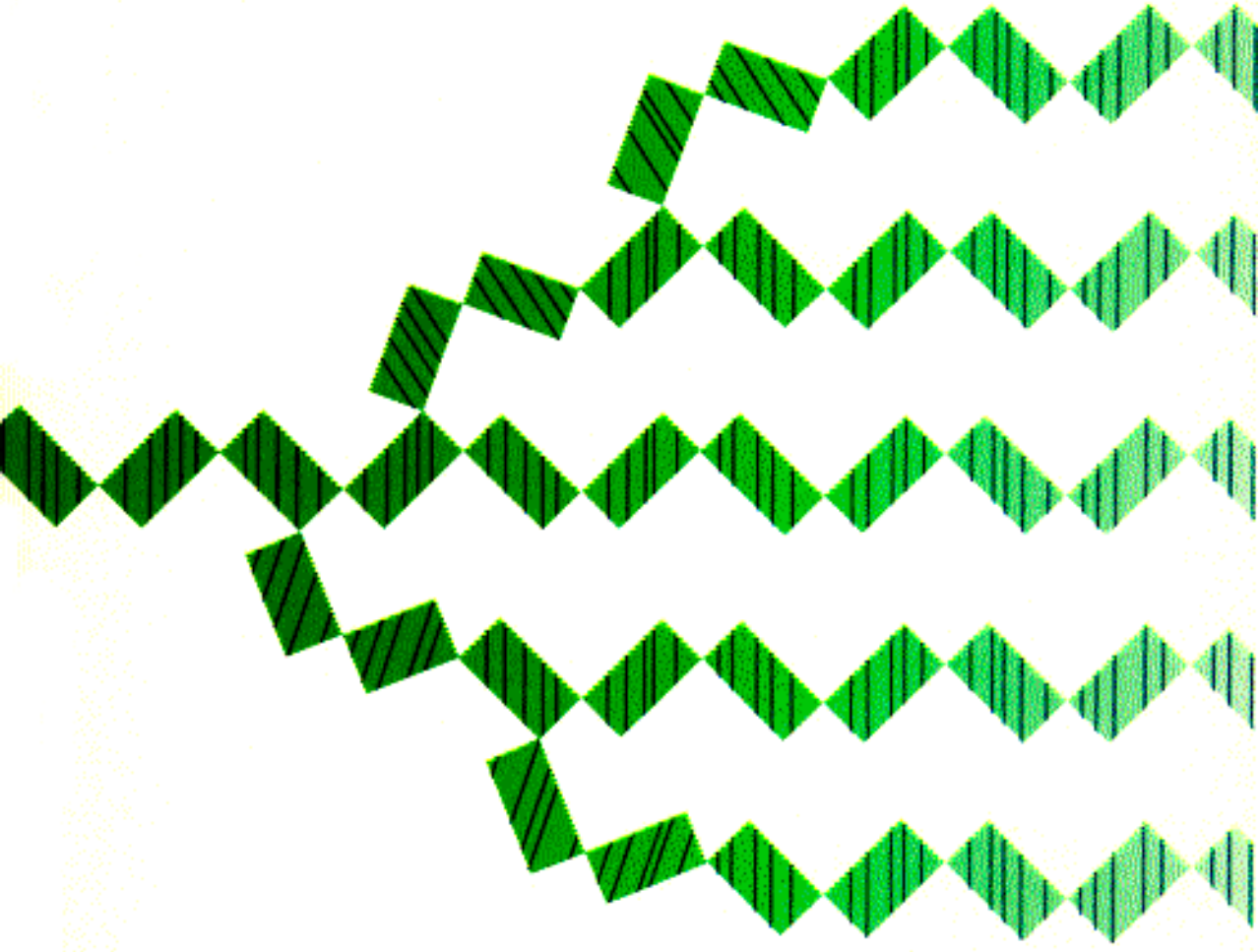
Molecular Biology and Evolution

Volume 1, Number 2, February 1984



The University of Chicago Press

Comparison of Regulatory and Structural Regions of Genes of Tryptophan Metabolism[1]

Charles Yanofsky
Stanford University

The genes of tryptophan biosynthesis are arranged and regulated differently in many microorganisms. Comparison of the transcription regulatory regions of the *trp* operons of several species of enterobacteria reveals that those sequences and structures believed to be essential for repression and attenuation control are conserved. Examples of divergent and convergent evolutionary change are presented. Rearrangements involving the homologous *trpG* and *pabA* genes and their presumed ancestral bi-specific gene are described. Alignment of homologous sequences of *trp* polypeptides encoded by fused and nonfused genes from various species reveals short connecting amino acid sequences at fusion junctions. These connecting sequences may be relics of gene fusion events and/or they may facilitate the proper folding of neighboring polypeptide domains.

Introduction

All organisms capable of synthesizing the amino acid tryptophan use seven identical enzymatic functions in the same biochemical pathway. In a comparison of the *trp* genes of these organisms, the most striking features are the many gene arrangements that exist, the different gene fusions that have occurred, and the variety of regulatory mechanisms that are employed in the control of gene expression. Why are there so many solutions to this single metabolic need?

The genes of tryptophan metabolism have been the subject of numerous studies carried out with a wide range of microbial species (Crawford 1975). As a consequence, appreciable structural information and regulatory information may be scrutinized. Much of this knowledge has been provided by the studies of Irving Crawford, Howard Zalkin, Brian Nichols, Terry Platt, their co-workers, and members of my group, although others have been excited by the wealth of existing natural *trp* gene variation and have contributed to our understanding of this diversity. For the most part, evolutionary investigations with *trp* genes have been

1. Key words: *trp* regulatory regions; *trp* genes; structural gene rearrangements. This paper is based on a talk presented at a joint meeting of the Genetics Society of America, the Society for the Study of Evolution, and the American Society of Naturalists in June 1983.

Address for correspondence and reprints: Department of Biological Sciences, Stanford University, Stanford, California 94305.

Mol. Biol. Evol. 1(2):143–161. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0001$02.00

comparative or directed toward elucidating structure-function relationships. Only now do we have the tools to manipulate the genome in an effort to explain the selective forces behind specific evolutionary events.

In this article I describe the evolutionary implications of several observations. I begin by comparing the structural features of the regulatory regions controlling expression of the *trp* genes of several enterobacterial species. The *trp* genes of these organisms are organized identically—in a single operon. In most of these species, expression of the *trp* operon is controlled by two transcription regulatory mechanisms, repression and attenuation (Rose et al. 1973; Blumenberg and Yanofsky 1982*a*, 1982*b*; Kelley and Yanofsky 1982; Kolter and Yanofsky 1983). Comparisons of the operator region sequences that are targets for repressor recognition identify conserved, essential contact sites as well as structural differences that explain observed functional variation. Comparisons of *trp* leader regulatory regions—responsible for attenuation control—reveal conservation of the functionally important structures crucial to attenuation. These structures are used to sense the extent of charging of $tRNA^{Trp}$ and to regulate transcription termination at the attenuator. Comparison of the primary structures of *trp* proteins also has provided information of evolutionary interest. Several instances of *trp* gene fusions have been analyzed at the molecular level (Miozzari and Yanofsky 1979; Nichols et al. 1980; Zalkin and Yanofsky 1981; Schechtman and Yanofsky 1983). The structural differences observed between homologous fused and separate genes invite speculation on both the genetic events that led to these fusions and the explanations for their occurrence. Finally, I describe studies performed to assess the functional significance of the amino acid sequence variation seen in homologous *trpA* proteins.

Comparison of the *trp* Operon Operators of Various Enterobacteria and the *aroH* and *trpR* Operators of *E. coli*

The nucelotide sequences of the operator regions of the *trp* operons of several enterobacterial species are shown in figure 1 along with the corresponding regions of the *aroH* and *trpR* operators of *E. coli* (Brown 1968; Miozzari and Yanofsky 1978; Gunsalus and Yanofsky 1980; Singleton et al. 1980; Blumenberg and Yanofsky 1982*a*). Mutational studies with the *trp* operator of *E. coli* have shown

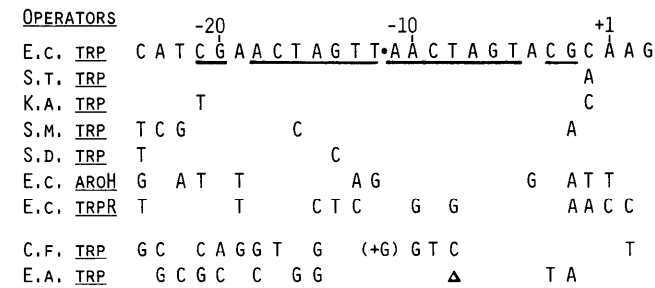


FIG. 1.—Homology of *trp*, *aroH*, and *trpR* operators. With the exception of the sequence of the *E. coli trp* operator, one strand of which is shown, only sequence differences are indicated. E.c. = *Escherichia coli;* S.t. = *Salmonella typhimurium;* K.a. = *Klebsiella aerogenes;* S.m. = *Serratia marcescens;* S.d. = *Shigella dysenteriae;* C.f. = *Citrobacter freundii;* E.a. = *Erwinia amylovora.* The extra G of the C.f. sequence is between the T and A at positions −12 and −11 of *E. coli.* The region of dyad symmetry is underlined.

that nucleotides in the central 14 base pair symmetrical segment of the operator are most essential for operator function; i.e., operator constitutive mutations that alter *trp* repressor binding occur only within this segment (Bennett and Yanofsky 1978). In addition there is a single base pair change in this segment in the *Shigella dysenteriae* promoter (fig. 1) that renders this operator partially constitutive; associated with this partial loss of operator function is a 10-fold reduction in promoter efficiency (Miozzari and Yanofsky 1978). This single mutational change in the promoter/operator region therefore assures modest expression of the operon in media containing excess tryptophan but limits high-level expression when tryptophan is lacking. There also are at least two missense mutations in *trpE* of our *Shigella* strain that drastically reduce the catalytic activity of the *trpE* protein (Manson and Yanofsky 1976). Apparently both regulatory and structural alterations have reduced this strain's capacity to synthesize tryptophan and regulate its formation. This strain has an absolute requirement for cysteine and is partially defective in the synthesis of several amino acids in addition to tryptophan (Miozzari and Yanofsky 1978). We conclude from these observations that our strain of *Shigella* is in the process of losing the capacity to synthesize several amino acids. We presume that this loss reflects an adjustment to the natural nutritional environment of this strain of *Shigella*.

The central core sequence of the *trp* operator is entirely conserved in *Salmonella typhimurium* and *Klebsiella aerogenes*, and there is a one base pair difference in the operator of *Serratia marcescens*. This difference does not appear to influence repressor binding activity. The *Citrobacter freundii* (Blumenberg and Yanofsky 1982*a*) and *Erwinia amylovora* (C. Yanofsky, unpublished) *trp* operators appear to have diverged considerably from those of the other enterics. The *Citrobacter* promoter does not bind *E. coli trp* repressor, whereas the *Erwinia* promoter does have binding activity. Further work is required to establish whether these highly divergent sequences are cloning artifacts.

The *E. coli aroH* and *trpR* operator sequences differ in the central symmetrical segment when compared with *trp* operon operators (fig. 1). Some or all of the differences probably account for the reduced affinity of these two operators for the *trp* repressor. In vivo, the *aroH* operator is regulated over about a 10–20-fold range (Brown 1968; Kelley and Yanofsky, unpublished), the *trpR* operator is regulated over only about a fivefold range (Kelley and Yanofsky 1982), whereas the *trp* operon is regulated over a 70-fold range (Jackson and Yanofsky 1973). In addition, in vivo, the amount of *trp* repressor appears to be in excess with respect to *trp* operon regulation, whereas it is limiting for *trpR* regulation (Kelley and Yanofsky 1982). Apparently, the affinities of the three operators for the *trp* repressor and the intracellular level of functional repressor are carefully balanced to allow differential regulation of the three operons. The structures of the operators that respond to the *trp* repressor therefore reflect the physiological roles these operators play in their respective organisms.

The operators of the *trp*, *aroH*, and *trpR* operons are located in different segments of their promoters (fig. 2) (Gunsalus and Yanofsky 1980; Singleton et al. 1980; Zurawski et al. 1981). However, the *trp* repressor presumably regulates transcription initiation in each operon by the same mechanism—excluding RNA polymerase from an essential recognition site in the promoter. It seems likely that the three operator sequences evolved independently from unrelated sequences. Thus these operator region sequences illustrate how convergent evolution has

provided specific recognition sites, each of which is used to regulate transcription initiation by responding to the same repressor.

Comparison of the Leader Regions of the *trp* Operons of Enterobacteria

The *trp* leader regions that have been sequenced have essentially the same overall structure; each specifies a transcript containing a peptide coding region followed by a segment that can form a structure, termed the "terminator," that is recognized by RNA polymerase as a transcription termination signal (fig. 3)

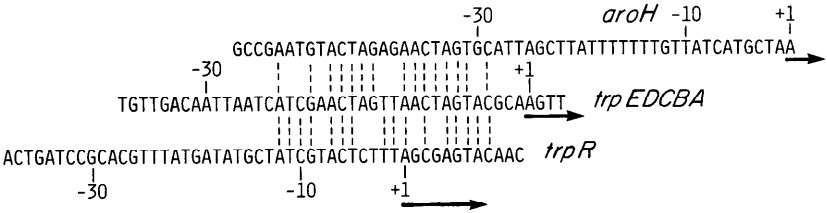


FIG. 2.—The promoter/operator regions of the *aroH*, *trpEDCBA*, and *trpR* operons of *Escherichia coli*. The operators are located in different segments of the respective promoters, suggesting that they arose by convergent evolutionary events. The transcription start sites are indicated by +1 and a horizontal arrow. Vertical dashed lines denote operator sequence identities.

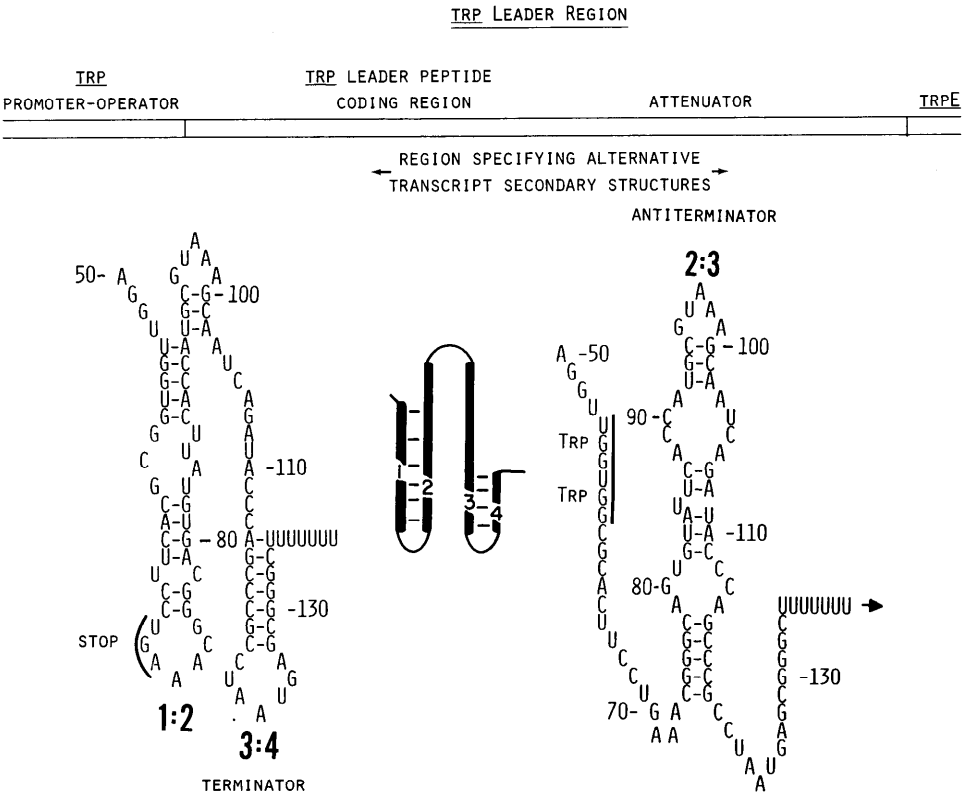


FIG. 3.—Organization of the *trp* leader region of *Escherichia coli*, and the alternative RNA secondary structures, 1:2, 2:3, and 3:4, that are believed to regulate transcription termination at the attenuator.

(Yanofsky 1981; Kolter and Yanofsky 1982; Farnham and Platt 1982; Ryan and Chamberlin 1983). The organization of the leader regions of other operons regulated by attenuation is very similar (Kolter and Yanofsky 1982).

Our current model of the details of attenuation control in the *trp* operon (Yanofsky et al., accepted) is as follows: As transcription proceeds over the initial portion of the leader region, the transcribing RNA polymerase molecule pauses when it completes the synthesis of the first RNA secondary structure, structure 1:2 (Farnham and Platt 1981; Winkler and Yanofsky 1981). (This structure in fact appears to be the pause signal; Fisher and Yanofsky [1983].) Immediately before or during the polymerase pause, a ribosome binds at the leader ribosome binding site and begins synthesis of the leader peptide. Polymerase then resumes transcription either spontaneously or in response to the approach of the translating ribosome. When RNA polymerase resumes transcription, the translating ribosome continues translation until it reaches the peptide stop codon, or, if there is a deficiency of charged $tRNA^{Trp}$, until it reaches one of the tandem Trp codons, where it stalls. When the latter occurs, structure 1:2 is disrupted, RNA segment 1 is masked by the stalled ribosome, and the antiterminator forms. Formation of the antiterminator precludes formation of the terminator, thereby preventing transcription termination at the attenuator. However, if there is adequate tryptophan, the translating ribosome moves to the stop codon and promptly dissociates from the transcript (Yanofsky et al., accepted; Kolter and Yanofsky, in preparation). When this occurs, either structure 1:2 or structure 2:3 forms the choice being discussed further on. If 1:2 forms, then 3:4 forms immediately thereafter, and transcription terminates at the attenuator. If 2:3 forms, then the terminator does not form, and transcription continues into the structural genes of the operon. In either case—whether tryptophan is limiting or not—the leader ribosome binding site pairs with a distal portion of the leader segment of the transcript (as detailed in fig. 8), and subsequent rounds of translation initiation are prevented (Das et al. 1983). Thus, we believe that the basis of attenuation is RNA polymerase's ability to recognize structure 3:4 as a termination signal; whenever this structure forms, termination is the consequence. Other features of the leader transcript and its translation must merely serve to regulate formation of this termination structure.

The *trp* operon leader region therefore is packed with essential structural information. The transcript of this region must be capable of forming structures 1:2, 2:3, and 3:4. The peptide coding region and the tandem Trp codons must be located appropriately to allow selection between alternative potential RNA secondary structures. The distal portion of the leader segment of the transcript must have the requisite sequence to pair with and shut down the leader ribosome binding site. In view of this richness of functionally essential sequences, it is of interest to compare the leader regions of various enterobacterial species to determine whether structure conservation reflects perceived events in the mechanism of attenuation.

The amino acid sequences of the *trp* leader peptides of various enterobacteria are compared in figure 4. The *S. dysenteriae* sequence is not presented; it is identical with the *E. coli* sequence. It is apparent that the crucial tandem Trp residues and their immediate neighbors are the most highly conserved residues in the peptide. We believe that the explanation for this conservation is that the RNA transcript segment encoding these residues forms part of a structure that

E.c.	Met	Lys	Ala	Ile	Phe	Val		Leu	Lys	Gly	Trp	Trp	Arg	Thr	Ser	END					
S.t.	Met	*ALA*	Ala	*THR*	Phe	*ALA*		Leu	*HIS*	Gly	Trp	Trp	Arg	Thr	Ser	END					
C.f.	Met	Lys	Ala	*THR*	Phe	Val		Leu	*HIS*	Gly	Trp	Trp	Arg	Thr	Ser	END					
S.m.	Met	*ASN*		*THR*	*TYR*	*ILE*	*SER*	Leu	*HIS*	Gly	Trp	Trp	Arg	Thr	Ser	*LEU*	*LEU*	*ARG*	*ALA*	*VAL*	END
K.a.	Met	Lys	*MET*	*HIS*	Phe	*ILE*	*THR*	Leu	*HIS*	*SER*	Trp	Trp	Arg	Thr	Ser	END					
E.a.	Met	*VAL*	*TYR*	*LEU*	Phe	*ASN*	*SER*	*ILE*	*THR*	Gly	Trp	Trp	Arg	*ILE*	Ser	*PRO*	END				

FIG. 4.—Predicted amino acid sequences of *trp* leader peptides. Residues that differ from those of *Escherichia coli* are in italics and underlined.

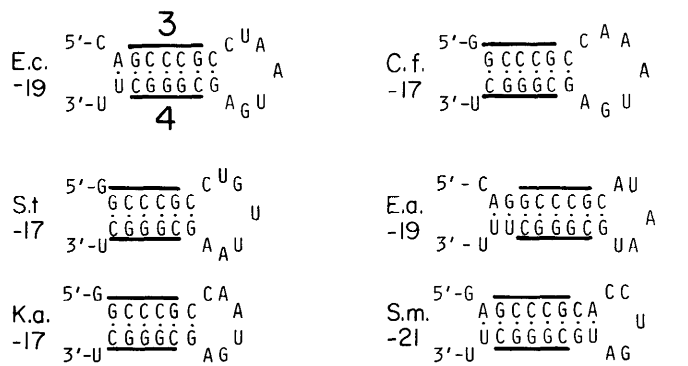


FIG. 5.—Comparison of the 3:4 terminator structures of six enterobacterial species. Calculated ΔG values are given below the species abbreviation. The crucial CGGGC and GCCCG sequences are under- or overlined by bold lines. Dots between bases signify base pairs. The pairing and stabilities of all the structures presented in this paper were predicted using the RNAFLD program of M. Zuker (Zuker and Steigler 1981).

participates in the regulation of transcription termination (see discussion of fig. 7).

In figure 5, I have compared the 3:4 secondary structures of six bacterial species. Bold lines demark five contiguous GC base pairs of a six base pair segment that is conserved. The five base sequence is also common to the 1:2 and 2:3 structures that participate in attenuation. When we compare the six predicted 3:4 structures, it is apparent that they are very similar and are approximately equally stable. However, notice that the loop base sequences vary widely. Thus, the sequence of a central, stable, base-paired segment and the overall spacing are conserved in the various structures. By contrast, when we compare the 2:3 secondary structures, we see appreciable sequence and structure variation (fig. 6). However, each of the transcripts can form a stable 2:3 secondary structure, and each structure can prevent formation of structure 3:4 by preempting the CGGGC segment that must be free for 3:4 formation to occur. Therefore, the overall stability of the structure is conserved, as well as a paired GC-rich segment that interferes with formation of structure 3:4. It is also apparent that the predicted stabilities of these structures vary greatly. The significance of this probably relates to the basal read-through level, as I shall discuss.

Considering next the 1:2 secondary structures (fig. 7), we see here as well appreciable sequence variation. What is conserved in these structures is the spacing between the tandem tryptophan codons and the CGGGC segment that forms part of structure 2:3 (marked by bold lines). This spacing is 14 or 15 nucleotides in each of the transcripts. Thus a ribosome stalled over the first or second Trp codon, if it masked approximately 10 nucleotides downstream from this codon,

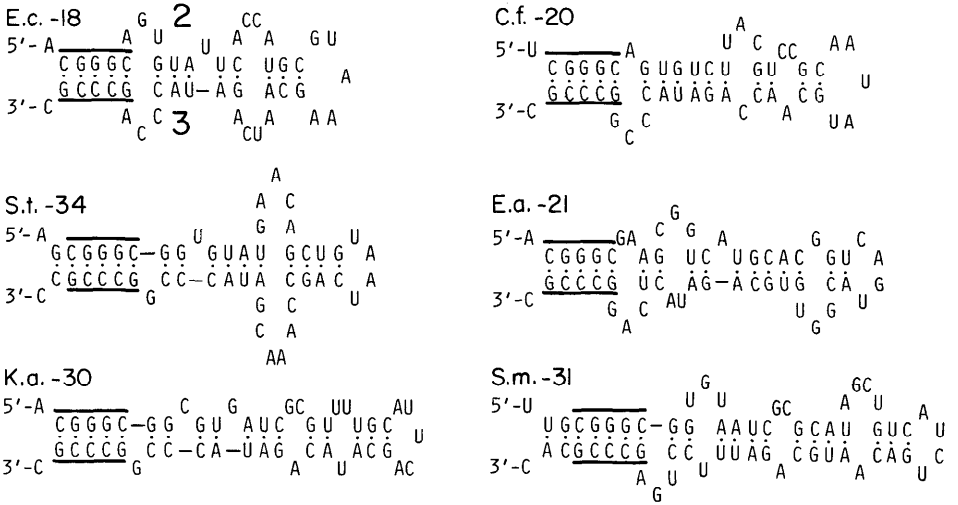


FIG. 6.—Comparison of the 2:3 antiterminator structures of six enterobacterial species. Calculated ΔG values are given alongside the species abbreviation. The CGGGC and GCCCG sequences are marked with bold lines.

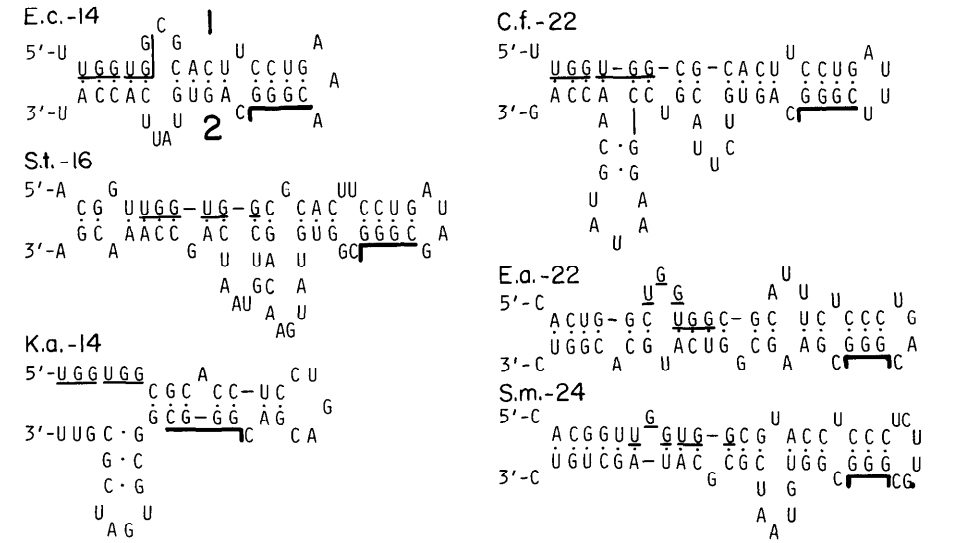


FIG. 7.—Comparison of the 1:2 structures of six enterobacterial species. Calculated ΔG values are given below the species abbreviation. The Trp codons and CGGGC sequence are underlined.

would free the CGGGC portion of segment 2, thereby allowing it to participate in the formation of the 2:3 secondary structure. Another feature of the 1:2 structures is that the relative locations of the Trp and translation stop codons vary somewhat; the leader peptide stop codon of *Serratia marcescens* follows the CGGGC sequence, whereas in the other bacterial transcripts the stop codon precedes this sequence. Other than the spacing between the Trp codons and the CGGGC sequence, conservation of the CGGGC sequence, and the moderate stabilities of the 1:2 structures, the structures do not have features in common. In each of the 1:2 structures, the nucleotide sequence immediately 3′ to the Trp

codons can theoretically pair with a complementary region in RNA segment 2. This required pairing may explain the conservation of the amino acid residues just beyond the Trp residues. The sequence of amino acids preceding the tandem Trp residues is not as highly conserved (fig. 4); this segment of the transcript does not form any important secondary structure. Thus, the leader transcript of each species can potentially form three alternative secondary structures, each containing the same CGGGC sequence.

When we compare the predicted stabilities of the various RNA secondary structures (table 1), it is evident that the relative stabilities of the 1:2 and 2:3 structures vary greatly from species to species, whereas the 3:4 structures are approximately equally stable. We believe that the relative stabilities of the 1:2 and 2:3 structures play a role in setting the different steady-state basal levels of read-through transcription that are appropriate to each bacterial species. Thus, as mentioned, when the translating ribosome dissociates at the leader peptide stop codon, we believe that either structure 1:2 or 2:3 can form (Yanofsky et al., accepted; Kolter and Yanofsky, in preparation). The ratio of these two structures—determined by their relative stabilities and by the position of the transcribing polymerase—presumably establishes how much read through will occur.

One additional feature of the leader region is conserved, the capacity of the leader ribosome binding site sequence to base pair with a slightly distal, complementary segment of the transcript (Yanofsky 1981; Blumenberg and Yanofsky 1982*a*; Das et al. 1983). As shown in figure 8, in *E. coli, Klebsiella aerogenes, Salmonella typhimurium,* and *Citrobacter freundii,* both the sequence of the Shine-Dalgarno region and a 6–7 base distal sequence immediately preceding structure 3:4 are highly conserved. In an organism in which there is significant sequence variation in these segments, *Serratia marcescens,* the corresponding regions have complementary changes that preserve perfect base pairing. We have shown in in vitro studies that base pairing of these segments in *S. typhimurium* can shut down synthesis of the leader peptide (Das et al. 1983). We assume that the purpose of this pairing is to stop synthesis of the leader peptide once the initial, translating ribosome has signaled the transcribing polymerase to terminate or continue transcription. The leader peptide itself has no known function.

Two additional aspects of attenuation raise questions of evolutionary interest. First, why do enteric bacteria use both repression and attenuation to control *trp* operon transcription—why would either not suffice? This question deserves se-

Table 1
Predicted Stabilities of *trp* Leader RNA Secondary Structures

	−ΔG's		
	1:2	2:3	3:4
Escherichia coli	14	18	19
Salmonella typhimurium	16	34	17
Klebsiella aerogenes	14	30	17
Citrobacter freundii	22	20	17
Erwinia amylovora	22	21	19
Serratia marcescens	24	31	21

NOTE.—Secondary structures and their stabilities were predicted using the RNAFLD program of M. Zuker (Zuker and Steigler 1981).

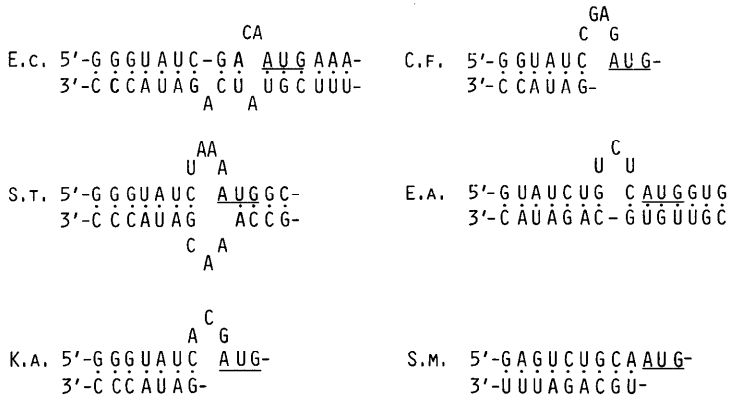


FIG. 8.—Ribosome binding site sequences paired with a complementary RNA segment. The start codon for each leader peptide is underlined.

rious attention now that we know that some biosynthetic operons are regulated by attenuation alone (Kolter and Yanofsky 1982). We have argued that perhaps, in the organism ancestral to the enterobacteria, *trp* operon transcription was regulated exclusively by attenuation, and *aroH* expression was regulated by repression. The existence of the tryptophan activated repressor could provide the selective basis for the gradual evolution of a repressor binding site within the *trp* operon promoter. According to this explanation, repression was added to the *trp* operon regulatory repertoire because the *trp* repressor was already present and performing an essential regulatory role in controlling *aroH* transcription.

