution following gene duplication.

In both cases the acceleration appears to have occurred only in functionally less important regions of the genes (table 5). In the $\beta$ chain of hemoglobin, the positions with the most crucial functions are, first, the heme contacts (HC) and,

<table>
<thead>
<tr>
<th>GENES COMPARED</th>
<th>DIVERGENCE TIME (Myr)</th>
<th>HC</th>
<th>Coop</th>
<th>$\alpha_1$</th>
<th>IP</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>H$\beta$ vs. M$\beta^{maj}$</td>
<td>80</td>
<td>.4 ± 0.2</td>
<td>.1 ± .1</td>
<td>1.0 ± .5</td>
<td>1.0 ± .4</td>
<td>1.1 ± .2</td>
</tr>
<tr>
<td>G$\beta^A$ vs. S$\beta^A$</td>
<td>7</td>
<td>.0</td>
<td>.0</td>
<td>2.6 ± 2.6</td>
<td>2.1 ± 2.1</td>
<td>.5 ± .5</td>
</tr>
<tr>
<td>G$\gamma$ vs. S$\gamma$</td>
<td>15</td>
<td>2.5 ± 2.5</td>
<td>.0</td>
<td>.0</td>
<td>.0</td>
<td>1.9 ± 1.0</td>
</tr>
<tr>
<td>G$\gamma^C$ vs. G$\beta^A$, S$\beta^A$</td>
<td>15</td>
<td>.0</td>
<td>.0</td>
<td>3.4 ± 1.4</td>
<td>1.3 ± .8</td>
<td>3.4 ± .6</td>
</tr>
<tr>
<td>G$\gamma$, S$\gamma$ vs. G$\beta^A$, S$\beta^A$, G$\beta^C$</td>
<td>30</td>
<td>.2 ± 0.1</td>
<td>.0</td>
<td>.9 ± .3</td>
<td>5.4 ± 1.0</td>
<td>3.3 ± .3</td>
</tr>
</tbody>
</table>

**Note.**—See the footnotes in table 4.

*The functional groups of codon positions follow the scheme used in Goodman (1981). HC = heme contacts; Coop = the cooperative sites including sites for $\alpha_1$-$\beta_1$ contacts, Bohr effect, and 2,3-diphosphoglycerate (2,3-DPG) binding; $\alpha_1$-$\beta_1$ = $\alpha_1$-$\beta_1$ contacts; IP = interior positions; Other = remaining positions. As the goat and sheep hemoglobins have lost the ability to bind 2,3-DPG, we have included the four codon positions for the 2,3-DPG binding in the Other instead of in the Coop.*
second, those concerned with the interchain cooperativity (Coop) that facilitates oxygen delivery. These two regions in $\beta^A$, $\beta^C$, and $\gamma$ are highly conservative because no nonsynonymous substitution, except the one in the HC region of Gy, has occurred in these two regions. The third most important functional group in the $\beta$ chain is the positions for the $\alpha$, $\beta$, contacts. In this region the nonsynonymous substitution rate has been accelerated in $\beta^A$ and $\beta^C$ but not in $\gamma$ (table 5). The fourth most important functional group is the remaining interior positions (IP). The rate in this region has been accelerated in $\beta^A$ and $\gamma$. The least important parts of the $\beta$ chain are the remaining exterior positions (Other). In this region acceleration has occurred in $\beta^A$, $\beta^C$, and $\gamma$. Thus, acceleration occurred only in less important functional groups of the $\beta$ chain, particularly the exterior positions.

**Discussion**

The analysis presented indicates that the rates of nonsynonymous substitution in $\beta^A$, $\beta^C$, and $\gamma$ have been greatly accelerated after the duplication separating $\gamma$ and the ancestor of $\beta^A$ and $\beta^C$ and the duplication separating $\beta^A$ and $\beta^C$. It also indicates that acceleration occurred only in the less important parts of these genes. We believe that this pattern of acceleration can best be explained by relaxation of purifying selection because some degree of relaxation would allow acceleration to occur in the less important parts but not in the most important parts of the gene. This conclusion is also supported by the observation that the proportions of transitions among nonsynonymous differences between $\gamma$ and the $\beta^A$-$\beta^C$ pair and between $\beta^A$ and $\beta^C$ are quite similar to the expected value or, in other words, have not been significantly disturbed by natural selection (table 3). We do not exclude the possibility that a substantial fraction of the nonsynonymous substitutions in these genes are advantageous substitutions. The latter might have been responsible for the development of differential oxygen affinities among these genes (Huisman 1974; Kitchen and Brett 1974).

As we mentioned in the Introduction, Goodman (1976, 1981) and his associates (Czelusniak et al. 1982) found exceptionally high rates of amino acid substitution in the early stages of divergence between hemoglobin and myoglobin and between the $\alpha$ and $\beta$ chains and took their finding as evidence against the neutral mutation hypothesis. Wilson et al. (1977) and Kimura (1981) have questioned the reliability of the high rates estimated. Here we point out that the alleged high rates are not necessarily incompatible with the neutral mutation hypothesis because they could have arisen mainly or partly from relaxation of purifying selection. (Actually, Goodman [1981] admits the possibility of relaxation of purifying selection as a cause of the acceleration but considers positive selection for advantageous mutations to be more important.) Goodman et al.'s conclusion was based mainly on their finding that the cooperative sites have higher substitution rates than the other parts of the $\alpha$ and $\beta$ chains. However, the early substitutions at the cooperative sites might not have had any significant advantage because the cooperativity would not be effective until a certain number of sites had the right amino acids (see also Goodman 1981). Indeed, there are examples in which a single amino acid substitution in the $\alpha$ or $\beta$ chain greatly reduces the cooperativity (Bellingham 1976). Thus, random drift might have played a significant role even in the evolution of this crucial region of the $\alpha$ and $\beta$ chains, not to mention the less important parts. At any rate, the cooperative region is but a small part of the $\alpha$ and $\beta$ chains, and thus, even if all of the substitutions in this
region were advantageous ones, it does not follow that the majority of amino acid substitutions in the evolution of the \( \alpha \) and \( \beta \) chains were due to advantageous mutations. We believe that advantageous substitutions have played a vital role in the improvement of the function of hemoglobin and myoglobin, but we do not think that they can account for the majority of the amino acid substitutions in the evolution of these proteins.

Acknowledgments

We are very grateful to Drs. A. W. Nienhuis and A. Davis for kindly providing us with their unpublished sequences for the second introns of sheep \( \beta^+ \) and \( \gamma \) globin genes. This work was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED


W. M. FITCH, reviewing editor

Received July 1, 1983; revision received September 2, 1983.
Role of Cryptic Genes in Microbial Evolution

Barry G. Hall
University of Connecticut

Shozo Yokoyama
Washington University School of Medicine

David H. Calhoun
Mount Sinai School of Medicine

Cryptic genes are phenotypically silent DNA sequences, not normally expressed during the life cycle of an individual. They may, however, be activated in a few individuals of a large population by mutation, recombination, insertion elements, or other genetic mechanisms. A consideration of the microbial literature concerning biochemical evolution, physiology, and taxonomy provides the basis for a hypothesis of microbial adaptation and evolution by mutational activation of cryptic genes. Evidence is presented, and a mathematical model is derived, indicating that powerful and biologically important mechanisms exist to prevent the loss of cryptic genes. We propose that cryptic genes persist as a vital element of the genetic repertoire, ready for recall by mutational activation in future generations. Cryptic genes provide a versatile endogenous genetic reservoir that enhances the adaptive potential of a species by a mechanism that is independent of genetic exchange.

Introduction

Recent advances in nucleic acid technology have led to the detection in higher eukaryotes of DNA sequences that appear to be phenotypically silent. Because some of these silent sequences may spread within the genome by forming additional copies of themselves, and because they make no obvious contribution to the fitness of the host organism, they have been termed “selfish DNA” (Dawkins 1976; Doolittle and Sapienza 1980; Orgel and Crick 1980). Another class of non-coding DNA called pseudogenes is defined as a region of DNA that is homologous to a coding sequence but which contains mutations that would prevent its expression (Jacq et al. 1977; Lauer et al. 1980; Nishioka et al. 1980; Li et al. 1981).

The detection of pseudogenes and the concept of selfish DNA have resulted from the application of recombinant DNA technology to studies of higher eukaryotes. Silent genes are also present among prokaryotic and eukaryotic microorganisms, but in most cases their existence was not initially revealed by direct examination of the genome. Instead, these genes were first detected as the result of phenotypic changes that occurred when the silent, or “cryptic,” genes were

1. Key words: microorganisms, cryptic genes, adaptation.

Address for correspondence and reprints: Barry G. Hall, Biological Sciences Group, U-44, University of Connecticut, Storrs, Connecticut 06268.
reactivated. We define "cryptic genes" as phenotypically silent DNA sequences not normally expressed during the life cycle of an individual but capable of activation as a rare event in a few members of a large population by mutation, recombination, insertion elements, or other genetic mechanisms.

In this report we first describe two cases in which the molecular basis for the activation of a cryptic gene is understood. We next review what is known about the activation of other cryptic genes in the context of short-term adaptations and long-term evolutionary development. A unifying hypothesis of the evolutionary significance of cryptic genes in microorganisms is presented.

**Activation of the ilvG Pseudogene of Escherichia coli K-12**

The _ilvGEDA_ transcription unit of _Escherichia coli K-12_ has recently been shown to contain a naturally occurring frameshift site near the middle of the _ilvG_ gene (Lawther et al. 1981, 1982). Mutant derivatives with one-base-pair deletions or two-base-pair insertions that correct the translational alignment of the _ilvG_ gene are readily isolated. The wild-type cell produces an amino terminal fragment of the _ilvG_ gene product that appears to be rapidly degraded. The mutant _ilvG_ gene codes for α-acetohydroxyacid synthase II, an isozyme that differs in some respects from isozymes I and III that are coded by unlinked genes (Bachman and Low 1980) in separate transcription units. Isozymes I and III, but not II, are sensitive to inhibition of catalytic activity by the pathway’s end product, valine (Umbarger 1978; DeFelice et al. 1979). For this reason, the wild-type and mutant cells are readily distinguished owing to the valine-resistant growth phenotype of the mutants expressing the valine-resistant _ilvG_ coded α-acetohydroxyacid synthase II.

The _ilvG_ isozyme II is not required under most growth conditions, since isozymes I and III have adequate biosynthetic capacities. In fact, the cells containing a mutationally activated _ilvG_ gene overproduce the _ilvEDA_ gene products and wastefully excrete valine into the medium (Rowley 1953; Leavitt and Umbarger 1962). However, when a moderate concentration of valine (1 μg/ml or more) is present in the medium, the growth of the wild-type cell, but not the mutant, is inhibited. Thus, whether the wild-type or the mutant cells exhibit the advantageous phenotype depends on the environmental circumstances. Since transcription of the _ilvGEDA_ genes is from _ilvG_ to _ilvA_ (Berg et al. 1979; Subrahmanyan et al. 1980), the naturally occurring frameshift site in _ilvG_ is polar on, and therefore reduces, _ilvEDA_ expression. Since the _ilvA_ gene product is an end product inhibitable allosteric enzyme which may participate in control of _ilvGEDA_ expression (Calhoun and Hatfield 1975), the effects of the naturally occurring frameshift mutation in _ilvG_ are amplified.

In a survey of over 300 independent isolates of _E. coli_, three valine-sensitive and 353 valine-resistant strains were observed (Rowley 1953). This observation excludes a trivial explanation, namely, that the _E. coli_ K-12 valine-sensitive phenotype is artifactual, arising only from a mutation acquired in the laboratory. The notion that microbial "mutants" are laboratory freaks is firmly ingrained in our thought processes. The possibility that mutant types isolated in the laboratory might accurately resemble strains occurring either commonly or infrequently in nature is not usually given much consideration. And yet, genetic diversity is the hallmark of microbial populations. The distinction between "wild type" (by definition the beginning strain, typically taken directly from nature), and "mutant"
(by definition derived in the laboratory from the wild type) is semantically essential. The discussion at hand is more than semantic, however. In this case the valine-resistant "mutant" strain resembles the more common natural isolate. The phenotypic variability of the $ilvG$ gene observed in the laboratory parallels the phenotypic variability observed in strains freshly isolated from nature. It is, in addition, similar to the occurrence in nature of mutants and wild types among many species and for many genetic systems, as discussed in detail below.

In summary, the presence of the naturally occurring frameshift in the wild type $E. coli$ K-12 $ilvG$ gene has potentially beneficial consequences, including (i) elimination of the $ilvG$ coded isozyme II that is not essential under many growth circumstances; (ii) the efficient allosteric control by valine, a pathway end product, since isozymes I and III are inhibited by valine; (iii) expression of $ilvEDA$ at less than maximum levels that are, nevertheless, adequate for maximal growth rate; (iv) altered levels of production of the $ilvA$ gene product, threonine deaminase, a protein with both allosteric and genetic regulatory properties; and (v) production of pathway end products at levels that are ample for growth but below levels that lead to futile overproduction and excretion. When the $ilvG$ pseudogene carried by the wild-type $E. coli$ K-12 cells is activated by spontaneous mutations, isozyme II is produced and the cell is phenotypically valine resistant. This selective advantage in the presence of valine is at the expense of economical and balanced amino acid synthesis in the absence of valine as detailed in ii–v above.

**Activation of the bglB SRC Operon of Escherichia coli K-12**

In several cases cryptic genes have been identified when bacterial species have acquired new metabolic functions by mutation.

A well-studied example is the phospho-β-glucosidase system of the family Enterobacteriaceae. The β-glucosides are metabolized by a complex system of permeases and phospho-β-glucosidases with different substrate specificities. *Klebsiella* species possess a complete set of permeases and hydrolases and consequently metabolize the aryl β-glucosides arbutin and salicin and the disaccharide cellobiose. Most *Citrobacter* strains utilize cellobiose but do not utilize the aryl β-glucosides, while most *Proteus vulgaris* strains utilize arbutin and salicin but not cellobiose. *Salmonella* and *Escherichia coli* are unable to utilize any β-glucoside sugars (Schaeffer and Mintzer 1959; Schaeffer and Malamy 1969). Classically these phenotypes would be explained by the loss of genetic information: partial loss in the case of *Proteus* and *Citrobacter* and complete loss in the cases of *Salmonella* and *E. coli*. Both *Salmonella* and *E. coli*, however, mutate easily to β-glucoside positive phenotypes, in the case of *E. coli* at spontaneous frequencies as high as $10^{-5}$ (Schaeffer and Mintzer 1959; Schaeffer and Schenkien 1968; Reynolds et al. 1981). Interestingly, the β-glucoside positive mutants do not acquire the ability to utilize the full range of β-glucoside sugars. Mutants of *Salmonella* utilize cellobiose but not the aryl β-glucosides, while mutants of *E. coli* utilize arbutin and salicin but not cellobiose (Schaeffer and Mintzer 1959; Schaeffer and Schenkien 1968; Schaeffer and Malamy 1969; Reynolds et al. 1981).

