

PART I

GENOME FEATURES

CHAPTER 1

PROKARYOTES

1.1 INTRODUCTION

While this book is being written, complete sequences of bacterial genomes are being produced at a rate of about two genomes per month, and the National Center for Biotechnology Information (NCBI) Web site (see the URL in Table 5.1) reports about 60 completely sequenced prokaryotic genomes. Data reported in this chapter refer to the status of completely sequenced genomes, summarized in Table 1.1. Obviously, by the time you read this book, many more will have been sequenced and perhaps some of the aspects dealt with could be viewed differently, although we do not expect dramatic changes in our knowledge unless technology speeds its pace considerably.

Table 1.1 reports the prokaryotic genomes completely sequenced up to now and includes such features as species name, EMBL data library accession number, size, shape, presence of extrachromosomal elements, and bibliographic references. From a look at this list, one can gain an appreciation of the diverse reasons for promoting the sequencing of one species rather than another. Bacterial species are sequenced according to their research interest in basic or applied science: their importance for phylogenetic investigations, to shed light into the metabolic machinery (mainly Archaea) as well as for their importance as human and/or animal pathogens, and for their role as a source of industrial enzymes. In other words, priority has been given to species already well known or species presenting attractive opportunities in applied fields; thus from a phylogenetic point of view, the choice turns out to be very random.

We know we are at the infancy of the genomic era; despite the fact that completely sequenced organisms are still tiny in number, they have already turned out to be full of surprises. In this chapter we summarize the principal sequencing achievements that have improved our knowledge of the prokaryotic genomes and have contributed to outlining methods and approaches to be used in such studies.

TABLE 1.1. Prokaryotic Genomes Completely Sequenced

Species	Main Chromosome		Extrachromosomal Elements		References
	Accession Number	Size (bp)	Accession Number	Size (bp)	
<i>Archaea</i>					
<i>Aeropyrum pernix</i>	BA000002	1,669,695			Kawarabayasi, Hino et al. (1999)
<i>Archaeoglobus fulgidus</i>	AE000782	2,178,400			Klenk, Clayton et al. (1997)
<i>Halobacterium</i> sp. NRC-1 (3 chromosomes)	AE004437	2,014,239	AF016485	191,346	Ng, Kennedy et al. (2000)
	AE004438	365,425	AE004438	365,425	
<i>Methanobacterium thermoautotrophicum</i>	AF016485	191,346			Smith, Doucette-Stamm et al. (1997)
	AE000666	1,751,377			
<i>Methanococcus jannaschii</i>	L77117	1,664,970	L77118	58,407	Bult, White et al. (1996)
			L77119	16,550	
<i>Methanococcus kandleri</i> AV19	AE009439	1,694,969			Slesarev, Mezhevaya et al. (2002)
<i>Methanosarcina acetivorans</i> str. C2A	AE010299	5,751,492			Galagan, Nusbaum et al. (2002)
<i>Methanosarcina mazei</i> Goe1	AE008384	4,096,345			Deppenmeier, Johann et al. (2002)
<i>Pyrobaculum aerophilum</i>	AE009441	2,222,430			Fitz-Gibbon, Ladner et al. (2002)
<i>Pyrococcus abyssii</i>	AL096836	1,765,118			Lecompte, Ripp et al. (2001)
<i>Pyrococcus furiosus</i> DSM 3638	AE009950	1,908,256			Robb, Maeder et al. (2001)
<i>Pyrococcus horikoshii</i>	AP000001-	1,738,505			Kawarabayasi, Sawada et al. (1998)
	AP000007				
<i>Sulfolobus solfataricus</i>	AE006641	2,992,245			She, Singh et al. (2001)
<i>Sulfolobus tokodaii</i>	BA000023	2,694,765	AJ010405	41,229	Kawarabayasi, Hino et al. (2001)
	AL445063-	1,564,905			
<i>Thermoplasma acidophilum</i>	AL445067				Ruepp, Graml et al. (2000)
	AP000991-	1,584,799			
<i>Thermoplasma volcanium</i>	AP000996				Kawashima, Amano et al. (2000)
<i>Bacteria</i>					
<i>Agrobacterium tumefaciens</i> str. C58(Cereon)	AE007869	2,841,581			Goodner, Hinkle et al. (2001)
<i>Agrobacterium tumefaciens</i> str. C58(U. Washington)	AE008688	2,841,490	AE008687	542,780	Wood, Setubal et al. (2001)
			AE008690	214,234	

<i>Aquifex aeolicus</i>	AE000657	1,551,335	AE000667	39,456	Deckert, Warren et al. (1998)
<i>Bacillus halodurans</i>	BA000004	4,202,353			Takami, Nakasone et al. (2000)
<i>Bacillus subtilis</i>	AL009126	4,214,814			Kunst, Ogasawara et al. (1997)
<i>Borrelia burgdorferi</i> ^a	AE000783	910,725	AE000791	9,386	Fraser, Casjens et al. (1997); Casjens, Palmer et al. (2000)
			AE000792	26,498	
			AE001575	30,750	
			AE001576	30,223	
			AE001577	30,299	
			AE001578	29,838	
			AE001579	30,800	
			AE001580	30,885	
			AE001581	30,651	
			AE001583 ^a	5,228	
			AE000793 ^a	16,823	
			AE001582 ^a	18,753	
			AE000785 ^a	24,177	
			AE000794 ^a	26,921	
			AE000786 ^a	29,766	
			AE000784 ^a	28,601	
			AE000789 ^a	27,323	
			AE000788 ^a	36,849	
			AE000787 ^a	38,829	
			AE000790 ^a	53,561	
			AE001584 ^a	52,971	
<i>Brucella melitensis</i>	AE008917	2,117,144			DeVecchio, Kapatral et al. (2002)
<i>Buchnera</i> sp. APS	AP000398	640,681	AP001070	7,258	Shigenobu, Watanabe et al. (2000)
			AP001071	7,786	
<i>Campylobacter jejuni</i>	AL111168	1,641,481			Parkhill, Wren et al. (2000)
<i>Caulobacter crescentus</i>	AE005673	4,016,947			Nierman, Feldblyum et al. (2001)
<i>Chlamydia pneumoniae</i> AR39	AE002161	1,229,853		4,524	Read, Brunham et al. (2000)
<i>Chlamydia pneumoniae</i> CWL029	AE001363	1,230,230			Kalman, Mitchell et al. (1999)
<i>Chlamydia trachomatis</i> MoPn	AE002160	1,069,412	AE002162	7,501	Read, Brunham et al. (2000)
<i>Chlamydia trachomatis</i> serovar D	AE001273	1,042,519			Stephens, Kalman et al. (1998)
<i>Chlamydophila pneumoniae</i> J138	BA000008	1,226,565			Shirai, Hirakawa et al. (2000)
<i>Clostridium acetobutylicum</i>	AE001437	3,940,880	NC_001988	192,000	Nolling, Breton et al. (2001)

(Continued)

TABLE 1.1. Continued

Species	Main Chromosome		Extrachromosomal Elements		References
	Accession Number	Size (bp)	Accession Number	Size (bp)	
<i>Clostridium perfringens</i>	BA000016	3,031,430			Shimizu, Ohtani et al. (2002)
<i>Corynebacterium glutamicum</i>	AX114121	3,309,400			Tauch, Homann et al. (2002)
<i>Deinococcus radiodurans</i> R1 (2 chromosomes)	AE000513	2,648,638	AE001826	177,466	White, Eisen et al. (1999)
	AE001825	412,348	AE001827	45,704	
<i>Escherichia coli</i> K-12	U00096	4,639,221			Blattner, Plunkett et al. (1997)
<i>Escherichia coli</i> O157:H7 EDL933	AE005174	5,528,970			Perna, Plunkett et al. (2001)
<i>Escherichia coli</i> O157:H7 Sakai	BA000007	5,498,450	AB011549	92,721	Hayashi, Makino et al. (2001)
			AB011548	3,306	
<i>Fusobacterium nucleatum</i> subsp.nucleatum ATCC25586	AE009951	2,174,500			Kapatral, Anderson et al. (2002)
<i>Haemophilus influenzae</i> Rd	L42023	1,830,138			Fleischmann, Adams et al. (1995)
<i>Helicobacter pylori</i> 26695	AE000511	1,667,867			Tomb, White et al. (1997)
<i>Helicobacter pylori</i> J99	AE001439	1,643,831			Alm, Ling et al. (1999)
<i>Lactococcus lactis</i> subsp. <i>lactis</i>	AE005176	2,365,589			Bolotin, Wincker et al. (2001)
<i>Listeria innocua</i>	AL592022	3,011,208	AL592102	81,900	Glaser, Frangeul et al. (2001)
<i>Listeria monocytogenes</i> EGD-e	NC_003210	2,944,528			Glaser, Frangeul et al. (2001)
<i>Mesorhizobium loti</i>	BA000012	7,036,074	AP003015-16	351,911	Kaneko, Nakamura et al. (2000)
			AP003017	208,315	
<i>Mycobacterium leprae</i>	AL450380	3,268,203			Cole, Eiglmeier et al. (2001)
<i>Mycobacterium tuberculosis</i>	AL123456	4,411,529			Cole, Brosch et al. (1998)
<i>Mycoplasma genitalium</i>	L43967	580,074			Fraser, Gocayne et al. (1995)
<i>Mycoplasma pneumoniae</i>	U00089	816,394			Himmelreich, Hilbert et al. (1996)
<i>Mycoplasma pulmonis</i>	AL445566	963,879			Chambaud, Heilig et al. (2001)
<i>Neisseria meningitidis</i> MC58	AE002098	2,272,351			Tettelin, Saunders et al. (2000)
<i>Neisseria meningitidis</i> Z2491	AL157959	2,184,406			Parkhill, Achtman et al. (2000)
<i>Nostoc</i> sp. PCC 7120	NC_003272	6,413,771	NC_003240	186,614	Kaneko, Nakamura et al. (2001)
			NC_003267	101,965	
			NC_003273	55,414	
			NC_003270	40,340	
			NC_003241	5,584	

<i>Pasteurella multocida</i>	AE004439	2,257,487		AL646053	2,094,509	May, Zhang et al. (2001)
<i>Porphyromonas gingivalis</i>	NC_002950	2,343,478				Unpublished
<i>Pseudomonas aeruginosa</i> PA01	AE004091	6,264,403				Stover, Pham et al. (2000)
<i>Ralstonia solanacearum</i>	AL646052	3,716,413				Salanoubat, Genin et al. (2002)
<i>Rickettsia conorii</i>	AE006914	1,268,755				Ogata, Audic et al. (2001)
<i>Rickettsia prowazekii</i>	AJ235269	1,111,523				Andersson, Zomorodipour et al. (1998)
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i>	AL513382	4,809,037				Parkhill, Dougan et al. (2001)
<i>Salmonella enterica</i> serovar <i>typhimurium</i> LT2	AE006468	4,857,432				McClelland, Sanderson et al. (2001)
<i>Sinorhizobium meliloti</i>	AL591688	3,654,135		AE006469	1,354,226	Galibert, Finan et al. (2001)
				AL911985	1,683,333	
<i>Staphylococcus aureus</i>	BA000017	2,878,040				Kuroda, Ohta et al. (2001)
<i>Streptococcus pneumoniae</i> R6	AE007317	2,038,615				Hoskins, Alborn et al. (2001)
<i>Streptococcus pneumoniae</i> TIGR4	AE005672	2,160,837				Tettelin, Nelson et al. (2001)
<i>Streptococcus pyogenes</i>	AE004092	1,852,441				Ferretti, McShan et al. (2001)
<i>Streptomyces coelicolor</i> A3(2)	AL645882	8,667,507		NC_003903	356,023	Bentley, Chater et al. (2002)
				NC_003904	31,317	
<i>Synechocystis</i> PCC6803	AB001339	3,573,470				Kaneko, Sato et al. (1996)
<i>Thermoanaerobacter tengcongensis</i>	AE008691	2,689,445				Bao, Tian et al. (2002)
<i>Thermotoga maritima</i>	AE000512	1,860,725				Nelson, Clayton et al. (1999)
<i>Treponema pallidum</i>	AE000520	1,138,011				Fraser, Norris et al. (1998)
<i>Ureaplasma urealyticum</i>	AF222894	751,719				Glass, Lefkowitz et al. (2000)
<i>Vibrio cholerae</i> (2 chromosomes)	AE003852	2,961,151				Heidelberg, Eisen et al. (2000)
	AE003853	1,072,914				
<i>Xanthomonas axonopodis</i> pv.citri str. 306	NC_003919					da Silva, Ferro et al. (2002)
<i>Xanthomonas campestris</i> pv.campestris str.ATCC33913	NC_003902					da Silva, Ferro et al. (2002)
<i>Xylella fastidiosa</i>	AE003849	2,679,305		AE003851	51,158	Simpson, Reinach et al. (2000)
				AE003850	1,285	
<i>Yersinia pestis</i>	AL590842	4,653,728		NC_003132	9,612	Parkhill, Wren et al. (2001)
				NC_003131	70,305	
				NC_003134	96,210	

^a Linear chromosome.

