ENZYME EVOLUTION:

I. THE IMPORTANCE OF UNTRANSLATABLE INTERMEDIATES

ARTHUR L. KOCH

Department of Microbiology, Indiana University, Bloomington, Indiana 47401

Manuscript received March 6, 1972

Revised copy received June 10, 1972

ABSTRACT

The proportion of two allelic mutant forms of a haploid species can be calculated from six quantities, the original proportion, the two effective growth rate constants, the two mutation rate constants for the transition from one to the other, and the elapsed time. With a program for carrying out this calculation in a mini-computer, we set about investigating the rates of exploration of multiple intragene changes in populations of enormous numbers over a large number of cell generations. It seemed evident from the outset that unless in the past very different proportions of the different types of mutational changes occurred, then neither the numbers nor the times could have been of sufficient magnitude to account for enzyme evolution by multiple simultaneous changes within a gene to yield a superior protein. Rather, we propose that the more rapid path of enzyme evolution consisted of the sequence: 1) gene duplication, 2) inactivation of a gene copy by a chain-terminating mutation or other mutational change which suppresses transcription and/or translation, 3) multiple base-pair changes and/or small deletions and additions within the non-functioning gene, and finally, 4) reversion of the mutation which was suppressing translation. It is shown that this path will lead to faster advance, if the selective advantages of different forms are dependent on the function of the gene products at certain times in history, while at other times, although this function may be necessary, it is not limiting for growth.

THE first living creatures on this planet must have endowed their descendants with catalysts that bind poorly and cause chemical changes very slowly, compared with the abilities of the enzymes found in today's organisms. Here we propose to study the kinetic processes important during that much earlier "Epoch of the Enzymes".

During this time, archaic protista must have undergone an extremely large number of divisions during which competition was dependent on enzymatic capability needed for anabolic and catabolic processes, and not on the myriad of more esoteric processes functioning later in evolution, such as those needed for dormancy, mobility, etc.

The suggestions made here as to the processes effective during the phase of enzyme evolution are similar to those important for the development of species later on, but differ importantly in regard to detail. These differences are due primarily to the inefficiency or absence of processes leading to recombination within a cistron, and therefore the necessity of accumulating multiple mutations

within a gene. The differences are also due to the large number of individual haploid organisms involved in the selective process.

OUTLINE OF THE PROPOSED EVOLUTIONARY PROCESS

When circumstances both within and without an organism have evolved to a point where the functional activity of one type of protein is limiting for growth, a variety of processes can be called into action. These can range from profound mechanisms that cause a decrease in the multiplication rate of the species (a process humans have not achieved, although mice have), to complicated regulatory mechanisms at the physiological level of induction or derepression, to some basic mutational change of genetic material coding for the protein whose function limits growth. In the laboratory today (HEGEMAN and ROSENBERG 1970; KOCH 1971) and presumably in nature today or in the precambrian era, limitation will select for constitutive mutations in repressor and operator genes which cause the organisms to produce high levels of the gene product in limiting supply. Simultaneously and/or subsequently, selection will act to favor improvement of the kinetic constants by all possible single simple point changes and single minor deletions or insertions that individually improve the function of the protein. Once the succession of such changes had led to the point where further improvement through any single change was not possible, subsequent progress would be slow or virtually halted because it would require multiple, simultaneous changes.

However, the process of gene duplication is also available to the organism. Duplication serves as an immediate way to augment the amount of any limiting protein species, and no matter how rare duplication events are, organisms bearing multiply genes will be selected and will become abundant in the population. In the long term, this is a dangerous solution to the biological problem because organisms possessing multiply genes for a given function face genetic stasis. They are at an impasse in that any further qualitative improvement in any one of the gene copies will no longer afford the organism with a high selective advantage. Therefore, organisms following this route may find themselves in an evolutionary *cul-de-sac*. Nonetheless, duplication provides a mechanism whereby multiple copies of a gene for a limiting function co-exist in individual organisms. At some later time, the function in question may be no longer limiting either because of changes in the environment or other genetic changes within the organism. Then, either one of the gene can serve the passive role of the recessive gene of eucaryotic genetics and be subject to non-selective "non-Darwinian" mutational changes. Thus, one of the gene copies may be subject to deletion, to elaboration to serve some different function or to be altered into a form which is superior in the original function. In the mathematical part of this paper, formulae are developed and then numerical computations presented that give a qualitative view of the relative speed at which deletions would eliminate a duplicate gene from the population compared to the rate at which further changes in the polypeptide structure would be explored. If a potentially superior form of the gene arises and if it is present at the propitious time when its protein product is again

critical to the organism, translatable revertants will outgrow and supplant their antecedent form.

Once this has happened, the original form of the gene becomes vestigial. Since it is no longer important for maintaining the enzymatic function, it would inevitably be deleted. It is proposed that this complex sequence of mutational and selection processes is kinetically much faster than multiple simultaneous changes within one organism. It is furthermore proposed that this process serves in haploid evolution the role of speeding the process that intermediate complex formation serves in chemical or biochemical kinetics. Ternary, quaternary, and higher simultaneous events are rare, so if mechanisms exist to conserve binary events long enough in the form of a complex to engage in another event, the rate of the overall process is greatly enhanced and dominated by that path, compared to the rate expected if the simultaneous multiple process be the only mechanism.

THE MATHEMATICS OF HAPLOID EVOLUTION

The following section is largely paraphrased and extended from the literature on bacterial growth, on mutation rates in microorganisms, and on continuous culture of microorganisms.

