# Part I
# Key Organisms

*Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.*
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

# 1
# Genome Projects on Model Organisms

*Alfred Pühler, Doris Jording, Jörn Kalinowski, Detlev Buttgereit, Renate Renkawitz-Pohl, Lothar Altschmied, Antoin Danchin, Agnieszka Sekowska, Horst Feldmann, Hans-Peter Klenk, and Manfred Kröger*

## 1.1 Introduction

Genome research enables the establishment of the complete genetic information of organisms. The first complete genome sequences established were those of prokaryotic and eukaryotic microorganisms, followed by those of plants and animals (see, for example, the TIGR web page at http://www.tigr.org/). The organisms selected for genome research were mostly those which were already important in scientific analysis and thus can be regarded as model organisms. In general, organisms are defined as model organisms when a large amount of scientific knowledge has been accumulated in the past. For this chapter on genome projects of model organisms, several experts in genome research have been asked to give an overview of specific genome projects and to report on the respective organism from their specific point of view. The organisms selected include prokaryotic and eukaryotic microorganisms, and plants and animals.

We have chosen the prokaryotes *Escherichia coli*, *Bacillus subtilis*, and *Archaeoglobus fulgidus* as representative model organisms. The *E. coli* genome project is described by M. Kröger (Giessen, Germany). He gives an historical outline of the intensive research on microbiology and genetics of this organism, which cumulated in the *E. coli* genome project. Many of the technological tools currently available have been developed during the course of the *E. coli* genome project. *E. coli* is without doubt the best-analyzed microorganism of all. The knowledge of the complete sequence of *E. coli* has confirmed its reputation as the leading model organism of Gram$^-$ eubacteria.

A. Danchin and A. Sekowska (Paris, France) report on the genome project of the environmentally and biotechnologically relevant Gram$^+$ eubacterium *B. subtilis*. The contribution focuses on the results and analysis of the sequencing effort and gives several examples of specific and sometimes unexpected findings of this project. Special emphasis is given to genomic data which

*Handbook of Genome Research. Genomics, Proteomics, Metabolomics, Bioinformatics, Ethical and Legal Issues.*
Edited by Christoph W. Sensen

ISBN: 3-527-31348-6

support the understanding of general features such as translation and specific traits relevant for living in its general habitat or its usefulness for industrial processes.

*A. fulgidus* is the subject of the contribution by H.-P. Klenk (Feldafing, Germany). Although this genome project was started before the genetic properties of the organism had been extensively studied, its unique lifestyle as a hyperthermophilic and sulfate-reducing organism makes it a model for a large number of environmentally important microorganisms and species with high biotechnological potential. The structure and results of the genome project are described in the contribution.

The yeast *Saccharomyces cerevisiae* has been selected as a representative eukaryotic microorganism. The yeast project is presented by H. Feldmann (Munich, Germany). *S. cerevisiae* has a long tradition in biotechnology and a long-term research history as a eukaryotic model organism *per se*. It was the first eukaryote to be completely sequenced and has led the way to sequencing other eukaryotic genomes. The wealth of the yeast's sequence information as useful reference for plant, animal, or human sequence comparisons is outlined in the contribution.

Among the plants, the small crucifer *Arabidopsis thaliana* was identified as the classical model plant, because of simple cultivation and short generation time. Its genome was originally considered to be the smallest in the plant kingdom and was therefore selected for the first plant genome project, which is described here by L. Altschmied (Gatersleben, Germany). The sequence of *A. thaliana* helped to identify that part of the genetic information unique to plants. In the meantime, other plant genome sequencing projects were started, many of which focus on specific problems of crop cultivation and nutrition.

The roundworm *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* have been selected as animal models, because of their specific model character for higher animals and also for humans. The genome project of *C. elegans* is summarized by D. Jording (Bielefeld, Germany). The contribution describes how the worm - despite its simple appearance - became an interesting model organism for features such as neuronal growth, apoptosis, or signaling pathways. This genome project has also provided several bioinformatic tools which are widely used for other genome projects.

The genome project concerning the fruitfly *D. melanogaster* is described by D. Buttgereit and R. Renkawitz-Pohl (Marburg, Germany). *D. melanogaster* is currently the best-analyzed multicellular organism and can serve as a model system for features such as the development of limbs, the nervous system, circadian rhythms and even for complex human diseases. The contribution gives examples of the genetic homology and similarities between *Drosophila* and the human, and outlines perspectives for studying features of human diseases using the fly as a model.

## 1.2 Genome Projects of Selected Prokaryotic Model Organisms

### 1.2.1 The Gram$^-$ Enterobacterium *Escherichia coli*

#### 1.2.1.1 The Organism

The development of the most recent field of molecular genetics is directly connected with one of the best described model organisms, the eubacterium *Escherichia coli*. There is no textbook in biochemistry, genetics, or microbiology which does not contain extensive sec-

tions describing numerous basic observations first noted in *E. coli* cells, or the respective bacteriophages, or using *E. coli* enzymes as a tool. Consequently, several monographs solely devoted to *E. coli* have been published. Although it seems impossible to name or count the number of scientists involved in the characterization of *E. coli*, Tab. 1.1 is an attempt to name some of the most deserving people in chronological order.

The scientific career of *E. coli* (Fig. 1.1) started in 1885 when the German pediatrician T. Escherich described isolation of the first strain from the feces of new-born babies. As late as 1958 this discovery was recognized internationally by use of his name

**Table 1.1.** Chronology of the most important primary detection and method applications with *E. coli*.

| | |
|---|---|
| 1886 | "bacterium coli commune" by T. Escherich |
| 1922 | Lysogeny and prophages by d'Herelle |
| 1940 | Growth kinetics for a bacteriophage by M. Delbrück (Nobel prize 1969) |
| 1943 | Statistical interpretation of phage growth curve (game theorie) by S. Luria (Nobel prize 1969) |
| 1947 | Konjugation by E. Tatum and J. Lederberg (Nobel prize 1958)<br>Repair of UV-damage by A. Kelner and R. Dulbecco (Nobel prize for tumor virology) |
| 1954 | DNA as the carrier of genetic information, proven by use of radioisotopes by M. Chase and A. Hershey (Nobel prize 1969) |
| 1959 | Phage immunity as the first example of gene regulation by A. Lwoff (Nobel prize 1965)<br>Transduction of *gal*-genes (first isolated gene) by E. and J. Lederberg<br>Host-controlled modification of phage DNA by G. Bertani and J.J. Weigle |
| 1959 | DNA-polymerase I by A. Kornberg (Nobel prize 1959)<br>Polynucleotide-phosphorylase (RNA synthesis) by M. Grunberg-Manago and S. Ochoa (Nobel prize 1959) |
| 1960 | Semiconservative duplication of DNA by M. Meselson and F. Stahl |
| 1961 | Operon theory and induced fit by F. Jacob and J. Monod (Nobel prize 1965) |
| 1964 | Restriction enzymes by W. Arber (Nobel prize 1978) |
| 1965 | Physical genetic map with 99 genes by A.L. Taylor and M.S. Thoman<br>Strain collection by B. Bachmann |
| 1968 | DNA-ligase by several groups contemporaneously |
| 1976 | DNA-hybrids by P. Lobban and D. Kaiser |
| 1977 | Recombinant DNA from *E. coli* and SV40 by P. Berg (Nobel prize 1980)<br>Patent on genetic engineering by H. Boyer and S. Cohen |
| 1978 | Sequencing techniques using *lac* operator by W. Gilbert and *E. coli* polymerase by F. Sanger (Nobel prize 1980) |
| 1979 | Promoter sequence by H. Schaller<br>Attenuation by C. Yanowsky<br>General ribosome structure by H.G. Wittmann |
| 1979 | Rat insulin expressed in *E. coli* by H. Goodmann<br>Synthetic gene expressed by K. Itakura and H. Boyer |
| 1980 | Site directed mutagenesis by M. Smith (Nobel prize 1993) |
| 1985 | Polymerase chain reaction by K.B. Mullis (Nobel prize 1993) |
| 1988 | Restriction map of the complete genome by Y. Kohara and K. Isono |
| 1990 | Organism-specific sequence data base by M. Kröger |
| 1995 | Total sequence of *Haemophilus influenzae* using an *E. coli* comparison |
| 1999 | Systematic sequence finished by a Japanese consortium under leadership of H. Mori |
| 2000 | Systematic sequence finished by F. Blattner |
| 2000 | Three-dimensional structure of ribosome by four groups contemporaneously |

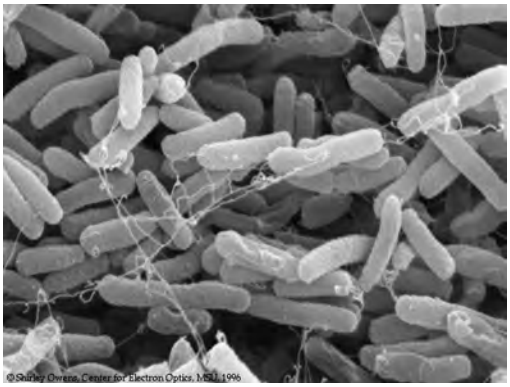


**Fig. 1.1** Scanning electron micrograph (SEM) of *Escherichia coli* cells. (Image courtesy of Shirley Owens, Center for Electron Optics, MSU; found at http://commtechab.msu.edu/sites/dlc-me/zoo/ zah0700.html#top#top)

to classify this group of bacterial strains. In 1921 the very first report on virus formation was published for *E. coli*. Today we call the respective observation "lysis by bacteriophages". In 1935 these bacteriophages became the most powerful tool in defining the characteristics of individual genes. Because of their small size, they were found to be ideal tools for statistical calculations performed by the former theoretical physicist M. Delbrück. His very intensive and successful work has attracted many others to this area of research. In addition, Delbrück's extraordinary capability to catalyze the exchange of ideas and methods yielded the legendary Cold Spring Harbor Phage course. Everybody interested in basic genetics has attended this famous summer course or at least came to the respective annual phage meeting. This course, which was an ideal combination of joy and work, became an ideal means of spreading practical methods. For many decades it was the most important exchange forum for results and ideas, and strains and mutants. Soon, the so called "phage family" was formed, which interacted almost like one big laboratory; for example, results were communicated preferentially by means of preprints. Finally, 15 Nobel prize-winners have their roots in this summer-school (Tab. 1.1).

The substrain *E. coli* K12 was first used by E. Tatum as a prototrophic strain. It was chosen more or less by chance from the strain collection of the Stanford Medical School. Because it was especially easy to cultivate and because it is, as an inhabitant of our gut, a nontoxic organism by definition, the strain became very popular. Because of the vast knowledge already acquired and because it did not form fimbriae, *E. coli* K12 was chosen in 1975 at the famous Asilomar conference on biosafety as the only organism on which early cloning experiments were permitted [1]. No wonder that almost all subsequent basic observations in the life sciences were obtained either with or within *E. coli*. What started as the "phage family", however, dramatically split into hundreds of individual groups working in tough competition. As one of the most important outcomes, sequencing of *E. coli* was performed more than once. Because of the separate efforts, the genome finished only as number seven [2–4]. The amount of knowledge acquired, however, is certainly second to none and the way this knowledge was acquired is interesting, both in the history of sequencing methods and bioinformatics, and because of its influence on national and individual pride.

Work on *E. coli* is not finished with completion of the DNA sequence; data will be continuously acquired to fully characterize the genome in terms of genetic function and protein structures [5]. This is very important, because several toxic *E. coli* strains are known. Thus research on *E. coli* has turned from basic science into applied medical research. Consequently, the human toxic strain O157 has been completely sequenced, again more than once (unpublished).

1.2.1.2

**Characterization of the Genome and Early Sequencing Efforts**

With its history in mind and realizing the impact of the data, it is obvious that an ever growing number of colleagues worldwide worked with or on *E. coli*. Consequently, there was an early need for organization of the data. This led to the first physical genetic map, comprising 99 genes, of any living organism, published in by Taylor and Thoman [6]. This map was improved and was refined for several decades by Bachmann [7] and Berlyn [8]. These researchers still maintain a very useful collection of strains and mutants at Yale University. One thousand and twenty-seven loci had been mapped by 1983 [7]; these were used as the basis of the very first sequence database specific to a single organism [4]. As shown in Fig. 2 of Kröger and Wahl [4], sequencing of *E. coli* started as early as 1967 with one of the first ever characterized tRNA sequences. Immediately after DNA sequencing had been established numerous laboratories started to determine sequences of their personal interest.

1.2.1.3

**Structure of the Genome Project**

In 1987 Isono's group published a very informative and incredibly exact restriction map of the entire genome [9]. With the help of K. Rudd it was possible to locate sequences quite precisely [8, 10]. But only very few saw any advantage in closing the sometimes very small gaps, and so a worldwide joint sequencing approach could not be established. Two groups, one in Kobe, Japan [3] and one in Madison, Wisconsin [2] started systematic sequencing of the genome in parallel, and another laboratory, at Harvard University, used *E. coli* as a target to develop new sequencing technology. Several meetings, organized especially on *E. coli*, did not result in a unified systematic approach, thus many genes have been sequenced two or three times. Although specific databases have been maintained to bring some order into the increasing chaos, even this type of tool has been developed several times in parallel [4, 10]. Whenever a new contiguous sequence was published, approximately 75 % had already previously been submitted to the international databases by other laboratories. The progress of data acquisition followed a classical e-curve, as shown in Fig. 2 of Kröger and Wahl [4]. Thus in 1992 it was possible to predict the completeness of the sequence for 1997 without knowledge of the enormous technical innovations in between [4].

Both the Japanese consortium and the group of F. Blattner started early; some people say they started too early. They subcloned the DNA first and used manual sequencing and older informatic systems. Sequencing was performed semi-automatically, and many students were employed to read and monitor the X-ray films. When the first genome sequence of *Haemophilus influenzae* appeared in 1995 the science foundations wanted to discontinue support of *E. coli* projects, which received their grant support mainly because of the model character of the sequencing techniques developed.

Three facts and truly international protest convinced the juries to continue financial support. First, in contrast with the other completely sequenced organisms, *E. coli* is an autonomously living organism. Second, when the first complete very small genome sequence was released, even the longest contiguous sequence for *E. coli* was already longer. Third, the other laboratories could only finish their sequences because the *E. coli* sequences were already publicly available. Consequently, the two main competing laboratories were allowed to purchase several of the sequencing machines already developed and use the shotgun approach to complete their efforts. Finally, they finished almost at the same time. H. Mori and his colleagues included already published sequences from other laboratories in their sequence data and sent them to the international databases on December 28th, 1996 [3] and F. Blattner reported an entirely new sequence on January 16th, 1997 [2]. They added the last changes and additions as late as October, 1998. Very sadly, at the end *E. coli* had been sequenced almost three times [4]. Nowadays, however, most people forget about all the other sources and refer to the Blattner sequence.