The β-glucoside system was initially investigated in *E. coli* K-12 by Schaeffer and his colleagues (Schaeffer 1967; Schaeffer and Maas 1967; Prasad and Schaeffer 1974) and has recently been investigated by others (Defez and DeFelice 1981; Reynolds et al. 1981; DiNardo et al. 1982). The gene for phosphoglucosidase A ($bgLA$) is expressed constitutively in wild-type cells, but the genes for phospho-
glucosidase B (bglB), for the transport system (bglC), and for the β-glucoside dependent positive regulatory protein (bglS) are not expressed. Since the hydrolytic enzymes act only on phosphorylated β-glucosides, only those cells that express the phosphoenolpyruvate dependent β-glucoside transport system specified by the bglC gene can utilize β-glucosides. Mutations in the cis-acting bglR site are required for the inducible expression of the bglBSRC operon. It appears that the bglR mutations create a site for the activation of transcription of the operon (Prasad and Schaefler 1974). Spontaneous bglR mutations arise at a frequency of about $10^{-5}$ and are the consequence of integration of insertion sequence IS1 or IS5 into a specific region of the chromosome. Two models may explain the insertional activation of the bgl operon (Defez and DeFelice 1981; Reynolds et al. 1981). According to the first model, the operon lacks a functional promoter, and the insertion elements contribute to the formation of a promoter. Neither of these elements is thought to contain a recognizable promoter itself; however, this conclusion has recently been questioned (Ciampi et al. 1982). According to the second model, the operon contains an operator site which is disrupted by the insertions. This model implies the existence of a repressor which is refractory to β-glucosides as inducers. The second model is supported by the discovery of a class of mutations which are sufficient to activate the operon even in cells with a wild-type bglR gene (Defez and DeFelice 1981). These mutations are located at 27 min on the E. coli map, on the opposite side of the chromosome from the bgl operon (83 min). Two such mutations were shown to be amber mutations and to be recessive to the wild-type allele. The new gene is designated bglY, and the authors suggest that it specifies a repressor which prevents transcription of the bgl operon.

The bgl operon is thus cryptic, but it can be activated by mutations in a variety of sites. In addition to mutations in bglR and bglY, it has recently been reported that the operon can also be activated by mutations in gyrA and gyrB, the genes for DNA gyrase (DiNardo et al. 1982). The authors suggest that expression of the bgl operon may be controlled by the degree of local supercoiling of the DNA.

The cryptic β-glucoside system is not limited to laboratory strains of E. coli. When 17 “wild” E. coli strains, isolated from both humans and a variety of animals from several locales in the United States (Selander and Levin 1980), were tested, all were unable to utilize β-glucosides. All of those strains, many of which had been maintained at −70 °C since their isolation, yielded spontaneous arbutin-utilizing mutants (Hall, unpublished observations). This observation, together with earlier reports (Schaefer and Schenkien 1968; Schaefer and Malamy 1969), indicates that the cryptic gene was present at the time of, and has been retained since, the evolutionary divergence that led to the modern E. coli species.

Naturally Occurring Auxotrophs

The serological and fermentation patterns exhibited by clinical isolates of Neisseria gonorrhoeae are adequate to identify the species but inadequate for detailed epidemiological studies. It was observed, however, that the nutritional growth requirements for amino acids, vitamins, purines, and pyrimidines vary widely among clinical isolates. Accordingly an “auxotyping” method based on these differing growth requirements is currently used to identify strains that are indistinguishable by other methods (Carfio and Catlin 1973; Eisenstein et al. 1977;
Knapp et al. 1978; Juni and Heym 1980). It is of interest that "revertants" to prototrophy for most of these growth factors (e.g., isoleucine, thiamine pyrophosphate) are observed in the laboratory (Carfio and Catlin 1973; Catlin 1973; Eisenstein et al. 1977; Knapp et al. 1978; Juni and Heym 1980). The occurrence of these revertants provides evidence that the essential genes were present in a cryptic state, inactivated by mutation. The auxotype of clinical isolates is almost always stable during repeated laboratory passage and during human infections and human-to-human transmission. So, even though a few revertants are commonly observed, the predominant auxotype initially present is only rarely observed to vary in nature.

The genus *Lactobacillus* is well known for its multiple nutritional requirements, including amino acids, purines, pyrimidines, and vitamins. In a study of the genetic basis for these requirements, Morishita et al. (1974) began with a strain of *L. casei* that required 12 amino acids and four vitamins, and they isolated mutants that could grow in the absence of a specific nutrient. They were able to isolate such mutants with respect to seven of the 12 amino acids and three of the four vitamins at frequencies expected for single-step mutations. Similar results were obtained with a strain of *L. acidophilus*, indicating the generality of the phenomenon. The nature of the mutations is unknown, but their existence makes it clear that the original auxotrophy did not result from multiple lesions or from irreversible loss of information via deletions.

Reversion of naturally occurring auxotrophic characters has been seen with other microbial species, including *Salmonella* sp. (Lederberg 1947) (several amino acids) and *Pasteurella pestis* (Englesberg and Ingram 1957) (several amino acids). The tryptophan requirement of *Shigella dysenteriae* was demonstrated by DNA sequencing to be due to a "down" promoter mutation (Miozzari and Yanofsky 1978) and two mutations (detected genetically) in the *trpE* structural gene coding for anthranilate synthase (Manson and Yanofsky 1976). The remainder of the *S. dysenteriae trp* operon was shown to be functionally intact (Manson and Yanofsky 1976).

### Acquisition of New Metabolic Capabilities

Metabolic phenotypes play a key role in the taxonomy of microorganisms, and one of the classical properties of *Escherichia coli* is its inability to utilize citric acid as a sole carbon and energy source (Bergery's manual 1974). Although citrate-utilizing *E. coli* have occasionally been isolated from nature (Isiguro et al. 1978), the genes for citrate utilization have invariably been plasmid borne, usually on drug resistance plasmids. A spontaneous citrate-utilizing mutant of *E. coli K-12* has recently been isolated (Hall 1982a). Citrate utilization requires mutations in two chromosomal genes, one on each side of the *gal* operon; and the mutant possesses a semiconstitutive citrate transport system. The properties of that transport system are quite distinct from those of at least one plasmid specified citrate transport system (C. Reynolds, personal communication). The parental strain, like other *E. coli*, is unable to transport citrate. Thus, the information required for a citrate transport system has probably been retained in a cryptic form since *E. coli* diverged from its Cit⁺ progenitor.

*Klebsiella* sp. normally possess two *lac* operons, one chromosomal and the other plasmid borne (Reeve and Braithwaite 1974). A mutant strain that had irreversibly lost both of its *lac* operons mutated spontaneously to Lac⁺ (Hall
114 Hall, Yokoyama, and Calhoun

1979). In the mutant strain lactose utilization involved a phosphoenolpyruvate-dependent transport system (Imai and Hall 1981), with the resulting lactose-6-phosphate being metabolized by a phospho-β-galactosidase (Hall 1979). It now appears that the critical mutation to Lac⁺ involved decryptification of a gene specifying a lactose specific enzyme II of the PEP-dependent transport system (Hall et al. 1982).

The cases above should not be confused with cases in which new metabolic capabilities arose via point mutations in regulatory genes or via point mutations that altered the catalytic specificities of enzymes (Clarke 1978; Hall 1983).

Retention of Cryptic Genes

The phenomenon of cryptic genes seems to be very general. It seems likely that most microbial species carry genes that are not expressed at a physiologically functional level and which therefore do not make a positive contribution to fitness. It is expected that mutations in such genes will eventually lead to permanent inactivation (Dykhuizen 1978). Depending on the length of time since the gene became cryptic, an increasing portion of the population is expected to lose the ability to decryptify the fully functional gene. Indeed, there is an example of such a scenario: *Shigella dysenteriae*, a close relative of *E. coli*, is typically Lac⁻ owing to absence of the gene for a lactose permease (*lacY* gene). It does possess a *lacZ* (β-galactosidase) gene that is homologous with the *lacZ* gene of *E. coli*, and a fully functional repressor (*lacI*) gene. In the absence of the permease gene, the β-galactosidase gene plays no functional role in lactose utilization, and mutations in *lacZ* are expected to accumulate. In fact, the *lacZ* gene of *S. dysenteriae* produces levels of β-galactosidase from 0% to 20% of that of *E. coli*, and the enzymes are more thermolabile than *E. coli* β-galactosidase (Luria 1965).

In view of the above we might ask why cryptic genes are retained within populations in functional forms. (The *bglC* gene has apparently been cryptic since the divergence of *E. coli*, yet a majority of the population has clearly retained the functional gene.)

Several possible mechanisms for retention are considered below, including (i) a growth advantage that derives from a superior metabolic regulatory system; (ii) occasional strong environmental selection for population members that express a cryptic gene, leading to survival only of population members that have retained the cryptic gene in a functional form; (iii) an unknown advantage to the cryptic state as demonstrated for auxotrophy; (iv) toxic intermediates that are present in special circumstances, for example, *galE* mutants, unless another gene is made cryptic; and (v) the possible role of the cryptic gene as a metabolic control mechanism that operates at the level of a population of cells rather than within each cell of a population.

Potential Advantages Due to a Cryptic Gene

The specific detailed biochemical and genetic information available for the cryptic *ilvG* gene provides evidence indicating the selective advantages of the cryptic and activated states of *ilvG* depending on environmental conditions. It seems clear that it is advantageous for a population of *Escherichia coli* to retain the *ilvG* pseudogene for activation during conditions of valine-induced growth inhibition. The spontaneous frequency of this mutational activation (10⁻⁷—10⁻⁸) is high enough to ensure that *E. coli* populations of ordinary size would include
members that express the ilvG coded isozyme. The disadvantages of expressing ilvG would appear to favor a return to the cryptic state in the absence of valine. The observation that the majority of natural isolates are valine resistant, and thus presumably express the ilvG gene, may indicate that under natural conditions the equilibrium favors the active state; or it may well be that the populations tested are not at equilibrium with respect to this trait and that the frequency of the cryptic gene is currently increasing. In either case, the frequency of valine-sensitive strains in nature is high enough (≈1%) to indicate that neither the cryptic nor the active state of ilvG is simply a laboratory artifact. Cryptification of the active ilvG gene in response to an environmental selection pressure is not merely an assumption. Selection for mutants without ilvG activity (by counterselection against the valine-resistant phenotype) yielded two types (Smith et al. 1979). One type has high ilvEDA expression at levels identical to the parent and therefore had nonpolar ilvG mutations. The other type had low ilvEDA expression at levels equal to the wild-type K-12 strain containing the pseudogene and can be presumed to have had polar mutations, as is true of the original ilvG pseudogene.

The case of the Lactobacilli provides perhaps the best example of the retention of nonfunctional genes. The evidence suggests that the genes for many of the biosynthetic pathways for amino acids, purines, and so on contain lesions in the form of single point mutations. Under these conditions the remaining genes in the defective pathways perform no useful function and should be lost as additional mutations occur. For one such pathway such mutations have indeed accumulated: strict tryptophan prototrophs could not be isolated from L. casei; however, it was possible to isolate mutants that used the precursors anthranilate or indole (Morishita et al. 1974). This suggests that there were multiple or irreversible lesions prior to anthranilate synthesis. How have the other genes escaped inactivation? We propose that periodically in the history of the population there has been strong selection for decryptification of the genes or pathways described. This selection would presumably be similar to the laboratory selection that revealed their existence. During such periods those members of the population that had accumulated inactivating mutations in the now-vital genes would be lost.

If there is occasional strong selection for the expression of normally cryptic genes, we should expect that the genes would remain decryptified even after those selective conditions have ceased to exist. We are thus confronted by an apparent paradox: if there is occasional selection for expression of "cryptic" genes, why are they cryptic at all? To resolve this paradox it is necessary to propose strong selection for cryptification of those genes under circumstances where the gene product is not required.

Formally, then, we propose that under one set of conditions members of the population with a cryptic gene are more fit than those members who express the gene in question, while under some alternative set of conditions those members who express the gene are at a strong selective advantage.

In addition to the intuitive basis above, there is a mathematical basis for the proposal that some type of selective advantage is operating to retain cryptic genes in the population.

**Equilibrium Frequencies of Cryptic Genes**

We distinguish three classes of genes in a haploid population: (i) cryptic genes, \( A_1 \); (ii) functional genes, \( A_2 \); and (iii) nonfunctional or irreversibly inactivated
genes, $A_3$. Let us denote the frequencies of $A_1$, $A_2$, and $A_3$ by $x_1$, $x_2$, and $x_3$, respectively. Assume that the rates of mutation from $A_1$ to $A_2$ and from $A_2$ to $A_1$ in each generation are $v_1$ and $v_2$, respectively. Furthermore, we assume that the rates of irreversible mutation from $A_2$ to $A_3$ and from $A_1$ to $A_3$ are the same and $\mu$ per generation. Let $1 - s$, and $1 - t$ be the relative fitnesses of $A_1$, $A_2$, and $A_3$, respectively. Thus, we assume constant fitness. As already noted, the fitness may vary from time to time owing to environmental change. Under that condition, the fitness in the present model can be interpreted as the mean values of the fitnesses through time.

After the processes of mutation and selection, the changes in the gene frequencies are given by

$$
\Delta x_1 = x_1 (s x_2 + t x_3) / W - (\mu + v_1) x_1 + v_2 x_2,
$$

$$
\Delta x_2 = x_2 [-s(1-x_2 + t x_3)] / W + v_1 x_1 - (\mu + v_2) x_2,
$$

$$
\Delta x_3 = x_3 [s x_2 - t(1-x_3)] / W + \mu x_1 + \mu x_2,
$$

where $W = 1 - sx_2 - tx_3$.

Before going into detail, we should note that if the frequency of nonfunctional genes in a population becomes unity, no further evolutionary changes occur unless there is migration or other forces which introduce functional or cryptic alleles into that population. Thus, it is important to study the existence of nontrivial equilibria and stability of these points.