Imagine that there are at time t , N organisms of one kind and M organisms of a second kind; that these grow with effective growth rate constants, λ_n and λ_m respectively; that they mutate at rates μ_m and μ_n to convert $N \rightarrow M$ and $M \rightarrow N$ respectively; then we can write:

$$\frac{dN}{dt} = \lambda_n N + \mu_n M - \mu_m N \quad (1)$$

and

$$\frac{dM}{dt} = \lambda_m M + \mu_m N - \mu_n M \quad (2)$$

The effective growth rate constants, λ_n and λ_m can each be split into a sum of terms: the reproductive growth rate or birth process, β ; a death process, δ ; and an emigration process, ϵ . Oftentimes the birth process is the dominant term, but even more often, the positive process will be almost counter-balanced by the death process. However in general, we can write:

$$\lambda_n = \beta_n - \delta_n - \epsilon_n \quad (3)$$

and

$$\lambda_m = \beta_m - \delta_m - \epsilon_m \quad (4)$$

With only two provisos, these four equations describe completely the way the two populations will change. Firstly, immigration into the population has been neglected, but could be introduced. Secondly, the equations apply only to a large enough population to be treated "deterministically". The equation would also apply to the mean behavior of a collection of smaller populations. But it will be most useful to us to apply this treatment for those situations where mutations are taking place often enough to have a reasonably small sampling error.

If for a period of time, all the rate constants, identified by greek letters, are constant, then the system of differential equations can be solved yielding equations of the form:

$$N = A + Be^{at} + Ce^{bt} \quad (5)$$

and

$$M = A' + B'e^{at} + C'e^{bt} \quad (6)$$

where a and b are the roots of the quadratic equation:

$$R^2 - (\lambda_n + \lambda_m - \mu_n - \mu_m) R + \lambda_n \lambda_m - \lambda_m \mu_n - \lambda_n \mu_m = 0 \quad (7)$$

The constants A , A' , B , etc. can be evaluated with only two additional pieces of information, i.e., N_0 and M_0 , the numbers of the two types at the time taken as zero.

This evaluation is tedious and leads to bulky expressions for the constants. The equations themselves do not allow one to grasp how selection is affected as the rates of the different mutations and growth processes are altered. Moreover, the really important variable is not the absolute number of individuals, but the ratio of the two types:

$$X = \frac{M}{N} \quad (8)$$

Studying how X changes allows us to encompass uncontrolled growth, continuous but limited growth, and even periodic catastrophies that lower the population of both types indiscriminantly. Therefore, equations 1 and 2 are rewritten as a single equation in X .

$$\frac{dX}{dt} = d\left(\frac{M}{N}\right)/dt = \frac{NdM - MdN}{N^2 dt} = \frac{(\lambda_m M + \mu_n N - \mu_n M)N - (\lambda_n N + \mu_n M - \mu_m N)M}{N^2 dt} \quad (9)$$

$$\frac{dX}{dt} = \frac{-\mu_n M^2 + (\lambda_m - \mu_n - \lambda_n + \mu_m) MN + \mu_m N^2}{N^2 dt} = -\mu_n \left[X^2 - \frac{(\lambda_m - \mu_n - \lambda_n + \mu_m)}{\mu_n} X - \frac{\mu_m}{\mu_n} \right] \quad (10)$$

This equation may be solved to yield a single equation describing how the proportions of the two types change. If either of μ_m or μ_n is zero, the equation greatly simplifies. This would be the case if either type of mutational event were sufficiently rare or if mutation in one direction was a more likely process than that in the other. This is frequently the case and has to be the case when the change considered is a deletion mutation. The solution of equation 10 is effected as follows: consider the expression in brackets on the right hand side of equation 10 as an algebraic expression. Set it equal to zero and solve for X . Let the two roots of the quadratic be r_1 and r_2 . Then, rewrite equation 10:

$$\frac{dX}{dt} = -\mu_n (X - r_1) (X - r_2) . \quad (11)$$

Therefore,

$$\frac{dX}{(X - r_1) (X - r_2)} = -\mu_n dt .$$

This equation can be integrated by the artifice suggested by the rule in algebra for the addition of fractions. This procedure, operated in reverse, allows one to decompose an expression into two "partial fractions". The left hand member can be rewritten as the difference of terms, each containing a single factor in X :

$$\frac{dX}{(r_2 - r_1) (X - r_2)} - \frac{dX}{(r_2 - r_1) (X - r_1)} = -\mu_n dt . \quad (12)$$

After cross multiplying by $(r_2 - r_1)$, each term can be integrated, and the constant of integration evaluated by the fact that $X = X_0$, when $t = 0$. This results in:

$$\ln \frac{X - r_2}{X_0 - r_2} - \ln \frac{X - r_1}{X_0 - r_1} = -\mu_n (r_2 - r_1) t \quad (13)$$

By taking antilogarithms, this can be rewritten as:

$$\frac{X - r_2}{X - r_1} = \frac{X_0 - r_2}{X_0 - r_1} e^{-\mu_n (r_2 - r_1) t} . \quad (14)$$

