CHAPTER I

Introduction:
Structural and functional genomics

I.1. General introduction

A famous researcher once said: “Give me your DNA sequence and I know you”. How wrong he was! Of course, the genomic DNA sequence of an organism specifies the characteristics of its owner, but knowing this string of letters by no means tells you the exact nature of the corresponding organism. In reality, a long and difficult road separates the determination of the last base of a genome and the understanding of how the organism in question evolved, grows, propagates, interacts with its environment and, eventually, dies. This understanding will be the Holy Grail of generations of geneticists to come.

The term genomics, introduced in 1986 by T. H. Roderick to describe the study of complete genomes, refers to the kind of research that is used to get answers to the above questions. However diverse the interpretation of this term in the scientific community may be, the following definitions more or less cover the subject. Structural genomics is the scientific discipline of mapping, sequencing and analysing genomes, while functional genomics refers to analysis of genome function. The structural genomics phase has a clear end-point with the completion, annotation (and publication) of a genome sequence. The fundamental strategy of functional genomics is expanding the scope of research from studying single genes or proteins to studying all genes or proteins simultaneously in a systematic fashion, in order to obtain a panoramic view of the organisms’ genetic potentials (Hieter & Boguski, 1997).

Like the term, the field of research now referred to as genomics is very young. In fact, until the completion of the first genome sequence in 1995, that of the bacterium *Haemophilus influenzae*, performing genomics research was not really feasible. However, nineteen genomes are already known today, januari 1999, including the genomes of *Bacillus subtilis* and *Escherichia coli*, the model organisms for Gram-positive and Gram-negative bacteria, respectively. The genomes of the model organisms for genetic research, the nematode *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster*, the plant *Arabidopsis thaliana*, the human, and the mouse genome, will have been determined within the next decade. This awesome avalanche of information yields, together with technological advances of the past few years, such as DNA chip technology, unprecedented opportunities for the scientific community
Chapter I

to study genetic organisations and regulation processes at a genomic scale. Moreover, the comparative study of complete genomes has, and will undoubtedly continue to do so in the future, uncovered biological phenomena that were not even thought of in the past. The genome sequence of the Gram-positive bacterium *Bacillus subtilis*, the subject of this thesis, is a very valuable asset in this study, since it has already been the subject of intense genetic research for several decades.

One of the important lessons geneticists have learned by now is that a chromosome, or genome, is not merely a carrier of a physically bundled collection of genes of an organism. Instead, it reflects the superimposition of a myriad of biological phenomena that have evolved during four billion years or so of the struggle of life on earth.

### I.2. Genome sequencing: structural genomics

#### Completed genomes

The past few years have brought a new revolution in genetics, since it caused a reversion of the way genetic research is being done. Rather than attempting to isolate a gene on the basis of a phenotype, investigators now take the sequence of a gene responsible for the sought-for function from one organism and search its counterpart in the genome of the organism of interest, and then proceed to confirm this function. The year 1995 represents an important landmark of this revolution with the first publication of the DNA sequence of an entire genome of a free-living organism. This concerned the 1,830,137 bps chromosome of the bacterium *Haemophilus influenzae* (Fleischmann *et al.*, 1995). Since then, the number of entirely sequenced genomes has expanded with astonishing speed. The second genome, from the bacterium *Mycoplasma genitalium*, was published 3 months later in that same year. This genome represents the smallest known genome of a free-living organism, consisting of only 580,070 bps and comprising no more than 470 predicted genes (Fraser *et al.*, 1995). The next year brought the completion of the genomes of the archaeon *Methanococcus jannaschii* (Bult *et al.*, 1996), the cyanobacterium *Synechocystis* sp. strain PCC6803 (Kaneko *et al.*, 1996), the pathogenic bacterium *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), and -the first eukaryote- *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996; Mewes *et al.*, 1997). In December 1998, the first complete genome sequence of a multicellular eukaryote, that of the 97 Mb genome of *Caenorhabditis elegans*, was published. Table I.1 summarises some basic characteristics of the genomes, the sequences of which have been determined so far. To date, nineteen genome sequences are completely known, and many more are in the process of being determined. Sequencing efforts are presently primarily directed to pathogenic organisms, such as *Enterococcus faecalis*, *Legionella pneumophila*, *Mycobacterium leprae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Streptococcus pneumoniae*, and *Vibrio cholerae*. However, most classical subjects of biological study are also being determined, such as the human genome and those of *A. thaliana* and *D. melanogaster*.

