17

Bacterial and Bacteriophage Evolution

One of the exciting changes that has occurred in the field of bacterial genetics in the past decade is the development of methods and information permitting the study of real evolutionary genetics with respect to bacteria and their genomes. Sufficient data are now available to permit researchers to address problems similar to those that have held the interest of geneticists studying eukaryotic organisms for many years. Although the bacterial evolutionary situation is not totally understood, our present comprehension has shed new light on classic problems in genetics and offers the promise of even more interesting results in the years to come.

Major topics include:

- Some of the possible measures of evolutionary change
- Uncertainties in assessing relatedness of organisms
- Changes in genomes with time; the emergence of new functions

- Identifying the function of unknown proteins
- The organization of genes on chromosomes
- One theory about how present-day chromosomes arose

What Is Evolution?

In the most literal sense, evolution is change occurring over time. For a geneticist, the change in question is, of course, base changes in the nucleic acid comprising the genome of a cell or a virus—in other words, a mutation of the sort discussed in Chapter 3. Most people's understanding of evolution is that it is a gradual process. However, that cannot be absolutely true. Mutations occur in an all-or-none, discontinuous manner. Either a genetic locus is mutant or it is not. Whether that mutation has any effect on the cell phenotype is an issue discussed in Chapter 3.

Implications of Strain Differences

The classic view of evolution is the one that has been proposed by Charles Darwin—the "survival of the fittest." This concept implies that as mutations accumulate in a population of individuals, some individuals will be better able to survive and to reproduce than others. The survivors are, therefore, better fitted to the current set of environmental conditions. An important point that is often forgotten, however, is that better fitted is not the same thing as best fitted. The end product of evolution is not necessarily the best-adapted organism for a particular ecological niche that could possibly be developed; it is merely better adapted than any of its competitors.

In many cases, experiments have shown that a particular cellular process is optimized rather than maximized. A good example of optimization is seen in the case of bacterial ribosomes. Charles Kurland and coworkers have demonstrated that ribosomes in the standard laboratory strain of *Escherichia coli* can mutate to have greater translational fidelity or greater speed, but they cannot maximize both the traits at the same time. Mikkola and Kurland tested seven natural isolates of *E. coli* and showed that their growth rates were initially relatively variable and generally slower than laboratory strains, although the mass of cells in a stationary phase culture was roughly constant for all strains tested. This observation indicates that the efficiency of glucose utilization is essentially

maximized in all strains. Mikkola and Kurland tested their strains using a chemostat, a continuous culture device in which fresh medium is continuously added and old medium removed. After 280 generations, the growth rates for the natural isolates converged on the value seen in laboratory strains, although the cell mass per culture volume remained unchanged. In vitro tests demonstrated that the ribosomes had mutated to better optimize mRNA translation, but the nature of the mutations was not determined. The conclusion to be drawn from these results is that current growth conditions apply selection to cultures, and thus extensive laboratory studies do not necessarily reflect what is happening in the natural environment.

Another point to be considered is the founder effect. When a population arises from a few individuals, the distribution of alleles within the founding members may not reflect the distribution of alleles within the worldwide population. Thus, some unusual genes may be present just by chance. These unusual genes may or may not contribute to the overall fitness of the final population, but their rarity in the world at large guarantees that geneticists examining the isolated population will notice them. Thus, the mere presence of a gene in a successful population does not necessarily indicate a beneficial effect from that gene.

THINKING AHEAD

What is the advantage to a cell in having DNA that cannot be expressed? If there is no advantage, why do cells keep untranscribed DNA in their genomes instead of reducing the genome size so that they can replicate faster?

Cryptic Genes

It is also important not to substitute human judgment for observation of natural processes. For example, it seems intuitively obvious that **cryptic genes**, genes that are never expressed, should be undesirable. However, cryptic genes are commonly found in bacteria. *Lactobacillus*, *Bacillus*, *Escherichia*, *Salmonella*, and many other genera contain inactive genes encoding functions that prove useful under particular circumstances. These genes may code for amino acid biosynthesis, sugar degradation, or more unusual metabolic functions. They can be activated by mutation and appropriately selected, but in some cases when the selection is removed the genes soon become cryptic again.

The best-studied examples occur in *E. coli* and involve genes for transporting and metabolizing rare sugars. They are the *bgl*, *cel*, and *asc* operons. The *bgl* operon codes for three proteins (Fig. 17.1). The first is an antiterminator protein (BglG) that can be phosphorylated. In its dephosphorylated form, it is an antiterminator of transcription. The second is a member of the phosphoenolpyruvate-dependent sugar transport system (BglF), an example of PTS enzyme II. This particular enzyme transports salicin or arbutin, and phosphorylates the sugar at the same time. When inducing substrates are absent, it phosphorylates BglG. The third Bgl protein is a phospho-β-glucosidase (BglB) that degrades the sugars. Standard laboratory strains of *E. coli* do not express the *bgl* operon because of a defective promoter. Insertion of IS*1* or IS*5*, 78–125 bp upstream of the promoter or base substitutions in the CRP-cAMP binding site of the promoter, can activate the operon. While it might seem that activation by insertion would be a useless characteristic, in fact it is not particularly rare. Manna et al. (2001) have shown that the relevant area is a hot spot for insertion of phage Mu. Tn*1* and Tn*5* also preferentially insert here. Transcription always begins in the same location, so the mechanism of activation is not simple creation of a new promoter. There are two transactivators of the *bgl* operon known, *bglJ* and *leuO*. Neither is normally expressed.

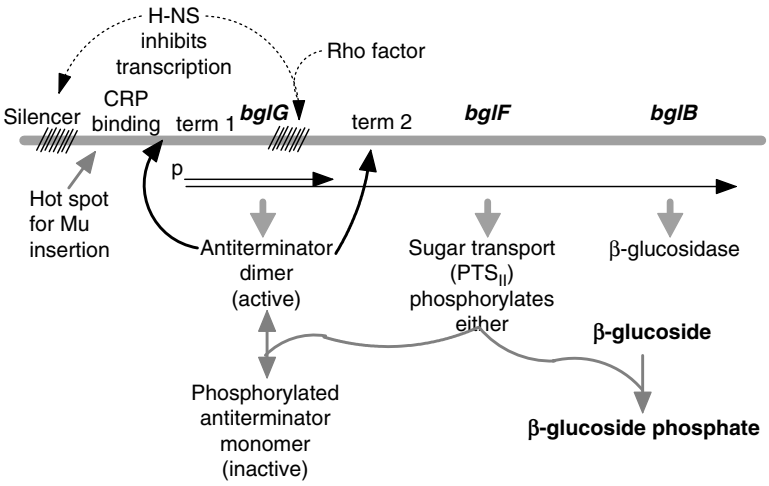

Figure 17.1. The *bgl* operon of *E. coli*. This operon is normally cryptic because the promoter is inactive. Insertions near the CRP binding site can activate the promoter as can certain proteins synthesized elsewhere on the genome. The operon includes a gene for sugar transport, degradation, and regulation.

Barry Hall and coworkers detected the *asc* operon during a series of experiments on directed evolution. They began with a strain deleted for the *bgl* operon and a related operon called *cel*. Sequentially, they selected for strains that could first use arbutin, then arbutin and salicin, and finally arbutin, salicin, and cellobiose. Each selected event could be either the return of function to a pre-existing cryptic gene or an alteration in substrate specificity of the protein product of an already expressed gene. Hall and Xu (1992) have shown that there are strong sequence homologies between the Asc active proteins and those of the *bgl* operon. Interestingly, the Bgl regulatory protein is a positive regulator, while the Asc regulatory protein is a repressor with strong homology to *galR*.