A computer program was set up which, given λ_n , λ_m , μ_n , μ_m , X_0 and t , solves the quadratic, then substitutes the roots into equation 14 and solves for X . It also automatically takes care of the situation where $r_1 = r_2$ (and equation 14 breaks down), and of certain other possible contingencies; then the computer uses the equation appropriate to these cases,

$$\frac{1}{X_0 - r} - \frac{1}{X - r} = -\mu_n t . \quad (15)$$

AN ILLUSTRATIVE NUMERICAL EXAMPLE

Evolution of Enzymes in Enteric Organisms: I chose to take mutational rate data relevant to *Escherichia coli* studied under laboratory conditions, and apply it to the evolution of an enzyme in an enteric bacterium under conditions that might be typical of conditions under which this organism evolved to become adapted to its mammalian hosts. I partitioned the estimate of the total mutation rate into a variety of components, based on simple statistical properties of the genetic code and the experimental data concerning different types of mutations observed during growth of coliphage T4 in the production of spontaneous mutations in the *rII* cistrons. Arbitrarily, I assumed that the sum of the processes causing the duplication of at least a complete gene are one-tenth as frequent as major deletions of a typical gene. Some mechanisms that have been discussed would appear to lead equally to both deletions and additions; some mechanisms uniquely produce deletions. A deletion, to be effective, need only cover a portion

TABLE 1

Estimates of quantities needed for the calculations of the in entero course of evolution of an enzyme of Escherichia coli

	Collective molecular weight	Fraction of total dry weight
Composition of growing organisms		
DNA (5×10^6 base pair per genome)	5×10^9 daltons	3.3×10^{-2}
Total dry weight	1.5×10^{11} daltons	1
Composition of hypothetical gene		
DNA (1,000 base pairs)	7×10^5 daltons	6×10^{-6}
Gene product (3.7×10^4 daltons each times 2×10^4 copies per cell)	7.4×10^8 daltons	5×10^{-3}
Mutational rates of the hypothetical gene		
All base pair substitutions (including chain terminations)	4.7×10^{-6} per generation	
All chain terminations	6.2×10^{-8} per generation	
All short deletions and additions	2.5×10^{-7} per generation	
All long deletions	4.8×10^{-8} per generation	
All long additions yielding duplication	4.8×10^{-9} per generation	
Doubling time	0.5 day	

Compositional data have been taken from McQUILLEN (1965) and KOCH (1971). The natural doubling time in a mammalian intestine is taken from GIBBONS and KAPSIMALIS (1967). It is assumed that protein synthesis is continuous and that any gaps in DNA synthesis are such that G_1 is about equal to G_2 . The mutational data are taken from DRAKE (1970). He gave an average value of 4×10^{-10} observable mutations per base pair per replication. Based on his discussion of spontaneous mutations in the T4 rII system of BENZER (1961), the following proportions have been assumed:

observable amino acid substitutions: 27::
 chain termination events : 44::
 short additions and deletions : 179

It has also been assumed, based on data with the T4 rII system, that 12% of the observable mutations are long deletions. As mentioned in the text, duplication to yield a functional gene is assumed to be 0.1 as probable as long deletion events that eliminate production of functional gene product. On the approximation that all codons are equally probable and that all base pair substitutions are equally probable, it can be calculated that chain termination events constitute:

$$\frac{7 + 8 + 8}{3 \times 3 \times 3 \times 64} \times 100 = 1.33\%$$

of the total base pair changes. With this value, the mutation rate for all base pair substitutions (including chain terminations) is given by:

$$4 \times 10^{-10} \frac{\text{Observable changes}}{\text{B.P.} \times \text{Generations}} \times \frac{1000 \text{ B.P.}}{\text{Gene}} \times \frac{44 \text{ Chain-terminations}}{(27+44+179) \text{ Observable non-deletions}} \times$$

$$\frac{(100-12) \text{ Observable non-deletion}}{100 \text{ Observable changes}} \times \frac{100 \text{ Base-pair-changes}}{1.33 \text{ Chain-terminations}} =$$

$$4.7 \times 10^{-6} \frac{\text{Base-pair-changes}}{\text{Generation} \times \text{Gene}}$$

Other assumptions (mentioned in text) are that the "hot spots" observed in the spontaneous spectra of T4 rII are due to an increase in number of certain small deletion or addition mutations, and that slower growth such as occurs under natural conditions does not alter the mutation rate when expressed on per generation basis.

of a gene, but a duplication, to be effective, must cover the entire gene and possibly some associated controlling elements such as operators and promoters. Another major arbitrary choice is that I assumed that the higher spontaneous rate of mutation in T4 *rII* cistrons is due to the higher frequency of occurrence of mutations at the so-called "hot spots". These are known to be small additions and deletions (DRAKE 1970), and I have arbitrarily assumed that this extraordinary abundance is decreased in the bacterium to the rates typical for other micro-additions and deletions. The final arbitrary choice for this calculation is to assume that the mutation rate per generation observed in the laboratory applies equally well to more slowly growing cultures associated with growth in the mammalian alimentary canal. The citations of the review literature and the detailed calculations of these values and the other needed parameters are given in Table 1 and its footnotes. Although better estimates of all these quantities are needed, and the assumptions stated above should be replaced with experimental fact, the important conclusions do not depend on the numerical values of the mutation rates except where stressed below.

From these values and some assumptions about growth rates, the time course of M/N can be calculated. Below, it is usually assumed that the specific growth rate is inversely proportional to the dry weight (biomass) of an organism, when organism growth is limited by some other nutrient than the substrate of the enzyme under consideration, i.e. the burden to the organism synthesizing an unneeded protein is proportional to the extra synthesis of macromolecules. When one nutrient is limiting to growth it is assumed that growth is proportional to the content and activity of the critical enzyme.

