PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The Evolution of DNA Sequences in Escherichia coli

D. L. Hartl, Meetha Medhora, L. Green and D. E. Dykhuizen

Phil. Trans. R. Soc. Lond. B 1986 **312**, 191-204
doi: 10.1098/rstb.1986.0001

Phil. Trans. R. Soc. Lond. B **312**, 191–204 (1986)
Printed in Great Britain

The evolution of DNA sequences in *Escherichia coli*

By D. L. Hartl, Meetha Medhora, L. Green and D. E. Dykhuizen

Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110-1095, *U.S.A.*

It is proposed that certain families of transposable elements originally evolved in plasmids and functioned in forming replicon fusions to aid in the horizontal transmission of non-conjugational plasmids. This hypothesis is supported by the finding that the transposable elements Tn*3* and γδ are found almost exclusively in plasmids, and also by the distribution of the unrelated insertion sequences IS*4* and IS*5* among a reference collection of 67 natural isolates of *Escherichia coli*. Each insertion sequence was found to be present in only about one-third of the strains. Among the ten strains found to contain both insertion sequences, the number of copies of the elements was negatively correlated. With respect to IS*5*, approximately half of the strains containing a chromosomal copy of the insertion element also contained copies within the plasmid complement of the strain.

Introduction

This paper considers the evolution of DNA sequences from the point of view of prokaryotes, in particular *Escherichia coli*. As presently conceived, the evolution of bacterial populations is influenced by the characteristics and interrelationships of three rather distinct types of DNA sequences: the bacterial chromosome, plasmids, and transposable elements. Theoretical views pertaining to the roles and interrelationships of these DNA sequences do exist. Present views were derived largely from laboratory studies of a limited number of isolates, for example *E. coli* K12, and they were strongly influenced by clinical studies of the spread of multiple antibiotic resistance among pathogenic bacteria. However, modern bacteriology and population genetics place great emphasis on studies of representative samples of bacteria taken from their natural environment, and sufficient information of this type has accumulated over the past ten years that it is now possible to evaluate whether conventional views of bacterial molecular evolution hold up in the light of more recent information. Our purpose is therefore to review briefly some of the recent data that are pertinent to this evaluation.

The conventional view of bacterial evolution that emerged from classical studies was that the chromosome is the repository of most housekeeping genes, and it evolves relatively slowly under the joint effects of mutation, recombination, and natural selection. Recombination was assigned an important role in bacterial populations because of the demonstrated ability of certain conjugational plasmids of *E. coli* to mobilize and transfer chromosomal genes.

In the conventional view, plasmids are envisaged as the agent of rapid evolutionary change, infectious across a wide range of species and genera, and capable of coming or going as the selection pressure for the genetic elements they carry increases or decreases (Davey & Reanney 1980). Plasmid-borne genetic determinants have often been implicated in situations in which bacteria have become adapted to rapid and drastic changes in their environment. Examples include plasmid-borne resistance to antibiotics and heavy metals, and plasmid-determined genes coding for new catabolic pathways for novel substrates or environmental pollutants (reviewed in Davey & Reanney 1980).

The exchangeability of plasmids among bacterial strains and species has been described in a picturesque metaphor by Reanney (1978), who refers to a bacterial population as comprising a 'commonwealth' of clones, each endowed with its own chromosome but potentially capable of sharing plasmid genetic resources with other clones. The commonwealth he envisages goes far beyond the individual members of a single species.

In the conventional view of bacterial molecular evolution, the role of transposable elements is envisaged as facilitating the transfer of DNA sequences between chromosome and plasmids by means of transposition or replicon fusion (Saedler *et al.* 1980). Transposable elements are also postulated to create novel favourable mutations in the chromosome by virtue of the ability of transposable elements to create deletions or other chromosomal rearrangements, or to contribute promoter sequences contained within them to chromosomal genes resulting in gene expression or altered regulation (Reynolds *et al.* 1981; Ciampi *et al.* 1982).

In the following sections, these conventional views will be compared with currently available data from natural isolates of *E. coli*. A more detailed review of the data may be found in Hartl & Dykhuizen (1984).

Chromosomal DNA

The bacterial chromosome needs little comment by way of definition. It is the DNA molecule that contains all of the genes that are required for growth, as well as genes for many non-essential metabolic functions. However, the chromosome is not by any means identical in different isolates. For example, the genome size of *E. coli* isolates ranges from 3400 to 4500 kilobases (Brenner *et al.* 1972). The extreme difference is 1100 kilobases, or about 700 average-sized genes. Not all of the difference is ascribable to plasmids. The genome size of *E. coli* K12, which contains no plasmids, is 3800 kilobases (Brenner *et al.* 1972). Thus, strains with a smaller genome than *E. coli* K12 must have a smaller chromosome. Virtually nothing is known about genetic variation of this type in natural populations.

The study of chromosomal genes in isolates of bacteria from natural populations was revolutionized by the application of protein electrophoresis. Because the vast majority of electrophoretic variants of enzymes (allozymes) result from alternative allelic forms of the structural gene, electrophoresis made it possible for the first time to identify multiple alleles of a sufficient number of chromosomal genes in a large number of isolates. To our knowledge, electrophoresis was first used in a bacterial population study by Milkman (1973), and very extensive surveys have been performed by Selander and collaborators (Selander & Levin 1980; Selander & Whittam 1983; Ochman *et al.* 1983; Whittam *et al.* 1983*a*, *b*; Ochman & Selander 1984*a*, *b*; Caugant *et al.* 1984). These studies were carried out with natural isolates of *E. coli*.

