Chimia 53 (1999) 558–561

ISSN 0009-4293

Genomics for Biotechnology

Václav Pačes

Abstract. A wealth of information has been gained from completely sequenced genomes. So far, most of the sequenced genomes are bacterial genomes. In addition to the basic metabolic pathways, various bacterial genomes encode pathogenicity, degradation of xenobiotics, synthesis of unusual compounds, or photosynthesis. The knowledge of the complete DNA sequence of bacterial genomes can facilitate considerably studies of these features as well as their practical applications in biotechnology. Many open reading frames (ORFs) found in bacterial genomes are identified with their function by a similarity search of standard databases. However, some of the bacterial genome projects are concluded by depositing the nucleotide sequence in a database with no simple means to study functions of those ORFs for which a similarity search did not allow convincing functional assignments. It is desirable to develop systems for easy functional analysis of these ORFs. *Rhodobacter capsulatus* is a bacterium that has the potential for developing such systems. Its genome harbors a defective phage called Gene Transfer Agent (GTA) that enables systematic deletions of DNA regions of various sizes. This unique feature, together with photosynthesis, nitrogen fixation, degradation of several pollutants, and synthesis of biodegradable plastic encoded by the *R. capsulatus* genome, make this bacterium an attractive subject of biotechnological applications.

1. Introduction

One of the major advances of molecular genetics in recent years is the development of methods for DNA sequencing. They became so efficient that determination of the complete DNA sequence of simple organisms is now possible. This development led to the establishment of a new branch of biology – genomics.

Genomics consists of three major fields: sequencing of a genome (structural genomics), complex analysis of the nucleotide sequence obtained (bioinformatics), and experimental assignment of genes and regulatory elements to their functions (functional genomics). Genomics yields unprecedented information about life. However, it not only gives answers to many basic questions about life and its evolution, but it also supplies us with tools for practical applications. Many projects of genetic and protein engineering, based on results of genomics, aim at medical, pharmaceutical, agricultural, and other applications.

Since 1995, twenty three different genomes were completely sequenced, and their nucleotide sequences have been published. These genomes are all microbial genomes, 22 are bacterial genomes, and one is the genome of a unicellular eukaryotic microorganism, the yeast *Saccharomyces cerevisiae*. In addition, the 97-megabase genomic sequence of the nematode *Caenorhabditis elegans* is essentially complete and was described recently by a consortium of specialists (*Table 1*).

Many more genome projects are at various stages of elaboration. Most of them are bacterial genome projects especially important for biotechnological applications. However, when searching various databases, one can see that most of the current bacterial genome projects are aimed at pathogenic bacteria (*Table 2*). The field of 'genomics for biotechnology' is less frequented and remains open to intensive research.

Detailed information on genome projects with complete references to publications in which reports originally appeared can be most readily obtained from specialized websites such as www.genome.ad.jp/kegg/kegg2.html, www-c.mcs.anl.gov/home/gaasterl/genomes.html, and www.tigr.org/tdb/mdb/mdb.html.

2. Sequencing Bacterial Genomes

The most common current strategy for sequencing bacterial genomes is based on unspecific fragmentation of the whole chromosome, sequencing of the individual cloned DNA fragments, and assembling longer contiguous sequences (contigs) using various computer programs. With this strategy, redundant information is obtained for most of the nucleotide sequences. Usually each base pair is sequenced six to eight times, and the sequence is confirmed from both DNA strands. This results in high reliability of the sequence data generated. However, when redundancy becomes too high, a more direct approach must be applied. Often, this happens when 80 to 90% of the sequence has been determined. At this stage, the chromosome is usually covered by several tens of contigs. Direct cloning of specific restriction fragments and/or primer walking are then often used to finish the genome sequencing.