Another case is also being considered below; this is designated "nutrient non-limiting" and could arise if there are control mechanisms that lead to the synthesis of reserve substances (such as polysaccharide and unused ribosomes) when excess nutrients are available. Such synthesis might have a long-term advantage, but a short-term disadvantage to the cell. However, such a possibility leads to growth independent of the amount of unneeded proteins made by the cell, if they do serve as reserve material so that less of the cell's capacity is used for the usual reserve substance. This will be further discussed below.

The first question to be considered is how fast a duplication would overtake the prototype once the function of the enzyme became growth-limiting. It can be seen from Figure 1 that no matter how infrequent mutational events or how rare mutants in the population were, at the time that enzyme function becomes limiting, organisms with duplicate genes would be established in the population in an extremely short time—a time to be measured only in a few weeks. The various panels include, besides the precise case corresponding to Table 1 (shown in Figure 1a); the effect of decreasing the frequency of events producing the duplicating to zero (Figure 1b), or decreasing the deletion mutation rate to zero (Figure 1c), or both (Figure 1d).

Even if the production rate of duplications is very low or even zero, an initial frequency of 1 in 10^{10} is enough to give rise to a population of 50% duplicates in 16 days, 1 in 10^{20} would do the same in 36 days, 1 in 10^{30} in 56 days for the con-

FIGURE 1.—The accumulation of cells containing duplication of genes for a limiting function. The time course of the replacement of the original monogenic population by cells with a duplication of the limiting gene is shown as a function of X_0 , μ_m and μ_n . In Figure 1a, μ_m and μ_n , the values from Table 1, have been used. The calculation assumes that back-mutation by deletion could occur in either duplicate. The cells bearing two genes are assumed to grow twice as fast, initially doubling at 4 times per day. As the population bearing a duplication increases in the total population, the limiting nutrient in the environment becomes further depleted. Then both growth rates (λ_n and λ_m) would decline so that the total population doubles two times in one day. For this reason, the computer program was modified so that λ_n and λ_m are suitably scaled down, progressively over a factor of two, as the population changes character.

ditions of Figure 2d. Rarer events are not relevant because 10^{30} is large compared to the number of *E. coli* in the world at any one time. Clearly the speed of the takeover is dependent almost entirely upon the assumed selective advantage of two fold.

When the selection is relaxed, the duplication will be eliminated from the population because of the slight selective advantage of the organism with only one copy of the gene in comparison to those with more copies. The saving in the synthesis of no longer needed gene product and of the gene itself is a small factor compared with the selective advantage producing the population containing duplicate genes but still a highly important one. A number of simultaneous processes of mutation and selection would come into play. These would include selection for a series of deletions. Eventually of course, all extra DNA will be eliminated.

For the moment, we consider the entire class of deletions combined together and assume that the average deletion saves half the cost of DNA production of

FIGURE 2.—The loss of cells bearing duplications when gene function is no longer limiting. The mutation rates were calculated from Table 1 assuming either gene copy may undergo deletion for those cases where no prior mutation that had inactivated one copy has been assumed. It was assumed for Figure 2a that the deletion has a selective advantage proportional to the savings of 50% of the DNA of the gene and $\frac{3}{4}$ of the protein coded by that gene (see text). Figure 2b is based on the assumption that a chain termination event previously had taken place and become dominant in the population. It is only that population which is considered. The chain termination itself was presumed to have saved half of the cost of unnecessary protein synthesis. In Figure 2c, it is assumed that the cells have avoided all the cost of protein synthesis; this would be the case if protein degradation of incomplete proteins were a highly efficient process yielding amino acids for resynthesis. The curves in Figure 2d are for the nutrient non-limited case where extra DNA and protein do not affect the growth rate, and mutation pressure is the only factor. It must be stressed that the calculations in the figure apply approximately to populations composed of deletion mutations covering ranges of starting and stopping points and to chain-termination events taking place at all sites on either gene. The average selective effect has been assumed in the calculations, but as pointed out in the text, certain classes of such mutations will have much stronger selective pressures and will replace other classes, and the final population will be eventually enriched in the most economical variety.

the gene and three-fourths the cost of protein production. The figure of three-fourths is assumed since half of the mutants deleting a sizable stretch of DNA will not make any protein at all because the operator regions have been deleted, and the remaining mutants will make only fractions of the total. If the termini of the deletions are evenly spaced, those that make protein at all, on the average would make half. Thus, on the average, one-fourth of the amount of polypeptide will be synthesized initially in the populations of newly produced deletion mutations.

Either gene duplicate could be eliminated, but the one added by the duplication event to the genome is more likely to be deleted than the copy located in its original position. This is because the inserted DNA had to be larger than the functional gene and on the average would contain part of adjacent genes. Thus, there is a larger distance between vital functions on the average than for the original. This not only increases the chance of viable deletion being produced but also increases the selective advantage to the deletion of such duplications which may already be coding for portions of other protein molecules of no purpose to the cell.

The significant conclusion is that a considerably longer time is needed to get rid of a duplication than to incorporate it originally (Figure 2a). It takes years instead of days in the case used here for illustration. Nonetheless, the time is short enough that an unneeded function would be eliminated from an intestinal organism in but a few generations of the host in which it lives.

Other changes than deletion would have most of these same selective effects. For example, a promoter mutation or initiator codon mutation could decrease the number of transcriptional events. They, like highly polar mutations, would save the cost of translating not only the gene in question, but other genetic elements associated with it.