The questions that obviously arise concerning cryptic genes are how they originate and why are they maintained in the population. Cryptic genes that resemble other genes in the same genome are taken to be **paralogous**, meaning that they result from DNA duplication followed by independent evolution of the two sets of genes. In the present discussion, the *bgl* and *asc* operons would be considered paralogous, given the sequence similarities demonstrated by Hall and Xu. Their simultaneous maintenance in the population raises some additional issues.

Directly repeated sequences are recombinationally unstable, tending to form loops that are excised like a λ prophage. This instability can be counteracted by sequence divergence within the paralogous genes and by physical distance between the duplications. Because the intervening DNA would be lost during recombination, recombinants formed from widely separated duplicated genes are likely to be inviable. In the case of *bgl* and *asc*, they are separated by about 26 minutes of DNA, including genes for all the ribosomal proteins. Therefore, recombinational loss is unlikely.

Note that all the cryptic genes just discussed code for normal proteins that can be immediately functional. The defects in the operon are all at the level of transcription, not in the protein-coding regions. Mutations do occur in the coding regions (the sequences are not identical), but harmful mutations must be disadvantageous. The only way that could happen would be if the cryptic genes are occasionally expressed. Khan and Isaacson (1998) reported that just because an operon is not expressed in the lab does not mean that it is never expressed. They used a reporter gene fused to the *bgl* promoter to show that even though the *bgl* operon is not transcribed in laboratory cultures, it is expressed during infection of mouse liver. One possible explanation would be that something in the mouse liver induces production of LeuO protein that subsequently activates *bgl* transcription.

There is, however, evidence that for everyday metabolism, it is critical that the *bgl* operon should NOT be expressed. In addition to the inactive promoter and the requirement for inducer and antitermination by BglG, the histonelike protein H-NS binds to two silencer sites (Fig. 17.1). The one upstream prevents promoter activity, and the one downstream facilitates an interaction with Rho protein that causes transcription termination (Dole et al. 2004). Nevertheless, there is a report that a sigma-factor-like protein from *Streptomyces* can activate the *bgl* operon by itself (Baglioni et al. 2003).

Arthur Koch (2004) has argued that most mutations that occur to activated cryptic genes are of an easily revertible type. If so, that would constitute another indication of the importance of keeping cryptic genes turned off except in times of significant need. Koch also considered how that effect may interact with the observations of directed mutation (see Chapter 3).

Another example of paralogous genes occurs in *Borrelia*. The organism has a linear chromosome, 12 linear plasmids, and 9 circular plasmids. Many of the plasmid genes are either paralogs of chromosomal genes or **pseudogenes**, nonfunctional genetic structures that have a similar base sequence to other chromosomal genes.

Expression of Evolutionary Relationships

Discussions on evolution invariably result in the necessity for graphical representations of relationships between organisms. The most usual type of diagram is the evolutionary tree (Fig. 17.2). In such a tree, distances along lines represent evolutionary time. Closely related organisms occur as physically close branches of the tree. Distantly related organisms are more widely separated. If the tree

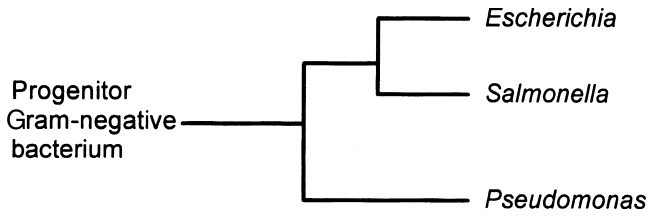

Figure 17.2. An evolutionary tree. This sample tree suggests that there was a single bacterium that gave rise to three different genera. *Pseudomonas* diverged from others earlier than *Escherichia* diverged from *Salmonella*. Divergence of the latter two was not a recent event because the branches have significant length.

includes an organism as the progenitor of all others as in Fig. 17.2, the tree is rooted. The toothlike arrangement of some trees has led scientists to refer to them as dendrograms.

Preparation of an evolutionary tree requires data that can be used to measure relatedness of organisms. Originally, these data would include such variables as physical size, shape, habitat, and so forth. However, these variables are unsatisfactory because of the limited number of possible types compared to the number of known bacteria. For example, there are only four or five fundamental cell shapes among bacteria. With the advent of macromolecular sequencing techniques, evolutionary biologists now speak of sequence similarities and use them as the measure of relatedness. The more precise method is to use nucleic acid sequence data because these will show all the changes that have occurred. However, the amino acid sequence is also informative because conserved sequence suggests functional importance. These sequencing techniques for classification are more satisfactory than the earlier methods but they still have potential pitfalls for the unwary.

Gene sequences can be similar for several reasons. One is that the genes are **orthologous**, descended from an unknown, common ancestor. Alternatively, one member of the set is the progenitor for all other members, meaning that they are **homologous**. Orthologous genes can potentially provide information about evolutionary distance. Another possibility, however, is that there is basically only one way to enzymatically perform a certain task. Therefore, if an organism makes an enzyme with that property, the shape of the enzyme will be essentially similar to all other such enzymes regardless of evolutionary lines of descent, making the genes **homoplastic**. This is an example of convergent evolution.

Molecular biologists now have sequence data from a wide variety of organisms available to them. These data are most useful when the genes being compared are highly conserved so that the changes accumulate very slowly over the time interval being examined. Many studies involve the smaller ribosomal RNA molecules because they are easy to obtain in large quantities and can be readily sequenced. Carl Woese and coworkers have published 16S rRNA sequences for a variety of organisms and expressed their data as evolutionary trees. While RNA sequences can be absolute, controversy has developed over the exact (computer-driven) methodology to be used in preparing the evolutionary trees. The Woese group uses distance measurements, estimating the fraction of positions at which the sequences differ. The assumption is that larger fractions mean greater distances on the tree. James Lake has proposed alternative

calculations called evolutionary parsimony that lead to different taxonomic structures for the Bacteria and the Archaea. These calculations attempt to reconstruct step-by-step changes that might have occurred to give present-day sequences.

The fundamental principle in tree building is that of parsimony—developing the simplest tree that accommodates all of the data. Although any sort of sequence data can be used, the most desirable is DNA sequence. Protein sequence data suffer from the existence of synonymous codons, meaning that some mutations are not apparent in the amino acid sequence. One of the first issues is having DNA sequence data with enough informative bases. Sequence in and of itself is not sufficient to provide evolutionary information. There must be sequence differences between molecules from different organisms that can be used to choose between trees. If all possible trees require the same number of base changes to achieve a particular sequence difference, the difference is not phylogenetically informative (Fig. 17.3).

The issue of informative bases is one of balance. If organisms have diverged relatively recently (in an evolutionary sense), only a gene that accumulates mutations quickly will offer enough informative sites. On the other hand, if two organisms diverged a long time ago, only relatively stable genes will retain enough of their original sequence to provide information about the relatedness of the organisms. Major taxonomic trees are based on rRNA sequences because they offer a good mix of properties. Overall ribosome structure is strongly constrained by the variety of tasks that must be performed for a cell to

	↓ ↓ ↓
Progenitor sequence	**A A G G C C T T**
Descendant 1	**A T G G G C T T**
Descendant 2	**A T C G G C T T**
Descendant 3	**A T C G C C T T**

Figure 17.3. The phylogenetic information content of nucleotide sequence changes is variable. *Arrows* indicate sequence differences. The first difference is not informative, however, because all descendants have it. The second two differences begin to establish lines of descent. Descendant 2 is less closely related to the progenitor than the other two descendants because it has two sequence differences while the others have only one each. The data are not sufficient to explain how descendant 2 is related to either descendant 1 or 3.

survive and be competitive; hence there are regions of rRNA that change very slowly. These relatively unchanging regions provide information about relationships corresponding to geologic time. There are, however, other regions of rRNA that are less critical and do accumulate mutations, and these regions offer data about evolutionarily recent changes.