<p>| Table I.1. Overview of completely sequenced genomes (in order of completion) |</p>
<table>
<thead>
<tr>
<th>Organism Strain (% (G/C))</th>
<th>size (kbp)</th>
<th>Number of ORFs (and RNA species)</th>
<th>c.d.(^5)</th>
<th>% unknown (% w.o. DB match; % sim. to hypoth.)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haemophilus influenzae Rd KW20 (38%)</td>
<td>1,830</td>
<td>1,743 (72)</td>
<td>85</td>
<td>42 (22; 20)</td>
<td>(Fleischmann et al., 1995)</td>
</tr>
<tr>
<td>Mycoplasma genitalium G-37 (32%)</td>
<td>580</td>
<td>470 (36)</td>
<td>88</td>
<td>32 (20; 12)</td>
<td>(Fraser et al., 1995)</td>
</tr>
<tr>
<td>Methanooccus jannaschii (^1) DSM 2661 (31%)</td>
<td>1,664</td>
<td>1,738 (43)</td>
<td>??</td>
<td>38 (22; 16)</td>
<td>(Bult et al., 1996)</td>
</tr>
<tr>
<td>Synechocystis sp. PCC6803</td>
<td>3,573</td>
<td>3,168 (??)</td>
<td>??</td>
<td>56 (45; 11)</td>
<td>(Kaneko et al., 1996)</td>
</tr>
<tr>
<td>Mycoplasma pneumoniae M129 (40%)</td>
<td>816</td>
<td>677 (39)</td>
<td>89</td>
<td>26 (16; 10)</td>
<td>(Himmelreich et al., 1996)</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (^2) S288C</td>
<td>12,068</td>
<td>5,885 (455)</td>
<td>70</td>
<td>29 (16; 13)</td>
<td>(Goffeau et al., 1996), (Mewes et al., 1997)</td>
</tr>
<tr>
<td>Helicobacter pylori 26695 (39%)</td>
<td>1,668</td>
<td>1,590 (44)</td>
<td>91</td>
<td>(31; ??)</td>
<td>(Tomb et al., 1997)</td>
</tr>
<tr>
<td>Escherichia coli K-12 (51%)</td>
<td>4,639</td>
<td>4,288 (107)</td>
<td>89</td>
<td>38 (32; 6)</td>
<td>(Blattner et al., 1997)</td>
</tr>
<tr>
<td>Methanobacterium thermoautotrophicum ΔH (50%)</td>
<td>1,751</td>
<td>1,855 (47)</td>
<td>92</td>
<td>54 (27; 28)</td>
<td>(Smith et al., 1997)</td>
</tr>
<tr>
<td>Bacillus subtilis 168 (44%)</td>
<td>4,215</td>
<td>4,221 (121)</td>
<td>87</td>
<td>42 (26; 16)</td>
<td>(Kunst et al., 1997)</td>
</tr>
<tr>
<td>Archaeoglobus fulgidus VC-16, DSM4304 (49%)</td>
<td>2,178</td>
<td>2,436 (56)</td>
<td>93</td>
<td>53 (26; 27)</td>
<td>(Klenk et al., 1997)</td>
</tr>
<tr>
<td>Borrelia burgdorferi (^1) B31 (29%)</td>
<td>911</td>
<td>853 (41)</td>
<td>94</td>
<td>41 (29; 12)</td>
<td>(Fraser et al., 1997)</td>
</tr>
<tr>
<td>Aquifex aeolicus (^4) VF5 (43%)</td>
<td>1,551</td>
<td>1,512 (51)</td>
<td>93</td>
<td>44 (27; 17)</td>
<td>(Deckert et al., 1998)</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis (^5) H37Rv (66%)</td>
<td>4,412</td>
<td>3,924 (50)</td>
<td>91</td>
<td>60 (16; 44)</td>
<td>(Cole et al., 1998)</td>
</tr>
<tr>
<td>Treponema pallidum pallidum (53%)</td>
<td>1,138</td>
<td>1,041 (50)</td>
<td>93</td>
<td>45 (28; 17)</td>
<td>(Fraser et al., 1998)</td>
</tr>
<tr>
<td>Pyrococcus horikoshii OT3 (42%)</td>
<td>1,739</td>
<td>2,061 (50)</td>
<td>91</td>
<td>80 (58; 22)</td>
<td>(Kawarabayasi et al., 1998)</td>
</tr>
<tr>
<td>Chlamydia trachomatis (41%) (^6) serovar D (D/UW-3/Cx)</td>
<td>1,043</td>
<td>894 (??)</td>
<td>??</td>
<td>32 (28; 4)</td>
<td>(Stephens et al., 1998)</td>
</tr>
<tr>
<td>Rickettsia prowazekii Madrid E (29%)</td>
<td>1,112</td>
<td>834 (37)</td>
<td>76</td>
<td>38 (25; 13)</td>
<td>(Andersson et al., 1998)</td>
</tr>
<tr>
<td>Caenorhabditis elegans (^7)</td>
<td>97,000</td>
<td>19,099 (&gt;1000)</td>
<td>27</td>
<td>?? (58; ?)</td>
<td>(C. elegans sequencing consortium (see genome, wustl.edu/gsc/C_elegans), 1998)</td>
</tr>
</tbody>
</table>

