Origins of metabolic diversity: Evolutionary divergence by sequence repetition

(enzymes/mutation/protein homology/protein sequences/recombination)

L. N. ORNSTON AND W. K. YEH

Department of Biology, Yale University, New Haven, Connecticut 06520

Communicated by Irwin C. Gunsalus, May 14, 1979

ABSTRACT Recurring patterns of primary structure have been observed in enzymes that mediate sequential metabolic reactions in bacteria. The enzymes, muconolactone Δ -isomerase [(+)-4-hydroxy-4-carboxymethylisocrotonolactone Δ^2 - Δ^3 -isomerase, EC 5.3.3.4] and β -ketoadipate enol-lactone hydrolase [4-carboxymethylbut-3-enolide(1,4)enol-lactone-hydrolase, EC 3.1.1.24], have been coselected in bacterial populations because the isomerase can confer no nutritional advantage in the absence of the hydrolase. Similar amino acid sequences recur within the structure of the isomerase, and the amino-terminal amino acid sequence of the isomerase from *Pseudomonas putida* appears to be evolutionarily homologous with the corresponding sequence of a β -ketoadipate enol-lactone hydrolase from *Acinetobacter calcoaceticus*. One interpretation of the sequence repetitions is that they reflect tandem duplication mutations that took place early in the evolution of the proteins. According to this view, the mutations caused elongation of structural genes and the creation of duplicated genes as the metabolic pathways evolved. A review of the sequence data calls attention to a different hypothesis: repeated amino acid sequences were introduced in the course of the proteins' evolution by substitution of copies of DNA sequences into structural genes. Our observations are interpreted on the basis of a model proposing genetic exchange between misaligned DNA sequences. The model predicts that misalignments in one chromosomal region can influence the nature of mutations in another region. Thus, as often has been observed, the mutability of a base pair will be determined by its location in a DNA sequence. Furthermore, the intrachromosomal recombination of DNA sequences may account for complex genetic modifications that occur as new pathways evolve. The model provides an interpretation of an apparent paradox, the rapid creation of new metabolic traits by bacterial genomes that are remarkably resistant to genetic drift.

Nowhere is biological diversity more evident than in the catabolic pathways of bacteria. The menu of potential growth substrates seems limitless, and the courses are subject to rapid change in an industrialized society. Often evolution has produced many different pathways for the dissimilation of a single growth substrate (1, 2), and additional genetic diversity is reflected in the distinctive induction mechanisms used by members of different bacterial genera (3). Analysis of rapidly evolving systems has caused bacterial genes to be regarded in increasingly fluid terms: illegitimate recombination may shift the moorings of genes (4, 5), and conjugation may cast them into genetically distant populations (6). Additional diversification of bacterial genomes might be expected to arise from genetic drift as the rapidly growing organisms wane and wax.

In view of the seemingly endless opportunities for genetic divergence, it is remarkable that bacterial taxonomy works: examination of a limited number of phenotypic properties allows the construction of classification schemes that may be confirmed at the level of nucleic acid hybridization (7). Resis-

tance of individual bacterial genes to genetic drift has been illustrated by Ambler and Wynn (8), who compared the primary structures of cytochromes c_{551} from different fluorescent *Pseudomonas* species. Against a background of wide structural variation in cytochromes c_{551} from the bacterial group as a whole, the primary structures of cytochromes c_{551} from independently isolated strains of *P. aeruginosa* were nearly invariant (9). Thus, there was no clear functional barrier to the divergence of the cytochrome c_{551} structural gene in distant ancestral lines of *P. aeruginosa*, but genetic divergence did not take place. Genetic regulation also appears to be conserved within well-circumscribed bacterial taxa. Transcriptional controls exerted over a set of homologous structural genes may vary widely in different bacterial genera, yet control patterns within a single taxon are the same (10).

The conservative properties of established bacterial genomes would seem to make them ill-suited bases for genetic improvisation, but novel bacterial genes do evolve. The underlying processes, mutation and selection, generally have been regarded in the context of growing cultures. The acquisition of novel nutritional traits has been studied in laboratory populations (11-15), and the results have fortified the view that new structural genes evolve in two stages: first, pre-existing genes are modified by random point mutations created by replication errors and, second, natural selection leads to the enrichment of mutant strains forming structural gene products with altered specificity. Genetic duplication often is regarded as a source of new gene copies, but little emphasis has been placed on the contribution sequence duplication may make to the evolutionary divergence of genes.