In an equilibrium population, it is sufficient to consider the following two equations:

$$\hat{x}_1(s \hat{x}_2 + t \hat{x}_3)/\hat{W} - (\mu + v_1)\hat{x}_1 + v_2\hat{x}_2 = 0 \quad (2a)$$

and

$$\hat{x}_3[s \hat{x}_2 - t(1-\hat{x}_3)]/\hat{W} + \mu(\hat{x}_1 + \hat{x}_3) = 0 \quad (2b)$$

where $\hat{x}_1$, $\hat{x}_2$, $\hat{x}_3$, and $\hat{W}$ are the equilibrium values of $x_1$, $x_2$, $x_3$, and $W$, respectively.

When $s = t = 0$, it is clear that $\hat{x}_1 = \hat{x}_2 = 0$, and $\hat{x}_3 = 1$ from equations (2a) and (2b). Thus, without selection, nonfunctional genes will eventually fix in the population.

To solve the equations (2a) and (2b), it is convenient to write (2b) as

$$(1 - \hat{x}_3)[\mu - (t - \hat{a})]/\hat{W} = 0 \quad (3)$$

where $\hat{a} = s \hat{x}_2 / (\hat{x}_1 + \hat{x}_2)$. Thus, two sets of equilibrium gene frequencies exist. One is the trivial equilibrium point $\hat{x}_3 = 1$. A nontrivial equilibrium gene frequency can be obtained by solving

$$\mu - (t - \hat{a})\hat{x}_3 / \hat{W} = 0 \quad (4)$$

where $\hat{W} = 1 - s\hat{x}_2 - t\hat{x}_3$. Putting (4) into (2a), we obtain

$$s\hat{a}\hat{q}/\hat{W} - v_1\hat{q} + v_2\hat{q} = 0 \quad (5)$$

where $\hat{q} = \hat{x}_2 / (\hat{x}_1 + \hat{x}_2)$ and $\hat{q} = \hat{x}_3 / (\hat{x}_1 + \hat{x}_2)$. Furthermore, by using equation (4) and definitions of $\hat{a}$, $\hat{p}$, and $\hat{q}$, we can reduce $\hat{W}$ to

$$\hat{W} = (1 - s\hat{q})/(1 + \mu) \quad (6)$$
From (5) and (6),
\[ A\dot{q}^2 - B\dot{q} + v_1 = 0, \]  
where \( A = s(1 + \mu + v_1 + v_2) \) and \( B = s(1 + \mu + v_1) + v_1 + v_2. \) Thus, with the condition that \( 0 < \dot{q} < 1, \)
\[ \dot{q} = \begin{cases} 
\frac{v_1}{(v_1 + v_2)} & \text{if } s = 0, \\
\frac{(B - \sqrt{B^2 - 4v_1A})/(2A)}{v_1} & \text{if } s \neq 0. 
\end{cases} \]

Once we obtain \( \dot{q}, \) from (4),
\[ \dot{x}_3 = \mu(1 - s\dot{q})/[{(1 + \mu)(t - s\dot{q})}]. \]
Then, with the definitions of \( \dot{p} \) and \( \dot{q}, \)
\[ \dot{x}_1 = (1 - \dot{q})(1 - \dot{x}_3), \]
and
\[ \dot{x}_2 = \dot{q}(1 - \dot{x}_3). \]

From (8a) to (11), we can derive nontrivial equilibrium gene frequencies depending on the values of \( s. \) Note that there is only one equilibrium point. When \( s = 0 \) and \( t > \mu/(1 + \mu), \)
\[ \dot{x}_1 = v_2/(v_1 + v_2)[1 - \mu/{[(1 + \mu)r]}], \]
\[ \dot{x}_2 = v_1/(v_1 + v_2)[1 - \mu/{[(1 + \mu)r]}], \]
\[ \dot{x}_3 = \mu/[{(1 + \mu)r}]. \]

Thus, when there is no selective difference between the functional and cryptic genes but these genes have higher fitness than nonfunctional genes, and when \( t > \mu/(1 + \mu), \) then nontrivial equilibrium frequencies of cryptic genes depend mainly on the forward and backward mutation rates between the functional and cryptic genes. Under this condition, the frequency of cryptic genes can be close to unity if the mutation rate from \( A_2 \) to \( A_1 \) is much higher than that from \( A_1 \) to \( A_2. \)

It may be useful to study formulas (8a)-(11) under special circumstances for \( s \neq 0. \) When \( s \) and \( t \gg \mu, v_1, \) and \( v_2, \) then \( \dot{q} \approx v_1/s, \) and, therefore, \( \dot{x}_1 \approx (v_1/s)(1 - \mu/t), \dot{x}_2 \approx (v_1/s)(1 - \mu/t), \) and \( \dot{x}_3 \approx \mu/t. \) Thus when cryptic genes have higher fitness than functional and nonfunctional genes, the frequency of the cryptic genes is given by \( (1 - v_1/s)(1 - \mu/t), \) and this can be close to unity, as expected. When \( s \gg \mu, v_1, \) and \( v_2, \) then \( \dot{x}_1 \approx (1 - v_1/s)\mu/(t - v_1), \dot{x}_2 \approx (v_1/s)(1 - \mu/\mu/\mu/(t - v_1), \) and \( \dot{x}_3 \approx \mu/(t - v_1) \) as long as \( t > v_1. \) Thus, as the value of \( t \) gets smaller, the frequency of cryptic genes becomes smaller. Let us now consider the case of \( t = 0, \) that is, the case where there is no selective difference between cryptic and nonfunctional genes. In this case, equation (4) reduces to \( \mu + \delta \dot{x}/\dot{W} = 0, \) or \( \mu + s\dot{x}_2\dot{x}_3/[(\dot{x}_1 + \dot{x}_3)\dot{W}] = 0. \) Thus, when \( s > 0, \) then there is no nontrivial equilibrium point. When \( s < 0, \) however, a nontrivial equilibrium frequency of cryptic genes exists. For example, by setting \( s' = -s, \dot{x}_1 \approx (v_3/s')[1 - \mu/(1 + s')/s'], \dot{x}_2 \approx (1 - v_3/s')\mu/(1 + s')/s', \) and \( \dot{x}_3 \approx \mu/(1 + s')/s' \) as long as \( s' > \mu/(1 - \mu). \)
Some numerical examples of the equilibrium gene frequencies are given in table 1, where \( \mu = v_1 = v_2 = 10^{-3} \) were used. These numerical examples show that the frequency of cryptic genes becomes high when the values of \( s \) and \( t \) are large, as expected. We should, however, point out that even when selection coefficients \( s \) and \( t \) differ by only an order of magnitude from the mutation rates, the frequency of cryptic genes is still high. For example, when \( s = t = 10^{-4} \) and \( \mu = v_1 = v_2 = 10^{-5} \), then \( \hat{x}_1 \approx 0.81 \) and \( \hat{x}_3 \approx 0.11 \) (see table 1). When \( s = 0 \) and \( t > 0 \), the frequency of cryptic genes is given approximately by \( v_2/(v_1 + v_2) \) (see eq. [12]). Thus, if \( v_1 = v_2 \), that frequency is 0.5 and still substantial. Although the frequency of cryptic genes is much lower for \( s < 0 \) (i.e., cryptic alleles are less fit than functional alleles), cryptic alleles do persist in the population. For example, the values of \( x_i \) are given by 0.123, 0.080, and 0.010 for \( s = -0.00005 \), \(-0.0001 \), and \(-0.001 \), respectively (table 1).

Stability of the equilibrium values in table 1 can be studied numerically using equation (1). The analyses show that these equilibrium points are locally stable. Thus, when the gene frequencies deviate from an equilibrium point they will return to that point. For example, consider the case of \( s = 0.01 \) and \( t = 0.001 \) with \( \mu = v_1 = v_2 = 10^{-5} \). Suppose that \( x_1 = 0.01 \), \( x_2 = 0.9 \), and \( x_3 = 0.09 \) in the initial population, then \( (x_1, x_2, x_3) = (0.1494, 0.0834, 0.7671), (0.9316, 0.0008, 0.0674), \) and \( (0.9885, 0.0010, 0.0105) \) after 500, 5,000, and 10,000 generations. At generation 12,065, \( x_1 = 0.9889 \), \( x_2 = 0.0010 \), and \( x_3 = 0.0101 \), which is very close to the equilibrium values given in table 1.

These deterministic analyses show that the frequency of cryptic genes can be close to unity even when the cryptic genes have only slightly higher fitnesses than the functional genes. The analyses also show that all three classes of genes can be maintained even if the functional genes have higher fitness than the cryptic and nonfunctional genes.

### Table 1

**Numerical Examples of Equilibrium Gene Frequencies**

<table>
<thead>
<tr>
<th>( s )</th>
<th>( t )</th>
<th>( \hat{x}_1 )</th>
<th>( \hat{x}_2 )</th>
<th>( \hat{x}_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>.01</td>
<td>.9980</td>
<td>.0010</td>
<td>.0010</td>
</tr>
<tr>
<td>.0001</td>
<td>.9889</td>
<td>.0010</td>
<td>.0101</td>
<td></td>
</tr>
<tr>
<td>.0001</td>
<td>.8880</td>
<td>.0009</td>
<td>.1111</td>
<td></td>
</tr>
<tr>
<td>.0005</td>
<td>.7493</td>
<td>.0008</td>
<td>.2499</td>
<td></td>
</tr>
<tr>
<td>.00003</td>
<td>.4997</td>
<td>.0006</td>
<td>.4997</td>
<td></td>
</tr>
<tr>
<td>.001</td>
<td>.9890</td>
<td>.0010</td>
<td>.0010</td>
<td></td>
</tr>
<tr>
<td>.001</td>
<td>.9801</td>
<td>.0099</td>
<td>.0100</td>
<td></td>
</tr>
<tr>
<td>.001</td>
<td>.8802</td>
<td>.0088</td>
<td>.1110</td>
<td></td>
</tr>
<tr>
<td>.0001</td>
<td>.9110</td>
<td>.0880</td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>.001</td>
<td>.9007</td>
<td>.0892</td>
<td>.0101</td>
<td></td>
</tr>
<tr>
<td>.0001</td>
<td>.8099</td>
<td>.0802</td>
<td>.1099</td>
<td></td>
</tr>
<tr>
<td>.00005</td>
<td>.6879</td>
<td>.0681</td>
<td>.2440</td>
<td></td>
</tr>
<tr>
<td>.0003</td>
<td>.4764</td>
<td>.0472</td>
<td>.4764</td>
<td></td>
</tr>
<tr>
<td>-.00005</td>
<td>.1230</td>
<td>.6385</td>
<td>.2385</td>
<td></td>
</tr>
<tr>
<td>-.0001</td>
<td>.082</td>
<td>.8099</td>
<td>.1099</td>
<td></td>
</tr>
<tr>
<td>-.001</td>
<td>.0098</td>
<td>.9801</td>
<td>.0101</td>
<td></td>
</tr>
</tbody>
</table>

**Note.** In the computation \( \mu = v_1 = v_2 = 10^{-3} \) were used.
The present mathematical analyses and the available data strongly suggest that some types of selective advantage of the cryptic gene must be operating over functional genes. This theoretical treatment is supported by a number of concrete observations.

There are examples of selection against organisms that express unneeded functions, although most are anecdotal. Morishita et al. (1974) point out that when the mutant strains of *Lactobacillus casei* that had become independent of various nutrients were grown for many generations on a complex medium, they often reverted to the original phenotype. Likewise, Baumann and Baumann (1981) have pointed out that many *Vibrio* species acquire nutritional requirements when stored on complex media. Beckwith (personal communication) has found that when Lac constitutive strains of *E. coli* are stored for long periods on rich medium they accumulate point mutations in the *lacZ* gene. Indeed, it is a common word-of-mouth observation that bacteria which grew quite well on minimal media when originally isolated from nature often acquire auxotrophies when stored or repeatedly transferred on a rich medium.

### The Advantage of Auxotrophy

In 1944 Lwoff predicted that in the presence of the end product an auxotroph would have a selective advantage over the prototroph (Lwoff 1944). Later, Zamenhoff and Eichorn (1967) studied this question by competing prototrophs with auxotrophic mutants of *Bacillus subtilis* in glucose-limited chemostats. They found that in the presence of the required amino acid, histidine and tryptophan auxotrophs had a strong selective advantage over prototrophs. They attributed the selective advantage of auxotrophy to energy and resource conservation, but in 1978 Dykhuizen (1978) cast serious doubt on this explanation for auxotrophic advantage. Placing tryptophan auxotrophs of *Escherichia coli* in competition with isogenic prototrophs in glucose-limited chemostats, he also found that auxotrophs had a strong selective advantage. He calculated the percentage of the energy budget devoted to tryptophan biosynthesis in repressed feedback-inhibited cells as 0.01%. He thus predicted, first, that if energy saving is the basis of auxotrophic advantage, the selective advantage should not exceed 0.01%. Second, he predicted that a nonsense mutation in *trpE*, which permits synthesis of neither tryptophan nor any of the *trp* proteins, should have an advantage over a *trpE* missense mutation which permits synthesis of the proteins but not tryptophan. Third, he predicted that the selective advantage of the auxotroph should be greatest when supplied with tryptophan, less when supplied with indole, and least when supplied with anthranilate. When tested by appropriate competition experiments in chemostats, none of the predictions was borne out: the selection for the auxotrophs was three orders of magnitude greater than could be explained on the basis of the energy budget; the selective advantage of the polar mutations was indistinguishable from that of the missense mutations; and there was no difference in the selection for the missense mutant on indole or anthranilate. In short, all of the tests failed to support the energy conservation hypothesis.

Although it does not appear to be based on energy conservation, there is a clear and consistent selective advantage to auxotrophy under conditions where the nutrient is in the environment. How, then, might auxotrophy be advantageous? We can imagine that some intermediates in anabolic pathways might be slightly toxic or might slightly perturb the regulation of some other pathway. The presence
of such an intermediate when the pathway end product is not required would slightly reduce growth rates below those that would be achieved in a mutant that no longer expressed the pathway.

An example from a laboratory situation suggests that in some circumstances it could be advantageous to cryptify catabolic genes. Escherichia coli mutants that are defective in the galE gene, the gene for UDP-galactose epimerase, are killed by galactose owing to the accumulation of highly toxic UDP-galactose (Ippen et al. 1971). In galE− strains, lactose negative and melibiose negative mutations are strongly selected in the presence of lactose or melibiose because metabolism of these substrates results in the generation of internal galactose. Although this could be viewed as simply an artifactual laboratory situation, it must be realized that the wide variety of metabolic capabilities that we now regard as “wild type” and as defining the various bacterial species might well be viewed as genetic deficiencies if we were aware of the phenotypes of the ancestral organisms from which they derived.