We must stress that promoter mutations, or mutations blocking initiation or creating new chain termination sites, can slow, but do not stop, the eventual elimination of unneeded genetic material. Figures 2b and 2c show the thousands of years needed to eliminate most of the genetic material after a random chain termination. If the deletion saves the cost of translation in part, Figure 2b applies. For the promoter mutation, chain initiation, or for the case of chain termination mutation in organisms able to recycle amino acids from incompleting protein, the calculation shown in Figure 2c applies. Figure 2d shows the loss if no selective advantage (including the cost of DNA synthesis) at all accrues after a deletion. This is for comparison and to point out that genetic elimination of this kind still occurs under "nutrient non-limited" growth.

In the next sections, we will discuss the applicability of these calculations to the evolution of *E. coli* and Precambrian organisms. But here it is necessary to point out that the DNA of tandem duplicate genes can circle back on itself in such a way that an internal recombination event can result in the elimination of an integral number of gene units. With effective recombination systems and without special mechanisms preventing such internal recombination, this should be a probable event. Consequently, the deletion rate would be higher than the one assumed in the above calculation which is consistent with data for the deletion of non-repetitious DNA. Moreover, such internal recombination deletion events can happen as often as necessary to eliminate *all* repetitious sequences. For this reason, we have carried out additional calculations to show the effect of higher deletion rates than assumed in Figure 2 on the length of time required to reduce the proportion of the population bearing duplicate genes to 50% (see Table 2).

The results of Figure 2 indicate that when a gene is no longer needed it will be

TABLE 2

Effect of increased deletion rate on the time to eliminate half the duplications

Mutation rate for deletion Fold increased*	Nutrient limiting conditions Case Figure 2a	Selection based on DNA cost & part of the protein Case Figure 2b	Selection based on DNA cost Case Figure 2c	Nutrient non-limiting conditions Case Figure 2d
1×	5.35 yr	15.46 yr	3,998 yr	11,070 yr
10×	4.14	11.82	1,288	990.5
100×	2.93	8.16	187.7	98.9
1000×	1.70	4.48	19.7	9.88

* Over the value given in Table 1 and assumed in other calculations in this paper.

omitted, or made non-functional and non-translated, initially by an admixture of deletion, promoter, initiator codon, and chain termination mutants. These multiple possibilities will lead relatively quickly to a population that has lost the gene function but has a mixed genetic character that will be quasi-stable since the subsequent mutation and selection leading to a population of full deletions of all non-essential genetic material is a very slow process.

This mixed population is the only case that I have been able to imagine with populations of haploid organisms which result in phenomena similar to the polymorphisms characteristic of natural populations of higher organisms. The quasi-steady-state proportions of the deletions, the promoter mutations, initiator codon mutations, and the chain termination mutations are dependent upon the original proportions and mutation rates in the initial phases, just after the removal of the selective pressure which had maintained the original duplication in the populations.

Consider for the moment competition between deletions and chain-termination mutations in their race to replace the duplicate gene. Since the time to eliminate half the original duplications by deletion alone (Figure 2a) is comparable to the time to eliminate half of the duplications by chain-terminated organisms alone (Figure 3), there will be about equal proportions of both in the quasi-steady-state polymorphic population, if neither type is present when selection is relaxed.

Although both mutation and selection play a role in eliminating genes for unneeded functions, the considerations of the previous section show that the selection against the synthesis of unneeded protein to be the quantitatively major factor. With an amplified revised program which can handle 3 species simultaneously, it was quickly ascertained that if we calculate the incidence of deletions which prevented translation relative to deletions that still presented the cell with a burden of protein synthesis, the former quite rapidly replaced the latter after both together had replaced the parental type. Likewise, chain terminations near the origin of protein translation quickly supplant more distal chain terminations if translation of a polypeptide fragment resulted in a burden to the cells. Because these subsets of the mutational types will arise in the long run even though they may represent only a minor contribution to the population at the time the original duplication is just removed or inactivated in the population, it is these two sub-

FIGURE 3.—Accumulation of chain termination mutations when gene function is no longer limiting. This graph corresponds to Figure 2a but assumes that instead of only deletion events, the only mechanism leading to loss of the duplicate gene function is via the accumulation of chain-termination mutations. It is also based on the approximation that the entire class of chain termination mutants have the selective advantage characteristic of a chain termination precisely in the middle of the gene. It is also assumed that reversion can only take place by base pair change at the codon for the original chain-termination mutation. This assumption is non-critical and the reversion mutation rate would have to be 10^6 times larger to give detectable deviation from the curve shown here.

classes which will co-exist for the longest period of time. We therefore chose to consider the sub-class of all those chain terminations in the first 10% of the peptide chain and the sub-class of all those deletions that eliminated the start of protein synthesis. We neglect all other classes of changes and prorate mutation rates and selective advantages for this calculation in the same way we did in the previous section. With neither type initially present, the proportion of the total population made up by this sub-class of chain-termination mutation is designated by Curve E in Figure 4. If initially chain termination mutants are present, this greatly enhances the chain-termination population as exemplified by Curve A. If initially deletions or organisms of the original type which had never possessed the duplicate are present, the maximum proportion of the chain-terminated cells in the population (Curve F) will be greatly decreased. Decreasing the deletion mutation rate increases the maximum proportion of the chain-terminated organisms (Curve C). Curve D shows the consequence if the system is "non-nutrient limited".