Genetic variation

The first conclusion derived from electrophoretic studies was that *E. coli* populations contain impressive amounts of genetic variation in the form of multiple alleles at many loci. Indeed, natural populations of *E. coli* contain more genetic variation than do populations of eukaryotes (Selander & Levin 1980). As a corollary, natural isolates of *E. coli* represent a large number of possible genotypes. For example, in one study of 20 enzymes among 109 isolates derived from diverse human and animal sources, 98 distinguishable multiple-locus electrophoretic phenotypes (referred to as e.ts) were identified (Selander & Levin 1980). However, strains with identical electrophoretic types, presumably reflecting identical or nearly identical genotypes for the relevant loci, may be found repeatedly in independent samples. For example, in the same collection of 109 isolates, seven groups of two to four isolates each were electrophoretically identical for all 20 enzymes. Interestingly, one e.t. was found to be indistinguishable from the e.t. of *E. coli* K12. Finally, identical or virtually identical e.ts are often isolated from populations that are very widely separated geographically or temporally (Whittam *et al.* 1983*a*). This reflects the fact that geographical differentiation accounts for only about 2% of the total genetic diversity in *E. coli* (Whittam *et al.* 1983*b*). Thus, *E. coli* represents a very diverse collection of genotypes, many or most of which are worldwide in distribution.

Recombination

A second conclusion derived from electrophoretic studies is that recombination between chromosomal genes is greatly restricted. This conclusion emerges from the finding that combinations of alleles at different loci are found with frequencies that are very different from those expected simply by multiplying together the frequencies of the relevant alleles (Whittam *et al.* 1983*a*). This situation of non-random association between the alleles of different loci is known as linkage disequilibrium. Ordinarily, recombination has the effect of dissipating linkage disequilibrium, and the rate at which the linkage disequilibrium disappears is a simple function of the recombination fraction between the loci. The finding of great and persistent linkage disequilibrium implies that recombination between chromosomal genes among natural isolates may be quite rare. This model has become known as the clonal model of population structure because each chromosomal genotype in a natural population effectively represents an independent lineage or clone (Ochman & Selander 1984*a*). An alternative model, that the linkage disequilibrium occurs because certain particular combinations of alleles are strongly favoured by natural selection while other combinations are strongly disfavoured, can be rejected, based on the finding that, in about 80% of the cases, the direction of linkage disequilibrium is not consistent in different localities (Whittam *et al.* 1983*b*). That is, certain pairs of alleles that occur together more often than would be expected by chance in one population are found to occur together less often than expected in some other population.

Infraspecific population structure

As expected with a mode of reproduction that is predominantly clonal, strains of *E. coli* can be grouped with respect to common ancestry. The most thorough approach is that of Whittam *et al.* (1983*a*), who used factor analysis to analyse the 12-locus e.ts of 1705 strains of *E. coli* and *Shigella*. This analysis led to identification of three large groups of *E. coli* strains, with great genetic variation within each group, but closer relationships within groups than between groups. One representative subset of 72 of the strains, which is called the ECOR reference collection of *E. coli* (Ochman & Selander 1984*b*), has been analysed with respect to an independent set of characters (biotype characters) related to the ability of the isolates to use particular nutrients as growth substrates (Miller & Hartl 1985). Unweighted pair-group cluster analysis of the biotype data again revealed a small number of related strain clusters, which corresponded well to those identified by electrophoresis. The clusters of strains can also be shown to differ in genetic attributes unrelated to both allozymes and biotypes. For example, the group designated Group I includes significantly more strains containing element IS*5* than other groups (Green *et al.* 1984).

Theoretical implications

In terms of chromosomal genes, the conventional view of the population genetics of *E. coli* fell short on two counts. First, the amount of genetic variation in the species was grossly underestimated. Considering the true amount of genetic variation that is now known to occur in this organism, it is interesting to recall that *E. coli* strains K12, B, and C were occasionally supposed to represent the full range of diversity among non-pathogenic *E. coli*. Secondly, the conventional view of *E. coli* population genetics tended to overestimate the frequency of recombination, an understandable error because of the great importance and ease of obtaining recombination in the laboratory. This view, supported by electrophoretic studies, implies that recombination is sufficiently rare that great linkage disequilibrium is maintained. However, rare recombination probably does play a role in the evolution of *E. coli*. Rare recombination occurring at a rate of the order of the mutation rate would have a negligible effect in dissipating linkage disequilibrium, but it would have major evolutionary implications in the long term (Hartl & Dykhuizen 1984).

Near neutrality of allozyme variants

The widespread occurrence of allozyme variation in *E. coli* has provided an opportunity to evaluate experimentally the hypothesis of Kimura (1968) that most electrophoretic variants of enzymes are selectively neutral or nearly neutral. The experimental design was straightforward: bacteriophage-mediated transduction was used to construct strains that contained alternative allozyme alleles from natural isolates, but the strains were otherwise isogenic for the genetic background of *E. coli* K12. Genetically marked pairs of such strains were then inoculated into chemostats in which metabolism of the limiting nutrient required the enzyme in question, and the rate of change in the relative frequencies of the strains over time was used to estimate the relative fitness of the competing strains under these conditions (Dykhuizen & Hartl 1983*a*). By using these procedures, the smallest selection coefficient that can be detected is approximately 0.002 (per hour). Details of the experiments have been discussed elsewhere (Dykhuizen & Hartl 1980, 1983*b*; Dykhuizen *et al.* 1984*a*, *b*; Hartl & Dykhuizen 1981), but the overall picture is that most allozyme alleles are selectively nearly neutral when growth is limited by a nutrient such as glucose that is common in the natural habitat. However, evidence of selection is often observed when the limiting nutrient is an unusual substrate (Dykhuizen *et al.* 1984*b*; Hartl & Dykhuizen 1985). Based on this evidence, it is probably legitimate to regard naturally occurring allozymes of *E. coli* as convenient genetic markers that are themselves nearly neutral. Our data are also consistent with the view that most electrophoretic mutations are slightly deleterious (Ohta 1973), with selection coefficients below the limit of resolution of chemostat techniques.