We include some basic knowledge of prokaryotes, such as morphology, classification, and main features regarding the organization, replication, and expression of genetic material. Such descriptions are far from exhaustive. Since we focus our attention on the main aspects that emerged from knowledge of complete genome sequencing, we ask the reader to refer to more in-depth studies by specialists in the field and to the numerous reviews and papers available in the literature.

1.2 MORPHOLOGY AND CLASSIFICATION

Prokaryotes are unicellular organisms and are the most numerous organisms on Earth (4 to 6×10^{30} cells, 3 to 5×10^{17} g of C; Whitman, Coleman et al. 1998); commonly known as bacteria, they include both Archaea and Bacteria. Their morphology is quite simple: the prokaryotic cell (ranging from 0.2 to $10\mu\text{m}$ in diameter) can be considered as one unit, a single compartment with no membrane-bound organelles inside.

Typically, prokaryotes have a cell wall containing peptidoglycan (except mycoplasmas and Archaea) which surrounds the cell and confers rigidity and protection. The way that peptidoglycan is arranged is the basis for the identification of bacterial organisms; two distinct cell wall types are revealed by a commonly used staining procedure, the Gram stain (see below), according to the relative content of peptidoglycan. Many pathogenic bacteria also have capsules, structures made of polysaccharides or proteins that are external to the cell wall. Such coatings are useful for both adhesion and resistance to host immune response.

The cytoplasm of prokaryotic cells is enclosed in the plasma membrane, a phospholipid bilayer, whose function is not only to control what enters and leaves the cell but also to provide a site for protein attachment and enzyme activity. Indeed, the size limit of prokaryotes is greatly influenced by the surface-to-volume ratio, since there are no internal membranes. The cytoplasm is diffuse and granular, due to the presence of many ribosomes—sites for protein synthesis, smaller than those of eukaryotes. Very often, there are inclusion bodies whose function is material storage. The genetic material is confined to the nucleoid, a region not delimited by a membrane, but visibly distinct from the rest of the cell by electron transmission microscopy. Sometimes, one or more additional circular DNA molecules (plasmids) are also present.

Even though only about 4000 species are described, it is estimated that the true number could range between 400,000 and 4,000,000. It is evident that a definition of *species* in small unicellular organisms such as prokaryotes is not easy. In eukaryotes, individuals belong to the same species if they are capable of fertile interbreeding, share specific morphological traits, and form a monophyletic group. In prokaryotes, instead, the classic species concept, based primarily on morphological traits, cannot be used reliably (Rossello-Mora and Amann 2001). For this reason, novel approaches have been adopted for microbial classification based on a variety of evidence, including chemotaxonomic markers (e.g., cell wall, polyamines, quinones, etc.), DNA properties (e.g., G + C content, extent of DNA hybridization, etc.), and rRNA sequences. In particular, classification based on comparative analysis of rRNA sequences, mostly 16S rRNA, is the one used most at present. Furthermore, in Bacteria there are three taxonomic ranks below the species: (1) the

type/group, a group of isolates or strains sharing a single characteristic (e.g., *Chlamydia trachomatis* serovar D); (2) the isolated pure culture, a clonal population from a single cell or isolate strain having a known set of characteristics (e.g., *Neisseria meningitidis* MC58); and (3) the single cell, that is, a single individual.

In general, two main types of classification can be adopted: the phylogenetic classification, grouping Bacteria according to their evolutionary relationships; and the phenetic classification, based on similarity in bacterial features, without considering origin or evolution. The main characteristics that define the major bacterial groups are the nature of the cell wall (gram-positive, gram-negative, no cell wall), cell shape (cocci, rods, helical), physiology (aerobes, thermophiles, chemolithotrophs, intracellular parasites, etc.), and motility (presence/absence of flagella, corkscrew motion).

There are many different classifications of Bacteria, based on one or more than one of the above-mentioned features. No official classification is available, but bacteriologists have fixed some rules for naming new and old species of bacteria. These rules are collected in the *International Code of Nomenclature of Bacteria (Bacteriological Code)*. The 1980 Approved Lists of Bacterial Names [see Pittman, Walczak et al. (1991) for an update] contained 2212 names of genera, species, or subspecies, and 124 names of higher taxa. A continuously updated list of validly published bacterial names, including more than 5500 taxa, can be found at the LBSN Web site (see the URL in the Appendix; Euzéby 1997).

Following a classic taxonomy scheme the bacterial species can be grouped into four divisions:

1. *Firmicutes* (Gibbons and Murray 1978; see also Pittman, Walczak et al. 1991) are gram-positive Bacteria with thick cell walls.
2. *Gracilicutes* (Gibbons and Murray 1978; see also Pittman, Walczak et al. 1991) are gram-negative Bacteria with thin cell walls.
3. *Mendosicutes* (Murray 1984; see also Woese and Fox 1977) enclose the single class Archaeobacteria, which represents the Archaea domain in phylogenetic classification (Woese, Kandler et al. 1990).
4. *Tenericutes* (Murray 1984) comprise bacterial species lacking the cell wall.

A remarkable breakthrough in the classification of bacteria was achieved with the advent of molecular data, particularly the sequence analysis of 16S rRNA. The classification based on molecular features divides prokaryotes into two domains, Bacteria and Archaea. This classification dates back to as early as the mid-1970s, when Woese and Fox (1977), using small subunit ribosomal RNA (SSU rRNA) data, first described a number of features characterizing a distinct group of unicellular organisms, the Archaeobacteria renamed Archaea in 1990 by Woese, Kandler et al. (1990), thus abandoning the classical bipartite division of living organisms into prokaryotes and eukaryotes. Ever since, the classical *tree of life* is usually illustrated with three main branches: Bacteria, Archaea, and Eukarya. There are, however, recent conflicting claims regarding this classification, and several fundamental questions regarding the time and mode of evolution, as well as the phylogeny of the three domains, are discussed in Sec. 8.5.

Although the rRNA-based classification of prokaryotes may have a limited resolving power in some cases and can be questioned for the peculiar evolutionary dynamics of rRNA genes (see Sec. 8.1), it remains a stable and operationally satis-

factory framework for prokaryotic classification. Furthermore, recent evolutionary analyses based on differences in gene content (Snel, Bork et al. 1999) or on protein sequence comparison (Brown, Douady et al. 2001; Brochier, Baptiste et al. 2002) were remarkably consistent with rRNA trees. This indicates that processes of lateral gene transfer (see Sec. 7.4), although frequent in bacterial evolution, have not completely erased the phylogenetic signal.

rRNA-derived phylogenetic trees as well as aligned and annotated rRNA sequences are provided by the Ribosomal Database Project (RDP; see the URL in the Appendix). The availability of completely sequenced genomes represents a powerful tool for shedding light on these fundamental questions and will provide the opportunity not only to better understand prokaryotic organisms at the molecular level, but also to discover new and unexpected features of their evolution. Figure 1.1 shows the classification of bacterial species whose complete genome has been fully sequenced, provided by the NCBI Taxonomy Browser (see the URL in Table 5.1).

In this chapter we describe the main features of the prokaryotic genome, treating the two domains, Archaea and Bacteria, separately whenever possible.

1.3 GENOME SHAPE AND SIZE

Until recently, bacterial genomes were believed always to be circular. If this is true for most bacteria, there are several species whose chromosomes are linear; actually, there may be a natural interchange between the linear and the circular geometry (Volf and Altenbuchner 1998), such as in *Streptomyces*. In the list of completely sequenced organisms reported in Table 1.1, only one species, *Borrelia burgdorferi*, has a linear chromosome.

Telomeres in linear replicons have been described in only a few species. In *Borrelia*, telomeres are covalently closed hairpins, where one DNA strand loops around, becoming the other (Hinnebusch, Bergstrom et al. 1990; Hinnebusch and Barbour 1991; Casjens, Murphy et al. 1997). In *Streptomyces*, instead, the chromosome is open and ends with specific proteins covalently bound to the 5' ends of the DNA (Sakaguchi 1990; Chen 1996). Linear replicons have been described on several branches of the bacterial phylogenetic tree, which would suggest that linearity arose more than once from circular progenitors (Casjens 1998).

The presence of a single large chromosome in prokaryotes also appears to be a general rule, although several bacterial genomes contain two or three large replicons (chromosomes) several hundred kilobase pairs long. This feature is a stable property of some genera [e.g., *Borrelia* (Barbour 1988) and *Rhizobium* (Honeycutt, McClelland et al. 1993)], which suggests these replicons are essential for their lifestyles. In addition, extrachromosomal DNA elements are present in many species.

The haploidy of bacteria is indeed an oversimplification (Casjens 1998). Many fast-growing species (such as *Azotobacter vinelandii* and *Borrelia hermsii*) contain more than one complete chromosome copy per cell during the exponential growth phase; *Deinococcus radiodurans* also has four copies of its replicon in the stationary phase. *M. jannaschi* has one to five chromosome copies per cell during the stationary growth phase but can harbor more than 10 copies in the exponential

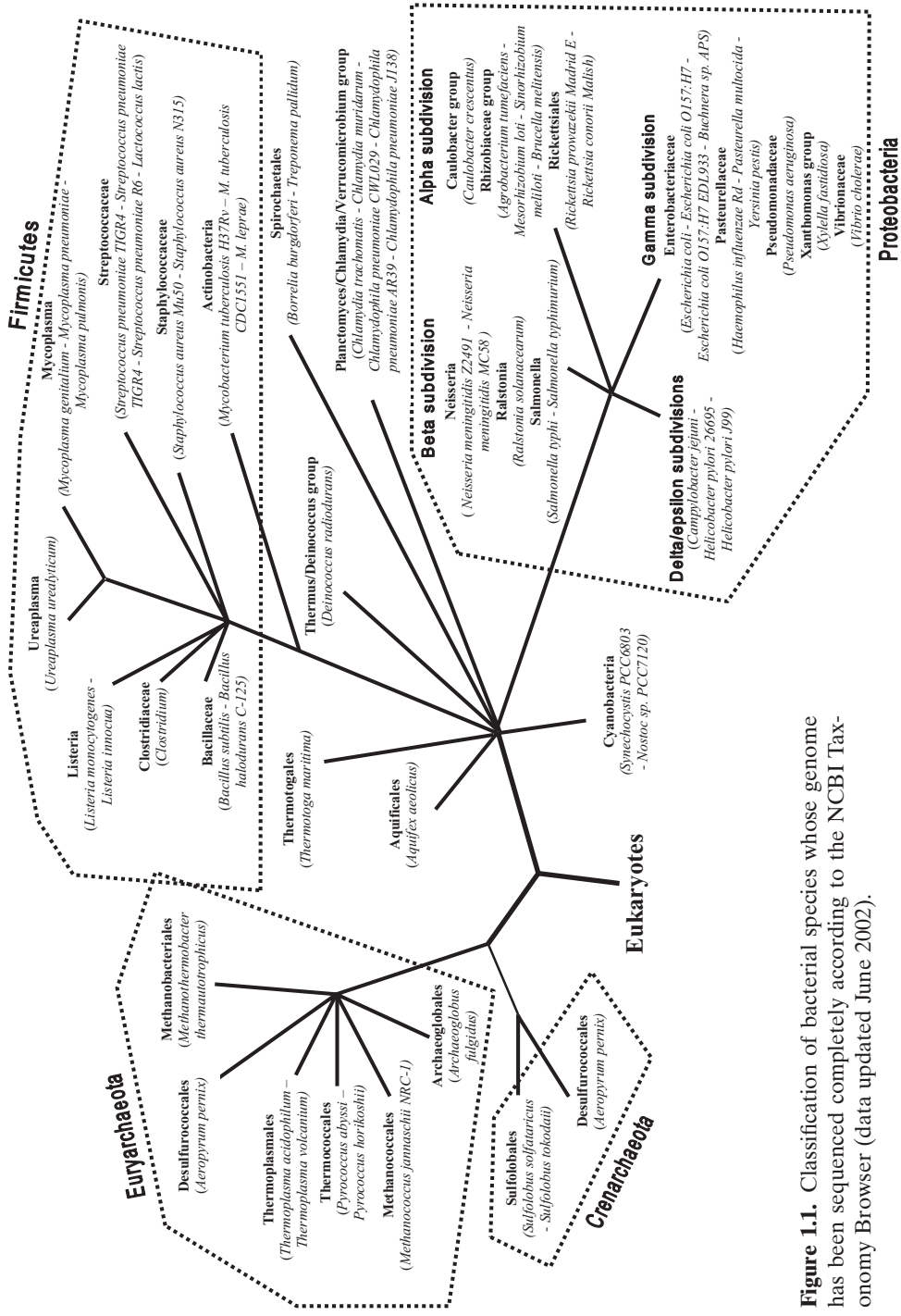


Figure 1.1. Classification of bacterial species whose genome has been sequenced completely according to the NCBI Taxonomy Browser (data updated June 2002).

phase, in contrast with what occurs in other archaeal species, such as *Sulpholobus* (Bernander 2000). As far as the Archaea are concerned, the genomes sequenced to date are circular in shape, and only few species have extrachromosomal elements.