3. *Rhodobacter capsulatus* Genome Project and Sequencing of DNA of Other Bacteria with Biotechnological Potential

In 1996, a collaborative *Rhodobacter capsulatus* genome project was launched at the University of Chicago, the Institute of Molecular Genetics of the Academy of Sciences in Prague, and the Institute of Chemical Technology in Prague. In this project, a specific sequencing strategy was

Correspondence: Prof. Dr. V. Pačes
Institute of Molecular Genetics
Academy of Sciences of the Czech Republic, and
Institute of Chemical Technology
CZ–16637 Prague, Czech Republic
Phone: (420-2) 20183541
Fax: (420-2) 24311019
E-Mail: vpaces@img.cas.cz

Table 1. *Selected Genome Projects*

Category	Species	Completed	Genome size (bp)	Genes
Actinobacteria	*Mycobacterium tuberculosis*	yes	4,411,529	4,397
Chlamydia	*Chlamydia trachomatis*	yes	1,042,519	937
Cyanobacteria	*Synechocystis sp. PCC6803*	yes	3,573,470	3,215
Gram-positive bacteria	*Bacillus subtilis*	yes	4,214,814	4,221
	Mycoplasma genitalium	yes	518,073	503
	Mycoplasma pneumoniae	yes	816,394	707
	Staphylococcus aureus	no		
Oxygen-reducing bacteria	*Aquifex aeolicus*	yes	1,551,335	1,572
Proteobacteria	*Escherichia coli*	yes	4,639,221	4,397
	Haemophilus influenzae	yes	1,830,135	1,791
	Helicobacter pylori	yes	1,667,867	1,609
	Rhodobacter capsulatus	no		
	Rickettsia prowazekii	yes	1,111,523	834
	Salmonella typhimurium	no		
Spirochetes	*Borrelia burgdorferi*	yes	910,724	1,279
	Treponema pallidum	yes	1,138,011	1,082
Archaea	*Archaeoglobus fulgidus*	yes	2,178,400	2,456
	Methanobacterium thermoautotrophicum	yes	1,751,377	1,914
	Methanococcus jannaschii	yes	1,664,987	1,813
	Pyrococcus horicoshii	yes	1,738,505	2,027
Fungi	*Candida albicans*	no		
	Saccharomyces cerevisiae (16 chromosomes)	yes	12,069,313	6,548
	Schizosaccharomyces pombe	no		
Cellular slime mold	*Dictyostelium discoideum*	no		
Higher plants	*Arabidopsis thaliana*	no		
	Oryza sativa	no		
	Zea mays	no		
Nematodes	*Caenorhabditis elegans* (6 chromosomes)	yes	97,000,000	19,000
Insects	*Drosophila melanogaster*	no		
Rodents	*Mus musculus*	no		
Human	*Homo sapiens*	no		

applied (see below) and the project is now close to completion (http://titan.img.cas.cz/rhodo/).

R. capsulatus is a purple, nonsulfur facultative photosynthetic bacterium. It can grow both phototrophically and heterotrophically. In spite of the relatively small genome, consisting of one 3.6-Mb chromosome and one 133-kb plasmid, the bacterium harbors a number of interesting metabolic pathways, such as two independent systems for nitrogen fixation, photosynthesis, CO_2 assimilation, poly(hydroxyalkanoic acid) (PHA) metabolism and degradation of a wide range of organic compounds, among them several pollutants.

The sequencing strategy used in the *R. capsulatus* genome project is based on the relatively elaborate construction of a cosmid encyclopedia [1]. This encyclopedia consists of 186 ordered overlapping cosmids covering the chromosome and six cosmids extending over the plasmid. Individual cosmids are sequenced, and the complete chromosome and plasmid sequences can be obtained without additional cloning of specific DNA fragments. Moreover, this approach enables systematic gene-deletion analysis using the transduction system [2] described below.

In the *R. capsulatus* genome, we found genes for several synthetic pathways with possible practical applications. They include the pathways for cobalamin (vitamin B_{12}) biosynthesis and for poly(hydroxyalkanoic acid) (PHA) metabolism [3]. PHA has recently received attention as a potentially biodegradable plastic [4]. In addition, the genome harbors genetic determination of metabolic pathways for photosynthesis and nitrogen fixation, processes that can now be studied in this organism with a high level of genetic background information. It also harbors genes for biotechnologically important degradative processes, *e.g.*, degradation of organic pollutants such as phenolic and benzoic compounds.