Plasmids

Plasmids are, of course, extrachromosomal genetic elements. In natural isolates of *E. coli*, plasmids range in size from a few hundred base pairs to a few hundred kilobase pairs, but the distribution of sizes is bimodal. One group of plasmids consists of elements that are smaller than 7.5 kilobases, and another group consists of plasmids that are larger than 40 kilobases, although there are a few plasmids of intermediate size (Silver *et al.* 1980). In addition to a replication origin, which is essential, plasmids contain genes for various functions. Some genes are related to plasmid functions, such as the formation of pili or plasmid transfer during conjugation. Other genes are related to host metabolism, and the best known of these are genes that determine antibiotic resistance. Some plasmids contain no known genes that affect the host. They are said to be 'cryptic plasmids', which is a tentative designation meaning simply that we do not know their function.

In some instances plasmids can contain genes that are normally located in the chromosome, such as in the F′ plasmids of *E. coli* that contain a segment of chromosomal DNA. On the other hand, some genes that are normally contained in plasmids are occasionally found in the bacterial chromosome (Davey & Reanney 1980). Here we seem to have a prokaryotic analogue of the relationship that exists in eukaryotic oncogenes between copies of the genes found in retroviruses (*v*-genes) and their very similar chromosomal counterparts (*c*-genes). One might designate prokaryotic plasmid-borne genes and their chromosomal counterparts as *p*-genes and *c*-genes. Additional complexity results from the fact that some plasmids can themselves become part of the chromosome. The best known example in *E. coli* is the integration of the F plasmid into the chromosome resulting in the creation of an Hfr bacterial strain that undergoes effective chromosome transfer upon conjugation.

In any event, bacterial genes can shuttle between the chromosome and plasmids, which in principle could lead to difficulty in determining which circular replicating elements are chromosomes and which are plasmids. However, operationally we can define the chromosome as that DNA molecule whose presence is essential for viability under all conditions of growth.

Distribution of numbers

Natural isolates of non-pathogenic *E. coli* contain many plasmids. One example is shown in figure 1, which shows the distribution of large and small plasmids in 50 strains of the ECOR

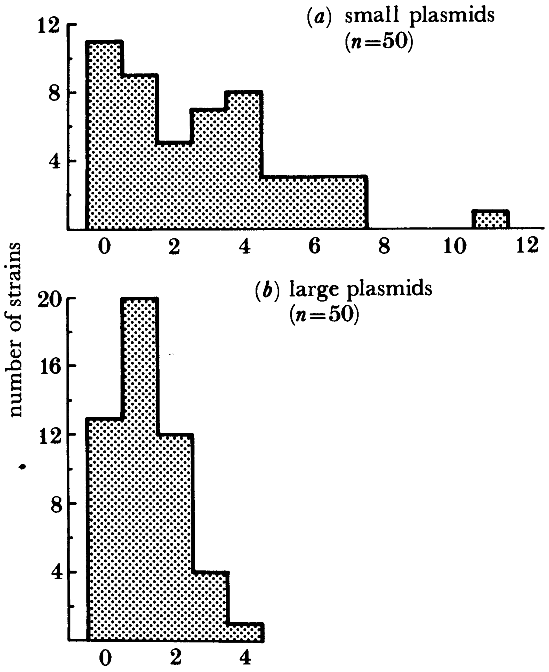

FIGURE 1. Distribution of small and large plasmids among strains in the ECOR collection. The means are 2.7 and 1.2, respectively; standard deviations are 2.5 and 1.0, respectively.

reference collection. DNA was prepared and subjected to agarose electrophoresis as described in Anderson & McKay (1983), which leaves circular DNA molecules largely intact. With this procedure, 'small plasmids' (less than about 25 kilobases) migrate faster than the band of chromosomal DNA, and 'large plasmids' (larger than about 25 kilobases) migrate slower than the chromosomal band. However, in this collection of strains, as in others, most plasmids are either smaller than about 7.5 kilobases or larger than about 40 kilobases. Most strains in the ECOR collection contain at least one large plasmid and several small ones. The means for these strains are 1.2 ± 0.1 large plasmids (range 0–4) and 2.7 ± 0.4 small plasmids (range 0–11). There is no significant correlation between the number of large and small plasmids in a strain.

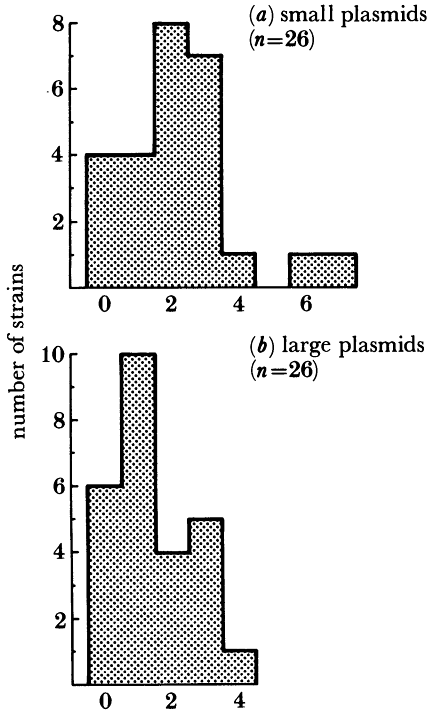

FIGURE 2. Distribution of small and large plasmids among strains in the Murray collection. The means are 2.2 and 1.4, respectively; standard deviations are 1.7 and 1.2, respectively.