Another approach to microbial evolution is to investigate structural relationships among genes or gene products that evolved in natural populations. In this article we describe the beginnings of one such investigation. The results suggest that, in the natural environment, the introduction of sequence repetitions may make a major contribution to the establishment of new genes within a cell line. In addition, natural selection may favor divergence of segments of enzyme structure that are far removed from the active site. Our observations suggest a model that may resolve the paradox of rapid genetic diversification in populations in which genes tend to be conserved.

Homologous primary structures in enzymes catalyzing sequential metabolic reactions

Our studies have focused on two enzymes, muconolactone Δ -isomerase [(+)-4-hydroxy-4-carboxymethylisocrotonolactone Δ^2 - Δ^3 -isomerase, EC 5.3.3.4] and β -ketoadipate enol-lactone hydrolase [4-carboxymethylbut-3-enolide(1,4)enol-lactone-hydrolase, EC 3.1.1.24], that catalyze sequential catabolic steps (Fig. 1). The enzymes are widely distributed in the microbial world, where they mediate essential steps in the conversion of aromatic growth substrates to tricarboxylic acid cycle intermediates (3). The enzymes are inducible, and the regulation of their syntheses varies widely (3, 16). For example, repre-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

FIG. 1. Sequential metabolic reactions mediated by muconolactone isomerase and β -ketoadipate enol-lactone hydrolase. The isomerase mediates the intralactonic rearrangement of a double bond, and enol-lactone hydrolase catalyzes the hydrolysis of a lactone to a dicarboxylic acid.

sentatives of *P. putida* form a single enol-lactone hydrolase that is induced by its product, β -ketoadipate (17); in contrast, members of *Acinetobacter calcoaceticus* form isofunctional enzymes, enol-lactone hydrolase I and enol-lactone hydrolase II, that are induced by different metabolic precursors (18). The muconolactone isomerases and enol-lactone hydrolases from these bacteria are oligomers with small subunits, each containing about 100 amino acids in its primary structure (19, 20). Meagher (21) cleaved *Pseudomonas* muconolactone isomerase with cyanogen bromide and used the dansyl-Edman procedure to determine the amino acid sequence of the 19-residue carboxyl-terminal fragment. More recently, we (22) used automated procedures to determine the first 50 residues of the amino-terminal amino acid sequences of the isomerases from *Pseudomonas* and *Acinetobacter*. Thus, most of the amino acid sequence of the *Pseudomonas* isomerase has been established, and half of the amino acid sequence of the *Acinetobacter* isomerase is known. Automated procedures also were used to determine the 29-residue amino-terminal amino acid sequence of enol-lactone hydrolase II from *Acinetobacter* (22). As described below, comparison of the sequences reveals similarities suggesting that all three proteins share a complex evolutionary history.

Direct alignment of the amino-terminal amino acid sequences of the *Acinetobacter* and *Pseudomonas* isomerases leads to the matching of identical residues in 28 of 50 positions, indicating that the enzymes share a common ancestral gene (22). The amino-terminal amino acid sequences of *Pseudomonas* muconolactone isomerase and *Acinetobacter* enol-lactone hydrolase II also are similar: introduction of a single gap into the 29-residue *Acinetobacter* enol-lactone hydrolase II sequence aligns 10 residues with identical residues in the *Pseudomonas* isomerase sequence (22). This degree of similarity is not observed when the *Acinetobacter* enol-lactone hydrolase II sequence is aligned with the amino-terminal amino acid sequence of muconolactone isomerase from the same organism: introduction of a single gap allows alignment of only 5 identical residues in the 29-residue comparison (22). A plausible interpretation of these observations is that duplication of an ancestral sequence gave rise to the amino-terminal amino acid sequences of the isomerase and the hydrolase, and that

duplicated sequences diverged markedly as they were coselected in *Acinetobacter* (22).