A second explanation is based on whether repression and attenuation are physiologically redundant control mechanisms. We have shown that repression regulates *trp* operon transcription in the physiological range from excess tryptophan to mild tryptophan starvation, whereas attenuation regulates transcription in the range from mild tryptophan starvation to severe tryptophan starvation (C. Yanofsky, R. Kelley, and V. Horn, in preparation). Thus the two regulatory mechanisms function independently to expand the range of expression of the operon. This realization provides a plausible explanation for the existence of only two Trp codons in the *trp* leader transcript whereas there are many regulatory codons (seven, in the *his* operon) in the leader transcripts of other biosynthetic operons. The *trp* transcript apparently was designed to permit a response to *severe* tryptophan starvation only.

A further interesting question is, What is the genetic source of the DNA segments that evolved into the leader regions of operons regulated by attenuation? Ames et al. (1983) have noted that there is striking homology and structure conservation between the *his* leader transcript and $tRNA^{His}$ and have suggested that leader regions evolved from tRNA structural genes. This suggestion seems very plausible since tRNA structural genes contain regions of dyad symmetry that could be altered and added to, to allow specific regulation by attenuation. It remains to be seen whether other leader regions of operons regulated by attenuation have structural similarities to any tRNAs.

trp *Gene Arrangements in Various Microorganisms*

The genes of tryptophan biosynthesis of a wide range of microorganisms have been studied and their organization in the genomes of their respective organisms

has been established (Crawford 1975). In figure 9, I have presented the *trp* gene arrangements that have been deduced for the set of microorganisms I wish to discuss. The structural gene segments corresponding to the seven *trp* enzymatic functions, designated A through G, can be seen to be organized differently in these microorganisms. There is, in addition, an eighth gene of concern, *pabA*, that is homologous to *trpG* (Reiners et al. 1978; Kaplan and Nichols 1983). Genes *trpG* and *pabA* code for glutamine amidotransferase subunits that transfer an amido group from glutamine to chorismate in the synthesis of o- and p-amino benzoate, respectively. In some organisms, e.g., *Bacillus subtilis* (Kane et al. 1972) and *Acinetobacter calcoaceticus* (Sawula and Crawford 1972), a single bi-specific amidotransferase replaces the *trpG* and *pabA* proteins. This protein interacts with two different polypeptides to form distinct enzyme complexes that catalyze synthesis of one or the other of the two amino benzoates. In *B. subtilis*, this gene is not in the *trp* operon (Kane 1977), whereas, in *A. calcoaceticus*, it is adjacent to *trpD* (Sawula and Crawford 1972). When a monofunctional *trpG* is associated with other *trp* genes, it may exist as a separate gene, as in *S. marcescens* (Zalkin and Hwang 1971; Miozzari and Yanofsky 1979), or is fused to *trpD*, as in *S. typhimurium* and *E. coli* (Hwang and Zalkin 1971; Grieshaber and Bauerle 1972; Yanofsky et al. 1981), or is fused to *trpC*, as in *Saccharomyces cerevisiae* (Zalkin, personal communication) and *Neurospora crassa* (Schechtman and Yanofsky 1983) (fig. 9).

The various *trp* gene arrangements that exist in these microorganisms lead one to wonder why there have been so many genetic solutions to the problem of synthesizing tryptophan. One constraint in the possible evolution of a eukaryote gene from prokaryote genes certainly derives from the fact that the mature eukaryotic messenger generally contains a single coding region. Thus, whenever two functional polypeptide domains are required for a particular enzymatic activity and these domains are present in separate polypeptides in prokaryotes, it would not be surprising to find the corresponding structural genes fused in eukaryotes (Bonner et al. 1965). The separate *trpB* and *trpA* genes of *E. coli* and the fused *trpA·B* gene in yeast fulfill this expectation (Zalkin and Yanofsky 1981). However, the *trpE* and *trpG* functions also are normally present in an enzyme complex, yet in bacteria, yeast, and *Neurospora* these functions do not reside on the same polypeptide chain. Rather, in the latter organisms the *trpG* domain is fused to the *trpC* domain, and the *trpE* function is provided by a separate polypeptide chain (Zalkin, personal communication; Schechtman and Yanofsky 1983). The question also has been raised, Why are related genes organized in a single operon in some bacterial species, whereas in others they are separate or in several clusters? A

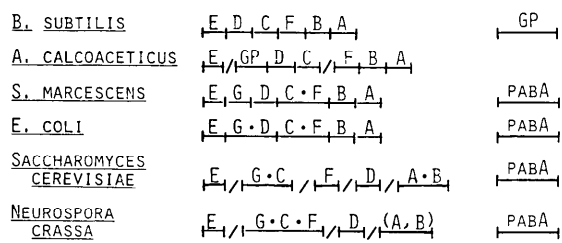


FIG. 9.—*trp* gene arrangements in various microorganisms. GP = the bi-specific *trpG-pabA* polypeptide.

reasonable explanation offered for this variation is that one of the biosynthetic intermediates in the pathway is used for some other purpose. This would necessitate genetic separation in order to achieve independent regulation or expression. The evolution of degradative pathways as well could provide the basis for the establishment of separate gene clusters. Alternatively, the selective pressures of the need for coordinate regulation may have led to the gene associations common in prokaryotes.

Consider the special case of gene fusion: There are two obvious potential advantages of fused genes over separate genes. If two polypeptides contain separate domains, both of which function in one enzyme complex, then it might be advantageous to join the domains in a single polypeptide to assure equal production and efficient localization. Moreover, if sequential reactions were performed by such a bifunctional protein, then the efficiency of the reaction sequence might be increased if the product of the first reaction, synthesized at one domain, were directly channeled to the second domain. Other regulatory considerations undoubtedly could also contribute to the selective pressures that lead to new gene arrangements. Our challenge is to explain the particular genomic configurations we observe today.

The nucleotide sequences of many of the *trp* genes of the organisms included in figure 9 have been determined. In addition, the sequences of the related *pabA* gene and the structural gene for the bi-specific *trpG-pabA* gene of *A. calcoaceticus* are now known (Kaplan and Nichols 1983; Nichols, personal communication). The deduced *trpG, pabA,* and *trpG-pabA* polypeptide amino acid sequences are highly conserved, with the *pabA* and *trpG-pabA* polypeptides most alike (Nichols, personal communication). We could imagine, as outlined in figure 10, that in a primitive prokaryotic ancestor, a single bi-specific G-P polypeptide existed and functioned in both the p-amino benzoate and o-amino benzoate pathways. This bi-specific gene then duplicated, with one copy invading the tryptophan operon, thus facilitating the evolution of a *trpG* dedicated solely to the tryptophan pathway, and the other *pabA* becoming specific for PABA synthesis. Subsequently, the *trpG* domain was fused to the *trpD* domain to give the *E. coli* arrangement. Along an independent line of descent, an already duplicated bi-specific G-P gene could have invaded an ancestral *trp* operon, locating itself immediatcly before *trpC*. Subsequently, this *trpG* could have become a dedicated gene, ultimately becoming fused to *trpC,* producing the arrangement we find in yeast and *Neurospora*. Why would some organisms have a bi-specific G-P protein, whereas in others specifically

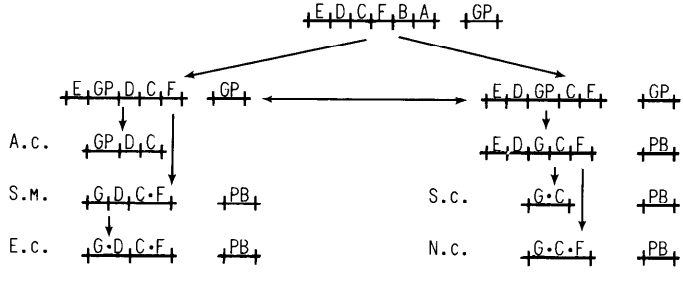


FIG. 10.—Hypothetical rearrangements in the evolution of some of the *trpG* and *pabA* associations we see today. GP = the structural gene for the bi-specific amidotransferase; PB = *pabA*, the structural gene for the p-amino benzoate pathway amidotransferase.

dedicated *trpG* and *pabA* proteins have evolved? Conceivably, organisms that spend much of their lifetime in a tryptophan-rich medium could not tolerate dependence on *trp* gene or *trp* operon expression for formation of a needed *pabA*-like activity; a separate *pabA* would overcome this difficulty. However, in organisms that rarely experience a tryptophan-rich environment, a bi-specific G-P polypeptide would suffice, and insertion of this gene into a *trp* operon would not result in a PABA deficiency. In fact, it might even be more efficient to coordinate regulation of G-P polypeptide production with *trp*-specific polypeptide formation.

In *S. marcescens, trpG* and *trpD* are adjacent genes, whereas in *E. coli* they are fused. The *trpG-trpD* junction of *S. marcescens* and the G·D fusion region of *E. coli* and several other enteric bacteria have been sequenced (Miozzari and Yanofsky 1979; Nichols et al. 1980). Inspection of the fusion region suggests how fusion may have occurred. We see that the spacer region of *S. marcescens* has become an amino acid connecting sequence in *E. coli* (fig. 11). The *trpG* stop codon has changed to a codon for leucine, and a deletion has placed the *trpG* and *trpD* sequences in the same reading frame. The Shine-Dalgarno region for *trpD* of *S. marcescens* has been replaced by nucleotides that probably are not recognized as a ribosome binding site. This presumably became necessary because retention of the Shine-Dalgarno region in such a fusion would permit secondary translation initiations at the internal ribosome binding site (Das and Yanofsky, in preparation).

If we compare the *E. coli* G·D and C·F amino acid sequences with those of the G·C·F polypeptide of *Neurospora* (fig. 12), we see homology throughout these genes. However, there are extra short polypeptide segments in each polypeptide that are not present in the other. Conceivably, these extra segments represent

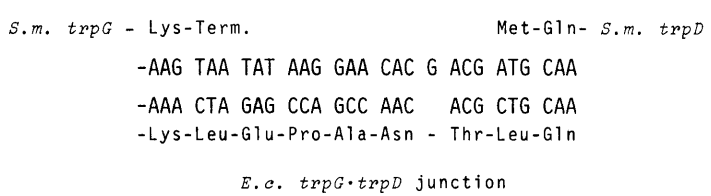


FIG. 11.—Comparison of the punctuation region between *trpG* and *trpD* of *Serratia marcescens* and the corresponding fusion region of *Escherichia coli trpG·D*.

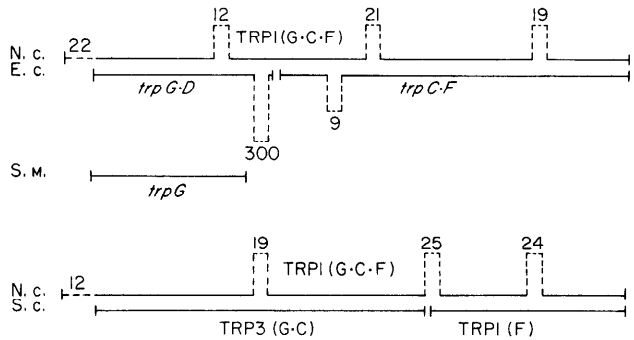


FIG. 12.—Amino acid sequence homology between TRP1 of *Neurospora crassa, trpG·D* and *trpC·F* of *Escherichia coli, trpG* of *Serratia marcescens*, and TRP2 (*G·C*) and TRP1 (F) of *Saccharomyces cerevisiae*. Heavy horizontal lines indicate segments with greater than 25% amino acid sequence identity in which there are relatively few additions or deletions. Dashed lines indicate nonhomologous segments. Elevated or lower dashed bars indicate extra internal amino acid residues. Numbers above or below the bars indicate the number of residues in each segment.

evolutionary relics of gene fusion events. They also may be essential spacer regions needed to separate domains. In addition, *trpD* of *E. coli* is not represented by a homologous segment in TRP1 of *Neurospora*. If we compare the G·C·F sequence of *Neurospora* with G·C and F of yeast (Zalkin, personal communication; Tschumper and Carbon 1980) (fig. 12), there also are connecting amino acid sequences between G and C and between C and F of *Neurospora* that are not present in yeast. There also is an additional sequence of amino acids in the F segment in *Neurospora* which, interestingly, is missing from both the *E. coli* *trpF* segment and yeast TRP1. Alignment of the bifunctional TRP5 polypeptide of yeast with the monofunctional *trpB* and *trpA* polypeptides of *E. coli* also reveals a connecting amino acid sequence between the A and B domains of the yeast polypeptide (fig. 13).

In these comparisons of *trp* genes, we see several instances of segments containing extra amino acids at presumed fusion junctions. These connecting sequences may be the translated equivalent of the untranslated spacer introns of the genes of higher eukaryotes. Thus, these segments may correspond to coding regions that were adjacent to a functionally selected segment prior to a gene rearrangement. Such segments may be retained as introns in higher eukaryotes where length may be irrelevant and where they subsequently would be excised at the messenger level, or they may be reduced in size by deletion in prokaryotes and lower eukaryotes, leaving a connecting amino acid sequence that either facilitates folding of adjacent polypeptide domains or remains functionally neutral.

The trpB-trpA *Gene Fusion*

In many of the bacterial species that have been studied to date, *trpB* and *trpA* are adjacent in the same transcriptional unit with *trpB* preceding *trpA* (fig. 13). Where nucleotide sequence information has been obtained, *trpB* and *trpA* are separated by the overlapping stop/start signals, UGAUG. Studies with *E. coli* suggest that this special punctuation sequence ensures equimolar production of the polypeptides specified by these adjacent genes (Oppenheim and Yanofsky

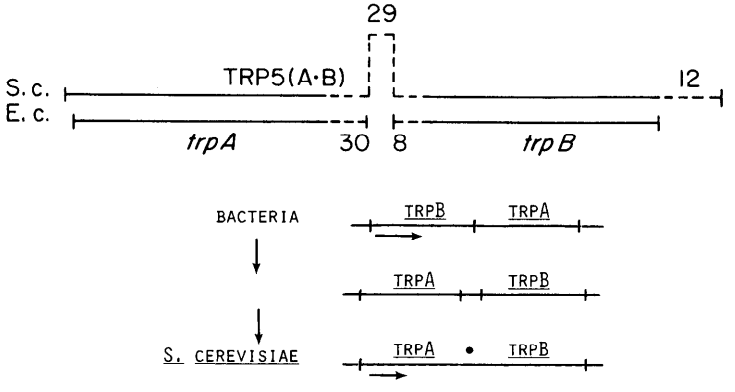


FIG. 13.—Amino acid sequence homology between TRP5 of *S. cerevisiae* and *trpA* and *trpB* of *Escherichia coli*. Heavy horizontal lines indicate segments with greater than 25% amino acid sequence identity in which there are relatively few additions or deletions. Dashed lines indicate nonhomologous segments. The elevated dashed bar indicates a segment of extra internal amino acid residues. The number above the bar indicates the number of residues in the extra segment. A hypothetical order of events in the evolution of the yeast sequence from the bacterial sequence is indicated below.

1980). A single base pair insertion or a double base pair deletion near the end of *trpB* or within the *trpB-trpA* punctuation region is all that is needed to evolve a fused *trpB-trpA* gene from the adjacent genes of prokaryotes. Yet when we compare the nucleotide sequence of such a fused gene, TRP5 of yeast, with *trpB* and *trpA* of *E. coli,* we find that the *trpA* segment is located before rather than after the *trpB* segment, and a connecting region now separates the A and B domains (Zalkin and Yanofsky 1981) (fig. 13). Why did not the *trpA-trpB* fusion occur by one of the simple events mentioned? It is conceivable that such a simple fusion did not occur because the fused gene would produce a nonfunctional protein. Thus, if either the C-terminal segment of the *trpB* polypeptide or the N-terminal segment of the *trpA* polypeptide must exist free so that it can contribute to a catalytically essential protein conformation, then restricting this freedom would result in enzyme inactivity. To explore this possibility, we have fused *trpB* in phase with *trpA* by deleting ~90 base pairs within the distal segment of *trpB* (Das and Yanofsky, in preparation). The deletion changes the amino acid sequence at the end of the *trpB* portion but leaves unaltered the amino acid sequence of the *trpA* portion. In this fusion, the Shine-Dalgarno region (for *trpA* translation) at the distal end of *trpB* is unaffected, and a normal *trpA* polypeptide is synthesized. The expected B·A fusion polypeptide also is produced. This polypeptide was separated from free A protein by column chromatography and was found to be deficient in both *trpA* enzymatic activity and the ability to complex with the *trpB* polypeptide (Das and Yanofsky, in preparation). Therefore, we have tentatively concluded that tethering the amino terminal end of the *trpA* polypeptide reduces its biological activity. We have one reservation regarding this experiment, namely, that the foreign amino acid sequence present at the end of the *trpB* portion of the fusion protein may be responsible for the observed effect. We plan to fuse *trpB* and *trpA* intact, in phase, and produce a fusion protein that has only *trpB* and *trpA* amino acid sequences.

Assessing the Functional Significance of Amino Acid Replacements in Homologous *trpA* Polypeptides

The *trpA*s and their encoded polypeptides of *E. coli, S. typhimurium,* and *K. aerogenes* have been sequenced and compared (Li and Yanofsky 1973*a,* 1973*b;* Nichols and Yanofsky 1979; Nichols et al. 1981). Approximately one-fourth of the base pairs differ in any pairwise comparison of these *trpA*s; however, the majority of the differences generate synonymous codons (table 2). (For comparison with other genes and proteins, see chaps. 1, 2, 7, 9, 11, and 12 in Nei and Koehn [1983].) Thus the *trpA* polypeptides of *E. coli* and *S. typhimurium* have only 40 amino acid differences (table 2). We have examined the functional significance of these amino acid replacements by producing interspecies hybrid pro-

Table 2
Species Differences: *trpA*s and Their Polypeptides

	Base Pairs (of 804–807)	Synonymous Codons	Amino Acid Residues (of 268–269)
Escherichia coli vs. *Salmonella typhimurium*	199	125	40
E. coli vs. *Klebsiella aerogenes*	189	124	34
S. typhimurium vs. *K. aerogenes*	201	111	44

teins containing contiguous segments derived from the parental *trpA* proteins (Schneider et al. 1981). The interspecies hybrid *trpA*s were produced by generating recombinants between the two *trpA*s contained on compatible, multicopy plasmids (fig. 14). Recombinants were selected that had a functional *trpB* protein; most of these also had a crossover in *trpA*. Many thousands of recombinants were screened for any that produced a partially or completely inactive *trpA* protein, but none was detected. A topological representation of the distribution of amino acid replacements in the parental proteins is shown in figure 15 along with a schematic diagram of the sites of exchange for several interspecies hybrid proteins that were examined further (Schneider et al. 1981). These hybrid proteins were catalytically indistinguishable from their parental proteins. The only difference noted was the somewhat greater thermal lability of a few of the hybrid proteins in crude extracts (Schneider et al. 1981). Recently, guanidine denaturation analyses were performed with one of the hybrid proteins (strain 14-26; fig. 15) and the parental proteins (Yutani et al., accepted). It was observed that the hybrid protein was readily distinguishable from its parental proteins. However, the behavior of the hybrid protein was explained by the realization that its response to denaturation reflected

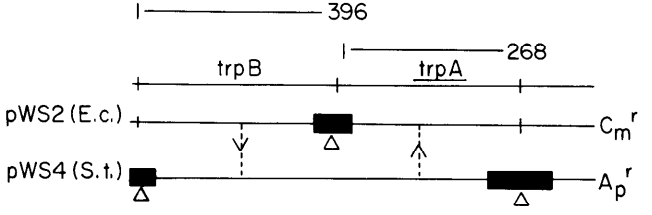


FIG. 14.—Scheme used to produce recombinants between *trpA* of *Escherichia coli* and *trpA* of *Salmonella typhimurium*. The *trpB-trpA* segments of the two organisms were present on compatible multicopy plasmids contained in the same bacterium. The black bars denote deletions that were introduced that inactivated both *trpB* and *trpA* of each plasmid. pWS2 carried chloramphenicol resistance (Cm^r) while pWS4 carried ampicillin resistance (Ap^r). For further details see Schneider et al. (1981).

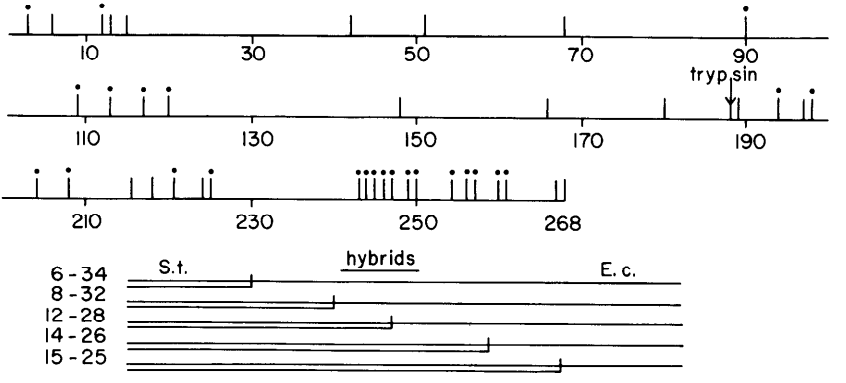


FIG. 15.—Amino acid differences in the *trpA* polypeptides of *Escherichia coli* and *Salmonella typhimurium* and the structures of several interspecies hybrids. Vertical lines mark the position of each amino acid difference. A dot over a vertical line indicates that the respective amino acid residues have different properties, e.g., an acidic replacing a neutral. Each hybrid is designated by the number of differences that are derived from *S. typhimurium* followed by the number from *E. coli*, e.g., 6-34. The arrow labeled "trypsin" is the site of cleavage of the *trpA* protein that produces two independently folding polypeptide domains (Higgins et al. 1979).

the parental source of the two domains of the *trpA* polypeptide (Higgins et al. 1979); these domains are known to fold independently (fig. 15) and to respond differently to guanidine denaturation (Yutani et al., accepted). All the findings with the interspecies hybrids therefore suggest that none of the residues that differ in the two *trpA* proteins depends for its contribution to function on the presence of any of the other amino acid replacements in the protein. Apparently, many of the amino acid replacements in the *trpA* proteins are functionally inconsequential.

Conclusions

Among microorganisms that synthesize tryptophan, we observe variation in regulatory and structural gene regions that obviously are compatible with expression of these genes and their specified polypeptides in their respective organisms. Clearly, each evolutionary solution has been successful, and presumably each solution reflects the natural history of the organism and its ancestors. However, we cannot yet confidently conclude that the particular gene arrangements or regulatory mechanisms that we observe in nature today have survived because they are extraordinarily effective as opposed to being merely acceptable. With the application of current technology in the manipulation of these genes and their regulatory regions, it may be possible to devise experiments that explain the molecular events involved in their evolution and the reasons for their occurrence.

Acknowledgments

The author greatly appreciates the valuable comments and suggestions of Irving Crawford, Howard Zalkin, and Brian Nichols, and the permission of Howard Zalkin, Brian Nichols, and Edith Miles to quote unpublished observations. The author is indebted to Michael Zuker and the MOLGEN project for the use of the RNAFLD secondary structure predict program. The studies performed in the author's laboratory that are described in this article were supported by grants from the National Science Foundation, the U.S. Public Health Service, the American Heart Association, and the American Cancer Society. The author is a career investigator of the American Heart Association.

This article is dedicated to the memory of David M. Bonner, who was one of the first to ponder the evolutionary significance of gene fusions.

LITERATURE CITED

AMES, B. N., T. TSANG, M. BUCK, and M. F. CHRISTMAN. 1983. The leader mRNA of the histidine attenuator region resembles $tRNA^{His}$: possible general regulatory implications. Proc. Nat. Acad. Sci. **80**:5240–5242.

BENNETT, G. N., and C. YANOFSKY. 1978. Sequence analysis of operator constitutive mutants of the tryptophan operon of *Escherichia coli*. J. Mol. Biol. **121**:179–192.

BLUMENBERG, M., and C. YANOFSKY. 1982*a*. Regulatory region of the *Klebsiella aerogenes* tryptophan operon. J. Bacteriol. **152**:49–56.

———. 1982*b*. Evolutionary divergence of the *Citrobacter freundii* tryptophan operon regulatory region: comparison with other enteric bacteria. J. Bacteriol. **152**:57–62.

BONNER, D. M., J. A. DEMOSS, and S. E. MILLS. 1965. The evolution of an enzyme. Pages 305–318 *in* V. BRYSON and H. J. VOGEL, eds. Evolving genes and proteins. Academic Press, New York.

BROWN, K. D. 1968. Regulation of aromatic amino acid biosynthesis in *Escherichia coli* K12. Genetics **60**:31–48.

CRAWFORD, I. P. 1975. Gene arrangements in the evolution of the tryptophan synthetic pathway. Bacteriol. Rev. **39**:87–120.

DAS, A., J. URBANOWSKI, H. WEISSBACH, J. NESTOR, and C. YANOFSKY. 1983. *In vitro* synthesis of the tryptophan operon leader peptides of *Escherichia coli, Serratia marcescens* and *Salmonella typhimurium*. Proc. Nat. Acad. Sci. **80**:2879–2883.

FARNHAM, P. J., and T. PLATT. 1981. Rho-dependent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. Nucleic Acids Res. **9**:563–577.

———. 1982. Effects of DNA base analogs on transcription termination at the tryptophan operon attenuator in *Escherichia coli*. Proc. Nat. Acad. Sci. **79**:998–1002.

FISHER, R., and C. YANOFSKY. 1983. A complementary DNA oligomer releases a transcription pause complex. J. Biol. Chem. **258**:9208–9212.

GRIESHABER, M., and R. BAUERLE. 1972. Structure and evolution of a bifunctional enzyme of the tryptophan operon. Natur. New Biol. **236**:232–235.

GUNSALUS, R. P., and C. YANOFSKY. 1980. Nucleotide sequence and expression of *Escherichia coli trpR*, the structural gene for the *trp* aporepressor. Proc. Nat. Acad. Sci. **77**:7117–7121.

HIGGINS, W., T. FAIRWELL, and E. W. MILES. 1979. An active proteolytic derivative of the α subunit of tryptophan synthase. Identification of the site of cleavage and characterization of the fragments. Biochemistry **18**:4827–4835.

HWANG, L. H., and H. ZALKIN. 1971. Multiple forms of anthranilate synthetase-anthranilate 5-phosphoribosylpyrophosphate phosphoribosyl transferase from *Salmonella typhimurium*. J. Biol. Chem. **246**:2338–2345.

JACKSON, E. N., and C. YANOFSKY. 1973. The region between the operator and first structural gene of the tryptophan operon of *Escherichia coli* may have a regulatory function. J. Mol. Biol. **76**:89–101.

KANE, J. F. 1977. Regulation of a common amidotransferase subunit. J. Bacteriol. **132**:419–425.

KANE, J. F., W. M. HOLMES, and R. A. JENSEN. 1972. Metabolic interlock: the dual function of a folate pathway gene as an extra-operonic gene of tryptophan biosynthesis. J. Biol. Chem. **247**:1587–1596.

KAPLAN, J. B., and B. P. NICHOLS. 1983. Nucleotide sequence of *Escherichia coli pabA* and its evolutionary relationship to *trp(G)D*. J. Mol. Biol. **168**:451–468.

KELLEY, R. L., and C. YANOFSKY. 1982. *trp* aporepressor production is controlled by autogenous regulation and inefficient translation. Proc. Nat. Acad. Sci. **79**:3120–3124.

KOLTER, R., and C. YANOFSKY. 1982. Attenuation in amino acid biosynthetic operons. Annu. Rev. Genet. **16**:113–134.

LI, S. L., and C. YANOFSKY. 1973*a*. Amino acid sequence studies with the tryptophan synthetase α chain of *Salmonella typhimurium*. J. Biol. Chem. **248**:1830–1836.

———. 1973*b*. Amino acid sequence studies with the tryptophan synthetase α chain of *Aerobacter aerogenes*. J. Biol. Chem. **248**:1837–1843.

MANSON, M., and C. YANOFSKY. 1976. Naturally occurring sites within the *Salmonella dysenteriae* operon severely limit tryptophan biosythesis. J. Bacteriol. **126**:668–678.

MIOZZARI, G., and C. YANOFSKY. 1978. A naturally occurring promoter down mutation: nucleotide sequence of the *trp* promoter/operator/leader region of *Shigella dysenteriae* 16. Proc. Nat. Acad. Sci. **75**:5580–5584.

———. 1979. Gene fusion during the evolution of the *trp* operon in Enterobacteriaceae. Nature **277**:486–489.

NEI, M., and R. K. KOEHN. 1983. Evolution of genes and proteins. Sinauer, Sunderland, Mass. 331 pages.

NICHOLS, B. P., M. BLUMENBERG, and C. YANOFSKY. 1981. Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the *trp* operon of *Klebsiella aerogenes, Salmonella typhimurium* and *Escherichia coli*. Nucleic Acids Res. **9**:1743–1755.

NICHOLS, B. P., G. F. MIOZZARI, M. VAN CLEEMPUT, G. N. BENNETT, and C. YANOFSKY. 1980. Nucleotide sequences of the *trpG* regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens*. J. Mol. Biol. **142:**503–517.

NICHOLS, B. P., and C. YANOFSKY. 1979. Nucleotide sequences of *trpA* of *Salmonella typhimurium* and *Escherichia coli:* an evolutionary comparison. Proc. Nat. Acad. Sci. **76:**5244–5248.

OPPENHEIM, D. S., and C. YANOFSKY. 1980. Translational coupling during expression of the tryptophan operon of *Escherichia coli*. Genetics **95:**785–795.

REINERS, J. J., L. J. MESSENGER, and H. ZALKIN. 1978. Immunological cross-reactivity of *Escherichia coli* anthranilate synthetase, glutamate synthetase and other proteins. J. Biol. Chem. **253:**1226–1233.

ROSE, J. K., C. L. SQUIRES, C. YANOFSKY, H. L. YANG, and G. ZUBAY. 1973. Regulation of *in vitro* transcription of the tryptophan operon by purified RNA polymerase in the presence of partially purified repressor and tryptophan. Natur. New Biol. **245:**133–137.

RYAN, T., and M. CHAMBERLIN. 1983. Transcription analyses with heteroduplex *trp* attenuator templates indicate that the transcript stem and loop structure serves as the termination signal. J. Biol. Chem. **258:**4690–4693.

SAWULA, R. V., and I. P. CRAWFORD. 1972. Mapping of the tryptophan genes of *Acinetobacter calcoaceticus* by transformation. J. Bacteriol. **112:**797–805.

SCHECHTMAN, M. G., and C. YANOFSKY. 1983. Structure of the trifunctional *trp-1* gene from *Neurospora crassa* and its aberrant expression in *Escherichia coli*. J. Mol. Appl. Genet. **2:**89–93.

SCHNEIDER, W., B. NICHOLS, and C. YANOFSKY. 1981. Procedure for production of hybrid genes and proteins and its use in assessing significance of amino acid differences in homologous tryptophan synthetase α polypeptides. Proc. Nat. Acad. Sci. **78:**2169–2173.

SINGLETON, C. K., W. D. ROEDER, G. BOGOSIAN, R. L. SOMERVILLE, and H. L. WEITH. 1980. DNA sequence of the *E. coli trpR* gene and prediction of the amino acid sequence of Trp repressor. Nucleic Acids Res. **8:**1551–1560.

TSCHUMPER, G., and J. CARBON. 1980. Sequence of a yeast DNA fragment containing chromosomal replicator and the TRP1 gene. Gene **10:**157–166.

WINKLER, M. E., and C. YANOFSKY. 1981. Pausing of RNA polymerase during *in vitro* transcription of the tryptophan operon leader region. Biochemistry **20:**3738–3744.

YANOFSKY, C. 1981. Attenuation in the control of expression of bacterial operons. Nature **289:**751–758.

