MINIREVIEW

Concept of a Bacterial Species

STEVEN KRAWIEC
Department of Biology, Lehigh University, Bethlehem, Pennsylvania 18015

The sequence of genes is a highly evolved feature of genomic organization in bacteria. Regions of the genome appear to be delimited by recurrent secondary structures in deoxyribonucleic acid. The demarcated sequence of genes can serve as a means of establishing bacterial species.

INTRODUCTION

The prevailing view of a eucaryotic species is expressed in a definition proposed in 1942 by Mayr, which states that “species are groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups” (13). The applicability of this definition to procaryotic species is seemingly challenged by the sexual character of bacteria. On the one hand, sexual activity in bacteria is so infrequent that sex is not seen to be a uniformly occurring species-specific character. On the other hand, when a strain does have the capacity to donate genetic information, the recipients are often not restricted to the same so-called species. Indeed, transmission of genetic elements among disparate procaryotes is well-documented (22). Either the absence of sex or promiscuity suggests that sexual compatibility is not a discriminatory feature among procaryotic species.

Our present knowledge of the organization of the genomes of some members of the Enterobacteriaceae may permit a further sense of the nature of a bacterial species. The thesis which is presented and examined here states that the essence of a species is not merely sexual compatibility, gene content, or phenotypic manifestation of gene content. Rather, the nature of a species emerges from a more fundamental feature of all these conditions, viz., the sequence of loci.

SEQUENCE OF CHROMOSOMAL LOCI AS AN IDENTIFYING CHARACTERISTIC OF BACTERIAL SPECIES

It has been asserted that procaryotes constitute a major portion of the extant biomass of the world (18). The individual cells of the diverse representatives in this vast domain of life are remarkably similar in colloidal size. The size of bacteria may be an expression of the surface required to transport nutrients into and waste products from metabolically active cells. The distribution of nutrients and wastes within the cytoplasmic volume is entirely by passive diffusion. Accordingly, the cell cannot become too large or parts of the cytoplasmic volume will become too distant from the surface (i.e., the area that refreshes the living mass). This limitation on the physical dimensions of procaryotic cells may, in turn, impose a physical limit on the size of the bacterial genome. The typical bacterial genome has (with little variation) a molecular weight of $1.0 \times 10^9$ to $3.0 \times 10^9$ (26). Two consequences of this restriction are that only genes of great selective value will be conserved and that the physical organization of the genome will be optimized. Both of these circumstances, genetic conservation and physical optimization, are manifest in the sequence of loci. Accordingly, this feature may be very informative about the fundamental nature of any procaryote.

This perspective has been clearly presented by Riley and Anilonis, who have written that “differences in gene arrangement may be more important in defining the identity of a bacterium than are differences either in nucleotide sequences of structural genes or in amino acid sequences of proteins” (24). These reviewers showed that many chromosomal arrangements of genes have been conserved despite the numerous mechanisms that might scramble the sequence of loci. The most detailed evidence comes from a comparison of the gene orders in Escherichia coli K-12 and Salmonella typhimurium. When the chromosomal maps of these two organisms are aligned, the vast majority of common genes are found to be in nearly equivalent positions. When corresponding loci are displaced from one another by more than 0.6 map unit, the two positions can be arbitrarily adjusted to create alignment. This process creates an excess length of genetic material in the map of one of the organisms. Such excesses are displayed as loops or bulges. The formation of the loops allows precise comparison of the extent of similarity between the two organisms. E. coli has 12 loops with respect to S. typhimurium, and S. typhimurium has 13 loops with respect to E. coli. The typical loop contains 0.6 to 1.5 map units of genetic material, although two loops contain more than 3 map units. In total, approximately 14% of each of the genomes is located in loops. Frequently, the loops have few identifiable genes or genes that are not common to the two bacteria. For example, the lac locus of E. coli is found on a loop. All in all, comparisons strongly indicate that the sequence of loci constitutes an adaptation that has been preserved during the evolution of these distinct but related enteric bacteria.