A more subtle problem is one of sequence alignment. Most DNA sequences do not align perfectly. Each organism has accumulated its share of small deletions and insertions. Therefore, a simple sequence alignment that begins at the 5′-end is unlikely to be satisfactory (Fig. 17.4). Instead, spaces need to be added to maximize the similarity, but that addition presents its own problems. Which space should be added first? A space added near the 5′-end will shift all sequences downstream, whereas a space added near the 3′-end has a much smaller effect. If the sequence in question encodes a protein, the proper reading frame must be preserved. If the sequence is for an rRNA molecule, the proper secondary and tertiary structures must be maintained. Once again, special computer programs are available to prepare and compare possible sequence alignments. The basic strategy is to align the sequences so that only a minimum number of changes are needed to produce the diverged sequences under study. Qi et al. (2004) have presented a new computer modeling program for evolutionary tree construction using protein sequence that attempts to answer some of the earlier criticisms.

Even the base composition of the DNA molecules being compared can have an impact on tree analysis. Steel and coworkers (1993) have presented a mathematic analysis showing that if two organisms have the same (G + C) content in their DNA, problems with evolutionary trees can arise. Standard methods of calculating trees assume that similar (G + C) content implies relatedness, whereas in fact the similarity may be due to convergent evolution. Steel and coworkers provide equations to discount random similarities resulting from similarities in base composition. The effect is most dramatic when the (G + C) content, which can range from 25% to 75%, is either very high or very low.

Organisms can and do differ in the % (G + C) in their DNA, mainly due to wobble in the third base of codons. Lawrence and Ochman (1997) present data to show that the % (G + C) of the third base in a codon increases faster than the increase in the overall % (G + C) of different organisms, reaching as high as 95% in *Micrococcus luteus*. On the other hand, the % (G + C) in the second position increases very slowly, ranging from a low of 33% to a high of 45%. The first base in the codon falls in between the other two ranges. Therefore, base changes in position 3 of the codons have little taxonomic meaning because that position

```
Ec  ATGAAAGCGT TAACGGCCAG GCAACAAGAG GTGTTTGATC TCATCCGTGA 50
St  .......... .......... .......... .......... .......G.. 50
Ecc .......TA. ....A..A.. ...G..GC.. ..T.A...C. .......C.. 50
Pp  ...TTGAAAC G...C.AC.. C...GCC..A A.TC.C.CGT .....AAGCG 50
Pa  ...C.GAA.C G...C..C.. C...GCC... A.CC.CTCCT .....AAGCG 50

Ec  TCACATCAGC CAGACAGGTA TGCCGCCGAC GCGTGCGGAA ATCGCGCAGC 100
St  .......... .......... ....A..... .......... .......... 100
Ecc ...T..TGCG ..A..C..A. .......A.. ...G...... ..T..T..A. 100
Pp  CTG.C.GGAA G.C.AC..CT .C.....T.. C..C..C..G ..T..T...G 100
Pa  CTG.C.GGAA G.CCAC..CT .T........ C..G...... .....C...G 100

EC  GTTTGGGGTT CCGTTCCCCA AACGCGGCTG AAGAACATCT GAAGGCGCTG 150
St  .......... .......... ........G. .......C.. T..A...... 150
Ecc AAC.A..... T..C..T..C ..T....... .......... ...A..A... 150
Pp  AGC....C.. .AAG..G..C ..T..C..C. .G..G..C.. C.....C..T 150
Pa  AAC.C..C.. .AAG..G..G .....C..C. .G..G..C.. C......... 150

Ec  GCACGCAAAG GCGTTATTGA AATTGTTTCC GGCGCATCAC GCGGGATTCG 200
St  ..G....... .G..GC.... ...C...... .........G A..T..C... 200
Ecc ..G..T.... .T..G..... ......A..G .....G..T. .T..T..... 200
Pp  ..C.....G. ...CG..C.. ...GACGC.G .....T..C. ....C..C.. 200
Pa  ........G. ...CC..C.. ...GAC.C.G .....C..T. ....C..C.. 200

Ec  TCTG------ ---------- --TTGCAGGA AGAGGAAGAA GGGTTGCCGC 232
St  ....------ ---------- --..A..... ...A..G..C ..A..A.... 232
Ecc ....------ ---------- --C..AT... ...A..GACG ..TA.T..T. 232
Pp  CA.CCCTGGC CTGGAAGCCA AGGCT----- -..A....CC ..CC....CA 244
Pa  CA.TCCCGGC TTCGAACCGC ATGCCGCCA. C..C..T..G ..CC.....G 250

Ec  TGGTAGGTCG TGTGGCTGCC GGTGAACCAC TTCTGGCGCA ACAGCATATT 282
St  .T..C..G.. ...C..G..G ........G. .......... G......... 282
Ecc ....T..... C.....C..A ........G. .G........ GG.A..C..C 282
Pp  .CA.C..C.. G..C...... ....CG..GA ....T..CG. ......C..C 294
Pa  ..A.C..A.. G..C..C... ..C.C...GA .C..C..CG. ....A.C..C 300

Ec  GAAGGTCATT ATCAGGTCGA TCCTTCCTTA TTCAAGCCGA ATGCTGATTT 332
St  .....C.... .C........ ......GC.G .....A..C. GC........ 332
Ecc ...T...GC. .......T.. C...G.GA.G .....A..C. GC........ 332
Pp  ..GCAATCC. GCA.CA..A. C...G....C ...C.C..CC .A..C..C.A 344
Pa  ..G.AATCC. GC.G.A..A. ...CG....C .....T..TC GC..C..C.A 350

Ec  CCTGCTGCGC GTCAGCGGGA TGTCGATGAA AGATATCGGC ATTATGGATG 382
St  .......... ..G.....T. .......... ...C.....T .......... 382
Ecc TT.......G ..G.....C. .......... .A.......T .......... 382
Pp  TT........ ..ACA...C. ..AGC..... G..CG..... ..CT.C..C. 394
Pa  ....T..... ..GC....C. ..AGC..... G..C...... ...C.C..C. 400

Ec  GTGACTTGCT GGCAGTGCAT AAAACTCAGG ATGTACGTAA CGGTCAGGTC 432
St  .G..T..... ...G..A... .....G.... .C..C..... T..C.....G 432
Ecc .C..T..A.. ...C...... .....AG.A. ....G..A.. T......A.T 432
Pp  .C...C.... ...G.....C .CCTGC.GT. .A.CC..C.. ...C...A.. 444
Pa  .C...C.... ...C..C..C GTC..C.GC. .A.CG..C.. ...C.....G 450

Ec  GTTGTCGCAC GTATTGATGA CGAAGTTACC GTTAAGCGCC TGAAAAAACA 482
St  ........T. .C........ T.....G..A .A..A..... .......G.. 482
Ecc ..C.....G. .......C.. T.....G..G .G........ .......G.. 482
Pp  .....G..G. .C..C.G... ......C... ..C......T .C..GCGCG. 494
Pa  ..G.....G. .G..C.GC.. G..G..C..G ..G..A...T .C..GCGCG. 500

EC  GGGCAATAAA GTCGAACTGT TGCCAGAAAA TAGCGAGTTT AAACCAATTG 532
St  .......... ..G..G...C ....G..... C......... .C...G..A. 532
Ecc ........CG ..AC.C..CC .CG.T..... .GAA...... GCC..G.... 532
Pp  A....GC... ...TGG...C .TG.C..... CCC...A..C GCC..C..C. 544
Pa  A....GC..G ...TGG...C ..G.G..... CCCT.....C GCT..G..C. 550

Ec  TCGTTGACCT TCGTCAGCAG AGCTTCACCA TTGAAGGGCT GGCGGTTGGG 582
St  .G..G..T.. G..CG.A..A .....T..G. .......... ......A..C 582
Ecc .T..C..... G........A .....TT.G. .A.....CT. A........C 582
Pp  AA..C..... GAAAG.A... GAGC.GGTG. .C..G..CT. .AGC..C..C 594
Pa  AA..C..T.. GAAGG..... GAAC.G.T.. .C.....CT. .AGC..C..C 600

Ec  GTTATTCGCA ACGGCGACTG GCTGTAACAT ATCTCTGAGA CCGCGTGCCG 632
St  ..C....... .......A.. .T..TAGTCT CTTTTTAATC TCCTTGTAAG 632
Ecc ..G.....T. ..A..... .AGCTAACCA TTCGCAGAGA TGCACTGCTC 632
Pp  ..C......C G.TGA----- -------TCC AGGAGGCGTC ATGCAGCAGT 632
Pa  ..G..C..AC G.TGA----- -------CAG GAGATACCAT GCAGACCTCC 638

Ec  CCTGGCGTCG CGGTTTGTTT TTCATCTCTC TTCATCAGGC TTGTCTGCAT 682
St  CCGCCATCCG GCAATCGTGT AGCCTGATGG CGCTGCGCTT ATCAGGCCTA 682
Ecc GTCTGTCTGG GTCAGTGCAT CATGTCCTGT ATTCATCGTT TAGCGTGTTA 682
Pp  TCATTCACGC ACCCGAGCAA GCCCAGTTGC CCCTGTTCGA AGCATTCCTC 682
Pa  CACTCGCTGC CCAGCGCCCA GTTGCCACTG TTCCAGGAAG CGTTCTGGGC 688

Ec  GGCATTCCTC ACTTCATCTG ATAAAGCACT CTGGCATCTC GCCTTACCCA 732
St  CGGGAATGCA GTTCCTGAGA TGATTAATTT GTAGGCCGGA TAAGGCGTTA 732
Ecc ATCTGCTAAC CATATATATT TAGTTACATT TCGCGCGCAT TTTCTACGAT 732
Pp  GCCCAGCCCG TGCTGCCAGG CCTGAAAGCC AGGGAACCGG CGCGCAAGAG 732
Pa  CAGCAACGGC GCTCCCTTGC TCGACGATGT CATCGACAGC CCTTCCAGCG 738

Ec  TGATTTTCTC CAATATCACC GTTC 756
St  CGTCGCCATC CGGCAATGCG CTCG 756
Ecc CCCTATTACC CTCTGTTTTT TCAC 756
Pp  CAGCCAGCCC GAGCTGTTCA GCGA 756
Pa  CCTCCATCGA GGAACCCGCT GCCT 760
```