1.2.1.4

**Results from the Genome Project**

When the sequences were finally finished, most of the features of the genome were already known. Consequently, people no longer celebrate the *E. coli* sequence as a major breakthrough. At that time everybody knew the genome was almost completely covered with genes, although fewer than half had been genetically characterized. Tab. 1.2 illustrates this and shows the counting differences. Because of this high density of genes, F. Blattner and coworkers defined "gray holes" whenever they found a noncoding region of more than 2 kb [2]. It was found that the termination of replication is almost exactly opposite to the origin of replication. No special differences have been found for either direction of replication. Approximately 40 formerly described genetic features could not be located or supported by the sequence [4, 8]. On the other hand, there are several examples of multiple functions encoded by the same gene. It was found that the multifunctional genes are mostly involved in gene expression and used as a general control factor. M. Riley determined the number of gene duplications, which is also not unexpectedly low when neglecting the ribosomal operons [10].

Everybody is convinced that the real work is starting only now. Several strain differences might be the cause of the deviations between the different sequences available. Thus the numbers of genes and nucleotides differ slightly (Tab. 1.2). Everybody would like to know the function of each of the open reading frames [5], but nobody has received the grant money to work on this important problem. Seemingly, other model organisms are of more public interest; thus it might well be that research on other organisms will now help our understanding of *E. coli*, in just the same way that *E. coli* provided information enabling understanding of them. In contrast with yeast, it is very hard to produce knock-out mutants. Thus, we might have the same situation in the postgenomic era as we had before the genome was finished. Several laboratories will continue to work with *E. coli*, they will constantly characterize one or the other open reading frame, but there will be no mutual effort [5]. A simple and highly efficient method using PCR products to inactivate chromosomal genes was recently developed [11]. This method has greatly facilitated systematic mutagenesis approaches in *E. coli*.

**Table 1.2** Some statistical features of the *E. coli* genome.

| ***Total size*** | ***4,639,221 bp***[1)] | | |
|---|---|---|---|
| | | ***According to Regulon***[4)] | ***According to Blattner***[5)] |
| Transcription units | Proven | 528 | |
| | Predicted | 2328 | |
| Genes | Total found | 4408 | 4403 |
| | Regulatory | 85 | |
| | Essential | 200 | |
| | Nonessential[2)] | 2363 | 1897 |
| | Unknown[3)] | 1761 | 2376 |
| | tRNA | 84 | 84 |
| | rRNA | 29 | 29 |
| Promoters | Proven | 624 | |
| | Predicted | 4643 | |
| Sites | | 469 | |
| Regulatory interactions | Found | 642 | |
| | Predicted | 275 | |
| Terminators | Found | 96 | |
| RBS | | 98 | |
| Gene products | Regulatory proteins | 85 | |
| | RNA | 115 | 115 |
| | Other peptides | 4190 | 4201 |

**1)** Additional 63 bp compared with the original sequence
**2)** Genes with known or predicted function
**3)** No other data available other than the existence of an open reading frame with a start sequence and more than 100 codons
**4)** Data from http://tula.cifn.unam.mx/Computational_Genomics/regulondb/
**5)** Data from http://www.genome.wisc.edu

1.2.1.5

**Follow-up Research in the Postgenomic Era**

Today it seems more attractive to work with toxic *E. coli* strains, for example O157, than with *E. coli* K12. This strain has recently been completely sequenced; the data are available via the internet. Comparison of toxic and nontoxic strains will certainly help us to understand the toxic mechanisms. It was, on the other hand, found to be correct to use *E. coli* K12 as the most intensively used strain for biological safety regulations [1]. No additional features changed this. This *E. coli* strain is subject to comprehensive transcriptomics and proteomics studies. For global gene expression profiling different systems like an Affymetrix GeneChip and several oligonucleotide sets for the printing of microarrays are available. These tools have already been extensively

used by researchers during recent years. Proteomics studies resulted in a comprehensive reference map for the *E. coli* K-12 proteome (SWISS-2DPAGE, Two-dimensional polyacrylamide gel electrophoresis database, http://www.expasy.org/ch2d). The "Encyclopedia of *Escherichia coli* K-12 Genes and Metabolism" (EcoCyc) (www.ecocyc.org) is a very useful and constantly growing *E. coli* metabolic pathway database for the scientific community [12].

Surprisingly, colleagues from mathematics or informatics have shown the most interest in the bacterial sequences. They have performed all kinds of statistical analysis and tried to discover evolutionary roots. Here another fear of the public is already formulated – people are afraid of attempts to reconstruct the first living cell. So there are at least some attempts to find the minimum set of genes for the most basic needs of a cell. We have to ask again the very old question: Do we really want to "play God"? If so, *E. coli* could indeed serve as an important milestone.

### 1.2.2
### The Gram$^+$ Spore-forming *Bacillus subtilis*

#### 1.2.2.1
#### The Organism

Self-taught ideas have a long life – articles about *Bacillus subtilis* (Fig. 1.2) almost invariably begin with words such as: "*B. subtilis*, a soil bacterium ...", nobody taking the elementary care to check on what type of experimental observation this is based. *Bacillus subtilis*, first identified in 1885, is named *ko so kin* in Japanese and *laseczka sienna* in Polish, or "hay bacterium", and this refers to the real biotope of the organism, the surface of grass or low-lying plants [13]. Interestingly, it required its genome to be sequenced to acquire again its right biotope. Of course, plant leaves fall on the soil surface, and one must naturally find *B. subtilis* there, but its normal niche is the surface of leaves, the phylloplane. Hence, if one wishes to use this bacterium in industrial processes, to engineer its genome, or simply to understand the functions coded by its genes, it is of fundamental importance to understand where it normally thrives, and which environmental conditions control its life-cycle and the corresponding gene expression. Among other important ancillary functions, *B. subtilis* has thus to explore, colonize, and exploit local resources, while at the same time it must maintain itself, dealing with congeners and with other organisms: understanding *B. subtilis* requires understanding the general properties of its normal habitat.

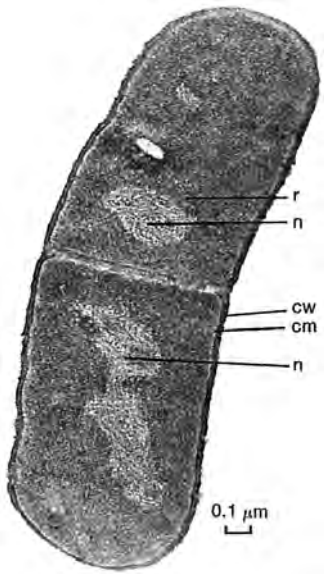


**Fig. 1.2** Electron micrograph of a thin section of *Bacillus subtilis*. The dividing cell is surrounded by a relatively dense wall (CW), enclosing the cell membrane (cm). Within the cell, the nucleoplasm (n) is distinguishable by its fibrillar structure from the cytoplasm, densely filled with 70S ribosomes (r).

1.2.2.2

### A Lesson from Genome Analysis: The *Bacillus subtilis* Biotope

The genome of *B. subtilis* (strain 168), sequenced by a team in European and Japanese laboratories, is 4,214,630 bp long (http://genolist.pasteur.fr/SubtiList/). Of more than 4100 protein-coding genes, 53 % are represented once. One quarter of the genome corresponds to several gene families which have probably been expanded by gene duplication. The largest family contains 77 known and putative ATP-binding cassette (ABC) permeases, indicating that, despite its large metabolism gene number, *B. subtilis* has to extract a variety of compounds from its environment [14]. In general, the permeating substrates are unchanged during permeation. Group-transfer, in which substrates are modified during transport, plays an important role in *B. subtilis*, however. Its genome codes for a variety of phosphoenolpyruvate-dependent systems (PTS) which transport carbohydrates and regulate general metabolism as a function of the nature of the supplied carbon source. A functionally-related catabolite repression control, mediated by a unique system (not cyclic AMP), exists in this organism [15]. Remarkably, apart from the expected presence of glucose-mediated regulation, it seems that carbon sources related to sucrose play a major role, via a very complicated set of highly regulated pathways, indicating that this plant-associated carbon supply is often encountered by the bacteria. In the same way, *B. subtilis* can grow on many of the carbohydrates synthesized by grass-related plants.

In addition to carbon, oxygen, nitrogen, hydrogen, sulfur, and phosphorus are the core atoms of life. Some knowledge about other metabolism in *B. subtilis* has accumulated, but significantly less than in its *E. coli* counterpart. Knowledge of its genome sequence is, however, rapidly changing the situation, making *B. subtilis* a model of similar general use to *E. coli*. A frameshift mutation is present in an essential gene for surfactin synthesis in strain 168 [16], but it has been found that including a small amount of a detergent into plates enabled these bacteria to swarm and glide extremely efficiently (C.-K. Wun and A. Sekowska, unpublished observations). The first lesson of genome text analysis is thus that *B. subtilis* must be tightly associated with the plant kingdom, with grasses in particular [17]. This should be considered in priority when devising growth media for this bacterium, in particular in industrial processes.

Another aspect of the *B. subtilis* life cycle consistent with a plant-associated life is that it can grow over a wide range of different temperatures, up to 54–55 °C – an interesting feature for large-scale industrial processes. This indicates that its biosynthetic machinery comprises control elements and molecular chaperones that enable this versatility. Gene duplication might enable adaptation to high temperature, with isozymes having low- and high-temperature optima. Because the ecological niche of *B. subtilis* is linked to the plant kingdom, it is subjected to rapid alternating drying and wetting. Accordingly, this organism is very resistant to osmotic stress, and can grow well in media containing 1 M NaCl. Also, the high level of oxygen concentration reached during daytime are met with protection systems – *B. subtilis* seems to have as many as six catalase genes, both of the heme-containing type (*katA*, *katB*, and *katX* in spores) and of the manganese-containing type (*ydbD*, PBX phage-associated *yjqC*, and *cotJC* in spores).

The obvious conclusion from these observations is that the normal *B. subtilis* niche is the surface of leaves [18]. This is consistent with the old observation that

*B. subtilis* makes up the major population of the bacteria of rotting hay. Furthermore, consistent with the extreme variety of conditions prevailing on plants, *B. subtilis* is an endospore-forming bacterium, making spores highly resistant to the lethal effects of heat, drying, many chemicals, and radiation.

1.2.2.3

**To Lead or to Lag: First Laws of Genomics**

Analysis of repeated sequences in the *B. subtilis* genome discovered an unexpected feature: strain 168 does not contain insertion sequences. A strict constraint on the spatial distribution of repeats longer than 25 bp was found in the genome, in contrast with the situation in *E. coli*. Correlation of the spatial distribution of repeats and the absence of insertion sequences in the genome suggests that mechanisms aimed at their avoidance and/or elimination have been developed [19]. This observation is particularly relevant for biotechnological processes in which one has multiplied the copy number of genes to improve production. Although there is generally no predictable link between the structure and function of biological objects, the pressure of natural selection has adapted together gene and gene products. Biases in features of predictably unbiased processes is evidence of prior selective pressure. With *B. subtilis* one observes a strong bias in the polarity of transcription with respect to replication: 70 % of the genes are transcribed in the direction of the replication fork movement [14]. Global analysis of oligonucleotides in the genome demonstrated there is a significant bias not only in the base or codon composition of one DNA strand relative to the other, but, quite surprisingly, there is a strong bias at the level of the amino-acid content of the proteins. The proteins coded by the leading strand are valine-rich and those coded by the lagging strand are threonine and isoleucine-rich. This first law of genomics seems to extend to many bacterial genomes [20]. It must result from a strong selection pressure of a yet unknown nature, demonstrating that, contrary to an opinion frequently held, genomes are not, on a global scale, plastic structures. This should be taken into account when expressing foreign proteins in bacteria.

Three principal modes of transfer of genetic material – transformation, conjugation, and transduction – occur naturally in prokaryotes. In *B. subtilis*, transformation is an efficient process (at least in some *B. subtilis* species such as the strain 168) and transduction with the appropriate carrier phages is well understood.

The unique presence in the *B. subtilis* genome of local repeats, suggesting Campbell-like integration of foreign DNA, is consistent with strong involvement of recombination processes in its evolution. Recombination must, furthermore, be involved in mutation correction. In *B. subtilis*, MutS and MutL homologs occur, presumably for the purpose of recognizing mismatched base pairs [21]. No counterpart of MutH activity, which would enable the daughter strand to be distinguished from its parent, has, however, been identified. It is, therefore, not known how the long-patch mismatch repair system corrects mutations in the newly synthesized strand. One can speculate that the nicks caused in the daughter strands by excision of newly misincorporated uracil instead of thymine during replication might provide the appropriate signal. Ongoing fine studies of the distribution of nucleotides in the genome might substantiate this hypothesis.

The recently sequenced genome of the pathogen *Listeria monocytogenes* has many features in common with that of the genome of *B. subtilis* [22]. Preliminary analysis

suggests that the *B. subtilis* genome might be organized around the genes of core metabolic pathways, such as that of sulfur metabolism [23], consistent with a strong correlation between the organization of the genome and the architecture of the cell.

1.2.2.4

**Translation: Codon Usage and the Organization of the Cell's Cytoplasm**

Exploiting the redundancy of the genetic code, coding sequences show evidence of highly variable biases of codon usage. The genes of *B. subtilis* are split into three classes on the basis of their codon usage bias. One class comprises the bulk of the proteins, another is made up of genes expressed at a high level during exponential growth, and a third class, with A + T-rich codons, corresponds to portions of the genome that have been horizontally exchanged [14].

When mRNA threads are emerging from DNA they become engaged by the lattice of ribosomes, and ratchet from one ribosome to the next, like a thread in a wiredrawing machine [24]. In this process, nascent proteins are synthesized on each ribosome, spread throughout the cytoplasm by the linear diffusion of the mRNA molecule from ribosome to ribosome. If the environmental conditions change suddenly, however, the transcription complex must often break up. Truncated mRNA is likely to be a dangerous molecule because, if translated, it would produce a truncated protein. Such protein fragments are often toxic, because they can disrupt the architecture of multisubunit complexes. A process copes with this kind of accident in *B. subtilis*. When a truncated mRNA molecule reaches its end, the ribosome stops translating, and waits. A specialized RNA, tmRNA, that is folded and processed at its 3′ end like a tRNA and charged with alanine, comes in, inserts its alanine at the C-terminus of the nascent polypeptide, then replaces the mRNA within a ribosome, where it is translated as ASFNQNVALAA. This tail is a protein tag that is then used to direct the truncated tagged protein to a proteolytic complex (ClpA, ClpX), where it is degraded [25].