It has been demonstrated on countless occasions that microorganisms may retain their viability if genetic loci are rearranged. The hypothesis that sequences of loci have been optimized predicts that some adverse effect is, nonetheless, engendered by the rearrangement. A stringent examination of this claim has been performed with inverted and transposed chromosomal segments in E. coli (7, 8). This bacterium has seven redundant ribosomal ribonucleic acid (rrn) genes distributed in its chromosome. Because the rrnD locus has a polarity opposite to that of most other rrn loci, it is possible to generate inversions of the chromosomal lengths between rrnD and other oppositely oriented rrn loci. Accordingly, inversions between rrnD and rrnB or between rrnD and rrnE have been generated. In the former inversion, approximately 17% of the genome is inverted, and in the latter circumstance 18% is inverted. With these inversions,
there is no loss of information because the newly created
junctures generated by the inversion are in reiterated nucle-
otide sequences (7). The recurrence of the highly homolo-
gous rrn loci also allows the construction of transpositions,
i.e., translocation of chromosomal segments between two
adjacent rrn loci to the chromosomal position of some third
rrn locus. (Transpositions of this sort are sometimes called
transitions [see below].) The region between rrnB and rrnE,
which constitutes less than 1% of the E. coli genome, has
been transposed to four other rrn loci around the genome (8).
The fitness of the organisms having rearrangements in their
genomes has been assessed by comparing their growth rates
with the growth rate of organisms with unperturbed ge-
nomes. In particular, mixed cultures of organisms with and
without chromosomal aberrations were diluted \(10^8\)
and allowed to grow to maximum crop, at which time
dilution and growth were repeated. Five cycles of this
protocol (equal to approximately 100 generations) were
effected. All of the organisms with inversions exhibited
decreased growth rates. Furthermore, the greater the size of
the inversion, the greater was the reduction in growth rate.
Similar effects were seen with transpositions. Transposition of
the rrnB-rrnE segment to the nearby rrnC locus had a
minimal effect on growth rate, whereas transposition to a
distant locus, rrnG, had a maximal effect. Transposition to
loci intermediate between these extremes had intermediate
effects on the growth rates. Duplications in which the
segment between rrnB and rrnE was retained at the original
site as well as transposed to another rrn locus had even more
severe effects on the growth rate. These results, based on an
important criterion, show that strains of E. coli with inver-
sions, transpositions, or duplications suffer a competitive
disadvantage compared with strains having the same loci in
the wild-type arrangement. These observations indicate that
gene sequence, like gene content, is an important selective
feature and, hence, a fundamental character of an evolving
population.

A confounding feature of the genetic organization of
procaryotes is the prevalence of plasmids. Genes may be
located on the chromosome, or, alternatively, genes may be
on extrachromosomal elements. Many plasmids are conju-
gative or mobile and can be transmitted from one bacterium
to another. The flow of genes on extrachromosomal ele-
ments has led to the claim that the gene pools for bacteria
(which can accommodate plasmids) are several orders of
magnitude greater than would occur if genes were confined
to the chromosome alone. This circumstance has created the
notion that the bacterial genome is extremely plastic and,
therefore, difficult to define. However, Campbell (1) has argued
that the sorts of genes that are found on the chromosome and
the type present on the extrachromosomal elements are
different in character. The former (which he calls
euchromosomal) are a collection of coadapted genes that are
characteristic of a particular type of bacterium. Genes on
extrachromosomal elements are selected to function in the
cytoplasmic environment of diverse types of bacteria. This
interpretation indicates that the collection of euchromosomal
genes is more or less constant in character, whereas the
genomes on extrachromosomal elements are transiently asso-
ciated with a particular type of procaryote and, thus, can be
quite varied in nature without altering the basic identity of the
organism. This interpretation appears to conform to what is
observed in nature. For example, laboratory experi-
ments have demonstrated that genes from a wide variety of
organisms can be transferred to E. coli (22). However,
isolates of E. coli from various habitats lack these loci,
indicating that the extensive introduction of genes is more
artificial than natural. All of these observations suggest
that the identity of a bacterium should be established on the
basis of euchromosomal genes rather than the genes that
might be present on extrachromosomal elements as an
expression of some environmental stress.

PREDICTIONS GENERATED BY THE SEQUENCE OF
LOCI PREMISE

The premise that sequence of loci is an adaptive feature
and (literally) an inherent character of a species predicts that
bacteria have some way of monitoring the extent of any one
gene. If a bacterium can determine that the genes are in a
proper sequence, it will have had to determine where any
gene began, where it ended, and where the next gene began.

Because deoxyribonucleic acid (DNA) is a double-
stranded molecule, secondary structures can abound. The
common form of DNA, B-DNA, is in itself a type of
secondary structure. The rigidity of the B-DNA organization
varies with the primary structure (i.e., the nucleotide
sequence). DNA molecules are not static; rather, they flex or
breathe (21). The extent to which two complementary strands
are dissociated from one another in situ is an expression of
the composition of that region. Other secondary structures
are an absolute expression of primary structure; for ex-
ample, the occurrence of cruciform structures in DNA is an
expression of palindromes or dyads, i.e., regions where the
complementary nucleotide sequences exhibit twofold rota-
tional symmetry (11, 20). These and other organizational
features of double-stranded DNA could serve as signals that
allow organisms to detect the beginnings and ends of indi-
vidual genes. In other words, the structural organization
would demarcate functional units of the genome. Recogniz-
ing the precise extent of genes is a useful, if not an absolute,
precondition for monitoring the sequence of loci.