Figure 17.4. DNA sequence comparison of the coding region of the *lexA* genes and the sequences immediately downstream from the termination codons of *E. coli* (Ec), *S. enterica serovar* Typhimurium (St), *Erwinia carotovora* (Ecc), *Pseudomonas putida* (Pp), and *Pseudomonas aeruginosa* (Pa). The nucleotide sequence of the *lexA* gene of *E. coli* is shown for comparison. The nucleotides are numbered starting from the first nucleotide of the ATG initiation codon. The termination codon is at about residue 606. Both these codons are underlined. Identical residues in the open reading frame are depicted by a *dot* and nucleotide substitutions are indicated by the appropriate letter. *Dashes* indicate insertions needed to maintain the alignment. (From Garriga, X., Calero, S., Barbe, J. [1992]. Nucleotide sequence analysis and comparison of the *lexA* genes from *Salmonella typhimurium*, *Erwinia carotovora*, *Pseudomonas aeruginosa*, and *Pseudomonas putida*. *Molecular and General Genetics* 236: 125–134.)

is constrained to conform by the overall % (G + C) in the organism, and they should not be heavily weighted in the construction of phylogenetic trees. Changes in the second base should be the most informative, if there is a method to identify them. For a review of the genetic code, see Table 17.1.

Two methods have been proposed for resolving the problem of sorting evolutionary fluctuation from self-imposed restrictions on the code. The first depends on identifying **signature sequences**, regions in the aligned proteins where sequence changes are very informative (i.e., present in all members of one group and not in members of another). Gupta (1998) argues forcefully that only comparisons of protein sequences can avoid the problems inherent in DNA composition.

However, there is another possible approach (Karlin et al. 1998). This method takes advantage of the enormous amount of DNA sequence now available. The authors have prepared a mathematical analysis showing how the relative abundance of dinucleotides in each DNA strand is characteristic of each organism, a **genome signature**. Interestingly, the dinucleotide statistics are sufficient to reflect nearly all of the nonrandomness in genomes. There is little advantage to trinucleotide or tetranucleotide analysis according to their data.

Assuming that genetic relatedness has taxonomic significance, many evolutionary biologists have developed trees that attempt to graphically illustrate both the relatedness of particular genera or species and their common ancestry. Carl Woese and coworkers have been particularly active in this regard, and a sample of their work is presented in Fig. 17.5. Based on studies of 16S rRNA, they have proposed that all living organisms should be categorized into three domains: *Archaea*, *Bacteria*, and *Eukarya*, with eukaryotes descended from archaea.

Table 17.1. Differences in codon usage in *Rhodobacter capsulatus*: Comparison of codon usage in *Rhodobacter* and *E. coli*.

Amino Acid	Codon	Fractional Codon Usage for Each Amino Acid					
		Rhodobacter	*E. coli*	Rc-Fru	R-Pho	R-Nif	R-Crt
Gly	GGG	0.22	0.02	0.31	0.14	0.18	0.29
Gly	GGA	0.03	0.00	0.01	0.03	0.02	0.04
Gly	GGU	0.09	0.59	0.07	0.12	0.10	0.08
Gly	GGC	0.66	0.38	0.61	0.70	0.71	0.59
Glu	GAG	0.52	0.22	0.61	0.44	0.52	0.56
Glu	GAA	0.48	0.78	0.39	0.56	0.48	0.44
Asp	GAU	0.38	0.33	0.48	0.25	0.43	0.41
Asp	GAC	0.62	0.67	0.52	0.75	0.57	0.59
Val	GUU	0.09	0.51	0.08	0.09	0.08	0.09
Val	GUC	0.47	0.07	0.44	0.50	0.45	0.42
Ala	GCG	0.46	0.26	0.52	0.37	0.46	0.50
Ala	GCA	0.03	0.28	0.02	0.04	0.02	0.03
Ala	GCU	0.05	0.35	0.02	0.11	0.03	0.04
Ala	GCC	0.46	0.10	0.44	0.48	0.49	0.43
Lys	AAG	0.76	0.26	0.74	0.80	0.67	0.80
Lys	AAA	0.24	0.74	0.26	0.20	0.33	0.20
Asn	AAU	0.18	0.06	0.44	0.05	0.18	0.30
Asn	AAC	0.82	0.94	0.56	0.95	0.82	0.70
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00
Ile	AUA	0.01	0.00	0.02	0.00	0.00	0.00
Ile	AUU	0.07	0.17	0.06	0.05	0.10	0.07
Ile	AUC	0.92	0.83	0.91	0.95	0.90	0.93
Thr	ACG	0.32	0.07	0.38	0.24	0.36	0.42
Thr	ACA	0.02	0.04	0.03	0.01	0.02	0.01
Thr	ACU	0.04	0.35	0.03	0.04	0.02	0.05
Thr	ACC	0.62	0.55	0.57	0.71	0.60	0.52
Trp	UGG	1.00	1.00	1.00	1.00	1.00	1.00
Cys	UGU	0.12	0.49	0.25	0.00	0.13	0.10
Cys	UGC	0.88	0.51	0.75	1.00	0.88	0.90
Tyr	UAU	0.42	0.25	0.58	0.31	0.42	0.54
Tyr	UAC	0.58	0.75	0.42	0.69	0.58	0.46