YANOFSKY, C., A. DAS, R. FISHER, R. KOLTER, and V. BERLIN. Accepted. Attenuation control of *trp* operon expression. *In* D. HAMER and M. ROSENBERG, eds. UCLA symposium on gene expression. Academic Press, New York.

YANOFSKY, C., T. PLATT, I. CRAWFORD, B. NICHOLS, G. CHRISTIE, H. HOROWITZ, M. VAN CLEEMPUT, and A. WU. 1981. The complete nucleotide sequence of the tryptophan operon of *Escherichia coli*. Nucleic Acids Res. **9:**6647–6668.

YUTANI, K., T. SATO, K. OGASAHARA, and E. W. MILES. Accepted. Comparison of denaturation of tryptophan synthase α subunits from *Escherichia coli,* from *Salmonella typhimurium,* and from an interspecies hybrid. Biochem. Biophys. Acta.

ZALKIN, H., and L. H. HWANG. 1971. Anthranilate synthetase from *Serratia marcescens:* on the properties and relationships to the enzyme from *Salmonella typhimurium*. J. Biol. Chem. **246:**6899–6907.

ZALKIN, H., and C. YANOFSKY. 1981. Yeast *trp5:* structure, function, regulation. J. Biol. Chem. **257:**1491–1500.

ZUKER, M., and P. STEIGLER. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9:**133–148.

ZURAWSKI, G., R. P. GUNSALUS, K. D. BROWN, and C. YANOFSKY. 1981. Structure and regulation of *aroH*, the structural gene for the tryptophan repressible 3-deoxy-D-arabino-heptulosonic-acid-7-phosphate synthase of *Escherichia coli*. J. Mol. Biol. **145:**47–73.

WALTER M. FITCH, reviewing editor

Received October 25, 1983.

Selective Neutrality of Glucose-6-Phosphate Dehydrogenase Allozymes in *Escherichia coli*[1]

Daniel E. Dykhuizen, Jean de Framond, and Daniel L. Hartl
Washington University School of Medicine

Six naturally occurring alleles representing four electromorphs of the enzyme glucose-6-phosphate dehydrogenase were transferred by P1-mediated transduction from natural isolates of *Escherichia coli* into the genetic background of *E. coli* K12 and were studied in pairwise competition in chemostats limited for glucose in order to estimate differences in growth rate associated with the alleles. Although the level of resolution of such experiments is a growth rate differential of approximately 0.002 h^{-1}, no significant differences among the strains were found. Studies of apparent K_m and V_{max} in crude enzyme extracts of the strains also failed to reveal any significant differences among the electromorphs. These results support the view that the alleles are selectively neutral or nearly neutral under these conditions.

Introduction

This paper reports the third in a series of loci in which naturally occurring allozyme alleles are transferred by transduction into the genetic background of *E. coli* K12 and examined for their effects on growth rate in chemostats. The objective of the study has been to determine by direct experiment the selective effects, if any, of such alleles and to define the nutritional or environmental conditions in which selection might be detected. Previous reports have focused on alleles of the *gnd* locus, which codes for 6-phosphogluconate dehydrogenase (Dykhuizen and Hartl 1980; Hartl and Dykhuizen 1981), and on alleles of the *pgi* locus, which codes for phosphoglucose isomerase (Dykhuizen and Hartl 1983*a*). The present report concerns alleles of the *zwf* locus, which is located at 41 min on the standard genetic map (Bachmann 1983) and codes for glucose-6-phosphate dehydrogenase (D-glucose-6-phosphate: nicotinamide adenine dinucleotide phosphate [NADP] oxidoreductase, EC 1.1.1.49). The rationale of the study is that conclusions about the selective effects of allozymes may have little generality unless based on studies of a number of loci. The *zwf* locus was chosen for consideration because glucose-6-phosphate dehydrogenase (G6PD) catalyzes the first

1. Key words: glucose-6-phosphate dehydrogenase, neutrality, allozymes.

Address for correspondence and reprints: Daniel E. Dykhuizen, Department of Genetics, Box 8031, Washington University School of Medicine, St. Louis, Missouri 63110.

Mol. Biol. Evol. 1(2):162–170. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0008$02.00

reaction in one branch of the two principal pathways of glucose-6-phosphate catabolism, the other branch being catalyzed by phosphoglucose isomerase, an enzyme studied previously (Dykhuizen and Hartl 1983*a*).

Glucose-6-phosphate dehydrogenase plays a key role in the oxidative branch of the pentose phosphate pathway. In *E. coli,* this pathway is the route through which an estimated 10%–30% of total utilized glucose is metabolized (Cohen 1951; Wang et al. 1958; Model and Rittenberg 1967); it accounts for approximately 50% of the total ribose produced by the cells (Johnson et al. 1973); and it is generally agreed to be an important source of reducing power (Katz and Rognstad 1967). Although the enzymes of the pathway, including G6PD, are produced constitutively in *E. coli* (Fraenkel and Levisohn 1967; Fraenkel 1968; Fraenkel and Banerjee 1971), little is known about the overall regulation of the pathway other than that it is indeed regulated, most likely at the level of G6PD itself (Orthner and Pizer 1974).

Given the metabolic importance of G6PD, it would not be surprising if selective constraints on the functional properties of the molecule were so stringent as to allow little functionally significant genetic polymorphism to remain segregating in natural populations. However, structural polymorphism does occur. Among 109 isolates of *E. coli* from diverse natural sources, five electromorphs of G6PD were recovered (Milkman 1973; Selander and Levin 1980). Here we provide evidence that the structural variants are selectively neutral (functionally equivalent) insofar as can be determined by competition experiments in glucose-limited chemostats.

Material and Methods

Strains

Strains used in the study are listed in table 1. Those used in the constructions or as controls are in the upper part of the table, the natural isolate sources of the *zwf* alleles are in the middle, and the strains used in the experiments are at the bottom. A zwf^+ allele from a natural isolate is designated as zwf^+ followed in parentheses by its strain of origin; zwf^+(*RM73C*) thus indicates the zwf^+ allele originally present in strain RM73C. Most strains in table 1 come in pairs consisting of the original $T5^S$ (bacteriophage T5-sensitive) strain and one (or in two cases, two) spontaneous $T5^R$ (*fhuA*) mutant derivatives. *fhuA,* formerly called *tonA,* is located at 4 min on the standard *E. coli* map and is concerned with ferric hydroxamate uptake, mutants also being resistant to bacteriophage T5 (Bachmann 1983). *fhuA* is usually a selectively neutral marker convenient for chemostat studies, but in the chemostat-adapted strain DD1296 and its derivatives, the marker is nonneutral. This potential problem can be overcome by means of the proper controls. The strains bearing naturally occurring zwf^+ alleles in the genetic background of DD1296 were derived by means of the method outlined for DD1298 (and its DD1330 derivative) in table 2. Three sequential transductions of zwf^+ into DD725 were followed by one into DD1296, and at the two final steps the electromorph type was verified by electrophoresis. Since there is no direct selection for Zwf^+, recipient strains DD725 and DD1296 carried an *eda-edd-zwf* deletion (*edd-1*), permitting transfer of zwf^+ by cotransduction with *eda,* selection for Eda^+ being on minimal glucuronate. The *edd* locus, at 41 min on the standard map (Bachmann 1983), codes for phosphogluconate dehydratase (EC 4.2.1.12) and is important in gluconate metabolism but not in glucose metabolism. The closely

Table 1
***Escherichia coli* Strains**

Strain (and Spontaneous T5^R Derivatives)[a]	Relevant Genotype (or Origin)	Zwf	Source
DD725	*edd-1*,[b] *rpsL*	△[b]	Dykhuizen and Hartl 1983[a]
DD1296	*edd-1*,[b] *rpsL*	△[b]	From 250-h glucose-limited DD725 chemostat
DF1071	*Hfr*[c] *gnd-1, relA1, pit-10, spoT1, fhuA22*, T2^R	4[d]	Fraenkel (1968) via Coli Genetics Stock Center
DD1144 (DD1157)	*zwfA2, rpsL* (*fhuA*)	0[e]	Dykhuizen and Hartl 1983[a]
DD1310 (DD1336)	*zwfA2, rpsL* (*fhuA*)	0[e]	P1 from DD1144 × DD1296, Eda$^+$ selection[f]
DD1146 (DD1159)	*zwf*$^+$(*K12*), *rpsL* (*fhuA*)	4[d]	Dykhuizen and Hartl 1983[a]
DD1312 (DD1337)	*zwf*$^+$(*K12*), *rpsL* (*fhuA*)	4[d]	P1 from DD1146 × DD1296, Eda$^+$ selection[f]
DD1196 (DD1290)	*edd-1*,[b] *rpsL* (*fhuA*)	△[b]	Dykhuizen and Hartl 1983[a]
DD1287 (DD1288)	*zwf*$^+$(*K12*), *rpsL* (*fhuA*)	4[d]	P1 from DF1071 × DD1196, Eda$^+$ selection[f]
DD1137 (DD1150)	*zwf*$^+$(*RM73C*), *rpsL* (*fhuA*)	2	RM73C and DD725 by method in table 2
RM183E	(Elephant)	1	R. Milkman via B. Levin [1[g]]
RM73C	(Orangutan, female)	2	R. Milkman via B. Levin [3[g]]
RM77C	(Human, female)	3	R. Milkman via B. Levin [15[g]]
RM66A	(Human, male)	4a[h]	R. Milkman via B. Levin [86[g]]
RM72B	(Gorilla, female)	4b[h]	R. Milkman via B. Levin [86[g]]
RM20	(Red wolf, female)	5a[h]	R. Milkman via B. Levin [4[g]]
RM182A	(Rabbit)	5b[h]	R. Milkman via B. Levin [4[g]]
DD1298 (DD1330)	*zwf*$^+$(*RM73C*), *rpsL* (*fhuA*)	2	RM73C and DD1296 by method in table 2
DD1301 (DD1331)	*zwf*$^+$(*RM77C*), *rpsL* (*fhuA*)	3	RM77C and DD1296 by method in table 2
DD1302 (DD1332, DD1357)	*zwf*$^+$(*RM66A*), *rpsL* (*fhuA*)	4a[h]	RM66A and DD1296 by method in table 2
DD1304 (DD1333, DD1358)	*zwf*$^+$(*RM72B*), *rpsL* (*fhuA*)	4b[h]	RM72B and DD1296 by method in table 2
DD1306 (DD1334)	*zwf*$^+$(*RM20*), *rpsL* (*fhuA*)	5a[h]	RM20 and DD1296 by method in table 2
DD1308 (DD1335)	*zwf*$^+$(*RM182A*), *rpsL* (*fhuA*)	5b[h]	RM182A and DD1296 by method in table 2

[a] All DD strains are F$^-$.

[b] *edd-1* and △ refer to a deletion △*eda-edd-zwf*.

[c] Hfr C of Cavalli.

[d] K12 allozyme is electrophoretically indistinguishable from type 4 electromorphs.

[e] *zwfA2* gene product has no detectable G6PD activity.

[f] Eda$^+$ selection is on minimal glucuronate.

[g] Number of strains in the corresponding electromorph class found among 109 natural isolates (Milkman 1973; Selander and Levin 1980). The observed distribution of electromorphs is not significantly different from that expected with selective neutrality.

[h] a and b refer to independent isolates of electromorphs in the same electrophoretic class.

Table 2
Method of Strain Construction

Strain	Relevant Genotype	Source
RM73C	*zwf*$^+$(*RM73C*)	Natural isolate
DD938.....	*zwf*$^+$(*RM73C*), *rpsL*	P1 from RM73C × DD725, Eda$^+$ selection
DD1121.....	*zwf*$^+$(*RM73C*), *rpsL*	P1 from DD938 × DD725, Eda$^+$ selection
DD1137.....	*zwf*$^+$(*RM73C*), *rpsL*	P1 from DD1121 × DD725, Eda$^+$ selection
DD1298.....	*zwf*$^+$(*RM73C*), *rpsL*	P1 from DD1137 × DD1296, Eda$^+$ selection
DD1330.....	*zwf*$^+$(*RM73C*), *rpsL fhuA*	Spontaneous T5^R in DD1298

linked locus *eda* codes for 2-keto-3-deoxygluconate 6-phosphate aldolase (EC 4.1.2.14) (Bachmann 1983). The symbolism 4a and 4b in table 1 designates two independent naturally occurring *zwf*$^+$ alleles of electromorph type 4, and similarly for 5a and 5b. In brackets at the right of the RM strains in table 1 are the number of strains, among 109 natural isolates, found to carry the corresponding electromorph.

Strain DD1296 warrants some comment because it was selected as a clone from a chemostat population of DD725 that had been maintained for 250 h in limiting glucose. This strain was chosen as the genetic background for the wild-type alleles in hope that such a chemostat-adapted strain would not be as susceptible to periodic selection as unadapted strains and would therefore allow longer experiments to be carried out. Periodic selection (Atwood et al. 1951) refers to the hitchhiking of genetic markers with favorable mutations that occur in chemostat culture. Although strain DD1296 is demonstrably different from strain DD725 in its ability to grow in chemostats (data not shown), the characteristics of periodic selection in the two strains are indistinguishable.

Chemostats

Chemostat medium consists of Davis salts (40 mM K_2HPO_4, 15 mM KH_2PO_4, 7.6 mM $[NH_4]_2SO_4$, 1.7 mM sodium citrate, and 0.8 mM $MgSO_4$) with either limiting glucose (0.1 g/liter) or limiting gluconate (0.1 g/liter). Inoculation and sampling procedures were as described previously (Dykhuizen and Hartl 1983*a*) except that the total number of colonies counted for each sampled point was approximately 4,000. The dilution rate was adjusted to an average generation time of 1.69 h, and the standard deviation among experiments was ± 0.09 h. The difference in specific growth rate between two competing strains, A and B, is estimated as the slope of the linear regression of ln (A[*t*]/B[*t*]) against *t;* A(*t*) and B(*t*) are their relative densities at time *t* hours after inoculation (Dykhuizen and Hartl 1983*b*). This slope, conventionally denoted *s*, measures the selection coefficient per hour. The approximate selection coefficient per generation can be calculated as (ln 2)*s*/*D*, where *D* is the fraction of the chemostat volume replaced per hour (Dykhuizen and Hartl 1983*b*). The average value of *D* in the present experiments is $D = 0.41$. Statistical significance of a slope was assessed by means of the *F*-statistic in an analysis of variance of the regression (Snedecor and Cochran 1967). Statistical significance of the difference between two slopes was assessed by means of the *t*-statistic calculated as in Snedecor and Cochran (1967). Type I error refers to the rejection of a null hypothesis when it is, in fact, true.

Enzyme Studies

Overnight cultures were washed and sonicated in buffer (10 mM Tris hydrochloride, 10 mM $MgCl_2$, and 1 mM dithiothreitol, pH 7.8). Reactions were carried out in 1 mM glucose-6-phosphate and 0.4 mM NADP in incubation buffer (50 mM Tris hydrochloride, 10 mM $MgCl_2$, and 0.3 mM dithiothreitol, pH 7.6). K_m and V_{max} were estimated from equations described previously (Dykhuizen and Hartl 1983*a*).

Results

Controls

Figure 1 exhibits the positive and negative control experiments involving strains that carry either a *zwf* point mutation or the *eda-edd-zwf* deletion in competition with their nonmutant counterparts. In each case the *fhuA*-bearing strain ($T5^R$) is listed first; the open symbols correspond to competition in limiting glucose, the closed symbols to limiting gluconate. (Neither the *edd* locus nor the *eda* locus plays a role in glucose metabolism.) When cells are competing for glucose, the selection coefficient against the deletion-bearing strain is 0.028 ± 0.002/h (open squares), which is almost the same as the 0.022 ± 0.002/h selection against a point mutation-bearing strain (open circles). Growth of comparable strains in gluconate yields the expected result that the deletion is strongly disfavored in gluconate (because of the included *edd* locus) but that the *zwf* point mutation is selectively neutral in gluconate.

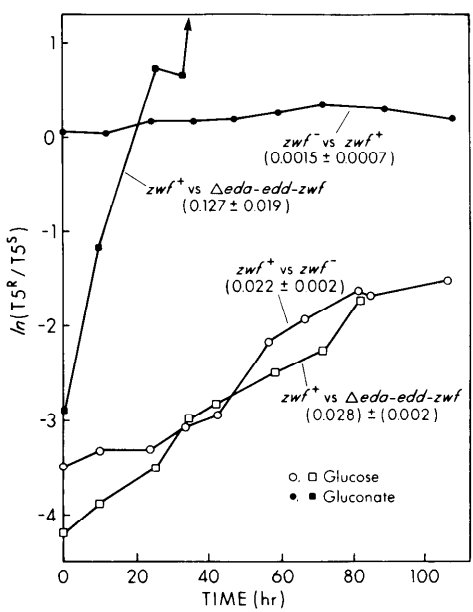


FIG. 1.—Controls showing competition in glucose (open symbols) or gluconate (solid symbols) between a zwf^+(G6PD$^+$) strain and one carrying a *zwf* point mutation (G6PD$^-$; circles) or one carrying an *eda-edd-zwf* deletion (squares). In all cases the strain listed first carries the *fhuA* (bacteriophage $T5^R$) marker. Specific strains are: ○ = DD1337 vs. DD1310, ● = DD1157 vs. DD1146, □ = DD1288 vs. DD1196, ■ = DD1288 vs. DD1196. Generation times were set at 1.69 ± 0.09 h^{-1} (standard error among experiments).

Experiments

Experimental results pertaining to six naturally occurring alleles are summarized in table 3. The experiments come in pairs that differ according to which allozyme-bearing strain is marked with *fhuA*. Were both the *fhuA* and the associated allozyme alleles neutral, none of the slopes beyond the 5% expected from type I error would be significant. Were the *fhuA* marker neutral but the associated allozymes nonneutral, the slopes in the paired experiments would be expected to be the negatives of one another, and the difference between the slopes would estimate twice the selective effect of the allozyme difference. However, in the genetic background in table 3, which originated as a clone from a chemostat population of DD725 that had been maintained for 250 h in limiting glucose, the *fhuA* marker is not selectively neutral as it is in many other strains. This marker effect is indicated by the consistently positive slope, significant in many cases, which corresponds to selection in favor of *fhuA*. Nevertheless, since the experiments are paired, the difference between corresponding slopes still estimates twice the selective difference due to the associated *zwf* alleles, because in one combination the selective effect of an allozyme will help the progress of the selectively favored marker, but in the other case it will hinder it. Indeed, a paired comparison is more powerful than one based on individual slopes because it has more degrees of freedom. The relevant *t*-tests are shown in the last column of table 3, where it can be observed that none of the comparisons indicates significant divergence of allele frequency in any pair of experiments. Consequently, the data support the view that the naturally occurring *zwf* allozymes are selectively neutral

Table 3
Selection Observed among Strains Differing in G6PD Allozymes

Strains	Zwf Phenotype		Selection Coefficient ± SE	F	Paired Comparison t[a]
	$T5^S$	$T5^R$			
DD1298 vs. DD1331.....	2	3	.0015 ± .0010	2.5 (1,8)	.09 (16)
DD1301 vs. DD1330.....	3	2	.0016 ± .0009	3.0 (1,8)	
DD1298 vs. DD1332.....	2	4a	.0010 ± .0011	.9 (1,8)	.21 (16)
DD1302 vs. DD1330.....	4a	2	.0013 ± .0006	4.4 (1,8)	
DD1298 vs. DD1334.....	2	5a	.0026 ± .0014	3.3 (1,7)	.39 (15)
DD1306 vs. DD1330.....	5a	2	.0021 ± .0005	14.5 (1.8)**	
DD1298 vs. DD1334.....	2	5a	.0052 ± .0012	20.0 (1,8)***	.69 (16)
DD1306 vs. DD1330.....	5a	2	.0033 ± .0016	4.3 (1,8)	
DD1304 vs. DD1332.....	4b	4a	.0041 ± .0006	46.4 (1,8)****	1.28 (16)
DD1302 vs. DD1333.....	4a	4b	.0057 ± .0011	25.1 (1,8)****	
DD1304 vs. DD1357.....	4b	4a	.0006 ± .0008	.6 (1,8)	1.73 (16)
DD1302 vs. DD1358.....	4a	4b	.0031 ± .0011	7.5 (1.8)*	
DD1308 vs. DD1334.....	5b	5a	.0025 ± .0006	20.6 (1,8)****	.48 (16)
DD1306 vs. DD1335.....	5a	5b	.0019 ± .0011	2.9 (1,8)	

NOTE.—Results of paired competition experiments involving several naturally occurring *zwf* alleles in reciprocal combinations with *fhuA* (bacteriophage T5 resistance). Although *fhuA* is nonneutral in this chemostat-evolved genetic background, the difference between slopes of the paired experiments estimates two times the selective effect of the *zwf* alleles in the strains. The relevant *t*-tests for the difference of slopes are shown in the last column. Numbers in parentheses are df.

[a] $P = .05$ when $t(16) = 1.746$.

* $P < .05$.

** $P < .01$.

*** $P < .005$.

**** $P < .001$.

at the level of resolution of chemostat competition experiments, which we have estimated elsewhere (Dykhuizen and Hartl 1983*a*) to be a selection coefficient of 0.002/h.

To determine whether the selective neutrality of *zwf* allozymes might occur in spite of differences in elementary kinetic parameters of the enzymes, assays of apparent K_m and V_{max} were carried out in extracts of the isogenic strains used in the competition experiments. In two experiments carried out at different times, no heterogeneity in K_m or V_{max} among the allozymes was detected (data not shown). This supports the view that the allozymes are functionally equivalent under the assay conditions, but differences smaller than about 20% in the kinetic parameters of the enzymes would escape detection in these types of experiments.

Discussion

The situation regarding the selective effects of *zwf* allozymes is somewhat different from that found previously for *gnd* and *pgi* allozymes. In the case of *gnd,* the 6-phosphogluconate dehydrogenase enzyme is used in both glucose and gluconate metabolism. Although no significant selective effects were observed with glucose limitation, selection could be observed in some cases in limiting gluconate, depending on the particular allele and, in one case, the genetic background (Dykhuizen and Hartl 1980; Hartl and Dykhuizen 1981). In the *pgi* case, the phosphoglucose isomerase enzyme is used in both glucose and fructose metabolism, and, whereas selective neutrality was found with glucose, one allele was associated with a selective effect in fructose (Dykhuizen and Hartl 1983*a*). With *zwf,* however, the enzyme is used only in the metabolism of glucose or substrates that are converted to glucose-6-phosphate, so the opportunity to examine the strains when cultured in a variety of essentially different substrates does not arise.

The situation with *gnd* and *pgi* has been described in terms of a potential for selection among the alleles—a potential that is unexpressed when glucose is the limiting substrate but that is expressed with one or more alternative substrates. We cannot exclude other types of selection potentials among the *zwf* alleles. Although the alleles are selectively neutral under the conditions we have examined, there remains the formal speculative possibility that selective differences could perhaps be revealed under alternative environmental conditions, which represents a possibility that can never be ruled out completely. Selection always involves an interaction between genotype and environment, and the findings with *gnd* and *pgi* provide ample precedent for neutrality in some conditions but selection in others. While such potentials for selection are a formal possibility, we have previously argued that competition for a single limiting substrate in chemostats may represent a selection intensity operating on the relevant enzymes that is greater than the selection intensity typically encountered in heterogeneous, fluctuating natural environments. Particularly for enzymes like G6PD, which utilize a single substrate and play a singular metabolic role, the failure to find selection in chemostats supports the hypothesis that the allozymes are neutral or nearly neutral under natural conditions.

Whittam et al. (1983) and Selander and Whittam (1983) have studied 1,705 natural isolates of *Escherichia coli* by means of protein electrophoresis of 12 loci. They identified 302 distinct electrophoretic types but found that the *E. coli* strains could be classified into three groups of strains that were more similar to one

another than to members of other such groups. These groups seem to represent common ancestry, perhaps remnants of past major selective changes in the *E. coli* population. Such clustering raises the possibility that some of the genetic variation among groups is not altogether selectively neutral but is genetically coadapted with other genes in the genetic background. Some of the strains used in the present study were also included in this extensive electrophoretic study, and Selander and Ochman (personal communication) have informed us that strain RM20 belongs to group II of the strains whereas RM73C, RM77C, and RM66A belong to group I. Since *E. coli* K12 belongs to group I, the strain DD1306 (and DD1334) represents a case in which the *zwf* allele from one group has been transferred into the genetic background of another. Our failure to find selective differences associated with these strains suggests that, whatever amount of genetic coadaptation there may be within the groups, some of the genetic variation within the groups is selectively neutral and has arisen during the expansion of the original, presumably selectively favored, clone.

Acknowledgment

This work was supported by National Institutes of Health grant GM30201.

LITERATURE CITED

ATWOOD, K. C., L. K. SCHNEIDER, and F. J. RYAN. 1951. Periodic selection in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **37:**146–155.

BACHMANN, B. J. 1983. Linkage map of *Escherichia coli* K-12, edition 7. Microbiol. Rev. **47:**180–230.

COHEN, S. S. 1951. Utilization of gluconate and glucose in growing and virus-infected *Escherichia coli*. Nature **168:**746.

DYKHUIZEN, D. E., and D. L. HARTL. 1980. Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics **96:**801–817.

———. 1983*a*. Functional effects of PGI allozymes in *Escherichia coli*. Genetics **105:**1–18.

———. 1983*b*. Selection in chemostats. Microbiol. Rev. **47:**150–168.

FRAENKEL, D. G. 1968. Selection of *Escherichia coli* mutants lacking glucose-6-phosphate dehydrogenase or gluconate-6-phosphate dehydrogenase. J. Bacteriol. **95:**1267–1271.

FRAENKEL, D. G., and S. BANERJEE. 1971. A mutation increasing the amount of a constitutive enzyme in *Escherichia coli*, glucose-6-phosphate dehydrogenase. J. Mol. Biol. **56:**183–194.

FRAENKEL, D. G., and S. R. LEVISOHN. 1967. Glucose and gluconate metabolism in an *Escherichia coli* mutant lacking phosphoglucose isomerase. J. Bacteriol. **93:**1571–1578.

HARTL, D. L., and D. E. DYKHUIZEN. 1981. Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **78:**6344–6348.

JOHNSON, R., A. KRASNA, and D. RITTENBERG. 1973. ^{18}O studies on the oxidative and nonoxidative pentose phosphate pathways in wild-type and mutant *Escherichia coli* cells. Biochemistry **12:**1969–1977.

KATZ, J., and R. ROGNSTAD. 1967. The labeling of pentose phosphate from glucose-C^{14} and estimation of the rates of transaldolase, transketolase, the contribution of the pentose cycle, and ribose phosphate synthesis. Biochemistry **6:**2227–2247.

MILKMAN, R. 1973. Electrophoretic variation in *Escherichia coli* from natural sources. Science **182:**1024–1026.

MODEL, P., and D. RITTENBERG. 1967. Measurements of the activity of the hexose monophosphate pathway of glucose metabolism with the use of [^{18}O] glucose. Biochemistry **6**:69–80.

ORTHNER, C. L., and L. I. PIZER. 1974. An evaluation of regulation of the hexose monophosphate shunt in *Escherichia coli*. J. Biol. Chem. **249**:3750–3755.

SELANDER, R. K., and B. R. LEVIN. 1980. Genetic diversity and structure in *Escherichia coli* populations. Science **210**:545–547.

SELANDER, R. K., and T. S. WHITTAM. 1983. Protein polymorphism and the genetic structure of populations. Pp. 89–114 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

SNEDECOR, G. W., and W. G. COCHRAN. 1967. Statistical methods. 6th ed. Iowa State University Press, Ames.

WANG, C. H., I. STERN, C. M. GILMOUR, S. KLUNGSOYR, D. J. REED, J. J. BIALY, B. E. CHRISTENSEN, and V. H. CHELDELIN. 1958. Comparative study of glucose catabolism by the radiorespirometric method. J. Bacteriol. **76**:207–216.

WHITTAM, T. S., H. OCHMAN, and R. K. SELANDER. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **80**:1751–1755.

MASATOSHI NEI, reviewing editor

Received July 1, 1983; revision received August 20, 1983.

Directed Evolution of Cellobiose Utilization in *Escherichia coli* K12[1]

Maja Kricker and Barry G. Hall
University of Connecticut

The cellobiose catabolic system of *Escherichia coli* K12 is being used to study the role of cryptic genes in evolution of new functions. *Escherichia coli* does not use β-glucoside sugars; however, mutations in several loci can activate the cryptic *bgl* operon and permit growth on the β-glucoside sugars arbutin and salicin. Such Bgl$^+$ mutants do not use cellobiose, which is the most common β-glucoside in nature. We have isolated a Cel$^+$ (cellobiose-utilizing) mutant from a Bgl$^+$ mutant of *E. coli* K12. The Cel$^+$ mutant grows well on cellobiose, arbutin, and salicin. Genes for utilization of these β-glucosides are located at 37.8 min on the *E. coli* map. The genes of the *bgl* operon are not involved in cellobiose utilization. Introduction of a deletion covering *bgl* does not affect the ability to utilize cellobiose, arbutin, or salicin, indicating that the new Cel$^+$ genes provide all three functions. Spontaneous cellobiose negative mutants also become arbutin and salicin negative. Analysis of β-glucoside positive revertants of these mutants indicates that there are separate loci for utilization of each of the β-glucoside sugars. The genes are closely linked and may be activated from a single locus. A fourth gene at an unknown location increases the growth rate on cellobiose. The *cel* genes constitute a second cryptic system for β-glucoside utilization in *E. coli* K12.

Introduction

Bacterial systems frequently provide the most explicit models for studying mechanisms that organisms use to evolve new metabolic functions. They can be manipulated genetically, produce numerous generations in a short time span, and allow comparisons of evolved mutant strains with unevolved antecedents. There is substantial evidence from these systems that one mechanism for the evolution of new functions is the acquisition of mutations altering regulation and catalytic activity of enzymes in existing pathways. Examples include the amidase system of *Pseudomonas putida* (reviewed in Clarke 1978), the utilization of unusual pentoses and pentitols by *Klebsiella pneumoniae* (reviewed in Mortlock 1981), acrobic metabolism of 1,2-propanediol by *E. coli* via the fucose pathway (Cocks et al.

1. Key words: *Escherichia coli,* cellobiose, cryptic genes, β-glucosides.

Address for correspondence and reprints: Barry G. Hall, Biological Sciences Group, U-44, University of Connecticut, Storrs, Connecticut 06268.

Mol. Biol. Evol. 1(2):171–182. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0006$02.00

1974), and the evolved β-galactosidase (EBG) system in *Escherichia coli* (reviewed in Hall 1983).

However, organisms need not modify existing systems to increase physiologic versatility. Reactivation of silent sequences by mutation provides a second, though not the only other, means of acquiring new metabolic capabilities. Cryptic genes have been detected in a multitude of bacterial species, including a variety of amino acid synthetic pathways in *Neisseria gonorrhoeae* (Juni and Heym 1980), *Salmonella* sp. (Lederberg 1947), *Pasteurella pestis* (Englesberg and Ingram 1957), and genes for citrate catabolism in *E. coli* (Hall 1982) and for β-glucoside catabolism in *E. coli* (Prasad and Schaefler 1974). Cryptic sequences are not limited only to prokaryotes. Investigators have also identified mutations activating cryptic genes coding for sucrose catabolic functions in *Saccharomyces cerevisiae* (Carlson et al. 1981). These discoveries serve not only to confuse taxonomists (Buchanan and Gibbons 1974) but also to identify potentially important elements of adaptive change. We have chosen the well-characterized β-glucoside system of *E. coli* to examine the role of cryptic genes in adaptive evolution.