Another photosynthetic bacterium with biotechnological potential is *Rhodopseudomonas palustris* [5]. A strain of *R. palustris* is able to utilize CO_2 and it grows on a number of organic xenobiotics, specifically a broad range of aromatic acids. Selected portions of its genome are being sequenced in our laboratory.

Table 2. *Microbial Genome Projects* (completed or in progress). 1 to 3 stands for possible to strong pathogenicity; biotechnological applications of several strains are indicated.

Actinobacillus actinomycetem-comittans	2	
Aquiflex aeolicus		
Archaeoglobus fulgidus		
Bacillus subtilis	1	Enzyme and antibiotics production
Bartonella henselae	2	
Bordetella pertussis	2	
Borrelia burgdorferi		
Campylobacter jejuni	2	
Candida albicans	2	
Caulobacter crescentus	1	
Chlamydia pneumoniae	2	
Chlamydia trachomatis	2	
Chlorobium tepidum	?	
Clostridium acetobutylicum	1	Production of acetone and butanol
Clostridium difficile	2	
Deinococcus radiodurans	1	
Dehalococcoides ethenogenes	?	
Desulfovibrio vulgaris	1	Metal removal from sewage
Enterococcus faecalis	2	Degradation of xenobiotics
Escherichia coli		
Francisella tularensis	3	
Fusobacterium nucleatum	2	
Halobacterium salinarium	1	
Haemophilus influenzae	2	
Helicobacter pylori	2	
Legionella pneumophila	2	
Listeria monocytogenes	2	
Methanobacterium thermoautotrophicum		
Methanococcus jannashii		
Methanosarcina mazei	1	
Mycobacterium avium	2	
Mycobacterium leprae	3	
Mycobacterium tuberculosis	3	
Mycoplasma genitalium	2	
Mycoplasma mycoides	2	
Mycoplasma pneumoniae	3	
Neisseria gonorrhoeae	2	
Naisseria meningitis	2	
Plasmodium falciparum	2	
Porphyromonas gingivalis	2	
Pseudomonas aeruginosa	2	Degradation of xenobiotics
Pseudomonas putida	1	Degradation of xenobiotics
Pyrobaculum aerophilum	1	
Pyrococcus furiosus	1	
Pyrococcus horicoshii		
Rhodobacter capsulatus	1	
Rickettsia prowazekii	3	
Saccharomyces cerevisiae		Fermentation
Salmonella typhimurium	2	
Schizosaccharomyces pombe		
Shewanella putrefaciens	1	
Staphylococcus aureus	2	
Streptococcus pneumoniae	2	
Streptococcus pyogenes	2	
Streptomyces coelicolor	1	Production of lincomycin
Sulfolobus solfataricus	1	
Synechocystis sp.		
Thermoplasma acidophilum	1	
Thermotoga maritima	1	
Thermus thermophilus	1	
Thiobacillus ferrooxidans	1	Ore bio-leaching
Treponema denticola	2	
Treponema pallidum		
Trypanosoma rhodesiense	3	
Ureaplasma urealyticum	2	
Vibrio cholerae	2	
Xylella fastidiosa	1	
Yersinia pestis	3	
Zymomonas mobilis		Ethanol production

Many prokaryotes contain plasmids in addition to chromosomes. Plasmids are normally circular DNA molocules ranging in size between a few and several hundred kilobases. An important feature of plasmids is their horizontal transfer, which serves adaptations of bacterial populations to changing environmental conditions. Recent challenges are the introduction of antibiotics and man-made chemicals used in agriculture and industry (xenobiotics). Plasmids can integrate and transmit foreign DNA, thereby endowing bacterial populations with genetic variability and flexibility needed for coping with the environmental stresses. As a result, many bacterial strains that grow on organic pollutants harbor the corresponding metabolic pathways on plasmids [6]. Projects are now underway to isolate bacterial strains from polluted soil in the Czech Republic and to screen them for presence of plasmid DNA. These plasmids are sequenced and individual genes or operons are subjected to further biochemical studies.