The size of bacterial chromosomes is notably variable even within a genus, as in *Treponema* (from 1040 to 3000 kbp), *Mycoplasma* (from 580 to 1350 kbp), or *Streptomyces* (from 6400 to 8400 kbp) (Casjens 1998). In the bacterial genomes completely sequenced at present, dimensions range from 580,074 bp (*Mycoplasma genitalium*) to 7,036,074 bp (*Mesorhizobium loti*) (see Table 1.1). However, bacterial genomes can be much larger: restriction analyses have shown that the largest known genome is *Myxococcus xanthus* (9200 kbp; He, Chen et al. 1994). Most sequenced bacterial genomes have genome size around 1 to 1.8 Mbp; the largest, beside *M. loti*, being, *Pseudomonas aeruginosa* (6,264,403 bp), *Escherichia coli*, 4,639,221 bp; *Bacillus subtilis*, 4,214,814 bp; *Mycobacterium tuberculosis*, 4,411,529 bp; *Synechocystis* sp., 3,573,470 bp; *Deinococcus radiodurans*, 2,648,638 bp; and *Neisseria meningitidis* 2,272,351 bp.

Figure 1.2 shows the distribution of genome sizes for Bacteria and Archaea completely sequenced so far, whose averages are 2.9 and 2.3 Mbp, respectively. Quite interestingly, it can be noticed that two peaks can be observed for eubacterial genome sizes, with the first one at about 1 Mbp, corresponding to *specialist species* (very small genomes, very specific niches), and the second one, around 2–4 Mbp, much more flattened out, for *generalist species* (larger genomes, a wide range of places to live). It can be also noted that Archaea show a narrower genome size distribution than that of Bacteria.

The large variation in chromosome size of bacterial genomes, which is much more evident when comparing higher taxonomic positions, may be due to rapid gene loss when a species chooses to live in a very specific ecological niche or when a gene is

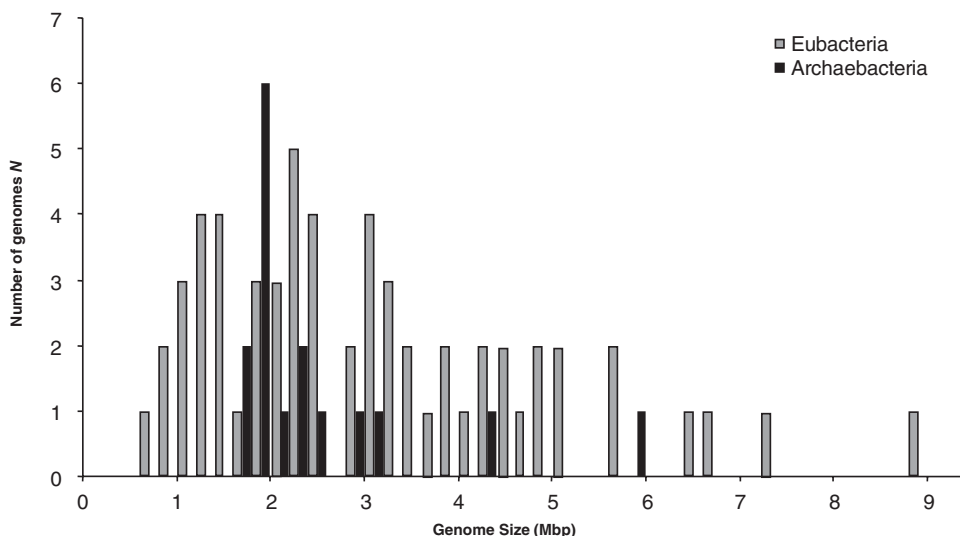


Figure 1.2. Distribution of genome sizes for the 76 bacterial and archeal completely sequenced genomes listed in Table 1.1.

gained via gene duplication or horizontal gene transfer (Casjens 1998). Gene duplication and the presence of repeated elements is undoubtedly at the basis of an increase in genome size; the genomes of *E. coli* and *B. subtilis* contain repetitive elements and cryptic prophage and phage remnants; *M. tuberculosis* contains repetitive DNA, insertion sequences (ISs), and duplicated housekeeping genes. Repeats in bacterial genomes seem to be important for genome plasticity, since they are involved in several processes, such as recombination, inversion, deletion, translocation, and transposition (Romero, Martinez-Salazar et al. 1999). Lateral gene transfer, also from distantly related organisms, is a fundamental process in bacterial genome evolution, since it is responsible for the acquisition and deletion of extensive amounts of DNA in the chromosome, thus producing extremely dynamic genomes (Ochman, Lawrence et al. 2000).

Gene content can be distributed in more than just a single chromosome, and extrachromosomal elements contribute greatly to gene equipment. *Deinococcus radiodurans* contains four genetic elements: two main circular chromosomes (total length: 3,060,986 bp), the megaplasmid, and the plasmid (total extrachromosomal length: 223,170 bp), which together amount to more than 3.2 Mbp. *Borrelia burgdorferi* harbors 17 linear and circular plasmids, whose combined size is about 533,000 bp, more than half the length of the main chromosome (910,725 bp).

An interesting question one might ask regards the minimum size compatible with independent life for a prokaryotic genome. It appears that the smallest genomes, like *Mycoplasma genitalium* (the smallest), *M. pneumoniae*, and *M. pulmonis*, belong to organisms that have a parasitic lifestyle and thus are adapted to specific niches and host–pathogen interactions. Indeed, among factors preventing evolution to a smaller size are the need for an adequate genome and the limited efficiency of the transcriptional and translational machinery (Koch 1996; see also Sec. 7.2).

A remarkable property clearly emerging from prokaryotic genome data is the striking correlation between the fraction of genes coding for regulatory proteins (e.g., 5.8% in *E. coli* and 9.4% in *P. aeruginosa*) and the genome size. This trend is particularly prominent for generalist species that can survive in diverse environments (Stover, Pham et al. 2000).

An interesting aspect of genome size is its content in coding versus noncoding regions. It should be recalled that most protein genes have only been predicted by the methods described in Sec. 6.8.7, and thus they should be regarded as hypothetical, especially those with no homologs in other species. In Bacteria, on average $85 \pm 4.7\%$ of the genome encodes for proteins (see Table 1.2), although a conspicuous part (on average $37.2 \pm 10.6\%$) of the coding portion consists of unidentified open-reading frames. Although the functions performed by these unidentified proteins are unknown to date, they are most likely to represent specific and necessary products for the organism.

In some bacterial species, a process of reductive evolution of the number of genes in the genome has been documented. In obligate intracellular parasites such as *Rickettsia* and *Chlamydia* and some endosymbionts (Andersson and Andersson 1999), genes become inactive when their function is no longer required by the organism. To date, the most extensive genome degradation has been reported for the leprosy bacillus *Mycobacterium leprae*, in which only 49.5% of the genome codes for proteins, with over 1000 pseudogenes (27% of the potential coding capacity; Cole, Eiglmeier et al. 2001).

TABLE 1.2. Content of Protein Coding Genes in Completely Sequenced Microbial Genomes

Species	ORFs	Percent Functionally Identified	Protein Coding Regions (%)	Operons ^d	Lateral Transfer ^a
<i>Archaea</i>					
<i>Aeropyrum pernix</i>	2694	23.5	88.8		
<i>Archaeoglobus fulgidus</i>	2436	46.0	92.2	+	-
<i>Halobacterium</i> sp. NRC-1 (3 replicons)					
<i>Methanobacterium thermoautotrophicum</i>	1855	46.0	92.0	+	+
<i>Methanococcus jannaschii</i>	1743	36.0		+	+
<i>Pyrococcus abyssi</i>	1765				
<i>P. horikoshii</i>	2061	35.0	90.7		
<i>Sulfolobus solfataricus</i>	2977	32.2	83.9		
<i>S. tokodaii</i>	2826			+	-
<i>Thermoplasma acidophilum</i>	1509	55.0	87.0	+	+
<i>T. volcanium</i>					
<i>Bacteria</i>					
<i>Aquifex aeolicus</i>	1512		93.0	+	+
<i>Bacillus halodurans</i>	4066	52.7	85.0	+	+
<i>B. subtilis</i>	4100	58.0	87.0	+	+
<i>Borrelia burgdorferi</i>	853	59.0	93.0	+	-
<i>Buchnera</i> sp. APS	583	85.8	88.0	+	
<i>Campylobacter jejuni</i>	1654	77.8	94.3	+	-
<i>Caulobacter crescentus</i>	3767	53.9	90.6	-(few)	
<i>Chlamydia pneumoniae</i> AR39	1052	60.0		+	-
<i>C. pneumoniae</i> CWL029	1052	60.0		+	-
<i>C. trachomatis</i> MoPn	924				+
<i>C. trachomatis</i> serovar D	894	68.0			+
<i>Chlamydomytila pneumoniae</i> J138	1072				+
<i>Clostridium acetobutylicum</i>	3740			+	
<i>Deinococcus radiodurans</i> R1	2633	69.0			
(2 chromosomes)	369				
<i>Escherichia coli</i> K-12	4288	62.0	87.8	+	+
<i>E. coli</i> O157:H7 EDL933	5349			+	+
<i>E. coli</i> O157:H7 Sakai	5361		88.1	+	+

A comparison of the two completely sequenced *E. coli* strains [i.e., the laboratory strain K-12 and the enterohemorrhagic strain O157:H7 (EDL933 and Sakai)] shows their complex relationships. They differ by about 860 kbp in length and share a clearly homologous backbone of 4.1 Mbp (with 75,168 scattered single nucleotide polymorphisms). Of the remaining genome, 1.34 Mbp represent islands specific to the pathogenic strain, nine of which encode putative virulence factors; 0.53 Mbp represent islands specific to the laboratory strain. Their atypical base composition suggests that most differences in overall gene content are attributable to horizontal transfer of relatively recent origin (Perna, Plunkett et al. 2001).

Noncoding regions represent the most variable part of bacterial genomes and they contain mostly repeated elements whose origin may be tracked in some cases. Prokaryotes, though to a lesser extent than eukaryotes, contain repeated sequences organized both in tandem and interspersed in the genome. The function of such repeated sequences is still largely unknown, yet data seem to suggest that they may represent multiple regulatory signals or serve other biological purposes mostly related with pathogenicity (van Belkum, Scherer et al. 1998; van Belkum 1999).

Tandem repeated elements found in microbial genomes might be prone to the same genetic variability encountered in the eukaryotic genome and could be used profitably as markers for identification and genotyping of bacterial strains. A simple way to detect tandem repeats in genomic sequences is the application of linguistic methodologies, in particular those that measure the linguistic complexity (LC) of DNA sequences (see Sec. 6.8.2). The LC profile along a genome, which can be computed by several methods, clearly highlights genome regions containing tandem repeats (see also Fig. 6.25), which correspond to local LC minima. Other bioinformatic approaches can be used, such as the dot-matrix plot (see Sec. 6.2) or algorithms specifically devoted to the search for tandem repeats (see Sec. 6.8.3).