Wild-type *E. coli* cannot utilize any β-glucoside sugars as carbon or energy sources. Mutants of *E. coli* have been isolated which utilize the aryl β-glucosides arbutin and salicin, but these do not utilize the most common β-glucoside in nature, the dissacharide cellobiose (Schaefler 1967). Other members of the family Enterobactericiae do utilize cellobiose (Schaefler and Malamy 1969), suggesting that it might be possible for *E. coli* exposed to appropriate selective pressures to evolve this capacity. Arbutin and salicin positive (Bgl^+) *E. coli* mutants express the *bgl* operon which is normally cryptic in wild-type strains (Schaefler 1967). The *bgl* operon is located at 83 min on the *E. coli* map (Bachmann 1983) and cotransduces with the *ilv* gene cluster at 84 min (Schaefler and Maas 1967). Spontaneous Bgl^+ mutations that occur at a high frequency (10^{-5} to 10^{-7}) are caused by insertions of the mobile DNA sequences IS1 or IS5 at a regulatory site *bglR* (Reynolds et al. 1981). The insertions cause inducible expression of two structural genes, *bglB and bglC*. The gene *bglC* codes for a transport protein that is a β-glucoside specific enzyme II of the phosphoenolpyruvate-dependent phosphotransferase system (Fox and Wilson 1968); *bglB* specifies phosphoglucosidase B, a hydrolase for phosphorylated β-glucosides. A third gene, *bglS*, codes for a positive regulator of the operon (Prasad and Schaefler 1974).

Mutations outside of the *bgl* operon also allow expression of the *bgl* enzymes. *bglY* mutations map at 28 min and cotransduce with the *trp* operon (DeFez and DeFelice 1981). Mutations in *gyrA* located at 46 min or *gyrB* located at 82 min also activate the *bgl* genes (Dinardo et al. 1982). All of these mutants are inducible for expression of the *bgl* operon.

Material and Methods

Culture Media and Growth Conditions

Cultures were grown at 37 C with aeration. Minimal media (Hall and Hartl 1974) contained .2% of the appropriate sugar as a carbon source. When required, amino acids were added to a concentration of 100 μg/ml, purines were added to a concentration of 40 μg/ml, spectinomycin or ampicillin was added to a concentration of 50 μg/ml. Solid media contained 1.5% agar.

Indicator Media

All mutants were isolated on either tetrazolium or MacConkey solid indicator medium. Salicin and cellobiose tetrazolium plates were prepared according to Miller (1972, p. 54). Salicin, arbutin, and cellobiose MacConkey media were prepared from MacConkey agar base according to instructions provided by Difco. Colonies unable to ferment the added sugar are white on this medium, whereas fermenting colonies are pink or red.

Matings and Transductions

Matings and transductions were carried out according to Miller (1972). Transductions were mediated by P1 cm clr100 (Rosner 1972).

In Vivo Alkaline Phosphatase Determinations

Alkaline phosphatase phenotypes were detected by plate assay (Willsky et al. 1973). Colonies grown on glucose minimal solid medium and glucose solid medium limited for phosphate (5×10^{-5} M) and containing .05 M tris (hydroxymethyl) amino-methane (Tris) buffer pH 8.0, were overlaid with 1 mM *p*-nitrophenyl phosphate in .5 M tris pH 8.0 and 1% agar. Constitutive colonies turn a bright yellow color in 10 min on both types of media. Inducible colonies remain white on glucose minimal medium but turn yellow in a few minutes on limited phosphate medium. Alkaline phosphatase negative mutants remain white on both types of media.

Growth Rates

Cultures were grown in cellobiose or salicin minimal medium to a density of approximately 5×10^8 cells/ml, washed once, then distributed to flasks containing β-glucoside or glucose minimal medium so that the density was approximately 1×10^8 cells/ml. Turbidity was monitored in a Gilford spectrophotometer at 600 nm for at least two doublings of each culture. The growth rates are reported as the first-order growth rate constant, calculated by the slope of the least squares fit of $\ln(A_{600})$ versus time (hours), from a minimum of eight points for each of at least three independent growth rate determinations, for each strain on each sugar.

Reversion Frequencies

Strains were grown overnight in glucose minimal medium. Cultures were washed once, concentrated 10-fold, and plated at appropriate dilutions on arbutin, salicin, cellobiose, and glucose minimal plates. Reversion frequencies are reported as the number of revertants per cell plated.

Results

Isolation of Cellobiose Utilizing Mutants

Cellobiose- (Cel^+) utilizing mutants were selected in three steps from the wild-type parent 1011A (table 1). In the first step, strain 1011A was streaked on salicin tetrazolium indicator plates (Lederberg 1948), then incubated at 30 C until salicin fermenting papillae appeared on the parent colonies. A salicin positive, arbutin positive (Sal^+, Arb^+) strain, MK1, was purified from one of these papillae. In the second step, MK1 was streaked onto cellobiose tetrazolium plates, and the same procedure was used to obtain papillae. The papillae formed pink colonies

Table 1
***Escherichia coli* Strains**

Strains	Relevant Genotype and Phenotype	Source
1011	*HfrC spc lacZ* (deletion W4680) *ebgA* (deletion 11)	Hartl and Hall 1974
1011A	*HfrC spc lacZ* (deletion W4680) *ebgA* (deletion 11) *amp* Spontaneous mutant of 1011	This study
MK1	*HfrC spc lacZ* (deletion W4680) *ebgA* (deletion 11) *amp bglR*$^+$ Spontaneous mutant of 1011A	This study
MK2	*HfrC spc lacZ* (deletion W4680) *ebgA* (deletion 11) *amp bglR*$^+$ Cel$^+$ Spontaneous mutant of MK1	This study
CSH75	*F*$^-$ *rpsL ilv ara leu lacY purE trp his argG malA xyl mtl metA or B thi proC gal*	Cold Spring Harbor Laboratory
SJ30	*F*$^-$ *rpsL ilv ara leu purE trp his argG metA or B thi malA xyl mtl gal lacZ* (deletion W4680) *lacY* from 1011A × CSH75 conjugation	This study
SJ30A1	*bglR*$^+$ Spontaneous mutant of SJ30	This study
AE341	*F′ lac* (deletion x74) *tna::Tn10 bglY supE*$^+$	A. Wright
JF201	*lac* (deletion x74) *bgl-pho* (deletion 201) *ara gyrA thi*	A. Wright
JF201T	*bgl-pho* (deletion 201) *tna::Tn10* transductant of JF201 (donor AE341)	This study
MKI201	*ilv bgl-pho* (deletion 201) EMS induced Ilv$^-$ mutant of JF201	This study
MK9	*F*$^-$ *ilv rpsL trp his argG metA or B ara leu lacZ* (deletion W4680) *lacY* Cel$^+$ (from MK2 × SJ30 conjugation)	This study
MK91	*bgl-pho* (deletion 201) Cel$^+$ Transductant of MK9 (donor JF201)	This study
MK93	*ilv bgl-pho* (deletion 201) Cel$^+$ EMS induced Ilv$^-$ mutant of MK91	This study
MK94	*bglR*$^+$ Cel$^+$ Transductant of MK93 (donor JF201)	This study
MK79	*bgl-pho* (deletion 201) Transductant of SJ30 (donor JF201)	This study
MK797	*bgl-pho* (deletion 201) Cel$^+$ Transductant of MK79 (donor MK9)	This study
CSH62	*HfrH thi*	Miller 1972
CSH62T	*HfrH thi bgl-pho* (deletion 201) *tna::Tn10* Transductant of CSH62 (donor JF201T)	This study
CSH62TC	*HfrH thi bgl-pho* (deletion 201) *tna::Tn10* Cel$^+$ Transductant of CSH62T (donor MK91)	This study
GMS343	*F*$^-$ *aroD6 argE3 lacY1 galK2 man-4*	B. Bachmann
(CGSC 5496)	*mtl-1 rpsL700 tsx-29? supE44?* λ-	*E. coli* Genetic Stock Center Yale University

NOTE.—EMS = ethyl methane sulfonate.

when restreaked to MacConkey cellobiose plates (Miller 1972). In the third step, one pink colony was restreaked on cellobiose minimal medium. A single large colony was reisolated and designated MK2; MK2 was red on MacConkey cellobiose medium. Both MK1 and MK2 were, like the parent 1011A, *HfrC*, Lac$^-$, as well as ampicillin and spectinomycin resistant, and were therefore not contaminants. Strain 1011A did not utilize any β-glucosides. Both MK1 and MK2 utilized arbutin and salicin, but only MK2 utilized cellobiose.

Genetic Analysis

To determine whether cellobiose utilization arose as a consequence of mutations in the *bgl* operon, the locus conferring the Cel$^+$ phenotype was mapped. If the *cel* mutation was in the *bgl* operon it should cotransduce with the *ilv* locus at 84 min. When MK1 was used as a donor to transduce MKI201 (*ilv bgl* deletion) to Ilv$^+$, 25% of the Ilv$^+$ transductants were arbutin and salicin positive. The maximum length of a chromosomal region carried by a P1 phage transducing particle is equivalent to 2 min on the *Escherichia coli* genetic map (Bachmann 1983). The distance between *ilv* and the β-glucoside markers is therefore 1 min by the mapping function of Wu (1966), confirming that MK1 is a bglR$^+$ mutant. When a Cel$^+$ strain was used as a donor in the same type of experiment, none of the Ilv$^+$ transductants were Cel$^+$, indicating that the *cel* locus lay outside of the *bgl* operon.

Conjugation experiments indicated that the *cel* marker was located in the region near *aroD* and *manA* (data not shown). To locate the *cel* mutation more precisely, CSH62TC (*bgl* deletion Cel$^+$) was used as a donor to transduce GMS343 to Aro$^+$, and 500 Aro$^+$ colonies were scored for mannose and cellobiose utilization (table 2). Among the 17 Aro$^+$ Man$^+$ cotransductants, only three were Cel$^+$, showing that *cel* does not lie between *aroD* and *manA*. Similarly, the fact that only three of the 144 Aro$^+$ Cel$^+$ cotransductants were Man$^+$ shows that *manA* does not lie between *cel* and *aroD*. The gene order is thus *cel, aroD, manA*. The cotransduction frequency of *aroD* with *cel* is .294; thus the distance between *cel* and *aroD* is .67 min by the mapping function of Wu (1966). The *cel* locus is therefore at 37.8 min on the *E. coli* map.

The observation that the gene for cellobiose utilization lies on the opposite side of the map from the *bgl* operon does not preclude the possibility that the *bgl* operon provides a necessary function for cellobiose metabolism. To explore this possibility we introduced a deletion of the entire *bgl* operon into a cellobiose

Table 2
Mapping of the *cel* Locus by Transduction

Donor	Recipient	Selected Marker	Recombinant Class	Number of Recombinants
CSH62TC	GMS343	Aro$^+$	Cel$^+$ Man$^-$	144
bgl Cel$^+$	*aroD manA*		Cel$^+$ Man$^+$	3
			Cel$^-$ Man$^+$	14
			Cel Man	339
			Total	500
Cotransduction		Aro$^+$Cel$^+$	Aro$^+$Man$^+$	
Frequency		.294	.034	

positive strain and asked if the strain could still utilize cellobiose. The donor, JF201, carries a deletion covering the *bgl* operon and the neighboring *phoS* and *phoT* genes and therefore exhibits an alkaline phosphatase constitutive phenotype (Reynolds 1983). When the deletion was introduced into strain MK9 (Cel$^+$) by cotransduction with *ilv,* 34 out of 100 transductants were alkaline phosphatase constitutive and therefore deleted for the *bgl* operon, yet all alkaline phosphatase constitutive transductants continued to grow on cellobiose. In contrast, when the same deletion was introduced into the *ilv*$^-$ *bglR*$^+$ mutant SJ30A1, 38% of the Ilv$^+$ transductants were alkaline phosphatase constitutive, and none of these utilized arbutin and salicin, demonstrating that the deletion did in fact eliminate the Bgl$^+$ phenotype. These results showed that the *bgl* operon did not provide any function necessary for cellobiose catabolism. To determine if the *bgl* operon could contribute to faster growth on cellobiose, though the operon was not required for cellobiose utilization, the growth rate of the Bgl$^+$ Cel$^+$ strain MK94 was compared with that of the *bgl* deletion Cel$^+$ strain MK91 on cellobiose (table 3). MK91 grew more slowly than MK94 on cellobiose. However, MK91 grew more slowly than MK94 on glucose as well, indicating that the difference in growth rates is not specific to cellobiose. When normalized to the growth rates on glucose, there was no significant difference in growth rates between the two strains. The *bgl* operon does not, therefore, contribute in any way to growth on cellobiose.

Both MK91 and MK94 were derived from MK9, an exconjugant of a mating between MK2 and a Cel$^-$ recipient (SJ30). MK91 and MK94 thus received a number of loci from MK2 which were not linked to the *cel* locus at 38 min. However, MK797 was constructed by a transductional cross between a Cel$^+$ donor and a *bgl* deletion Cel$^-$ recipient (MK79) and therefore received only those markers that were linked to the *cel* locus. Both MK91 and MK94 were red on MacConkey cellobiose medium, but Cel$^+$ transductants, such as MK797, were pink on this medium, suggesting that the transductants ferment the sugar poorly. To determine if there was a growth rate difference between "red" and "pink" strains, the growth rates on cellobiose and glucose of the "red" strains MK91 and MK94 were compared with that of the "pink" strain MK797. The normalized growth rates of both "red" strains were significantly higher than that of MK797, suggesting that a second mutation enhances the growth rate on cellobiose. The second mutation was present in strain MK2, was transferred to the recipient during the construction of MK9 by conjugation, and was therefore present in the direct descendants of MK9. Because it was *not* transferred to the recipient during the

Table 3
Growth Rates of Representative Strains on Cellobiose and Salicin

Strain	Cellobiose	Salicin	Glucose	Cellobiose/ Glucose	Salicin/ Glucose
MK94	.240 ± .029	.626 ± .024	.837 ± .049	.287 ± .035	.748 ± .047
MK91	.197 ± .006	.139 ± .001	.616 ± .033	.321 ± .020	.227 ± .014
SJ30A1	ND	.508 ± .013	.914 ± .010	...	.556 ± .014
MK797	.126 ± .006	ND	.611 ± .014	.198 ± .009	...

NOTE.—Mean growth rates on the substances above each column are reported as the first-order growth rate constant ±95% confidence limits. Normalized growth rates are the means of the ratios of growth rates on cellobiose or salicin to growth rates on glucose. ND = not determined.

construction of MK797 by transduction, the second mutation (enhancer) is apparently unlinked to the *cel* locus.

The *cel* System Specifies Multiple Functions

The ability of strain MK2 to utilize arbutin and salicin did not depend on expression of the *bgl* operon. MK91, and other Cel$^+$ strains deleted for the *bgl* operon, remained fully capable of growth on arbutin and salicin. The Cel system thus provides all of the transport and hydrolase functions required for the metabolism of three β-glucoside sugars, arbutin, salicin, and cellobiose.

If MK94 expresses both the *bgl* operon and the *cel* locus allowing salicin utilization, then MK94 should grow more rapidly on salicin than either the *bgl*R$^+$ Cel$^-$ mutant SJ30A1 or the *bgl* deletion Cel$^+$ strain MK91. Table 3 (last column) shows that when growth rates on salicin were normalized to glucose, MK94 grew more rapidly on salicin than either MK91 or SJ30A1. The combined growth rates of MK91 and SJ30A1 equaled the growth rate of MK94 on salicin, indicating that the systems are additive and function independently.

Clearly, the *cel* system specifies three identifiable functions: utilization of arbutin, salicin, and cellobiose. We did not observe any segregation of the cellobiose, arbutin, and salicin phenotypes in a variety of transductions, indicating that either these functions are tightly linked or a single genetic element specifies all three functions. To determine whether these functions were specified by a single genetic element or by three separate genetic elements, spontaneous cellobiose negative mutants were selected. MK91 was streaked onto a cellobiose MacConkey plate and incubated until white papillae appeared on the colonies. Fifteen papillae were picked from different regions of the plate to avoid isolation of siblings. All 15 isolates were cellobiose negative and were also arbutin and salicin negative. The mutants carried the same five auxotrophic markers as MK91, indicating they were indeed derivatives of the parent. In some strains the alkaline phosphatase structural gene was no longer expressed due to unidentified mutations. In those cases presence of the *bgl* deletion was verified by transducing out the deletion and showing that the recipient had gained the alkaline phosphatase constitutive phenotype.

Isolation of single-step mutants which failed to utilize any of the three β-glucosides suggested that all three functions were specified by a single genetic element. To explore this further, spontaneous cellobiose, arbutin, and salicin positive revertants were selected from various β-glucoside negative mutants. One of these β-glucoside negative mutants, MK912, was streaked on arbutin, salicin, or cellobiose MacConkey plates, and papillae were re-isolated using the same procedure as for isolation of cellobiose negative mutants (fig. 1). Twenty-five revertants were isolated following selection on cellobiose. Although all of these revertants used cellobiose, none utilized arbutin or salicin. From one such revertant, MK9123, several arbutin revertants were isolated. One class of arbutin revertants grew on all three β-glucosides. A second class of revertants grew on arbutin and cellobiose but not salicin. From one Arb$^+$ Cel$^+$ revertant (MK912301), several salicin-utilizing revertants were isolated, and these revertants (e.g., MK9123011) used all three β-glucosides. From the mutant MK912 which utilized no β-glucosides, revertants utilizing all three β-glucosides could be isolated in three sequential steps, indicating that three separate genes are responsible for utilization of the three β-glucoside sugars, arbutin, salicin, and cellobiose.

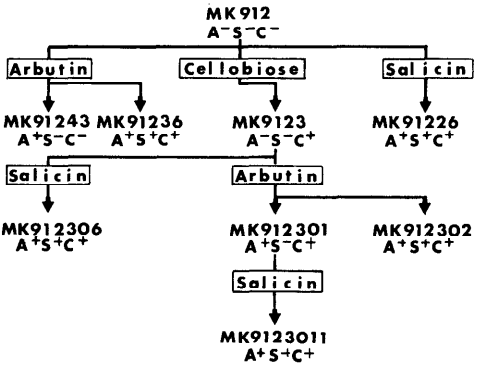


FIG. 1.—Pedigree of spontaneous β-glucoside positive revertants. Sugars used for selection are enclosed in boxes. Phenotypes are shown below strain names. A = arbutin, S = salicin, C = cellobiose. In each case the strain shown is representative of a class containing multiple independent isolates.

When 10 single-step arbutin revertants were isolated directly from MK912, two classes were obtained. The first class (MK91243) utilized only arbutin, but not cellobiose or salicin, consistent with a separate arbutin gene. The second class (MK91236) utilized all three β-glucosides. Twenty single-step salicin revertants were similarly isolated from MK912, and all of these belonged to a single class which utilized all three β-glucosides. All functions were regained in a single step, suggesting coordinate control at a single locus of the separate β-glucoside genes. Nine second-step salicin revertants were isolated from MK9123, and again, the same class was obtained, the class which utilized all three β-glucosides. All revertants which utilize salicin utilize the other β-glucosides whether they are selected in one step (MK91226), two steps (MK912302), or three steps (MK9123011), suggesting that salicin utilization is associated with a locus for activation of the *cel* genes.

Revertants with all three β-glucoside functions could be selected in either a single step (MK91226) or by three sequential steps (MK9123011). If three separate mutational events were required for reversion of the three functions, the reversion frequencies would probably be very low. The spontaneous rate of point mutations in *E. coli* is typically 10^{-8} to 10^{-9}; however, the highest mutation frequency observed for the single *bglR* mutation is 10^{-5} (Reynolds et al. 1981). If the *cel* mutations occur at a similar frequency, the expected frequency of three concerted *cel* mutations is 10^{-15}. The frequency of revertants as determined by direct plating on minimal medium (see Material and Methods) was approximately 10^{-8} whether these were selected on arbutin, salicin, or cellobiose. The overwhelming majority of these utilized all three β-glucosides, further indicating that the *cel* cluster can be activated by a single mutation. The reversion frequencies do not depend on *rec* function. Introduction of a *recA*⁻ allele into MK912 did not alter reversion frequencies for any of the β-glucosides, indicating these reversions result from either a point mutation or genetic rearrangement involving illegitimate recombination mechanisms.

Discussion

The observation that introduction of a deletion of the *bgl* operon into Cel⁺ strains does not prevent growth on any of the three β-glucosides shows that the

cel system expresses all three functions independently of the *bgl* operon. Analysis of mutants and revertants indicates that functions for utilization of arbutin, salicin, and cellobiose are specified by separate genetic elements. We designate the arbutin gene *celA*, the salicin gene *celS*, and the cellobiose gene *celC*. Expression of all three functions can be lost by a single mutational step and can be regained at a frequency consistent with activation at a single locus. Nonetheless, components can be activated individually. Because of this, the genetic state of specific *cel* alleles in any strain cannot be deduced from the phenotype, and gene designations do not indicate allele assignments. These designations are not meant to imply enzyme functions, which are unknown, but to indicate individual genetic components identified by revertant analysis. The lack of recombination between these genes in transductional crosses suggests the genes are tightly linked. The *cel* genes thus comprise a second cryptic gene cluster for β-glucoside utilization which can be activated in *Escherichia coli* K12.

Single-step mutants from MK1 were pink on MacConkey cellobiose medium. Two sequential selections were required to obtain a mutant which fermented cellobiose well and exhibited a red phenotype on MacConkey cellobiose medium. Cel$^+$ transductants also exhibit the pink phenotype and have a decreased growth rate on cellobiose compared with "red" strains, indicating that single-step Cel$^+$ mutants from MK1 and the Cel$^+$ transductants probably express only one of two genetic elements involved in cellobiose utilization. One element, within the *cel* gene cluster, appears to be sufficient for growth on cellobiose. The second element, unlinked to the *cel* locus and therefore not present in Cel$^+$ transductants, apparently enhances cellobiose utilization. Since Cel$^+$ transductants exhibit the pink phenotype, there is still another *cel* gene, which we call *celM* (for *M*odifier of cellobiose-utilizing activity), at an unknown location, separated from the *cel* cluster by a distance too large for cotransduction.

Two lines of evidence indicate that the *cel* mutations do not alter the substrate specificity of a functioning system but do decryptify a silent gene cluster. First, three separable functions can be ascribed to the *cel* genes, and expression of these functions does not require separate mutational events. It is improbable that genes of another pathway could gain all of the new functions simultaneously. Second, other members of the family Enterobactericiae display the same phenotypes as *E. coli* Cel$^+$ mutants. There is wide variation in the utilization of β-glucosides among Enterobactericiae (Schaefler and Malamy 1974). Wild-type *Klebsiella* sp. ferment the full range of β-glucosides, arbutin, salicin, and cellobiose. *Citrobacter* sp. ferments cellobiose well and aromatic β-glucosides poorly or not at all, while wild-type *Proteus vulgaris* metabolizes aromatic β-glucosides but not cellobiose. Wild-type *Salmonella* sp. are similar to *E. coli* in that they do not use any of the β-glucosides. *Salmonella* mutants have been isolated which grow on cellobiose but not on either arbutin or salicin (Schaefler and Scheinken 1968). Arbutin-utilizing *Salmonella* mutants have been obtained in a second step from cellobiose positive strains (Schaefler and Malamy 1974). From β-glucoside negative mutants we isolated revertants that grew only on cellobiose and that may be genetically equivalent to cellobiose positive *Citrobacter* and *Salmonella* strains. Similarly, we isolated second-step revertants which grow on cellobiose and arbutin, paralleling *Salmonella* arbutin- and cellobiose-utilizing mutants. Most of our strains express the same range of functions found in *Klebsiella* sp. These observations suggest that the *cel* genes of *E. coli* did not evolve independently, from an existing

pathway, but are silent homologues of β-glucoside genes found in other Enterobactericiae.

Two functions common to the *cel* cluster and the *bgl* operon are the utilization of arbutin and salicin. The *cel* gene cluster is located nearly 180° from the *bgl* operon on the *E. coli* map. A number of functionally related gene pairs involved in central metabolism lie approximately either 90° or 180° apart in *E. coli*, and it has been proposed that the ancestral chromosome may have undergone two sequential duplications (Zipkas and Riley 1975; Riley and Anilionis 1978). These observations raise the question of whether the two β-glucoside systems arose independently or from an ancient genome duplication.

The maintenance of two cryptic β-glucoside systems in *E. coli* would appear to have no selective advantage. Yet either can be activated by a single mutation, indicating that the structural genes are intact though not expressed in both operons. Both β-glucoside systems might be activated when aryl β-glucosides are a primary carbon source. Mutants expressing both the *cel* and *bgl* genes grow faster on salicin than mutants expressing a single system. Cel$^+$ revertants expressing all β-glucoside utilization functions were isolated in a single step on media containing arbutin or salicin, suggesting that selection for expression of both cryptic systems could occur simultaneously or sequentially on the same carbon source.

There are a number of reports suggesting that unneeded functions reduce fitness (Zamenhoff and Eichorn 1967; Dykhuizen 1978). We obtained spontaneous β-glucoside negative mutants easily by growing Cel$^+$ mutants on rich (MacConkey) medium, indicating that there was considerable selective advantage for cryptification of the *cel* genes when other carbon sources are available. The *bgl* operon is inducible, and there should be little advantage to cryptification of genes under transcriptional regulation. However, it has been suggested that since toxic cyanogenic β-glucosides are found in nature, transcriptional control may be inadequate to protect the cell against poisonous compounds if these compounds are inducers (Reynolds et al. 1981). Thus, there are at least two plausible natural conditions that might favor silencing of these systems—the presence of multiple carbon sources or the presence of toxic β-glucosides. Hall et al. (1983) have recently presented a systematic discussion of the role of cryptic genes in microbial evolution. They have suggested that repeated cryptification and decryptification of genes may be a means of long-term regulation of rarely utilized functions and may account for the retention of those genes in the population. Mathematical analyses of this question support their model (Hall et al. 1983; Li 1984). In particular, Li (1984) has shown that a gene that spends an average of 25,000 generations in the cryptic state and only 200 generations in the decryptified state will not be lost owing to irreversible mutational inactivation, even when the rate of decryptification is only 10^{-7} per generation. Numerous examples of cryptic systems in microorganisms capable of reactivation under selective conditions (Lederberg 1947; Englesberg and Ingram 1957; Juni and Heym 1980; Hall 1982) support such a model. During this study we have isolated *E. coli* strains that exhibit every phenotype for β-glucoside utilization found in naturally occurring populations of Enterobactericiae. The *cel* genes provide a specific system for testing the hypothesis that there is selection for mutations which maintain silent genes for pathways that are unneeded, or even deleterious, under the most prevalent natural conditions.

Acknowledgments

We are grateful to Andrew Wright for the gift of several strains and for many helpful discussions. Barry G. Hall is supported by Research Career Development Award 1 K04 AI00366 from the National Institute of Allergy and Infectious Diseases of the U.S. Public Health Service. This work was supported by grants from the National Institutes of Health and the University of Connecticut Research Foundation.

LITERATURE CITED

BACHMANN, B. J. 1983. Linkage map of *Escherichia coli* K12, edition 7. Microbiol. Rev. **47**:180–230.

BUCHANAN, R. E., and N. E. GIBBONS, eds. 1974. Bergey's manual of determinative bacteriology. 8th ed. Williams & Wilkins, Baltimore.

CARLSON, M., B. C. OSMOND, and D. BOTSTEIN. 1981. Genetic evidence for a silent *SUC* gene in yeast. Genetics **94**:41–54.

CLARKE, P. H. 1978. Experiments in microbial evolution. Pp. 137–218 *in* L. N. ORNSTON and J. R. SOKATCH, eds. The bacteria. Academic Press, New York.

COCKS, G. T., AGUILAR, J., and LIN, E. C. C. 1974. Evolution of L-1,2-propanediol catabolism in *Escherichia coli* by recruitment of enzymes for L-fucose and L-lactate metabolism. J. Bacteriol. **118**:83–88.

DEFEZ, R., and M. DEFELICE. 1981. Cryptic operon for β-glucoside metabolism in *Escherichia coli* K12: genetic evidence for a regulatory protein. Genetics **97**:11–25.

DINARDO, S., K. A. VOELKEL, R. STERNGLANZ, A. E. REYNOLDS, and A. WRIGHT. 1982. *Escherichia coli* DNA topoisomerase I mutants have compensatory mutations in DNA gyrase genes. Cell **31**:43–51.

DYKHUIZEN, D. 1978. Selection for tryptophan auxotrophy of *Escherichia coli* in glucose limited chemostats as a test of the energy conservation hypothesis. Evolution **32**:125–150.

ENGLESBERG, E., and L. INGRAM. 1957. Meiotrophic mutants of *Pasteurella pestis* and their use in the elucidation of nutritional requirements. Proc. Natl. Acad. Sci. USA **43**:369–372.

FOX, C. F., and G. WILSON. 1968. The role of a phosphoenolpyruvate-dependent kinase system in β-glucoside catabolism in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **59**:988–994.

HALL, B. G. 1982. Chromosomal mutation for citrate utilization in *Escherichia coli* K12. J. Bacteriol. **151**:269–273.

———. 1983. Evolution of new metabolic functions in laboratory organisms. Pp. 200–230 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

HALL, B. G., and D. L. HARTL. 1974. Regulation of newly evolved enzymes. I. Selection of a novel lactase regulated by lactose in *Escherichia coli*. Genetics **76**:391–400.

HALL, B. G., S. YOKOYAMA, and D. H. CALHOUN. 1983. Role of cryptic genes in microbial evolution. Mol. Biol. Evol. **1**:109–124.

HARTL, D. L., and B. G. HALL. 1974. Second naturally occurring β-galactosidase in *E. coli*. Nature **248**:152–153.

JUNI, E., and G. A. HEYM. 1980. Studies of some naturally occurring auxotrophs of *Neisseria gonorrhoeae*. J. Gen. Microbiol. **121**:85–92.

LEDERBERG, J. 1947. The nutrition of *Salmonella*. Arch. Biochem. **13**:287–290.

———. 1948. Detection of fermentative variants with tetrazolium. J. Bacteriol. **56**:695.

LI, W.-H. 1984. Retention of cryptic genes in microbial populations. Mol. Biol. Evol. **1**:212–218.

MILLER, J. H. 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, New York.

MORTLOCK, R. P. 1981. Regulatory mutations and the development of new metabolic pathways in bacteria. Evol. Biol. **14:**205–267.

PRASAD, I., and S. SCHAEFLER. 1974. Regulation of the β-glucoside system in *Escherichia coli* K12. J. Bacteriol. **120:**638–650.

REYNOLDS, A. E. 1983. Characterizations of mutations which activate the *bgl* operon. Ph.D. thesis. Tufts University School of Medicine.

REYNOLDS, A. E., Y. FELTON, and A. WRIGHT. 1981. Insertion of DNA activates the cryptic *bgl* operon in *E. coli* K12. Nature **293:**625–629.

RILEY, M., and A. ANILIONIS. 1978. Evolution of the bacterial genome. Annu. Rev. Microbiol. **32:**519–560.

ROSNER, J. L. 1972. Formation, induction, and curing of bacteriophage P1 lysogens. Virology **49:**679–689.

SCHAEFLER, S. 1967. Inducible system for the utilization of β-glucosides in *Escherichia coli.* I. Active transport and utilization of β-glucosides. J. Bacteriol. **93:**254–263.

SCHAEFLER, S., and A. J. MALAMY. 1969. Taxonomic investigations on expressed and cryptic phospho-β-glucosidases in Enterobactericiae. J. Bacteriol. **99:**422–433.

SCHAEFLER, S., and W. K. MAAS. 1967. Inducible system for the utilization of β-glucosides in *Escherichia coli.* II. Description of mutant types and genetic analysis. J. Bacteriol. **93:**264–272.

SCHAEFLER, S., and I. SCHEINKEN. 1968. β-glucoside permeases and phospho-β-glucosidases in *Aerobacter aerogenes:* relationship with cryptic phospho-β-glucosidases in Enterobactericiae. Proc. Nat. Acad. Sci. **59:**285–292.