1.2.2.5

**Post-sequencing Functional Genomics: Essential Genes and Expression-profiling Studies**

Sequencing a genome is not a goal *per se*. Apart from trying to understand how genes function together it is most important, especially for industrial processes, to know how they interact. As a first step it was interesting to identify the genes essential for life in rich media. The European–Japanese functional genomics consortium endeavored to inactivate all the *B. subtilis* genes one by one [26]. In 2004, the outcome of this work are still the first and only result in which we can list all the essential genes in bacteria. In this genome counting over 4100 genes, 271 seem to be essential for growth in rich medium under laboratory conditions (i.e. without being challenged by competition with other organisms or by changing environmental conditions). Most of these genes can be placed into a few large and predicable functional categories, for example information processing, cell envelope biosynthesis, shape, division, and energy management. The remaining genes, however, fall into categories not expected to be essential, for example some Embden–Meyerhof–Parnas pathway genes and genes involved in purine biosynthesis. This opens the perspective that these enzymes can have novel and unexpected functions in the cell. Interestingly, among the 26 essential genes that belongs to either "other functions" or "unknown genes" categories, seven belong to or carry the signature for (ATP/)GTP-

binding proteins – several now seem to code for tRNA modifications [27], but some could be in charge of coordination of essential processes listed above, as seems to be true for eukaryotes. A remarkable outcome of this project was the discovery that essential genes are grouped along the leading replication strands in the genome [28]. This feature is general and indicates that genes cannot, at least under strong competitive conditions, be shuffled randomly in the chromosome. This has important consequences for genetic manipulation of organisms of biotechnological interest.

Beside identification of *B. subtilis* essential genes, the European–Japanese project produced an almost complete representative collection of mutants of this bacterium. This collection is freely available to the scientific community, in particular for biotechnology-oriented studies. This strategy is a good example of genome-wide approaches that would have been unthinkable two decades ago. The obvious continuation in this line was the use of transcriptome analysis (identification of all transcripts on DNA arrays under a variety of experimental conditions). Several dozen reports appeared in the literature in those years dealing with data obtained by this global approach. With different technical solutions (from commercially available macroarrays with radioactive labeling through custom-made glass microarrays with fluorescent labeling) they offer an almost exhaustive point of view at the level of transcription answering a given question, assuming particular attention has been devoted to controlling all upstream experimental steps (RNA preparation, cDNA synthesis) and to making use of well-chosen statistical analyses. Many reports have been devoted to the study of heat-shock proteins but not much work was devoted to the equally important cold-shock proteins of the bacteria. The two-component system *desKR* was recently identified; this regulates expression of *des* gene coding for desaturase, which participates in cold adaptation through membrane lipid modification. To discover whether the *desKR* system is exclusively devoted to *des* regulation or constitutes a cold-triggered regulatory system of global relevance, macroarray studies seemed to be the method of choice [29]. A major outcome of this study was, it seemed, that the *desKR* system controls *des* gene expression only. Unexpectedly, this work uncovered many novel partners involved in cold shock response, with almost half of these genes annotated as carrying an unknown function. The categories of genes affected by the cold-shock response were, as expected, the cold shock protein genes that were already known, but also heat-shock protein genes and genes involved in translation machinery, amino acid biosynthesis, nucleoid structure, ABC permeases (for acetoin in particular), purine and pyrimidine biosynthesis and glycolysis, the citric acid cycle, and ATP synthesis. Because these last categories are found in most transcriptome studies, it remains to be seen whether they are indeed specific to cold shock. Among genes of unknown function particularly interesting in biotechnology one can mention the *yplP* gene, which codes for a transcriptional regulator that belongs to the NtrC/NifA family and which, when inactivated, causes a cold-specific late-growth phenotype. However, the exact role of YplP protein remains to be understood.

An essential complement to transcriptome studies is exploration of the bacterial proteome, which gives a detailed look at the behavior of the final players. Because what really counts for a cell is the final level of its mature proteins, and because many post-transcriptional modifications can alter the fate of mRNA translation products into mature proteins, transcriptome analysis alone

provides only an approximation of what is going to really happen in the cell. Studying the proteome expressed under specific conditions can help uncover interesting links between different parts of metabolism. This has been explored under conditions important for biotechnology, for example the salt-stress response and iron metabolism [30]. This recent work aiming at establishing the network of proteins affected by osmotic stress has shown that *B. subtilis* cells growing under high-salinity are subjected to iron limitation, as indicated by the increase of expression of several putative iron-uptake systems (*fhuD*, *fhuB*, *feuA*, *ytiY* and *yfmC*) or iron siderophore bacillibactin synthesis and modification genes (*dhbABCE*). The derepression of the *dhb* operon seems to be more a salt-specific effect than a general osmotic effect, because it is not produced by addition of iso-osmotic non-ionic osmolites (sucrose or maltose) to the growth medium. Some high-salinity growth-defect phenotypes could, furthermore, be reversed by supplementation of the medium with excess iron. This work shows that two distinct factors important in fermentors – iron limitation and high-salinity stress – hitherto regarded as separate growth-limiting factors, are indeed not so separate.

1.2.2.6

**Industrial Processes**

*Bacillus subtilis* is *g*enerally *r*ecognized *a*s *s*afe (GRAS). It is much used industrially both for enzyme production and for food-supply fermentation. Riboflavin is derived from genetically modified *B. subtilis* by use of fermentation techniques. For some time high levels of heterologous gene expression in *B. subtilis* was difficult to achieve. In contrast with Gram-negatives, A + T-rich Gram-positive bacteria have optimized transcription and translation signals; although *B. subtilis* has a counterpart of the *rpsA* gene, this organism lacks the function of the corresponding ribosomal S1 protein which enables recognition of the ribosome-binding site upstream of the translation start codons [31]. Traditional techniques (e.g. random mutagenesis followed by screening; *ad hoc* optimization of poorly defined culture media) are important and will continue to be used in the food industry, but biotechnology must now include genomics to target artificial genes that follow the sequence rules of the genome at a precise position, adapted to the genome structure, and to modify intermediary metabolism while complying with the adapted niche of the organism, as revealed by its genome. As a complement to standard genetic engineering and transgenic technology, knowing the genome text has opened a whole new range of possibilities in food-product development, in particular enabling "humanization" of the content of food products (adaptation to the human metabolism, and even adaptation to sick or healthy conditions). These techniques provide an attractive means of producing healthier food ingredients and products that are currently not available or are very expensive. *B. subtilis* will remain a tool of choice in this respect.

1.2.2.7

**Open Questions**

The complete genome sequence of *B. subtilis* contains information that remains underutilized in the current prediction methods applied to gene functions, most of which are based on similarity searches of individual genes. In particular, it is now clear that the order of the genes in the chromosome is not random, and that some hot spots enable gene insertion without much damage whereas other regions are forbidden. For production of small molecules one must use higher-level information on meta-

bolic pathways to reconstruct a complete functional unit from a set of genes. The reconstruction *in silico* of selected portions of metabolism using the existing biochemical knowledge of similar gene products has been undertaken. Although the core biosynthetic pathways of all twenty amino acids have been completely reconstructed in *B. subtilis*, many satellite or recycling pathways, in particular the synthesis of pyrimidines, have not yet been identified in sulfur and short-carbon-chain acid metabolism. Finally, there remain some 800 genes of completely unknown function in the genome of strain 168, including a few tens of "orphan" genes that have no counterpart in any known genome, and many more in the genome of related species. It remains to be understood whether they play an important role for biotechnological processes.

### 1.2.3 The Archaeon *Archaeoglobus fulgidus*

#### 1.2.3.1 The Organism

*Archaeoglobus fulgidus* is a strictly anaerobic, hyperthermophilic, sulfate-reducing archaeon. It is the first sulfate-reducing organism for which the complete genome sequence has been determined and published [32]. Sulfate-reducing organisms are essential to the biosphere, because biological sulfate reduction is part of the global sulfur cycle. The ability to grow by sulfate-reduction is restricted to a few groups of prokaryotes only. The Archaeoglobales are in two ways unique within this group:

1. they are members of the Archaea and therefore unrelated to all other sulfate reducers, which belong to the Bacteria; and
2. the Archaeoglobales are the only hyperthermophiles within the sulfate-reducers, a feature which enables them to occupy extreme environments, for example hydrothermal fields and sub-surface oil fields.

The production of iron sulfide as an end product of high-temperature sulfate reduction by *Archaeoglobus* species contributes to oil-well "souring", which causes corrosion of iron and steel in submarine oil- and gas-processing systems. *A. fulgidus* is also a model for hyperthermopilic organisms and for the Archaea, because it is only the second hyperthermophile whose genome has been completely deciphered (after *Methanococcus jannaschii*), and it is the third species of Archaea (after *M. jannaschii* and *Methanobacterium thermoautotrophicum*) whose genome has been completely sequenced and published.

*A. fulgidus* DSM4304 (Fig. 1.3) is the type strain of the Archaeoglobales [33]. Its glycoprotein-covered cells are irregular spheres (diameter 2 µm) with four distinct monopolar flagella. It grows not only organoheterotrophically, using a variety of carbon and energy sources, but also lithoautotrophically on hydrogen, thiosulfate, and carbon dioxide. Within the range 60–95 °C it grows best at 83 °C.

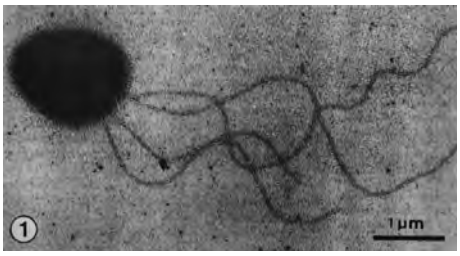


**Fig. 1.3** Electron-micrograph of *A. fulgidus* DSM4303 (strain VC-16), kindly provided by K.O. Stetter, University of Regensburg. The bar in the lower right corner represents 1 µm.

Before genome sequencing very little was known about the genomic organization of *A. fulgidus*. The first estimate of its genome size, obtained by use of pulsed field gel electrophoresis, was published after final assembly of the genome sequences had already been achieved. Because extra-chromosomal elements are absent from *A. fulgidus*, it was determined that the genome consists of only one circular chromosome. Although data about genetic or physical mapping of the genome were unknown before the sequencing project, a small-scale approach to physical mapping was performed late in the project for confirmation of the genome assembly. Sequences of only eleven genes from *A. fulgidus* had been published before the sequencing project started; these covered less than 0.7 % of the genome.

1.2.3.2

**Structure of the Genome Project**

The whole-genome random sequencing procedure was chosen as sequencing strategy for the *A. fulgidus* genome project. This procedure had previously been applied to four microbial genomes sequenced at The Institute for Genomic Research (TIGR): *Haemophilus influenzae*, *Mycoplasma genitalium*, *M. jannaschii*, and *Helicobacter pylori* [34]. Chromosomal DNA for the construction of libraries was prepared from a culture derived from a single cell isolated by means of optical tweezers and provided by K.O. Stetter. Three libraries were used for sequencing – two plasmid libraries (1.42 kbp and 2.94 kbp insert size) for mass sequence production and one large insert λ-library (16.38 kbp insert size) for the genome scaffold. The initial random sequencing phase was performed with these libraries until 6.7-fold sequence coverage was achieved. At this stage the genome was assembled into 152 contigs separated by sequence gaps and five groups of contigs separated by physical gaps. Sequence gaps were closed by a combined approach of editing the ends of sequence traces and by primer walking on plasmid- and λ-clones spanning the gaps. Physical gaps were closed by direct sequencing of PCR-fragments generated by combinatorial PCR reactions. Only 0.33 % of the genome (90 regions) was covered by only one single sequence after the gap-closure phase. These regions were confirmed by additional sequencing reactions to ensure a minimum sequence coverage of two for the whole genome. The final assembly consisted of 29,642 sequencing runs which cover the genome sequence 6.8-fold.

The *A. fulgidus* genome project was financed by the US Department of Energy (DOE) within the Microbial Genome Program. This program financed several of the early microbial genome-sequencing projects performed at a variety of genome centers, for example *M. jannaschii* (TIGR), *M. thermoautotrophicum* (Genome Therapeutics), *Aquifex aeolicus* (Recombinant BioCatalysis, now DIVERSA), *Pyrobaculum aerophilum* (California Institute of Technology), *Pyrococcus furiosus* (University of Utah), and *Deinococcus radiodurans* (Uniformed Services University of the Health Sciences). Like the *M. jannaschii* project which was started one year earlier, the *A. fulgidus* genome was sequenced and analyzed in a collaboration between researchers at TIGR and Carl R. Woese and Gary J. Olsen at the Department of Microbiology at the University of Illinois, Champaign-Urbana. The plasmid libraries were constructed in Urbana, whereas the λ-library was constructed at TIGR. Sequencing and assembly was performed at TIGR using automated ABI sequencers and a TIGR assembler, respectively. Confirmation of the assembly by mapping with large-size restriction fragments was performed in Urbana. Open reading frame (ORF) prediction

and identification of functions, and the data mining and interpretation of the genome content was performed jointly by both teams.

Coding regions in the final genome sequence were identified with a combination of two sets of ORF generated by programs developed by members of the two teams – GeneSmith, by H.O. Smith at TIGR and CRITICA, by G.J. Olsen and J.H. Badger in Urbana. The two sets of ORF identified by GeneSmith and CRITICA were merged into one consensus set containing all members of both initial sets. The amino acid sequences derived from the consensus set were compared with a non-redundant protein database using BLASTX. ORFs shorter than 30 codons were carefully inspected for database hits and eliminated when there was no significant database match. The results of the database comparisons were first inspected and categorized by TIGR's microbial annotation team. This initial annotation database was then further analyzed and refined by a team of experts for all major biological role categories.

The sequencing strategy chosen for the *A. fulgidus* genome project has some advantages compared with alternative strategies applied in genome research:

1. Given the relatively large set of automated sequencers available at TIGR, the whole-genome random sequencing procedure is much faster than any strategy that includes a mapping step before the sequencing phase;
2. Within the DOE Microbial Genome Program the TIGR strategy and the sequencing technology used for the *M. jannaschii* and *A. fulgidus* genome projects proved to be clearly superior in competition with projects based on multiplex sequencing (*M. thermoautotrophicum* and *P. furiosus*), by finishing two genomes in less time than the competing laboratories needed for one genome each; and
3. The interactive annotation with a team of experts for the organism and for each biological category ensured a more sophisticated final annotation than any automated system could achieve at that time.