Interspersed repeat elements have been identified in numerous bacterial species (Lupski and Weinstock 1992). Most of these elements are shorter than 200 bp, evenly distributed in the genome, and located primarily in noncoding regions. For example, REP (repetitive extragenic palindrome) and ERIC (enterobacterial repetitive intergenic consensus) sequence motifs have been found to be widespread in numerous enterobacterial genomes. The identification of interspersed repeats in genomic sequences can be accomplished either by similarity searches against collections of known repeat elements or by the application of bioinformatic tools specifically designed for such a task (see Sec. 6.8.3).

1.4 GENE CONTENT AND ORGANIZATION

Gene identification is one of the first steps in the analysis of microbial genomes and is usually performed with computer-aided methods that employ statistical gene prediction models. Microbial genomes tend to be gene-rich, typically containing about 85% coding sequences; thus gene discovery is much easier than in eukaryotic genomes, especially in higher eukaryotes such as humans whose genome has less than 2% coding sequences. In addition, unlike eukaryotic genes, microbial genes are not interrupted by introns. A survey of all entirely sequenced genomes has revealed

the complete absence of spliceosomal introns and genes coding for components of the spliceosomal machinery from a wide range of bacterial and archaeal species (Logsdon 1998).

In microbial genomes, the identification of a significantly long ORF gives quite reliable proof of protein-coding regions. However, the most reliable way to identify a gene in a new genome is to find a close homolog from another organism. This can be done very effectively using database searching programs such as BLAST and FASTA (see Sec. 6.4 for details on these programs). If significant similarity is not found with known proteins, the search for protein signatures specific for one or more protein families can be carried out by comparing the unknown protein against databases of protein signatures such as ProSite (Hofmann, Bucher et al. 1999), ProDom (Corpet, Servant et al. 2000), and Pfam (Bateman, Birney et al. 2000) (see also Secs. 5.6 and 6.9). Unfortunately, many of the genes in newly sequenced genomes have no significant similarities to known genes or domains. For these genes we must rely on computational methods for their identification. Several algorithms have been devised that provide a very high prediction accuracy in microbial genomes. They are generally based on the observation of base compositional heterogeneity at the three codon positions of the predicted gene or on the recognition of a specific codon use strategy. These methods, whose detailed description can be found in Sec. 6.8.7, may reach over 97% prediction accuracy.

In discriminating between coding and noncoding regions, the analysis of the pattern of linguistic complexity (LC; see Sec. 6.8.2) may also reveal useful, since noncoding regions are, overall, less complex than coding ones. Furthermore, LC pattern analysis can help detect potential regulatory sites when typical LC patterns are displayed. Figure 1.3 shows the number of genes in the available microbial genomes that hit any combination of NCBI protein families (cluster of orthologous

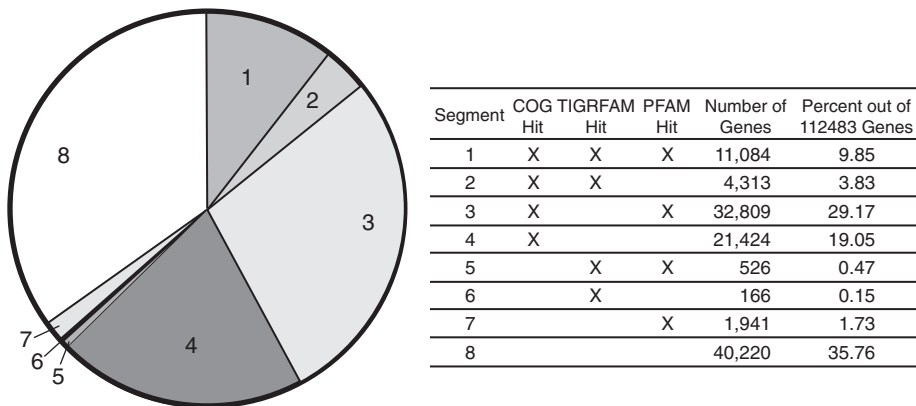


Figure 1.3. Number of protein-coding genes in the complete microbial genomes (a total of 112,483 based on the TIGR-assigned annotation) showing hits (marked with ×) with any combination of the NCBI protein families (cluster of orthologous genes, COGs), the TIGR protein families (TIGRfam), and the Pfam domain database. The original color code is rendered here as a gray scale. (Data from the TIGR Comprehensive Microbial Resource, July 2001; see the URL in the Appendix.)

groups of proteins, COGs), the TIGR protein families (TIGRfam), or Pfam. It is striking to note that more than 35% of proteins predicted are still completely unknown, as no hit has been found to known proteins or protein domains. Microbial genes identified so far in completely sequenced genomes range from a minimum of 470 genes in *Mycoplasma genitalium* to a maximum of 6752 genes in *Mesorhizobium loti*. Figure 1.4 shows the distribution of genes in the diverse functional families.

Gene number seems to reflect bacterial lifestyle. Specialized parasites (e.g., *Mycoplasma*) have about 500 to 600 genes, while *Myxococcus xanthus*, which has a complex life cycle (with a sporulation phase, etc.) encodes about 10,000 proteins (Casjens 1998; Shimkets 1998). In *M. loti*, *P. aeruginosa*, *E. coli*, and *B. subtilis*, the number of genes exceed 4000, to cover more than 87% of the genome. In the latter cases the size of the genome reflects evolutionary events and certainly cannot be considered as the minimum size for independent life.

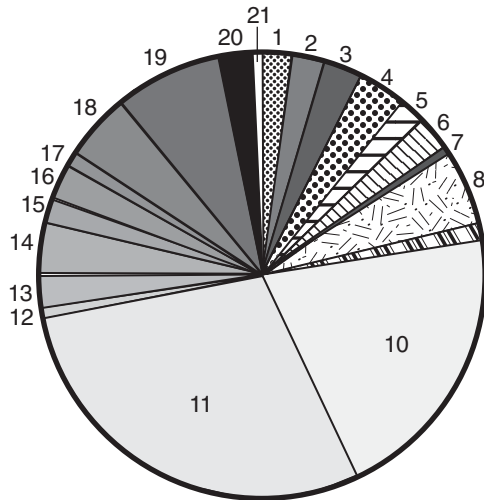
Another unexpected property is the extreme variability not only in gene content but also in gene order, even between closely related species and among independent isolated genomes of the same species (e.g., *Salmonella typhi*; Liu and Sanderson 1995).

Archaeal genomes have a coding capacity of about 90%, and unidentified proteins account for about 40% of the ORFs predicted, whose gene number ranges between 1700 and 2900 (see Table 1.2). A gene number around 2000 already exceeds what is necessary for a lithoautotrophic lifestyle.

In Archaea, duplicated regions are generally present even in the smallest organisms, such as *Methanobacterium thermoautotrophicum*, and it has been suggested that gene duplication could provide metabolic flexibility. Lateral transfer has been detected in the two methanogenic species (*M. thermoautotrophicum* and *Methanococcus jannaschi*) and in *Thermoplasma acidophilum*. Table 1.2 reports the number of open-reading frames (ORFs) for each completely sequenced genome, the percentage of proteins identified and of protein-coding regions, the presence or absence of lateral transfer events, and the degree of conservation of operonal gene organization (the latter topic is discussed below).

Comparative analysis based on data available for completely sequenced prokaryotic genomes has revealed a picture of generally well conserved protein sequences as opposed to the very scarce conservation of gene organization. Indeed, the structure of bacterial genomes has a high level of plasticity. This can be appreciated from the great variation in shape and size of genomes, from the number of replicons in the various organisms, and to a greater extent, from the different genome structures found even between closely related species.

Jacob and Monod (1961) were the first scientists to study a transcriptionally regulated system, the lactose metabolism system in *E. coli*. They discovered that to maximize gene regulation efficiency in prokaryotes, the enzymes for a particular metabolic pathway are often grouped in a cluster that is transcribed into a polycistronic mRNA from a single promoter sequence. This unit of bacterial gene expression and regulation, which includes structural genes and their control elements, is called an *operon*. An operon is made up of several elements: an *operator*, which is the binding site for repressor molecules; a *promoter*, which contains the binding site for RNA polymerase; and a *repressor*, a gene encoding a protein that binds to DNA at the operator and blocks the binding of RNA polymerase at the promoter.



Segment	Gene Role	Number of Genes	Percent out of 112,483 Genes
1	Amino acid biosynthesis	2,398	2.13
2	Biosynthesis of cofactors, prosthetic groups, and carriers	2,444	2.17
3	Cell envelope	3,107	2.76
4	Cellular processes	3,707	3.29
5	Central intermediary metabolism	2,100	1.86
6	DNA metabolism	2,769	2.46
7	Disrupted reading frame	107	0.09
8	Energy metabolism	6,909	6.14
9	Fatty acid and phospholipid metabolism	1,614	1.43
10	Hypothetical proteins, conserved	23,544	20.9
11	Hypothetical proteins, not conserved	31,941	28.3
12	Other categories	999	0.88
13	Protein fate	2,763	2.45
14	Protein synthesis	4,670	4.15
15	Purines, pyrimidines, nucleosides, and nucleotides	1,744	1.55
16	Regulatory functions	3,301	2.93
17	Transcription	1,191	1.05
18	Transport and binding proteins	5,502	4.89
19	Unclassified	8,365	7.43
20	Unknown function	2,986	2.65
21	Viral functions	322	0.28

Figure 1.4. Distribution of genes from complete microbial genomes in the various functional categories. The original color code is rendered here as a gray scale. (Data from the TIGR Comprehensive Microbial Resource, July 2001; see the URL in the Appendix.)

At present, gene order appears not conserved over long evolutionary time, except for some essential operons, such as ribosomal protein operons (Mushegian and Koonin 1996; Koonin and Galperin 1997). There are several examples of differences in genome structure between closely related species. Rearrangement of gene order is frequently observed and can be due to deletions and insertions. Clearcut examples are *Mycoplasma genitalium* and *M. pneumoniae*, where seven blocks with the same gene order can be recognized, but they are shuffled in the two genomes. More intriguing still, differences occur in comparing two isolates of the same species; the genome of *Helicobacter pylori* strain J99 is about 24,000bp smaller than that of strain 26695, and 6 to 7% of the genes are specific to each strain (Alm, Ling et al. 1999). These “unique” genes are mostly located in a *plasticity zone* (probably a pathogenicity island). The same occurs in *Salmonella typhi* isolates, which show extensive chromosomal rearrangements. This lack of synteny can be due to homologous recombination events or to other dynamic processes whose nature remains to be understood.

Because of this fluidity, it is very difficult to establish general rules in bacterial genome organization. In many cases, genes belonging to operons in some genomes are found scattered in other genomes. This observation, although needing experimental validation, may suggest that the operonlike structure is not as diffuse as it was believed to be. However, the general finding of clusters of genes that perform similar or related functions, presumably operons, demonstrates that this is an important advantageous property which also applies in the case of horizontal transfers.

Another common feature is the presence of genes in a higher copy number near the origin of replication, and gene orientation in the same direction as replication (but *E. coli* is an exception), to minimize a head-on collision between transcription and replication (Casjens 1998).

1.5 BASE COMPOSITION

The base composition of prokaryotic genomes is extremely variable. In Bacteria the mean guanine and cytosine (GC) content of the genome ranges from 67% in *Deinococcus radiodurans* to 25.5% in *Ureaplasma urealyticum*. In Archaea such a range is closer, from 56.2% in *Aeropyrum pernix* to 31.3% in *Methanococcus jannaschi* (Table 1.3). Figure 1.5 shows the %GC distribution of completely sequenced genomes listed in Table 1.3.

Average base composition in protein coding genes is not very interesting per se, except in extremely biased cases, since owing to the peculiar properties of the genetic code (i.e., degeneracy) it should not affect expression patterns. In this respect, Dayhoff (1978) has shown that, in general, average proteins can be made by RNAs having the most variable base compositions. In this respect, Lobry (1997) has shown that the average amino acid composition of bacterial proteins is greatly influenced by genomic G + C content. However, this influence was often found to be lower than expected, assuming neutrality in the evolutionary pattern.