Table 17.1. (*Continued*)

Amino Acid	Codon	Fractional Codon Usage for Each Amino Acid					
		Rhodobacter	*E. coli*	Rc-Fru	R-Pho	R-Nif	R-Crt
Phe	UUU	0.14	0.24	0.33	0.06	0.21	0.16
Phe	UUC	0.86	0.76	0.67	0.94	0.79	0.84
Ser	UCG	0.56	0.04	0.57	0.59	0.45	0.55
Ser	UCA	0.01	0.02	0.00	0.01	0.02	0.01
Ser	UCU	0.04	0.34	0.03	0.03	0.02	0.05
Ser	UCC	0.17	0.37	0.20	0.15	0.20	0.13
Ser	AGU	0.02	0.03	0.02	0.02	0.01	0.03
Ser	AGC	0.21	0.20	0.18	0.19	0.30	0.23
Arg	CGG	0.28	0.00	0.27	0.14	0.32	0.38
Arg	CGA	0.02	0.01	0.00	0.01	0.02	0.03
Arg	CGU	0.11	0.74	0.07	0.20	0.08	0.13
Arg	CGC	0.55	0.25	0.63	0.63	0.54	0.44
Arg	AGG	0.02	0.00	0.03	0.01	0.02	0.01
Arg	AGA	0.01	0.00	0.00	0.00	0.01	0.01
Gln	CAG	0.80	0.86	0.79	0.85	0.84	0.77
Gln	CAA	0.20	0.14	0.21	0.15	0.16	0.23
His	CAU	0.44	0.17	0.69	0.19	0.45	0.57
His	CAC	0.56	0.83	0.31	0.81	0.55	0.43
Leu	CUG	0.64	0.83	0.76	0.54	0.66	0.61
Leu	CUA	0.00	0.00	0.01	0.00	0.00	0.00
Leu	CUU	0.13	0.04	0.10	0.09	0.13	0.19
Leu	CUC	0.19	0.07	0.11	0.35	0.16	0.12
Leu	UUG	0.04	0.03	0.03	0.02	0.05	0.08
Leu	UUA	0.00	0.02	0.00	0.00	0.00	0.00
Pro	CCG	0.61	0.77	0.60	0.76	0.55	0.58
Pro	CCA	0.01	0.15	0.00	0.00	0.03	0.01
Pro	CCU	0.03	0.08	0.01	0.02	0.04	0.04
Pro	CCC	0.35	0.00	0.39	0.23	0.38	0.37

Abbreviations: Rc, *R. capsulatus*; R, *Rhodobacter*; Fru, fructose catabolic genes; Pho, photosynthetic genes; Nif, nitrogen utilization genes; Crt, carotenoid biosynthetic genes. (From Wu, L.-F., Saier, Jr., M.H. [1991]. Differences in codon usage among genes encoding proteins of different function in *Rhodobacter capsulatus*. *Research in Microbiology* 142: 943–949.)

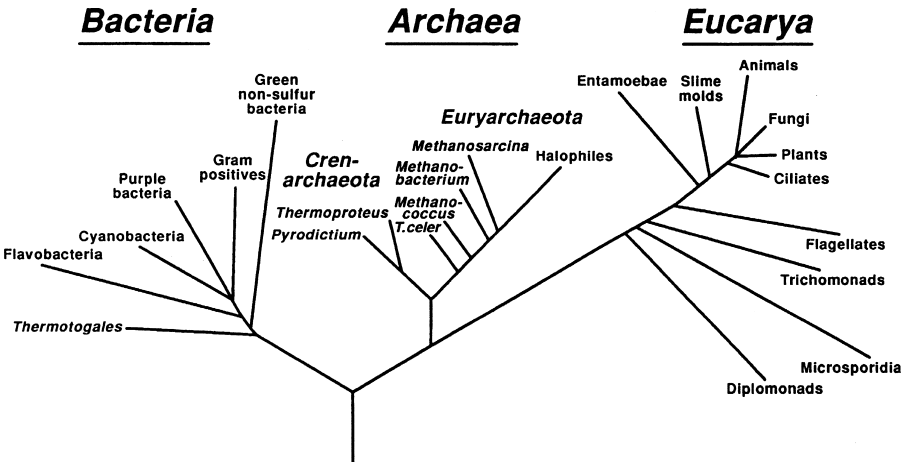

Figure. 17.5. Universal phylogenetic tree in rooted form, showing the three domains: *Archaea*, *Bacteria*, and *Eukarya*. The position of the root was determined by using the paralogous gene couple translation elongation factors EFTu and EFG. (From Woese, C.R. [1994]. There must be a prokaryote somewhere: Microbiology's search for itself. *Microbiological Reviews* 58: 1–9.)

However, Gupta (1998) argues that his analysis suggests that eukaryotes developed from a fusion of an Archaean and a Gram-negative bacterium. Fusion hypotheses have long been used to explain the origins of mitochondria and chloroplasts, so this theory merely extends those ideas. Simonson et al. (2005) have refined the fusion hypothesis and have proposed that the universal phylogenetic tree does not begin with a root but rather with a ring that represents the fusion event.

One current area of controversy is the contribution to evolution of **horizontal** (lateral) gene transfer. In a horizontal transfer, an organism receives a gene not by inheritance from its parent or by mutation, but by any of the genetic transfer processes discussed earlier. Lake and coworkers argue that the genes most likely to successfully transfer horizontally are the "operational" genes, those not involved in transcription, translation, or DNA replication. In support of their claim, they examined 312 sets of orthologous genes in six completely sequenced bacterial genomes and found evidence that horizontal transfer is more or less constant over time and did not occur at substantially higher frequencies in the remote past. Boucher et al. (2003) have reviewed the impact of horizontal transfer of genes on the taxonomy of prokaryotes. Novichkov et al. (2004) have developed a computer model that attempts to identify when the rate of evolu-

tionary change does not match the amount of change seen in individual genes. From their data, they believe they can identify the minority of changes due to horizontal transfer and estimate that 70% of the genes studied fit a model of vertical transfer. However, some gene families show large anomalies. Ochman et al. (2005) show that horizontal transfer may actually maintain species integrity and does not prevent phylogenetic reconstruction at many levels.

Genomics and Proteomics

Bacteria

Genomics is the study of patterns in the genome of an organism, and **proteomics** is the study of the proteins it produces and the conditions under which they are produced. Genomics follows logically from the collection of DNA sequence data. Proteomics became possible for the same reason. Once the DNA sequence of an organism is known, computer programs can identify open reading frames using fairly simple assumptions. Initially, the correlation of open reading frames with real proteins requires a good genetic map, and for this reason *E. coli* is one of the best-studied organisms. However, as the data accumulate, other approaches are possible. For example, x-ray crystallography of proteins provides a three-dimensional structure to the protein that can be correlated with the DNA sequence. Knowing which part of the protein binds to a ligand like ATP means that DNA sequences coding for ATP binding can be identified. With that information in hand, now the computer programs can search known open reading frames for those that appear to bind ATP. It therefore becomes possible to predict probable functions of a protein without ever isolating it.