1.2.3.3

**Results from the Genome Project**

Although the initial characterization of the genome revealed all its basic features, annotation of biological functions for the ORF will continue to be updated for new functions identified either in *A. fulgidus* or for homologous genes characterized in other organisms. The size of the *A. fulgidus* genome was determined to be 2,178,400 bp, with an average G + C content of 48.5 %. Three regions with low G + C content (<39 %) were identified, two of which encode enzymes for lipopolysaccharide biosynthesis. The two regions with the highest G + C content (>53 %) contain the ribosomal RNA and proteins involved in heme biosynthesis. With the bioinformatics tools available when genome characterization was complete, no origin of replication could be identified. The genome contains only one set of genes for ribosomal RNA. Other RNA encoded in *A. fulgidus* are 46 species of tRNA, five of them with introns 15–62 bp long, no significant tRNA clusters, 7S RNA and RNase P. All together 0.4 % of the genome is covered by genes for stable RNA. Three regions with short (<49 bp) non-coding repeats (42–60 copies) were identified. All three repeated sequences are similar to short repeated sequences found in *M. jannaschii* [35]. Nine classes of long, coding repeats (>95 % sequence identity) were identified within the genome, three of them might represent IS elements, and three other repeats encode conserved hypothetical proteins found previously in other genomes. The consensus set of ORF contains 2436 members with an average length of

822 bp, similar to *M. jannaschii* (856 bp), but shorter than in most bacterial genomes (average 949 bp). With 1.1 ORF per kb, the gene density seems to be slightly higher than in other microbial genomes, although the fraction of the genome covered by protein coding genes (92.2 %) is comparable with that for other genomes. The elevated number of ORF per kbp might be artificial, because of a lack of stop codons in high G + C organisms. Predicted start codons are 76 % ATG, 22 % GTG, and 2 % TTG. No inteins were identified in the genome. The isoelectric point of the predicted proteins in *A. fulgidus* is rather low (median p*I* is 6.3); for other prokaryotes distributions peak between 5.5 and 10.5. Putative functions could be assigned to about half of the predicted ORF (47 %) by significant matches in database searches. One quarter (26.7 %) of all ORF are homologous to ORF previously identified in other genomes ("conserved hypotheticals"), whereas the remaining quarter (26.2 %) of the ORF in *A. fulgidus* seem to be unique, without any significant database match. *A. fulgidus* contains an unusually large number of paralogous gene families: 242 families with 719 members (30 % of all ORF). This might explain why the genome is larger than most other archaeal genomes (average approximately 1.7 Mbp). Interestingly, one third of the identified families (85 out of 242) have no single member for which a biological function could be predicted. The largest families contain genes assigned to "energy metabolism", "transporters", and "fatty acid metabolism".

The genome of *A. fulgidus* is neither the first archaeal genome to be sequenced completely nor is it the first genome of a hyperthermophilic organism. The novelties for both features had already been reported together with the genome of *M. jannaschii* [35]. *A. fulgidus* is, however, the first sulfate-reducing organism whose genome was completely deciphered. The next genome of a sulfate reducer followed almost seven years later with that of *Desulfotalea psychrophila*, an organism whose optimum growth temperature is 75 ° lower than that of *A. fulgidus* [36]. Model findings in respect of sulfur and sulfate metabolism were not expected from the genome, because sulfate metabolism had already been heavily studied in *A. fulgidus* before the genome project. The genes for most enzymes involved in sulfate reduction were already published, and new information from the genome confirmed only that the sulfur oxide reduction systems in Archaea and Bacteria were very similar. The single most exciting finding in the genome of *A. fulgidus* was identification of multiple genes for acetyl-CoA synthase and the presence of 57 β-oxidation enzymes. It has been reported that the organism is incapable of growth on acetate [37], and no system for β-oxidation has previously been described in the Archaea. It appears now that *A. fulgidus* can gain energy by degradation of a variety of hydrocarbons and organic acids, because genes for a least five types of ferredoxin-dependent oxidoreductases and at least one lipase were also identified. Interestingly, at about the same time as the unexpected genes for metabolizing enzymes were identified, it was also reported that a close relative of *A. fulgidus* is able to grow on olive oil (K. O. Stetter, personal communication), a feature that would require the presence of the genes just identified in *A. fulgidus*. On the other hand, not all genes necessary for the pathways described in the organism could be identified. Glucose has been described as a carbon source for *A. fulgidus* [38], but neither an uptake-transporter nor a catabolic pathway for glucose could be identified in the genome. There is still a chance that the required genes are hidden in the pool of functionally

uncharacterized ORFs. Other interesting findings in respect of the biology of *A. fulgidus* concern sensory functions and regulation of gene expression. *A. fulgidus* seems to have complex sensory and regulatory networks – a major difference from results reported for *M. jannaschii* – consistent with its extensive energy-producing metabolism and versatile system for carbon utilization. These networks contain more than 55 proteins with presumed regulatory functions and several iron-dependent repressor proteins. At least 15 signal-transducing histidine kinases were identified, but only nine response regulators.

1.2.3.4

**Follow-up Research**

Almost 30 papers about *A. fulgidus* were published between the initial description of the organism in 1987 and the genome sequence ten years later. In the six years since the genome was finished *A. fulgidus* was mentioned in the title or abstract of more than 165 research articles. Although functional genomics is now a hot topic at many scientific meetings and discussions, *A. fulgidus* seems to be no prime candidate for such studies. So far not a single publication has dealt with proteomics, transcriptome, or serial mutagenesis in this organism. The recently published papers that refer to the *A. fulgidus* genome sequence fall into three almost equally represented categories: comparative genomics, structure of *A. fulgidus* proteins, and characterization of expressed enzymes. One of the most interesting follow up stories is probably that on the flap endonucleases. In October 1998 Hosfield et al. described newly discovered archaeal flap endonucleases (FEN) from *A. fulgidus, M. jannaschii* and *P. furiosus* with a structure-specific mechanism for DNA substrate binding and catalysis resembling human flap endonuclease [39]. In Spring 1999 Lyamichev et al. showed how FEN could be used for polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes [40]. In May 2000, Cooksey et al. described an invader assay based on *A. fulgidus* FEN that enables linear signal amplification for identification of mutatons [41]. This procedure could eventually become important as a non-PCR based procedure for SNP detection. The identification of 86 candidates for small non-messenger RNA by Tang et al. [42] and identification of the unusual organization of the putative replication origin region by Maisnier-Patin et al. [43] also mark noteworthy progress in our knowledge about the model organism *A. fulgidus*.

## 1.3 Genome Projects of Selected Eukaryotic Model Organisms

### 1.3.1 The Budding Yeast *Saccharomyces cerevisiae*

1.3.1.1

**Yeast as a Model Organism**

The budding yeast, *Saccharomyces cerevisiae* (Fig. 1.4), can be regarded as one of the most important fungal organisms used in biotechnological processes. It owes its name to its ability to ferment saccharose, and has served mankind for several thousand years in the making of bread and alcoholic beverages. The introduction of yeast as an experimental system dates back to the 1930s and has since attracted increasing attention. Unlike more complex eukaryotes, yeast cells can be grown on defined media giving the investigator complete control over environmental conditions. The elegance of yeast genetics and the ease of manipulation of yeast have substantially con-

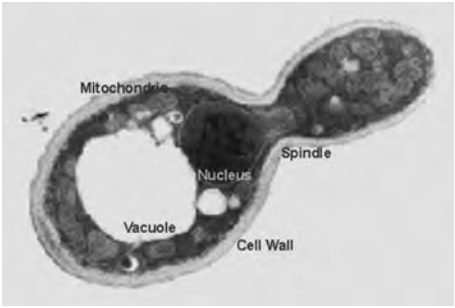


**Fig. 1.4** Micrograph of the budding yeast *Saccharomyces cerevisiae* during spore formation. The cell wall, nucleus, vacuole, mitochondria, and spindle are indicated.

tributed to the explosive growth in yeast molecular biology. This success is also a consequence of the notion that the extent to which basic biological processes have been conserved throughout eukaryotic life is remarkable and makes yeast a unique unicellular model organism in which cell architecture and fundamental cellular mechanisms can be successfully investigated. No wonder, then, that yeast had again reached the forefront in experimental molecular biology by being the first eukaryotic organism of which the entire genome sequence became available [44, 45]. The wealth of sequence information obtained in the yeast genome project was found to be extremely useful as a reference against which sequences of human, animal, or plant genes could be compared.

The first genetic map of *S. cerevisiae* was published by Lindegren in 1949 [46]; many revisions and refinements have appeared since. At the outset of the sequencing project approximately 1200 genes had been mapped and detailed biochemical knowledge about a similar number of genes encoding either RNA or protein products had accumulated [47]. The existence of 16 chromosomes ranging in size between 250 and ~2500 kb was firmly established when it became feasible to separate all chromosomes by pulsed-field gel electrophoresis (PFGE). This also provided definition of "electrophoretic karyotypes" of strains by sizing chromosomes [48]. Not only do laboratory strains have different karyotypes, because of chromosome length polymorphisms and chromosomal rearrangements, but so do industrial strains. A defined laboratory strain (αS288C) was therefore chosen for the yeast sequencing project.

1.3.1.2

**The Yeast Genome Sequencing Project**

The yeast sequencing project was initiated in 1989 within the framework of EU biotechnology programs. It was based on a network approach into which 35 European laboratories initially became involved, and chromosome III – the first eukaryotic chromosome ever to be sequenced – was completed in 1992. In the following years and engaging many more laboratories, sequencing of further complete chromosomes was tackled by the European network. Soon after its beginning, laboratories in other parts of the world joined the project to sequence other chromosomes or parts thereof, ending up in a coordinated international enterprise. Finally, more than 600 scientists in Europe, North America, and Japan became involved in this effort. The sequence of the entire yeast genome was completed in early 1996 and released to public databases in April 1996.

Although the sequencing of chromosome III started from a collection of overlapping plasmid or phage lambda clones, it was expected that cosmid libraries would subsequently have to be used to aid large-scale sequencing [49]. Assuming an average insert length of 35–40 kb, a cosmid library containing 4600 random clones would represent the yeast genome at approximately twelve times the genome equivalent. The

advantages of cloning DNA segments in cosmids were at hand – clones were found to be stable for many years and the small number of clones was advantageous in setting up ordered yeast cosmid libraries or sorting out and mapping chromosome-specific sublibraries. High-resolution physical maps of the chromosomes to be sequenced were constructed by application of classical mapping methods (fingerprints, cross-hybridzation) or by novel methods developed for this program, for example site-specific chromosome fragmentation [50] or a high-resolution cross-hybridization matrix, to facilitate sequencing and assembly of the sequences.

In the European network chromosome-specific clones were distributed to the collaborating laboratories according to a scheme worked out by the DNA coordinators. Each contracting laboratory was free to apply sequencing strategies and techniques of its own provided that the sequences were entirely determined on both strands and unambiguous readings were obtained. Two principle approaches were used to prepare subclones for sequencing:

1. generation of sublibraries by use of a series of appropriate restriction enzymes or from nested deletions of appropriate subfragments made by exonuclease III; and
2. generation of shotgun libraries from whole cosmids or subcloned fragments by random shearing of the DNA.

Sequencing by the Sanger technique was either performed manually, labeling with [$^{35}$S]dATP being the preferred method of monitoring, or by use of automated devices (on-line detection with fluorescence labeling or direct blotting electrophoresis system) following a variety of established procedures. Similar procedures were applied to the sequencing of the chromosomes contributed by the Sanger laboratory and by laboratories in the USA, Canada, and Japan. The American laboratories largely relied on machine-based sequencing.

Because of their repetitive substructure and the lack of appropriate restriction sites, the yeast chromosome telomeres were a particular problem. Conventional cloning procedures were successful for a few exceptions only. Telomeres were usually physically mapped relative to the terminal-most cosmid inserts using the chromosome fragmentation procedure [50]. The sequences were then determined from specific plasmid clones obtained by "telomere trap cloning", an elegant strategy developed by Louis and Borts [51].

Within the European network, all original sequences were submitted by the collaborating laboratories to the Martinsried Institute of Protein Sequences (MIPS) who acted as an informatics center. They were kept in a data library, assembled into progressively growing contigs, and, in collaboration with the DNA coordinators, the final chromosome sequences were derived. Quality controls were performed by anonymous resequencing of selected regions and suspected or difficult zones (total of 15–20 % per chromosome). Similar procedures were used for sequence assembly and quality control in the other laboratories. During recent years further quality controls were carried and resulted in a nearly absolute accuracy of the total sequence.

The sequences of the chromosomes were subjected to analysis by computer algorithms, identifying ORF and other genetic entities, and monitoring compositional characteristics of the chromosomes (base composition, nucleotide pattern frequencies, GC profiles, ORF distribution profiles, etc.). Because the intron splice site/branchpoint pairs in yeast are highly conserved, they could be detected by using defined search patterns. It was finally found that only 4 % of the yeast genes contain (mostly

short) introns. Centromere and telomere regions, and tRNA genes, sRNA genes, or the retrotransposons, were sought by comparison with previously characterized datasets or appropriate search programs. All putative proteins were annotated by using previously established yeast data and evaluating searches for similarity to entries in the databases or protein signatures detected by using the PROSITE dictionary.

1.3.1.3

**Life with Some 6000 Genes**

With its 12.8 Mb, the yeast genome is approximately a factor of 250 smaller than the human genome. The complete genome sequence now defines some 6000 ORFs which are likely to encode specific proteins in the yeast cell. A protein-encoding gene is found every 2 kb in the yeast genome, with nearly 70 % of the total sequence consisting of ORF. This leaves only limited space for the intergenic regions which can be thought to harbor the major regulatory elements involved in chromosome maintenance, DNA replication, and transcription. The genes are usually rather evenly distributed among the two strands of the single chromosomes, although arrays longer than eight genes that are transcriptionally oriented in the same direction can be found eventually. With a few exceptions, transcribed genes on complementing strands are not overlapping, and no "genes-in-genes" are observed. Although the intergenic regions between two consecutive ORF are sometimes extremely short, they are normally maintained as separate units and not coupled for transcription. In "head-to-head" gene arrangements the intervals between the divergently transcribed genes might be interpreted to mean that their expression is regulated in a concerted fashion involving the common promoter region. This, however, seems not to be true for most genes and seems to be a principle reserved for a few cases. The sizes of the ORFs vary between 100 to more than 4000 codons; less than 1 % is estimated to be below 100 codons. In addition, the yeast genome contains some 120 ribosomal RNA genes in a large tandem array on chromosome XII, 40 genes encoding small nuclear RNA (sRNA) and 275 tRNA genes (belonging to 43 families) which are scattered throughout the genome. Overall, the yeast genome is remarkably poor in repeated sequences, except the transposable elements (Tys) which account for approximately 2 % of the genome, and, because of to their genetic plasticity, are the major source of polymorphisms between different strains. Finally, the sequences of non-chromosomal elements, for example the 6 kb of the 2μ plasmid DNA, the killer plasmids present in some strains, and the yeast mitochondrial genome (ca.75 kb), must be considered.

On completion of the yeast genome sequence it became possible for the first time to define the proteome of a eukaryotic cell. Detailed information was laid down in inventory databases and most of the proteins could be classified according to function. It was seen that almost 40 % of the proteome consisted of membrane proteins, and that an estimated 8 to 10 % of nuclear genes encode mitochondrial functions. It came as an initial surprise that no function could be attributed to approximately 40 % of the yeast genes. However, even with the exponential growth of entries in protein databases and the refinement of *in silico* analyses, this figure could not be reduced substantially. The same was observed for all other genomes that have since been sequenced. As an explanation, we have to envisage that a considerable portion of every genome is reserved for species- or genus-specific functions.