Figure 2 shows the distribution of plasmids in a second collection of natural isolates of *E. coli*, a particularly interesting one because it was collected by Dr E. D. G. Murray 45–55 years ago and stored in stab tubes until they were opened in 1980 to obtain duplicates for the National Collection of Type Cultures (London) (Datta & Hughes 1983). There is little or no difference between the plasmid distribution in the Murray collection and that in the ECOR collection. The 26 strains in the Murray collection average 1.4 ± 0.2 large plasmids (range 0–4) and 2.2 ± 0.3 small plasmids (range 0–7). Although particular plasmids may be unstable in bacterial strains stored for long periods, the Murray collection gives no evidence for large-scale loss of plasmids under routine storage conditions.

Theoretical expectations

According to the conventional view, the distribution of plasmids among isolates results from their infectious horizontal transfer by means of conjugation. The evidence for this model derives from the spread of multiple drug resistance plasmids (resistance transfer factors), and the importance of the phenomenon in clinical settings is well documented (Datta & Hughes 1983). Moreover, the horizontal transfer of large conjugational plasmids has been observed in chemostats, and the rate of transfer of F is approximately 3.3×10^{-12} ml per cell per hour (Levin *et al.* 1979). Small plasmids, unable to spread by themselves, can be transferred by a sort of hitch-hiking when a large plasmid in the same cell undergoes conjugational transfer. The rate of cotransfer of the small plasmid pCR1 with a derivative of the large plasmid F in chemostats is approximately 2.1×10^{-12} ml per cell per hour (Levin & Rice 1980).

The conventional view also assumes that the genotype of the new host cell plays little part in the transmission of a plasmid or its ability to be maintained, unless the new host already happens to contain a plasmid within the same incompatibility group.

Test of the model

One implication of the model of the previous section is that the plasmids found in a collection of natural isolates of *E. coli* should be distributed at random with regard to bacterial genotype. In the long run a critical evaluation of this prediction will require development of a method of identification of individual plasmids that is suitable for large-scale population screening. However, the prediction can be evaluated crudely by an examination of mere numbers of plasmids. If the number of plasmids in natural isolates is non-random with regard to bacterial genotype, then the model must be revised to incorporate additional important determinants of plasmid distribution.

Figure 3 provides the distribution of plasmid numbers among four collections of *E. coli* strains

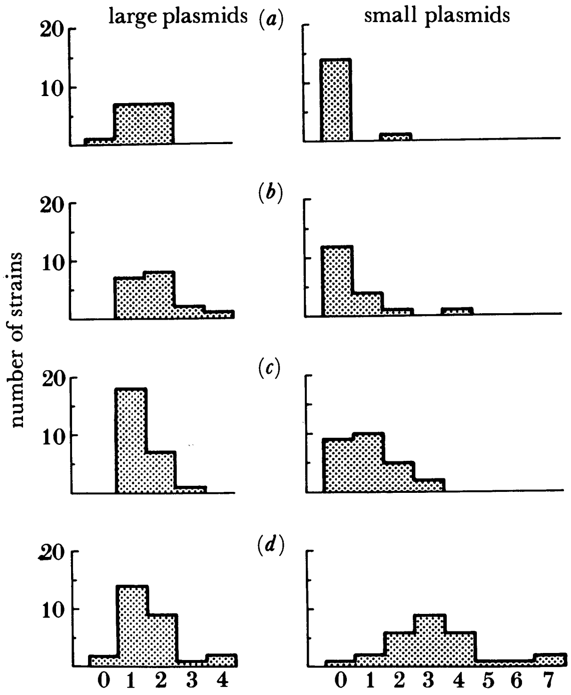

FIGURE 3. Distribution of large and small plasmids among strains sharing a recent common ancestor. (*a*) Serotype O18:K1, outer membrane protein (o.m.p.) 6, electrophoretic type (e.t.) f, $n = 15$; (*b*) O18:K1, o.m.p. 9, e.t. f, $n = 18$; (*c*) O1:K1, o.m.p. 9, e.t. e, $n = 26$; (*d*) O7:K1, o.m.p. 3, e.t. 1, $n = 28$. Data from Achtman *et al.* (1983) and Ochman & Selander (1984*a*).

with capsular antigen K1. The strains were collected at diverse places in Europe and North America, mainly in the 1970s, but a few isolates date back to the 1940s. Achtman *et al.* (1983) have studied the biotypes of the strains, the electrophoretic pattern of their outer membrane proteins (o.m.p.), and their plasmid profiles. Ochman & Selander (1984*a*) have determined the electrophoretic type (e.t.) of the strains with respect to ten polymorphic chromosomal enzymes. One finding of great interest is that strains of the same O:K serotype collected at different places and times often represent clones of the same genotype as judged by their o.m.p. pattern and e.t. (Achtman *et al.* 1983; Ochman & Selander 1984*a*). For present purposes, the strains are most useful for the information they provide regarding the distribution of plasmids among genotypes.

Large plasmids

The distributions of the number of large plasmids among the four types of clones in figure 3 are indistinguishable. However, the plasmids have not yet been identified, so it is possible that there will be significant differences among the large plasmids contained in strains of different genotypes. Although there is variation within each O:K serotype, the types of colicin produced by the clones are different (Achtman *et al.* 1983), which might be found to correlate with differences in the composition of large plasmids.

Small plasmids

The data in figure 3 are far more conclusive with regard to small plasmids because the small plasmids are not distributed at random with regard to genotype. Remarkably, all but one of the isolates of the O18:K1 clone in figure 3*a* contain no small plasmids, and 12 out of the 18 isolates of the related O18:K1 clone in figure 3*b* lack small plasmids as well. Furthermore, the clone in figure 3*a* was isolated in such diverse places as England and Sweden in Europe, and Seattle, San Francisco, New York, and Texas in the United States. In contrast, isolates of the O1:K1 clone in figure 3*c* contain an average of 1.0 ± 0.2 small plasmids, and isolates of the O7:K1 clone in figure 3*d* contain an average of 3.2 ± 0.3 small plasmids.