The average value of base composition as well as its variation along the genome is an extremely interesting feature that can shed light on the evolutionary history of a given organism and also be used in applications of genetic engineering. It is generally believed that the mean GC content is related to bacterial phylogeny

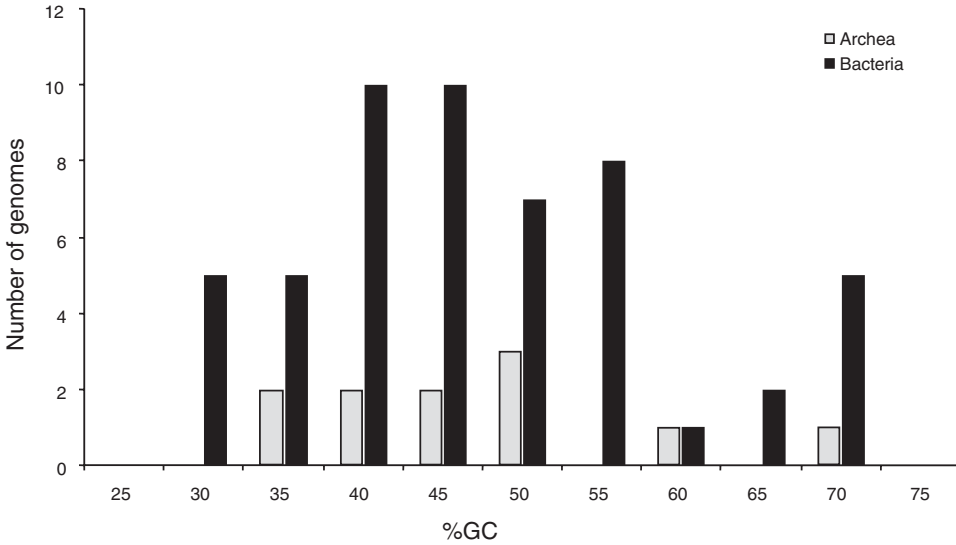


Figure 1.5. Distribution of %GC for the 64 bacterial and archaeal completely sequenced genomes listed in Table 1.1.

(Muto and Osawa 1987) as supported by the phylogenetic tree based on 5S rRNA sequences (Hori and Osawa 1986). Among gram-positive bacteria, those with high or low genomic GC content cluster together. Other authors have also suggested the existence of a biased mutation pressure, called *AT/GC pressure* by Muto and Osawa (1987), which causes differences in the genomic GC content among different lineages. This AT/GC pressure could have played a major role in diversification of bacterial genomic sequences and codon use during evolution.

Completely sequenced genomes reveal that there is very often a nonhomogeneous pattern of base composition along the genome which may reflect important aspects of genomic organization: (1) formation of compartments (i.e., AT- or GC-rich traits of the genome, due to neutral or selective events; see below), (2) acquisition of new material from outside (e.g., horizontal genes transfer), and (3) insertion/deletion events. In all cases, this behavior interferes with the normal divergent pattern of vertical evolution and complicates the phylogenetic analysis of organisms and, even more so, the measurements of their genetic distances (see Sec. 7.3).

The presence of distinct regions having different %GC values has been described in several microbial genomes, such as the archaeon *Archaeoglobus fulgidus* (five regions) and in the bacterial species *Helicobacter pylori* (strain 26695: five regions; strain J99: nine regions; see Table 1.3). In some species, a region with GC content higher than in the remaining genome has been found: for example, in the rRNA operons of *Aquifex aeolicus* and *Mycoplasma genitalium*. Other organisms have a lower GC value in some regions: for example, in genes for polysaccharide production of the archaeon *Methanobacterium jannaschi* [probably a mark of a lateral transfer event (Bult, White et al. 1996)], and in *Bacillus subtilis*, where many (A + T)-rich islands are present, probably a remnant of bacteriophage lysogens or other inserted elements (Kunst, Ogasawara et al. 1997). Figure 1.6 shows the %GC

TABLE 1.3. Compositional Features of Completely Sequenced Microbial Genomes

Species	Nucleotide Composition						
	%A	%T	%C	%G	GC Skew	AT Skew	%G + C
<i>Archaea</i>							
<i>Aeropyrum pernix</i>	21.50	22.10	28.30	27.90	-0.0071	-0.0138	56.2
<i>Archaeoglobus fulgidus</i>	25.80	25.60	24.20	24.30	0.0021	0.0039	48.5
<i>Halobacterium</i> sp. NRC-1 (3 replicons)	16.00	16.00	34.00	33.90	-0.0015	0.0000	67.9
<i>Methanobacterium thermoautotrophicum</i>	25.00	25.30	24.70	24.80	0.0020	-0.0060	49.5
<i>Methanococcus jannaschii</i>	34.40	34.10	15.50	15.80	0.0096	0.0044	31.3
<i>Pyrococcus abyssi</i>	27.50	27.70	22.40	22.20	-0.0045	-0.0036	44.6
<i>P. horikoshii</i>	28.90	29.10	21.20	20.60	-0.0144	-0.0034	41.8
<i>Sulfolobus solfataricus</i>	31.90	32.20	17.80	17.90	0.0028	-0.0047	35.7
<i>Thermoplasma acidophilum</i>	27.10	26.80	22.90	23.00	0.0022	0.0056	45.9
<i>T. volcanium</i>	30.10	29.90	19.90	20.00	0.0025	0.0033	39.9
<i>Sulfolobus tokodaii</i>	33.43	33.78	16.29	16.50	0.0066	-0.0052	32.8
<i>Bacteria</i>							
<i>Aquifex aeolicus</i>	28.40	28.10	21.60	21.70	0.0023	0.0053	43.3
<i>Bacillus halodurans</i>	28.20	28.00	21.60	22.00	0.0092	0.0036	43.6
<i>B. subtilis</i>	28.10	28.30	21.80	21.70	-0.0023	-0.0035	43.5
<i>Borrelia burgdorferi</i>	35.40	35.90	14.32	14.20	-0.0042	-0.0070	28.5
<i>Buchnera</i> sp. APS	37.00	36.60	13.00	13.20	0.0076	0.0054	26.2
<i>Campylobacter jejuni</i>	34.80	34.60	15.30	15.20	-0.0033	0.0029	30.5
<i>Caulobacter crescentus</i>	16.40	16.30	33.60	33.50	-0.0015	0.0031	67.1
<i>Chlamydia pneumoniae</i> AR39	29.50	29.80	20.20	20.30	0.0025	-0.0051	40.5
<i>C. pneumoniae</i> CWL029	29.80	29.50	20.30	20.20	-0.0025	0.0051	40.5
<i>C. trachomatis</i> MoPn	29.80	29.80	20.10	20.10	0.0000	0.0000	40.2
<i>C. trachomatis</i> serovar D	29.40	29.20	20.60	20.60	0.0000	0.0034	41.2
<i>Chlamydia pneumoniae</i> J138	29.86	29.56	20.33	20.26	-0.0017	0.0051	40.6
<i>Clostridium acetobutylicum</i>	34.57	34.51	15.42	15.50	0.0026	0.0009	30.9
<i>Deinococcus radiodurans</i> R1 (chromosome I)	16.90	16.30	33.30	33.30	0.0000	0.0181	66.6
<i>D. radiodurans</i> R1 (chromosome II)	16.40	16.50	33.50	33.40	-0.0015	-0.0030	66.9
<i>Escherichia coli</i> K-12	24.60	24.50	25.40	25.30	-0.0020	0.0020	50.7
<i>E. coli</i> O157:H7 EDL933	24.50	24.40	24.90	24.90	0.0000	0.0020	49.8

<i>E. coli</i> O157:H7 Sakai	24.70	24.70	25.20	25.30	0.0020	0.0000	50.5
<i>Haemophilus influenzae</i> Rd	31.00	30.80	19.10	18.90	-0.0053	0.0032	38.0
<i>Helicobacter pylori</i> 26695	30.30	30.80	19.60	19.20	-0.0103	-0.0082	38.8
<i>H. pylori</i> 26696	16.80	16.50	33.50	32.90	-0.0090	0.0090	66.4
<i>H. pylori</i> J99	30.30	30.40	19.60	19.40	-0.0051	-0.0016	39.0
<i>Lactococcus lactis</i> subsp. <i>lactis</i>	32.30	32.20	17.50	17.70	0.0057	0.0016	35.2
<i>Listeria innocua</i>	31.27	31.29	18.87	18.57	-0.0080	-0.0003	37.4
<i>L. monocytogenes</i>	31.05	30.97	19.13	18.85	-0.0074	0.0012	38.0
<i>Mesorhizobium loti</i>	18.60	18.60	31.60	31.10	-0.0080	0.0000	62.7
<i>Mycobacterium leprae</i>	21.00	21.10	28.70	29.00	0.0052	0.0024	57.7
<i>M. tuberculosis</i>	17.10	17.10	32.80	32.70	-0.0015	0.0000	65.5
<i>Mycoplasma genitalium</i>	34.50	33.70	15.70	15.90	0.0063	0.0117	31.6
<i>M. pneumoniae</i>	29.40	30.50	20.00	19.90	-0.0025	-0.0184	39.9
<i>M. pulmonis</i>	36.90	36.30	13.30	13.30	0.0000	0.0082	26.6
<i>Neisseria meningitidis</i> MC58	24.20	24.20	25.50	25.90	0.0078	0.0000	51.4
<i>N. meningitidis</i> Z2491	23.90	24.10	25.80	25.90	0.0019	-0.0042	51.7
<i>Nostoc</i> sp. PCC7120	29.29	29.36	20.64	20.71	0.0016	-0.0013	41.3
<i>Pasteurella multocida</i>	29.80	29.70	19.90	20.40	0.0124	0.0017	40.3
<i>Porphyromonas gingivalis</i> W83	25.80	25.80	24.00	24.20	0.0041	0.0000	48.2
<i>Rickettsia conorii</i>	33.68	33.88	16.12	16.32	0.0063	-0.0030	32.4
<i>R. prowazekii</i>	35.30	35.60	14.30	14.60	0.0104	-0.0042	28.9
<i>Salmonella enterica</i> thyphi	23.92	23.99	26.01	26.09	0.0015	-0.0015	52.1
<i>S. typhirium</i> LT2	23.90	23.88	26.11	26.11	-0.0001	0.0004	52.2
<i>Sinorhizobium meliloti</i>	18.64	18.63	31.49	31.24	-0.0039	0.0001	62.7
<i>Staphylococcus aureus</i>	33.40	33.60	16.30	16.40	0.0031	-0.0030	32.7
<i>Streptococcus pneumoniae</i>	30.18	30.10	19.80	19.92	0.0029	0.0013	39.7
<i>S. pneumoniae</i> TIGR4	30.26	30.04	19.75	19.95	0.0049	0.0036	39.7
<i>S. pyogenes</i>	30.80	30.50	19.00	19.40	0.0104	0.0049	38.4
<i>Synechocystis</i> PCC6803	26.00	26.10	23.80	23.80	0.0000	-0.0019	47.6
<i>Thermotoga maritima</i>	26.90	26.70	22.70	23.40	0.0152	0.0037	46.1
<i>Treponema pallidum</i>	23.50	23.60	26.20	26.50	0.0057	-0.0021	52.7
<i>Ureaplasma urealyticum</i>	37.20	37.20	12.50	12.90	0.0157	0.0000	25.4
<i>Vibrio cholerae</i> (chromosome I)	25.90	26.30	23.70	23.90	0.0042	-0.0077	47.6
<i>V. cholerae</i> (chromosome II)	26.40	26.60	23.20	23.60	0.0085	-0.0038	46.8
<i>Xylella fastidiosa</i>	22.50	24.70	24.90	27.70	0.0532	-0.0466	52.6
<i>Yersinia pestis</i>	26.21	26.16	23.69	23.94	0.0052	0.0009	47.6

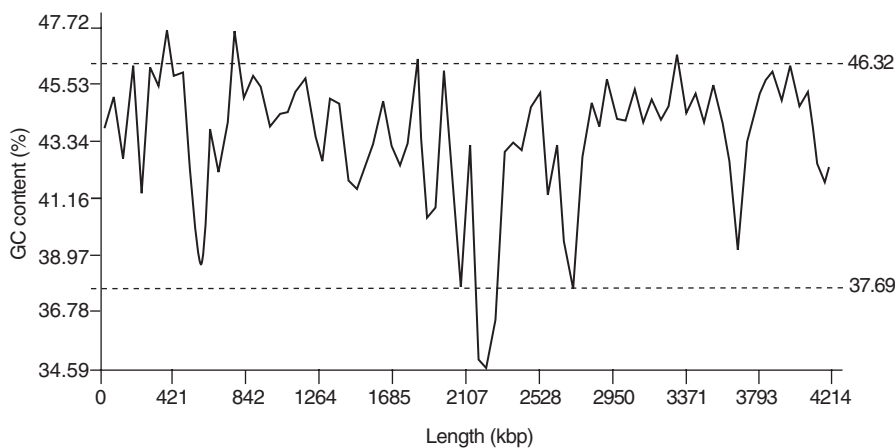


Figure 1.6. %GC content display for the *Bacillus subtilis* 168 DNA molecules generated by the TIGR facility. The genome is represented on the *X*-axis from left to right. Along the *Y*-axis of the plot is the average %GC for a “window” of 50kbp. The two dashed lines represent the 5% lower limit and the 95% upper limit average GC content for the DNA molecule shown; 90% of the GC content of this genome falls within these two dashed lines.

variation along the *B. subtilis* genome where GC-low and GC-rich regions can be distinguished.