If the observed homology between muconolactone isomerase and enol-lactone hydrolase extends throughout the entire proteins, the evidence will provide support for the hypothesis that a single ancestral protein can give rise to enzymes that catalyze different reactions in a metabolic pathway (23, 24). On the basis of present evidence, we must entertain the possibility that genetic rearrangements introduced the homologous regions of the isomerase and the hydrolase after the catalytic sites of the enzymes had been established. Precedent for such rearrangements is suggested by the appearance of amino acid sequences from dihydrofolate reductase in the carboxyl-terminal section of β -galactosidase from *Escherichia coli* (25). Opportunity for genetic exchange between the muconolactone isomerase and enol-lactone hydrolase structural genes has existed throughout their evolution because the isomerase cannot have been selected in an organism that lacked the hydrolase (16). Evidence summarized below indicates that repeated recombinational events between misaligned DNA sequences took place during the evolution of the isomerase. If, as physiological data (18) suggest, the structural genes for muconolactone isomerase and enol-lactone hydrolase II are linked within an operon, there would have been no physical obstacle to exchange between the genes.

Repeated amino acid sequences in muconolactone isomerases

The primary structure of muconolactone isomerase contains a number of internal sequence repetitions (Fig. 2), indicating that exchange between misaligned DNA sequences recurred during the evolution of the structural gene for the protein. No doubt varying phases of misalignment preceded genetic exchange, but a simple harmonic of displacement is suggested by the representation of misaligned muconolactone isomerase sequences shown in Fig. 3. In this comparison residue 1 of the *Pseudomonas* isomerase has been aligned with residue 5 of the *Acinetobacter* isomerase. Nevertheless, the misaligned sequences share 10 identical residues in 46 positions. The representation depicted in Fig. 3 suggests that a phase shift corresponding to four amino acid residues preceded genetic exchange between copies of an ancestral DNA sequence during evolution of the isomerase structural gene. Identical residues, occurring singly or as dipeptides, are fairly evenly distributed in the alignment shown in Fig. 3, suggesting that several independent genetic exchanges took place between the misaligned sequences during the evolution of the isomerase structural gene.

Genetic divergence by sequence repetition

The high frequency of sequence similarity in the misaligned isomerase sequences (Fig. 3) suggests a mechanism whereby sequence repetition could contribute to genetic divergence (Fig. 4). A distinctive feature of the model shown in Fig. 4 is that it proposes double, rather than single, crossover events between misaligned genetic sequences. Double crossover events usually are required for donor DNA to become part of a recipient replicon and, therefore, are fundamental genetic processes. Repeated sequences generated by double crossover events need not be tandem, and the duplicated regions represent substitutions, rather than additions, of genetic material. The mechanism proposed in Fig. 4 is radical in that it would allow oligonucleotides in DNA and peptides in proteins to be substituted by a single genetic event. Yet the mechanism is conservative because the deleted genetic material is quantitatively replaced with a duplicate of a nearby sequence. Hence, the overall phase of translation is not altered, and the length of the peptide chain

<u>Acinetobacter</u> Isomerase	1 <u>Met</u> Leu Phe <u>Gln</u> <u>Val</u> Arg <u>Met</u> <u>Asp</u> --- Val His
<u>Pseudomonas</u> Isomerase	1 <u>Met</u> Leu Phe His <u>Val</u> <u>Lys</u> <u>Met</u> --- Thr Val <u>Lys</u>
<u>Acinetobacter</u> Isomerase	11 <u>Leu</u> <u>Pro</u> <u>Val</u> <u>Ser</u> <u>Met</u> Pro Thr <u>Asp</u> <u>Gln</u> <u>Ala</u>
<u>Pseudomonas</u> Isomerase	11 <u>Leu</u> <u>Pro</u> <u>Val</u> <u>Asp</u> <u>Met</u> <u>Asp</u> Pro Ala <u>Lys</u> <u>Ala</u>
<u>Acinetobacter</u> Isomerase	21 Asn <u>Gln</u> Ile <u>Lys</u> --- Ser Val Glu <u>Lys</u> <u>Ala</u> Tyr Ser
<u>Pseudomonas</u> Isomerase	21 Thr <u>Gln</u> Leu <u>Lys</u> Ala <u>Asp</u> --- Glu <u>Lys</u> Glu <u>Leu</u> Ala
<u>Acinetobacter</u> Isomerase	32 Gln <u>Glu</u> <u>Leu</u> <u>Gln</u> Arg Gln <u>Gly</u> Lys --- Trp <u>Arg</u> <u>His</u> Ile Trp
<u>Pseudomonas</u> Isomerase	32 Gln Arg <u>Leu</u> <u>Gln</u> Arg Glu <u>Gly</u> --- Thr Trp <u>Arg</u> <u>His</u> <u>Leu</u> Trp
<u>Pseudomonas</u> Isomerase	77 <u>Met</u> <u>Glx</u> Ile <u>Glx</u> <u>Val</u> <u>Asx</u> <u>Gly</u> Leu --- Cys <u>Arg</u> <u>His</u>
<u>Pseudomonas</u> Isomerase	88 Pro Ser Ser Ile <u>Asx</u> Ser <u>Asx</u> <u>Arg</u> <u>His</u> Lys end