Two explanations might be offered for the absence of small plasmids in the O18:K1 clone in figure 3*a*. First, the replication or segregation of small plasmids in this genotype could be defective, so small plasmids might be unstable. This possibility is subject to direct experimental test. Host genotype has been observed to affect the stability of small plasmids in bacterial chemostats (Noack *et al.* 1982). Alternatively, the infectious transmission of small plasmids might actually occur at such a low rate that the clone could have spread throughout Europe and the United States without having become infected.

Transposable elements

The third component in the evolution of bacterial DNA consists of transposable elements. Transposable elements are DNA sequences that are capable of changing their location from one position in a replicon to a different position in the same replicon or to another replicon. Many transposable elements include DNA sequences called insertion sequences. Insertion sequences are themselves transposable elements which are usually smaller than 2 kilobases and which code for no known functions other than those associated with their transposition and the regulation of transposition (Calos & Miller 1980).

The typical structure of an insertion sequence includes a sequence of nucleotides that is present in inverted orientation at opposite ends. The sequence of this inverted repeat is characteristic of each element and provides the recognition sequence for the transposase enzyme that is encoded in the central region of the element and that promotes the transposition reaction. The inverted repeats permit two appropriately spaced and identical elements to mobilize the transposition of the entire DNA sequence between the elements, because the transposase can recognize the repeated sequences at the extreme ends. Composite transposable elements consisting of a central region of unique DNA sequence flanked by a pair of insertion sequences are well known. Such transposons are often discovered because the central region contains a genetic determinant that confers antibiotic resistance on the host. The element Tn*5*, for example, consists of a 2.6 kilobase central region containing a kanamycin–neomycin resistance gene, flanked by two virtually identical copies of the 1.5 kilobase insertion sequence IS*50* (Berg *et al.* 1975, 1982; Rothstein *et al.* 1981). Tn*5* was originally isolated from *Klebsiella*, and the central region of unique DNA not only contains the kanamycin resistance gene but also a gene that is unexpressed in *E. coli* but which confers streptomycin resistance in *Rhizobium* species (Putnoky *et al.* 1983). This finding is remarkable, because it suggests that transposons may become very widely disseminated among bacterial species. Similar conclusions had been reached earlier from studies of the distribution of antibiotic resistance. Temperate bacteriophages such as λ are currently interpreted as highly specialized transposable elements.

Insertion sequences are a normal constituent of the bacterial chromosome. The chromosome of *E. coli* K12 contains from one to ten copies of each of six identified insertion sequences: IS*1*, IS*2*, IS*3*, IS*4*, IS*5*, and IS*30*. In addition, the insertion sequence γδ is present in the F plasmid but not in the chromosome of *E. coli* K12. However, studies of the distribution of IS*4* and IS*5* among natural isolates suggest that natural isolates may contain several additional insertion sequences that have so far escaped identification because they are not present in the K12 strain (Green *et al.* 1984).

Theories of evolution

The role of insertion sequences and other transposable elements in bacterial evolution is controversial because their effects upon the fitness of the host are unclear (Hartl *et al.* 1984). At one end of the spectrum is the model of selfish DNA (Doolittle & Sapienza 1980; Orgel & Crick 1980), in which insertion sequences are proposed to be of no positive benefit to the host. Selfish DNA is maintained solely because of its capacity for transposition and transfer among strains. Many other models are possible, such as the one proposed by Campbell (1981), and defended forcefully, that extrachromosomal DNA elements have probably earned their keep since the earliest stages of their evolution by contributing to an increase in fitness of the host. A different type of model for the evolution of transposable elements has been put forward by Syvanen (1984), who argues that the elements have evolved because they can bring about genomic rearrangements, some fraction of which might be favourable for fitness. However, the evolution of such complex entities as transposable elements was surely a long, slow process, and it is hard to imagine how some future benefit accruing to such elements could be the driving force behind their evolution. In considering the evolution of complex DNA sequences, we believe that it is essential to follow the logic of Darwin in thinking about complex morphological adaptations: whatever theory is proposed, it is necessary that each successive step in the process be favourable in itself.

Fitness effects of insertion sequences

Evidence for the Campbell model of the evolution of insertion sequences has been provided by competition studies of IS*50* in isogenic strains by using chemostats (Biel & Hartl 1983; Hartl *et al.* 1983). We find that the fitness of a strain containing IS*50* is increased by about 5% per hour compared with an otherwise isogenic counterpart, and that the increased fitness is not attributable to actual transposition of the element, although a functional transposase enzyme is required. The mechanism of selection of IS*50* is not known in great detail, but similar favourable effects in chemostats have also been reported for lysogens of λ, Mu, P1, and P2 (Edlin *et al.* 1975, 1977; Lin *et al.* 1977). On the other hand, the view of transposable elements as mutators has received support from Chao *et al.* (1983), who have found that strains containing Tn*10* are favoured by selection in chemostats, and the selection appears to result from the transposition of Tn*10* to a specific favourable chromosomal site.

Distribution of insertion sequences

The distributions of the number of copies of two insertion sequences, IS*4* and IS*5*, have been studied among the ECOR strains (Green *et al.* 1984; Dykhuizen *et al.* 1985), and the data are summarized in figure 4. These two elements were chosen for analysis because they represent the extreme numbers present in *E. coli* K12. The sequence IS*4* is found in one copy, and IS*5* is found in approximately ten copies. Surprisingly, the distributions observed among the ECOR strains are virtually indistinguishable. With both elements, approximately two-thirds of the strains do not contain the IS element. The absence of the element in a strain was determined