Two technologies have provided significant proteomic information. They are two-dimensional gel electrophoresis (separation of proteins by molecular weight and charge) and DNA microarray technology. Both these techniques allow the experimenter to determine which genes are active under a given set of conditions. In one case, the actual proteins are observed as spots on the gel, and in the other case, mRNA is assayed as a surrogate for detecting the protein itself.

Subtle variations are also possible. For example, Lee and Lee (2003) discuss the advantages of using time-of-flight mass spectrometry as an adjunct to two-dimensional gels. They looked at the effect of starvation, temperature shock, and oxidative damage on gene expression.

Despite all this effort, much remains to be done, even in *E. coli*. Matte et al. (2003) point out that only 50% of the *E. coli* open reading frames have an experimentally verified protein associated with them. Another 30% have a function attributed to them based on similarity of DNA sequence to known proteins in other organisms.

Bacteriophage

For many years the role of bacteriophages in bacterial evolution was ignored. Recently, however, more careful studies have shown that there is a significant impact. Chibani-Chennoufi et al. (2004) have reviewed the evidence that in estuarine waters in the summer the concentration of bacteria is about 10^6/ml while that of their phages is 10^7/ml. This tenfold excess of virus over host cells means that there is the enormous potential for transduction of DNA from one cell to another. One estimate is that there are roughly 10^{30} phages on the planet, making them the most numerous biological entity.

Other estimates are in accord with these ideas. One strain of *E. coli* O157 may have 18 different prophages, accounting for 16% of its genome (Canchaya et al. 2003). The question then arises, what is the selective advantage that keeps a prophage in the population? One possibility is superinfection immunity, protection for the host cell. However, there are other positive advantages. The λ *bor* gene confers serum resistance on pathogens. There are also many cases of lysogenic conversion where the prophage encodes a toxin that enhances the pathogenicity of the host.

Specific Examples of Evolution

Evolution does not occur in the abstract. Changes in DNA sequence can be reflected by changes in RNA sequence, protein amino acid sequence, and other more subtle changes. Some of these changes offer additional information to the evolutionary biologist.

THINKING AHEAD

Bacteria do not use all codons with equal frequency. Which organisms would you expect to have the most biased codon usage? Why?

Evolution of Genomes

One intriguing question about bacterial genomes is how bacteria can develop two chromosomes both carrying essential genes. One suggested answer comes from the work of Itaya and Tanaka (1997), who performed "genetic surgery" on the *Bacillus subtilis* 168 chromosome. They inserted two partial neomycin resistance cassettes into the chromosome and inserted a low copy number plasmid origin of replication in between. The neomycin cassettes contained enough sequence similarity to recombine. Successful recombination between the cassettes would excise a fragment 310 kb in length, and the recombination event would generate a neomycin resistance gene to provide a selectable marker. When they applied selection, they obtained neomycin-resistant bacteria. Pulsed field gel electrophoresis showed that the original chromosome had lost 310 kb and a new, circular subgenome was present. This type of event could presumably occur in nature, especially given the tendency of plasmids to integrate into chromosomes.

Another issue in genome evolution is what amount of variability is the result of mutations passed on to the progeny cells (vertical transmission) and how much is the result of genetic transfer processes like conjugation, transduction, or genetic transformation (horizontal transfer). Lawrence and Ochman (1997) and Ochman et al. (2005) have looked at variation in the (G + C) content of *E. coli* and *Salmonella enterica* DNA, and they estimate that 31 kb of DNA accumulate due to horizontal transfer over the course of one million years. Much of this DNA presumably comes from virus infections, either as transduction or as new genes acquired with a prophage.

Richard Lenski's research group has an interesting, ongoing experiment in evolution. They have been propagating a set of 12 populations of *E. coli* B, all derived from a common ancestor, for over 10 years (Lenski et al. 2003). Each population is subcultured daily in a glucose-limited minimal medium, so that approximately 20,000 generations have elapsed for each population. Some populations have evolved mutator mutations and some have not. DNA sequencing of small regions of the genome has allowed them to estimate that nonmutator populations have about three synonymous base substitution mutations and mutator populations have about 250 in the 20,000 generations. Given that these mutations should be neutral in their effect, they allow an estimate of mutation rate that is independent of selection. For the duration of the experiment, the fitness of the populations increased by about 70%.

Brüssow et al. (2004) review the contribution of bacteriophages to genetic evolution. Phage studies are difficult because of a lack of fossil record

and concomitant molecular clock. Nevertheless, it is clear (see later) that phage modularity is a significant evolutionary tool. Phages also have a significant impact on their host bacteria in addition to the obvious effects of transduction. This is due to the extra genes that may be carried by prophages to enhance the pathogenicity of their host cells. These genes can encode toxins, proteins that change the antigenicity of the host, or enzymes that protect the host from mammalian defenses like superoxide or white blood cells.

Evolution of the Genetic Code

The genetic code itself is not invariant. Minor differences in the genetic code are discussed in Chapter 3. Each time a tRNA suppressor mutation appears, the genetic code changes slightly for that particular organism. In nearly all cases, the change does not become fixed in the population, but it could theoretically do so. Certain mitochondria (presumably derived from early bacteria) routinely translate the RNA codon UAA as tryptophan instead of termination.

The synonymous codons also offer opportunities for change. Individual organisms do not use synonymous codons with equal frequency. Instead, they exhibit characteristic patterns of codon preference (Table 17.1) that are reflected in the relative abundance of individual tRNA types. It is possible for mutations to occur that will not alter the amino acid sequence of a protein but that will contribute to a change in the (G + C) ratio of an organism's DNA.

Changes in codon usage that favor the more abundant tRNA species improve the translation of a particular mRNA molecule, and the converse is also true. The effect of these changes is to make it difficult for viruses infecting one organism to function equally well in its evolved relative. Similarly, the accumulating sequence differences can contribute to a genetic isolation of an evolving population. For example, one can predict the ability of particular DNA sequences to participate in recombination following genetic transformation based on knowledge of the sequence of the homologous gene in the recipient cell. This approach offers an additional method for attempting to define bacterial species, one based on sexual isolation.

Evolution of Proteins

There are three possible effects of mutations on proteins. The changes may be silent and not affect the amino acid sequence. The protein function may remain

but the amino acid sequence may change, or both the function and the amino acid sequence may change. Each effect contributes different information to the evolutionary geneticist.

Silent changes are advantageous because they presumably result in minimal change in the selective pressure on the organism. The DNA (G + C) content may be trivially altered, but that effect should be small, at least for any single mutation. The lack of selection means that silent changes should accumulate steadily with time. Presumably, the more the silent changes seen, the greater the time since two different organisms diverged from their common ancestor.

Changes that preserve function but alter amino acid sequence are valuable to the structural biologist. They offer clues as to which regions of the protein molecule are critical and which are dispensable. They also provide an indication of which regions of the molecule must remain hydrophobic or hydrophilic. An example of diverged protein sequences is shown in Fig. 17.6. As more protein sequences become available in the computer databases, it becomes possible for scientists to identify the physiologic function of unknown proteins by looking for common modules. For example, ATP binding sites, DNA binding sites, and membrane-embedded proteins all have easily recognized structural motifs in their amino acid sequences.