GC variations along the genome could be explained in many different ways, and it should be assumed that the concept of isochores in prokaryotes is probably different from that proposed for eukaryotes; this issue is discussed in Part III. Beside nucleotide frequency variation along the genome, we can have an asymmetric distribution of the two complementary base pairs in DNA strands, called *AT* or *GC skews*. These can be measured in terms of two parameters, defined as *AT* and *GC skew*:

$$AT_{skew} = \frac{A - T}{A + T} \quad GC_{skew} = \frac{G - C}{G + C}$$

where A, C, G, and T are the occurrences of the relevant bases in the sequence under investigation. In the case of intrastrand parity of complementary bases (e.g., $A \approx T$ and $C \approx G$), the skew index is close to zero and indicates the absence of strand compositional asymmetry. On the contrary, if one of the bases significantly outnumbers its complementary base (e.g., $A \gg T$), a remarkable positive or negative skew index is observed. In general, the skew index may range between -1 ($A = 0$ or $G = 0$) and $+1$ ($T = 0$ or $C = 0$).

Figure 1.7 shows the distribution of global *AT* and *GC skew* calculated on microbially complete genomes, listed in Table 1.3. It is evident that both *AT* and *GC skew* do not significantly deviate from zero for archaeal and bacterial genomes, thus generally obeying the *second parity rule*, first put forward by Erwin Chargaff (1950), whereby there is intrastrand equivalence for complementary bases (i.e., $A = T$ and $G = C$). The genome of *Xylella fastidiosa* represents the only notable exception, as a deviating *GC* and *AT skew* is observed (Fig. 1.7).

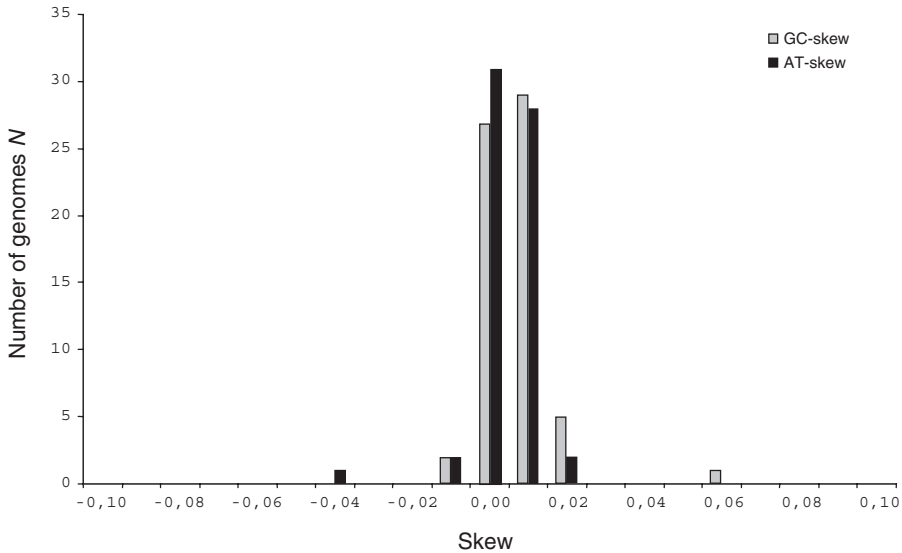
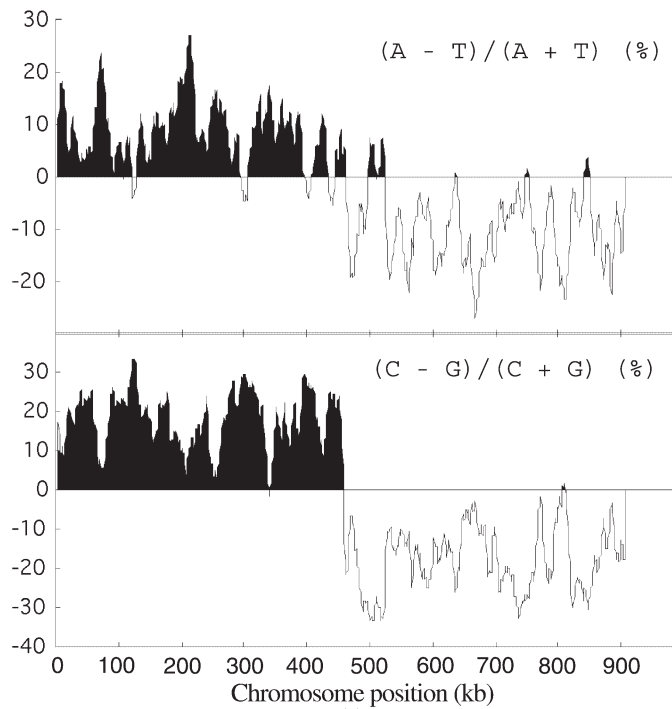


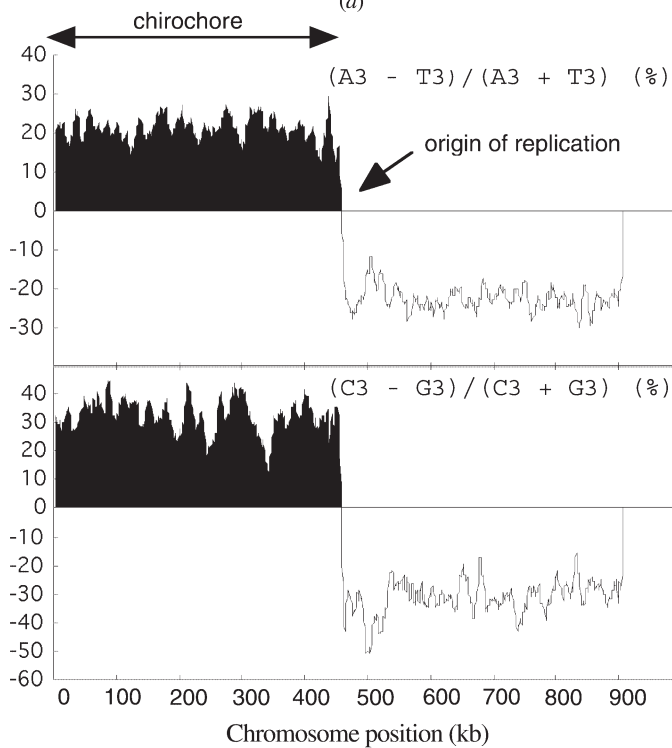
Figure 1.7. Distribution of global AT and GC skew calculated on microbial complete genomes listed in Table 1.3.

On the contrary, traits with positive or negative AT or GC skews have been found along the prokaryotic genomes. This has been correlated with regions involved in processes that generate single-stranded tracts of the genome, such as DNA replication or transcription. Lobry (1996) first demonstrated the existence of an asymmetric substitution patterns in the leading and lagging strands of three bacterial genomes (*E. coli*, *Bacillus subtilis*, and *Haemophilus influenzae*). This asymmetry divides the bacterial chromosome into two segments (chirocores) whose boundaries coincide with the origins of replication. The author suggests that this asymmetry could be due to a mutational bias (different mutation rates in the two strands as a consequence of asymmetries in replication or repair) rather than to a selective bias. Since then, asymmetries in base composition between the two strands have been observed in many prokaryotic genomes (Blattner, Plunkett et al. 1997; McLean, Wolfe et al. 1998; Mrazek and Karlin 1998), as in Fig. 1.8, where the base skew transition around the replication origin is shown for the *Borrelia burgdoferi* genome. The existence of mutational differences between the leading and lagging strands has been confirmed by Tillier and Collins (2000) to be the cause of base composition skew in bacterial genomes. The authors have demonstrated that the sequence of a gene product could be influenced by whether it is encoded on the leading or lagging strand.

The genome linguistic analysis (see Sec. 6.8) can also represent a powerful tool to investigate general genome properties. The analysis of w -mer use in complete bacterial genomes has revealed several peculiar properties. In *E. coli* the most frequent oligomers in the leading strand of replication form a family containing the trimer CTG; the palindromic tetramer CTAG is underrepresented (Karlin, Mrazek et al. 1997), and this could be explained by the hypothesis by some authors that



(a)



(b)

Figure 1.8. Base skew transition around the experimentally mapped replication origin of *Borrelia burgdoferi*: (a) genome sequence; (b) third codon positions where the chirochore structure is enhanced. Adapted with permission from CNRS UMR 5558, *Biométrie-Génétique et Biologie des Populations: Rapport d'activités, 1994-1998*, Université Claude Bernard, Lyon, France, June 1998. <http://biomserv.univ-lyon1.fr/unrfranc/sommaire.html>.

CTAG may “kink” DNA and thereby interfere with function (Medigue, Viari et al. 1991).

In *B. subtilis*, analysis of the abundance of oligonucleotides has revealed the existence of a dinucleotide bias: The most overrepresented ones are AA, TT, and GC, and those less represented are TA, AC, and GT. The distribution of words of 4, 5, and 6 nt seems to be significantly correlated to replication, as several of these words are very significantly overrepresented in one strand and underrepresented in the other (Kunst, Ogasawara et al. 1997).

1.6 CODON USE

Taking into account the genome compositional constraints reported above and the properties of the genetic code, it is conceivable that synonymous codons are not used equally in Bacteria. Until recently it was generally accepted that the most important factors affecting codon use variation in prokaryotic genomes were GC-base compositional bias and selection at the translational level; indeed, these were the criteria allowing discrimination between high- and low-expressed genes (Sharp, Stenico et al. 1993). According to Shpaer (1986) and Gouy (1987) in Bacteria, synonymous codon use may also be influenced by base composition at neighboring sites; these context-dependent codon biases are widespread but not conserved among the different bacterial species (McVean and Hurst 2000).

A compilation of codon use tables for several prokaryotic genomes is given in the CUTG database (Nakamura, Gojobori et al. 2000) and at the TIGR Comprehensive Microbial Resource (CMR; see the URL in the Appendix). Codon use tables can also be obtained by considering subsets of genes belonging to specific functional categories or falling within user-selected coordinates for a chosen chromosome. Figure 1.9 shows the codon use chart for all protein-coding genes of the *E. coli* O157:H7 genome.

In comparative genome studies, some authors have defined several indexes to investigate codon use strategy (see Sec. 6.8.9). The preference of a specific codon within its family can be expressed by the ratio between the number of its occurrences and the total occurrences of the corresponding amino acid. If no codon preference is exhibited within a family of synonymous codons, the codon use percentage is simply equal to $100/n$, where n is the number of synonymous codons within the family.

The *effective number of codons* used in a gene (ENC index) and the *Codon Adaptation Index* (CAI) are the most used indexes to measure bias in codon use strategy (i.e., how far the codon use of a gene departs from equal use of synonymous codons). The effective number of codons (ENC), proposed by Wright (1990), is calculated through a method used in population genetics to determine the effective number of alleles segregating in a given population. The larger the variety of synonymous codons used by a gene, the larger the ENC, with a minimum expected value of 20 and a maximum of 61; if synonymous codon use in a given gene is random, the ENC value approaches 61 (see also Sec. 6.8.9).