WILLSKY, G. R., R. L. BENNETT, and M. H. MALAMY. 1973. Inorganic phosphate transport in *Escherichia coli:* involvement of two genes which play a role in alkaline phosphatase regulation. J. Bacteriol. **113:**529–539.

WU, T. T. 1966. A model for three point analysis of random general transduction. Genetics **54:**405–410.

ZAMENHOFF, S., and H. H. EICHORN. 1967. Studies of microbial evolution through loss of biosynthetic functions: establishment of defective mutants. Nature **216:**455–458.

ZIPKAS, D., and M. RILEY. 1975. Proposal concerning mechanism of evolution of the genome of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **72:**1354–1358.

WALTER M. FITCH, reviewing editor

Received August 19, 1983; revision received October 24, 1983.

Major Morphological Effects of a Regulatory Gene: *Pgm1-t* in Rainbow Trout[1]

Robb F. Leary, Fred W. Allendorf, and Kathy L. Knudsen
University of Montana

We have investigated the morphological effects of a genetic locus, *Pgm1-t*, that affects the expression of a phosphoglucomutase locus (*Pgm1*) in liver of rainbow trout (*Salmo gairdneri*). We have previously shown that embryos with liver *Pgm1* expression hatch earlier than those without liver *Pgm1* expression. We predicted that this difference in developmental rate should cause a reduction in meristic counts in the more rapidly developing fish with liver *Pgm1* expression. Eight meristic (countable) characters in nine full-sib groups segregating for the presence or absence of liver *Pgm1* expression are in agreement with this prediction. In eight of the nine families, there is a significant difference in the multivariate distribution of the eight meristic counts between full sibs with and without liver *Pgm1* expression. This separation in multivariate space is based on a tendency for lower meristic counts in fish with liver *Pgm1* expression. The magnitude of these morphological differences is similar to that between two subspecies of cutthroat trout (*Salmo clarki*) that show substantial genetic divergence at structural loci encoding enzymes (Nei's $D = 0.34$). These data support the view that small changes in the developmental process caused by genetic differences at regulatory genes can have large effects on morphology.

Introduction

Increasing emphasis is being placed on the potential importance of changes in regulatory genes to bring about evolutionary change. It has been proposed that such changes can result in large organismal effects by altering the rate and timing of developmental events (Britten and Davidson 1969; King and Wilson 1975; Wilson 1976). These alterations in early development are believed to result in significant changes in the life history characteristics and the morphology of individuals (Frazetta 1970; Gould 1980). These authors have suggested that changes in regulatory genes may be of more evolutionary importance than changes in structural genes. The evidence supporting the evolutionary importance of regulatory genes in eukaryotic organisms has been mainly indirect, based largely on differences in the rate of divergence of morphology and structural loci in closely

1. Key words: phosphoglucomutase, regulatory genes, morphology, development, rainbow trout.

Address for correspondence and reprints: Robb F. Leary, Department of Zoology, University of Montana, Missoula, Montana 59812.

Mol. Biol. Evol. 1(2):183–194. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0002$02.00

related organisms (King and Wilson 1975; Wilson 1976). However, direct evidence for the adaptive importance of regulatory genetic differences in eukaryotes is accumulating (see the recent review of MacIntyre [1982]).

We have been studying the organismal effects of a regulatory gene that affects the tissue-specific expression of a phosphoglucomutase (PGM; E.C. 2.7.5.1) locus in rainbow trout (*Salmo gairdneri*). Most rainbow trout have no product (PGM1) of the structural locus *Pgm1* in the liver. Allendorf et al. (1982) have described allelic variation at a regulatory locus, *Pgm1-t,* that affects the presence of PGM1 in the liver. Individuals homozygous for the common allele, *Pgm1-t(a/a)*, show little or no PGM1 enzyme in liver tissue. Heterozygotes, *Pgm1-t(a/b)*, show a greater than 100-fold increase in the amount of PGM1 enzyme in liver tissue. Homozygotes for the variant allele, *Pgm1-t(b/b)*, have approximately twice as much PGM1 in liver tissue as heterozygotes. The presence of PGM1 in the liver has been shown to be associated with increased developmental rate, increased developmental stability, and decreased age at sexual maturity in these fish (Allendorf et al. 1983; Leary et al. 1983). These effects would almost certainly influence the ability of individuals to survive and reproduce in natural situations.

In this paper, we present the results of our investigation of the morphological effects of the *Pgm1-t* locus. Individuals with liver PGM1 have faster developmental rates than their full sibs without liver PGM1 activity (Allendorf et al. 1983). Environmentally induced increases in the developmental rate of fishes, including rainbow trout, have been shown to generally result in a decrease in meristic counts (Gabriel 1944; Taning 1950; Barlow 1961; Garside 1966; MacCrimmon and Kwain 1969; Lindsey and Harrington 1972; Ali and Lindsey 1974; Kwain 1975; MacGregor and MacCrimmon 1977). Thus, we predicted that full sibs with liver PGM1 should tend to have lower meristic counts than their full sibs without liver PGM1. We have tested this prediction by comparing eight meristic traits of full sibs in nine families, segregating for the presence or absence of liver PGM1 activity.

Methods

We crossed 18 individuals of the Arlee strain of rainbow trout maintained by the Montana Department of Fish, Wildlife, and Parks (see Leary et al. [1983] for the history of this strain) to create nine full-sib families, segregating for the presence or absence of liver PGM1. The fertilized eggs were incubated and the progeny raised at 9 C at the Jocko River State Trout Hatchery, Arlee, Montana, of the Montana Department of Fish, Wildlife, and Parks. In six of these families, the male parent had liver PGM1 activity, *Pgm1-t(a/b)*, and the female parent did not, *Pgm1-t(a/a)*. The reciprocal cross was made in one family. Both the parents were heterozygous at *Pgm1-t* in the remaining two families. We could not compare the morphology of heterozygotes and homozygotes with liver PGM1 activity because these genotypes are electrophoretically indistinguishable.

The counts of the following eight meristic characters were taken on the first 25 fish identified in each family: rays in the anal fin, rays in the dorsal fin, rays in the pectoral fins, rays in the pelvic fins, gill rakers on the upper first branchial arches, gill rakers on the lower first branchial arches, mandibular pores, and vertebrae. The counts of the bilateral characters were taken on the left and right sides of each individual. The total count (left + right) was used in the statistical analyses of the data. Meristic counts were taken from as many as five additional

fish so that we had nearly equal sample sizes of both *Pgm1-t* phenotypes in each family.

All of the progeny in six families and half of the progeny in the other families were sacrificed 182 days after fertilization (approximately 150 days after hatching) and frozen for subsequent electrophoretic and meristic analyses. The remaining progeny were sampled again 394 days after fertilization. There are no significant differences in the distributions of any of the meristic traits between fish in the same family sampled at 182 and 394 days after fertilization (Wilcoxon two-sample test). We concluded that these meristic characters are determined before 182 days and combined these two samples. This conclusion is in agreement with previous studies that have shown that meristic counts in fishes are determined early in development, often before hatching (Barlow 1961; Eisler 1961; MacCrimmon and Kwain 1969; Lindsey and Harrington 1972; Ali and Lindsey 1974).

Sample preparation and electrophoresis in starch gels followed Utter et al. (1974) with the stains and buffer systems of Allendorf et al. (1977). Isozyme loci are designated with the nomenclature described in Allendorf et al. (1983). The following 19 enzymes encoding 42 loci were screened in the electrophoretic examination of rainbow and cutthroat trout: adenylate kinase (ADK; E.C. 2.7.4.8), alcohol dehydrogenase (ADH; E.C. 1.1.1.1), aspartate aminotransferase (AAT; E.C. 2.6.1.1), creatine kinase (CK; E.C. 2.7.3.2), esterase (EST; E.C. 3.1.1.1), glucosephosphate isomerase (GPI; E.C. 5.3.1.9), glyceraldehyde-3-phosphate dehydrogenase (GAP; E.C. 1.2.1.12), glycerol-3-phosphate dehydrogenase (G3P; E.C. 1.1.1.8), glycyl-leucine peptidase (GL; E.C. 3.4.11), isocitrate dehydrogenase (IDH; E.C. 1.1.1.42), lactate dehydrogenase (LDH; E.C. 1.1.1.27), lycyl-glycyl-glycine peptidase (LGG; E.C. 3.4.11), malate dehydrogenase (MDH; E.C. 1.1.1.37), malic enzyme (ME; E.C. 1.1.1.40), phosphoglucomutase (PGM; E.C. 2.7.5.1), 6-phosphogluconate dehydrogenase (6PG; E.C. 1.1.1.44), sorbitol dehydrogenase (SDH; E.C. 1.1.1.14), superoxide dismutase (SOD; E.C. 1.15.1.1), and xanthine dehydrogenase (XDH; E.C. 1.2.3.2).

Results and Discussion

Effects of *Pgm1-t* on Morphological Differences within Families

As predicted, individuals with liver PGM1 tend to have lower meristic counts than their full sibs without liver PGM1 (table 1). There are nine significant differences in the pairwise comparisons, and, in each case, the individuals with liver PGM1 have a lower meristic distribution than their full sibs without it (Wilcoxon two-sample test; table 1). In 43 of the 69 full-sib comparisons, the individuals with liver PGM1 have a lower mean meristic count than their full sibs without liver PGM1 (sign test, $\chi^2 = 4.19$; $P < .05$). Three of the T2 comparisons had the same meristic count for both PGM1 phenotypes.

The meristic differences between the liver PGM1 phenotypes appear to be especially evident for four of the eight meristic characters: rays in the anal fin, rays in the dorsal fin, rays in the pectoral fins, and gill rakers on the lower first branchial arches. There are seven significant differences among these characters. Furthermore, the mean meristic count of the full sibs with liver PGM1 is lower than those without liver PGM1 in 27 of the 34 pairwise comparisons for these characters ($\chi^2 = 11.77$; $P < .001$).

Meristic characters of fish are modifiable by the environment only during certain "critical periods" of development (Taning 1950; Lindsey 1954; Mac-

Table 1
Summary of Comparisons of Counts

Character	Means		Significant Differences*	
	$A > B$	$A < B$	$A > B$	$A < B$
Anal rays	6	1	1	0
Dorsal rays	7	2	2	0
Pectoral rays	6	3	2	0
Pelvic rays	4	4	0	0
Upper gill rakers	4	5	1	0
Lower gill rakers	8	1	2	0
Mandibular pores	4	5	1	0
Vertebrae	4	5	0	0
Total	43	26	9	0

NOTE.—At eight meristic characters between full sibs without (*A*) and with (*B*) liver PGM1 activity in nine families of rainbow trout.

*$P < .05$ (Wilcoxon two-sample test).

Table 2
Morphological Differences

Family	Sample Size		
	A	*B*	*D*
I1	18	12	1.47*
I3	13	17	1.31**
I6	15	15	1.58**
I7	13	13	.58
I10	16	14	1.32**
I12	13	13	1.94**
I13	15	15	1.92**
I14	14	16	2.80**
I15	13	13	1.40**
Mean	...	...	1.59

NOTE.—*D*, square root of Mahalanobis's distance; between full sibs without (*A*) and with (*B*) liver PGM1 activity in nine families of rainbow trout, based on discriminant analysis of eight meristic characters.

*$P < .05$ (Wilcoxon two-sample test).

**$P < .01$ (Wilcoxon two-sample test).

Crimmon and Kwain 1969; Fowler 1970; Lindsey and Harrington 1972; Ali and Lindsey 1974). The number of vertebrae in rainbow trout raised at 9 C is not affected by light after the formation of eye pigment (about 17 days after fertilization), but anal and dorsal fin rays are affected after this time (MacCrimmon and Kwain 1969). Liver organogenesis in rainbow trout at this temperature occurs at approximately 15 days after fertilization (Ballard 1973). We, therefore, did not expect the number of vertebrae to be correlated with the presence or absence of liver PGM1. We have not been able to determine from the literature the critical periods for any of the other characters.

Multivariate comparison of all full sibs with and without liver PGM1 using discriminant analysis (Pimentel 1979) provides a measure of morphological difference at all eight characters simultaneously. In eight of the nine families there is a significant difference of the distribution in discriminant space between full

sibs with and without liver PGM1 (Wilcoxon two-sample test; table 2; fig. 1). In all families where a significant difference is observed, the individuals with liver PGM1 are distinguished from their full sibs without liver PGM1 by lower overall meristic counts.

In contrast to the large morphological effect of the *Pgm1-t* regulatory locus, we did not observe any evidence that structural variation at isozyme loci has a detectable effect on these meristic characters. There are 10 loci that are segregating in at least three of these nine and in five other families that were raised with them (*Est1; Idh2; Idh3,4; Ldh4; Mdh3,4; Pgm2; Sdh; Sod1*). Out of a total of 313 pairwise comparisons, there are only 15 differences at the level $P < .05$ between the meristic distributions of genotypes at these loci. This is no more than one would expect by chance alone. Furthermore, these significant differences are not grouped into any one locus or meristic character.

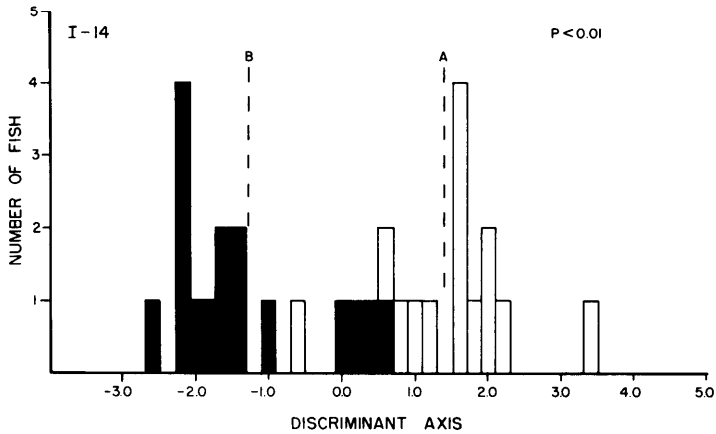
Comparative Magnitude of Morphological and Isozymic Differences between Species

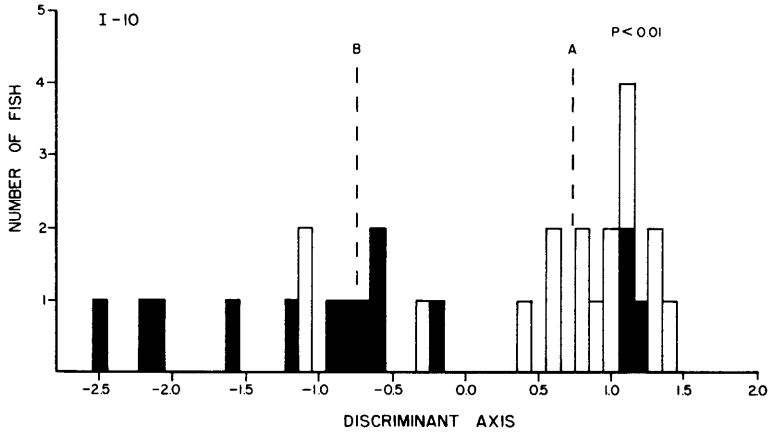
We have placed the magnitude of the morphological differences between sibs with and without liver PGM1 into an evolutionary perspective by comparing it with the amount of isozymic and morphological divergence detected between rainbow trout and a closely related species. The cutthroat trout (*Salmo clarki*) is a polytypic species that has been divided into several subspecies (Loudenslager and Gall 1980). We have estimated morphological divergence at the same eight meristic characters and isozymic divergence at 42 loci encoding enzymes between rainbow trout, west slope cutthroat trout (*S. c. lewisi*), and Yellowstone cutthroat trout (*S. c. bouvieri*). We used the McBride Lake strain of Yellowstone cutthroat trout maintained by the Montana Department of Fish, Wildlife, and Parks (n, the number of fish analyzed, = 39), westslope cutthroat trout collected from O'Keefe Creek, Missoula County, Montana ($n = 51$), and the Arlee strain of rainbow trout ($n = 160$ for isozyme analysis and $n = 29$ for morphological analysis) for these comparisons.

The following 16 loci are monomorphic for the same allele in our samples of these three taxa: *Aat2; Adh; Adk; Ck3; Gap3; Gl2; Gpi2; G3p2; Ldh1,2,3,5; Mdh1,2; 6Pg;* and *Xdh*. Eleven loci with fixed interspecific differences are presented in table 3. The allele frequencies for those 15 loci that showed intraspecific variation are given in table 4. The alleles at each locus are designated by their relative mobility to the common allele at the homologous locus in rainbow trout. Because of their ancient tetraploid ancestry (Ohno 1974), many loci in salmonid fishes are still functionally duplicated; that is, the two loci show no evidence of structural or regulatory divergence. Residual tetrasomic inheritance has been observed at some of these loci (May et al. 1982; Allendorf and Thorgaard, accepted). The duplicated loci, *Aat3,4, Idh3,4, Mdh1,2, Mdh3,4,* and *Me1,2,* were treated as single tetrasomic loci for the calculation of allele frequencies.

Isozymic divergence between these taxa was estimated using Nei's standard genetic distance, D (Nei 1975) (table 5). Morphological divergence was estimated as in the full-sib comparisons, using the distance between group centroids in discriminant space, which is the square root of Mahalanobis's distance, D^2 (Sneath and Sokal 1973, p. 405) (table 5; fig. 2).

The average morphological distance between the groups of full sibs with and without liver PGM1 (1.59; table 2) is nearly identical with that observed for the





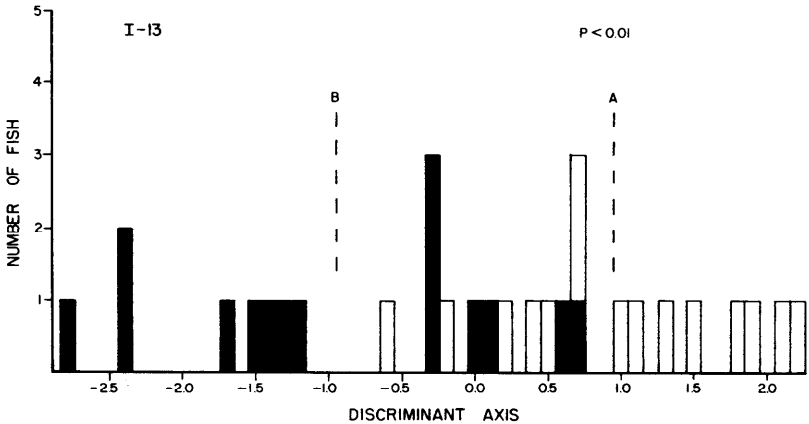


FIG. 1.—Discriminant analysis of eight meristic characters comparing three families of full sibs with (shaded) and without (unshaded) liver PGM1. *A* = centroid of fish without liver PGM1; *B* = centroid of fish with liver PGM1.

Table 3
Relative Mobilities of 11 Diagnostic Loci

Loci	Alleles		
	Rainbow	Westslope	Yellowstone
Aat1	100	200	165
Ck2	100	84	84
Gll	100	100	101
Gpi3	100	92	100
Idh1	100	100	−75
Lgg	100	100	135
Me1,2	100	88	100
Me3	100	100	84
Me4	100	100	110
Pgm1	100	100	Null

NOTE.—Differentiate rainbow trout, westslope cutthroat trout, and Yellowstone cutthroat trout.

same characters between the Yellowstone and westslope cutthroat trout (1.50; table 5). Although these two taxa are recognized as subspecies, the amount of genetic divergence between them is equal to or greater than that observed between most congeneric species of fishes (Utter et al. 1973; Avise 1974; Avise and Smith 1977; Buth and Burr 1978). Thus, the morphological effect of the *Pgm1-t* regulatory variation is comparable to the amount of morphological differentiation that can occur between taxa with substantial structural genetic divergence. These data directly support the view that morphological differences between taxa may be due to changes at one or a few regulatory loci with major effects rather than the gradual accumulation of genetic differences at many loci with small effects. This is further supported by the relatively large amount of morphological divergence observed between the rainbow and westslope cutthroat trout—but the comparatively small amount of structural genetic divergence (table 5).

Evolutionary Importance of Regulatory Genes

The morphological effects of the *Pgm1-t* locus are almost certainly a consequence of the accelerated developmental rate of individuals with liver PGM1, as similar effects can be produced by environmentally increasing the developmental rate of fishes (see Introduction for pertinent literature). We believe that the increased developmental rate of individuals with liver PGM1 is due to the large increase in the amount of PGM in the liver of these fish (Allendorf et al. 1982) and is, thus, a consequence of the variation at the *Pgm1-t* locus and not of variation at the chromosomal segment marked by this locus. The liver is responsible for the release of metabolic energy stored in the yolk in the form of glycogen, which is the major energy source of fishes during early development (Terner 1968; Boulekbache 1981). We suspect that the increase in liver PGM results in a more constant or efficient flux of energy production during development and thus increased developmental rate. This view is supported by preliminary data (Aronson and Allendorf, unpublished) that juvenile rainbow trout with liver PGM1 can access energy stored in the liver as glycogen more rapidly than their full sibs without liver PGM1.

The increased developmental rate of individuals with liver PGM1 not only influences their meristic counts but also increases their size and decreases their

Table 4
Allele Frequencies at 15 Isozyme Loci

Locus and Alleles	Allele Frequencies		
	Rainbow	Yellowstone	Westslope
Aat3,4:			
100	1.000	.575	.863
110	...	.200	...
90	...	.225	...
77	...	...	.137
Ck1:			
100	.912	1.000	1.000
76	.088	...	...
Gap4:			
100	1.000	1.000	.961
75	...	...	.039
Gpi1:			
100	1.000	1.000	.931
150	...	...	.069
G3p1:			
100	.991	1.000	1.000
140	.009	...	...
Idh2:			
100	.763	1.000	1.000
140	.237	...	...
Idh3,4:			
100	.705	.500	.417
114	.070	...	...
86	...	...	.500
71	.034	.500	...
40	.191	...	.083
Ldh4:			
100	.981	1.000	1.000
76	.019	...	...
Mdh3,4:			
100	.875	1.000	1.000
83	.125	...	...
Pgm2:			
100	.959	1.000	.922
90	.041	...	...
85	...	...	.078
Sdh:			
100	.931	1.000	...
200	.019	...	...
40	.050	...	1.000
Sod1:			
100	.772	1.000	1.000
152	.228	...	...

Note.—With intraspecific variation in Arlee rainbow trout, McBride Yellowstone cutthroat trout, or westslope cutthroat trout from O'Keefe Creek. The duplicated loci *Aat*, *Idh*, and *Mdh* are each treated as a single locus.

age at sexual maturity (Allendorf et al. 1982, 1983). These latter effects would certainly influence the fitness of individuals in natural populations. The timing of hatching and subsequent emergence from the gravel has been found to have important effects on the survival of salmonid fishes (Mason and Chapman 1965). Predation on salmonid fishes is known to be size specific (Parker 1971). Age at sexual maturity will affect the reproductive value of individuals.

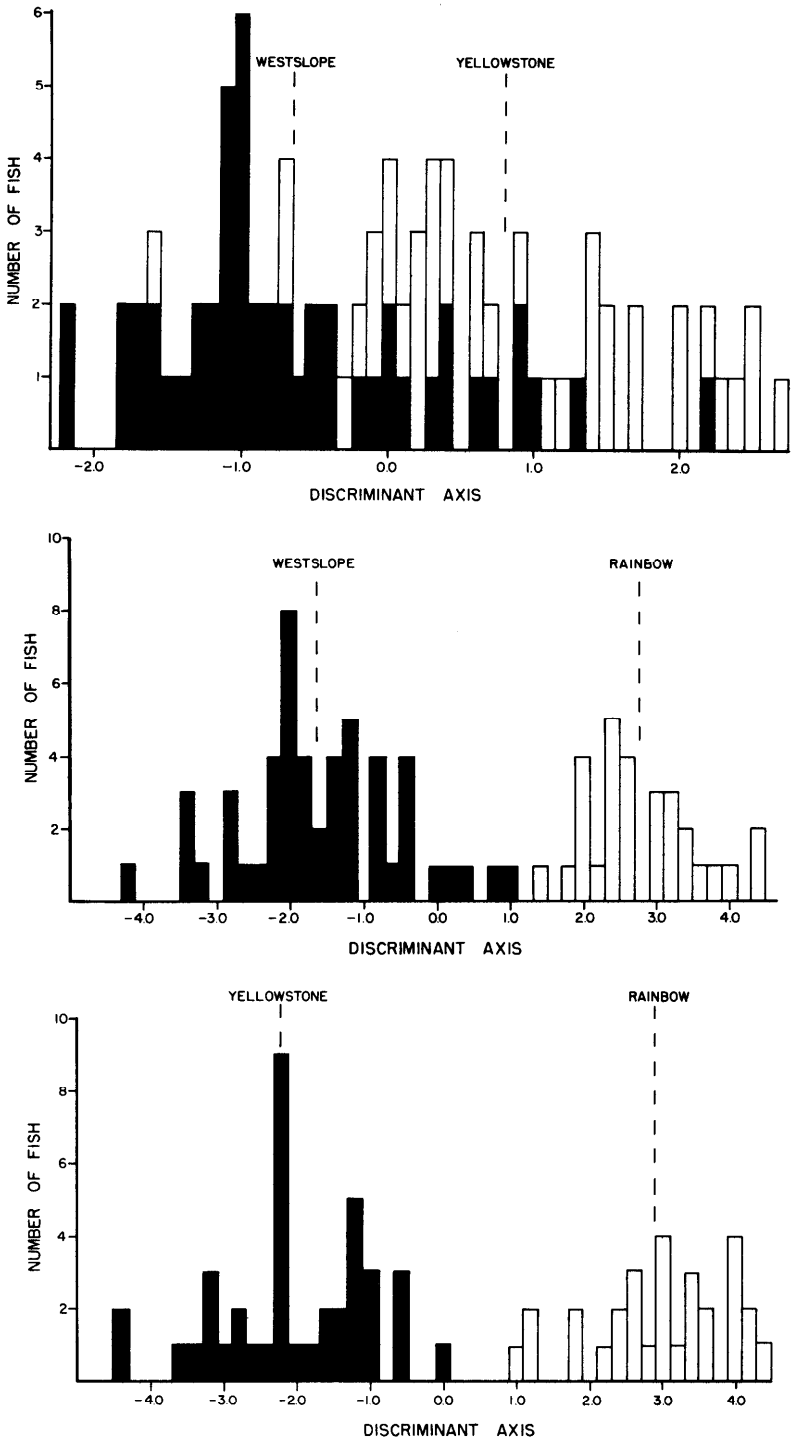


FIG. 2.—Discriminant analysis of rainbow trout, Yellowstone cutthroat trout, and westslope cutthroat trout using eight meristic characters. Vertical dashed lines represent centroids in discriminant space.

Table 5
Morphological and Genetic Divergences

	Rainbow	Yellowstone	Westslope
Rainbow.....	...	5.17	4.36
Yellowstone.....	.27	...	1.50
Westslope.....	.16	.34	...

NOTE.—Above diagonal, square root of Mahalanobis's D^2. Below diagonal, Nei's D. Among rainbow trout, Yellowstone cutthroat trout, and westslope cutthroat trout based on eight meristic characters and 42 enzyme loci.

Our studies with *Pgm1-t* demonstrate that a small increase in the developmental rate associated with a single regulatory locus has a variety of phenotypic effects; some of these are undoubtedly of adaptive importance. These findings support the view that small changes in the developmental process can have major effects upon the morphology and life history characteristics of individuals (Frazetta 1970; Gould 1977, 1980). In contrast, we found no evidence of any morphological effect of allelic variation at 10 structural loci encoding enzymes. These data thus support the view that such changes are more likely the result of changes in gene regulation than changes at structural loci.

Acknowledgments

This work was supported by National Science Foundation grants DEB-8004681, ISP-8011449, and BSR-8300039 to F.W.A. Appreciation is extended to the Montana Department of Fish, Wildlife, and Parks for financial support awarded to R.F.L., and especially to Jim Crepeau and Jack Boyce of the Jocko River State Trout Hatchery for their help. H. McPherson was instrumental in stimulating conversation.

LITERATURE CITED

ALI, M. Y., and C. C. LINDSEY. 1974. Heritable and temperature induced meristic variation in the medaka, *Oryzias latipes*. Can. J. Zool. **52:**959–976.

ALLENDORF, F. W., K. L. KNUDSEN, and R. F. LEARY. 1983. Adaptive significance of differences in the tissue specific expression of a phosphoglucomutase gene in rainbow trout. Proc. Nat. Acad. Sci. USA **80:**1397–1400.

ALLENDORF, F. W., K. L. KNUDSEN, and S. R. PHELPS. 1982. Identification of a gene regulating the tissue expression of a phosphoglucomutase locus in rainbow trout. Genetics **102:**259–268.

ALLENDORF, F. W., N. MITCHELL, N. RYMAN, and G. STAHL. 1977. Isozyme loci in brown trout (*Salmo trutta* L.): detection and interpretation from population data. Hereditas **86:**179–190.

ALLENDORF, F. W., and G. H. THORGAARD. Accepted. Polyploidy and the evolution of salmonid fishes. *In* B. TURNER, ed. Evolutionary genetics of fish. Plenum, New York.

AVISE, J. C. 1974. Systematic value of electrophoretic data. Syst. Zool. **23:**465–481.

AVISE, J. C., and M. H. SMITH. 1977. Gene frequency comparisons between sunfish (Centrachidae) populations at various stages of evolutionary divergence. Syst. Zool. **26:**319–335.

BALLARD, W. W. 1973. Normal embryonic stages for salmonid fishes, based on *Salmo gairdneri* and *Salvelinus fontinalis*. J. Exp. Zool. **184:**7–26.

BARLOW, G. W. 1961. Causes and significance of morphological variation in fishes. Syst. Zool. **10:**105–117.

BOULEKBACHE, H. 1981. Energy metabolism in fish development. Amer. Zool. **21:**377–389.

BRITTEN, R., and E. DAVIDSON. 1969. Gene regulation for higher cells: a theory. Science **165:**349–357.

BUTH, D. G., and B. M. BURR. 1978. Isozyme variability in the cyprinid genus *Campostoma*. Copeia **1978**(2):298–311.

EISLER, R. 1961. Effects of visible radiation on salmonid embryos and larvae. Growth **25:**281–346.

FOWLER, J. A. 1970. Control of vertebral number in teleosts: an embryological problem. Quart. Rev. Biol. **45:**148–167.

FRAZETTA, T. H. 1970. From hopeful monsters to bolyerine snakes. Amer. Natur. **104:**55–72.

GABRIEL, M. L. 1944. Factors affecting the number and form of vertebrae in *Fundulus heteroclitus*. J. Exp. Zool. **95:**105–147.

GARSIDE, E. T. 1966. Developmental rate and vertebral number in salmonids. J. Fisheries Res. Board Can. **23:**1537–1549.

GOULD, S. J. 1977. Ontogeny and phylogeny. Harvard University Press, Cambridge, Mass.

———. 1980. Is a new and general theory of evolution emerging? Paleobiology **6:**119–130.

KING, M. C., and A. C. WILSON. 1975. Evolution at two levels in humans and chimpanzees. Science **188:**107–116.

KWAIN, W. 1975. Embryonic development, early growth, and meristic variation in rainbow trout (*Salmo gairdneri*) exposed to combinations of light intensity and temperature. J. Fisheries Res. Board Can. **32:**397–402.

LEARY, R. F., F. W. ALLENDORF, and K. L. KNUDSEN. 1983. Developmental stability and enzyme heterozygosity in rainbow trout. Nature **301:**71–72.

LINDSEY, C. C. 1954. Temperature-controlled meristic variation in the paradise fish *Macropodus opercularis* (L.). Can. J. Zool. **30:**87–98.