If mechanisms exist to degrade protein and efficiently reutilize amino acids, then a chain termination at any location will be competitive with deletions covering all or part of the gene. This case is shown in Curve B and shows that chain termination mutants as a class will accumulate and will be maintained for a long time; a time to be measured in the thousands of years. Similar to the early chain-termination mutations, promotor mutations and initiator mutations will arise in the population and will be selected during the initial phase and then remain in

FIGURE 4.—Accumulation and loss of chain-termination mutations. Starting from a population of cells containing duplicate genes at a time when selection for that gene's function is relaxed, the accumulation of chain-termination mutations in one or the other of the genes is presented. The accumulation of deletion and chain-termination mutation are competitive processes. The balance determines maximum accumulation of chain-termination mutation. Selective growth of full deletions eventually displaces the population of chain-terminated cells. Curve E is the standard case of which the others are modifications. The particular cases considered are as follows:

Curve A: Conditions as in Curve E, except that initially, there is one chain-termination containing cell per 10,000 cells bearing a duplication.

Curve B: The full cost of synthesizing unneeded protein is assumed to be saved by the chain termination mutation. This case would apply if degradation of incomplete proteins and reutilization of the amino acids are efficient processes.

Curve C: Like E, except deletion rate decreased 10-fold.

Curve D: For comparison with Curve E; only difference is that the cost of 100% instead of 95% of the protein is assumed saved by the chain termination.

Curve E: Only chain termination events in the first 10% of gene and 50% of the deletions (i.e., those covering the operator region) are considered in the calculation. The mutation rates from Table 1 are prorated accordingly. It is assumed the 95% of the cost of protein synthesis is saved by the chain termination mutations. It is assumed that initially the population consisted purely of cells containing duplication, i.e., no deletions and chain terminations.

Curve F: Conditions as in Curve E, but at time zero, there is one cell with only a single gene per 10,000 cells bearing duplicate genes.

the population for many thousands of years before they are finally eliminated by deletion.

Clearly then, there will be non-functional DNA not subject to selection which provides an opportunity for a variety of genetic changes to arise by further point mutations, frame-shift mutations, and small deletions and additions. Focusing

FIGURE 5.—Accumulation of secondary changes in a gene already rendered non-functional by a chain termination or initiator codon mutation. In Figure 5a, the accumulation of the classes of mutants containing the indicated number of changes drawn from any possible base-pair substitution or additional chain termination or short deletions or additions is shown on a double logarithmic plot against the time under non-selective conditions. In Figure 5b, the production of cells with time is shown with any particular mutant with any possible 2,3,4, or 5 base-pair specific changes of any of the three possible transitions and transversions at any of the 1,000 base-pair sites. The results are expressed on the assumption of a base population of 10^{15} individuals.

attention on this combined population of chain-termination mutations and promoter mutations, etc. the next step is to compute the extent of the accumulation in this population collectively of *all* single base pair changes taken together (see upper curve in Figure 5). The collective class of all small additions and deletions is shown as the lowest-lying curve. We also calculated the number of all single changes of both types, but it is not detectably different from the first curve. The important conclusion to be drawn from Figure 5a is that in about 100 years, every organism in the population in question will contain at least one additional mutational change.

The time course for the introduction of a chain termination is shown in Figure 5a. This is relevant because a second mutation preventing full translation would be as effective as a deletion in preventing the later reversion to protein synthesis. Clearly this process takes a very long time, one to be measured in hundreds of thousands of years.

The relevant question is: given a gene which is subject to mutational pressures but not selective pressures, how long will it take for a very special set of multiple changes to occur by chance; a special set such that after reversion to production of protein, will yield a superior protein. The number of organisms with any specified particular combinations of base pair changes are shown in Figure 5b. The computational results are based on the assumption that we are dealing with a steady-state population of 10^{15} organisms and that further mutations arise and disappear by mutational processes with no selective pressure at all. It is from this figure that the major contention of this paper is made. *The evolution of polypeptide structure is favored by long periods of time during which a gene whose function is temporarily not needed can be maintained in the population.* The longer the period of time during which function is not needed, the more complicated the renovations in the gene which the mutational processes can consider. But all will be lost if that period is too long and the inevitable deletion (Figure 2c) or, the much less probable additional mutation in either a promoter, initiator, or chain-termination site (Figure 5a) eliminate the chance that the gene may become translatable once again by a single reversion mutation.

If conditions do revert in time to those where function of the enzyme is growth limiting, then additional new duplications of the original gene will proliferate just as they did at the beginning of the previous cycle. But they will be competition with revisions at the site of the mutation that prevent translation in the modified variation of that gene. Even if reversion is very rare compared with the duplication process, from the arguments made explicit in Figure 1, it is the reversion, if it codes for a better enzyme, which will take over the population by the selective virtue of its superior maximum activity or its more avid K_m value.

Further prolongation of the limitation for function would enable the population with the improved gene to again give rise to duplicates—to initiate a new cycle of this evolutionary process. Over eons of time this process would continue leading just as they did at the beginning of the previous cycle. But they will be in competition toward the physical and chemical limits of specificity inherent in polypeptide structure and reachable by successive evolution steps, each of which embodies, at most, four or five changes from previous functional forms.

REASONABLENESS OF THIS MODEL FOR EVOLUTION OF ENZYMES IN ENTERIC ORGANISMS

The calculations presented above are consistent with data about the ecology of an enteric organism like *Escherichia coli*. For example, one has a reasonable explanation of why constitutive producers of the enzyme β -galactosidase are not found in natural populations isolated from fecal samples of adult mammals.