Zuckerkandl and Pauling (27) introduced the notion that
certain biological macromolecules can serve as “documents
of evolutionary history.” Especially useful are “semantopho-
retic molecules or semantides—molecules that carry the
information of the genes or a transcription thereof.” A
corollary of this claim is that comparisons of present day
semantides allow a measure of phylogenetic relatedness.
This concept may apply equally well to structures within
DNA that delimit the organization of the genome. Features
that demarcate the extents of genomic regions should be
evident in two ways, first by their recurrence in clones of
bacteria and second by their persistence in evolutionarily
related types of bacteria.

Distinctive intercistronic nucleotide sequences are preva-
 lent in various members of the Enterobacteriaceae (6).
These sequences, which extend for 50 to 100 bases, have
several notable features. First, they recur frequently. Se-
quences with approximately one-half the bases in common
have been found in both S. typhimurium and E. coli. Those
in the former organism occur at the junctures of the hisJ-
hisQ genes of the histidine transport operon and at the
juncture of the hisG-hisD genes of the histidine biosynthetic
operon (6). Those in the latter organism occur at the juncture
of the lamB locus, which specifies the lambda receptor, and
the molA locus, which has an unknown function (6). In
addition, homologous sequences occur in E. coli after the
tril gene, which regulates the tryptophan operon (6); the
glyA gene, which specifies shikimate 5-dihydroxy-
lysine transaminase (19); glnS, which specifies glutaminyl
transfer ribonucleic acid synthetase (19); fol, which specifies dihydrofolate
reductase (19); gdhA, which specifies a nicotinamide adenine
dinitrocelate phosphate-specific glutamate dehydrogenase (14); and mtlA, which specifies mannitol-specific enzyme II of the bacterial phosphotransferase system (10). It is notable that the sequences occur after regulatory genes, at the juncture of structural genes in an operon, and after what is presumed to be the last gene of an operon. A particularly dramatic example of such a sequence has been observed starting 520 base pairs after the thrC gene; there the sequence recurs three times (16). A computer survey of established nucleotide sequences has revealed 67 occurrences of the consensus sequence, and a statistical argument suggests that hundreds of such sequences probably exist in the E. coli genome (4). The second notable feature is that these related sequences contain several dyads (6). Prominent in the potential secondary structures are intrastrand stems that may include as many as 70 or so bases. In addition, there may be alternate secondary structures which have smaller stem structures involving approximately 25 bases (6). A regulatory function, modulation of transcription, has been postulated for these structures. However, this interpretation is not inconsistent with their also being demarcations of the extents of genes. For example, among 10 inversion mutants of S. typhimurium with the hisD locus fused to a novel promoter, 7 of the inversions occurred in the intercistronic sequence between the hisD and hisG loci (25). Furthermore, relatively short cruciform structures have been identified as being formally equivalent to sites at which crossing over is postulated to occur (12). A definitive interpretation of whether these sequences establish the limits of genes depends on knowing how recurrent they are, where they occur among related genes, what their relationship is to sequences created by genetic exchange, what other sorts of similar sequences occur at the termini of genes, and what types of genes lack such distinctive sequences.