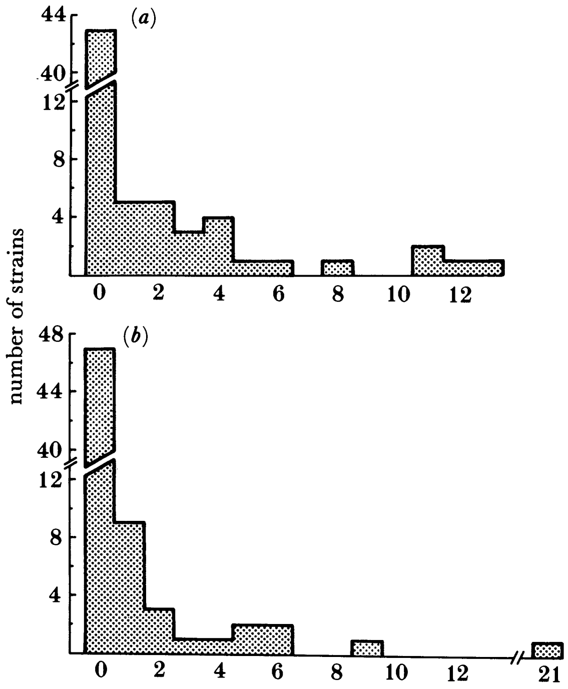

FIGURE 4. Distribution of number of chromosomal copies of IS*4* and IS*5* among strains in the ECOR collection. The means are 1.6 and 1.1, respectively; standard deviations are 3.1 and 3.0, respectively. (*a*) IS*4*, $n = 67$; (*b*) IS*5*, $n = 67$.

by means of hybridization with total DNA, plasmid DNA included, so the isolates in the 0 class in figure 4 do not contain the insertion sequence anywhere in the genome. Among strains that do contain the element, the average number of copies of IS*4* is 4.4 ± 0.8 and that of IS*5* is 3.7 ± 1.0.

Theoretical results regarding the expected distribution of insertion sequences among bacteria have been calculated under a variety of assumptions (Sawyer & Hartl 1985). The results in figure 4 give a satisfactory fit to a number of models in which fitness is a decreasing function of copy number, and individuals that fail to reproduce are replaced with uninfected individuals that may later become infected by means of horizontal transfer. The rates of horizontal transfer required to result in distributions quantitatively like those in figure 4 are approximately one-quarter to one-fifth of the rate of transposition.

Insertion sequences in plasmids

If insertion sequences new to a strain are first recruited in plasmids and later transpose to the chromosome, then insertion sequences contained in the chromosome of a species would also be expected to be found in plasmids. The occurrence of IS*2* and IS*3* in the F plasmid provides a good illustration of this point, and we have carried out hybridization experiments to determine the plasmid distribution of IS*4* in 30 IS*4*-containing strains of the ECOR collection and the distribution of IS*5* in 27 IS*5*-containing strains.

With respect to the 30 strains containing IS*4*, 18 strains contained chromosomal IS*4* sequences only, 3 strains contained plasmid IS*4* sequences only, and 2 strains contained IS*4* in both types of molecule. These findings are consistent with the hypothesis of statistical independence in the distribution of IS*4* among plasmids and in the chromosome. Among the 30 strains there were 40 large plasmids and 82 small ones, and IS*4* was found in 3 large plasmids (8%) and in 2 small ones (2%). In two isolates judged to be very closely related because of their identical e.ts, IS*4* was contained in both a large plasmid and a small plasmid, but there were no chromosomal copies. This kind of distribution would be expected as a consequence of the transfer of the small plasmid by means of cointegrate formation between the small plasmid and the large self-transmissible plasmid brought about by IS*4*.

With respect to the 27 strains containing IS*5*, 13 strains contained chromosomal copies only, 2 strains contained plasmid copies only, and 12 strains contained both chromosomal and plasmid copies. These findings depart highly significantly from independence, primarily because there are far too many strains with the IS sequence in both plasmids and the chromosome. Among the 27 strains there were 38 large plasmids and 89 small ones, and IS*5* was found in 13 large plasmids (34%) but in no small ones. The absence of IS*5* in small plasmids is statistically significant.

The discovery that both IS*4* and IS*5* are abundantly distributed among large plasmids supports the conventional view that plasmids play an essential role in the dissemination of transposable elements. Indirectly, the finding also suggests that large plasmids may be rather easily exchanged among strains, lending additional credence to the commonwealth analogy (Reanney 1978). However, the finding in figure 3*a* that certain clones of worldwide distribution lack small plasmids, suggests that generalization of the commonwealth idea to include small plasmids may be premature. In addition, the absence of IS*5* in small plasmids suggests that we have a good deal to learn yet about the relationship between insertion sequences and small plasmids.

Discussion

Studies of natural isolates of non-pathogenic *E. coli* have added significant new information to that obtained from previous studies of the genetics, molecular biology, and medical microbiology of this organism. A synthesis of all currently available information about the population genetics of *E. coli* must include the following points.

(i) *E. coli* populations contain impressive amounts of genetic variation. This genetic variation can be detected by any of a variety of techniques, but electrophoresis is the most generally useful and powerful technique for studying the products of chromosomal genes. Under normal conditions, most electrophoretic enzyme variants appear to be nearly equal in fitness, although slight detrimental effects cannot be excluded.

(ii) Some genotypes of *E. coli* are worldwide in distribution, whereas others are more geographically restricted (Selander & Levin 1980; Caugant *et al.* 1984). Isolates sharing the same serotype usually represent one or at most several clones of genetically closely related organisms (Achtman *et al.* 1983; Ochman & Selander 1984*a*).

(iii) Chromosomal recombination among clones of *E. coli* is limited, as indicated by the observed high degree of linkage disequilibrium and the ability to assign isolates to a small number of relatively distinct clusters of strains (Whittam *et al.* 1983*a*; Miller & Hartl 1985).