Nor are they retained when introduced into a mammalian intestine. No natural isolate so far checked has been constitutive for this enzyme (KOYAMA 1971). A burden of about 3% of the total protein is too large for maintenance in the population. On the other hand, we can understand why *E. coli* isolated from many non-milk drinking organisms, including particular humans, retain the inducible *lac* operon. The calculations given above show that the inducible *lac* operon should be retained over thousands of years in the absence of lactose, and therefore be retained from host generation to host generation, even if it is not useful during the long periods of no milk ingestion, typical of most adult mammals (KOYAMA 1971). We can also explain why the photo repair system is maintained in a natural population of *E. coli* even though neither its substrate (DNA treated with short ultraviolet) nor the light to drive the photochemical reaction is available in the alimentary tract.

A number of reservations must be made in applying these calculations to the evolution of enteric organisms. We have assumed that mutational rates expressed per unit time as determined in the laboratory also apply under natural conditions where growth is slower than under the usual laboratory conditions. Although NOVICK and SZILARD's (1950) original results would be consistent with this assumption, later studies by KUBITCHEK are not. These studies, summarized in KUBITCHEK (1970), show that chemostat growth at low dilution rates of *E. coli* strain B_{1t} under tryptophan limitation results in organisms that are actually short polynucleate cells in which only about one out of four genomes serves in further replication (the functional one replicates about three times per cell generation). Studies of other growth conditions (such as glucose, succinate, or phosphorus limitation) lead to the observation of mutational rate proportional to the growth rate. If the mutational processes in natural conditions take place at a rate proportional to the growth rate constant, then they will take place at faster rates per generation than have been assumed above. Under these circumstances, although selection is a less effective way to eliminate genetically irrelevant material compared to mutational pressures than we have assumed, selection, even if 12-fold less effective, would still be the dominant force in leading to the deletion or loss of translatability of an unneeded gene.

The remaining point to be considered in this section is the effect of recombinational systems catalyzing the formation of duplications and deletions of redundant genetic material from the genome. Clearly, once a tandem duplication is formed, recombination between sister identical double stranded DNA molecules and/or internal recombination within one DNA molecule permits further changes to both augment and lessen the number of replicates in the organism.

Most genes in bacteria are probably present in single copies in the chromosome (THOMAS 1970), and in the case of the ribosomal RNA genes, selective forces retain the multipicates. Although duplication is rare in microorganisms, there is the case of duplicate genes for the minor tyrosine tRNA (RUSSELL *et al.* 1970). The selective forces maintaining the duplicate must be subtle in this case since the organism grows well under laboratory conditions without one or the other duplicate.

RUSSELL *et al.* (1970) provide us with a way to estimate the relative rate of deletion *vs.* base pair changes. They find that the deletion rate is $(15/20 \times 2.2 \times 10^{-6} = 1.6 \times 10^{-6})$ /generation. This is about 50 times the value for deletion of non-repetitious genetic material assumed in Table 1. Therefore, fewer changes will occur before deletion takes place with tandem repetitious DNA than we have calculated above.

Thus, the existence of recombination mechanisms will greatly inhibit evolution by the proposed mechanism for evolutionary advance. The only possible way of avoiding this difficulty is if the extant mechanisms which allow the formation of duplications function in such a way that the two duplicates are separated from each other by genetic information essential to the organism. Processes that insert DNA at random locations (generalized phage transduction) could serve this role in procaryotes, and translocations and inversions might serve this role in eucaryotes. This route should be very ineffective since most duplicates fixed in the population by the selective pressure will be in tandem, and at best, only a small proportion of original duplicates will be of the non-tandem kind. Moreover, tandem duplication with the possibility of higher multiples of reduplication would out-compete the rare single duplications at a remote site where further rapid duplications would not be possible.

In summary of this and previous sections, the genetic opportunity for occurrences of enzyme evolution in the enteric organisms under conditions inferred from measurements made in the laboratory on modern organisms would appear to depend on repeated cycles of high competition for a particular function followed by long intervals of non-limitation. When these latter intervals are either too short or too long, evolutionary progress by the proposed mechanism will be inhibited. Processes present in enteric organisms functioning in repair and recombination tend to slow and probably block intragene evolution not only by decreasing the marker point mutations, but more significantly, speeding the deletion of tandem duplicated genes by the process of internal recombination. Such processes as reutilization of incomplete protein fragments by special protein degradation mechanisms and the possible capacity of the organism to divert resources into reserves for the long-term advantage produce a greater opportunity for evolution.

THE EVOLUTION IN THE PRECAMBRIAN ERA

The considerations of the previous sections allow us to approach the question of enzyme evolution in the Precambrian era. After the first self-reproducing creature came into being, but prior to the time that blue-green algae and eucaryotes greatly modified the planet creating radically new niches for procaryotes, major evolutionary changes had to take place. This period endured some 10-fold longer than the time since the middle Jurassic that enteric organisms have had to become adapted to their mammalian intestinal habitat.

If we grant that the world ecosystem at this early time was effectively a collection of continuous cultures of mixed monera, then we can apply our above mathe-

mathematical treatment. It does not seem reasonable to assume that enzyme structure progressed simply because there were large numbers of organisms and long times and possibly high mutation rates. For example, to get one particular new mutant with a certain set of five particular multiple changes by a simultaneous set of 5 mutational changes within one cell generation for the conditions set in Figure 5 would happen only once in $(1.5 \times 10^{-10})^5 = 10^{-49}$. But to have 10^{49} organisms in hand would require a wet compact mass corresponding to a sphere whose radius is from here to somewhat beyond the next nearest star. Increased mutation rate would render such a change more probable, but also would increase the chances of lethal mutation in that one rare mutant organism.