Another possibility is that an unneeded protein might act directly to reduce fitness by interacting inappropriately with other cellular components. A transport protein that competed for a limited number of sites in the membrane could easily reduce fitness when it was no longer required. Dykhuizen and Davies (1980) have shown that lactose use interferes with the concurrent use of maltose and have provided evidence supporting the hypothesis that this interference is due to competition for a limited number of membrane sites by the lactose and maltose permeases. As another example, the lactose permease has been implicated in “lactose killing” under unusual environmental circumstances (Dykhuizen and Hartl 1978).

If expression of a gene could reduce fitness, why wouldn’t the organism simply evolve a regulatory system to prevent expression of that gene when it was not required? The results of various directed evolution studies (reviewed in Clark [1978]; Hall [1982b, 1983]) suggest that microorganisms do not evolve “optimal” solutions to problems but simply take the first solution that solves the problem. For many systems it is probably far easier to cryptify a gene than it is to evolve a regulatory system.

Cryptic Genes as a Metabolic Control System

The activation of cryptic genes by mutation is a regulatory event that permits the expression of an otherwise silent gene. Other mechanisms that regulate gene and enzyme activity, such as induction, repression, and allosteric control, are most effective for those circumstances requiring frequent or continuous modulation during the lifetime of most individual members of a species. This physiological regulation generates a homogeneous adaptation to an environmental change, that is, the genetic constitution of the population remains unchanged. Cryptic gene activation operates at a higher hierarchical level, affecting only a very few members of a species. Environmental changes may occur in a repetitive fashion over long evolutionary time spans. Indeed, such changes typify the evolutionary history of many extant species. We view cryptification and decryptification of genes as adaptations to these cyclic environmental changes. These adaptations occur more frequently than, and are superimposed on, long-term evolutionary changes that may occur during periods of relative stasis (Cronin et al. 1981). One consequence of an environmental change that selects for expression of a cryptic gene would be a sharp reduction in the genetic variability of the population, similar
to that which occurs during periodic selection in chemostats. Since microbial populations probably enjoy little genetic exchange in natural populations (Selander and Levin 1980), such cyclic changeovers are expected to result in populations with a very limited number of distinctive genotypes. That expectation appears to be fulfilled for natural *Escherichia coli* populations. In a survey of 109 clones from geographically dispersed natural populations, based on electrophoretic mobilities of 20 different enzymes, Selander and Levin (1980) detected three pairs of identical clones isolated from hosts that were separated by wide geographic distances. Since the mean genetic diversity of that population was very high (0.47), they concluded that the natural population of *E. coli* consists of a very limited number of clones. The activation of cryptic genes is, therefore, an event with both regulatory and evolutionary implications. The existence of various mechanisms for activating cryptic genes emphasizes the widespread nature of the phenomenon.

We view our proposed role for cryptic genes as being applicable primarily to microorganisms. First, null alleles are rare in populations of higher eukaryotes. Second, because the somatic tissue provides a buffer between gene transmission and environment, higher eukaryotes are less subject to the boom-or-bust of variation in substrates that affects microorganisms. Eukaryotic microorganisms such as yeasts, however, should be subject to the same selective pressures as those encountered by bacteria, and we fully expect that those species also contain cryptic genes. Consistent with this expectation, Carlson et al. (1981) have shown that *Saccharomyces cerevisiae* strains often possess silent genes for sucrose metabolism and that these alleles can be mutationally activated.

The ubiquity of cryptic gene activation in microorganisms raises questions about the possibility that genes in higher forms may similarly be activated in a small fraction of the individuals of a population. It is noteworthy in this regard that the activation of vertebrate genes as a result of environmental carcinogenesis or by genetic insertion of tumor viruses has recently been documented (Goldfarb et al. 1982; Reddy et al. 1982; Tabin et al. 1982). There is insufficient evidence available at this time, however, to propose that the mechanisms that lead to vertebrate gene activation serve the same positive evolutionary roles as the type we propose for microorganisms. However, there is no evidence to the contrary, and the possibility should not be excluded in the absence of further experimental evidence.

**Acknowledgments**

We are indebted to L. Sarokin for first raising the question about retention of cryptic genes in bacterial populations, to J. Antonovics for very helpful discussions, and to J. M. Calvo for his particularly insightful comments on an earlier version of this paper. We are also grateful to Wen-Hsiung Li, Bruce Levin, and Fumio Tajima for their comments on the theoretical aspects of this work. Fumio Tajima’s contribution on the derivation of exact solutions from equation (1) was particularly important.

Barry G. Hall is supported by Research Career Development Award 1 KO4 AI00366 from the National Institute of Allergy and Infectious Diseases. Research in his laboratory was supported by grants from the National Institutes of Health and the National Science Foundation. Shozo Yokoyama is supported by grants GM28672 and MH31302 from the U.S. Public Health Service with Washington University. David H. Calhoun is the recipient of an Irma T. Hirschl Career Scientist
Award. Research in his laboratory is supported by grant GM23182 from the National Institutes of Health.

LITERATURE CITED


MANSON, M. D., and C. YANOFSKY. 1976. Naturally occurring sites within the Shigella


Masatoshi Nei, reviewing editor

Received April 18, 1983; revision received July 12, 1983.
Estimates of DNA and Protein Sequence Divergence:
An Examination of Some Assumptions

G. B. Golding
National Institute of Environmental Health Sciences

Some of the assumptions underlying estimates of DNA and protein sequence divergence are examined. A solution for the variance of these estimates that allows for different mutation rates and different population sizes in each species and for an arbitrary structure in the initial population is obtained. It is shown that these conditions do not strongly affect estimates of divergence. In general, they cause the variance of divergence to be smaller than a binomial variance. Thus, the binomial variance that is usually assumed for these estimates is safely conservative. It is shown that variability in the mutation rate among sites can have an effect as large as or larger than variability in the mutation rate among bases. Variability in the mutation rate among bases and among sites causes the number of substitutions between two sequences to be underestimated. Protein and DNA sequences from several species are collected to estimate the variability in mutation rates among sites. When many homologous sequences are known, standard methods to estimate this variability can be used. The estimates of this variability show that this factor is important when considering the spectrum of spontaneous mutations and is strongly reflected in the divergence of sequences. Smaller variability is found for the third position of codons than for the first and second codon positions. This may be because of less selective constraints on this position or because the third position has been saturated with mutations for the sequences examined.

Introduction

Estimates of sequence divergence derived from DNA and amino acid analysis have allowed the differences between species to be quantified. A method to estimate sequence divergence was first derived by Jukes and Cantor (1969) and by Kimura and Ohta (1972). With these methods, estimates of the rates of substitution can be obtained. One can also look for clues to the functional implications of the relative rates of divergence in different regions.

1. Key words: population genetics, evolution, sequence divergence, nonrandom mutation, gamma distribution, negative binomial distribution.

Address for correspondence and reprints: G. B. Golding, Genetics Laboratory, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709.

Mol. Biol. Evol. 1(1):125–142. © 1983 by The University of Chicago. All rights reserved. 0737-4038/83/0101-0001$02.00
These estimates depend on several assumptions. One of these assumptions is that the relative mutation rates between bases are constant. There is, however, evidence that this is not true and that, generally, transitions occur more frequently than transversions (Topal and Fresco 1976; Vogel and Kopun 1977; Fitch 1980; Gojobori et al. 1982b). Several authors (e.g., Kimura 1980, 1981; Takahata 1980; Gojobori et al. 1982a; Kaplan and Risko 1982; Kaplan 1983) have examined the effects of variable mutation rates between bases. They have shown that, in general, the estimates of the number of substitutions increase (relative to those obtained by Jukes and Cantor's [1969] method) when transitions occur more frequently. This is because multiple substitutions occur at certain sites, and a larger number of total substitutions are required to explain a given observed number of differences. This effect is larger for more divergent species and may be ignored if the species are closely related (Kaplan 1983).

It has been pointed out by several authors (e.g., Gillespie and Langley 1979; Templeton et al. 1981) that the effects of the initial conditions on estimates of divergence are usually ignored. These initial conditions include whether the original population is polymorphic and to what extent. Templeton et al. (1981) show that for several species sufficient polymorphism exists to strongly affect estimates of genetic divergence. For DNA and protein divergence it is also necessary to know whether there is a nonrandom association between different sites in these sequences. The population sizes of the species considered may also affect the variance of divergence. The initial conditions will be more important when the species are closely related.

Furthermore, it is well known that mutational "hot spots" exist in many genes. These create a large amount of variability in the mutation rate among sites. The possibility that such variation may affect estimates of divergence has been recognized since the initial study of restriction patterns by Upholt (1977; see also Nei and Li 1979). These authors state that if such variability occurs, their methods will underestimate the actual time of divergence. The size of this effect was not examined by these authors.

In this paper, I examine the effects of initial conditions and of site variability in the mutation rates on estimates of divergence. A solution to the variance of divergence is obtained that allows for arbitrary initial conditions and different mutation rates and population sizes in each species, and assumes sites to be completely linked. It is shown that the initial conditions have only small effects for DNA and protein sequences and that the usual estimate of the variance of divergence is "conservative," that is, the assumed variance is too large.

The size of the effects of site variability in the mutation rates is examined. It is shown that site variability can have as large an effect on estimates of divergence as do biases in the transition/transversion rates. The effect of mutation rate biases among the bases can be mimicked by the effects of variability in the mutation rates among sites. Evidence is presented to show that such variability in the mutation rate among sites exists and is important in nature.

Theory

The method of identity coefficients (Jacquard 1974, p. 104) is used to derive the expected divergence and its variance. This method involves the construction of recursion relationships for the probabilities that specific samples of gametes
are, or are not, identical. Here, identity in state rather than identity by descent is considered.

Two populations that have been isolated for $t$ generations will be considered. These two populations consist of $N_1$ and $N_2$ randomly mating, diploid individuals. It is assumed that $N_1$ and $N_2 \gg 1$ and are constant over time. The generations are distinct and nonoverlapping. These assumptions can be relaxed if the effective population sizes are used (Crow and Kimura 1970, sec. 7.6). No selection is considered, and the sites are assumed to be completely linked. For the moment each site in the DNA or protein sequence is assumed to have a constant mutation rate of $\mu_1$ in population 1 and $\mu_2$ in population 2. At each site there are $k$ alleles possible (for DNA sequences $k = 4$).

Let $x_i$ be the frequency of identical alleles at site $i$ between populations 1 and 2. The expected divergence (the expected proportion of differences between the two sequences), $D$, among a sample of $n$ sites is

$$E(D) = E\left[\frac{1}{n} \sum_{i=1}^{n} (1 - x_i)\right],$$

and since all sites have the same mutation rate,

$$E(D) = E\left[\frac{1}{n} n(1 - x_i)\right] = 1 - E(x_i).$$

The variance of divergence is

$$\text{var} (D) = \text{var}\left[\frac{1}{n} \sum_{i=1}^{n} (1 - x_i)\right] = \left(\frac{1}{n}\right)^2 \left[\sum_{i=1}^{n} \text{var} (x_i) + \sum_{i \neq j} \text{cov} (x_i,x_j)\right].$$

Since all sites have the same mutation rate, the expected values for each site are equivalent, and thus

$$\text{var} (D) = \left(\frac{1}{n}\right)^2 [n \text{var} (x_i) + n(n - 1) \text{cov} (x_i,x_j)]$$

$$= \frac{1}{n} [E(x_i^2) - E^2(x_i) + (n - 1)E(x_i x_j) - (n - 1)E^2(x_i)].$$

Thus only two sites need to be considered jointly. Similarly, Takahata (1982) and Golding and Strobeck (1982) have shown that only two sites need to be considered for similar measures within populations.

A solution for these quantities, in a manner that retains the initial conditions and the effects of population variability, can be obtained. To do this, define $\Phi_1$ as the probability that two gametes drawn at random, one from population 1 and one from population 2, are identical at site 1. Thus $\Phi_1 = E(x_i)$. Define $\Delta_1$ as the probability that members of a pair of gametes (one from population 1 and one from population 2) are identical at site 1 and that members of another pair of gametes (drawn at random without replacement) are identical at site 1. Thus $\Delta_1 = E(x_i^2)$. Define $\Delta_2$ as the probability that two gametes from populations 1 and 2
are identical at site 1 and two other gametes from populations 1 and 2 are identical at site 2. Thus $\Delta_2 = E(x_1x_2)$.

To derive a system of recursion equations for these variables, another seven coefficients have to be defined. These are again probabilities that two (denoted by $\Phi$), three (denoted by $\Gamma$), or four (denoted by $\Delta$) gametes, drawn at random without replacement, are identical. With one exception ($\Phi_1$), these probabilities will not enter into the following discussion but are necessary to derive the quantities of interest here. Define $\Phi_2$ and $\Phi_3$ as the probability that two gametes from population 1 and population 2, respectively, are identical. These are the expected homozygosity (or gene diversity) of the two populations. Define $\Phi_4$ as the probability that two gametes (one from each population) are identical at both site 1 and site 2. Define $\Gamma_1$, $\Gamma_2$, $\Gamma_3$, as the probability that three gametes, two chosen from population 1 (from population 2) and one chosen from population 2 (from population 1), are identical at site 1. Define $\Gamma_4$ as the probability that a gamete from population 1 (from population 2) has site 1 identical to that on another gamete chosen from population 2 (from population 1) and has site 2 identical to that on a third gamete chosen from population 2 (from population 1). As given in Appendix 1, each of these probabilities can be expressed in terms of gene frequencies (the relation between coefficients of identity and gene frequencies was noted by Weir and Cockerham [1974] and has been independently discovered by several authors, e.g., Serant [1974]).