FIG. 2. Internal sequence repetitions in the primary structure of *Acinetobacter* isomerase and *Pseudomonas* isomerase. Numerals above amino acid residues indicate the position of the residues in the primary sequence of the proteins. Gaps have been introduced into the sequences at positions indicated by dashes. Underlined residues occur as fragments of repeated sequences within the primary structure of the proteins.

in the gene product is not changed. Because the double cross-over model demands the loss of genetic material, it is likely to represent, in the absence of recombination with other genomes, a point of irreversible divergence from the ancestral sequence. The repeated sequences are homologous with each other. In the terminology introduced by Fitch (27) they are orthologous

because they occupy different genetic loci (Fig. 4). In contrast, paralogous sequences (27) are homologous sequences encoded at the same genetic locus in separate genomes (Fig. 4).

Orthologous sequences within a genome may serve as the basis for homologous pairing that misaligns other genetic regions. Indeed, tandem duplications are known to be genetically

<u>Acinetobacter</u> Isomerase	1 Met Leu Phe Gln Val Arg Met Asp Val His Leu Pro Val Ser Met Pro Thr Asp Gln Ala Asn Gln Ile Lys Ser	25
<u>Pseudomonas</u> Isomerase	1 Met Leu Phe His Val Lys Met Thr Val Lys Leu Pro Val Asp Met Asp Pro Ala Lys Ala Thr	21
<u>Acinetobacter</u> Isomerase	26 Val Glu Lys Ala Tyr Ser Gln Glu Leu Gln Arg Gln Gly Lys Trp Arg His Ile Trp Arg Ile Thr Gly Gln Tyr	50
<u>Pseudomonas</u> Isomerase	22 Gln Leu Lys Ala Asp Glu Leu Glu Leu Ala Gln Arg Leu Gln Arg Glu Gly Thr Trp Arg His Leu Trp Arg Ile	46

FIG. 3. Similarities in the misaligned amino-terminal amino acid sequences of *Acinetobacter* isomerase and *Pseudomonas* isomerase. Direct alignment of the amino-terminal methionines results in the placement of 28 identical residues in matching positions in the 50-residue sequences (22). In the secondary alignment, depicted here, the amino-terminal methionine of *Pseudomonas* isomerase has been shifted so it is aligned with residue 5 in the *Acinetobacter* isomerase sequence. Despite the shift, 10 of 46 positions are occupied by identical residues (enclosed in boxes).

FIG. 4. Generation of nontandem duplicated sequences by crossover between misaligned genomes. The letters may be taken to represent either single nucleotides or sequence sets (e.g., codons). Unlike models for generation of tandem duplications by single crossover events (26), this model proposes a double crossover event that may give rise to nontandem duplicated sequences. The duplicated sequence (DE) substitutes for a deleted sequence (HI) of equal length. Hence, intragenic duplication would not alter the phase of translation of the gene or change the length of the gene product. The duplicate sequence (DE, underlined in the recombinant sequence) is homologous with the copied segment of the original sequence (DE, not underlined in the recombinant sequence) but occupies a different genetic locus. As defined by Fitch (27), the duplicate region (DE) and the region that was copied (DE) are orthologous because they occupy different genetic loci. Replicates of the sequence DE would be paralogous with DE, at the same genetic locus, but not with DE at a different genetic locus. Similarly, replicates of DE would be paralogous with DE and orthologous with DE.

unstable because pairing of misaligned, repeated sequences, followed by a crossover event, leads either to their loss or to their future repetition (26). Similarly, pairing of nontandem duplicated regions can favor misalignment, and exchange between misaligned, nonhomologous sequences can create additional sequence duplications (Fig. 5). Multiple recombinational events after pairing of orthologous and paralogous segments could rearrange sequences as a deck of cards is rearranged by shuffling (Fig. 6). Thus, marked potential for genetic variation springs from the fact that interactions in one genetic region may influence the structure of another genetic region. Yet, in different contexts, similar interaction could serve as the basis for repair mechanisms that would restrict the mutational options of highly evolved genomes (Fig. 7).