Sharp and Li (1987) proposed the Codon Adaptation Index (CAI) to measure how closely the codon use of a specific gene matches the optimum codon use strategy in a set of reference genes from the same organism. To determine whether

	T	C	A	G
T	TTT } Phe (F) 2.26% 36987 TTC } 1.82% 28545	TCT } 0.86% 14047 TCC } 0.89% 14614	TAT } Try (Y) 1.65% 26974 TAC } 1.23% 20114	TGT } Cys (C) 0.53% 8749 TGC } 0.66% 10906
	TTA } Leu (L) 1.39% 22692 TTG } 1.33% 21866	TCA } 0.81% 13232 TCG } 0.88% 14361	TAA } STOP 0.20% 3374 TAG } 0.02% 480	TGA } STOP 0.11% 1950 TGG } Trp (W) 1.51% 24795
C	CTT } 1.14% 18666 CTC } 1.05% 17222	CCT } 0.72% 11764 CCC } 0.55% 9109	CAT } His (H) 1.30% 21323 CAC } 0.94% 15500	CGT } 2.01% 32940 CGC } 2.09% 34167
	CTA } Leu (L) 0.38% 6316 CTG } 5.07% 82751	CCA } Pro (P) 0.84% 13709 CCG } 2.23% 36422	CAA } Gln (Q) 1.46% 23970 CAG } 2.94% 48113	CGA } Arg (R) 0.38% 6290 CGG } 0.63% 10392
A	ATT } Ile (I) 2.95% 48303 ATC } 2.39% 39062	ACT } 0.90% 14737 ACC } 2.28% 37266	AAT } Asn (N) 1.91% 31267 AAC } 2.17% 35456	AGT } Ser (S) 0.93% 15252 TGC } 1.81% 28341
	ATA } Ile (I) 0.55% 9054 ATG } Met (M) 2.69% 43913	ACA } Thr (T) 0.80% 13075 ACG } 1.50% 24511	AAA } Lys (K) 3.39% 55430 AAG } 1.09% 17841	AGA } Arg (R) 0.29% 4801 AGG } 0.19% 3115
G	GTT } 1.82% 29762 GTC } 1.48% 24309	GCT } 1.53% 25093 GCC } 2.51% 41065	GAT } Asp (D) 3.26% 53271 GAC } 1.89% 30952	GGT } 2.40% 39261 GGC } 2.79% 45685
	GTA } Val (V) 1.09% 17794 GTG } 2.59% 42375	GCA } Ala (A) 2.04% 33444 GCG } 3.21% 52503	GAA } Glu (E) 3.89% 63546 GAG } 1.83% 29982	GGA } Gly (G) 0.87% 14284 GGG } 1.15% 18885

Figure 1.9. Codon use chart for all protein-coding genes of the *E. coli* O157:H7 genome. For each codon triplet the percent codon frequency and number of codon occurrences are shown. (Data from the TIGR Comprehensive Microbial Resource; see the URL in the Appendix.)

codon use was correlated with gene expression, they took a set of highly expressed genes in the bacterium *E. coli* as a reference set and showed that this was in fact correlated with the gene expression level in *E. coli*.

The availability of a higher number of completely sequenced genomes has contributed greatly to this scenario with new and greater details. Pan, Dutta et al. (1998) have shown that codon use in highly expressed genes of the eubacterial species *Haemophilus influenzae* and *Mycobacterium tuberculosis* is biased. In particular, they demonstrated the existence of a preference for G-starting codons by highly expressed genes. This could be a general feature of bacteria, irrespective of their overall GC content. Kanaya, Yamada et al. (1999), by correlating codon use and tRNA abundance for several unicellular organisms, observed that codons preferred in highly expressed genes were related to the codons that are optimal for the translation process, as predicted by the composition of isoaccepting tRNA genes.

McInerney (1998) introduced the concept of the influence of replicational/transcriptional selection on the codon use of a given gene. He has analyzed the complete genome sequence of *Borrelia burgdorferi*, demonstrating that in this species the genes may adopt two different codon use strategies, depending on whether the gene is located on the leading or lagging strand of replication. The GC skew, which is due to an asymmetric replication mechanism, causes codon use variations. In other words, the mutational bias between the two strands, together with transcriptional selection, would be responsible for the location of the most highly expressed genes in the leading strand of replication. This, in turn, correlates to GC skew (see the previous paragraph).

Also, in *Chlamidia trachomatis* the choice of synonymous codons is the result of several factors, the most important being gene location, hydropathy, and the degree of conservation of the protein under examination (Romero, Zavala et al. 2000). In *E. coli* the codon bias is influenced by both replication (Bulmer 1990) and translation selection; the latter mechanism is responsible for two codon biases, both of them context-dependent: the avoidance of AGG motifs and the avoidance of out-of-frame stop codons (Smith and Smith 1996). Recently, McVean and Hurst (2000) have suggested that the underrepresentation of AGG motifs in the *E. coli* genome is the result of selective forces, while the avoidance of out-of-frame stop codons is the consequence of the mutational bias.

In *B. subtilis* three classes of genes have been recognized, according to different codon use (Kunst, Ogasawara et al. 1997): class 1, including the majority of the genes (3375 genes), among which are most of the genes for sporulation; class 2 (188 genes), including genes expressed under exponential growth conditions; and class 3 (537 genes), with codons enriched in AT residues.

In *M. tuberculosis*, a comparative analysis has revealed a statistically significant preference for the amino acids Ala, Gly, Pro, Arg, and Trp, which are all encoded by GC-rich codons, and a comparative reduction in the use of the amino acid encoded by AT-rich codons (Asn, Ile, Lys, Phe and Tyr; Cole, Brosch et al. 1998). Himmelreich, Hilbert et al. (1996), in their analysis of the codon use of *M. pneumoniae*, were able to distinguish gene subsets with low (below 35%) and high (50 to 56%) GC content. The codon use of the low- and high-(G + C)-content subfractions is influenced by the base composition, favoring codons with either G-C or A-T at the third position; and it is also related to genes that are frequently expressed like those coding for ribosomal proteins.

Campylobacter jejuni has a genome with a very low GC content (30.4%); this seems to affect the codon use, since codons ending in A or T are strongly preferred (Gray and Konkel 1999). In *Helicobacter pylori* the low GC (39%) partially reflects the synonymous codon use, but there is no evidence for translational selection or biased mutation patterns among synonymous codons (Lafay, Atherton et al. 2000).

1.7 REPLICATION AND EXPRESSION

In Bacteria, most studies have been aimed at the recognition of enzymes involved in replication, transcription, and in protein synthesis. These studies have been based mainly on present knowledge; that is, they refer to the most studied prokaryotes, primarily *E. coli*. It is well known in this bacterium that the replication of circular double-stranded DNA is initiated by the *initiator proteins* (e.g., dnaA protein) binding to the origin of replication (a specific point on the DNA, which in *E. coli* is oriC) and bending the double-stranded DNA around it. The following step is the binding of helicases and primases to this complex. Helicases are able to unravel and separate the double-stranded DNA helix, while primases (primosomes), synthesize the RNA primer. At this stage, the DNA polymerase enters the replication forks and starts synthesizing a DNA strand complementary to the template strand. The speed of replication is about 1000 bp per second and thus the *E. coli* genome is fully replicated in about 40 min.

E. coli contains three distinct enzymes capable of catalyzing the replication of DNA: DNA polymerases (pol) I, II, and III. DNA polymerases I and II appear to perform *proofreading/editing functions*, moving along the DNA and correcting mismatched base pairings. Indeed, normal bacteria show errors in replication at a rate of approximately 10^{-9} per base pair, while mutants with defective pol I and pol II show errors at a rate of approximately 10^{-5} per base pair. DNA polymerase III is the main polymerase for replication. This enzyme is much less abundant than pol I; however, its activity is nearly 100-fold that of pol I. DNA pol I was the first replication enzyme to be characterized in *E. coli*. It is made of a single polypeptide of 103 kDa coded by the locus *polA*. The chain can be cleaved by proteolysis in two fragments, a 68-kDa Klenow fragment, which possesses the 3'-5' exonuclease and the polymerase activities, and a 35-kDa small fragment containing the 5'-3' exonuclease activity.

The complete sequencing of bacterial genomes has allowed us to discover the features distinguishing each species from the general pattern of *E. coli*. In all the complete bacterial genomes sequenced so far, the primary DNA polymerase corresponds to the DNA polymerase III (pol III) holoenzyme in *E. coli*, although some variation may occur, particularly in the number of subunits of the core structure. The *E. coli* pol III is made up of several subunits, and the assemblage of the holoenzyme is a multistep process. The core enzyme contains three subunits: α (130 kDa), which synthesizes DNA; ϵ (25 kDa), which has a 3'-5' proofreading exonuclease activity; and θ (10 kDa), which probably is required for assembly. The addition of the τ subunit (71 kDa), leads the core enzyme to dimerize, generating the pol III* complex; then the addition of the complex γ made up of several subunits, of which the best characterized are γ (52 kDa, alternative product of the gene coding for τ subunit) and δ (32 kDa) creates the pol III' complex, which finally generates the holoenzyme after the addition of the β subunit (40 kDa).

In *Aquifex aeolicus*, an additional member of the γ - τ / δ' family is present among the pol III subunits; in *Borrelia burgdorferi* and *Mycoplasma genitalium*, only 4 of the 10 polypeptides (α , β , γ , τ) of the *E. coli* pol III have been identified. *Mycoplasma pneumoniae* codes for two potential α subunits; and only the subunits β (*dnaN*), δ' (*holB*), γ , and τ (*dnaX*) are present. This may indicate a simplified replication complex compared to those of other gram-negative bacteria, such as *E. coli* and *H. influenzae*. Similar to what happens in other minimal genomes (e.g., *M. pneumoniae* or *B. burgdorferi*), the genome of *Treponema pallidum* encodes for the α , β , ϵ , γ , and τ subunits of *E. coli* DNA polymerase III. *Rickettsia prowazekii*, an endocellular parasite, has a smaller set of genes involved in DNA replication; only four genes for the core structure of DNA polymerase III have been identified: α (*dnaE*), ϵ (*dnaQ*), β (*dnaN*), θ , and γ (*dnaX*) subunits. Further details on the replication mechanism in the genomes sequenced remain to be elucidated.

Bacterial transcription [i.e., the transfer of DNA genetic information to a complementary sequence of RNA nucleotides by the DNA-dependent RNA polymerase (RNA pol)] is also well described in *E. coli* and in a few other bacteria. According to these notions, the basic critical components for transcription are the subunits of RNA pol, the promoters (i.e., the DNA sequence of the operon recognized by the DNA-dependent RNA polymerase), and the σ factors, which stabilize the polymerase in order to start polymerization at specific sites.

In *E. coli*, RNA pol is a complex enzyme, made up of four kinds of subunits: the

α subunit (with a mass of 40kDa, encoded by the *rpoA* gene), which binds to the promoter; the β subunit (155kDa, encoded by the *rpoB* gene), whose function is the binding of ribose–triphosphate–organic base; the β' subunit (160kDa, encoded by the *rpoC* gene), which binds to DNA; and the σ subunit (70kDa, encoded by the *rpoD* gene), which is responsible for the initiation of transcription. In bacterial cells, the RNA pol exists in two forms: the core enzyme ($\alpha_2\beta\beta'$), which elongates RNA, and the holoenzyme ($\alpha_2\beta\beta'\sigma$), which initiates RNA synthesis.

Recognition of individual promoters is determined by the kind of σ factor present. In *E. coli* there are several σ factors, each named after the molecular weight (e.g., the 70-kDa σ in *E. coli* is called σ_{70}). Sigma size varies widely, from 32 to 92kDa.

Comparative genomics has shown that the transcription mechanism is mostly similar to that of *E. coli*, as also proved experimentally in some cases. The number and type of σ factors are extremely variable and appear to be a species-specific feature; for example, in *Bacillus halodurans*, 11 σ factors have been identified, of which 10 are unique to this species; they are probably needed for its hyperalkaline lifestyle; in *B. burgdorferi*, beside σ_{70} , two alternative σ factors (σ_{54} and *rpoS*) have been found; *R. prowazekii* encodes an alternative σ_{32} factor. In *B. subtilis*, 19 σ factors have been identified, which might be due to the sporulation phase present during its life cycle. Alternative σ factors play a key role in directing differentiation of the spore cell. During sporulation, many genes are expressed which are not expressed during vegetative growth, due to several alternative σ factors involved, each at a certain time, to initiate transcription of the gene subset needed at that time.