Recursion relationships can be derived for these coefficients by considering all possible ways that the desired probability could have arisen. For example, for $\Phi_1$, two sites will be identical if they were different in the previous generation but mutated to the same allele or if they were identical and did not mutate. Therefore,

$$\Phi_{1'} = (1 - \mu_1)(1 - \mu_2)\Phi_{1'} + (\mu_1 + \mu_2)\frac{1}{k - 1}(1 - \Phi_{1'})$$

$$= (\mu_1 + \mu_2)\frac{1}{k - 1} + \left[1 - (\mu_1 + \mu_2)\frac{k}{k - 1}\right] \Phi_{1'},$$

and

$$\Phi_1 = \frac{1}{k} + \left[1 - (\mu_1 + \mu_2)\frac{k}{k - 1}\right] \left(\Phi_1^{1'} - \frac{1}{k}\right).$$

The recursion relationships for the other coefficients are given in Appendix 2. Some of these equations (those for $\Phi_1$, $\Phi_2$, $\Phi_3$, $\Gamma_1$, $\Gamma_2$, and $\Delta_1$) were found by Li and Nei (1975) using a continuous model and with $\mu_1 = \mu_2$.

The formula for $\Phi_1$ with $k = 4$, $\mu_1 = \mu_2 = \mu$, with $\Phi_{1'} = 1$, when substituted into equation (1), is recognized as that given by Jukes and Cantor (1969):

$$E(D) = \frac{3}{4}[1 - \exp \{-\frac{2}{3}(2\mu t)\}]$$

when $k \gg 1$,

$$\Phi_1 = \exp (-2\mu t)$$

is Nei's (1972) identity measure, $I$, with genetic distance defined as $D = -\ln(I)$. Although these parameters are given in terms of generations, the mutation rates and effective population sizes can be altered to measure time in arbitrary units.
In any case, $2\mu t$ is the expected number of nucleotide substitutions per base since the divergence of the two sequences.

The variance of divergence from equation (2) is

$$\text{var}(D) = \frac{1}{n}[\Delta'_1 - (\Phi'_1)^2] + \frac{n-1}{n}[\Delta'_2 - (\Phi'_2)^2].$$

The approximate variance formula given by Kimura and Ohta (1972), with this terminology, is

$$\text{var}(D) = \frac{1}{n}\Phi'_1(1-\Phi'_1)$$

(a binomial variance). If these two formulas are to agree, then $\Delta'_1 = \Phi'_1$ and $\Delta'_2 = (\Phi'_2)^2$. It will now be shown that these two equalities are not strictly correct and hold only approximately. The assumptions necessary for these approximations are demonstrated and the size of the effects examined.

The recursion relationships, given in Appendix 2, can be used to determine $\Delta'_1$ and $\Delta'_2$. Although these relationships can be solved easily, the answers are complicated. Therefore, only the solutions with $\mu_1 = \mu_2 = \mu$, $N_1 = N_2 = N$, $k = 4$, and $\Theta = 4N\mu k/(k-1)$ are presented here. First, to show that $\Delta'_1 \approx \Phi'_1$, a sufficient condition is that $t \gg 2N$, so that $\exp(-t/2N) \approx 0$. This assumption is required to reduce many effects of the initial conditions and indicates the length of time necessary. Also, if $\Theta \ll 1$ (one-third the population size times the mutation rate per nucleotide), then

$$\Delta'_1 \approx \frac{1}{4}(1 - 6\Theta) + (\Phi'_1 - \frac{1}{4})(1 - 2\Theta) \exp(-8\mu t/3) \approx \frac{1}{4} + (\Phi'_1 - \frac{1}{4}) \exp(-8\mu t/3) = \Phi'_1. \quad (3)$$

Second, the formula for $\Delta'_2$ with $t \gg 2N$ gives

$$\Delta'_2 \approx (\frac{1}{4})^2 + \frac{1}{2}(\Phi'_1 - \frac{1}{4}) \exp(-8\mu t/3) + [\Phi'_1 - \frac{1}{2}\Phi'_0 + (\frac{1}{4})^2] \exp(-16\mu t/3).$$

and if $\Phi'_0 = (\Phi'_0)^2$, then

$$(\Phi'_1 - \frac{1}{4})^2 \approx (\frac{1}{4})^2 + \frac{1}{2}(\Phi'_1 - \frac{1}{4}) \exp(-8\mu t/3) + (\Phi'_1 - \frac{1}{4})^2 \exp(-16\mu t/3) = (\Phi'_1)^2. \quad (4)$$

This illustrates that three assumptions are implicitly necessary for a binomial variance: (i) $t \gg 2N$, the time of divergence should be much larger than the population size; (ii) $\Theta \ll 1$, the mutation rate per nucleotide should be much smaller than the population size; and (iii) $\Phi'_0 \approx (\Phi'_0)^2$. It should be noted that the last assumption is, again, not strictly correct. In words, it states that the probability of joint identity at both site 1 and site 2 must be equal to the squared probability of identity at one site. That is, there can be no association of alleles between two completely linked sites (i.e., no linkage disequilibrium). For an initial population in equilibrium and with small $\Theta$, $\Phi'_0 - (\Phi'_0)^2 \approx \Theta^2(k - 1)^2$. Therefore, there is a nonzero, positive correlation between sites which creates a positive covariance in equation (2). This covariance tends to make the variance larger than a binomial variance. The error is of order $\Theta^2$. The error in the approximation of equation (3) is of order $\Theta$ and tends to make the variance smaller than a binomial variance.
To compare the difference between the true variance and a binomial variance, consider the following numerical example. Two populations with $\mu_1 = \mu_2 = \mu$, $N_1 = N_2 = N$, separated from a single, random-mating, initial population are examined. Let the initial population be at equilibrium with 90% homozygosity for a gene of 300 nucleotides. To have this homozygosity at equilibrium implies that $4N\mu$ must be approximately 0.00037 (Golding and Strobeck 1982; Takahata 1982; see also Appendix 2). Table 1 gives the initial values for the probabilities using this $4N\mu$ and the equilibrium equations given in Appendix 2. Table 1 also gives values for the remaining parameters necessary for the equations of Appendix 2. In particular, it is assumed that the two populations were derived from the initial population a million generations ago and $2\mu t = 0.02$. Substituting these values into the equations of Appendix 2 gives the probabilities necessary to determine the mean and variance of divergence. If $n = 100$ nucleotides are examined, these are $E(D) = 0.02010$ and $n \text{ var } (D) = 0.018609$. However, Jukes and Cantor’s (1969) method gives $E(D) = 0.01974$ and the binomial variance is $n \text{ var } (D) = 0.019346$.

These initial conditions are the expected values for a random choice of one sequence for each of the two populations from the original population. Since the speciation process may involve substantial sampling error (Templeton 1980), the initial values of $\Phi$, and so forth may be larger or smaller than these values. If the actual initial conditions are known, they can be included using these results.

This numerical example shows that the true variance is actually smaller than a binomial variance. This was found to be generally true after examining many numerical cases. The size of the difference between the two variances increases as $\Theta$ increases and as $n$ decreases. Many different cases were examined with the population sizes and mutation rates different in each population, and, again, the binomial variance was usually larger than the variance given in equation (2). Thus the assumption of a binomial variance yields a safe, conservative estimate.

In conclusion, initial conditions do not strongly affect the estimate of divergence. The variance of divergence is affected by both the mutation rate and size of each population and by the initial conditions. In general, these cause the variance to be smaller than a binomial variance. This is a useful result, since most studies do not determine the values of $\Theta$ and $2N$, and ignoring them leads to a conservative estimate of the variance.

**Table 1**

<table>
<thead>
<tr>
<th>A Numerical Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Phi_1 = \Phi_2 = \Phi_1^* = .99963$</td>
</tr>
<tr>
<td>$\Gamma_1^* = \Gamma_2 = .99926$</td>
</tr>
</tbody>
</table>

Let $4N\mu = .00037$ $2\mu t = .02$ $t = 1 \times 10^6$

$\mu = 10^{-8}$ $2N = 1.85 \times 10^4$ $\frac{1}{2Nt} = 54.0$

Then $\Phi_1 = .97990$ $\Delta_1 = .97881$ $\Delta_1^* = .96021$

*Note:*—The choice of initial conditions is explained in the text.
Variable Mutation Rates among Sites

Since the work of Benzer (1961), the concept of “hot spots” of mutation has been accepted readily. If such hot spots exist within the sequence being used to study the divergence of two species, the estimates of divergence may be affected strongly. It is my purpose here to determine how strong these effects might be. It is assumed that each site has a characteristic mutation rate which remains constant over time and that the mutation rates for many sites have some density distribution. This distribution should (i) be continuous and (ii) have a nonnegative argument. There are two well-characterized probability distributions with these properties that will be examined: the gamma and the lognormal distribution. The results of the previous section suggest that initial conditions will have only small effects, and hence they are ignored here.

Let each site have a mutation rate $k_i$. From equation (1) the expected divergence (the expected proportion of differences between the two sequences) is

$$E(D) = E \left[ \frac{1}{n} \sum_{i=1}^{n} (1 - x_i) \right]$$

$$= 1 - \frac{1}{n} \sum_{i=1}^{n} E(x_i)$$

$$= 1 - \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1}{4} + \frac{3}{4} \exp \left[ -\frac{4}{3} (2\mu_i t) \right] \right\}$$

$$= \frac{3}{4} \left\{ 1 - \frac{1}{n} \sum_{i=1}^{n} \exp \left[ -\frac{4}{3} (2\mu_i t) \right] \right\}.$$  

The right-hand term is given by the gamma moment generating function (Handbook of tables for probability and statistics 1966, p. 18) and yields

$$E(D) = \frac{3}{4} \left[ 1 - (1 + \frac{4}{3}\lambda c^2)^{-c^{-2}} \right],$$

where $\lambda = 2\bar{\mu}t$ (the mean) and $c = \sqrt{\text{var} (2\bar{\mu}t/\lambda^2)}$ (the coefficient of variation). In this case $\lambda$ is not the expected number of substitutions per base but, rather, the expected number averaged over all sites. The coefficient of variation, $c$, is a measure of the differences in $\mu$, from site to site. Assuming that

$$\text{var} (D) = \left( \frac{1}{n} \right)^2 \sum_{i=1}^{n} [E(x_i^2) - E^2(x_i)]$$

and that, at time $t = 0$, $E(x_i) = 1$, then

$$\text{var} (D) = \frac{3}{16} \left( \frac{1}{n} \right) [1 + 2(1 + \frac{4}{3}\lambda c^2)^{-c^{-2}} - 3(1 + \frac{8}{3}\lambda c^2)^{-c^{-2}}].$$

From the results of the previous section, this is probably a conservative estimate. If an estimate of $\text{var} (D)$ is obtained, these equations can be solved to find $\lambda$ and $c$. Table 2 gives the expected divergence for several values of $\lambda$ and $c$. The importance of this distribution has been amply shown by Fitch (1980) and Holmquist and Pearl (1980) for another reason (see Discussion).
A similar analysis can be done when the mutation rates are distributed lognormally. Since the moment generating function is not known, the mean was evaluated numerically. Table 2 gives the expected divergence for several values of c with a lognormal distribution for the mutation rates. For comparison, the expected divergence is shown in table 2 with mutation rates constant among sites but with the rate of transitions one to 10 times the rate of transversions. This has been taken from the formulas given in Kimura (1981).

Table 2 shows that both a transition/transversion bias and variability in the mutation rates among sites (whether gamma or lognormally distributed) cause the expected divergence to decrease. Variability in the mutation rates among sites can have an effect as large as or larger than a transition/transversion bias. In each case the size of the effect is smaller for closely related species. Variability in the mutation rates among sites begins to affect the expected divergence before a bias in mutation rates among bases. A comparison of the gamma and lognormal distributions shows that the former has a larger effect for the same c.

**Observed Distributions of Mutations**

**Spontaneous Mutants in Genes**

In 1961, Benzer collected a large number of spontaneous mutants in the rII region of T4. He found two major hot spots containing 809 of the 1,612 mutants collected. It has been estimated (Drake 1970, p. 53) that the rII region contains approximately 2,000 sites in total. The number of sites which have had k mutations are presented in table 3, column a. The class with zero mutations uses the estimate.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Expected Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>COEFFICIENT OF VARIATION</td>
<td>0</td>
</tr>
<tr>
<td>2µt</td>
<td></td>
</tr>
<tr>
<td>Gamma Distribution</td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.047</td>
</tr>
<tr>
<td>.10</td>
<td>.088</td>
</tr>
<tr>
<td>.15</td>
<td>.125</td>
</tr>
<tr>
<td>.20</td>
<td>.158</td>
</tr>
<tr>
<td>Lognormal Distribution</td>
<td></td>
</tr>
<tr>
<td>.05</td>
<td>.047</td>
</tr>
<tr>
<td>.10</td>
<td>.089</td>
</tr>
<tr>
<td>.15</td>
<td>.126</td>
</tr>
<tr>
<td>.20</td>
<td>.160</td>
</tr>
</tbody>
</table>

**Transition Rate 1–10 x Transversion Rate**

<table>
<thead>
<tr>
<th></th>
<th>1 x</th>
<th>5 x</th>
<th>10 x</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>.048</td>
<td>.048</td>
<td>.048</td>
</tr>
<tr>
<td>.10</td>
<td>.093</td>
<td>.094</td>
<td>.092</td>
</tr>
<tr>
<td>.15</td>
<td>.134</td>
<td>.136</td>
<td>.132</td>
</tr>
<tr>
<td>.20</td>
<td>.170</td>
<td>.172</td>
<td>.170</td>
</tr>
</tbody>
</table>
of 2,000 sites minus the 250 sites observed to have one or more mutants. The average number of mutants per site is 0.806 (0.402 excluding the two hot spots) with a standard deviation of 13.42 (2.06). If mutation rates are distributed according to a gamma distribution, the number of observed mutations will follow a negative binomial distribution. Fitting a negative binomial distribution to these data (Kendall and Stuart 1977, sec. 5.13) gives a coefficient of variation (of the mutation rate among sites) of 16.61 (or 4.89 excluding the two hot spots). A Poisson fit to these data shows that a significant excess occurs in the zero class and in sites with five or more mutations. If a Poisson distribution is used to fit the data and a number for the zero class is found which makes the best fit, there is still an excess of sites with many substitutions. Approximately 10% of the observed mutants are not base substitutions (the two hot spots are known to be due to frameshifts). Therefore, it is not clear whether such mutants contribute significantly to evolutionary change since they are strongly deleterious.