Selection of duplicated sequences within structural genes

As described above, the structural genes for the isomerase and the hydrolase appear to have diverged more rapidly when they were coselected in a single cell line than when they evolved in separate organisms. Two selective forces favor the rapid divergence of gene copies as they become established within a genome. First, genetic regions carrying lengthy duplications are excised readily by legitimate recombination mechanisms (26). Substitution of short duplicated segments within longer

FIG. 5. Base pairing of orthologous sequences allows recombination to generate additional orthologous sequences within a genome. In this example, pairing of the orthologous sequences DE favors misalignment of the rest of the genome; recombination substitutes KL for OP and, thus, generates a new set of duplicates within the genome. As in Fig. 4, the duplicated sequences need not be tandem to each other.

FIG. 6. Base pairing of paralogous and orthologous sequences opens many opportunities for genetic diversification by recombination. In the example depicted here, pairing of the orthologous sequences DE (underlined in the recipient sequence) and the paralogous sequences MNOPQRST (connected by vertical lines) causes a looping out of the sequence DEJKL; replacement of the residue L with a residue from the looped out sequence would lead to substitution of D, E, J, or K for L. Excision of the looped out region would cause a deletion, and substitution of L with more than one residue from the looped out region could give rise to a recombinant carrying an insertion. It should be noted that spontaneous mutations giving rise to dysfunctional gene products tend to be small deletions or insertions (28).

regions of sequence identity would cause rapid divergence of homologous genes. Second, the accurate self association of homologous protein monomers into separate oligomers may require a number of amino acid substitutions at protein-protein interfaces (29). Under this selective pressure, substitution of duplicated sequences should provide more immediate variation than would be allowed by multiple point mutations.

The foregoing arguments propose that nontandem duplicated sequences may be introduced into structural genes as they diverge; mutations creating the duplications would be selected strongly as duplicated genes diverged to assume different functions within a cell line. Repeated amino acid sequences, observed in many proteins (30), generally are interpreted as the residuum left from tandem duplication mutations that lengthened the original structural genes of the proteins. Diversity and specificity in protein structure tend to be regarded as the product of point mutation, the consequence of replication errors, followed by selective pressure for a new catalytic function. In contrast to this view, we suggest that repeated sequences may be introduced by substitution at various points in protein's evolutionary history; we propose that some internally repeated sequences are stamps of self identity acquired as structural genes coevolved in a genome with their evolutionary homologs.

Genetic origin of biological diversity

The mechanisms by which duplicated genes diverge within a single ancestral line are poorly understood. Most existing evidence concerning molecular evolution has been obtained by comparison of the sequences of isofunctional proteins in divergent organisms. There is no *a priori* reason to believe [and there is reason to doubt (31)] that conclusions derived from these studies will have equal predictive value in the analysis of orthologous genes, duplicated genes that were coselected as part of the same genome. For example, point mutations, the presumed products of replication errors, often are regarded as the major sources of genetic diversity. Yet other kinds of mutation may make major contributions as duplicated genes become established within a genome. Dayhoff and Barker (32) suggested that short deletions and insertions are likely to become fixed at this point in a gene's evolutionary history, and we propose that genetic sequences duplicated from nearby regions may be substituted to provide additional sources of genetic

FIG. 7. Recombination between orthologous sequences as a mechanism of repair. The original sequence carries two orthologous sets, DE and KL. A mutation substitutes x for one K of the original sequence. Alignment of the orthologous sequences followed by recombination allows the mutant residue x to be substituted by the original residue K.

diversity. In principle, repeated sequences allow one genetic region to influence the structure of another, and we suggest that under some circumstances they may serve as the basis of repair mechanisms that conserve the structure of established genes.

It appears that the catabolic enzymes of bacteria will prove to resemble the immunoglobulins in many respects: repeated sequences, conserved regions, and a capacity for rapid genetic diversification. A comparative study of amino acid sequences led to the hypothesis that varied combinations of small DNA segments give rise to the variable region immunoglobulin genes (33). Repetition of sequences in cloned immunoglobulin genes fostered the proposal that a mismatch repair system might substitute nucleotides from one orthologous DNA segment into another within the same genome (34). It is reasonable to speculate that similar mechanisms may underlie genetic diversification in vertebrates and bacteria.

Our sequence data were gathered in collaboration with P. Fletcher of the Yale Cell Biology Unit. The work was supported by National Science Foundation Grants PCM 76-11008 and PCM 77-24884 and by National Institutes of Health Grants GM 21714 and GM 25487.