(iv) The majority of isolates of *E. coli*, including those maintained for many years in stab cultures, contain one or two large plasmids and several small plasmids. However, different clones of worldwide distribution can have very different numbers of small plasmids. Although many large plasmids can be transmitted between strains by means of conjugation, it is still unclear how frequently such transfer occurs among non-pathogenic strains in their natural environment (Levin *et al.* 1979). Even greater uncertainty attaches to the transmissibility of small plasmids (Levin & Rice 1980). In other words, it is still unclear how much real communication occurs within the 'commonwealth' (Reanney 1978).

(v) Insertion sequences can be found in multiple copies in the chromosome and in large (and sometimes small) plasmids. This finding suggests that insertion sequences are indeed disseminated among strains by means of plasmids, and the role of insertion sequences in the evolution of complex multiple-resistance plasmids is well known. However, with IS*4* and IS*5*, about two-thirds of all isolates do not contain the insertion sequence. Among strains that do contain the sequence in the chromosome, the number and size distribution of restriction fragments containing the sequence provide a useful index of recent common ancestry and genetic relatedness among strains (Green *et al.* 1984; Dykhuizen *et al.* 1985).

Much of what we know of the population genetics of *E. coli* is still at a rather unsatisfactory descriptive level. However, the description has reached the point where specific, concrete hypotheses can be proposed and subjected to rigorous experimental test. The new tools of molecular population genetics applied to the evolution of DNA sequences will inevitably result in new perceptions about the evolutionary process and perhaps will reveal new principles.

This work was supported by a grant from the U.S. National Institute of General Medical Sciences GM-30201.

References

Achtman, M., Mercer, A., Kusecek, B., Pohl, A., Heuzenroeder, M., Aaronson, W., Sutton, A. & Silver, R. P. 1983 Six widespread bacterial clones among *Escherichia coli* K1 isolates. *Infect. Immun.* **39**, 315–335.

Anderson, D. G. & McKay, L. L. 1983 Simple and rapid method for isolating large plasmid DNA from lactic Streptococci. *Appl. environ. Microbiol.* **46**, 549–552.

Berg, D. E., Davies, J., Allet, B. & Rochaix, J.-D. 1975 Transposition of R-factor genes to the bacteriophage lambda. *Proc. natn. Acad. Sci. U.S.A.* **72**, 3628–3632.

Berg, D. E., Johnsrud, L., McDivitt, L., Ramabhadran, R. & Hirschel, B. J. 1982 Inverted repeats of Tn*5* are transposable elements. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2632–2635.

Biel, S. W. & Hartl, D. L. 1983 Evolution of transposons: natural selection for Tn*5* in *Escherichia coli* K12. *Genetics* **103**, 581–592.

Brenner, D. J., Fanning, G. R., Skerman, F. J. & Falkow, S. 1972 Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J. Bacteriol.* **109**, 953–965.

Calos, M. P. & Miller, J. H. 1980 Transposable elements. *Cell* **20**, 579–595.

Campbell, A. 1981 Evolutionary significance of accessory DNA elements in bacteria. *A. Rev. Microbiol.* **35**, 55–83.

Chao, L., Vargas, C., Spear, B. B. & Cox, E. C. 1983 Transposable elements as mutator genes in evolution. *Nature, Lond.* **303**, 633–635.

Ciampi, M. S., Schmid, M. B. & Roth, J. R. 1982 Transposon Tn*10* provides a promoter for transcription of adjacent sequences. *Proc. natn. Acad. Sci. U.S.A.* **79**, 5016–5020.

Caugant, D. A., Levin, B. R. & Selander, R. K. 1984 Distribution of multilocus genotypes of *Escherichia coli* within and between host families. *J. Hyg.* **92**, 377–384.

Datta, N. & Hughes, V. M. 1983 Plasmids of the same Inc groups in Enterobacteria before and after the medical use of antibiotics. *Nature, Lond.* **306**, 616–617.

Davey, R. B. & Reanney, D. C. 1980 Extrachromosomal genetic elements and the adaptive evolution of bacteria. *Evol. Biol.* **13**, 113–147.

Doolittle, F. W. & Sapienza, C. 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature, Lond.* **284**, 601–603.

Dykhuizen, D. E. & Hartl, D. L. 1980 Selective neutrality of 6PGD allozymes in *Escherichia coli* and the effects of genetic background. *Genetics* **96**, 801–817.

Dykhuizen, D. E. & Hartl, D. L. 1983*a* Selection in chemostats. *Microbiol. Rev.* **47**, 150–168.

Dykhuizen, D. E. & Hartl, D. L. 1983*b* Functional effects of PGI allozymes in *Escherichia coli. Genetics* **105**, 1–18.

Dykhuizen, D. E., de Framond, J. & Hartl, D. L. 1984*a* Selective neutrality of glucose-6-phosphate dehydrogenase allozymes in *Escherichia coli. Molec. Biol. Evol.* **1**, 162–170.

Dykhuizen, D. E., de Framond, J. & Hartl, D. L. 1984*b* Potential for hitchhiking in the *eda-edd-zwf* gene cluster of *Escherichia coli. Genet. Res., Camb.* **43**, 229–239.

Dykhuizen, D. E., Sawyer, S. A., Green, L., Miller, R. D. & Hartl, D. L. 1985 Joint distribution of insertion elements IS*4* and IS*5* in natural isolates of *Escherichia coli. Genetics.* (In the press.)

Edlin, G., Lin, L. & Kudrna, R. 1975 λ lysogens of *E. coli* reproduce more rapidly than nonlysogens. *Nature, Lond.* **255**, 735–737.

Edlin, G., Lin, L. & Bitner, R. 1977 Reproductive fitness of P1, P2 and Mu lysogens. *J. Virol.* **21**, 560–564.

Green, L., Miller, R. D., Dykhuizen, D. E. & Hartl, D. L. 1984 Distribution of DNA insertion element IS*5* in natural isolates of *Escherichia coli. Proc. natn. Acad. Sci. U.S.A.* **81**, 4500–4504.