An interesting observation made for the first time was the occurrence of regional

variations of base composition with similar amplitudes along the chromosomes. Analysis of chromosomes III and XI revealed almost regular periodicity of the GC content, with a succession of GC rich and GC poor segments of ~50 kb each. Another interesting observation was that the compositional periodicity correlated with local gene density. Profiles obtained from similar analyses of chromosomes II and VIII again showed these phenomena, albeit with less pronounced regularity. Similar compositional variation has been found along the arms of other chromosomes, with pericentromeric and subtelomeric regions being AT-rich, though spacing between GC-rich peaks is not always regular. Usually, however, there is a broad correlation between high GC content and high gene density.

Comparison of all yeast sequences revealed there is substantial internal genetic redundancy in the yeast genome, which at the protein level is approximately 40 %. Whereas an estimate of sequence similarity (at both the nucleotide and the amino acid level) is highly predictive, it is still difficult to correlate these values with functional redundancy. Interestingly, the same phenomenon has since been observed in all other genomes sequenced. Gene duplications in yeast are of different type. In many instances, the duplicated sequences are confined to nearly the entire coding region of these genes and do not extend into the intergenic regions. Thus, the corresponding gene products share high similarity in terms of amino acid sequence, and sometimes are even identical, and, therefore, might be functionally redundant. As suggested by sequence differences within the promoter regions or demonstrated experimentally, however, expression varies. It is possible one gene copy is highly expressed whereas another is poorly expressed. Turning on or off expression of a particular copy within a gene family might depend on the differentiated status of the cell (for example mating type, sporulation, etc.). Biochemical studies also revealed that in particular instances "redundant" proteins can substitute each other, thus accounting for the observation that large amounts of single-gene disruption in yeast do not impair growth or cause "abnormal" phenotypes. This does not imply, however, that these "redundant" genes are *a priori* dispensable. Rather they might have arisen from the need to help adapt yeast cells to particular environmental conditions.

Subtelomeric regions in yeast are rich in duplicated genes which are of functional importance to carbohydrate metabolism or cell-wall integrity, but there is also much variety of (single) genes internal to chromosomes that seem to have arisen from duplications. An even more surprising phenomenon became apparent when the sequences of complete chromosomes were compared with each other. This revealed 55 large chromosome segments (up to 170 kb) in which homologous genes are arranged in the same order, with the same relative transcriptional orientations, on two or more chromosomes [52]. The genome has continued to evolve since this ancient duplication occurred – genes have been inserted or deleted, and Ty elements and introns have been lost and gained between two sets of sequences. If optimized for maximum coverage, up to 40 % of the yeast genome is found to be duplicated in clusters, not including Ty elements and subtelomeric regions. No observed clusters overlap and intra- and interchromosomal cluster duplications have similar probabilities.

The availability of the complete yeast genome sequence not only provided further insight into genome organization and evolution in yeast but also offered a reference to search for orthologs in other organisms. Of particular interest were those genes that

are homologs of genes that perform differentiated functions in multicellular organisms or that might be of relevance to malignancy. Comparing the catalog of human sequences available in the databases with the yeast ORF reveals that more than 30 % of yeast genes have homologs among human genes of known function. Approximately 100 yeast genes are significantly similar to human disease genes [53], and some of the latter could even be predicted from comparison with the yeast genes.

1.3.1.4

**The Yeast Postgenome Era**

It was evident to anyone engaged in the project that determination of the entire sequence of the yeast genome should only be regarded as a prerequisite for functional studies of the many novel genes to be detected. Thus, a European functional analysis network (EUROFAN) was initiated in 1995 and similar activities were started in the international collaborating laboratories in 1996. The general goal was to systematically investigate yeast genes of unknown function by use of the approaches:

1. improved data analysis by computer (*in silico* analysis);
2. systematic gene disruptions and gene overexpression;
3. analysis of phenotypes under different growth conditions, for example temperature, pH, nutrients, stress;
4. systematic transcription analysis by conventional methods; gene expression under different conditions;
5. *in situ* cellular localization and movement of proteins by the use of tagged proteins (GFP-fusions);
6. analysis of gene expression under different conditions by 2D gel-electrophoresis of proteins; and
7. complementation tests with genes from other organisms.

In this context, a most compelling approach is the genome-wide analysis of gene expression profiles by chip technology. High-density micro-arrays of all yeast ORF were the first to be successfully used in studying various aspects of a transcriptome [54, 55]. Comprehensive and regularly updated information can be found in the yeast Genome Database (http://www.yeastgenome.org).

Now the entire sequence of a laboratory strain of *S. cerevisiae* is available, the complete sequences of other yeasts of industrial or medical importance are within our reach. Such knowledge would considerably accelerate the development of productive strains needed in other areas (e.g. *Kluyveromyces*, *Yarrowia*) or the search for novel anti-fungal drugs. It might even be unnecessary to finish the entire genome if a yeast or fungal genome has substantial synteny with that of *S. cerevisiae*. A special program devoted to this problem, analysis of hemiascomycetes yeast genomes by tag-sequence studies for the approach of speciation mechanisms and preparation of tools for functional genomics, has recently been finalized by a French consortium [56].

In all, the yeast genome project has demonstrated that an enterprise like this can successfully be conducted in "small steps" and by teamwork. Clearly, the wealth of fresh and biologically relevant information collected from yeast sequences and from functional analyses has had an impact on other large-scale sequencing projects.

1.3.2

**The Plant *Arabidopsis thaliana***

1.3.2.1

**The Organism**

*Arabidopsis thaliana* (Fig. 1.5) is a small cruciferous plant of the mustard family first described by the German physician Johan-


**Fig. 1.5** The model plant *Arabidopsis thaliana*. (a) Adult plant, approx. height 20 cm [59]; (b) flower, approx. height 4 mm [60]; (c) chromosome plate showing the five chromosomes of *Arabidopsis* [57].

nes Thal in 1577 in his Flora Book of the Harz mountains, Germany, and later named after him. In 1907 *A. thaliana* was recognized to be a versatile tool for genetic studies by Laibach [57] when he was a student with Strasburger in Bonn, Germany. More than 30 years later in 1943 – then Professor of Botany in Frankfurt – he published an influential paper [58] clearly describing the favorable features making this plant a true model organism: (1) short generation time of only two months, (2) high seed yield, (3) small size, (4) simple cultivation, (5) self fertilization, but (6) easy crossing yielding fully fertile hybrids, (7) only five chromosomes in the haploid genome, and (8) the possibility of isolating spontaneous and induced mutants. An attempt by Rédei in the 1960s to convince funding agencies to develop *Arabidosis* as a plant model system was unsuccessful, mainly because geneticists at that time had no access to genes at the molecular level and therefore no reason to work with a plant irrelevant to agriculture.

With the development of molecular biology two further properties of the *A. thaliana* genome made this little weed the superior choice as an experimental system. Laibach had already noted in 1907 that *A. thaliana* contained only one third of the chromatin of related *Brassica* species [57]. Much later it became clear that this weed has: (1) the smallest genome of any higher plant [61] and (2) a small amount of repetitive DNA. Within the plant kingdom characterized by its large variation of genome sizes (see, for example, http://www.rbgkew.org.uk/cval/database1.html), mainly because of different amounts of repetitive DNA, these features support efficient map-based cloning of genes for a detailed elucidation of their function at the molecular level. A first set of tools for that purpose became available during the 1980s with a comprehensive genetic map containing 76 phenotypic markers obtained by mutagenesis, RFLP maps, *Agro-*

*bacterium*-mediated transformation, and cosmid and yeast artificial chromosome (YAC) libraries covering the genome severalfold with only a few thousand clones.

It was soon realized that projects involving resources shared by many laboratories would profit from centralized collection and distribution of stocks and related information. "The Multinational Coordinated *Arabidopsis thaliana* Genome Research Project" was launched in 1990 and, as a result, two stock centers in Nottingham, UK, and Ohio, USA, and the *Arabidopsis* database at Boston, USA [62] were created in 1991. With regard to seed stocks these centers succeeded an effort already started by Laibach in 1951 [59] and continued by Röbbelen and Kranz. The new, additional collections of clones and clone libraries provided the basis for the genome sequencing project later on. With increased use of the internet at that time the database soon became a central tool for data storage and distribution and has ever since served the community as a central one-stop shopping point for information, despite its move to Stanford and its restructuring to become "The *Arabidopsis* Information Resource" (TAIR; http://www.arabidopsis.org).

In subsequent years many research tools were improved and new ones were added – mutant lines based on insertions of T-DNA and transposable elements were created in large numbers, random cDNA clones were sequenced partially, maps became available for different types of molecular marker such as RAPD, CAPS, microsatellites, AFLP, and SNP which were integrated with each other, recombinant inbred lines were established to facilitate the mapping process, a YAC library with large inserts and BAC and P1 libraries were constructed, physical maps based on cosmids, YAC, and BAC were built, and tools for their display were developed.

1.3.2.2

**Structure of the Genome Project**

In August 1996 the stage was prepared for large-scale genome sequencing. At a meeting in Washington DC representatives of six research consortia from North America, Europe, and Japan launched the "*Arabidopsis* Genome Initiative". They set the goal of sequencing the complete genome by the year 2004 and agreed on the strategy, the distribution of tasks, and guidelines for the creation and publication of sequence data. The genome of the ecotype Columbia was chosen for sequencing, because all large insert libraries had been prepared from this line and it was one of the most prominent ecotypes for all kinds of experiment worldwide besides Landsberg erecta (Ler). Because Ler is actually a mutant isolated from an inhomogeneous sample of the Landsberg ecotype after X-ray irradiation [63], it was not suitable to serve as a model genome. The sequencing strategy rested on BAC and P1 clones from which DNA can be isolated more efficiently than from YAC and which, on average, contain larger inserts than cosmids. This strategy was chosen even though most attempts to create physical maps had been based on cosmid and YAC clones at that time and that initial sequencing efforts had employed mostly cosmids. BAC and P1 clones for sequencing were chosen *via* hybridization to YAC and molecular markers of known and well separated map positions. Later, information from BAC end sequences and fingerprint and hybridization data, created while genome sequencing was already in progress, were used to minimize redundant sequencing caused by clone overlap. This multinational effort has been very fruitful and led to complete sequences for two of the five *Arabidopsis* chromosomes, chromosomes 2 [64] and 4 [65], except for their rDNA repeats and the heterochromatic regions

around their centromers. The genomic sequence was completed by the end of the year 2000, more than three years ahead of the original timetable. Sequences of the mitochondrial [66] and the plastid genome [67] have also been determined, so complete genetic information was available for *Arabidopsis*.

1.3.2.3

**Results from the Genome Project**

The sequenced chromosomes have yielded no surprises with regard to their structural organization. With the exception of one sequenced marker, more than one hundred have been observed in the expected order. Repetitive elements and repeats of transposable elements are concentrated in the heterochromatic regions around the centromers where gene density and recombination frequency are below the average (22 genes/100 kbp, 1 cM/50 - 250 kbp). With a few minor exceptions these average values do not vary substantially in other regions of the chromosomes, which is in sharp contrast with larger plant genomes [68–70]. In addition, genomes such as that of maize do contain repetitive elements and transposons interspersed with genes [71], so *Arabidopsis* is certainly not a model for the structure of large plant genomes.

From the sequences available it has been calculated that the 120 Mbp gene containing part of the nuclear genome of *Arabidopsis* contains approximately 25,000 genes (chr. 2: 4037, chr. 4: 3744 annotated genes; [72]), whereas the mitochondrial and plastid genomes carry only 57 and 78 genes on 366,924 and 154,478 bp of DNA, respectively. Most of the organellar proteins must therefore be encoded in the nucleus and are targeted to their final destinations *via* N-terminal transit peptides. It has recently been estimated that 10 or 14 % of the nuclear genes encode proteins located in mitochondria or plastids, respectively [73]. Only for a fraction of the predicted plastid proteins could homologous cyanobacterial proteins be identified [64, 65]; even so, lateral gene transfer from the endosymbiotic organelle to the nucleus has been assumed to be the main source of organellar proteins. These data indicate that either the large evolutionary distance between plants and cyanobacteria prevents the recognition of orthologs and/or many proteins from other sources in the eukaryotic cell have acquired plastid transit peptides, as suggested earlier on the basis of Calvin cycle enzymes [74]. A substantial number of proteins without predicted transit peptides but with higher homology to proteins from cyanobacteria than to any other organism have also been recognized [64, 65], indicating that plastids, at least, have contributed a significant part of the protein complement of other compartments. These data clearly show that plant cells have become highly integrated genetic systems during evolution, assembled from the genetic material of the eukaryotic host and two endosymbionts [75]. That this system integration is an ongoing process is revealed by the many small fragments of plastid origin in the nuclear genome [76] and the unexpected discovery of a recent gene-transfer event from the mitochondrial to the nuclear genome. Within the genetically defined centromer of chromosome 2, a 270-kbp fragment 99 % identical with the mitochondrial genome has been identified and its location confirmed *via* PCR across the junctions with unique nuclear DNA [64]. For a comprehensive list of references on *Arabidopsis thaliana* readers are referred to Schmidt [77]. This contribution cites only articles not listed by Schmidt or which are of utmost importance to the matters discussed here.

1.3.2.4

**Follow-up Research in the Postgenome Era**

Potential functional assignments can be made for up to 60 % of the predicted proteins on the basis of sequence comparisons, identification of functional domains and motifs, and structural predictions. Interestingly, 65 % of the proteins have no significant homology with proteins of the completely sequenced genomes of bacteria, yeast, *C. elegans*, and *D. melanogaster* [64], clearly reflecting the large evolutionary distance of the plant from other kingdoms and the independent development of multicellularity accompanied by a large increase in gene number. The discovery of protein classes and domains specific for plants, e.g. $Ca^{2+}$-dependent protein kinases containing four EF-hand domains [65] or the B3 domain of *ABI3*, *VP1*, and *FUS3* [78], and the significantly different abundance of several proteins or protein domains compared with *C. elegans* or *D. melanogaster*, for example myb-like transcription factors [79], C3HC4 ring finger domains, and P450 enzymes [65], further support this notion. Already the larger number of genes in the *Arabidopsis* genome (approx. 25,000) compared with the genomes of *C. elegans* (approx. 19,000) and *D. melanogaster* (approx. 14,000) seems indicative of different ways of evolving organisms of comparable complexity. Currently, the underlying reason for this large difference is unclear. It might simply reflect a larger proportion of duplicated genes, as it does for *C. elegans* compared with *D. melanogaster* [80]. The large number of observed tandem repeats [64, 65] and the large duplicated segments in chromosomes 2 and 4, 2.5 Mbp in total, and in chromosomes 4 and 5, a segment containing 37 genes [65], seem to favor this explanation. But other specific properties of plants, for example autotrophism, non-mobile life, rigid body structure, continuous organ development, successive gametophytic and sporophytic generations, and, compared with animals, different ways of processing information and responding to environmental stimuli and a smaller extent of combinatorial use of proteins, or any combination of these factors, might also contribute to their large number of genes. Functional assignment of proteins is not currently sufficiently advanced to enable us to distinguish between all these possibilities.