Evolution of Regulatory Sequences

Chapters 4 and 14 carry extensive discussion of changes in regulatory sequences. New promoters can arise or old promoters can disappear. Similarly, enhancer sequences can be added or subtracted. Operators can mutate to be independent of repressors or repressors can mutate so that they always bind to operators and never to inducers. The general problem with evolutionary analysis of such structures is that they are of limited extent. The shorter the sequence being studied, the less likely the chance that there will be many phylogenetically informative changes. Moreover, promoter sequences are strongly constrained by functionality considerations. For these reasons, regulatory regions have not been extensively studied from the point of view of evolution.

Evolution of Mitochondria and Other Endosymbionts

Andersson et al. (1998) have reported the complete genome sequence for the obligate intracellular parasite *Rickettsia prowazekii*. It is 1100 kb and includes 834

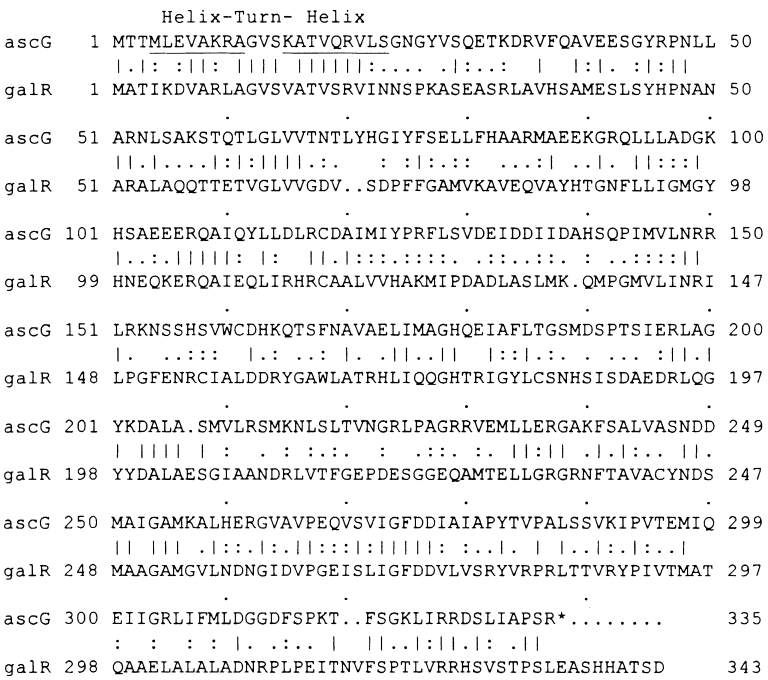

Figure 17.6. Alignment of the *ascG*-encoded *asc* repressor with the *galR*-encoded galactose operon repressor. The helix segments of the helix-turn-helix region (DNA-binding motif) are underlined. Perfect matches are indicated by a *vertical line*. Mismatches are assigned a similarity score. A pair of *dots* (:) indicates a score ≥0.5, while a *single dot* (.) indicates a score of 0–0.5. the individual letters indicate separate amino acids. A, alanine; B, asparagine or aspartic acid; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; Z, glutamine or glutamic acid. (From Hall, B.G., Xu, L. [1992]. Nucleotide sequence function, activation, and evolution of the cryptic *asc* operon of *Escherichia coli K-12*. *Molecular biology and Evolution* 9: 688–706.)

protein-coding genes. None of the genes codes for anaerobic glycolysis, but there is a fully functional tricarboxylic acid cycle and electron transport. In effect, it is a mitochondrion, and there are detectable similarities between mitochondrial genes and *R. prowazekii* genes of similar function. Roughly 24% of the genome is noncoding, suggesting that these are remnants of silenced genes that may be lost without harm to the bacterium or be mutated to new functions.

More recently the same group (Canback et al. 2002) has examined the relationships among glycolytic enzymes in Bacteria, Archaea, and Eukarya. They compared the results from *Bartonella henselae*, a rickettsial relative, and found that there is little evidence of exchange among the three groups (mitochondria excluded), with the exception of some transfer from cyanobacteria to eukarya.

Another source of information about endosymbionts is the bacteria required by certain insects for normal metabolic functions. Several of these organisms have sequenced genomes, and certain features are apparent (Wernegreen 2002). The genomes are very AT-rich and have accumulated significant numbers of deleterious mutations. The synonymous substitution rate in *Buchnera* is four times that in *E. coli*.

Genetic Structure of the Chromosome

Bacteria

There are several levels at which chromosome organization is visible. For example, there are clusters of genes in operons. How might such clusters arise? One model is the selfish operon model (Lawrence 2003) that argues for grouping of genes so that they can travel as a functional unit via transduction or other genetic transfer process. In addition, a group of genes that is coordinately regulated can use a cis-acting regulator, one that has a lower affinity binding constant. Boucher et al. (2003) argue that genes not so clustered will be unable to affect their new host in a positive sense because only the entire group of genes can carry out the pathway and participate in Darwinian selection.

In a related experiment, Audit and Ouzounis (2003) performed a computer analysis of the complete DNA sequence of 86 Bacteria and Archaea, looking for patterns in gene clustering and/or strand preference. They found what they describe as long-range correlations, meaning that at whatever scale greater than 2 kb they use to examine the chromosome, they find that genes tend to assemble and arrange themselves in the same orientation. The strand orientation bias is greater in organisms with analogs to PolC, suggesting that the mode of DNA replication may play a role in selecting for gene arrangements. The overall effect, regardless of scale, is operon-like.

Rocha (2004) points out that the presence of repeated elements leads to chromosome instability and selection for certain gene orders. For example, the

linear chromosome of *Streptomyces coelicolor* has its essential genes clustered in the middle near the origin of replication. Repeated elements tend to occur out on the arms, which are unstable.

Bacteriophage

Some people have been declaring this decade to be the "Age of Phage" owing to the discovery not only of the important ecological roles played by bacteriophages, but also of the lessons to be learned from their genomics. As reviewed by Hendrix (2003), most bacteriophage genomes appear to be modular (Fig. 17.7), with each module a mosaic of genes for a particular viral pathway. However, the distinction between temperate and lytic phages is an important one genetically. In the prophage state, phages are subject to all the processes that affect the host chromosome, including lateral gene transfer. The lytic phages, on the other hand, lyse their host cell so rapidly that it is difficult to imagine how significant lateral gene transfer could occur. Therefore, families such as the T4 phage are likely to represent essentially pure vertical gene transmission.

Hendrix and coworkers have identified genes within modules that seem to have no corresponding gene in otherwise similar modules. They have designated these genes as "morons," units of more DNA that include a promoter and a transcription terminator. Morons often have a different (G + C) content, which makes them correspondingly easier to identify. In pathogenic bacteria they are frequently associated with virulence factors and other functions not essential to phage growth.