LINDSEY, C. C., and R. W. HARRINGTON, JR. 1972. Extreme vertebral variation induced by temperature in a homozygous clone of the self-fertilizing cyprinodontid fish *Rivulus marmoratus*. Can. J. Zool. **50:**733–744.

LOUDENSLAGER, E. J., and G. A. E. GALL. 1980. Geographic patterns of protein variation and subspeciation in cutthroat trout, *Salmo clarki*. Syst. Zool. **29:**27–42.

MACCRIMMON, H. R., and W. KWAIN. 1969. Influence of light on early development and meristic characters in the rainbow trout, *Salmo gairdneri* Richardson. Can. J. Zool. **47:**631–637.

MACGREGOR, R. B., and H. R. MACCRIMMON. 1977. Evidence of genetic and environmental influences on meristic variation in the rainbow trout, *Salmo gairdneri* Richardson. Environmental Biol. Fishes **2:**25–33.

MACINTYRE, R. 1982. Regulatory genes and adaptation: past, present, and future. Evol. Biol. **15:**247–286.

MASON, J. C., and D. W. CHAPMAN. 1965. Significance of early emergence, environmental rearing capacity and behavioral ecology of juvenile coho salmon in stream channels. J. Fisheries Res. Board Can. **22:**173–190.

MAY, B., J. E. WRIGHT, and K. R. JOHNSON. 1982. Joint segregation of biochemical loci in salmonidae. III. Linkage associations in Salmonidae including data from rainbow trout (*Salmo gairdneri*). Biochem. Genet. **20:**29–40.

NEI, M. 1975. Molecular population genetics and evolution. Elsevier, New York.

OHNO, S. 1974. Protochordata, Cyclostamata, and Pisces. Animal cytogenetics. Gebruder-Borntraeger, Stuttgart.

PARKER, R. R. 1971. Size selective predation among juvenile salmonid fishes in a British Columbia inlet. J. Fisheries Res. Board Can. **28:**1503–1510.

PIMENTEL, R. A. 1979. Morphometrics, the multivariate analysis of biological data. Kendall/Hunt, Dubuque, Iowa.

SNEATH, P. H. A., and R. R. SOKAL. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

TANING, V. A. 1950. Influence of the environment on number of vertebrae in telostean fishes. Nature **165:**28.

TERNER, C. 1968. Studies of metabolism in embryonic development. III. Glucogenolysis and gluconeogenesis in trout embryos. Comp. Biochem. Physiol. **25:**989–1003.

UTTER, F. M., F. W. ALLENDORF, and H. O. HODGINS. 1973. Genetic variability and relationships in Pacific salmon and related trout based on protein variations. Syst. Zool. **22:**257–270.

UTTER, F. M., H. O. HODGINS, and F. W. ALLENDORF. 1974. Biochemical genetic studies of fishes: potentialities and limitations. Pp. 213–238 *in* D. C. MALINS and J. R. SARGENT, eds. Biochemical and biophysical perspectives in marine biology. Vol. **1.** Academic Press, New York.

WILSON, A. C. 1976. Gene regulation in evolution. Pp. 225–234 *in* F. J. AYALA, ed. Molecular evolution. Sinauer, Sunderland, Mass.

RICHARD K. KOEHN, reviewing editor

Received August 8, 1983; revision received October 3, 1983.

Concerted Evolution of the Immunoglobulin V_H Gene Family[1]

Takashi Gojobori and Masatoshi Nei
University of Texas at Houston

With the aim of understanding the concerted evolution of the immunoglobulin V_H multigene family, a phylogenetic tree for the DNA sequences of 16 mouse and five human germ line genes was constructed. This tree indicates that all genes in this family have undergone substantial evolutionary divergence. The most closely related genes so far identified in the mouse genome seem to have diverged about 6 million years (MY) ago, whereas the most distantly related genes diverged about 300 MY ago. This suggests that gene duplication caused by unequal crossing-over or gene conversion occurs very slowly in this gene family. The rate of occurrence of gene duplication in the V_H gene family has been estimated to be 5×10^{-7} per gene per year, which seems to be at least about 100 times lower than that for the rRNA gene family. This low rate of concerted evolution in the V_H gene family helps retain intergenic genetic variability that in turn contributes to antibody diversity. Because of accumulation of destructive mutations, however, about one-third of the mouse and human V_H genes seem to have become nonfunctional. Many of these pseudogenes have apparently originated recently, but some of them seem to have existed in the genome for more than 10 MY. The rate of nucleotide substitution for the complementarity-determining regions (CDRs) is as high as that of pseudogenes. This suggests that there is virtually no purifying selection operating in the CDRs and that germ line mutations are effectively used for generating antibody diversity.

Introduction

Immunoglobulins are composed of heavy and light chains each of which is composed of a variable region and a constant region. The variable region is responsible for antigen binding, whereas the constant region is responsible for effector function. The variable region consists of four framework regions (FRs) and three complementarity-determining regions (CDRs) (e.g., Tonegawa 1983). Both the heavy- and light-chain variable regions are controlled by multigene fam-

1. Key words: concerted evolution, immunoglobulin, variable region genes, unequal crossing-over, gene conversion, pseudogenes.

Address for correspondence and reprints: Dr. Masatoshi Nei, Center for Demographic and Population Genetics, P.O. Box 20334, Houston, Texas 77025.

Mol. Biol. Evol. 1(2):195–212. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0003$02.00

ilies. The heavy-chain variable region (V_H) family seems to have about 160 genes in the mouse (Kemp et al. 1981) and about 80 genes in the human (Rabbitts et al. 1980). The variable region gene families are apparently subject to concerted evolution, in which the turnover of genes occurs stochastically owing to unequal crossing-over or gene conversion (Smith et al. 1971; Hood et al. 1975; Ohta 1980; Arnheim 1983). Thus, the extent of variation of immunoglobulin genes seems to be controlled by unequal crossing-over and gene conversion as well as the classical evolutionary forces of mutation, selection, and genetic drift. However, the rate of turnover of genes is not known, though it is believed to be quite high (Hood et al. 1975; Honjo 1983; Ohta 1983*a*).

Recently a substantial number of V_H genes from mice and humans have been sequenced, and this gives us an opportunity to study the evolutionary history of these genes. Such a study will give us some idea about the pattern of concerted evolution of V_H genes as well as the mechanism of production of antibody diversity. Furthermore, when many genes exist as multiple copies in the genome, some of them are expected to become nonfunctional (Haldane 1933; Nei 1969). Indeed, recent studies of the V_H gene family indicate that it contains many pseudogenes (Bothwell et al. 1981; Givol et al. 1981; Rechavi et al. 1983). Nucleotide sequences of these pseudogenes will provide information on the pattern of nonfunctionalization of multiple genes (Li et al. 1981; Miyata and Yasunaga 1981). The purpose of this paper is to examine the pattern of concerted evolution in the V_H gene family and its significance for the production of immunoglobulin diversity.

Nucleotide Sequences Used

Surveying the literature, we collected nucleotide sequence data for 16 mouse and five human germ line V_H genes, which included the leader (signal peptide) region, the first three framework regions, and the first two CDRs (see table 1 and fig. 1). A large part of the third CDR (CDR3) is coded for by the D and J gene segments (e.g., Leder 1982), and thus CDR3 was not included in the present study. Seven of the 16 mouse genes, that is, VH3, VH6, VH23, VH102, VH145, VH186-1, and VH186-2, were identified by the same DNA probe (MP1 in table 1), and thus they are considered to be closely related. (In this paper we use the same gene notations as those of the original authors.) Gene VH6 seems to be a pseudogene, since it includes several termination codons caused by a single frameshift mutation (one nucleotide deletion from the twenty-second codon). Gene VH145 has a serine codon at the twenty-second position instead of a cysteine codon for a normal functional gene (see fig. 1). This mutation from cysteine to serine apparently disturbs the molecular structure of variable regions by impairing the disulfide bond between the cysteines at positions 22 and 98. This suggests that this gene is also a pseudogene, but there is no other clear-cut evidence. The rest of the genes detected by MP1 are all functional genes.

Another set of five mouse genes, that is, VH104, VH108A, VH108B, VH111, and VH105, was obtained by a different DNA probe (MP2 in table 1). VH104 and VH111 are apparently pseudogenes, since they contain a termination codon at the thirty-ninth and thirty-fourth positions, respectively. VH108B is also probably a pseudogene, since the initiation codon (ATG) has changed to an isoleucine codon (ATA). Of course, there might be some correction mechanism to make this gene functional (Givol et al. 1981). The remaining two genes (VH105 and VH108A) are functional. Table 1 includes four more mouse genes (i.e., VHSPT15, VH441,

Table 1
Immunoglobulin V_H Genes used for the Present Study

Gene	Codons (*n*)	DNA[a] Probe	Sub-group	Source
Mouse:				
VH3	116	MP1	II	Bothwell et al. (1981)
VH6(ψ)[b]	116[c]	MP1	II	Bothwell et al. (1981)
VH23	116	MP1	II	Bothwell et al. (1981)
VH102	116	MP1	II	Bothwell et al. (1981)
VH145(ψ)	116	MP1	II	Bothwell et al. (1981)
VH186-1	116	MP1	II	Bothwell et al. (1981)
VH186-2	116	MP1	II	Bothwell et al. (1981)
VH104(ψ)	116	MP2	II	Givol et al. (1981)
VH108A	116	MP2	II	Givol et al. (1981)
VH108B	116	MP2	II	Givol et al. (1981)
VH111(ψ)	116	MP2	II	Givol et al. (1981)
VH105	116	MP2	II	Cohen et al. (1982)
VHSPT15	118	MP3	III	Kim et al. (1981)
VH441	115	MP4	III	Ollo et al. (1981)
VHPJ14	115	MP5	?	Sakano et al. (1980)
VH101	115	MP6	?	Kataoka et al. (1982)
Human:				
VHHA2(ψ)	113	HP1	I	Rechavi et al. (1983)
VHHG3	116	HP1	I	Rechavi et al. (1983)
VHH11	116	HP2	III	Rechavi et al. (1982)
VHH16BR(ψ)	116	HP2	III	Rechavi et al. (1982)
VH26	116	HP2	III	Matthyssens and Rabbitts (1980)

[a] MP1, mouse S43 DNA (from MOPC104E cDNA). MP2, mouse MPC11 cDNA. MP3, mouse M167 DNA. MP4, mouse UPC10 DNA. MP5, mouse M141 DNA. MP6, mouse MC101 DNA. HP1, mouse VH104 DNA. HP2, mouse S107 DNA. MP3, mouse pμ/107 DNA.

[b] ψ, pseudogene.

[c] This contains a one-nucleotide deletion.

VHPJ14, and VH101). They were obtained by separate DNA probes. All of these genes are functional.

In the present study we used five human V_H genes, that is, VHHA2, VHHG3, VHH11, VHH16BR, and VH26. VHHA2 and VHH16BR contain termination codons at positions 26 and 29, respectively, which probably makes these genes nonfunctional. The other three are functional.

For studying the pattern of concerted evolution of V_H genes, it is desirable to use a random sample of genes. Obviously our sample of genes is not a random one, but since 16 of the 160 genes were sampled in the mouse, they should provide a rough idea about the concerted evolution of the mouse V_H gene family. It should be noted that the genes sampled by the same DNA probes are usually closely related, whereas those sampled by different probes are often remotely related. According to Kabat et al.'s (1979) classification, the genes identified by probe MP1 and MP2 produce polypeptides belonging to subgroup II, whereas the genes identified by MP3 and MP4 produce polypeptides belonging to subgroup III. It is not clear which group of polypeptides is produced by the genes identified by MP5 and MP6. In man only five genes are available; they are clearly insufficient for drawing any general conclusion about the pattern of concerted evolution. However, these genes are useful for getting a time scale of concerted evolution of the

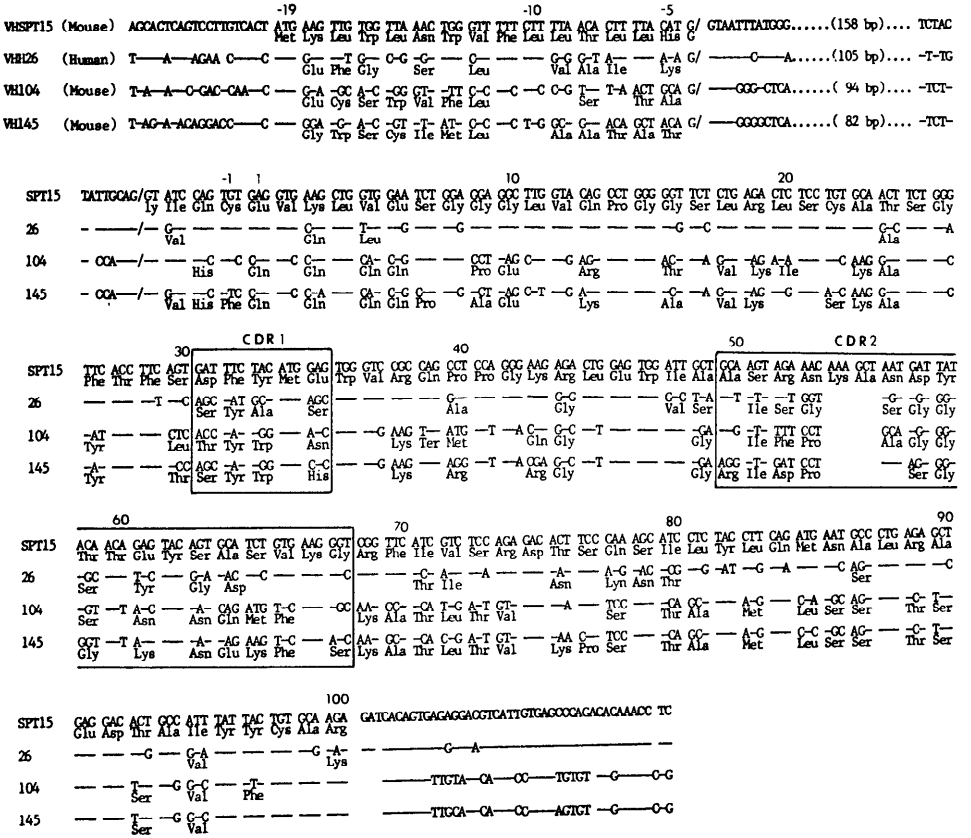


FIG. 1.—Four examples of the nucleotide and amino acid sequences for the V_H germ line genes used. VHSPT15, Mouse functional gene. VHH26, Human functional gene. VH104, Mouse pseudogene, which has a termination codon at the thirty-ninth codon position. VH145, Mouse pseudogene which has a serine codon rather than a cysteine codon at the twenty-second position. In the sequences for VHH26, VH104, and VH145, a dash stands for the same nucleotide or same amino acid as that for VHSPT15, and a blank indicates a gap. Codons are numbered in the positive direction from the beginning of the first framework region (FR1) and in the negative direction from the end of the leader region. The complementarity-determining regions (CDR1 and CDR2) are enclosed in boxes. FR2 lies between CDR1 and CDR2, while FR3 follows CDR2.

V_H genes in the mouse, as will be discussed later. The genes identified by probes HP1 and HP2 produce subgroup I and III polypeptides, respectively.

Phylogenetic Tree of V_H Genes

To see the evolutionary history of immunoglobulin V_H genes, we reconstructed a phylogenetic tree of the genes mentioned above, considering the evolutionary distances among them. In this case we used only the protein-coding regions, since alignment of DNA sequences for the noncoding regions was not easy. Alignment of the coding regions was made by using a slight modification of Needleman and Wunsch's (1970) method. In practice, alignment was easily established for all pairs of genes since only a few deletions or insertions existed in the coding regions. A few examples of the sequence alignments are presented in figure 1. The nucleotides involved in deletions or insertions were eliminated from the comparisons of DNA sequences.

The evolutionary distance between a pair of sequences was measured by the number of nucleotide substitutions per site. This number was estimated by Jukes and Cantor's (1969) formula

$$\delta = -(3/4)\ \ln[1-(4/3)\pi], \tag{1}$$

where π is the proportion of nucleotides that are different between the two sequences under consideration. The δ value was obtained for each of the three nucleotide positions of codons. For making the phylogenetic tree, however, the average of δ's for the first and second positions was used. We did not use third-position data, because the numbers of third-position changes for certain pairs of sequences were very large (see below), and in such a case as this Jukes and Cantor's (1969) method gives an underestimate of δ (Kimura 1981; Takahata and Kimura 1981; Gojobori et al. 1982). We also did not use the CDRs for this purpose because these regions are known to have a high rate of nucleotide substitution similar to that of pseudogenes (Nei 1983). The total number of codons used for tree making was therefore about 93 (see fig. 1).

The average δ values for the first and second nucleotide positions for all pairs of genes are presented in table 2. Using these values, we constructed two different phylogenetic trees, that is, one for all genes and the other for functional genes only. We did this because pseudogenes are known to evolve faster than functional genes (Kimura 1980; Proudfoot and Maniatis 1980; Li et al. 1981; Miyata and Yasunaga 1981), and this may distort the tree reconstructed. However, the topology of the functional genes within the all-gene tree was identical with that of the tree for functional genes only, though some branches of the all-gene tree were slightly longer than those of the latter tree, as expected (fig. 2). In the construction of the phylogenetic trees we used the unweighted pair-group method (UPGMA). It is known that, whatever the tree-making method used, a phylogenetic tree reconstructed is subject to large random errors, but UPGMA tends to give a better performance than other commonly used statistical methods in recovering the true tree (Tateno et al. 1982; Nei et al. 1983).

The phylogenetic tree obtained for all genes is presented in figure 2. In this figure the estimated number of nucleotide substitutions per site (δ) is related to evolutionary time by assuming that δ is proportional to evolutionary time and that the closest genes between man and mouse (VH441 and VHH26) diverged 80 MY ago, that is, the time when man and mouse diverged (see Nei 1975, p. 9). In the presence of concerted evolution, the latter assumption could be wrong, since these two genes may have separated well before the human-mouse divergence; then the time scale in figure 2 would give an underestimate of evolutionary time. In the present case, however, this assumption does not seem to be unreasonable, since the rates of nucleotide substitution for functional genes and pseudogenes obtained under this assumption are close to those for other gene families, as will be discussed later.

Figure 2 shows that some genes in the mouse or human genome are more closely related to each other than others are, as expected from the theory of concerted evolution. It is also noted that the genes identified by the same DNA probe are usually closely related. However, each gene has a surprisingly long history. The most closely related genes observed are VH186-2 and VH145, but even these genes seem to have diverged about 6 MY ago. All other genes diverged much earlier. The earliest separation of genes occurred between the VH101-VHPJ14 cluster and the other human and mouse genes, the time of separation being about

Table 2

Pairwise Evolutionary Distances (δ) for the Germ Line DNA Sequences of 16 Mouse and Five Human V_H Genes[a]

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. VH186-2	.01	.02	.03	.03	.03	.04	.12	.11	.19	.16	.15	.23	.23	.54	.39	.47	.43	.43	.58	.52
2. VH145.....		.03	.03	.04	.04	.06	.13	.12	.20	.17	.16	.24	.25	.55	.39	.48	.44	.43	.60	.54
3. VH186-1			.03	.04	.04	.05	.12	.10	.18	.15	.16	.22	.22	.53	.40	.49	.45	.43	.59	.53
4. VH6.....				.04	.04	.04	.11	.09	.17	.15	.15	.21	.21	.52	.38	.48	.44	.43	.57	.52
5. VH3.....					.02	.04	.11	.11	.18	.13	.13	.22	.22	.53	.39	.47	.43	.42	.54	.50
6. VH102.....						.04	.10	.10	.17	.14	.14	.21	.21	.53	.39	.47	.43	.41	.56	.50
7. VH23.....							.11	.10	.18	.16	.15	.21	.22	.51	.37	.45	.43	.41	.54	.49
8. VH108A								.06	.06	.11	.10	.24	.26	.48	.42	.48	.45	.43	.54	.49
9. VH105.....									.11	.08	.10	.23	.24	.50	.42	.49	.46	.44	.56	.51
10. VH108B										.15	.16	.29	.32	.51	.49	.54	.52	.50	.58	.53
11. VH104.....											.10	.26	.28	.51	.46	.53	.50	.48	.55	.53
12. VH111.....												.24	.26	.57	.45	.54	.50	.49	.56	.51
13. VHHG3.....													.09	.41	.34	.39	.35	.35	.50	.45
14. VHHAZ														.40	.37	.40	.35	.35	.56	.53
15. VHSPT15 ...															.21	.21	.16	.16	.53	.53
16. VH441.....																.18	.13	.14	.47	.44
17. VHH16BR ..																	.06	.08	.54	.52
18. VHH11																		.03	.49	.47
19. VHH26																			.48	.46
20. VHPJ14.....																				.06
21. VH101.....																				

[a] Genes are arranged in the same order as that in fig. 2.

300 MY ago. This corresponds approximately to the time when mammals and reptiles diverged (Nei [1975], p. 9). Separation of the VH441-VHSPT15 gene cluster and the gene clusters identified by probes MP1 and MP2 also seems to have occurred as early as 270 MY ago. Thus, many gene duplications occurred much earlier than the time of human-mouse divergence.

If we note that each of these gene duplications was probably mediated either by unequal crossing-over or by gene conversion, our results indicate that the concerted evolution of the V_H gene family occurs very slowly. (Gene conversion does not result in genuine gene duplication but has the same effect as gene duplication if the entire region rather than a part of a gene is "converted" by another gene.) The rate of *observable* gene duplication can be computed from the phylogenetic tree in figure 2 by tracing the evolutionary pathways of mouse genes. For example, genes VH186-2 and VH145 each experienced nine duplication events from the ancestral gene for all genes during a period of about 300 MY. Gene VH186-1 experienced eight duplication events. The average number of observable duplication events for all mouse genes is 5.9. Therefore, a mouse gene seems to have duplicated every 50 MY on average, or with a rate of 2×10^{-8} per gene per year. Of course, this rate of observable gene duplication is not equal to the rate of *occurrence of gene duplication,* since many duplicate genes that appeared in

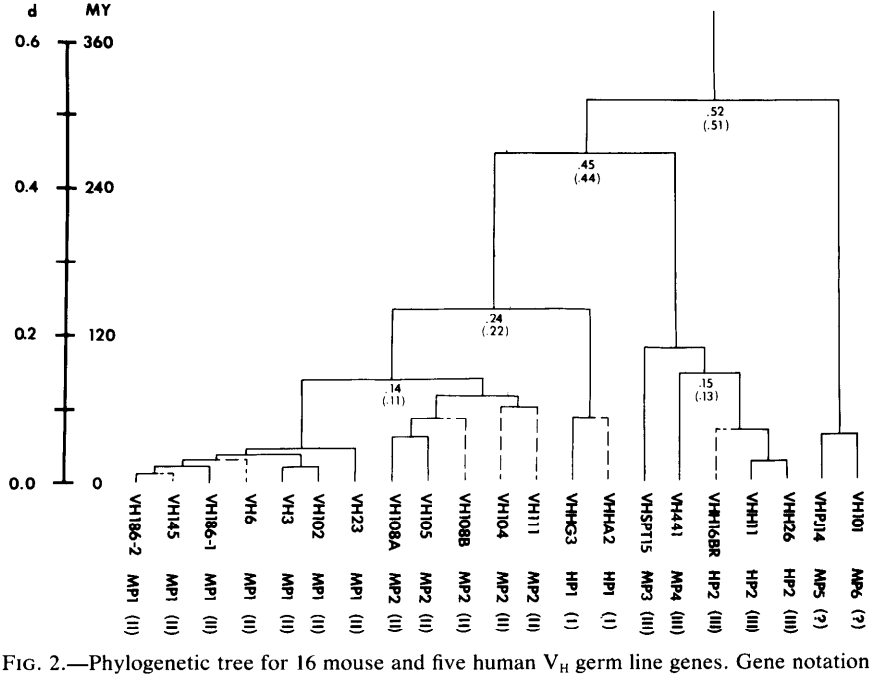


FIG. 2.—Phylogenetic tree for 16 mouse and five human V_H germ line genes. Gene notations are the same as those used by the original authors. MP1, MP2, etc. represent the DNA probe used, whereas I, II, and III are the subtypes of the polypeptide produced (see table 1). Dashed lines show pseudogenes. The phylogenetic tree constructed for the 14 functional genes has the same topology as that of the above tree if the pseudogene part is excluded. However, some branch lengths are shorter than those of the all-gene tree. The values in parentheses in the above tree diagram are the branch lengths for the functional-gene tree, whereas the values above them are those for the all-gene tree. The other branch lengths were nearly the same for the two trees. In this figure, d represents δ in the text.

the early evolutionary stage might have been lost or may remain undetected because of the limited number of genes sampled. If we assume that (i) there is no selection, (ii) any one of the n genes in the V_H family is subject to gene duplication with probability λ per year, and (iii) whenever a gene is duplicated, another gene is eliminated, so that the total number of genes (n) remains constant, then the expected time during which all of s genes sampled ($s \leqslant n$) are derived from a single common ancestral gene is given by

$$\bar{t}=\frac{n-1}{\lambda}\left(1-\frac{1}{s}\right) \tag{2}$$

(F. Tajima, personal communication). This is approximately equal to Ohta's (1983*b*) fixation time of a single mutant repeat gene if s is large. Furthermore, the expectation of the number of observable gene duplications for a sample of s genes is

$$\begin{aligned} r(s) &= 2\sum_{i=2}^{s}\frac{1}{i} \\ &\approx 2(\log_e s - 0.423) \end{aligned} \tag{3}$$

for a large s (F. Tajima, personal communication). Therefore, the expectation of the rate of observable gene duplication becomes

$$d\equiv\frac{r(s)}{\bar{t}}=\frac{2\lambda s(\log_e s-0.423)}{(n-1)(s-1)}. \tag{4}$$

Under the present assumption, the rate of occurrence of gene duplication can be obtained from (2) by putting $\bar{t}$ = 300 MY, n = 160, and s = 16. It becomes 5×10^{-7}. This is more than 20 times higher than the rate of observed gene duplication but still very low compared with the general view that the turnover of variable region genes is quite rapid (Hood et al. 1975; Honjo 1983; Ohta 1983*a*). (Ohta [1983*a*], e.g., speculated $\lambda = 10^{-5}$ per generation.) Although the number of human V_H genes examined is small, one may obtain another estimate of λ from these genes. In this case $\bar{t}$ = 270 MY, n = 80, and s = 5, so that $\lambda = 2.3 \times 10^{-7}$. This is again very small. Incidentally, the expected rate of observable gene duplication can be obtained from (4) by using $r(s)$ = 4.7 and $\bar{t}$ = 300 MY for the mouse. It becomes 1.6×10^{-8}. This is very close to the observed rate (2×10^{-8}).

The accuracy of our estimate of the rate of occurrence of gene duplication depends on the validity of our assumption of constancy of n. In practice, it is quite possible that the number of V_H genes in mammals steadily increased for the past 300 MY. If this is the case, the rate would be smaller than the λ value obtainable from (2). At the present time, however, it is difficult to know whether our assumption is correct since, except in mice and humans, the number of V_H genes is not known.

Our finding that gene duplication occurs very slowly in the V_H gene family is interesting in relation to immunoglobulin diversity. Immunoglobulin diversity is produced by several different mechanisms, one of which is DNA rearrangement. By this mechanism, each of the constant region genes is combined with one of many variable region genes in the genome, generating a great deal of immunoglobulin diversity (e.g., Leder 1982; Tonegawa 1983). We note that if there were little genetic variation among different variable region genes, as in the case of the tRNA and rRNA gene families, this mechanism would have been virtually useless

in producing immunoglobulin diversity. Both unequal crossing-over and gene conversion are mechanisms to reduce genetic variability among different copies of genes, so that a low frequency of their occurrence is preferable in order to have a large amount of immunoglobulin diversity. Probably for this reason, the frequency of gene duplication is kept low in the V_H gene family.

Rates of Nucleotide Substitution

In the study of the evolutionary times of gene duplication, we assumed that mouse gene VH441 and human gene VHH11 or VHH26 diverged 80 MY ago. To see whether this assumption is reasonable, we examined the rates of nucleotide substitution for the first, second, and third nucleotide positions of codons between these mouse and human genes for the FRs and CDRs separately. The results obtained are presented in table 3. It is seen that the substitution rates for FRs are close to those for globin genes, where the first-, second-, and third-position rates for functional genes have been estimated to be 0.85×10^{-9}, 0.7×10^{-9}, and 2.65×10^{-9} per site per year (Li et al. 1981). Particularly interesting is the similarity in the third-position rate between immunoglobulin genes and globin genes. Since the majority of third-position changes are synonymous and are known to occur at more or less the same rate for all genes (Kimura 1983), this similarity suggests that our assumption of the divergence time of 80 MY between genes VH441 and VHH11 is reasonable.

A quantity closely related to the third-position rate is the rate of silent nucleotide substitution. Miyata et al. (1980) have shown that this rate is about 5×10^{-9} for many different genes. We have therefore examined this rate by using Miyata and Yasunaga's (1980) method for computing silent substitutions. Our result for the FRs was 4.3×10^{-9}, which is very close to Miyata et al.'s estimate.

Another support for our assumption comes from the substitution rate for the CDRs. Nei (1983) estimated the rate of nucleotide substitution for the CDRs of the light-chain κ variable region genes (V_κ gene family) by using the most closely related genes available (K2 and HK101) between the mice and humans and assuming that the two genes diverged 80 MY ago. The results obtained showed that

Table 3
Number of Nucleotide Substitutions per Site (δ) and the Rates of Nucleotide Substitutions per Site per Year (λ)[a]

Region (codons used) and Nucleotide Position	δ			λ($\times 10^{-9}$)
	VH441-VHH11	VH441-VHH26	Average	
FR[b] (93):				
First	.15 ± .04	.17 ± .05	.16 ± .05	1.01 ± .28
Second	.10 ± .04	.10 ± .04	.10 ± .04	.65 ± .22
Third	.42 ± .09	.42 ± .09	.42 ± .09	2.63 ± .53
CDR (22):				
First	.41 ± .17	.59 ± .23	.50 ± .20	3.14 ± 1.28
Second	.50 ± .20	.70 ± .27	.60 ± .24	3.74 ± 1.48
Third	.97 ± .39	1.42 ± .68	1.19 ± .55	7.47 ± 3.45

[a] Obtained from the comparison of the mouse gene VH441 with the human genes VHH11 and VHH26 under the assumption that they diverged 80 MY ago.

[b] The leader region is included.

the rates for the first, second, and third nucleotide positions are 3.1×10^{-9}, 5.5×10^{-9}, and 4.8×10^{-9} per site per year, respectively. The corresponding rates for the V_H genes are 3.1×10^{-9}, 3.7×10^{-9}, and 7.5×10^{-9} (table 3). Considering the large standard errors associated with these estimates, the two sets of data are in good agreement. Although there is a possibility that both sets of genes used diverged earlier than 80 MY ago, the agreement of the rates of substitution suggests that the actual time of divergence between genes VH441 and VHH11 or K2 and HK101 would not be far above the time we assumed. It is also interesting to note that the substitution rates for the three nucleotide positions of codons are more or less the same for the V_H and V_κ genes and approximately equal to the pseudogene rate (4.6×10^{-9}; Li et al. 1981). It is therefore possible that most mutations at the CDRs are selectively neutral, and the rate of nucleotide substitution is approximately equal to the mutation rate.

Pseudogenes in the V_H Gene Family

In table 1 we have listed seven pseudogenes or probable pseudogenes, of which five have termination codons. The remaining two seem to be pseudogenes for the reasons mentioned earlier. If we consider three more partially sequenced genes (ψVH3, VH11, and VH13) from the T15 mouse immunoglobulin gene group (Crews et al. 1981), the total number of known pseudogenes becomes eight, because ψVH3 is known to have a termination codon and a four-base insertion. This brings the proportion of pseudogenes to 8/24, or one-third. If this proportion applies to the entire V_H family genes, about 53 of the mouse V_H genes and about 27 of the human V_H genes are expected to be pseudogenes.