Because life must have been grim, generation times should have been long and errors produced during replication rare per unit time. On the other hand, mutation rates through most of this period should have been high compared with our estimates for *E. coli*. Not only would there have initially been no repair mechanisms, but ionizing and ultraviolet radiation may have been more important than at present. Selection against cells synthesizing non-functional protein may have been intense since the total genome was simpler than today's *E. coli*, and the product of any one gene could be a larger proportion of the total. Possibly most important, there was initially no machinery for recombination. This renders the formation of duplicate genes linked in a chromosome less likely. But duplication would have to be an extremely rare phenomenon to significantly impede the takeover of the population under selective conditions by those cells bearing the duplicate genes. On the other hand, absence of any recombination mechanism protects the duplicates from the special process of internal recombination that would rapidly eject them from the population after the removal of the selective pressure.

CONCLUSIONS

From the numerical considerations presented above, it is clear that evolution would proceed through the outlined mechanism very rapidly compared to the route of simultaneous multiple changes. Of course, this kinetic mechanism will only function given an appropriate balance of mutation rates and an appropriate fluctuation in the internal and/or external environment of the organism in both kind and degree.

The role of gene duplication in evolution is not a new idea; it was originally suggested by LEWIS in 1951. It has recently received careful consideration in the monograph by OHNO (1970). Clearly, duplication followed by mutation has been important in evolution of new functions as well as elaboration of old. We second his statement: "Natural selection merely modified, while redundancy created", but we suggest that duplication for haploid organisms during earlier evolution was the sole way then available to achieve the necessary luxury of this redundancy for either invention or elaboration.

The only alternative to duplication as a mechanism to produce a temporarily unnecessary copy in the genome is to assume that circumstances altered so that

temporarily no gene function of the kind in question was necessary at all—the gene function is provided by some neighboring organism, or a substrate is temporarily available from the outside. This possibility is clearly more tenable for some kinds of genes than for others.

We have stressed the role of the chain-termination mutation as a way to render an unneeded gene less of a burden to the cells, and noted that promoter mutation or translation initiator codon mutation would do as well. Regulatory gene mutations of the kinds studied in procaryotes, e.g., the *i^s* mutations that lead to a repressor unresponsive to the inducer, would serve this role only for the special case mentioned in the last paragraph. Other special circumstances could operate for those systems subject to positive control. Perhaps there are control types of mutations in eucaryotes of a qualitatively different kind than known in procaryotes. At least the idea that control mutations of some kind would be needed for evolutionary progress was clearly presented in 1966 by DREW SCHWARTZ.

The ideas expressed here have evolved by quite similar mechanisms to those proposed here for protein evolution, though the time scale is much shorter (23 years). A number of people have caused pressures of either the selective or the mutational kind. Among those are KENNETH BERNSTEIN, JERRY BLUMBERG, CAROL DEPPE, THOMAS JUKES, JOEL KASOW, PAULINE KOYAMA, JOHN LEMMONS, ED MEREK, AARON NOVICK and LES REAL. The computer, a Wang 700, was bought with funds from the N.S.F. (Grant GB-7846). Research, both theoretical and of a more experimental kind, is supported in our laboratory by the United States Public Health Service under Grant AI-09337.

LITERATURE CITED

- BENZER, S., 1961 Genetic fine structure. *The Harvey Lectures* **56**: 1–21.
- DRAKE, J. W., 1970 *The Molecular Basis of Mutation*. Holden-Day, San Francisco.
- FRASER, A. and D. BURNELL, 1970 *Computer Models in Genetics*. McGraw-Hill, New York.
- GIBBONS, R. J. and B. KAPSIMALIS, 1967 Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *J. Bacteriol.* **93**: 510–512.
- HAGEMAN, G. D. and S. L. ROSENBERG, 1970 The evolution of bacterial enzyme systems. *Ann. Rev. Microbiol.* **24**: 429–462.
- KOCH, A. L., 1971 The adaptive responses of *Escherichia coli* to a feast and famine existence. *Adv. Microbiol. Physiol.* **6**: 147–217.
- KOYAMA, P., 1971 The adaptation of *Escherichia coli* to vertebrate hosts. Master's Thesis. Indiana University, Bloomington.
- KUBITSCHKE, H., 1970 *Introduction to Research with Continuous Cultures*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- LEWIS, E. B., 1951 Pseudoallelism and gene evolution: Cold Spring Harbor Symp. Quant. Biol. **16**: 159–174.
- MCQUILLEN, K., 1965 The physical organization of nucleic acid and protein synthesis. *Symp. Soc. Gen. Microbiol.* **15**: 134–158.
- NOVICK, A., 1955 Growth of bacteria. *Ann. Rev. Microbiol.* **9**: 97–110.
- NOVICK, A. and L. SZILARD, 1950 Experiments with the chemostat on spontaneous mutations of bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **36**: 708–719.
- OHNO, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, New York.

- RUSSELL, R. L., J. N. ABELSON, A. LANDY, M. L. GEFTER, S. BRENNER and J. D. SMITH, 1970 Duplicate genes for tyrosine transfer RNA in *Escherichia coli*. J. Mol. Biol. **47**: 1-13.
- SCHWARTZ, D., 1966 The genetic control of alcohol dehydrogenase in maize: Gene duplication and repression. Proc. Natl. Acad. Sci. U.S. **56**: 1431-1436.
- THOMAS, C. A., JR., 1970 The Theory of the Master Gene, pp. 973-998. In: *The Neurosciences: Second Study Program*. Edited by F. O. SCHMITT. The Rockefeller University Press, New York.