The spontaneous amber mutants that have been collected in the lacI gene of Escherichia coli by Coulondre et al. (1978) and Glickman (1982) are single base

Table 3
Distribution of the Number of Times a Site Was Changed in Several DNA and Protein Sequences

<table>
<thead>
<tr>
<th>Number of Changes (a)</th>
<th>lacI C (b)</th>
<th>lacI G (c)</th>
<th>Cytochrome (d)</th>
<th>Myoglobin (e)</th>
<th>Human mtDNA (f)</th>
<th>Human Globin β1 (g)</th>
<th>Human Globin β2 (h)</th>
<th>Human Globin β3 (i)</th>
<th>Primate mtDNA P41 (j)</th>
<th>Primate mtDNA P42 (k)</th>
<th>Primate mtDNA P43 (l)</th>
<th>Primate mtDNA trRNA (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 …….1,750</td>
<td>5</td>
<td>3</td>
<td>34</td>
<td>73</td>
<td>854</td>
<td>84</td>
<td>96</td>
<td>51</td>
<td>119</td>
<td>135</td>
<td>56</td>
<td>164</td>
</tr>
<tr>
<td>1 …….117</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>9</td>
<td>39</td>
<td>34</td>
<td>32</td>
<td>49</td>
<td>25</td>
<td>15</td>
<td>60</td>
<td>28</td>
</tr>
<tr>
<td>2 …….52</td>
<td>9</td>
<td>5</td>
<td>14</td>
<td>12</td>
<td>6</td>
<td>18</td>
<td>9</td>
<td>27</td>
<td>8</td>
<td>2</td>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>3 …….24</td>
<td>4</td>
<td>9</td>
<td>17</td>
<td>14</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4 …….13</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>10</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5 …….10</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6 …….6</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7 …….7</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8 …….2</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9 …….1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10 …….2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11 …….3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12 …….2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13 …….1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14 …….1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15 …….1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 …….2</td>
<td>5</td>
<td>1</td>
<td>21</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17 …….1</td>
<td>5</td>
<td>1</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20 …….1</td>
<td>1</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27 …….1</td>
<td>1</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32 …….1</td>
<td>1</td>
<td>71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>51 …….1</td>
<td>1</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>292 …….1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>517 …….1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16.61</td>
<td>1.60</td>
<td>1.50</td>
<td>1.04</td>
<td>1.14</td>
<td>1.78</td>
<td>.684</td>
<td>.735</td>
<td>.0</td>
<td>.689</td>
<td>.861</td>
<td>.0</td>
<td>.983</td>
</tr>
</tbody>
</table>
substitutions. Both of these studies show that the mutants do not follow a Poisson distribution (table 3, cols. b, c). A total of 36 sites can generate amber mutants by base changes. Coulondre et al.'s and Glickman's data show an average number of 6.14 (SD = 10.14) and 12.89 (SD = 19.65) mutants per site, respectively. These have coefficients of variation of 1.60 and 1.50, respectively. If the sites are separated on the basis of transition or transversion mutants to amber codons, the coefficients of variation in Glickman's data are 1.12 for transitions and 0.942 for transversions.

Changes Inferred from Amino Acid Sequences

The number of times that a codon has sustained a replacement can also be calculated from the amino acid sequences of a protein for several species. Uzzell and Corbin (1971) suggested a negative binomial distribution for the number of changes that have occurred in cytochrome c in 32 taxa (table 3, col. d). These authors showed that a negative binomial distribution provided a better fit to the data than did a Poisson distribution. In their study, they adjusted the number of sites with no replacements in order to maximize the fit of the negative binomial distribution. This was to allow for invariable residues in the protein. It has been shown by Fitch and Markowitz (1970) and by Coates and Stone (1981) that different sites may be invariable at different times during evolution. The negative binomial distribution has sufficient generality to allow for some sites to have very low replacement rates. Therefore, all of the amino acid positions that showed no replacements are included here in the zero class. This makes the fit of a negative binomial distribution worse; however, it also makes it impossible to fit a Poisson distribution. In this case the mean is 3.35 (SD = 3.93) with $c = 1.04$.

A similar analysis can also be done for the myoglobin amino acid sequence from 26 species given in Dayhoff (1978, pp. 235–239). The myoglobin gene does not have a large number of deletions/insertions and hence the analysis does not involve aligning the sequences. Using the phylogeny given in Dayhoff (1978), the minimum number of replacements that must have occurred are counted (Fitch and Margoliash 1967). This number may be much less than the actual number that occurred (Holmquist 1978; Holmquist and Pearl 1980), and, therefore, the coefficients of variation given here should be considered minimum estimates. Values of the inferred number of replacements at each site are given in table 3, column e, and show an average number of 1.58 (SD = 2.20) with $c = 1.14$. Again, all of the sites that have no replacements are included.

Inferred Changes in DNA Sequences

Aquadro and Greenberg (1983) have determined mtDNA sequences from seven humans. They observed 854 sites with no substitutions, 39 sites with one, and six sites with at least two substitutions (table 3, col. f). They conclude that there are too many sites with multiple substitutions to fit a Poisson distribution. The average number of mutants per site is 0.0567 (SD = 0.2587), and the coefficient of variation is 1.78. A negative binomial distribution can fit these data very well.

The recently published DNA sequences of β-globin allow an analysis at each position of the codon. Because of the redundancy of the genetic code, each of these positions is expected to have a different level of selective constraint. The sequences for human β (Lawn et al. 1980), human δ (Spritz et al. 1980), mouse β<sup>ma</sup>, β<sup>min</sup> (Konkel et al. 1979), two rabbit β genes (Hardison et al. 1979), goat β
and β\textsuperscript{c} (Schon et al. 1981), and chicken β genes (Richards et al. 1979) were used. The phylogeny is taken from Fitch and Langley (1976), with β\textsuperscript{mau} and β\textsuperscript{min} diverging after the split of rabbit and mouse lineages. The results for the first, second, and third codon positions of β-globin are shown in table 3, columns g, h, and i, respectively. They have means 0.617 (SD = 0.892), 0.440 (SD = 0.738), and 1.04 (SD = 0.996) with c = 0.684, 0.735, and 0.0, respectively.

The data of Brown et al. (1982) compare the sequences of two proteins and three tRNA genes in the mitochondria of five primates. The coefficients of variation for the first, second, and third codon positions of their protein 4 are 0.689, 0.861, and 0.0, respectively (table 3, cols. j, k, l). All codon positions of protein 5 had a c of 0.0, and the three tRNA genes had a c of 0.983 (table 3, col. m).

Evidence of Variability in Substitutions from Two Sequences

If an estimate of var (D) can be obtained, rough estimates of λ and c can be found from a comparison of just two sequences. Var (D) cannot be estimated in the usual way, since the comparison of two sequences leads to data that are encoded as 0 or 1. The sample variance for such data is a constant function of the mean, and thus an estimate of var (D) must be made from external data. Unfortunately, estimates depending on second-order moments are very sensitive to noise. In lieu of estimates of λ and c, it is possible to determine if any variability exists in different regions of two sequences. To do this, every possible subgroup of \( x \) adjacent nucleotides is examined and the number of differences between two sequences in this group of \( x \) nucleotides is counted. A \( \chi^2 \)-like statistic can be computed for each subgroup with an expected mean given by the overall number of differences between the two sequences and with a binomial variance. The sum for all possible subgroups is used as a test statistic. If all subgroups were independent, this would be the \( \chi^2 \) test for heterogeneity (Sokal and Rohlf 1969, p. 581) with degrees of freedom equal to the number of subgroups minus one. This statistic, as used here, is not distributed as a \( \chi^2 \) because different subgroups can contain some of the same nucleotides. Therefore, the distribution of this statistic was derived empirically from a computer simulation. The simulation takes the observed total number of differences and distributes them at random among all nucleotide positions; all possible subgroups with \( x \) adjacent nucleotides are examined, and the statistic is calculated. This is repeated 2,000 times. To illustrate the method, a comparison of the human ε-globin gene (Baralle et al. 1980) and the β-globin gene is presented. When I examined every five adjacent first-, second-, and third-position nucleotides, the statistics were 194.8, 100.4, and 134.3, respectively. Only 2.65% of the 2,000 random samples with a mean the same as that of the first position had a statistic as large as 194.8. This shows that there are groups of five adjacent first-position nucleotides that have too many and too few substitutions to assume that substitutions are distributed randomly along the sequence. The second and third positions did not show statistically significant “hot” or “cold” spots for substitution. The same analysis was done for groups of 10 adjacent nucleotides, and again the first position nucleotides had a statistic of 225.6, and only 1.35% of the 2,000 random samples deviated as much. Thus the hot and cold spots in the first position are detectable over at least 10 codons. In particular, a hot spot in the middle of the second exon is indicated. A group of 10 first-position nucleotides here had seven differences. The simulation showed that the maximum number of differences for any group of 10 is greater than or
equal to seven only 2.85% of the time. The data for the second and third positions did not show any variability in the distribution of substitutions along the sequence. None of the three positions showed any variability in the rate of substitution between the three exons. Again, these data suggest that substitution rates are variable among sites.

Discussion

Initial conditions do not strongly affect mean estimates of sequence divergence; they can, however, have a large effect on the variance of this estimate. The variance depends on the initial conditions, the population sizes and mutation rates within each species, and the time since divergence. In general, a binomial variance is assumed for the estimate of divergence, and it is shown that this variance is generally larger than the true value. This assumption, therefore, is justified since the binomial variance has a simple form and is a conservative estimate.

The assumption of equal mutation rates at each site is more important. If the mutation rate varies, there can be a large effect on the estimates of divergence. Both a gamma and a lognormal distribution of mutation rates cause the expected divergence to decrease. Due to Jensen's inequality (Feller 1971, p. 153), the expected divergence will decrease independent of the distribution of mutation rates. This effect can be larger than that caused by unequal mutation rates between bases. It also affects more closely related sequences. Since a transition/transversion bias can in effect create hot spots for mutation, it can have similar effects as variability in the mutation rate among sites. Hence, it is possible that site variability can mimic the effects of elevated mutation rates between sets of bases.

The estimated values of the coefficient of variation given in table 3, when compared with the coefficients of variation in table 2, indicate that there is sufficient variability in the mutation rate between sites to strongly affect estimates of divergence. It is expected that this variability will change from gene to gene since the factors promoting (inhibiting) mutation will depend on the structure of each gene and its surrounding environment. In some cases the fit of the negative binomial distribution is not very good. This does not mean that a Poisson distribution will fit the data, since this distribution is a limiting case of the negative binomial. Thus, even when \( c = 0 \), this does not necessarily imply that a Poisson distribution will fit the data.

Collections of spontaneous mutations clearly demonstrate that their distribution is not Poisson. There exist some sites where mutations are very likely to occur and other sites where mutations are unlikely to occur. They do not occur with equal probability at all sites. For example, Coulondre et al. (1978) suggest that the probability of mutation depends on neighboring bases. Both Coulondre et al. (1978) and Glickman (1982) group their results according to the type of base change (the majority are transitions) creating the mutation. Hot spots were observed for both transitions and transversions. Therefore, there is variability in mutation rates among sites, superimposed on a variability among bases.

Other authors have suggested a gamma distribution for the mutation rates (Uzzell and Corbin 1971; Holmquist and Pearl 1980; Holmquist et al. 1982), mainly to allow for invariable residues in the DNA. The results above suggest that the actual mutation rates are themselves better described by a gamma distribution. The effects of invariable residues are then compounded by the effects of nonrandom mutation rates among sites.
Uzzell and Corbin (1971) determined that the number of times a site underwent replacement can be fit to a negative binomial distribution, which is to be expected for Poisson events with gamma distributed means. When nonvariable residues are included in their analysis, the case against a Poisson distribution becomes even stronger. Fitch and Markowitz (1970) determined the number of nucleotide substitutions per codon for 29 species of cytochrome c. The variability in these data led them to use two Poisson distributions plus an invariable group to fit the data. Recently, Foster et al. (1982) have again used a mixture of two Poisson distributions to fit UV-induced mutations. With more extensive data for cytochrome c, Fitch (1976) showed that two Poisson distributions, or even three, along with an invariable group were not sufficient to fit the observed number of substitutions. Although this is probably a reasonable and biologically realistic approach, it does have disadvantages. There is no way to tell (before more study is done) how many and which of the sites belong to each class. Also, there is no reason to stop at a mixture of two distributions. Enough distributions with different means can be added until the fit is statistically significant. As an alternative, mixtures of Poisson distributions will “conform” to a negative binomial distribution if the two sequences are not too strongly divergent (Bliss and Fisher 1953). This is because the negative binomial distribution is skewed and allows for some sites to have very low mutation rates. If the means of two Poisson distributions are relatively close, a negative binomial distribution will give an approximate fit.

The data from DNA sequences again show variability between sites for the substitution rates. The trend toward a smaller variability in the third position of the codon may be an indication of fewer selective constraints on this position. If this were true, the first position should show less variability than the second position, as is the case. This could also be a reflection of the number of substitutions that have occurred. The small number of taxa examined here artificially minimizes the number of substitutions that can be detected. More data are required to determine if the substitutions in the third positions follow a Poisson distribution.

It is not suggested that mutation rates follow a gamma distribution; rather, it is suggested that mutation rates follow a complex distribution owing to a large number of causes. A gamma distribution has some features that enable it to mimic the effects of several of these causes. To include all possible assumptions about mutation rates would be very difficult, and the assumption of variable mutation rates among sites offers a simple alternative. A gamma or lognormal distribution is a first step toward a better description, and when more is known about the spectrum of mutations, better distributions could be used. The results above show that variability in mutation rates can have a large effect on the estimate of divergence, and it is important to include them.

Acknowledgments

I wish to thank C. Langley, N. Kaplan, and B. Margolin for their help and advice in the preparation of the manuscript. Financial support was provided by a Natural Sciences and Engineering Research Council of Canada Postdoctoral Fellowship.

APPENDIX 1

Relation between Frequency Moments and the Identity Coefficients

For population 1, let the frequency of gametes carrying the $l$th allele at site 1 and the $m$th allele at site 2 be $f_{lm}$. Let $p_l$ be the frequency of gametes carrying
the \( l \)th allele at site 1 and \( y_m \) the frequency of gametes carrying the \( m \)th allele at site 2. Let \( g_m, q_m, \) and \( z_m \) be the corresponding frequencies in population 2. Then, to an order of \((1/N)\), the identity coefficients can be expressed as

\[
\Phi_i = E\left( \sum_i p_i q_i \right), \quad \Phi_j = E\left( \sum_i p_i^j \right),
\]

\[
\Phi_3 = E\left( \sum_i q_i \right), \quad \Phi_4 = E\left( \sum_i \sum_m f_{im} g_m \right),
\]

\[
\Gamma_i = E\left( \sum_i p_i^2 q_i \right), \quad \Gamma_2 = E\left( \sum_i p_i q_i \right),
\]

\[
\Gamma_3 = E\left( \sum_i \sum_m f_{im} q_i z_m \right), \quad \Gamma_4 = E\left( \sum_i \sum_m g_{im} p_i y_m \right),
\]

\[
\Delta_1 = E\left( \sum_i \sum_n p_i q_i p_n q_n \right), \quad \Delta_2 = E\left( \sum_i \sum_n p_i q_i y_m z_m \right),
\]

where the sums are over all alleles and the expectations are over (conceptual) replicate populations.