Why are there so many pseudogenes in the V_H family? The answer seems to be that the occurrence of nonfunctional genes or pseudogenes is an inevitable consequence of destructive mutations that accumulate in duplicate genes as long as some of the duplicate genes are functional and sufficient for producing required proteins. Theoretically, if the product of one copy of a gene is sufficient for all the required physiological functions, the rest of the duplicate copies can be disabled. In the case of immunoglobulins, however, many different kinds of antibody must be produced, so many copies are required. If the number of functional genes is reduced to a certain value, the immunological system may become less efficient and consequently give a selective disadvantage to its carrier. It is therefore likely that the number of functional genes in the genome is determined by the balance between this functional requirement and the mutational destruction of duplicate genes. It is then possible that the proportion of pseudogenes is higher in organisms with a large number of V_H genes than in those with a small number.

In the presence of concerted evolution there are two ways in which the number of pseudogenes increases. One is the disabling of functional genes, and the other is the duplication of pseudogenes themselves. The latter mechanism apparently operated in the cases of *Xenopus* 5S RNA pseudogenes (Jacq et al. 1977) and goat globin pseudogenes (Cleary et al. 1981; Li and Gojobori 1983). However, the pseudogenes in the V_H family seem to have been produced mainly by the former mechanism, as is clear from the phylogenetic tree of genes in figure 2. The close relationship between mouse pseudogenes VH104 and VH111 suggests the possibility that they diverged after their ancestral gene became nonfunctional. However, these two genes have a termination codon at different nucleotide positions, as mentioned earlier, and it is therefore likely that they were disabled independently after their divergence.

It should be mentioned that the nonfunctionality of V_H pseudogenes is caused mostly by nucleotide substitution rather than by deletion or insertion. Only in one of the eight pseudogenes is nonfunctionality a result of deletion. This is quite different from the situation in other gene families such as the globin family (Proudfoot and Maniatis 1980; Li 1983). The reason for this difference is not clear.

Miyata and Yasunaga (1981) and Li et al. (1981) showed that the rate of nucleotide substitution for globin pseudogenes is much higher than the rate for functional globin genes. To see if this is also true with immunoglobulin genes, we compared the substitution rates for the pseudogenes and functional genes. In this case it is important to note that the disabling of a gene may occur some time after gene duplication, although theoretically a gene can be disabled at the time of gene duplication as well (Leder et al. 1981). A general scheme of evolution of pseudogenes is given in figure 3. In this figure, T_d is the time since gene duplication, and T_n is the time since disabling. To estimate T_d and T_n, we need another functional gene (C) of which the time of divergence (T) from pseudogene A and functional gene B is known.

In figure 3, l, m, and n stand for the number of nucleotide substitutions for the evolutionary branches O-A, O-B, and O-C, respectively. If a pseudogene has a higher rate of nucleotide substitution than a functional gene and T_n is sufficiently large, we expect l to be larger than m. Let us first examine this point, estimating l, m, and n from δ values. To estimate l, m, and n, we note the properties: $l + m = \delta_{AB}$, $l + n = \delta_{AC}$, and $m + n = \delta_{BC}$, where δ_{AB}, δ_{AC}, and δ_{BC} represent the values of δ between sequences A and B, A and C, and B and C, respectively. Therefore, l, m, and n can be estimated by

$$\begin{aligned} l &= \bar{\delta} - \delta_{BC}, \\ m &= \bar{\delta} - \delta_{AC}, \\ n &= \bar{\delta} - \delta_{AB}, \end{aligned} \tag{5}$$

where $\bar{\delta} = (\delta_{AB} + \delta_{AC} + \delta_{BC})/2$. We note that for estimating l, m, and n no

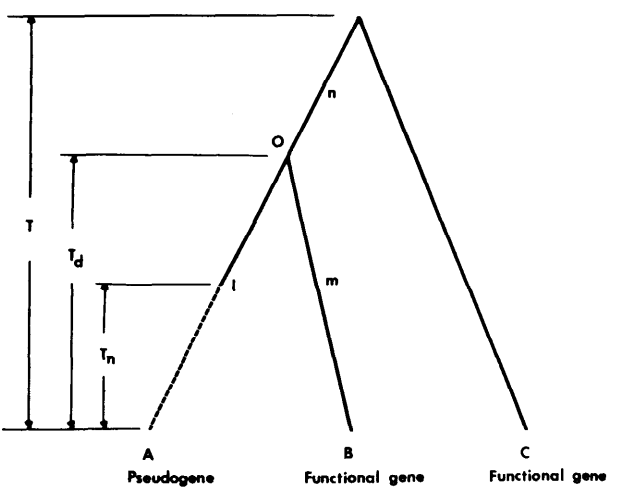


FIG. 3.—Schematic evolutionary relationship among pseudogene A and two related functional genes B and C; l, m, and n denote the number of nucleotide substitutions between O and A, between O and B, and between O and C, respectively. T, T_d, and T_n represent the time since divergence between B and C, time since divergence between A and B, and time since disabling of A, respectively.

information on T is required. Therefore, we used closely related genes as C for each pseudogene. The estimates of l, m, and n for the seven pseudogenes are presented in table 4. In the computation of these estimates, the CDRs were again excluded. It is clear that in the cases of pseudogenes VH108B and VHH16BR l is greater than m in all three nucleotide positions. This indicates that the rate of nucleotide substitution in these pseudogenes is higher than that in their functional counterparts. In other pseudogenes, however, l is not necessarily greater than m, and the difference between them is not statistically significant. This suggests that, in these pseudogenes, nucleotide substitution has not been accelerated or that the time since disabling is too short for the accelerated substitution rate to be detected.

Interestingly, pseudogenes VH104 and VHH11 do not show a clear-cut accelerated rate, although they apparently diverged quite a long time ago (see fig. 2). It is possible that these two genes were disabled relatively recently. Indeed, comparison of the DNA sequences of these two genes has indicated that the numbers of nucleotide substitutions between the first and second nucleotide positions of codons (0.117 and 0.074 per site, respectively) are smaller than the number at the third position (0.181). This suggests that the two genes were functional and subjected to purifying selection at the protein level until recently.

For the two pseudogenes (VH108B and VHH16BR) that showed an accelerated rate of nucleotide substitution, we estimated T_d and T_n in figure 3 by using Li et al.'s (1981) statistical method. Mouse functional gene VH441 served as gene C in figure 3 for the human pseudogene VHH16BR, assuming $T = 80$ MY. For the mouse pseudogene VH108B, we used the functional genes identified by probe MP1 as C, assuming $T = 65$ MY (see fig. 2). The results obtained are presented in table 5, together with the estimate of the rate of nucleotide substitution for pseudogenes (see Li et al. [1981] for the method for estimating this rate). The estimate (ca. 5×10^{-9} per site per year) of the rate of nucleotide substitution for pseudogenes is of the same order of magnitude as that for globin pseudogenes

Table 4
Estimates of the Number of Nucleotide Substitutions

Pseudogene	Functional Genes		Nucleotide Position								
			First			Second			Third		
A	B	C	l	m	n	l	m	n	l	m	n
VH108B	VH105	VH186-2	.13	−.01	.14	.07	.05	.05	.12	.05	.13
VH108B	VH105	VH102	.11	.00	.13	.07	.05	.03	.09	.03	.15
VHH16BR	VHH11	VH441	.06	.01	.15	.05	.01	.10	.12	.10	.32
VHH16BR	VHH26	VH441	.07	.03	.14	.05	.01	.10	.11	.09	.33
VH145	VH23	VH105	.04	.00	.10	.03	.04	.06	.01	.05	.15
VH145	VH3	VH105	.02	.02	.12	.03	.02	.06	.01	.04	.15
VH104	VH105	VH186-2	.08	.00	.13	.05	.03	.06	.02	.05	.13
VH104	VH105	VH3	.06	.02	.12	.04	.04	.04	.02	.05	.15
VH111	VH105	VH186-2	.10	.00	.12	.04	.05	.04	.15	.05	.13
VH111	VH105	VH3	.09	.01	.13	.03	.06	.02	.14	.07	.13
VH6	VH23	VH105	.03	.00	.10	.00	.05	.06	.02	.02	.19
VH6	VH3	VH105	.02	.03	.11	.00	.02	.06	.03	.02	.17
VHHA2	VHHG3	VH186-2	.08	.03	.20	.01	.06	.18	.08	.10	.31
VHHA2	VHHG3	VH102	.08	.03	.17	.01	.06	.16	.10	.08	.31

Note.—l, m, and n for the three branches in fig. 3; CDRs are excluded.

Table 5
Rates of Nucleotide Substitution

Pseudogene	Functional Gene			T_n	T_d	T
A	B	C	$\lambda(\times 10^{-9})$	(MY)	(MY)	(MY)
VHH16BR	VHH11	VH441	3.91	14.9	27.2	80.0
	VHH26	VH441	4.29	11.6	30.7	80.0
Average			4.10	13.3	29.0	80.0
		VH3	2.59	26.7[a]	26.7	65.0
		VH23	12.65	6.7	18.0	65.0
VH108B	VH105	VH102	4.67	17.1	26.5	65.0
		VH186-1	7.11	12.8	26.3	65.0
		VH186-2	5.08	19.1	26.6	65.0
Average			6.42	16.5	24.8	65.0

NOTE.—Per site per year for pseudogenes (λ), time since disabling (T_n), and time since duplication (T_d). T is the time since divergence between B and C in fig. 3. CDRs are excluded.

[a] Since $T_n > T_d$, T_n was assumed to be equal to T_d (Li et al. 1981).

(4.6×10^{-9}; Li et al. 1981) and higher than the rate for the third nucleotide position of codons for the FRs of V_H genes (tables 3 and 5). In the case of VHH16BR, the pseudogene and its closest functional counterparts (VHH11 and VHH26) seem to have diverged about 29 MY ago, and disabling of VHH16BR seems to have occurred about 13 MY ago. The mouse pseudogene VH108B seems to have diverged from its closest functional counterpart (VH105) about 25 MY ago and to have been disabled about 17 MY ago. (The functional gene VH108A was not used here because this gene seems to have experienced partial gene conversion.) These estimates have a large standard error. It is noted that the value (25 MY) of T_d for VH108B is substantially smaller than the divergence time (53 MY) between VH108B and VH105 given in figure 2. This difference occurred because the phylogenetic tree in figure 2 was constructed under the assumption of equal rates of nucleotide substitution for both functional genes and pseudogenes.

Discussion

We have seen that concerted evolution occurs very slowly in the V_H genes. This finding is different from what many immunologists have visualized. They (e.g., Hood et al. 1975; Honjo 1983) seem to believe that concerted evolution occurs quite rapidly in the immunoglobulin variable region genes. The main reason for this belief seems to be that "in portions of the V region that are highly conserved, coincidental (concerted) evolution is reflected by species-specific residues at certain positions that distinguish most of the immunoglobulin chains of one species from those of a second" (Hood et al. 1975). However, the existence of species-specific amino acid residues is not really inconsistent with the slow rate of concerted evolution. Namely, even if the turnover of genes occurs as slowly as found here, the existence of species-specific residues can be explained by the low rate of nucleotide substitution ($[1 - 7] \times 10^{-9}$ per site per year); by chance alone certain residues are expected to be conserved for a quite long time, particularly in the presence of gene clusters such as those in figure 2.

The slow rate of concerted evolution in the gene family is in sharp contrast to the evolution of ribosomal RNA (rRNA) genes or transfer RNA (tRNA) genes. Man is known to have about 400 rRNA genes, but the nucleotide sequences of the coding regions of these genes are virtually identical (Arnheim 1983). Recently, studying the restriction site sequences of the nontranscribable spacer (NTS) of rRNA genes in the human, chimpanzee, gorilla, orangutan, and gibbon, Arnheim (1983) and his associates discovered that the human genes have a distinct restriction site sequence and this sequence exists in all copies of rRNA genes. This suggests that this sequence appeared by mutation after the human-ape divergence and has spread through the entire rRNA gene array. Since the human-ape divergence apparently occurred about 5 MY ago (Wilson et al. 1977), concerted evolution must have occurred quite rapidly in the rRNA genes.

If we use equation (2) and assume $\bar{t} = 5 \times 10^6$, the rate of occurrence of gene duplication can be estimated. In the present case $n = s = 400$, so that $\lambda = 8 \times 10^{-5}$. This is more than 100 times higher than the value for the V_H gene family. This value could still be an underestimate because the human restriction site sequence in rRNA genes may have appeared later than the time of human-ape divergence. However, this agrees well with another estimate that can be obtained independently. Studying the frequency of occurrence of *bb* mutants (partial deficiencies of rRNA genes) in *Drosophila melanogaster,* Frankheim et al. (1980) estimated that the rate of occurrence of unequal crossing over for the rRNA gene cluster is 3×10^{-4} per X chromosome per generation. The X chromosome of *D. melanogaster* has about 250 copies of rRNA genes (Tartof 1975), and according to Szostak and Wu's (1980) study with yeast, a single unequal crossing-over event affects seven repeats of rRNA genes on average. Therefore, the rate of occurrence of unequal crossing-over is estimated to be $3 \times 10^{-4} \times (7/250) = 8 \times 10^{-6}$ per gene per generation. If *D. melanogaster* has 10 generations in a year in nature, this gives a rate of 8×10^{-5} per gene per year. This happens to be identical with the above estimate of λ.

These computations indicate that concerted evolution occurs much faster in the rRNA gene family than in the V_H gene family. Why is there so much difference in these two rates? One possible explanation is that the controlling mechanism of unequal crossing-over or gene conversion is not the same for the two gene families. Another explanation is that the selection schemes for the two families are different. As mentioned earlier, diversity is required for immunoglobulins, and the genetic heterogeneity among the V_H genes is one of the important sources of immunoglobulin diversity. In rRNA genes, however, homogeneity seems to be advantageous, since all rRNA molecules apparently have the same function. Since unequal crossing-over and gene conversion both have the effect of reducing the genetic variability among multigenes, they are apparently used effectively for maintaining the homogeneity of rRNAs. That is, a copy of an rRNA gene that deviates from the standard one seems to be disadvantageous and thus to be eliminated. In the case of V_H genes, however, an individual who has many copies of the same DNA sequence owing to unequal crossing-over or gene conversion seems to be at a selective disadvantage and is thus eliminated from the population. This difference in selection scheme would probably be sufficient to explain the difference in the rate of concerted evolution between the V_H and rRNA gene families, though some mathematical study is necessary to test this hypothesis quantitatively. In our view, the latter explanation is more reasonable than the

former since it requires no special mechanism for controlling the frequency of unequal crossing-over or gene conversion. If this view is correct, our estimate of λ for V_H genes refers to the effective rate of gene duplication that corresponds to the case of no selection.

It now seems clear that the genome of higher organisms contains many multigene families (see Li [1983] for a recent review). Indeed, most genes seem to exist in multiple copies in the genome, though the number of copies is not always large. All of these multigene families apparently undergo concerted evolution, but the pattern of evolution is expected to vary from gene family to gene family according to the type of requirement for multiple genes. The evolutionary patterns of the V_H and rRNA gene families are probably the two extremes in terms of the rate of fixation of duplicate genes. To understand the significance of concerted evolution for the entire genome, we must study the evolutionary patterns of many other gene families.

Antibody diversity is produced by various mechanisms such as germ line mutation, DNA rearrangement, somatic mutation, and deletion/insertion (Tonegawa 1983). In the present paper we are concerned only with germ line genes. However, we note that the genetic variation among the germ line V_H genes is very large, as is clear from table 2. This large genetic variation is translated into antibody diversity through DNA rearrangement, and this factor alone is very important for generating antibody diversity. It is interesting to note that the CDRs are particularly variable (hypervariable) and the rate of nucleotide substitution in these regions is as high as the pseudogene rate, which is supposed to be equal to the intrinsic mutation rate at the nucleotide level (table 3). This suggests that there is virtually no purifying selection operating in the CDRs, and germ line mutations are effectively used for generating antibody diversity (Nei 1983). This finding is similar to Ohta's (1978) earlier observation that the rate of amino acid substitution at the hypervariable regions (CDRs) is equal to that of fibrinopeptides, the highest known rate for proteins.

Addendum

After we completed our work, Litman et al. (1983) published the nucleotide sequence of a V_H germ line gene from *Caiman crocodylus*, a reptile. This sequence is relatively closely related to the mouse gene VHSPT15, the δ value being .35. This evolutionary distance corresponds to 210 MY if we use the evolutionary time scale given in figure 2. Since reptiles and mammals diverged about 300 MY ago, this indicates that our time scale could be erroneous. If we assume that $\delta = .35$ corresponds to an evolutionary time of 300 MY, all the mouse and human genes in figure 2 will become older by about one-third. Therefore, our conclusion that the concerted evolution of the V_H gene family occurs very slowly will remain unchanged. However, since the number of nucleotide substitutions in a single gene of about 93 codons is known to be subject to large stochastic errors, it is not clear whether the results obtained from the mouse-*Caiman* divergence are more reliable than those from the mouse-human divergence. It should also be noted that there is a possibility of horizontal gene transfer (Busslinger et al. 1982) in this case. We have therefore decided not to extend our analysis to these new sequence data at this moment.

Acknowledgments

We thank Tomoko Ohta and two reviewers for their valuable comments. This study was supported by research grants from the National Institutes of Health and the National Science Foundation.

LITERATURE CITED

ARNHEIM, N. 1983. Concerted evolution of multigene families. Pp. 38–61 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

BOTHWELL, A. L. M., M. PASKIND, M. RETH, T. IMANISHI-KARI, K. RAJEWSKY, and D. BALTIMORE. 1981. Heavy chain variable region contribution to the NP^b family of antibodies: somatic mutation evident in a γ2a variable region. Cell **24**:625–637.

BUSSLINGER, M., S. RUSCONI, and M. L. BIRNSTIEL. 1982. An unusual evolutionary behaviour of a sea urchin histone gene cluster. EMBO J. **1**:27–33.

CLEARY, M. L., E. A. SCHON, and J. B. LINGREL. 1981. Two related pseudogenes are the result of a gene duplication in the goat β-globin locus. Cell **26**:181–190.

COHEN, J. B., K. EFFRON, G. RECHAVI, Y. BEN-NERIAH, R. ZABUT, and D. GIVOL. 1982. Simple DNA sequences in homologous flanking regions near immunoglobulin V_H genes: a role in gene interaction? Nucleic Acids Res. **10**:3353–3370.

CREWS, S., J. GRIFFIN, H. HUANG, K. CALAME, and L. HOOD. 1981. A single V_H gene segment encodes the immune response to phosphorylcholine: somatic mutation is correlated with the class of the antibody. Cell **25**:59–66.

FRANKHEIM, R., D. A. BRISCOE, and R. K. NURTHEN. 1980. Unequal crossing over at the rRNA tandon as a source of quantitative genetic variation in Drosophila. Genetics **95**:727–742.

GIVOL, D., R. ZAKUT, K. EFFRON, G. RECHAVI, D. RAM, and J. B. COHEN. 1981. Diversity of germline immunoglobulin V_H genes. Nature **292**:426–430.

GOJOBORI, T., K. ISHII, and M. NEI. 1982. Estimation of the average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol. **18**:414–423.

HALDANE, J. B. S. 1933. The part played by recurrent mutation in evolution. Amer. Natur. **67**:5–19.

HONJO, T. 1983. Immunoglobulin genes. Annu. Rev. Immunol. **1**:499–528.

HOOD, L., J. H. CAMPBELL, and S. C. R. ELGIN. 1975. The organization, expression, and evolution of antibody genes and other multigene families. Annu. Rev. Genet. **9**:305–353.

JACQ, C., J. R. MILLER, and G. G. BROWNLEE. 1977. A pseudogene structure in 5S DNA of *Xenopus laevis*. Cell **12**:109–120.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KABAT, E. A., T. T. WU, and H. BILOFSKY. 1979. Sequences of immunoglobulin chains. NIH Pub. no. 80-2008.

KATAOKA, T., T. NIKAIDO, T. MIYATA, K. MORIWAKI, and T. HONJO. 1982. The nucleotide sequences of rearranged and germline immunoglobulin V_H genes of a mouse myeloma MC101 and evolution of V_H genes in mouse. J. Biol. Chem. **257**:277–285.

KEMP, D. J., B. TYLER, O. BERNARD, N. GOUGH, S. GERONDAKIS, J. M. ADAMS, and S. CORY. 1981. Organization of genes and spacers within the mouse immunoglobulin V_H locus. J. Mol. Appl. Genet. **1**:245–261.

KIM, S., M. DAVIS, E. SINN, P. PATTEN, and L. HOOD. 1981. Antibody diversity: somatic hypermutation of rearranged V_H genes. Cell **27**:573–581.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Nat. Acad. Sci. **78**:454–458.

———. 1983. The neutral theory of molecular evolution. Pp. 208–233 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

LEDER, A., D. SWAN, F. RUDDLE, P. D'EUSTACHIO, and P. LEDER. 1981. Dispersion of α-like globin genes of the mouse to three different chromosomes. Nature **293**:196–200.

LEDER, P. 1982. The genetics of antibody diversity. Sci. Amer. **246**:102–115.

LI, W.-H. 1983. Evolution of duplicate genes and pseudogenes. Pp. 14–37 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

LI, W.-H., and T. GOJOBORI. 1983. Rapid evolution of the goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1**:94–108.

LI, W.-H., T. GOJOBORI, and M. NEI. 1981. Pseudogenes as a paradigm of neutral evolution. Nature **292**:237–239.

LITMAN, G. W., L. BERGER, K. MURPHY, R. LITMAN, K. HINDS, C. L. JAHN, and B. W. ERICKSON. 1983. Complete nucleotide sequence of an immunoglobulin V_H gene homologue from *Caiman*, a phylogenetically ancient reptile. Nature **303**:349–352.

MATTHYSSENS, G., and T. H. RABBITTS. 1980. Structure and multiplicity of genes for the human immunoglobulin heavy chain variable region. Proc. Nat. Acad. Sci. **77**:6561–6565.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**:23–36.

———. 1981. Rapidly evolving mouse α globin-related pseudo gene and its evolutionary history. Proc. Nat. Acad. Sci. **78**:450–453.

MIYATA, T., T. YASUNAGA, and T. NISHIDA. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Nat. Acad. Sci. **77**:7328–7332.

NEEDLEMAN, S. B., and C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.

NEI, M. 1969. Gene duplication and nucleotide substitution in evolution. Nature **221**:40–41.

———. 1975. Molecular population genetics and evolution. North-Holland, Amsterdam and New York.

———. 1983. Genetic polymorphism and the role of mutation in evolution. Pp. 165–190 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

NEI, M., F. TAJIMA, and Y. TATENO. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. **19**:153–170.

OHTA, T. 1978. Sequence variability of immunoglobulins considered from the standpoint of population genetics. Proc. Nat. Acad. Sci. **75**:5108–5112.

———. 1980. Evolution and variation of multigene families. Springer, Berlin.

———. 1983*a*. On the evolution of multigene families. Theoret. Pop. Biol. **23**:216–240.

———. 1983*b*. Time until fixation of mutant belonging to a gene family. Genet. Res. **41**:47–55.

OLLO, R., C. AUFFRAY, J.-L. SIKORAV, and F. ROUGEON. 1981. Mouse heavy chain variable regions: nucleotide sequence of a germline V_H gene segment. Nucleic Acids Res. **9**:4099–4109.

PROUDFOOT, N. J., and T. MANIATIS. 1980. The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell **21**:537–544.

RABBITTS, T. H., G. MATTHYSSENS, and P. H. HAMLYN. 1980. Contribution of immunoglobulin heavy-chain variable-region genes to antibody diversity. Nature **284**:238–243.

RECHAVI, G., B. BIENZ, D. RAM, Y. BEN-NERIAH, J. B. COHEN, R. ZAKUT, and D. GIVOL. 1982. Organization and evolution of immunoglobulin V_H gene subgroups. Proc. Nat. Acad. Sci. **79**:4405–4409.

RECHAVI, G., D. RAM, L. GLAZER, R. ZAKUT, and D. GIVOL. 1983. Evolutionary aspects of immunoglobulin heavy chain variable region (V_H) gene subgroups. Proc. Nat. Acad. Sci. **80:**855–859.

SAKANO, H., R. MAKI, Y. KUROSAWA, W. ROEDER, and S. TONEGAWA. 1980. Two types of somatic recombination are necessary for the generation of complete immunoglobulin heavy-chain genes. Nature **286:**676–683.

SMITH, G. P., L. HOOD, and W. M. FITCH. 1971. Antibody diversity. Annu. Rev. Biochem. **40:**969–1012.

SZOSTAK, J. W., and R. WU. 1980. Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae*. Nature **284:**426–430.

TAKAHATA, N., and M. KIMURA. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics **98:**641–657.

TARTOF, K. D. 1975. Redundant genes. Annu. Rev. Genet. **9:**355–385.

TATENO, Y., M. NEI, and F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18:**387–404.

TONEGAWA, S. 1983. Somatic generation of antibody diversity. Nature **302:**575–581.

WILSON, A. C., S. S. CARLSON, and T. J. WHITE. 1977. Biochemical evolution. Annu. Rev. Biochem. **46:**573–639.

WALTER M. FITCH, reviewing editor

Received July 12, 1983; revision received July 28, 1983.

Retention of Cryptic Genes in Microbial Populations[1]

Wen-Hsiung Li
University of Texas at Houston

Cryptic genes are silenced genes that can still be reactivated by mutation. Since they can make no positive contribution to the fitness of their carriers, it is not clear why many cryptic genes in microbial populations have not degenerated into useless DNA sequences. Hall et al. (1983) have suggested that cryptic genes have persisted because of occasional strong environmental selection for reactivated genes. The present mathematical study supports their suggestion. It shows that a cryptic gene can be retained *without having any selective advantage over a useless DNA sequence,* if selection for the reactivated gene occasionally occurs for a substantially long time.

Introduction

A *cryptic gene* is a silenced gene that is a single nucleotide substitution, that is present at a high frequency in a population, and that can still be reactivated by mutation, recombination, insertion, deletion, or other genetic mechanisms (Hall et al. 1983). A nonfunctional gene is defined as a silenced gene that cannot be reactivated by a single mutational event. A cryptic gene cannot make a positive contribution to the fitness of its carriers, and it may become nonfunctional by the pressure of mutation causing further degeneration or deletion. Yet cryptic genes appear to be widespread in microbial populations (Hall et al. 1983). To explain this phenomenon, Hall et al. (1983) have proposed a model in which the cryptic state is advantageous under one set of environmental conditions, whereas the functional state is favored under another set of environmental conditions. In their mathematical treatment, however, they have not taken into account the environmental change but have assumed constant fitnesses for all three states. Under this assumption, the cryptic gene (and its functional allele) can be retained in the population only if the fitness of the functional allele is higher than those of the cryptic and nonfunctional alleles, or if the fitness of the cryptic allele is higher

1. Key words: cryptic genes, population dynamics, microbial evolution.

Address for correspondence and reprints: Dr. Wen-Hsiung Li, P.O. Box 20334, Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77025.

Mol. Biol Evol. 1(2):213–219. 1984.
© 1984 by The University of Chicago. All rights reserved.
0737-4038/84/0102-0004$02.00

than that of the nonfunctional allele. Under the former condition, the frequency of the functional allele would usually be high—a situation contrary to the observation that the frequency is usually very low. The latter condition does not seem realistic because the only difference between a cryptic and a nonfunctional allele is that the latter cannot be reactivated by a single mutation. In the following, I shall remove the assumption of constant fitnesses and show that neither of the above two conditions is necessary for the retention of cryptic genes in a population.

Mathematical Theory

Absence of Mutation

Hall et al. (1983) used A_1, A_2, and A_3 to denote the cryptic, the functional, and the nonfunctional genes, respectively. It is more natural to denote the functional gene by A_1 and the cryptic gene by A_2 because the functional gene was probably the original allele and because the cryptic state is the intermediate state between the other two states. Let $m_1(t)$, $m_2(t)$, and $m_3(t)$ be the Malthusian fitnesses of A_1, A_2, and A_3 at time t, respectively. The first problem to be understood is how fast the gene frequencies can change under different environments. For simplicity, let us assume that $m_2(t) = m_3(t)$. Let $x(t)$ be the frequency of A_1 at time t and a selective difference $s(t) = m_1(t) - m_2(t)$. It can then be shown that

$$\frac{dx}{dt} = s(t)x(1-x)$$

or (1)

$$\ln \frac{x(t)}{1-x(t)} - \ln \frac{x(0)}{1-x(0)} = \int_0^t s(\tau)d\tau.$$

(see Crow and Kimura 1970, pp. 190–192; Nagylaki 1975). As $\ln[x/(1-x)]$ is an increasing function of x, $x(t) \geq x(0)$ if the integral is zero or positive, and $x(t) < x(0)$ if otherwise.

Let us assume that there are two types of environments; in environment 1, $s(t) = s_1$, and in environment 2, $s(t) = -s_2$. We first consider the increase in x in environment 1. In this case,

$$\ln \frac{x(t)}{1-x(t)} - \ln \frac{x(0)}{1-x(0)} = s_1 t. \tag{2}$$

It will be seen from table 1 that, if s_1 is large, the increase in x is extremely rapid. Indeed, if $s_1 = 0.1$, it takes only 92 generations for x to increase from 0.01 to 0.99. If $s_1 = 0.001$, the time required is 100 times longer. This is, however, still a relatively short period, for the generation time in microorganisms is very short.

Table 1
Number of Generations Required for the Frequency of A_1 to Increase from y to z

s	$y = 10^{-6}, z = .01$	$y = .01, z = .99$	$y = .99, z = .999999$
.1.....	92.2	91.9	92.2
.001.....	9,220	9,190	9,220

NOTE.—Assuming that A_1 has an advantage of s over A_2 and A_3.

In environment 2 the frequency of A_1 will decrease. If $s_2 = 0.001$, it takes only 9,190 generations for x to decrease from 0.99 to 0.01.

When the environment varies with time, we can write $t = t_1 + t_2$, where t_i is the number of generations the environment is of the ith type. We can then write equation (1) as

$$\ln \frac{x(t)}{1-x(t)} - \ln \frac{x(0)}{1-x(0)} = s_1 t_1 - s_2 t_2. \tag{3}$$

One simple situation is that the environment varies cyclically. Let T be the period, and assume that in each cycle T_i generations are in environment i so that $T = T_1 + T_2$. If $s_1T_1 - s_2T_2$ is zero, x will vary cyclically with period T, but if the difference is positive (negative) x will increase (decrease) with the number of cycles. As an example, assume $s_1 = 0.1$, $s_2 = 0.001$, $T_1 = 10$, $T_2 = 990$, and $T = 1{,}000$. Then $s_1T_1 - s_2T_2 = 0.01$, and it will take 919 cycles or 919,000 generations for x to increase from 0.01 to 0.99.

Presence of Mutation

To understand the retention of cryptic genes in a population, we must also consider mutation. Let u_{ij} be the mutation rate per generation from A_i to A_j. It can be shown that

$$\frac{dx_1}{dt} = [m_1(t) - m_2(t)]x_1x_2 + [m_1(t) - m_3(t)]x_1x_3 - (u_{12} + u_{13})x_1 + u_{21}x_2, \tag{4}$$

$$\frac{dx_2}{dt} = -[m_1(t) - m_2(t)]x_1x_2 + [m_2(t) - m_3(t)]x_2x_3 - (u_{21} + u_{23})x_2 + u_{12}x_1, \tag{5}$$

$$\frac{dx_3}{dt} = -[m_1(t) - m_3(t)]x_1x_3 - [m_2(t) - m_3(t)]x_2x_3 + u_{13}x_1 + u_{23}x_2, \tag{6}$$

where x_1, x_2, and x_3 are the frequencies of A_1, A_2, and A_3 at time t. In the above equations, I assumed that reversible mutation cannot occur from A_3 to A_1 or A_2, that is, $u_{31} = u_{32} = 0$. Since $x_1 + x_2 + x_3 = 1$, it suffices to consider only two of the three equations. These equations are equivalent to equation (1) of Hall et al. (1983), except that I used different notations and Malthusian instead of Wrightian fitnesses.