The data currently available indicate that the function of many plant genes is different from those of animals and fungi. Besides the basic eukaryotic machinery which was present in the last common ancestor before endosymbiosis with cyanobacteria created the plant kingdom, and which can be delineated by identification of orthologs in eukaryotic genomes [65], all the plant-specific functions must be elucidated. Because more than 40 % of all predicted proteins from the genome of *Arabidopsis* have no assigned function and many others have not been investigated thoroughly, this will require an enormous effort using different approaches. New techniques from chip-based expression analysis and genotyping to high-throughput proteomics and protein–ligand interaction studies and metabolite profiling have to be applied in conjunction with identification of the diversity present in nature or created *via* mutagenesis, transformation, gene knock-out, etc. Because multicellularity was established independently in all kingdoms, it might be wise to sequence the genome of a unicellular plant. Its gene content would help us identify all the proteins required for cell-to-cell communication and transport of signals and metabolites. To collect, store, evaluate, and access all the relevant data from these various approaches, many of which have been performed in parallel and de-

signed for high-throughput analyses, new bioinformatic tools must be created. To coordinate such efforts, a first meeting of "The Multinational Coordinated *Arabidopsis* 2010 Project" (http://www.arabidopsis.org/workshop1.html), with the objective of functional genomics and creation of a virtual plant within the next decade, was held in January 2000 at the Salk Institute in San Diego, USA. From just two examples it is evident that the rapid progress in *Arabidopsis* research will continue in the future. A centralized facility for chip-based expression analysis has been set-up already in Stanford, USA, and a company provides access to 39,000 potential single-nucleotide polymorphisms, which will speed up mapping and genotyping substantially. At the end the question remains the same as in the beginning – why should all these efforts be devoted to a model plant irrelevant to agriculture? The answer can be given again as a question – which other plant system would provide better tools for tackling all the basic questions of plant development and adaptation than *Arabidopsis thaliana*?

### 1.3.3 The Roundworm *Caenorhabditis elegans*

#### 1.3.3.1 The Organism

The free-living nematode *Caenorhabditis elegans* (Fig. 1.6) is the first multicellular animal whose genome has been completely sequenced. This worm, although often viewed as a featureless tube of cells, has been studied as a model organism for more than 20 years. In the 1970s and 1980s, the complete cell lineage of the worm from fertilized egg to adult was determined by microscopy [81], and later the entire nervous system was reconstructed [82]. It has proved to have several advantages as an object of biological study, for example simple growth conditions, rapid generation time with an invariant cell lineage, and well developed genetic and molecular tools for its manipulation. Many of the discoveries made with *C. elegans* are particularly relevant to the study of higher organisms, because it shares many of the essential biological features, for example neuronal growth, apoptosis, intra- and intercellular signaling pathways, food digestion, etc. that are in the focus of interest of, for example, human biology.

**Fig. 1.6** The freeliving bacteriovorous soil nematode *Caenorhabditis elegans* which is a member of the Rhabditidae (found at http://www.nematodes.org).

A special review by the *C. elegans* Genome Consortium [17] gives an interesting summary of how "the worm was won" and examines some of the findings from the near-complete sequence data.

The *C. elegans* genome was deduced from a clone-based physical map to be 100 Mb. This map was initially based on cosmid clones using a fingerprinting approach. Later yeast artificial chromosome (YAC) clones were incorporated to bridge the gaps between cosmid contigs, and provided coverage of regions that were not represented in the cosmid libraries. By 1990, the physical map consisted of fewer than 20 contigs and was useful for rescue experiments that were able to locate a phenotype of interest to a few kilobases of DNA [83]. Alignment of the existing genetic and physical maps into the *C. elegans* genome map

was greatly facilitated by cooperation of the entire "worm community". When the physical map had been nearly completed the effort to sequence the entire genome became both feasible and desirable. By that time this attempt was significantly larger than any sequencing project before and was nearly two orders of magnitude more expensive than the mapping effort.

1.3.3.2

**The Structure of the Genome Project**

In 1990 a three-year pilot project for sequencing 3 Mb of the genome was initiated as a collaboration between the Genome-Sequencing Center in St Louis, USA and the Sanger Center in Hinxton, UK. Funding was obtained from the NIH and the UK MRC. The genome sequencing effort initially focused on the central regions of the five autosomes which were well represented in cosmid clones, because most genes known at that time were contained in these regions. At the beginning of the project, sequencing was still based on standard radioisotopic methods, with "walking" primers and cosmid clones as templates for the sequencing reactions. A severe problem of this primer-directed sequencing approach on cosmids were, however, multiple priming events because of repetitive sequences and efficient preparation of sufficient template DNA. To address these problems the strategy was changed to a more classic shotgun sequencing strategy based on cosmid subclones generally sequenced from universal priming sites in the subcloning vectors. Further developments in automation of the sequencing reactions, fluorescent sequencing methods, improvements in dye-terminator chemistry [84], and the generation of assembly and contig editing programs, led to the phasing out of the instrument in favor of four-color, single-lane sequencing. The finishing phase then used a more ordered, directed sequencing strategy as well as the walking approach, to close specific remaining gaps and resolve ambiguities. Hence, the worm project grew into a collaboration among *C. elegans* Sequencing Consortium members and the entire international community of *C. elegans* researchers. In addition to the nuclear genome-sequencing effort other researchers sequenced its 15-kb mitochondrial genome and performed extensive cDNA analyses that facilitated gene identification.

The implementation of high-throughput devices and semi-automated methods for DNA purification and sequencing reactions led to overwhelming success in scaling up of the sequencing. The first 1 Mb threshold of *C. elegans* finished genome sequences was reached in May 1993. In August 1993, the total had already increased to over 2 Mb [85], and by December 1994 over 10 Mb of the *C. elegans* genome had been completed. Bioinformatics played an increasing role in the genome project. Software developments made the processing, analysis, and editing of thousands of data files per day a manageable task. Indeed, many of the software tools developed in the *C. elegans* project, for example ACeDB, PHRED, and PHRAP, have become key components in the current approach to sequencing the human genome.

The 50 Mb mark was passed in August 1996. At this point, it became obvious that 20 % of the genome was not covered by cosmid clones, and a closure strategy was implemented. For gaps in the central regions, either long-range PCR was used, or a fosmid library was probed in search of a bridging clone. For the remaining gaps in the central regions, and for regions of chromosomes contained only in YAC, purified YAC DNA was used as the starting material for shotgun sequencing. All of these final regions have been essentially completed with

the exception of several repetitive elements. The final genome sequence of the worm is, therefore, a composite from cosmids, fosmids, YAC, and PCR products. The exact genome size was approximate for a long time, mainly because of repetitive sequences that could not be sequenced in their entirety. Telomeres were sequenced from plasmid clones provided by Wicky et al. [86]. Of twelve chromosome ends, nine have been linked to the outermost YAC on the physical map.

1.3.3.3

**Results from the Genome Project**

Analysis of the approximately 100 Mb of the total *C. elegans* genome revealed nearly 20,000 predicted genes, considerably more than expected before sequencing began [87, 88], with an average density of one predicted gene per 5 kb. Each gene was estimated to have an average of five introns and 27 % of the genome residing in exons. The number of genes is approximately three times the number found in yeast [89] and approximately one-fifth to one-third the number predicted for the human genome [17].

Interruption of the coding sequence by introns and the relatively low gene density make accurate gene prediction more challenging than in microbial genomes. Valuable bioinformatics tools have been developed and used to identify putative coding regions and to provide an initial overview of gene structure. To refine computer-generated gene structure predictions, the available EST and protein similarities and the genomic sequence data from the related nematode *Caenorhabditis briggsae* were used for verification. Although approximately 60 % of the predicted genes have been confirmed by EST matches [17, 90], recent analyses have revealed that many predicted genes needed corrections in their intron-exon structures [91].

Similarities to known proteins provide an initial glimpse into the possible function of many of the predicted genes. Wilson [17] reported that approximately 42 % of predicted protein products have cross-phylum matches, most of which provide putative functional information [92]. Another 34 % of predicted proteins match other nematode proteins only, a few of which have been functionally characterized. The fraction of genes with informative similarities is far less than the 70 % observed for microbial genomes. This might reflect the smaller proportion of nematode genes devoted to core cellular functions [89], the comparative lack of knowledge of functions involved in building an animal, and the evolutionary divergence of nematodes from other animals extensively studied so far at the molecular level. Interestingly, genes encoding proteins with cross-phylum matches were more likely to have a matching EST (60 %) than those without cross-phylum matches (20 %). This observation suggested that conserved genes are more likely to be highly expressed, perhaps reflecting a bias for "house-keeping" genes among the conserved set. Alternatively, genes lacking confirmatory matches might be more likely to be false predictions, although the analyses did not suggest this.

In addition to the protein-coding genes, the genome contains several hundred genes for noncoding RNA. 659 widely dispersed transfer RNA genes have been identified, and at least 29 tRNA-derived pseudogenes are present. Forty-four percent of the tRNA genes were found on the X chromosome, which represents only 20 % of the total genome. Several other noncoding RNA genes, for example those for splicosomal RNA, occur in dispersed multigene families. Several RNA genes are located in the introns of protein coding genes, which might indicate RNA gene transposition [17]. Other noncoding RNA genes are located in long tandem

repeat regions; the ribosomal RNA genes were found solely in such a region at the end of chromosome I, and the 5S RNA genes occur in a tandem array on chemomosome V.

Extended regions of the genome code for neither proteins nor RNA. Among these are regions that are involved in gene regulation, replication, maintenance, and movement of chromosomes. Furthermore, a significant fraction of the *C. elegans* genome is repetitive and can be classified as either local repeats (e.g. tandem, inverted, and simple sequence repeats) or dispersed repeats. Tandem and inverted repeats amount to 2.7 and 3.6 % of the genome, respectively. These local repeats are distributed non-uniformly throughout the genome relative to genes. Not surprisingly, only a small percentage of tandem repeats are found within the 27 % of the protein coding genes. Conversely, the density of inverted repeats is higher in regions predicted as intergenic [17]. Although local repeat structures are often unique in the genome, other repeats are members of families. Some repeat families have a chromosome-specific bias in representation. Altogether 38 dispersed repeat families have been recognized. Most of these dispersed repeats are associated with transposition in some form [93], and include the previously described transposons of *C. elegans*. In addition to multiple-copy repeat families, a significant number of simple duplications have been observed to involve segments that range from hundreds of bases to tens of kilobases. In one instance a segment of 108 kb containing six genes was duplicated tandemly with only ten nucleotide differences observed between the two copies. In another example, immediately adjacent to the telomere at the left end of chromosome IV, an inverted repeat of 23.5 kb was present, with only eight differences found between the two copies. There are many instances of smaller duplications, often separated by tens of kilobases or more that might contain a coding sequence. This could provide a mechanism for copy divergence and the subsequent formation of new genes [17].

The GC content is more or less uniform at 36 % throughout all chromosomes, unlike human chromosomes that have different isochores [94]. There are no localized centromers, as are found in most other metazoa. Instead, the extensive highly repetitive sequences that are involved in spindle attachment in other organisms might be represented by some of the many tandem repeats found scattered among the genes, particularly on the chromosome arms. Gene density is also uniform across the chromosomes, although some differences are apparent, particularly between the centers of the autosomes, the autosome arms, and the X chromosome.

More striking differences become evident on examination of other features. Both inverted and tandem repeat sequences are more frequent on the autosome arms than in the central regions or on the X chromosome. This abundance of repeats on the arms is probably the reason for difficulties in cosmid cloning and sequence completion in these regions. The fraction of genes with cross-phylum similarities tends to be smaller on the arms as does the fraction of genes with EST matches. Local clusters of genes also seem to be more abundant on the arms.

#### 1.3.3.4 Follow-up Research in the Postgenome Era

Although sequencing of the *C. elegans* genome has essentially been completed, analysis and annotation will continue and – hopefully – be facilitated by further information and better technologies as they become available. The recently completed genome

sequencing of *C. briggsae* enabled comparative analysis of functional sequences of both genomes. The genomes are characterized by striking conservation of structure and function. Thus, many features could be verified in general. Although several specific characteristics had to be corrected for *C. elegans*, however, for example the number of predicted genes, the noncoding RNA, and repetitive sequences [95], it is now possible to describe many interesting features of the *C. elegans* genome, on the basis of analysis of the completed genome sequence.

The observations and findings of the *C. elegans* genome project provide a preliminary glimpse of the biology of metazoan development. There is much left to be uncovered and understood in the sequence. Of primary interest, all of the genes necessary to build a multicellular organism are now essentially available, although their exact boundaries, relationships, and functional roles have to be elucidated more precisely. The basis for a better understanding of the regulation of these genes is also now within our grasp. Furthermore, many of the discoveries made with *C. elegans* are relevant to the study of higher organisms. This extends beyond fundamental cellular processes such as transcription, translation, DNA replication, and cellular metabolism. For these reasons, and because of its intrinsic practical advantages, *C. elegans* has proved to be an invaluable tool for understanding features such as vertebrate neuronal growth and pathfinding, apoptosis and intra- and intercellular signaling pathways.

### 1.3.4 The Fruitfly *Drosophila melanogaster*

#### 1.3.4.1 The Organism

In 1910, T.H. Morgan started his analysis of the fruit fly *Drosophila melanogaster* and identified the first white-eyed mutant fly strain [96]. Now, less than 100 years later, almost the complete genome of the insect has been sequenced, offering the ultimate opportunity to elucidate processes ranging from the development of an organism to its daily behavior.