- Dagley, S. (1975) in *Essays in Biochemistry*, eds. Campbell, P. N. & Aldridge, W. N. (Academic, New York), Vol. 11, pp. 81-138.
- Dagley, S. (1978) in *The Bacteria*, eds. Ornston, L. N. & Sokatch, J. R. (Academic, New York), Vol. 6, pp. 305-388.
- Stanier, R. Y. & Ornston, L. N. (1973) in *Advances in Microbial Physiology*, eds. Rose, A. H. & Tempest, D. W. (Academic, London), Vol. 9, pp. 89-151.
- Cohen, S. N. (1976) *Nature (London)* **263**, 731-738.
- Kleckner, N. (1977) *Cell* **11**, 11-23.
- Bennett, P. M. & Richmond, M. H. (1978) in *The Bacteria*, eds. Ornston, L. N. & Sokatch, J. R. (Academic, New York), Vol. 6, pp. 1-69.
- Jones, D. & Sneath, P. H. A. (1970) *Bacteriol. Rev.* **34**, 40-81.
- Ambler, R. P. & Wynn, M. (1973) *Biochem. J.* **131**, 485-498.
- Ambler, R. P. (1974) *Biochem. J.* **137**, 1-15.
- Crawford I. P. (1975) *Bacteriol. Rev.* **39**, 87-120.
- Clarke, P. H. (1974) *Symp. Soc. Gen. Microbiol.* **19**, 113-159.
- Clarke, P. H. (1978) in *The Bacteria*, eds. Ornston, L. N. & Sokatch, J. R. (Academic, New York), Vol. 6, pp. 137-218.
- Hegeman, G. D. & Roseberg, S. L. (1970) *Annu. Rev. Microbiol.* **24**, 429-462.
- Jenson, R. A. (1976) *Annu. Rev. Microbiol.* **30**, 409-426.
- Lin, E. C. C., Hacking, A. J. & Aguilar, J. (1976) *BioScience* **26**, 548-555.
- Ornston, L. N. & Parke, D. (1977) in *Current Topics in Cellular Regulation*, eds. Horecker, B. L. & Stadtman, E. R. (Academic, New York), Vol. 12, pp. 209-262.
- Ornston, L. N. (1966) *J. Biol. Chem.* **241**, 3800-3810.
- Canovas, J. L. & Stanier, R. Y. (1967) *Eur. J. Biochem.* **1**, 289-300.
- Patel, R. N., Mazumdar, S. & Ornston, L. N. (1975) *J. Biol. Chem.* **250**, 6567-6577.
- Patel, R. N., Meagher, R. B. & Ornston, L. N. (1974) *J. Biol. Chem.* **249**, 7410-7419.
- Meagher, R. B. (1977) *Biochim. Biophys. Acta* **494**, 33-47.
- Yeh, W. K., Davis, G., Fletcher, P. & Ornston, L. N. (1978) *J. Biol. Chem.* **253**, 4920-4923.
- Horowitz, N. H. (1945) *Proc. Natl. Acad. Sci. USA* **31**, 153-157.
- Horowitz, N. H. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, New York), pp. 15-23.
- Hood, J. M., Fowler, A. V. & Zabin, J. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 113-116.
- Anderson, R. P. & Roth, J. R. (1977) *Annu. Rev. Microbiol.* **31**, 473-505.
- Fitch, W. M. (1973) *Annu. Rev. Genet.* **7**, 343-380.
- Farabaugh, P. J., Schmeissner, U., Hofer, M. & Miller, J. H. (1978) *J. Mol. Biol.* **126**, 847-863.
- Ornston, L. N. & Parke, D. (1976) *Biochem. Soc. Trans.* **4**, 468-473.
- Barker, W. C., Ketcham, L. K. & Dayhoff, M. O. (1978) *J. Mol. Evol.* **10**, 265-281.
- Arnheim, N., Hindenberg, A., Begg, G. S. & Morgan, F. J. (1973) *J. Biol. Chem.* **248**, 8036-8042.
- Dayhoff, M. O. & Barker, W. C. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Silver Spring, MD), Vol. 5, pp. 41-46.
- Kabat, E. A., Wu, T. T. & Bilofsky, H. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2429-2433.
- Seidman, J. G., Leder, A., Nau, M., Norman, B. & Leder, P. (1978) *Science* **202**, 11-17.