APPENDIX 2

Recursion Relationships

For the model given in the text, the recursion relationships for the identity coefficients can be found using the same probability arguments as those used to derive \( \Phi_i \). The recursion relationships are

\[
\Phi_i^{*+1} = (\mu_1 + \mu_2) \frac{1}{k-1} + \left[ 1 - (\mu_1 + \mu_2) \frac{k}{k-1} \right] \Phi_i,
\]

\[
\Phi_2^{*+1} = \frac{1}{2N_i} + 2\mu_1 \frac{1}{k-1} + \left( 1 - \frac{1}{2N_i} - 2\mu_1 \frac{k}{k-1} \right) \Phi_2,
\]

\[
\Phi_3^{*+1} = \frac{1}{2N_2} + 2\mu_2 \frac{1}{k-1} + \left( 1 - \frac{1}{2N_2} - 2\mu_2 \frac{k}{k-1} \right) \Phi_3,
\]

\[
\Gamma_1^{*+1} = \left( \frac{1}{2N_i} + 2\mu_1 \frac{1}{k-1} \right) \Phi_1 + \mu_2 \frac{1}{k-1} \Phi_2 + \left( 1 - \frac{1}{2N_2} - 2\mu_1 \frac{k}{k-1} - \mu_2 \frac{k}{k-1} \right) \Gamma_1,
\]

\[
\Gamma_2^{*+1} = \left( \frac{1}{2N_2} + 2\mu_2 \frac{1}{k-1} \right) \Phi_1 + \mu_1 \frac{1}{k-1} \Phi_2 + \left( 1 - \frac{1}{2N_1} - 2\mu_2 \frac{k}{k-1} - \mu_1 \frac{k}{k-1} \right) \Gamma_2,
\]

\[
\Delta_1^{*+1} = \left( 2\mu_1 \frac{1}{k-1} + 2\mu_2 \frac{1}{k-1} \right) \Phi_1 + \frac{1}{2N_2} \Gamma_1 + \frac{1}{2N_1} \Gamma_2
\]

\[
+ \left( 1 - \frac{1}{2N_1} - \frac{1}{2N_2} - 2\mu_1 \frac{k}{k-1} - 2\mu_2 \frac{k}{k-1} \right) \Delta_1,
\]

\[
\Phi_4^{*+1} = \left( 2\mu_1 \frac{1}{k-1} + 2\mu_2 \frac{1}{k-1} \right) \Phi_1 + \left( 1 - 2\mu_1 \frac{k}{k-1} - 2\mu_2 \frac{k}{k-1} \right) \Phi_4,
\]
\[
\Gamma_3^{i+1} = \left( 2\mu_i \frac{1}{k-1} + 2\mu_i \frac{1}{k-1} \right) \Phi_1^{i} + \frac{1}{2N_2} \Phi_2^{i+1} + \left( 1 \frac{1}{2N_2} \frac{k}{k-1} + 2\mu_i \frac{k}{k-1} \right) \Gamma_3^{i},
\]
\[
\Gamma_4^{i+1} = \left( 2\mu_i \frac{1}{k-1} + 2\mu_i \frac{1}{k-1} \right) \Phi_1^{i} + \frac{1}{2N_1} \Phi_2^{i} + \left( 1 - \frac{1}{2N_1} - 2\mu_i \frac{k}{k-1} - 2\mu_i \frac{k}{k-1} \right) \Gamma_4^{i},
\]
\[
\Delta_1^{i+1} = \left( 2\mu_i \frac{1}{k-1} + 2\mu_i \frac{1}{k-1} \right) \Phi_1^{i} + \frac{1}{2N_1} \Gamma_1^{i} + \frac{1}{2N_2} \Gamma_4^{i} + \left( 1 - \frac{1}{2N_1} \frac{1}{k-1} - 2\mu_i \frac{k}{k-1} - 2\mu_i \frac{k}{k-1} \right) \Delta_1^{i}.
\]

When \( \mu_1 = \mu_2 = \mu, \ N_1 = N_2 = N, \ \Theta = 4N\mu_1(k-1) \) and \( k = 4 \), then
\[
\Delta_1^{i} = \frac{1}{4} \left( \frac{1 + 2\Theta + 4\Theta^2}{1 + 2\Theta} \right) + \left( \Phi_1^{i} - \frac{1}{4} \right) + \frac{1}{2} \left( \frac{1}{1 + 4\Theta} \right) \left( \Phi_2^{i} - \frac{1 + \Theta}{1 + 4\Theta} \right) \lambda_i^{i},
\]
\[
+ \left( \frac{2}{1 + 2\Theta} \right) \left( \Gamma_1^{i} - \frac{1}{4} \Phi_2^{i} - \Phi_1^{i} \frac{1 + \Theta}{1 + 2\Theta} + \frac{1}{2} \frac{1 + \Theta}{1 + 2\Theta} \right) \lambda_i^{i},
\]
\[
+ \left[ \Delta_1^{0} - \left( \frac{2}{1 + 2\Theta} \right) \left( \Gamma_1^{0} - \frac{1}{4} \Phi_2^{0} \right) - \frac{1}{2} \left( \frac{1}{1 + 4\Theta} \right) \Phi_2^{0}
\right.
\]
\[
+ \frac{1 - 2\Theta^2}{(1 + 2\Theta)^2} \left( \Phi_1^{0} - \frac{1}{4} \right) + \frac{1}{4} \frac{1 - 4\Theta^2}{(1 + 4\Theta)^2} \lambda_i^{i},
\]
\[
\Delta_2^{i} = \left( \frac{1}{4} \right)^2 + \frac{1}{2} \left( \Phi_1^{0} - \frac{1}{4} \right) \lambda_i^{i} + \left[ \Phi_1^{0} - \frac{1}{2} \Phi_1^{0} + \left( \frac{1}{4} \right)^2 \right] \lambda_i^{i} + 2 \left( \Gamma_1^{0} - \Phi_2^{0} \right) \lambda_i^{i}
\]
\[
+ \left( \Delta_1^{0} - 2\Gamma_1^{0} + \Phi_2^{0} \right) \lambda_i^{i},
\]

where
\[
\lambda_i^{i} = \left( 1 - \frac{1}{2N} \right) \frac{k}{k-1},
\]
\[
\lambda_i^{i} = \left( 1 - \frac{2\mu_i}{k-1} \right) ,
\]
\[
\lambda_i^{i} = \left( 1 - \frac{1}{2N} \right) \frac{k}{k-1},
\]
\[
\lambda_i^{i} = \left( 1 - \frac{2\mu_i}{k-1} \right) ,
\]
\[
\lambda_i^{i} = \left( 1 - \frac{4\mu_i}{k-1} \right) ,
\]
\[
\lambda_i^{i} = \left( 1 - \frac{1}{2N} \right) \frac{k}{k-1}.
\]
\[ \lambda_i = \left(1 - \frac{2}{2N} - 4\mu \frac{k}{k-1}\right). \]

For a single, randomly mating, initial population in equilibrium (with \( \mu_1 = \mu_2 = \mu, N_1 = N_2 = N, \) and \( \Theta = 4N\mu(k-1) \)), the initial values of the identity coefficients are

\[
\begin{align*}
\Phi_1^0 &= \Phi_2^0 = \Phi_3^0 = \frac{1 + \Theta}{1 + k\Theta}, \\
\Phi_4^0 &= \frac{1 + 2\Theta\Phi_1^0}{1 + k\Theta}, \\
\Gamma_1^0 &= \Gamma_2^0 = \frac{(2 + \Theta)\Phi_1^0}{2 + k\Theta}, \\
\Gamma_3^0 &= \Gamma_4^0 = \frac{2(1 + \Theta)\Phi_1^0 + \Phi_4^0}{3 + 2k\Theta}, \\
\Delta_1^0 &= \frac{(1 + \Theta)\Phi_1^0 + 2\Gamma_1^0}{3 + k\Theta}, \\
\Delta_2^0 &= \frac{(1 + \Theta)\Phi_1^0 + 2\Gamma_2^0}{3 + k\Theta},
\end{align*}
\]

(These values can be found in many sources and are not new results). Finally, the equilibrium probability that two randomly chosen sequences of length \( n \) sites differ at zero sites (the expected gene diversity) is

\[
\sum_{j=0}^{n} \binom{n}{j} \left(\frac{1}{k}\right)^{n-j} \left(1 - \frac{1}{k}\right) (1 + jk\Theta)^{-1}
\]

(Golding and Strobeck 1982; Takahata 1982). The example in the text with \( 4N\mu = 0.00037 \) was found by setting this probability at 0.9, \( n = 300 \), and solving for \( 4N\mu \).

**LITERATURE CITED**


WALTER M. FITCH, reviewing editor

Received May 24, 1983; revision received August 5, 1983.
Information for Contributors

*Molecular Biology and Evolution* is a bimonthly journal devoted to the interdisciplinary science between molecular biology and evolutionary biology. The journal emphasizes experimental papers, but theoretical papers are also published if they have a solid biological basis. Although this journal is primarily for original papers, review articles and book reviews normally written by solicited authors are also published. Brief discussion and comment on material published in this journal or on issues particularly relevant to readers of this journal will be published as "Letters to the Editor." Letters that refer to a paper handled by an Associate Editor should be sent to that editor or to the Editor in Chief.

To minimize publication delays, authors should follow the instructions given here and should also provide their telephone numbers.

Submission of Manuscripts

Send manuscripts (one original and two high-quality copies) to the Editor in Chief, Managing Editor, or any Associate Editor (addresses below). Any manuscript or any part of a manuscript which has been published or submitted for publication elsewhere cannot be accepted for publication. Anyone who wishes to write a review article should contact the Managing Editor. Correspondence about book reviews should also be addressed to the Managing Editor. Decision on acceptance of papers will be made as rapidly as possible. Papers that are not suitable to the journal will be returned immediately to authors without detailed review.

After a manuscript is accepted, its author will be requested to sign an agreement transferring copyright to the publisher. No published material may be reproduced or published elsewhere without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken to imply that the authors are prepared to make available to the public any unpublished sequences on which the paper is based and any clone of cells, DNA, or antibodies used in the experiments reported. This principle also applies to computer programs.

Preparation of Manuscripts

Papers must be written in English and organized in the sequence described below. Each section must be typed double-spaced on heavyweight, nonerasable bond; the page margins should be 1 1/2 inches wide to allow for corrections and manuscript editor’s notes and queries. Special typefaces (e.g., italic or sans serif) should not be used, and right-hand margins should not be justified. Word-processing output on dot matrix printers is acceptable only if it is the quality of the standard typewriter. Handwritten items (e.g., Greek letters) must be identified in the margin. Non-English words must have correct diacritics. Although each major part of the paper (e.g., Literature Cited) must begin on a new page, the pages should be numbered consecutively throughout, beginning with the title page and continuing through the abstract, text, Appendix, Literature Cited, footnotes, tables, and ending with figure legends.

Title page.—This page should contain the paper’s title, the names of all authors, the institution(s) at which the research was done, the current affiliations of all authors, the name and address for correspondence, and a footnote on nonstandard abbreviations used, if any (see below). Finally, the title page should also provide a running head (maximum of 50 characters and spaces).

Abstract.—The abstract should be a one-manuscript-page factual condensation of the entire paper, including a statement of purpose, a clear description of observations and findings, and a concise presentation of conclusions. It should not assert that the findings are discussed.
Information for Contributors

Key words.—A list of three to six words or phrases should be provided that will accurately index the subject matter of the article.

Text.—The text should comprise the following sections: (1) Introduction, (2) Material and Methods, (3) Results, (4) Discussion, and (5) Acknowledgments (if any). Papers should be concise but will not be restricted in length.

All organisms mentioned must be identified by their scientific binomens and underlined. Symbols for genetic loci must also be underlined and should follow the established rules of genetic nomenclature for the various organisms (consult M. Demerec et al., Genetics 54 [1966]: 61–76). Include the formal IUB name and number of all enzymes mentioned.

Do not use abbreviations for words or phrases used less than five times. Abbreviations used by the Journal of Biological Chemistry will be regarded as standard; nonstandard abbreviations should be defined collectively in a footnote.

Mathematical equations must be carefully typewritten: spacing between characters should be correct as typed. It will be assumed that all characters in equations and their counterparts in the text will be set in italics unless the author specifies otherwise the first time a character appears. Equations should be numbered sequentially, in arabic numerals in parentheses, on the right-hand side of the page.

These and other guidelines can be found in the Council of Biology Editors Style Manual (4th ed., 1978). In general, all material should conform to the CBE format. See also recent issues of this journal.

Terminology.—A satisfactory interdisciplinary communication requires using words with careful attention to their precise meaning in both disciplines. Authors may use any word they choose provided only that its meaning is clear, consistent, and serves to increase the paper’s comprehensibility. The following preferred usages are not prescriptive but will be assumed unless authors define them otherwise.

Where the alignments disagree, they are differences rather than changes since there may have been multiple changes to create a single difference. Differences or changes are replacements if the sequences are amino acids, substitutions if they are nucleotides. Mutations should be restricted to changes before selection has operated. Homology must be defined since it has two common meanings: (1) observed similarity and (2) inferred common ancestry. The term similarity is preferred for meaning 1 because sequences may have similarity acquired by convergence (analogy) rather than retained after divergence (homology). When homology arises via a gene duplication (all or part), it is properly called paralogy; when it arises via speciation, it is properly called orthology. Gaps are introduced into sequences to increase their similarity rather than to optimize similarity (homology), unless an algorithm is employed that guarantees an optimized result according to the way similarity (homology) is defined (e.g., as maximum matches—a third meaning of homology). Similarity should not be asserted to be significant unless patently obvious or accompanied by a probability statement and its method of determination ($\chi^2$, standard measure, binomial, etc.).

As recommended to the IUB, the preferred single letter code for nucleotide bases including ambiguity is: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, R = A/G (purine), Y = C/T (pyrimidine), M = A/C, W = A/T, S = C/G, K = G/T, B = C/G/T (not A), D = A/G/T (not C), H = A/C/T (not G), V = A/C/G (not T), N = X = A/C/G/T (any or unknown). For ambiguous nucleotides, T and U are equivalent.