The consensus sequence recognized by the σ subunit of bacterial RNA polymerase is composed of two conserved boxes located at positions -35 and -10 with respect to the transcription start site. The distance between these two conserved boxes, whose consensus are TTGACA and TATAAT for the -35 and the -10 box, respectively, may range between 15 and 19bp. Figure 1.10 shows the base frequency matrices (see Sec. 6.8.4) for the *rpoD* recognition elements calculated from the known *E. coli* binding sites as well as the corresponding logo (see Sec. 6.8.6). Using pattern discovery methods (see Sec. 6.8.6), Thieffry, Salgado et al. (1998) predicted a large fraction of *E. coli* genome transcriptional regulatory sites. A comprehensive library of DNA binding sites for *E. coli* promoters can be found in the DPInteract Database (Robison, McGuire et al. 1998; see the URL in the Appendix).

In prokaryotes, transcription is coupled to translation—the synthesis of proteins by amino acid polymerization—because mRNAs are rapidly degraded, usually within a minute after transcription. Once a length of newly made RNAs has dissociated from the DNA, a ribosome can bind to it and protein synthesis can be initiated. So translation begins right after transcription and well before transcription termination. These processes have been studied extensively and their various steps are well known.

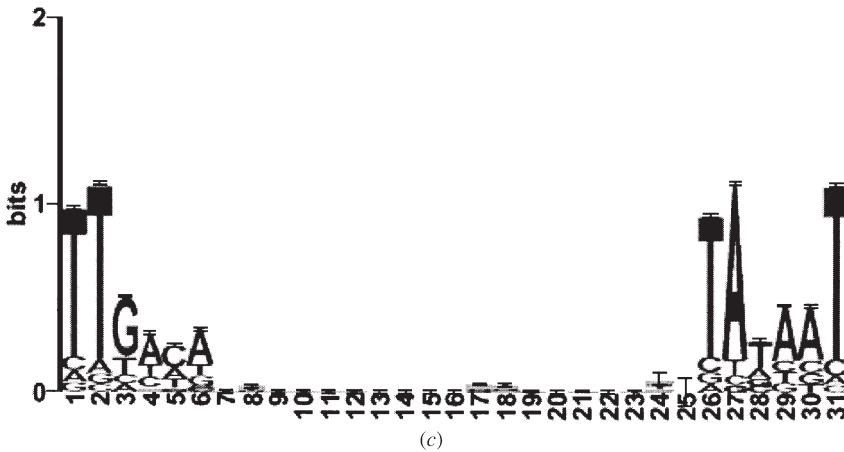
Genes encoding the 5S, 16S, and 23S rRNA are generally organized into an operon (called *rrn*) in the domain Bacteria. However, even if their clustering permits transcription of equimolar quantities of each rRNA gene, alternative organizations are observed in some prokaryotes, thus indicating that organization into operons is not a strict requirement. The number of rRNA operons per bacterial genome varies from 1 to as many as 15 copies, and this features seems to be required in species

Position	A	C	G	T	Consensus
01	16	19	13	202	T
02	18	10	12	210	T
03	21	24	161	44	G
04	137	46	20	47	A
05	55	126	20	49	C
06	143	24	41	42	A

(a)

Position	A	C	G	T	Consensus
01	12	22	17	199	T
02	210	12	9	19	A
03	46	32	36	136	T
04	159	33	29	29	A
05	158	37	31	24	A
06	11	21	9	209	T

(b)



(c)

Figure 1.10. Position weight matrices for (a) the -35 and (b) the -10 boxes from 250 *E. coli* promoters collected in the DPInteract Database. (c) Corresponding sequence logo. [(a, b) From Robison, McGuire et al. (1998).]

with a high growth rate. For example, the pathogenic bacteria *Mycoplasma pneumoniae* has one rRNA operon, while the enteric bacterium *E. coli* possesses seven copies per genome. The spore-forming bacteria *Bacillus subtilis* (10 copies per genome) and *Clostridium paradoxum* (15 copies) show the highest number of rRNA operons (Klappenbach, Dunbar et al. 2000).

Ribosomal RNA operons can be also involved in recombination events; indeed, inversions can occur between rRNA operons, but they are generally unstable. The *rrn* operons contain genes for both rRNA and tRNA, each functional sequence being separated from the next by transcribed spacer regions. The genes encoding transfer RNAs (tRNAs) are often located in the internally transcribed spacer (ITS) region and distal to the 5S rRNA gene. The ITS region is variable in length and is responsible for most sequence diversity between multiple rRNA operons in many

species of bacteria. The rRNA operon is transcribed as a single RNA molecule and then is cleaved by ribonucleases into mature rRNA and tRNA molecules. Very often the rRNA operons are localized around the origin of replication. This could provide a gene dosage effect during rapid growth; indeed, the effective number of rRNA operons in *E. coli* can be as high as 36 copies during exponential growth with multiple replication forks.

It is well known that tRNA number is species-specific. The most divergent pattern in protein synthesis in the various bacterial species regards tRNA synthetase genes. *Bacillus halodurans* lacks the glutamyl-tRNA synthetase gene (*glnS*), one of two threonyl-tRNA synthetase genes (*thrZ*), and one of two tyrosyl-tRNA synthetase genes (*tyrS*), while a member of the same genus, *B. subtilis*, lacks only the *glnS* gene; the same occurs in *Thermotoga maritima*, *Treponema pallidum*, *Mycoplasma genitalium*, and *M. pneumoniae*. Both strains of *Helicobacter pylori* sequenced up to now lack the asparaginyl-tRNA synthetase gene (*asnS*); in *Rickettsia prowazekii*, no *asnS* and *glnS* have been identified. In some cases the mechanisms used to overcome deficiencies are known, whereas in others they are still to be unraveled. In organisms lacking *glnS* gene, a single glutamyl tRNA synthetase aminoacylates both tRNA^{Glu} and tRNA^{Gln} with glutamate. Subsequently, the amidation of glutamate to glutamine leads to the formation of glutamyl tRNA synthetase (Freist, Gauss et al. 1997). Probably the lack of *asnS* gene may be overcome through a transamidation process forming Asn-tRNA^{Asn} from Asp-tRNA^{Asn} (Tomb, White et al. 1997).

In Archaea, many components of the replication and expression mechanisms are similar to those of eukaryotes. Indeed, this is the main reason why the Archaea were defined as a separate branch of prokaryotes in the tree of life, more closely related to eukaryotes (Woese, Kandler et al. 1990), even if from complete genome sequencing it appears that the resemblance of Archaea and Bacteria is more pronounced than expected (see below and Chapter 8).

The first studies of archaeal DNA replication *in vivo* showed that these species are sensitive to aphidicolin, a specific inhibitor of eukaryal but not bacterial replication polymerase [see Forterre and Elie (1993) for a review]. Since then, several eukaryotic-specific replication proteins have been identified in Archaea, including the origin recognition complex (ORC), the B DNA polymerase family, the Dna2 (3'-5' helicase), and the ATP-dependent DNA ligase (Edgell and Doolittle 1997). In some cases, bacterial and eukaryal proteins are also homologous, but the archaeal/eukaryal counterparts still show a higher degree of similarity (e.g., replication factor C, which is the sliding-clamp protein, or ribonuclease H).

Data reported for the completely sequenced Archaea have revealed that the archaeal replication apparatus does not have eukaryotic features only. In *Archaeoglobus fulgidus*, a homolog of the proofreading ϵ subunit of the *E. coli* pol III not previously recorded in Archaea has been observed; the *Halobacterium* NRC-1 genome encodes for three DNA polymerase types: two family B polymerases, a bacteriophage-like family A polymerase, and the heterodimeric family D polymerase. In *Methanobacterium thermoautotrophicum*, the presence of two Cdc6 homologs and three histones would support a DNA replication initiation and chromosome packaging with eukaryal features; however, the presence of an *ftsZ* gene indicates a bacterial type of cell division initiation. Very recently, comparative genomic studies on the hyperthermophilic archaeon *Pyrococcus abyssi* showed that this species has a bacterial mode of replication but uses eukaryotic-like machinery.

The eukaryotic-like replication genes lie around the origin of replication, which is strongly conserved, while (as in bacterial genomes) the replication terminus is a hot spot of genome shuffling (Myllykallio, Lopez et al. 2000).

Similarly, archaeal transcription and translation mechanisms seem to be a mosaic of eukaryal and bacterial features (Bell and Jackson 1998). The first studies on transcription in Archaea showed the existence of striking parallels with the eukaryal transcriptional apparatus: like Bacteria, Archaea have a single RNA polymerase that transcribes all genes, but it is very similar to that of eukaryotes in that it contains three or four large and many small subunits; moreover, archaeal RNA polymerases are similar in sequence and antigenicity to eukaryal RNA polymerase II. Promoters in Archaea are a -30 TATA-like sequence that is a binding site for a transcription factor (TFB), not σ -like subunits. These findings also contributed to the prevailing view that Archaea and Eukarya share a peculiar machinery that differs from that of their bacterial counterpart. However, further investigations have highlighted that transcription in Archaea has a mixed character. Indeed, the search for transcription-associated proteins in four completely sequenced genomes (*Methanococcus jannaschi*, *Archaeoglobus fulgidus*, *Methanobacterium thermoautotrophicum*, and *Pyrococcus horikoshii*) has revealed that out of 280 transcription-associated proteins predicted, only 51 have homologs in Eukarya only, while 168 have bacterial homologs as well. The remaining proteins have homologs in both phylogenetic domains, while some elements (such as GvpE, a transcriptional activator involved in the regulation of gas vesicle synthesis in halophilic archaea) are unique to Archaea (Kyrpides and Ouzounis 1999).

Deeper in detail, the homology with Eukarya is confined essentially to the initiation factors TFB (similar to the eukaryotic TFIIB, even if the two proteins bind to the DNA in inverse orientation in the two structures), the components of the RNA polymerase core enzyme (up to 15 components versus 4 in *E. coli*), and the TATA-box-binding protein (TBP, which displays about 40% identity with eukaryal TBP).

In Archaea translation also appears again to be a mixture of prokaryotic and eukaryotic features. There is evidence for a bacterial-like initiation process (i.e., leader sequence binding to ribosome), where several involved factors are related to bacterial analogs whereas others are of eukaryotic origin. The Archaea have 70S ribosomes (bacterial-sized), and the way they recognize the start codon resembles the bacterial process (Bell and Jackson 1998). However, archaeal translation is inhibited by diphtheria toxin, in a manner similar to what occurs for eukaryal ribosomes, but it is not inhibited by most bacterial-translation-inhibiting antibiotics. Archaeal pre-RNA processing resembles the bacterial pathway, but the Archaea possess a protein homologous to eukaryal fibrillarin (Dennis 1997). Translation is initiated with methionine (like Eukarya), not formylmethionine (like Bacteria); archaeal tRNAs contain introns like the eukaryotic introns. Finally, the Archaea have homologs of the eukaryotic elongation factors eEF-1a and eEF-2.

As to archaeal comparative genomics, as in Bacteria, the most divergent feature is the number of tRNA synthetases: *Archaeoglobus fulgidus* lacks both glutamine and asparagine synthetases; *Methanobacterium thermoautotrophicum* and *Methanococcus jannaschii* lack glutamine, cysteine, lysine, and asparagine synthetases. In *Halobacterium* sp., no asparagine and glutamine tRNA synthetases have been detected. *A. fulgidus* and *M. thermoautotrophicum* lack selenocysteine-tRNA gene; the former has only one rRNA operon and lacks the 5S rRNA gene.

From the data reported above, which are based on a limited set of genomes, several notions have already been confirmed and new ones have emerged, thanks to a comparison of the complete genomes sequenced until now. One of the most limiting factors in the exploitation of genomic data is that a large portion of the genome protein coding region remains to be elucidated. Thus functional genomics is becoming a rapidly growing field exploiting new methods of both dry and wet biology. Among the latter, RNA interference (RNAi) is giving good results.

The identification of gene function based on comparative analysis is often difficult since the similarity of homologous genes could be so low as to be hardly detectable and the setup of the genome may vary dramatically even between closely related species. Events of lateral gene transfer and/or genome fusion may be responsible for very important divergences in some tracts of related genomes. These events are more frequent than expected and have already created great arguments regarding the classification and evolution of prokaryotes.