The existence of precise duplications of genes suggests that bacteria have the capacity to assess the boundaries of a gene. Present day semantides serve as an ample record that whether contemporary bacteria have retained the ability to duplicate genes precisely and to use the exact duplications as adaptations. The prospect of genetic reiterations originating at specific sites appears to be well-founded since the highest degrees of similarity for some genes with functional related products occur upstream and downstream from the structural gene per se. In particular, among nucleotide sequences specifying the target site recognition capacity of restriction endonucleases, the greatest level of identity occurs on either side of the primary sequence that is translated into protein (5). Similarly, amplification of specific genes in a plasmid of Proteus mirabilis has been achieved through the introduction of identical nucleotide sequences on both sides of the gene which becomes reiterated. The sequences which are introduced into the genome are not related to naturally occurring insertion sequences. The flanking sites serve as positions at which recombinational events required for amplification occur (47). Furthermore, recent observations regarding the suppression of topoisomerase I-deficient mutants by a topoisomerase compensating locus (toc) have been interpreted to mean that "fusion joints" themselves are functional and can be selected (A. R. J. Purcell, E. Gamliel, D. J. Zabel, and R. E. Depew, Abstr. Annu. Meet. Am. Soc. Microbiol. 1984, H172, p. 120). Enzymatic evidence also supports the notion that one gene and not its immediate neighbor can be amplified, a circumstance which suggests that amplification is a structurally based process rather than an accident. In particular, mutants of Klebsiella pneumoniae which gain the ability to metabolize xylitol selectively amplify the rrbD gene, which specifies ribitol dehydrogenase, and not the adjacent gene, rrbK, which specifies D-ribulose kinase. The efficiency of this response is clear. The organism exploits a latitude in the dehydrogenase that allows it to respond to an epimer of its natural substrate. Effective use of xylitol is achieved by overproduction of the dehydrogenase from reiterated templates. A similar strategy is not required for ribulose kinase since the organism possesses a naturally occurring xylulose kinase (9). Finally, the most compelling evidence for selective reiteration of chromosomal segments comes from mutants of E. coli that are resistant to high levels of ampicillin because they hyperproduce β-lactamase (3). Data indicate that as many as "10 identical repeats [which, as it occurs, are larger than the selected gene] organized in tandem can be generated [at] a normal mutation frequency in E. coli" (3). Furthermore, "the presence of short DNA repeats flanking the selected resistance marker is apparently important for the generation of specific amplification" (3).

Genes present in one type of bacterium can often be transferred to a closely related strain that lacks these genes. Indeed, types of bacteria are often distinguished by the presence of a phenotype conferred by the presence of specific sets of genes in one type of bacterium and the absence of the phenotype and relevant genes in another type. It is informative to determine the locations at which distinctive sets of genes integrate into otherwise similar genetic organizations. The 17 nif genes that specify the nitrogen fixation phenotype in K. pneumoniae can be transferred by conjugation into E. coli. When this transfer occurs, the 24 kilobases that comprise the nif loci integrate at a position between the his and shiA loci, a position that is equivalent to that of the nif loci in K. pneumoniae (2). Likewise, the genes for arabitol and ribitol catabolism can be transferred by transduction from E. coli C, where they occur naturally, to E. coli K-12, which has no genes for pentitol utilization. The four structural genes and the four regulatory genes of the two operons integrate in E. coli K-12 between the metG and attP2 loci. This position is the ordinary location of the genes for pentitol utilization in E. coli C (23). These examples suggest that there are preferred sites for genes in Enterobacteriaceae chromosomes. Whether these sites are recognized by homology, site-specific sequences, secondary DNA structure, or some other genetic signal awaits nucleotide sequence analysis of the juncures in hybrid chromosomes.

**FUNDAMENTAL BIOLOGICAL MEANING OF SEQUENCE OF LOCI**

A bacterium that competes and succeeds in an environment is said to be adapted to that environment. Another way to describe this circumstance is to state that the organism has an optimal genotype for its niche. The emergence of a collection of genes well-suited to the environment of the organism is a consequence of selection. Populations of such organisms can perpetuate themselves if the assemblage of genes that confers the reproductive advantage is both preserved and not distributed to other similar populations. Having genes arranged in a particular order on a chromosome is a means of reducing exchange with organism
isms that do not have a homologous arrangement of loci. Sequence of loci, then, both establishes the identity of a species and preserves the identity of a species by creating a species barrier to the exchange of genes.

The species barrier is not absolute. As noted previously, promiscuous plasmids can sometimes be exchanged among gram-negative families with high frequencies. Euchromosomal genes, by contrast, are stable. It is the coadapted collection of euchromosomal genes which have been preserved in a characteristic sequence for eons that identifies a bacterial species by a biologically meaningful criterion.

**BASIC UNITY WITH RESPECT TO THE CONCEPT OF A SPECIES IN EUCARYOTES AND PROCARYOTES**

A eucaryotic species is defined to a great extent by sexual compatibility. What this criterion means is that gametes bearing equivalent sets of chromosomes fuse during the process of syngamy to generate a zygote. In the mature organism that arises from the zygote, gametes are generated once again as a prelude to establishing the next generation. Repetition of this cycle gives temporal continuity to the species. An essential event in this on-going sequence is meiosis, the process that produces gametes. During prophase I of meiosis, homologous chromosomes synapse; i.e., the loci of homologous chromosomes align in perfect register. Closely related organisms that are sexually incompatible fail in this process because the organizations of the chromosomes in the chromosome sets are not equivalent. The importance of synapsis reveals that sequence of loci in chromosomes is a fundamental feature of eucaryotes that is intimately related to the concept of a eucaryotic species. Thus, the interpretation presented here asserts that the concepts of species for procaryotes and eucaryotes are essentially identical.

**LITERATURE CITED**