Literature Cited.—Literature in the text should be cited by author and year and, where citation is to a book, the relevant pages thereof. Text citations of two or more works at a time should be given in chronological order. When a paper written by three or more authors is cited, write the name of the first author plus et al. The Literature Cited section at the end of the paper should be arranged alphabetically and then chronologically and should contain only works specifically cited in the text. References to papers that have not yet been published will be as for articles (see below), except that “accepted” (along with the journal name) will replace the volume and page numbers. “In press” will not be used.
Information for Contributors

(When such papers include authors of the submitted manuscript, copies of those papers must accompany the submitted manuscript.)

For the style of citations, please note the following examples:

Journal articles:

Books:

Book chapters:

The abbreviations of periodicals should be those used by the Council of Biology Editors. Periodical titles may also be written out.

Articles should include the name of the reviewing editor (the Editor or Associate Editor with whom author has corresponded) at the end of the Literature Cited section.

Footnotes.—Footnotes should be used sparingly. When necessary, they should be indicated in the text by superscript arabic numerals; the notes themselves should be typed on a page separate from the text. Footnotes to tables are referenced by superscript letters, except for significance levels, which use asterisks; table footnotes should be typed on the same page as the table to which they pertain.

Tables.—Each table must have a brief and self-explanatory title, be numbered with arabic numerals in order of its appearance in the text, and be typed on a separate page. Large, complex tables are discouraged. Guidelines for table format may be found in the CBE Style Manual and the Chicago Manual of Style or may be obtained by writing the editors of this journal.

Legends.—Figure legends should be typed on pages at the end of the manuscript, after tables. Each legend must be descriptive so that the illustration can be understood apart from the text and must define abbreviations used in the illustration.

Illustrations.—Each illustration (figure) should be an original, not a photocopy. Illustrations should be separate and have uniform lettering. They should be numbered consecutively, following the sequence in which they are mentioned in the text. The place where each illustration is to be inserted may be indicated by a circled note in the margin of the typescript. Names of authors, figure number, and an arrow indicating proper orientation should be written lightly in pencil on the back of each figure. Line drawings must be of high quality; typewritten or hand lettering is unacceptable. Photographs should be high-contrast, glossy prints. Magnifications may be indicated by a micron bar or in the legend. Photographs to be reproduced without further reduction must be so marked and may not exceed 4 1/4 inches wide by 7 5/8 inches long (122 by 194 mm) in order to fit the journal format. Please keep in mind that these dimensions are maxima. Because of the need for a figure legend, the illustration cannot be the maximum size in both dimensions.

Proofs and Reprints

Offprint order forms will be sent to the author (or in the case of multiple authors, to the senior author) with page proofs. There will be no page charge for publication.
Forthcoming

Comparison of Regulatory and Structural Regions of Genes of Tryptophan Metabolism
Charles Yanofsky

Concerted Evolution of the Immunoglobulin V$_{	ext{\gamma}}$ Gene Family
Masatoshi Nei and Takashi Gojobori

Major Morphological Effects of a Regulatory Gene: $Pgm1$ in Rainbow Trout
Robb F. Leary, Fred W. Allendorf, and Kathy L. Knudsen

Retention of Cryptic Genes in Microbial Populations
Wen-Hsiung Li

Selective Neutrality of Glucose-6-Phosphate Dehydrogenase Allozymes in *Escherichia coli*
Daniel E. Dykhuizen, Jean de Framond, and Daniel L. Hartl

Directed Evolution of Cellobiose Utilization in *Escherichia coli* K12
Maja Kricker and Barry G. Hall

Immunological Similarities between Specific Chloroplast Ribosomal Proteins from Chlamydomonas and Ribosomal Proteins from *Escherichia coli*
Robert J. Schmidt, Alan M. Myers, Nicholas W. Gillham, and John E. Boynton

Rabbit Globin Pseudogene $\psi\beta$2 Is a Product of Conversion between a $\delta$-Globin Gene and Rabbit Gene $\beta$1
Ross C. Hardison and Jean B. Margot

Comparison of the $\beta$-Like Globin Gene Families of Rabbits and Humans: The Gene Cluster 5'-$\xi$-$\gamma$-$\delta$-$\beta$-3' Predates the Mammalian Radiation and Each Gene Has Subsequently Diverged at a Different Rate
Ross C. Hardison

Estimation of Evolutionary Distance between Nucleotide Sequences
Masatoshi Nei and Fumio Tajima

Silencing of Duplicate Genes: A Null Allele Polymorphism for Lactate Dehydrogenase in Brown Trout (*Salmo trutta*)
Fred W. Allendorf, Gunnar Stahl, and Nils Ryman

The Relationship between Codon Boundaries and Multiple Reading Frame Preferences: Coding Organization of Bacterial Insertion Sequences
David J. Galas and Temple F. Smith
Molecular Biology and Evolution

Editor in Chief
Walter M. Fitch, University of Wisconsin

Managing Editor
Masatoshi Nei, University of Texas

Associate Editors
Roy J. Britten, California Institute of Technology
Wesley M. Brown, University of Michigan
Richard B. Flavell, Cambridge University
Ken W. Jones, University of Edinburgh
Richard K. Koehn, State University of New York at Stony Brook
Robert K. Selander, University of Rochester

Editorial Board
Fred W. Allendorf
University of Montana
Norman Amheim
SUNY at Stony Brook
John C. Avise
University of Georgia
Francisco J. Ayala
University of California, Davis
Anthony Hugh D. Brown
CSIRO, Canberra
L. L. Cavalli-Sforza
Stanford University
Elizabeth A. Craig
University of Wisconsin—Madison
Irving P. Crawford
University of Iowa
W. Ford Doolittle
Dalhousie University
Gabriel Dover
Cambridge University
Marshall H. Edgell
University of North Carolina
George Fox
University of Houston
Morris Goodman
Wayne State University
R. Grantham
Université de Lyon
Michael Grunstein
University of California, Los Angeles
Barry G. Hall
University of Connecticut
Daniel L. Hartl
Washington University
Alec Jeffreys
University of Leicester
Motoo Kimura
National Institute of Genetics, Japan
Costas Krimbas
Agricultural College of Athens, Greece
Charles H. Langley
National Institute of Environmental Health Sciences, North Carolina
Richard C. Lewontin
Harvard University
Wen-Hsiung Li
University of Texas at Houston
Alan Maxam
Harvard Medical School
Roger Milkman
University of Iowa
Takashi Miyata
Kyushu University, Japan
Tomoko Ohta
National Institute of Genetics, Japan
Dennis A. Powers
Johns Hopkins University
Nils Ryman
University of Stockholm, Sweden
Barbara Schaal
Washington University
Temple E. Smith
Northern Michigan University
Howard M. Temin
University of Wisconsin—Madison
Michael S. Waterman
University of Southern California
Sherman Weissman
Yale University Medical School
Gregory S. Whitt
University of Illinois at Urbana-Champaign
Eleutherios Zouros
Dalhousie University
Molecular Biology and Evolution

December 1983, Volume 1, Number 1

iii Introductory Statement

1 Species Adaptation in a Protein Molecule
   M. F. Perutz

29 Structure and Evolution of Human and African Ape rDNA Pseudogenes
   Elise Brownell, Mark Krystal, and Norman Arnheim

38 Mitochondrial DNA Differentiation during the Speciation Process in
   Peromyscus
   John C. Avise, John F. Shapira, Susan W. Daniel, Charles F. Aquadro, and
   Robert A. Lansman

57 Evolution of Antibiotic Resistance Genes: The DNA Sequence of a Kana-
   mycin Resistance Gene from Staphylococcus aureus
   Gary S. Gray and Walter M. Fitch

67 Geographic Components of Linkage Disequilibrium in Natural Populations
   of Escherichia coli
   Thomas S. Whittam, Howard Ochman, and Robert K. Selander

84 Rare Variant Alleles in the Light of the Neutral Theory
   Motoo Kimura

94 Rapid Evolution of Goat and Sheep Globin Genes Following Gene Duplication
   Wen-Hsiung Li and Takashi Gojobori

109 Role of Cryptic Genes in Microbial Evolution
   Barry G. Hall, Shozo Yokoyama, and David H. Calhoun

125 Estimates of DNA and Protein Sequence Divergence: An Examination of
   Some Assumptions
   G. B. Golding
Molecular Biology and Evolution (ISSN 0737-4038) is published six times a year in December, February, April, July, September, and November for the first volume and in January, March, May, July, September, and November for future volumes by the University of Chicago Press and sponsored by the Molecular Biology and Evolution Society and the American Society of Naturalists.

Subscription Rates U.S.A.: institutions, 1 year $60.00; individuals, 1 year $30.00. Student subscription rate, U.S.A.: 1 year $24.00 (copy of student ID must accompany subscription). Other countries add $4.00 for each year's subscription to cover postage. Single copy rates: institutions $10.00, individuals $5.00. Business correspondence should be addressed to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, Illinois 60637.

Change of Address Please notify the Press and your local postmaster immediately, giving both your old and new addresses. Allow four weeks for the change. Postmaster: Send address changes to Molecular Biology and Evolution, The University of Chicago Press, P.O. Box 37005, Chicago, Illinois 60637. Claims for missing numbers should be made within the month following the regular month of publication.

Editorial Correspondence Letters to the Editor should be addressed to the Editor in Chief of Molecular Biology and Evolution, 528A Service Memorial Institute, University of Wisconsin, 1300 University Avenue, Madison, Wisconsin 53706. For manuscript submission, see Information for Contributors at the end of this issue.

Copying beyond Fair Use The code on the first page of an article in this journal indicates the copyright owner's consent that copies of the article may be made beyond those permitted by Sections 107 or 108 of the U.S. Copyright Law provided that copies are made only for personal or internal use, or for the personal or internal use of specific clients, and provided that the copier pay the stated per copy fee through the Copyright Clearance Center, Inc., Operation Center, P.O. Box 765, Schenectady, New York 12301. To request permission for other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale, kindly write to the publisher.

Application to mail at second-class postage pending at Chicago, Illinois.

© 1983 by The University of Chicago. All rights reserved.
Introduction Statement

Welcome to the pages of Molecular Biology and Evolution. Since the rational first response to the news of any new scientific publication has to be "Why another journal?" let me begin by addressing that question. There is an urgent need for better communication between the molecular biologists and the evolutionists. This need can be filled only by a journal that (1) is devoted to this combination, (2) is of high quality, (3) has a rapid reviewing and publishing process, (4) is affordable and hence widely disseminated, (5) has a large international readership, and (6) is owned and controlled by the scientific community it serves. Since no existing journal meets all of these requirements, we decided to go ahead with MBE. We were particularly delighted that the University of Chicago Press shared our view that scholarly publishing meant putting the journal into the possession of scholars at a price each scholar could afford.

The scope of Molecular Biology and Evolution is broad, but the "and" in our title is informative in that it is not "or," that is, a paper must contain material about macromolecules and about their evolutionary implications. The journal will publish papers that critically examine the evolutionary significance of macromolecules—mechanisms of mutational change, processes of developmental control, mechanisms of natural selection, maintenance of genetic polymorphism, phylogenetic relationships of organisms, and theories that integrate these various aspects of molecular study.

Our first issue demonstrates MBE's breadth and interest. It begins with an Invited Paper by M. F. Perutz, "Species Adaptation in a Protein Molecule." Perutz's paper describes perhaps the largest collection of molecular changes ever correlated, in a historical perspective, with functional changes. It also contains a few puzzles.

In the second paper, "Structure and Evolution of Human and African Ape rDNA Pseudogenes," Elise Brownell, Mark Krystal, and Norman Arnheim detail the presence of an old ribosomal DNA (pseudogene) cloned from chimpanzees, gorillas, and humans that appears no longer to interact with the normal rDNA genes.

In "Mitochondrial DNA Differentiation during the Speciation Process in Peromyscus," John C. Avise, John F. Shapira, Susan W. Daniel, Charles F. Aquadro, and Robert A. Lansman show, using restriction maps of mitochondrial DNA, that the perfectly good species, P. polionotus from Florida, is more closely related to a group of P. maniculatus from the southern United States than many maniculatus deer mice are to each other. This shows that the nightmare of paraphyly enjoyed by some taxonomists simply cannot be avoided at the species level under the current definition of a species.

The fourth paper, "Evolution of Antibiotic Resistance Genes: The DNA Sequence of a Kanamycin Resistance Gene from Staphylococcus aureus," by Gary S. Gray and Walter M. Fitch, shows that this and three other aminocyclitol resistance genes are all distantly related with a phylogenetic relationship for the genes that is different from that of the taxa from which they derive, supporting the idea of their acquisition by interspecific transfer, possibly by plasmids.

According to Thomas S. Whittam, Howard Ochman, and Robert K. Selander in "Geographic Components of Linkage Disequilibrium in Natural Populations of
Escherichia coli," the amount of linkage disequilibrium is substantial and strongly associated with the reproductive system and population structure of this organism. It is suggested that periodic selection may play a prominent role in this.

"Rare Variant Alleles in the Light of the Neutral Theory," by Motoo Kimura, provides a novel way to estimate the fraction of newly arising alleles that are effectively neutral. The value is about one in seven for two primates, a fish, and the fruit fly.

In "Rapid Evolution of Goat and Sheep Globin Genes Following Gene Duplication," Wen-Hsiung Li and Takashi Gojobori show that the rates of amino acid-altering nucleotide substitution in globin genes are greatly accelerated after gene duplication, and that this acceleration is apparently due to relaxation of purifying selection rather than advantageous mutations.

"Role of Cryptic Genes in Microbial Evolution," by Barry G. Hall, Shozo Yokoyama, and David H. Calhoun, proposes a new concept of cryptic genes (genes normally silent but capable of reactivation) and examines how these genes may be maintained in a population.

The ninth and final paper, "Estimates of DNA and Protein Sequence Divergence: An Examination of Some Assumptions," by G. B. Golding, investigates the problem of models that permit variability in rates of nucleotide substitution, in terms of either particular base substitution or their sites. For comparative purposes, he shows that the statistically crucial parameter, the variance, is safely conservative using a model based on the negative binomial (of which the Poisson is a limiting case) and that, for real structural genes, the variance is least for the third of the three codon positions.

In addition to such research reports as these, future issues will contain book reviews, mini-reviews, and letters to the editor.

WALTER M. FITCH