Hartl, D. L. & Dykhuizen, D. E. 1981 Potential for selection among nearly neutral allozymes of 6-phosphogluconate dehydrogenase in *Escherichia coli. Proc. natn. Acad. Sci. U.S.A.* **78**, 6344–6348.

Hartl, D. L. & Dykhuizen, D. E. 1984 The population genetics of *Escherichia coli. A. Rev. Genet.* **18**, 31–68.

Hartl, D. L. & Dykhuizen, D. E. 1985 The neutral theory and the molecular basis of preadaptation. In *Population genetics and molecular evolution* (ed. T. Ohta & K. Aoki), pp. 107–124. Tokyo: Japan Scientific Societies Press.

Hartl, D. L., Dykhuizen, D. E. & Berg, D. E. 1984 Accessory DNAs in the bacterial gene pool: playground for coevolution. In *Origins and development of adaptations* (Ciba Foundation Symposium 102), pp. 233–245. London: Pitman.

Hartl, D. L., Dykhuizen, D. E., Miller, R. D., Green, L. & de Framond, J. 1983 Transposable element IS*50* improves growth rate of *E. coli* cells without transposition. *Cell* **35**, 503–510.

Kimura, M. 1968 Evolutionary rate at the molecular level. *Nature, Lond.* **217**, 624–626.

Levin, B. R. & Rice, V. A. 1980 The kinetics of transfer of nonconjugative plasmids by mobilizing conjugative factors. *Genet. Res., Camb.* **35**, 241–259.

Levin, B. R., Stewart, F. M. & Rice, V. A. 1979 The kinetics of conjugative plasmid transmission: fit of a simple mass action model. *Plasmid* **2**, 247–260.

Lin, L., Bitner, R. & Edlin, G. 1977 Increased reproductive fitness of *Escherichia coli* lambda lysogens. *J. Virol.* **21**, 554–559.

Milkman, R. 1973 Electrophoretic variation in *E. coli* from natural sources. *Science, Wash.* **182**, 1024–1026.

Miller, R. D. & Hartl, D. L. 1985 Biotyping confirms a nearly clonal population structure in *Escherichia coli*. *Evolution*. (In the press.)

Noack, D., Roth, M., Geuther, R., Muller, G., Undisz, K., Hoffmeier, C. & Gaspar, S. 1982 Maintenance and genetic stability of vector plasmids pBR322 and pBR325 in *Escherichia coli* K12 strains grown in a chemostat. *Molec. gen. Genet.* **184**, 121–124.

Ochman, H. & Selander, R. K. 1984*a* Evidence for clonal population structure in *Escherichia coli*. *Proc. natn. Acad. Sci. U.S.A.* **81**, 198–201.

Ochman, H. & Selander, R. K. 1984*b* Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**, 690–693.

Ochman, H., Whittam, T. S., Caugant, D. A. & Selander, R. K. 1983 Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella*. *J. gen. Microbiol.* **129**, 2715–2726.

Ohta, T. 1973 Slightly deleterious mutant substitutions in evolution. *Nature*, *Lond.* **246**, 96–98.

Orgel, L. & Crick, F. H. C. 1980 Selfish DNA: the ultimate parasite. *Nature*, *Lond.* **284**, 604–607.

Putnoky, P., Kiss, G. B., Ott, I. & Kondorosi, A. 1983 Tn*5* carries a streptomycin resistance determinant downstream from the kanamycin resistance gene. *Molec. gen. Genet.* **191**, 288–294.

Reanney, D. C. 1978 Coupled evolution: adaptive interactions among the genomes of plasmids, viruses, and cells. *Int. Rev. Cytol.* (suppl. 8), 1–68.

Reynolds, A. E., Felton, J. & Wright, A. 1981 Insertion of DNA activates the cryptic *bgl* operon in *E. coli* K12. *Nature*, *Lond.* **293**, 625–629.

Rothstein, S. J., Jorgensen, R. A., Yin, J. C. P., Yong-Di, Z., Johnson, R. C. & Reznikoff, W. S. 1981 Genetic organization of Tn*5*. *Cold Spring Harbor Symp. quant. Biol.* **45**, 99–105.

Saedler, H., Cornelis, G., Cullum, J., Schumacher, B. & Sommer, H. 1980 IS*1*-mediated rearrangements. *Cold Spring Harbor Symp. quant. Biol.* **45**, 93–98.

Sawyer, S. A. & Hartl, D. L. 1985 Distribution of transposable elements in prokaryotes. *Theor. Pop. Biol.* (In the press.)

Selander, R. K. & Levin, B. R. 1980 Genetic diversity and structure in *Escherichia coli* populations. *Science*, *Wash.* **210**, 545–547.

Selander, R. K. & Whittam, T. S. 1983 Protein polymorphism and the genetic structure of populations. In *Evolution of genes and proteins* (ed. M. Nei & R. K. Koehn), pp. 89–114. Sunderland, Massachusetts: Sinauer Associates.

Silver, R. P., Aaronson, W., Sutton, A. & Schneerson, R. 1980 Comparative analysis of plasmids and some metabolic characteristics of *Escherichia coli* K1 from diseased and healthy individuals. *Infect. Immun.* **29**, 200–206.

Syvanen, M. 1984 The evolutionary implications of mobile genetic elements. *A. Rev. Genet.* **18**, 271–293.

Whittam, T. S., Ochman, H. & Selander, R. K. 1983*a* Multilocus genetic structure in natural populations of *Escherichia coli*. *Proc. natn. Acad. Sci. U.S.A.* **80**, 1751–1755.

Whittam, T. S., Ochman, H. & Selander, R. K. 1983*b* Geographical components of linkage disequilibrium in natural populations of *Escherichia coli*. *Molec. Biol. Evol.* **1**, 67–83.