*Drosophila* (Fig. 1.7), a small insect with a short lifespan and very rapid, well-characterized development is one of the best analyzed multicellular organisms. During the last century, more than 1300 genes – mostly based on mutant phenotypes – were genetically identified, cloned, and sequenced. Surprisingly, most were found to have counterparts in other metazoa including even humans. Soon it became evident that not only are the genes conserved among species, but also the functions of the encoded proteins. Processes like the development of limbs, the nervous system, the eyes and the heart, or the presence of circadian rhythms and innate immunity are highly

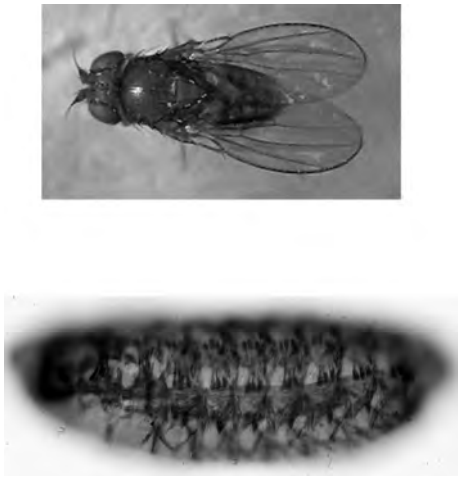

**Fig. 1.7** The fruit fly *Drosophila melanogaster*. (a) Adult fly. (b) Stage 16 embryo. In dark-brown, the somatic muscles are visualized by using an β3-tubulin-specific antibody; in red-brown, expression of β1-tubulin in the attachment sites of the somatic muscles is shown.

conserved, even to the extent that genes taken from human sources can supplement the function of genes in the fly [80]. Early in the 1980s identification of the HOX genes led to the discovery that the anterior–posterior body patterning genes are conserved from fly to man [97]. Furthermore, genes essential for the development and differentiation of muscles like twist, MyoD, and MEF2 act in most of the organisms analyzed so far [98]. The most striking example is the capability of a certain class of genes, the PAX-6 family, to induce ectopic eye development in *Drosophila*, irrespective of the animal source from which it was taken [99]. Besides these developmental processes, human disease networks involved in replication, repair, translation, metabolism of drugs and toxins, neural disorders like Alzheimer's, and also higher order functions like memory and signal transduction cascades have also been shown to be highly conserved [97, 100, 101]. It is quite a surprising conclusion that although simply an insect, Drosophila is capable of serving as a model system even for complex human diseases.

The genomic organization of *Drosophila melanogaster* has been known for many years. The fly has four chromosomes, three large and one small, with an early estimate of $1.1\times10^8$ bp. Using polytene chromosomes as tools, in 1935 and 1938 Bridges published maps of such accuracy they are still used today [96]. Making extensive use of chromosomal rearrangements he constructed cytogenetic maps that assigned genes to specific sections and even specific bands. With the development of techniques such as *in-situ* hybridization to polytene chromosomes, genes could be mapped with a resolution of less than 10 kb. Another major advantage of *Drosophila* is the presence of randomly shuffled chromosomes, the so-called balancers, which enable easy monitoring and pursuit of given mutations on the homologous chromosome and guarantee the persistence of such a mutation at the chromosomal place where it has initially occurred, by suppression of meiotic recombination.

1.3.4.2

**Structure of the Genome Project**

The strategy chosen for sequencing was the whole-genome-shotgun sequencing (WGS) technique. For this technique, the whole genome is broken into small pieces, subcloned into suitable vectors and sequenced. For the *Drosophila* genome project, libraries containing 2 kb, 10 kb, and 130 kb inserts were chosen as templates. They were sequenced from the ends, and assembled by pairs of reads, called mates, from the ends of the 2 kb and 16 kb inserts [102]. Assembly of the sequences was facilitated by the absence of interspersed repetitive elements like the human ALU-repeat family. The ends of the large BAC clones were taken to unequivocally localize the sequences into large contigs and scaffolds. Three million reads of ~500 bp were used to assemble the *Drosophila* genome. The detailed strategies, techniques, and algorithms used are described in an issue of Science [103] published in March 2000. The work was organized by the Berkeley *Drosophila* Genome Project (BDGP) and performed at the BDGP, the European DGP, the Canadian DGP and at Celera Genomics under the auspices of the federally funded Human Genome Project. The combined DGP produced all the genomic resources and finished 29 Mbp of sequences. In total, the *Drosophila* genome is ~180 Mbp in size, a third of which is heterochromatin. From the ~120 Mbp of euchromatin, 98 % are sequenced with an accuracy of at least 99.5 %. Because of the structure of the heterochromatin, which is mainly built up of re-

petitive elements, retrotransposons, rRNA clusters, and some unique sequences, most of it could not be cloned or propagated in YAC, in contrast with the *C. elegans* genome project [102].

1.3.4.3

**Results of the Genome Project**

As a major result, the number of genes was calculated to be ~13,600 using a *Drosophila*-optimized version of the program GENIE [100]. Over the last 20 years, 2500 of these have already been characterized by the fly community, and their sequences have been made available continuously during the ongoing of the fly project. The average density is one gene in 9 kb, but ranges from none to nearly 30 genes per 50 kbp, and, in contrast with *C. elegans*, gene-rich regions are not clustered. Regions of high gene density correlate with G + C-rich sequences. Computational comparisons with known vertebrate genes have led to both expected findings and surprises. So genes encoding the basic DNA replication machinery are conserved among eukaryotes; especially, all the proteins known to be involved in recognition of the replication start point are present as single-copy genes, and the ORC3 and ORC6 proteins are closely similar to vertebrate proteins but much different from those in yeast and, apparently, even more different from those in the worm. Focusing on chromosomal proteins, the fly seems to lack orthologs to most of the mammalian proteins associated with centromeric DNA, for example the CENP-C/MIF-2 family. Furthermore, because *Drosophila* telomeres lack the simple repeats characteristic of most eukaryotic telomeres, the known telomerase components are absent. Concerning gene regulation, RNA polymerase subunits and co-factors are more closely related to their mammalian counterparts than to yeast; for example, the promoter interacting factors UBF and TIF-IA are present in *Drosophila* but not in yeast. The overall set of transcription factors in the fly seems to comprise approximately 700 members, about half of which are zinc-finger proteins, whereas in the worm only one-third of 500 factors belong to this family. Nuclear hormone receptors seem to be more rare, because only four new members were detected, bringing the total to 20 compared with more than 200 in *C. elegans*. As an example of metabolic processes, iron pathway components were analyzed. A third ferritin gene has been found that probably encodes a subunit belonging to cytosolic ferritin, the predominant type in vertebrates. Two newly discovered transferrins are homologs of the human melanotransferrin p97, which is of especial interest, because the iron transporters analyzed so far in the fly are involved in antibiotic response rather than in iron transport. Otherwise, proteins homologous to transferrin receptors seem to be absent from the fly, so the melanotransferrin homologs might mediate the main insect pathway for iron uptake (all data taken from Adams et al. [100]). The sequences and the data compiled are freely available to the scientific community on servers in several countries (see: http://www.fruitfly.org). In addition, large collections of EST (expressed sequence tags), cDNA libraries, and genomic resources are available, as are data bases presenting expression patterns of identified genes or enhancer trap lines, for example flyview at the University of Münster (http://pbio07.uni-muenster.de), started by the group of W. Janning. Furthermore, by using more advanced transcription profiling approaches in combination with lower stringency gene prediction programs, approximately 2000 new genes could be detected, according to a recent report [104].

1.3.4.4

**Follow-up Research in the Postgenome Era**

Comparative analysis of the genomes of *Drosophila melanogaster, Caenorhabditis elegans*, and *Saccharomyces cerevisiae* has been performed [80]. It showed that the core proteome of *Drosophila* consists of 9453 proteins, which is only twice the number for *S. cerevisiae* (4383 proteins). Using stringent criteria, the fly genome is much closer related to the human genome than to that of *C. elegans* one. Interestingly, the fly has orthologs to 177 of 289 human disease genes examined so far and provides an excellent basis for rapid analysis of some basic processes involved in human disease. Furthermore, hitherto unknown counterparts were found for human genes involved in other disorders, for example, *menin*, *tau*, *limb girdle muscular dystrophy type 2B*, *Friedrich ataxia*, and *parkin*. Of the cancer genes surveyed, at least 68 % seem to have *Drosophila* orthologs, and even a p53-like tumor-suppressor protein was detected. All of these fly genes are present as single copies and can be genetically analyzed without uncertainty about additional functionally redundant copies. Hence, all the powerful *Drosophila* tools such as genetic analysis, exploitation of developmental expression patterns, loss-of-function, and gain-of-function phenotypes can be used to elucidate the function of human genes that are not yet understood. A very recent discovery, the function of a new class of regulatory RNA termed microRNA, has rapidly attracted attention to the non-coding parts of the genome [105]. Screens for miRNA targets in *Drosophila* [106] have already led to identification of a large collection of genes and will certainly be very helpful in elucidating the role of the miRNA during development. Furthermore, genome-wide protein–protein-interaction studies using the yeast two-hybrid system have been performed [107] and led to a draft map of about 20,000 interactions between more than 7,000 proteins. These results must certainly be refined but clearly emphasize the importance of genome-wide screening of both proteins and RNA in combination with elaborated computational methods.

## 1.4 Conclusions

This book chapter summarizes the genome projects of selected model organisms with completed, or almost completed, genomic sequences. The organisms represent the major phylogenetic lineages, the eubacteria, archaea, fungi, plants, and animals, thus covering unicellular prokaryotes with singular, circular chromosomes, and uni- and multicellular eukaryotes with multiple linear chromosomes. Their genome sizes range from ~2 to ~180 Mbp with estimated gene numbers ranging from 2000 to 25,000 (Tab. 1.3).

The organization of the genome of each organism has its own specific and often unexpected characteristics. In *B. subtilis*, for example, a strong bias in the polarity of transcription of the genes with regard to the replication fork was observed, whereas in *E. coli* and *S. cerevisiae* the genes are more or less equally distributed on both strands. Furthermore, although insertion sequences are widely distributed in bacteria, none was found in *B. subtilis*. In *A. thaliana*, an unexpectedly high percentage of proteins showed no significant homology with proteins of organisms outside the plant kingdom, and thus are obviously specific to plants. Another interesting observation resulting from the genome projects concerns gene density. In prokaryotic organisms, i.e.

**Table 1.3** Summary of information on the genomes of model organisms described in this chapter (status April 2004).

| *Organism* | *Genome structure* | *Genome size (kb)* | *Estimated no. of genes/ORF* |
|---|---|---|---|
| *Escherichia coli* | 1 chromosome, circular | 4600 | 4400 |
| *Bacillus subtilis* | 1 chromosome, circular | 4200 | 4100 |
| *Archaeoglobus fulgidus* | 1 chromosome, circular | 2200 | 2400 |
| *Saccharomyces cerevisiae* | 16 chromosomes, linear | 12,800 | 6200 |
| *Arabidopsis thaliana* | 5 chromosomes, linear | 130,000 | 25,000 |
| *Caenorhabditis elegans* | 6 chromosomes, linear | 97,000 | 19,000 |
| *Drosophila melanogaster* | 4 chromosomes, linear | 180,000 | 16,000 |

eubacteria and archaea, genome sizes vary substantially. Their gene density is, however, relatively constant at approximately one gene per kbp. During evolution of eukaryotic organisms, genome sizes grew, but gene density decreased from one gene in two kbp in *Saccharomyces* to one gene in 10 kbp in *Drosophila*. This led to the surprising observation that some bacterial species can have more genes than lower eukaryotes and that the number of genes in *Drosophila* is only approximately three times higher than the number in *E. coli*.

Another interesting question concerns the general homology of genes or gene products between model organisms. Comparison of protein sequences has shown that some gene products can be found in a wide variety of organisms.

Thus, comparative analysis of predicted protein sequences encoded by the genomes of *C. elegans* and *S. cerevisiae* suggested that most of the core biological functions are conducted by orthologous proteins that occur in comparable numbers in these organisms. Furthermore, comparing the yeast genome with the catalog of human sequences available in the databases revealed that a significant number of yeast genes have homologs among human genes of unknown function. *Drosophila* is of special importance in this respect, because many human disease networks have been shown to be highly conserved in the fruitfly. Hence, the insect is, in an outstanding manner, capable of serving as a model system even for complex human diseases. Because the respective genes in the fly are present as single copies, they can be genetically analyzed much more easily.

Genome research of model organisms has just begun. At the time when this article was written, more than 100 genome sequences, mainly from prokaryotes, have been published and more than 200 genomes are currently being sequenced worldwide. The full extent of this broad and rapidly expanding field of genome research on model organisms cannot be covered by a single book chapter. The World-Wide Web, however, is an ideal platform for making available the outcome of large genome projects in an appropriate and timely manner. Beside several specialized entry points, two web pages are recommended as an introduction with a broad overview of model-organism research – the WWW Virtual Library: Model Organisms (http://ceolas.org/VL/mo) and the WWW Resources for Model Organisms (http://genome.cbs.dtu.dk/gorm/modelorganism.html). Both links are excellent starting sites for detailed information on model-organism research.

## References


1 Berg, P., Baltimore, D., Brenner, S., Roblin, R.O. 3rd, Singer, M.F. (1975), Asilomar conference on recombinant DNA molecules. *Science* 188, 991–994.

2 Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V. et al. (1997), The complete genome sequence of *Escherichia coli* K-12, *Science* 277, 1453–1474.

3 Yamamoto, Y., Aiba, H., Baba, T., Hayashi, K., Inada, T. et al. (1997), Construction of a contiguous 874-kb sequence of the *Escherichia coli* -K12 genome corresponding to 50.0–68.8 min on the linkage map and analysis of its sequence features. *DNA Res.* 4, 91–113.

4 Kröger, M., Wahl, R. (1998) Compilation of DNA sequences of *Escherichia coli* K12: description of the interactive databases ECD and ECDC, *Nucl. Acids Res.* 26, 46–49.

5 Thomas, G.H. (1999), Completing the *E. coli* proteome: a database of gene products characterised since completion of the genome sequence, *Bioinformatics* 15, 860–861.

6 Taylor, A.L., Thoman, M.S. (1964) The genetic map of Escherichia coli K-12. *Genetics* 50, 659–677.

7 Bachmann, B.J. (1983) Linkage map of *Escherichia coli* K-12, edition 7. *Microbiol. Rev.* 47, 180 –230.

8 Neidhard, F.C. (1996) *Escherichia coli and Salmonella typhimurium – Cellular and Molecular Biology*, 2nd edn. American Society for Microbiology, Washington DC.

9 Kohara, Y., Akiyama, K., Isono, K. (1987), The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* 50, 495–508.

10 Roberts, R. (2000) Database issue of Nucleic Acids Research, *Nucl. Acids Res.* 28, 1–382.

11 Datsenko, K.A., Wanner, B.L. (2000), One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *PNAS* 97, 6640–6645.

12 Karp, P.D., Riley, M.R., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S. (2002), The EcoCyc Database. *Nucl. Acids Res.* 30, 56–58.

13 Sekowska, A. (1999), Une rencontre du métabolisme du soufre et de l'azote: le métabolisme des polyamines chez *Bacillus subtilis*, thesis in *Génétique cellulaire et moléculaire*. Versailles-Saint-Quentin: Versailles, France. p. 202.

14 Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G. et al. (1997), The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256.