4. Functional Genomics of Bacteria

Open reading frames (ORFs) are identified in the bacterial genome with the use of specialized software based on the presence of start and stop codons, ribosome binding sites, and codon use. As shown for *Rhodobacter*, about 30% of the ORFs defined so far either have no match in databases or are homologous to genes with unknown function. The major objective of functional genomics is to elucidate functions of these genes. This is difficult because individual genes have to be inactivated and the generated mutants tested for phenotype. Many of these inactivations are lethal, whereas others do not lead to regognizable phenotype changes. It is, therefore, desirable to develop methods for systematic functional analysis of bacterial genomes.

A unique system was developed for *R. capsulatus* that makes it possible to analyze functions of individual genes or gene clusters. The *R. capsulatus* genome contains a defective transducing phage called the Gene Transfer Agent (GTA). Kumar *et al.* [7] suggested that GTA could be used to delete specific *R. capsulatus* DNA regions efficiently and systematically. *Fig. 1* shows how GTA is used to delete whole-cosmid-sized regions of the chromosome. Experiments are under way to modify the system so that individual genes or ORFs can be deleted. This is an important feature of the *R. capsulatus* genome project,

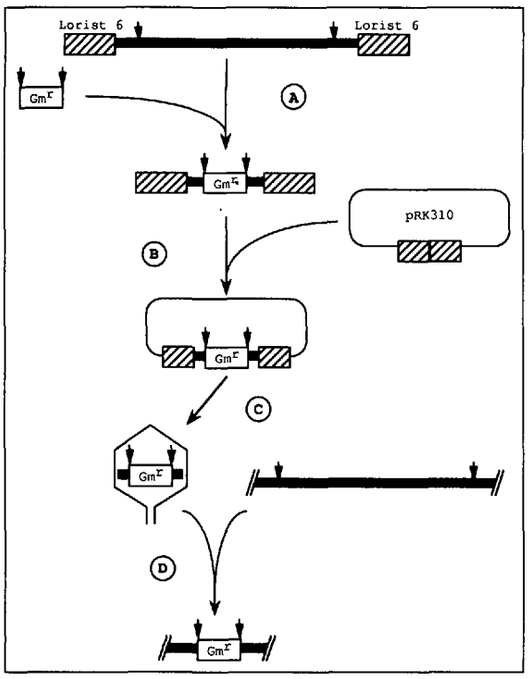

Fig. 1. *GTA-Induced deletions in* R. capsulatus *genome*. Cosmids (Lorist 6) containing *R. cap-sulatus* DNA are digested with a restriction enzyme (A) and the cassette carrying resistance to an antibiotic (Gm) is ligated into the cut cosmid (A). *Escherichia coli* strain with a plasmid (pRK310) carrying a region of the cosmid vector is transformed with the constructs (B). The co-integrants formed are transferred to a GTA-containing strain of *Rhodobacter* by conjugation. Some of the GTA particles produced contain the resistance cassette flanked by *Rhodobacter* DNA (C). These particles are added to wild-type *Rhodobacter*, and transductants with the DNA region replaced by the resistance cassette are selected on antibiotic-containing plates (D).

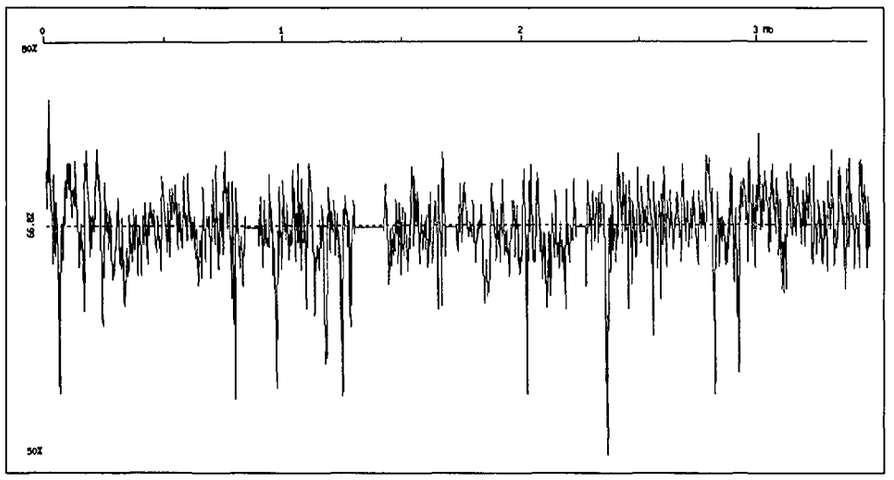

Fig. 2. *GC-Content of the* R. capsulatus *genome*. The AT-richer peaks are likely to belong to phage genomes.

since functions of many unidentified ORFs can be directly investigated.