Summary

Nucleic acid and protein sequence data are consistent with models that call for evolution by gene duplication and subsequent evolutionary divergence. Bacterial chromosomes exhibit clusters of genes regardless of the scale used in the examination. Relatedness measures allow experimenters to construct evolutionary trees to try to visually express the ways in which organisms or genes have evolved from each other. Two common methods of tree construction are distance measurements and evolutionary parsimony, although neither of them is perfectly satisfactory. Difficulties arise in locating informative base changes that are not so frequent so as to blur the evolutionary trail and in aligning the sequences

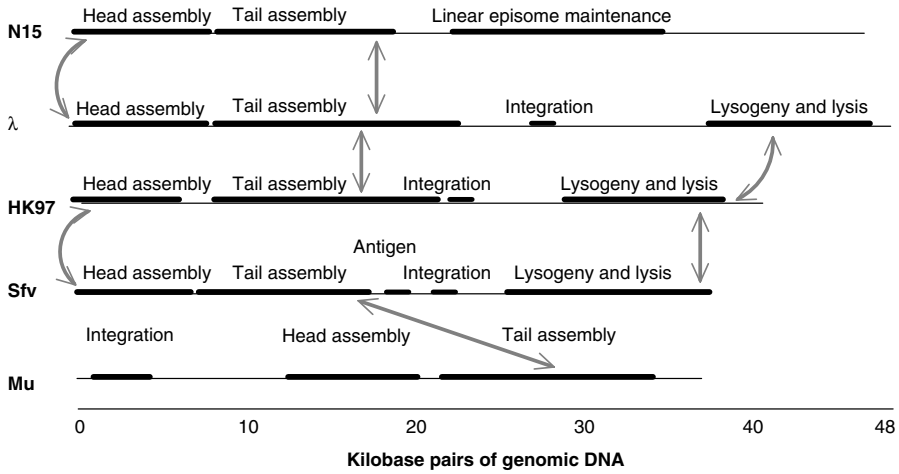

Figure 17.7. Modular design of bacteriophages. Shown in the picture are stylized genomes of five bacteriophages that infect enteric hosts. The *thin horizontal lines* represent the individual genomes. The *thicker lines* represent mosaic modules as labeled. *Gray arrows* link modules that belong to the same DNA sequence families. Note that the order of modules is often preserved. (Adapted from Lawrence, J.G., Hatfull, G.F., Hendrix, R.W. [2002]. Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *Journal of Bacteriology* 184: 4891–4905.)

so as to minimize the number of mutations needed to achieve present-day sequences. One potentially confusing aspect of bacterial evolution is that cryptic genes may be preserved intact over evolutionary time (many thousands of generations). Bacteriophages have the potential to contribute greatly to discussions on evolution because of their ubiquitous presence, ease of isolation, and modular construction.

Questions for Review and Discussion

1. How would you decide whether two bacteria are related? How would you quantify the degree of relatedness?
2. Will the possibility of convergent evolution affect your answer to question 1? Why or why not?
3. If a bacterium received a piece of DNA that coded for proteins using different codon preferences than its own, what would be the effect? Given that you

have a bacterium with pieces of DNA having different codon preferences, what might happen over evolutionary time?

4. How would you decide whether a particular gene that is presently expressed is normally cryptic?
5. What are the advantages of having a modular construction to a bacteriophage?

References

General

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., Nesbø, C.L., Case, R.J., Doolittle, W.F. (2003). Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics* 37: 283–328.

Brüssow, H., Canchaya, C., Hardt, W.-D. (2004). Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* 68: 560–602.

Casjens, S. (1998). The diverse and dynamic structure of bacterial genomes. *Annual Review of Genetics* 32: 339–377.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.-L., Brüssow, H. (2004). Phage–host interaction: An ecological perspective. *Journal of Bacteriology* 186: 3677–3686.

Dawkins, R. (1989). *The Selfish Gene*. New York: Oxford University Press. (A new edition of a classic book on evolutionary molecular biology. Most examples are not bacterial, but the point of view is an important one. For insights into how the scientific process works, don't miss the endnotes.)

Gupta, R.S. (1998). Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiology and Molecular Biology Reviews* 62: 1435–1491.

Hendrix, R.W. (2003). Bacteriophage genomics. *Current Opinion in Microbiology* 6: 506–511.

Karlin, S., Campbell, A.M., Mrázek, J. (1998). Comparative DNA analysis across diverse genomes. *Annual Review of Genetics* 32: 185–225.

Lake, J.A., Jain, R., Rivera, M.C. (1999). Mix and match in the tree of life. *Science* 283: 2027–2028.

Lawrence, J.G. (2003). Gene organization: Selection, selfishness, and serendipity. *Annual Review of Microbiology* 57:419–440.

Lee, P.S., Lee, K.H. (2003). *Escherichia coli*—A model system that benefits from and contributes to the evolution of proteomics. *Biotechnology and Bioengineering* 84: 801–814.

Matte, A., Sivaraman, J., Ekiel, I., Gehring, K., Jia, Z., Cygler, M. (2003). Contribution of structural genomics to understanding the biology of *Escherichia coli*. *Journal of Bacteriology* 185: 3994–4002.

Posada, D., Crandall, K.A., Holmes, E.C. (2002). Recombination in evolutionary genomics. *Annual Review of Genetics* 36: 75–97.

Riesenfeld, C.S., Schloss, P.D., Handelsman, J. (2004). Metagenomics: Genomic analysis of microbial communities. *Annual Review of Genetics* 38: 525–552.

Rocha, E.P.C. (2004). Order and disorder in bacterial genomes. *Current Opinion in Microbiology* 7: 519–527.

Wernegreen, J.J. (2002). Genome evolution in bacterial endosymbionts of insects. *Nature Reviews: Genetics* 3: 850–861.

Specialized

Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Ponten, T., Alsmark, U.C.M., Podowski, R.M., Naslund, A.K., Eriksson, A.S., Winkler, H.H., Kurland, C.G. (1998). The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396: 133–140.

Audit, B., Ouzounis, C.A. (2003). From genes to genomes: Universal scale-invariant properties of microbial chromosome organisation. *Journal of Molecular Biology* 332: 617–633.

Baglioni, P., Bini, L., Liberatori, S., Pallini, V., Marri, L. (2003). Proteome analysis of *Escherichia coli* W3110 expressing an heterologous sigma factor. *Proteomics* 3: 1060–1065.

Canback, B., Andersson, S.G.E., Kurland, C.G. (2002). The global phylogeny of glycolytic enzymes. *Proceedings of the National Academy of Sciences of the USA* 99: 6097–6102.

Dole, S., Nagarajavel, V., Schnetz, K. (2004). The histone-like nucleoid structuring protein H-NS represses the *Escherichia coli bgl* operon downstream of the promoter. *Molecular Microbiology* 52: 589–600.

Itaya, M., Tanaka, T. (1997). Experimental surgery to create subgenomes of *Bacillus subtilis* 168. *Proceedings of the National Academy of Sciences of the USA* 94: 5378–5382.

Khan, M.A., Isaacson, R.E. (1998). In vivo expression of the β-glucoside (*bgl*) operon of *Escherichia coli* occurs in mouse liver. *Journal of Bacteriology* 180: 4746–4749.

Lawrence, J.G., Ochman, H. (1997). Amelioration of bacterial genomes: Rates of change and exchange. *Journal of Molecular Evolution* 44:383–397.

Lenski, R.E., Winkworth, C.L., Riley, M.A. (2003). Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. *Journal of Molecular Evolution* 56: 498–508.

Manna, D., Wang, X., Higgins, N.P. (2001). Mu and IS*1* transpositions exhibit strong orientation bias at the *Escherichia coli bgl* locus. *Journal of Bacteriology* 183: 3328–3335.

Novichkov, P.S., Omelchenko, M.V., Gelfand, M.S., Mironov, A.A., Wolf, Y.I., Koonin, E.V. (2004). Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of Bacteriology* 186: 6575–6585.

Ochman, H., Lerat, E., Daubin, V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proceedings of the National Academy of Sciences of the USA* 102: 6595–6599.

Qi, J., Luo, H., Hao, B. (2004). CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research*, 32 (Suppl. 2): W45–W47.

Simonson, A.B., Servin, J.A., Skophammer, R.G., Herbold, C.W., Rivera, M.C., Lake, J.A. (2005). Decoding the genomic tree of life. *Proceedings of the National Academy of Sciences of the USA* 102: 6608–6613.