15 Moreno, M.S., Schneider, B.L., Maile, R.R., Weyler, W., Saier, M.H., Jr. (2001), Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: novel modes of regulation revealed by whole-genome analyses. *Mol. Microbiol.* 39, 1366–1381.

16 Fabret, C., Quentin, Y., Guiseppi, A., Busuttil, J., Haiech, J., Denizot, F. (1995), Analysis of errors in finished DNA sequences: the surfactin operon of *Bacillus subtilis* as an example. *Microbiology* 141, 345–350.

17 Wilson, R.K. (1999), How the worm was won. The *C. elegans* genome sequencing project. *Trends Genet.* 15, 51–58.

18 Arias, R.S., Sagardoy, M.A., van Vuurde, J.W. (1999), Spatio-temporal distribution of naturally occurring *Bacillus* spp. and other bacteria on the phylloplane of soybean under field conditions. *J. Basic Microbiol.* 39, 283–292.

19 Rocha, E.P. (2003), DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet*. 19, 600–603.

20 Rocha, E.P., Danchin, A., Viari, A. (1999), Universal replication biases in bacteria. *Mol. Microbiol*. 32, 11–16.

21 Schofield, M.J. and Hsieh, P. (2003), DNA mismatch repair: molecular mechanisms and biological function. *Annu. Rev. Microbiol*. 57, 579–608.

22 Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A. et al. (2001), Comparative genomics of Listeria species. *Science* 294, 849–852.

23 Rocha, E.P., Sekowska, A., Danchin, A. (2000), Sulphur islands in the *Escherichia coli* genome: markers of the cell's architecture? FEBS Lett. 476, 8–11.

24 Danchin, A., Guerdoux-Jamet, P., Moszer, I., Nitschke, P. (2000), Mapping the bacterial cell architecture into the chromosome. Philos. *Trans. R. Soc. Lond. B Biol. Sci*. 355, 179–190.

25 Gottesman, S., Roche, E., Zhou, Y., Sauer, R.T. (1998), The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev*. 12, 1338–1347.

26 Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K. et al. (2003), Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* 100, 4678–4683.

27 Soma, A., Ikeuchi, Y., Kanemasa, S., Kobayashi, K., Ogasawara, N. et al. (2003), An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Mol. Cell*. 12, 689–698.

28 Rocha, E.P. and Danchin, A. (2003), Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res*. 31, 6570–6577.

29 Beckering, C.L., Steil, L., Weber, M.H., Volker, U., Marahiel, M.A. (2002), Genomewide transcriptional analysis of the cold shock response in *Bacillus subtilis*. *J. Bacteriol*. 184, 6395–6402.

30 Hoffmann, T., Schutz, A., Brosius, M., Volker, A., Volker, U., Bremer, E. (2002), High-salinity-induced iron limitation in *Bacillus subtilis*. *J. Bacteriol*. 184, 718–727.

31 Danchin, A. (1997), Comparison between the *Escherichia coli* and *Bacillus subtilis* genomes suggests that a major function of polynucleotide phosphorylase is to synthesize CDP. *DNA Res*. 4, 9–18.

32 Klenk, H.-P., Clayton, R.A., Tomb, J.-F., White, O., Nelson, K.E. et al. (1997), The complete genome of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* 390, 364–370.

33 Stetter, K.O. (1988), *Archaeoglobus fulgidus* gen. nov., sp. nov.: a new taxon of extremely thermophilic archaebacteria. *System. Appl. Microbiol*. 10, 172–173.

34 Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G. et al. (1997), The complete genome sequence of the gastric pathogen *Heliobacter pylori*. *Nature* 388, 539–547.

35 Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D. et al. (1996), Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*, *Science* 273, 1058–1073.

36 Rabus, R., Ruepp, A., Frickey, T., Rattei, T., Fartmann, B. et al. (2004), The genome of Desulfotalea psychrophila, a sulphate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol*., in press.

37 Vorholt, J., Kunow, J., Stetter, K.O., Thauer, R.K. (1995), Enzymes and coenzymes of the carbon monoxide dehydrogenase pathway for autotrophic $CO_2$ fixation in *Archaeoglobus lithotrophicus* and the lack of carbon monoxide dehydrogenase in the heterotrophic *A. profundus*. *Arch. Microbiol*. 163, 112–118.

38 Stetter KO, Lauerer G, Thomm M, Neuner A (1987) Isolation of extremely thermophilic sulfate reducers: evidence for a novel branch of archaebacteria. *Science* 236, 822–824.

39 Hosfield, D.J., Frank, G., Weng, Y., Tainer, J.A., Shen, B. (1998), Newly discovered archaebacterial flap endonucleases show a structure-specific mechanism for DNA substrate binding and catalysis resembling human flap endonuclease-1. *J. Biol. Chem*. 273, 27154–27161.

40 Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W. et al. (1999), Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol*. 17, 292–296.

41 Cooksey, R.C., Holloway, B.P., Oldenburg, M.C., Listenbee, S., Miller, C.W. (2000), Evaluation of the invader assay, a linear signal amplification method, fro identification of mutations associated with resistance of rifampin and isoniazid in *Mycobacterium*

*tuberculosis*, *Antimicrob. Agents Chemother.* 44, 1296–1301.

**42** Tang, T.H., Bachellerie, J.P. Rozhdestvensky, T., Bortolin, M.L., Huber, H. et al., (2003), Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus, *Proc. Natl. Acad. Sci. USA* 99, 7536–7541.

**43** Maisnier-Patin, S., Malandrin, L., Birkeland, N.K., Bernander, R. (2002), Chromosome replication patterns in the hyperthermophilic euryarchaea *Archaeoglobus fulgidus* and *Methanocaldococcus (Methanococcus) janashii*, *Mol. Microbiol.* 45, 1443–1450.

**44** Goffeau, A. Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B. et al. (1996), Life with 6000 genes, *Science* 274, 546–567.

**45** Anonymous (1997), Dictionary of the yeast genome, *Nature* 387 (suppl.).

**46** Lindegren, C.C. (1949) The Yeast Cell, its Genetics and Cytology. Educational Publishers, St Louis, MI.

**47** Mortimer, R.K., Contopoulou, R., King, J.S. (1992), Genetic and physical maps of *Saccharomyces cerevisiæ*, Edition 11. *Yeast* 8, 817–902.

**48** Carle, G.F., Olson, M.V. (1985), An electrophoretic karyotype for yeast. *Proc. Natl. Acad. Sci. USA* 82, 3756.

**49** Stucka, R., Feldmann, H. (1994), Cosmid cloning of Yeast DNA, in: *Molecular Genetics of Yeast – A Practical Approach* (Johnston, J. Ed.) Oxford Univ. Press, pp. 49–64.

**50** Thierry, A., Dujon, B. (1992), Nested chromosomal fragmentation in yeast using the meganuclease *I-Sce* I: a new method for physical mapping of eukaryotic genomes. *Nucleic Acids Res.* 20, 5625–5631.

**51** Louis, E.J., Borts, R.H. (1995), A complete set of marked telomeres in *Saccharomyces cerevisiae* for physical mapping and cloning. *Genetics* 139, 125–136.

**52** Wolfe, K.H., Shields, D.C. (1997), Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.

**53** Foury, F., Roganti, T., Lecrenier, N., Pumelle, B. (1998), Yeast genes and human disease. *FEBS Lett.* 440, 325–331.

**54** DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997), Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686.

**55** Goffeau, A. (2000), Four years of postgenomic life with 6,000 yeast genes. *FEBS Lett.* 480, 37–41.

**56** Feldmann, H. (ed.) (2000), Genolevures: Genomic exploration of the hemiascomycetous yeasts. *FEBS Lett.* 487, no. 1 (http://cbi.labri.fr/genolevures).

**57** Laibach, F. (1907). Zur Frage nach der Individualität der Chromosomen im Pflanzenreich. *Beih. Bot. Centralblatt, Abt. I*, 22, 191–210.

**58** Laibach, F. (1943) *Arabidopsis thaliana* (L.) Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen. *Bot. Arch.* 44, 439–455.

**59** Laibach, F. (1951) Über sommer- und winterannuelle Rassen von *Arabidopsis thaliana* (L.) Heynh. Ein Beitrag zur Atiologie der Blütenbildung, *Beitr. Biol. Pflanzen* 28, 173–210.

**60** Müller, A. (1961) Zur Charakterisierung der Blüten und Infloreszenzen von *Arabidopsis thaliana* (L.) Heynh. *Die Kulturpflanze* 9, 364–393.

**61** Arumuganathan, K., Earle E.D. (1991), Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9, 208–218.

**62** Cherry, J.M., Cartinhour, S.W., Goodmann, H.M. (1992), AAtDB, an *Arabidopsis thaliana* database. *Plant Mol. Biol. Rep.* 10, 308–309, 409–410.

**63** Rédei, G.P. (1992), A heuristic glance at the past of *Arabidopsis* genetics, in: *Methods in Arabidopsis Research* (Koncz, C., Chua, N.-H., Schell, J. Eds.), World Scientific Publishing, Singapore, pp. 1–15.

**64** Lin, X., Kaul, S., Rounsley, S., Shea, T.P., Benito, M.I. et al. (1999), Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402, 761–768.

**65** Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G. et al. (1999), Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* 402, 769–777.

**66** Unseld, M., Marienfeld, J.R., Brandt, P., Brennicke, A. (1997), The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genet.* 15, 57–61.

**67** Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Tabata, S. (1999), Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* 6, 283–290.

**68** Gill, K.S., Gill, B.S., Endo, T.R., Taylor, T. (1996a), Identification and high density mapping of gene-rich regions in the chromosome group *5* of wheat. *Genetics* 143, 1001–1012.

**69** Gill, K.S., Gill, B.S., Endo, T.R., Taylor, T. (1996b), Identification and high density mapping of gene–rich regions in the chromosome group *1* of wheat. *Genetics* 144, 1883–1891.

**70** Künzel, G., Korzun, L., Meister, A. (2000), Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154, 397–412.

**71** SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., Bennetzen, J.L. (1996), Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274 765–768.

**72** Meyerowitz, E.M. (2000), Today we have the naming of the parts. *Nature* 402, 731–732.

**73** Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000), Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016.

**74** Martin, W., Schnarrenberger, C. (1997), The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* 32, 1–8.

**75** Herrmann, R.G. (2000), Organelle genetics – part of the integrated plant genome. *Vortr. Pflanzenzüchtg.* 48, 279–296.

**76** Bevan, M., Bennetzen, J.L., Martienssen, R. (1998), Genome studies and molecular evolution. Commonalities, contrasts, continuity and change in plant genomes. *Curr Opin Plant Biol.* 101–102.

**77** Schmidt, R. (2000), The *Arabidopsis* genome. *Vortr. Pflanzenzüchtg.* 48, 228–237.

**78** Suzuki, M., Kao, C.Y., McCarty, D.R. (1997), The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. *Plant Cell* 9, 799–807.

**79** Jin, H., Martin, C. (1999), Multifunctionality and diversity within the plant MYB-gene family. *Plant Mol. Biol.* 41, 577–585.

**80** Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R. et al. (2000), Comparative genomics of eukaryotes. *Science* 287, 2204–2215.

**81** Sulston, J. (1988), Cell Lineage; In: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 123–155, CSHL Press

**82** Chalfie, M., White, J. (1988), The Nervous system, In: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 337–391, CSHL Press.

**83** Coulson, A., Waterston, R., Kiff, J., Sulston, J., Kohara, Y. (1988), Genome linking with yeast artificial chromosomes. *Nature* 335, 184–186.

**84** Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D., Wilson, R.K. (1992), DNA sequencing with dye-labeled terminators and T7 DNA polymerase: Effect of dyes and dNTPs on incorporation of dye-terminators, and probability analysis of termination fragments. *Nucleic Acids Res.* 20, 2471–2483.

**85** Wilson, R., Ainscough, R., Anderson, K., Baynes, C., Berks, M. et al. (1994), 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature* 368, 32–38.

**86** Wicky, C., Villeneuve, A.M., Lauper, N., Codourey, L., Tobler, H., Muller, F. (1996), Telomeric repeats (TTAGGC) are sufficient for chromosome capping in *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* 93, 8983–8988.

**87** Herman, R.K. (1988), Genetics, in: *The Nematode Caenorhabditis elegans* (Wood, W.B. Ed.) pp. 17–45, CSHL Press.

**88** Waterston, R., Sulston, J. (1995), The genome of *Caenorhabditis elegans. Proc. Natl. Acad. Sci. USA* 92, 10836–10840.

**89** Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V. et al. (1998), Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022–2027.

**90** Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., Spieth, J. (2001), WormBase: network access to the genome and biology of *Caenorhabditis elegans. Nucleic Acids Res.* 29, 82–86.

**91** Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M. et al. (2003), C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nature Genet*. 34, 35–41.

**92** Green, P., Lipman, D., Hillier, L., Waterston, R., States, D., Claverie, J.M. (1993), Ancient conserved regions in new gene sequences and the protein database. *Science* 259, 1711–1716.

**93** Smit, A.F. (1996), The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev*. 6, 743–748.

**94** Bernardi, G. (1995), The human genome: organization and evolutionary history. *Annu. Rev. Genet*. 29, 445–476.

**95** Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R. et al. (2003), The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLOS Biology*, 1, 166–192.

**96** Rubin, G.M., Lewis, E.B. (2000), A brief history of *Drosophila´s* contributions to genome research. *Science* 287, 2216–2218.

**97** Veraksa, A., Del Campo, M., McGinnis, W. (2000), Developmental patterning genes and their conserved functions: From model organisms to humans. *Mol. Genet. Metab*. 69, 85–100.

**98** Zhang, J.M., Chen, L., Krause, M., Fire, A., Paterson, B.M. (1999), Evolutionary conservation of MyoD function and differential utilization of E proteins. *Dev. Biol*. 208, 465–472.

**99** Halder, G., Callaerts, P., Gehring, W. J. (1995), Induction of ectopic eyes by targeted expression of the *eyeless* gene in *Drosophila*. *Science* 267, 1788–1792.

**100** Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D. et al., (2000), The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195.

**101** Mayford, M., Kandel, E.R. (1999), Genetic approaches to memory storage. *Trends Genet*. 15, 463–470.

**102** Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P. et al. (2000), A whole-genome assembly of *Drosophila*. *Science* 287, 2196–2204.

**103** Anonymous (2000) The Drosophila Genome, *Science* 287, issue 5461.

**104** Hild, M., Beckmann, B., Haas, S.A., Koch, B., Solovyev, V. et al. (2003), An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol*. 5, R3.

**105** Lai Eric C. (2003), microRNAs: Runts of the genome assert themselves. *Curr. Biol*. 13, R925–R936.

**106** Stark, A., Brennecke, J., Russell, R.B., Cohen, S.M. (2003), Identification of *Drosophila* microRNA targets. *PLOS Biology*, 1, 397–409.

**107** Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B. et al. (2003), A protein interaction map of *Drosophila melanogaster*. *Science* 302, 1727–1736.