The ultimate goal of functional genomics of bacteria is the complete description of all metabolic pathways operating in individual bacterial strains.

5. Bacterial Bioinformatics

As more bacterial genomes are sequenced, new information can be generated by simple comparisons of the nucleotide and amino-acid sequences deposited in the databases. Functions can be found with a high degree of reliability for many ORFs by searching for similarities. However, as mentioned above, in most of the genomes the functions of approximately one third of the ORFs remain unknown or uncertain. They are either homologous to unidentified ORFs from other organisms, or have no match in databases at all. The unique ORFs with no match can either encode proteins with unknown functions or they can be subjects of very fast evolutionary change, encoding proteins with already known functions but with primary structures that diverge too much to allow identification by a similarity search. The ORFs encoding proteins with unknown functions found as orthologs in several organisms probably encode new proteins not yet characterized biochemically.

Without even knowing the function of a gene, it is possible to study its position in various genomes, the degree of its similarity with other genes, and its changes in evolution. Various types of restructuring are found in genes or operons when comparisons of related genomes are performed. When comparing the *R. capsulatus* chromosome and plasmid with several bacterial DNAs, we found a number of recombination sites and gene translocations. These studies can shed light on mechanisms of genome plasticity and evolution.

In addition to GTA, the *R. capsulatus* genome contains at least eight DNA regions belonging to other defective phages, transposons, or insertion sequences. The overall GC content of the *R. capsulatus* DNA is 68%. The defective phages present in this genome can be identified because their DNA has lower GC content, and the codon bias is different from that used by the cell (*Fig. 2*). Some of these structures may appear suitable for construction of vectors for gene transfer.

Individual bacterial genomes were analyzed with different criteria set for the gene identification. For instance, many genome projects set the lower limit for a gene at 70 to 100 codons, potentially missing smaller genes. The problem with small genes, *e.g.*, those genes encoding proteins consisting of less than 70 amino acids, is a serious one. We found in the *R. capsulatus* genome genes coding for 68 amino acids (cold-shock protein), 49 and 60 amino acids (subunits of light-harvesting protein) and 57 amino acids (pilus-assembly protein). Thus, we also screen bacterial genomes for the presence of small genes and analyze their function.

This work was supported by grants No. A5052706 of the Academy of Sciences, No. 204/97/0206 of the Grant Agency of the Czech Republic, and No. VS96074 of the Ministry of Education.

Received: September 11, 1999

[1] M. Fonstein, M. R. Haselkorn, *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 2522.
[2] H.C. Yen, N.T. Hu, B.L. Marrs, *J. Mol. Biol.* **1979**, *131*, 157.
[3] Č. Vlček, V. Pačes, N. Maltsev, J. Pačes, R. Haselkorn, M. Fonstein, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 9384.
[4] R.G. Kranz, K.K. Gabbert, T.A. Locke, T. Madigan, *Appl. Environ. Microbiol.* **1997**, *63*, 3003.
[5] a) J. Gibson, M. Dispensa, S. Harwood, *J. Bact.* **1997**, *179*, 634; b) S. Harwood, J. Gibson, *Appl. Environ. Microbiol.* **1988**, *54*, 712.
[6] a) A.M. Boronin, *FEMS Microbiol. Lett.* **1992**, *100*, 461; b) V. Brenner, J.J. Arensdorf, D.D. Focht, *Biodegradation* **1994**, *5*, 359; c) D.E. Crowley, M.V. Brennerová, C. Irwin, V. Brenner, D.D. Focht D.D, *FEMS Microbiol. Ecol.* **1996**, *20*, 79.
[7] V. Kumar, M. Fonstein, R. Haselkorn, *Nature* **1996**, *381*, 653.