We are particularly interested in the following question: *Can the cryptic allele be retained without having a selective advantage over the nonfunctional allele?* To answer this question, we assume that $m_2(t) = m_3(t)$. The second term on the right-hand side of equations (5) and (6) then disappears. It seems from equation (6) that, unless m_1 is on the average larger than m_3 (m_2), x_3 will eventually increase to 1, and A_1 and A_2 will be eliminated. It turns out that this condition is not necessary for the retention of A_1 and A_2. (As mutation can occur between A_1 and A_2, if one of them is retained, the other is also retained.)

Table 2 shows some numerical examples that were obtained by iteration on a computer, replacing equations (5) and (6) by difference equations (see Discussion for the parameter values used). In all examples I assume that the environment

Table 2
Changes in Gene Frequency under Selection and Mutation

Time	$4T + T_1$	$5T$	$9T + T_1$	$10T$
Case 1. $T_1 = 200$, $u_{12} = u_{13} = u_{23} = 10^{-5}$				
x_1	.99975	3.9×10^{-5}	.99975	3.9×10^{-5}
x_2	.00012	.39243	.00012	.39245
Case 2. $T_1 = 200$, $u_{12} = u_{23} = 10^{-5}$, $u_{13} = 10^{-6}$				
x_1	.99986	7.1×10^{-5}	.99986	7.1×10^{-5}
x_2	.00012	.71417	.00012	.71424
Case 3. $T_1 = 200$, $u_{12} = u_{23} = 10^{-6}$, $u_{13} = 10^{-7}$				
x_1	.99743	2.8×10^{-5}	.99997	8.9×10^{-5}
x_2	.00004	.27534	2.9×10^{-5}	.88680
Case 4. $T_1 = 100$, $u_{12} = u_{23} = 10^{-5}$, $u_{13} = 10^{-6}$				
x_1	5.3×10^{-7}	3.0×10^{-11}	1.6×10^{-7}	9.0×10^{-12}
x_2	3.7×10^{-7}	3.0×10^{-7}	1.1×10^{-7}	9.0×10^{-8}
Case 5. $T_1 = 150$, $u_{12} = u_{23} = 10^{-5}$, $u_{13} = 10^{-6}$				
x_1	.00111	1.4×10^{-9}	.03948	5.2×10^{-8}
x_2	6.9×10^{-6}	1.4×10^{-5}	.00024	.00052

NOTE.—The environment is assumed to vary cyclically, with period $T = T_1 + T_2$; in each cycle, the first T_1 generations are in environment 1 and the last T_2 generations in environment 2. In all cases, $s_1 = .1$, $s_2 = .001$, $T_2 = 25{,}000$, $u_{21} = 10^{-7}$, and the initial frequencies are $x_1 = 0$, $x_2 = 10^{-6}$, and $x_3 = .999999$.

varies cyclically and that in each cycle the first T_1 generations are in environment 1 and the last T_2 generations are in environment 2. I assume further that, in environment 1, $m_1 - m_2 = s_1 = 0.1$, and, in environment 2, $m_1 - m_2 = -s_2 = -0.001$. Under these assumptions, in each cycle the frequency of A_1 will be the highest after the first T_1 generations, that is, at $t = (n - 1)T + T_1$, and the lowest at the end of the cycle, that is, at $t = nT$, where n is the number of cycles. In all cases, $T_2 = 25{,}000$ generations, $u_{21} = 10^{-7}$, and the initial frequencies are $x_1 = 0$, $x_2 = 10^{-6}$, and $x_3 = 0.999999$.

In the first three cases, $T_1 = 200$. This T_1 was chosen because A_1 can increase from a very low frequency to a high frequency in 200 generations if $s_1 = 0.1$ (table 1) and because $s_1T_1 - s_2T_2 = -5$, so that m_1 is on the average smaller than m_2. In all of these three cases, A_1 and A_2 can be retained. In case 1, the mutation rate (u_{13}) from A_1 to A_3 is equal to that (u_{12}) from A_1 to A_2. Therefore, the frequency (x_2) of A_2 is expected to be lower than that (x_3) of A_3, for mutation occurs irreversibly from A_2 to A_3. Indeed, x_2 oscillates between 0.00012 and 0.39, whereas x_3 is between 0.00013 and 0.61. In case 2, $u_{13} = 10^{-6}$ is one order smaller than $u_{12} = 10^{-5}$, and x_2 can reach a value as high as 0.71; the maximum value of x_3 is only 0.29. In case 3, $u_{12} = u_{23} = 10^{-6}$ and $u_{13} = 10^{-7}$, and x_2 can become as large as 0.89. This value is higher than that in case 2 mainly because the mutation rate (u_{23}) from A_2 to A_3 in case 3 is one order lower than that in case 2 (10^{-6} vs. 10^{-5}), so that the decrease in x_2 resulting from irreversible mutation is slower in case 3 than in case 2.

In case 4, T_1 is only 100, so that A_1 cannot reach a substantially high frequency from a very low frequency (table 1). It is seen that x_1 and x_2 decrease with the number of cycles. Therefore, A_1 and A_2 will eventually be eliminated from the population. In case 5, T_1 is 150. Since x_1 and x_2 increase with the number of cycles, A_1 and A_2 can be retained in the population.

The last example suggests that, if $T_1 = 200$, A_1 and A_2 can be retained even if T_2 is considerably larger than 25,000. Indeed, if $T_2 = 250{,}000$ and the other parameter values are the same as those in case 3, x_2 increases from 0.000001 at $t = 0$ to 0.00014 at the end of the second cycle. That is, A_1 and A_2 can be retained in the population. Note that in this case $s_1T_1 - s_2T_2 = -230$, that is, A_1 is on the average much less fit than A_2 and A_3.

The constant-fitness model predicts that A_1 and A_2 cannot be retained in any of the above cases because $s_1T_1 - s_2T_2 < 0$ in every case. This prediction holds only for case 4. I shall explain later why the difference occurs between the two models.

Discussion

The biological rationale for the mutation rates used in the above examples is as follows. I assumed that the rate (u_{12}) of mutation from the functional allele (A_1) to the cryptic allele (A_2) is at least as high as that (u_{13}) from the functional allele to the nonfunctional allele (A_3). This assumption seems reasonable because the former type of mutation would include defects in regulatory sequences, nonsense mutations, and frameshifts owing to minor insertions or deletions, while the latter type would be more drastically destructive changes such as large deletions. Limited data suggest that, in bacteria, u_{12} is of the order of 10^{-5} or lower (table 23-1 in Strickberger [1976]). Thus, the assumption of $u_{12} = 10^{-5}$ or 10^{-6} seems reasonable. I also assumed that back mutation from A_2 to A_1 occurs at the rate of $u_{12} = 10^{-7}$. Available data suggest that this rate ranges from 10^{-5} to 10^{-8} (Strickberger [1976], table 23-1; Hall et al. 1983). Obviously a higher rate of back mutation is more favorable for the retention of A_1 and A_2. In addition, I assumed that the rate (u_{23}) of mutation from A_2 to A_3 is the same as that (u_{12}) from A_1 to A_2, reasoning that most mutations of the former type would be similar to mutations of the latter type, that is, regulatory defects, nonsense mutations, minor insertions, and minor deletions. A higher u_{23} value is obviously less favorable for the retention of A_1 and A_2. However, unless u_{23} is actually considerably higher than assumed, the conclusion drawn from the above examples will still hold.

The rationale for the selection coefficients used is as follows. I assumed that $s_1 = 0.1$ and $s_2 = 0.001$. In practice, the s_1 value may be larger than 0.1 because, as a result of their inability to utilize a resource available to A_1 individuals, individuals without the functional allele (A_1) may grow much more slowly or may not even be able to survive in environment 1. If s_1 is indeed larger, the number of generations required for a substantial increase in x_1 would be smaller than those assumed in table 2. Of course, the contrary would be true if s_1 were smaller than 0.1. A larger s_2 value is less favorable for the retention of A_1 and A_2 because in environment 2 the decrease in x_1 will become faster. It can, however, be shown that the retention of A_1 and A_2 is not strongly dependent on the magnitude of s_2. For example, if $s_2 = 0.01$ instead of 0.001 and the other parameter values are the same as those in case 2, A_1 and A_2 can still be retained.

In all the above examples, I have assumed that the environment changes cyclically. This assumption was made to simplify numerical computations. Actually, many general conclusions can be drawn from these examples. First, if x_1 or x_2 becomes high, then the retention of A_1 and A_2 is assured for a long period of time, say at least one million generations, regardless of how the environment changes. For instance, in case 2, x_2 was high (0.71) at the end of the fifth cycle,

so that A_1 and A_2 will still be retained even if the environment continues to be in type 2 for one million generations before switching to type 1. Second, if x_1 becomes very high before the environment switches to type 2, x_2 may increase to a high value (see cases 2 and 3). Third, A_1 and A_2 may persist in low frequencies for a long period of time without being eliminated from the population (case 5). Fourth, a long run of environment 1 is more effective for the retention of A_1 and A_2 than many short runs with a total duration equal to that of the long run. This follows from the observation that A_1 and A_2 cannot be retained if $T_1 = 100$ and $T_2 = 25{,}000$ (case 4) but can be retained if $T_1 = 200$ and $T_2 = 250{,}000$; the ratios of T_1 to T_2 for the two cases are 0.004 and 0.0008, respectively. In the extreme case where $T_1 = 1$, A_1 and A_2 cannot be retained if $T_2 \geq s_1/s_2$, because the selection for A_1 is too ineffective to compensate the effect of irreversible mutation. In this extreme case, the conditions for the retention of A_1 and A_2 are similar to those required under the constant-fitness model.

In essence, the above examples show that A_2 can be retained without having any selective advantage over A_3, if selection for A_1 occasionally occurs for a substantially long time. The reason for this is quite simple. Although in environment 2 the frequency of A_1 will decrease fairly rapidly, A_2 can persist for a long time because x_2 will first increase at a fast rate and then decrease at the rate of mutation. If the environment switches to type 1 for a substantially long period before x_2 becomes extremely small, a substantial increase in x_1—and, consequently, a substantial reduction in x_3—will occur. When the environment switches to type 2, x_2 will start to increase. In particular, should x_1 become almost 1, x_2 may increase to a high value, as was seen in cases 2 and 3 above. (A similar argument has been made by Hall et al. [1983], though no mathematical treatment was given.) By contrast, under the constant-fitness model, if A_1 is less fit than A_2 and A_3, and A_2 and A_3 are equally fit, then the frequency of A_1 will always decrease, eventually to zero, because of selection and irreversible mutation, and the frequency of A_2 will also eventually decrease to zero because of irreversible mutation. In conclusion, the constant-fitness model requires more stringent conditions than actually needed for the retention of cryptic genes.

So far I have not considered the effect of population subdivision. To see this effect, let us consider a simple example. Suppose that there are two populations which exchange 10% of their genes in each generation. We assume that the environment varies cyclically with $T = 25{,}200$ generations and that in each cycle population 1 is in environment 2, except for the first 100 generations, and population 2 is in environment 2, except for the second 100 generations. The other parameter values are the same as those in case 4 in table 2. Numerical computations show that at the end of the fifth cycle x_2 increases to 0.0002 in both populations. Thus, in contrast to case 4, the cryptic gene can be retained. This example indicates that the conditions required for the retention of cryptic genes are weaker in a subdivided population than in a population without subdivision.

We may conclude from these results that loss of a cryptic gene from a microbial population is an extremely rare event. This conclusion, however, does not hold if a population goes through a severe bottleneck, a situation that apparently often occurs at the time of speciation. This might have been the case in *Shigella dysenteriae,* which lost the gene for a lactose permease (*lac Y* gene) after its separation from *Escherichia coli* (Hall et al. 1983). The loss appears to have occurred rather quickly, because *Shigella* and *Escherichia* are close relatives and,

in fact, are considered by some authors to be one species (Whittam et al. 1983). The fact that the gene diversity is considerably lower in *S. dysenteriae* than in *E. coli* (Whittam et al. 1983) suggests that the former has gone through at least one bottleneck. Let us assume that a bottleneck has indeed occurred and see how quickly the nonfunctional allele can become fixed in the population, assuming fixation will eventually occur. Let N_e be the effective population size and p the frequency of A_3 right after the bottleneck. Neglecting the effect of mutation and the existence of A_1, we can treat the problem as the conditional fixation of a neutral allele in a population. Using formula (14) of Kimura and Ohta (1969), we can show that the mean conditional fixation time ($\bar{t}_1$) is between $2.8N_e$ and $4N_e$, if p is between .5 and $1/(2N_e)$. If N_e is 10,000, $\bar{t}_1$ is less than 40,000 generations, which is a short period, because the generation time in bacteria is very short. Therefore, the *lac Y* gene could have become lost quickly in *S. dysenteriae* if a fairly severe bottleneck had indeed occurred after its separation from *E. coli*.

Acknowledgments

This study was stimulated by Hall, Yokoyama, and Calhoun (1983). Thanks to B. G. Hall and M. Kricker for many suggestions. This study was supported by research grants from NSF and NIH.

LITERATURE CITED

CROW, J. F., and M. KIMURA. 1970. An introduction to population genetics theory. Harper & Row, New York.

HALL, B. G., S. YOKOYAMA, and D. H. CALHOUN. 1983. Role of cryptic genes in microbial evolution. Mol. Biol. Evol. **1**:109–124.

KIMURA, M., and T. OHTA. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics **61**:763–771.

NAGYLAKI, T. 1975. Polymorphisms in cyclically-varying environments. Heredity **35**:67–74.

STRICKBERGER, M. W. 1976. Genetics. Macmillan, New York.

WHITTAM, T. S., H. OCHMAN, and R. K. SELANDER. 1983. Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Nat. Acad. Sci. USA **80**:1751–1755.

ROBERT K. SELANDER, reviewing editor

Received July 22, 1983; revision received September 21, 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 1½ 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.

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 (χ^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.

(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:

BARRIE, P. A., A. J. JEFFREYS, and A. F. SCOTT. 1981. Evolution of the β-globin gene cluster in man and the primates. J. Mol. Biol. **149:**319–336.

Books:

INGRAM, V. M. 1963. The hemoglobins in genetics and evolution. Columbia University Press, New York.

Book chapters:

HALL, B. G. 1983. Evolution of new metabolic functions in laboratory organisms. Pp. 234–257 *in* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer Associates, Sunderland, Mass.

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 4¾ inches wide by 7⅝ 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.

Editor in Chief

Walter M. Fitch
Department of Physiological Chemistry
University of Wisconsin—Madison
1300 University Avenue
Madison, Wisconsin 53706

Managing Editor

Masatoshi Nei
Center for Demographic and Population Genetics
University of Texas Health Science Center at Houston
P.O. Box 20334
Astrodome Station
Houston, Texas 77025

Associate Editors

Roy J. Britten
Kerckhoff Marine Laboratory
California Institute of Technology
Corona del Mar, California 92625

Wesley M. Brown
Department of Cellular and Molecular Biology
University of Michigan
Ann Arbor, Michigan 48109

Richard B. Flavell
Department of Cytogenetics Plant Breeding Institute
University of Cambridge
Cambridge C82, 2LQ England

Ken W. Jones
Department of Genetics
University of Edinburgh
West Mains Road
Edinburgh EH9, 3JN Scotland

Richard K. Koehn
Department of Ecology and Evolution
State University of New York at Stony Brook
Stony Brook, New York 11794

Robert K. Selander
Department of Biology
University of Rochester
Rochester, New York 14627

Canadian Journal of Genetics and Cytology

EDITOR: **Peter B. Moens**, Department of Biology, York University, 4700 Keele Street, Downsview, Ontario, Canada M3J 1P3

The *Canadian Journal of Genetics and Cytology* publishes papers in applied and basic genetics and cytology. Traditionally a substantial number of the articles are of agricultural significance. In this respect the Journal enjoys international contributions and a large readership. Basic research articles are in the fields of molecular genetics, population genetics, mutagenesis, and chromosome structure and behaviour.

Indexed by *Biological Abstracts*, *Current Contents*, *Science Citation Index*, *Chemical Abstracts*, *Biological and Agricultural Index*, *CAB Abstracts*, and others

Canadian Journal of Biochemistry and Cell Biology

EDITORS: **Morris Kates and J. M. Neelin**, Canadian Journal of Biochemistry and Cell Biology, Faculty of Science and Engineering, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5

The *Canadian Journal of Biochemistry and Cell Biology*, formerly the *Canadian Journal of Biochemistry*, publishes papers in any field of general biochemistry and in experimental cell biology. These include full-length papers and rapid communications reporting original work or invited reviews on topical subjects, as well as certain symposia and special issues dedicated to a particular subject or occasion. Recent special issues include papers devoted to "Gene Replication and Expression: Organization and Function of Nucleus, Chromosomes, and DNA" (March 1982), and "Mechanism of Hormone Action" (July 1983); collections of symposia papers include "Biochemical Evolution of the Translation Apparatus" (April 1982), "Gene Regulation During Heat Shock" (June 1983), "Basement Membrane" (late 1983), "Calmodulin" (late 1983), and "Glycoprotein Biosynthesis" (late 1983). It is one of the leading international journals on general biochemistry and cell biology, and attracts subscriptions from more than 72 countries.

Indexed by *Biological Abstracts*, *Current Contents*, *Science Citation Index*, *Chemical Abstracts*, *Excerpta Medica*, *Index Medicus*, and others

1984 Subscription Rates (12 issues)

Canadian Journal of Genetics and Cytology

Institutional

CANADIAN: $68.00
FOREIGN: $78.00
SINGLE COPY: $12.50

Personal

CANADIAN: $31.00
FOREIGN: $41.00
SINGLE COPY: $7.50

Canadian Journal of Biochemistry and Cell Biology

Institutional

CANADIAN: $82.00
FOREIGN: $97.00
SINGLE COPY: $12.50

Personal

CANADIAN: $31.00
FOREIGN: $46.00
SINGLE COPY: $7.50

Payment must be enclosed with order

Cheques should be made payable to the Receiver General for Canada, credit NRCC

Send to:
Distribution, R-88
(Can. J. Genet. Cytol.) or
(Can. J. Biochem. Cell Biol.)
National Research Council of Canada, Ottawa, Ontario, Canada K1A 0R6

National Research Council Canada
Conseil national de recherches Canada

BIOCHEMICAL GENETICS

editor: **Hugh S. Forrest**, *University of Texas at Austin*

Reflecting the extraordinary scope, dynamics, and progress that presently characterize the field of modern biochemical genetics, this journal offers an interdisciplinary forum for the presentation of the latest developments in the field and provides the essential link between the sciences of biology, chemistry, and genetics. It presents original papers, letters, and occasional review articles on fundamental theoretical and experimental research in the biochemical genetics of all organisms, from virus to man.

CONTENTS: (Vol. 21, Nos. 7/8)

Genetics of alkaline phosphatase of the small intestine of the house mouse (*Mus musculus*). Genetic mapping and characterization of aldehyde oxidase of *Anopheles albimanus* (diptera: culicidae). Linkage relationships among the malic enzyme, hexokinase-1, and red loci on the X chromosome of *Tribolium confusum*. Inheritance, intracellular localization, and genetic variation of phosphoglucomutase isozymes in maize (*Zea mays* L.). Postnatal changes in canine erythrocyte pyruvate kinase isozymes. Genetic localization and sequential electrophoresis of glucose-6-phosphate dehydrogenase in *Drosophila melanogaster*. Genetics of insect hemolymph α-mannosidase in the silkworm, *Bombyx mori*. Characterization of a low-activity allele of $NADP^+$-dependent isocitrate dehydrogenase from *Drosophila melanogaster*. Human liver alcohol dehydrogenase: $ADH_{Indianapolis}$ results from genetic polymorphism at the ADH_2 gene locus. Implication of *Triticum searsii* as the B-genome donor to wheat using DNA hybridizations. Purification and properties of genetic variants of mouse trypsinogen. Genetics of two tissue esterase polymorphisms (Est-4 and Est-5) in the rabbit. The phosphorylase kinase activity of hearts from phosphorylase kinase-deficient mice. Hemoglobin bali (bovine): β^A 18 (B1) Lys ⟶ His: one of the "missing links" between β^A and β^B of domestic cattle exists in the bali cattle (Bovinae, *Bos banteng*). Svp-3, a third polymorphic locus for mouse seminal vesicle proteins. Linkage relationships of peptidase-7, *Pep-7*, in the mouse. Linkage analyses and biochemical genetics of sorbitol dehydrogenase-1 (*Sdh-1*) in the mouse. Hidden breaks in the ribosomal RNA of phylogenetically tetraploid fish and their possible role in the diploidization process.

Subscription: Volume 22, 1984 (12 issues) $295.00

Write to the Sample Copy Dept. for your free examination copy!

PLENUM PUBLISHING CORPORATION
233 Spring Street, New York, N.Y. 10013

If the key word is genetics—

behavioral *genetics*
clinical *genetics* and counseling
immuno*genetics*
molecular *genetics*
population *genetics*
cancer *genetics*
developmental *genetics*
biochemical *genetics*
cyto*genetics*
mutagenesis/somatic cell *genetics*

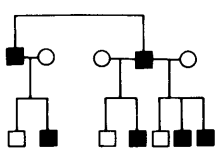
—the key research is in

The American Journal of Human Genetics

Established in 1948, the **AJHG** is one of the leading journals addressing central issues in contemporary genetics. The **Journal**—which is the official publication of The American Society of Human Genetics—provides a record of research and review relating to heredity in humans and to the application of genetic principles in medicine, psychology, anthropology, and the social sciences as well as in related areas of molecular and cell biology.

Editor: **Dr. David E. Comings.**

Unlock the information you need:



Published bimonthly by **The University of Chicago Press.** 1-year rates: Individuals $75; Institutions $95. Outside of the U.S.A. postage surcharges vary according to destination. Visa and MasterCard accepted. Mail complete charge card information, payment, or purchase order to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, IL 60637. **Please note:** Subscriptions entered on a calendar-year basis only. Students accepted as members of the society only. For student and all other membership rates, write to the Executive Office, American Society of Human Genetics, P.O. Box 6015, Rockville, MD 20850.

12/83

Subscribe to the AJHG.

Join!

THE AMERICAN INSTITUTE OF BIOLOGICAL SCIENCES

AIBS is the one organization whose sole purpose is to represent the biology profession as a whole through

- *BioScience*
- Scientific meetings
- Special Science Programs
- Public Responsibilities Program
- *Forum*
- Education

Through membership, you will receive *BioScience*, the monthly professional magazine for biologists. By reading *BioScience* it is possible to keep abreast of where we are and where we are going through

- Disciplinary and interdisciplinary articles
- Features and News/People and Places
- Reports of current research in biology
- Grants and awards
- Books available to members at discount prices
- Book reviews and new title lists
- Professional opportunities

Just complete this form and send to us with your payment. We will take care of the rest.

YES! I would like an AIBS Membership.

Name _____

Address _____

_____

_____

(Zip Code)

MEMBERSHIP DUES

☐ Individual @$32.50 per year

☐ Sustaining @ $47.00 per year
(Includes $20.00 per year for *BioScience*)

☐ Student @$17.00 per year

☐ Emeritus @$17.00 per year
(Includes $10.50 per year for *BioScience*)

Make checks payable to AIBS. Please remit in U.S. dollars only.
Mail to: AIBS, Box 9197, Arlington, VA 22209

AIBS

Another Munksgaard Journal

Clinical GENETICS

An International Journal of Genetics in Medicine

Editors

KÅRE BERG, Oslo
JAN ARVID BÖÖK, Uppsala
JAN MOHR, Copenhagen

Clinical Genetics publishes papers describing original research and new developments concerning genetics and clinical medicine. Papers are invited that deal with genetic etiology of diseases or defects, genecontrolled pathogenesis, biochemical genetics including inborn errors of metabolism, immunogenetics, pharmacogenetics, cytogenetics, population genetics, and genetic epidemiology, research methods in human genetics, and practical applications of genetics in diagnostics, counceling, and legal medicine. Case reports are invited if they bring new information on already known disases or report the discovery of inherited conditions. Comments on published papers may be accepted as Letters to the Editors.

Subscription

Clinical Genetics is published monthly. The price per year in 1983 for 2 volumes of 6 issues is D. kr. 1440,00 plus postage D. kr. 108.00 (total US $ 185.80, £ 108.40, DM 464.00) payable in advance. Please add D.kr. 60.00 ($ 7.20) for air freight to North American subscribers.
Prices are subject to exchange-rate fluctuations.

Please order from the Publisher at the address below – or through any bookseller.

MUNKSGAARD
International Publishers Ltd.
35 NÖRRE SÖGADE P.O. BOX 2148 DK-1016 COPENHAGEN K DENMARK

MOLECULAR BIOLOGY AND EVOLUTION

Volume 1, Number 1 December 1983

Two fields coming together

The almost daily discoveries of new facets of the genome have brought about a growing need for communication between molecular biologists and evolutionists. **MBE** will publish papers that critically examine the evolutionary significance of macromolecules—

- mechanisms of mutational change
- processes of developmental control
- maintenance of genetic polymorphism
- mechanisms of natural selection
- phylogenetic relationships of organisms
- theories that integrate these various aspects of molecular study

Walter M. Fitch, Editor-in-Chief

Sponsored by the Molecular Biology and Evolution Society and the American Society of Naturalists

Published bimonthly

Charter rates: ☐ Individuals $25 ☐ Institutions $50
☐ Students $20 *Valid until April 1, 1984.* **(H)**
Add $4.00 for subscriptions mailed outside the USA.

☐ Please send more information on **MBE.**

Name _____
Address _____
City _____ State _____ ZIP _____

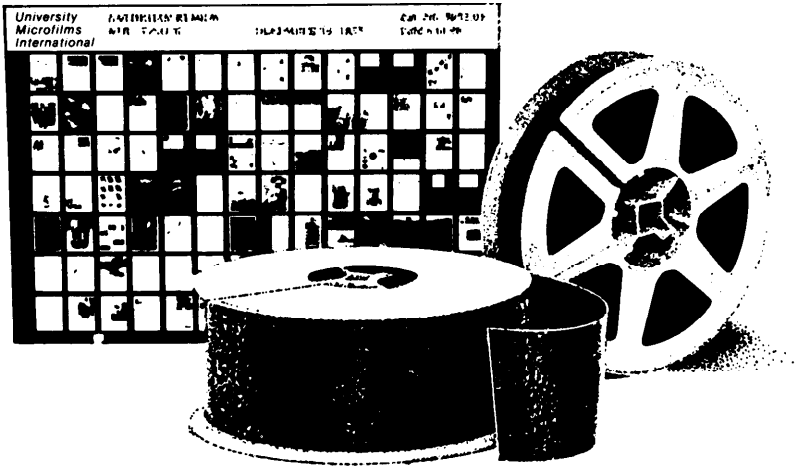
Visa and MasterCard accepted. Please mail this coupon with complete charge card information, purchase order, or payment to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, IL 60637.

11/83

L04MB

NEW FROM THE UNIVERSITY OF CHICAGO PRESS

this publication is available in microform



Please send me additional information.

Name _____

Institution _____

Street _____

City _____

State _____ Zip _____

University Microfilms International

300 North Zeeb Road
Dept. P.R.
Ann Arbor, MI 48106
U.S.A.

18 Bedford Row
Dept. P.R.
London, WC1R 4EJ
England

Published By The Biochemical Society, London

BIOSCIENCE REPORTS

VOLUME 4 OUT IN 1984

Short Papers & Reviews in Molecular & Cellular Biology

FAST

- Submission of manuscript directly to a member of the Board of Editors
- Decision in 2-3 weeks
- Publication 2 months thereafter

INTERNATIONAL

- Editorial offices in London and Houston
- Editors in North America and Europe
- Authors worldwide

QUALITY PRODUCTION

- Typeset on word processor in uniform format
- Printed on high-grade paper for clear halftones

HONARARY ADVISORY EDITORS

A Kornberg, Stanford
R R Porter, Oxford
F Sanger, Cambridge

EUROPE

W J Brammar, Leicester
A A Eddy Manchester
J R Griffiths, London
F W Hemming, Nottingham
H-D Klenk, Giessen
V Mutt, Stockholm
P-P Slonimski, Gif-sur-Yvette
R J P Williams, Oxford
A R Williamson, Greenford

BOARD OF EDITORS

C A Pasternak (Chairman), London
William J Lennarz (Deputy Chairman), Houston

NORTH AMERICA

Ralph A Bradshaw, Irvine
Gordon H Dixon, Calgary
Russell F Doolittle, La Jolla
Donald M Engelman, New Haven
Elias Lazarides, Pasadena
Hans J Müller-Eberhard, La Jolla
Alan Peterkofsky, Bethesda
Karl A Piez, Palo Alto

Subscriptions:

£95.00 US$215 for 12 issues to
The Biochemical Society Book Depot PO Box 32 Commerce Way Colchester [illegible] UK

Information to authors, sample list of contents, and other information:

Ms Suzanne Miller Editorial Secretary Bioscience Reports Department of Biochemistry St George's Hospital Medical School Cranmer Terrace London SW17 ORE, UK Telephone 01 672 1255 x 5010 international telephone +44 1 672 1255 Telex 945292 SAGEMS G

Dr William J Lennarz Department of Biochemistry and Molecular Biology University of Texas System Cancer Center M D Anderson Hospital and Tumor Institute Texas Medical Center 6723 Bertner Avenue Houston TX 77030 USA Telephone [illegible] Telex [illegible]

Forthcoming

Individual and Evolutionary Variation of Primate Ribosomal DNA Transcription Initiation Regions
Golder N. Wilson, Mechthilde Knoller, Roy D. Schmickel, and L. Lynne Szura
The Relationship between Codon Boundaries and Multiple Reading Frame Preferences; Coding Organization of Bacterial Insertion Sequences
David J. Galas and Temple F. Smith
Calibrating the Molecular Clock: Estimates of Ground Squirrel Divergence Using Fossil and Geological Time Markers
David Glenn Smith and Richard G. Coss
Estimation of Evolutionary Distance between Nucleotide Sequences
Fumio Tajima and Masatoshi Nei
Silencing of Duplicate Genes: A Null Allele Polymorphism for Lactate Dehydrogenase in Brown Trout (*Salmo trutta*)
Fred W. Allendorf, Nils Ryman, and Gunnar Stahl
Immunological Similarities between Specific Chloroplast Ribosomal Proteins from *Chlamydomonas reinhardtii* and Ribosomal Proteins from *Escherichia coli*
Robert J. Schmidt, John E. Boynton, Nicholas W. Gillham, and Alan M. Myers
Molecular Probes of Phylogeny and Biogeography in Toads of the Widespread Genus *Bufo*
Linda R. Maxson
Molecular Evolution of Chloroplast DNA Sequences
Stephanie E. Curtis and Michael T. Clegg
Review of *Macromolecular Sequences in Systematic and Evolutionary Biology*, Morris Goodman, ed.
Linda Maxson
Review of *Evolution of Genes and Proteins*, Masatoshi Nei and Richard K. Koehn, eds.
Francisco J. Ayala
Review of *Statistical Analysis of DNA Sequence Data*, B. S. Weir, ed.
Aravinda Chakravarti

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 Arnheim
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 F. 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

February 1984, Volume 1, Number 2

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.

© 1984 by The University of Chicago. All rights reserved.