\(^{5}\) c.d.: coding density: the percentage of DNA that is actually coding for proteins and RNA species.

\(^{1}\) M. jannaschii contains two extrachromosomal elements (ECEs) of 58 kb (28% G/C) and 16 kb (29% G/C), respectively. These data are the sum of the three genomic elements.

\(^{2}\) The yeast genome consists of 16 chromosomes and only 12.068 kb of the genome, totalling 13.389 kb, was sequenced. Not completely sequenced are Ty-elements, rDNA repeats, mtDNA and some highly repeated genes (e.g. CUP1 & ENA2). Information on similarities was obtained from: http://www.mips.biochem.mpg.de/yeast/

\(^{3}\) B. burgdorferi contains one linear chromosome of 910,725 bps and at least 17 linear and circular plasmids totalling over 533,000 bps. Presented information covers only the chromosome. The
plasmids, with a coding density of 71%, encode at least 430 polypeptides of which 58% has no d.b. match and 26% match only hypothetical proteins.

\(^4\) A. aeolicus contains one ECE of 39.5 kb with a G/C content of 36.4% and a coding density of 54%.

\(^5\) The 44% indicated under “similar to hypotheticals” also includes weak similarities.

\(^6\) C. trachomatis contains one ECE of 7.5 kb.

\(^7\) The genome size is an approximation, since some gaps remain in the sequence. 16,260 C. elegans genes have been reviewed for these data. On average, there are 5 introns per gene, and 27% of the DNA encodes exons.

Information on genome projects can be obtained at:


Genome sequencing centre (C. elegans):  http://genome.wustl.edu/gsc/C_elegans/

The A. thaliana genome is scheduled to be finished in 2000 (Meinke et al., 1998) and the human genome, comprising three billion basepairs, is scheduled to be entirely sequenced by the end of the year 2003 (Collins et al., 1998). Since the D. melanogaster genome will be finished by 2002, and the mouse genome project being started in the near future, the genomes of all important model organisms used in genetic research will be determined in the course of the next decade.

Viral (\& phage), mitochondrial, and plastid genomes will not be discussed here, although over 200 viral, 20 mitochondrial, and 11 plastid genomes have now been determined and published.

Representatives from all three major kingdoms, eukarya, bacteria, and archaea, have now completely been sequenced, presenting unprecedented opportunities for comparative genome analyses. From Table I.1 it is evident that bacterial genomes all have a gene-coding density of about one gene per kilobase of DNA and, except for Rickettsia prowazekii, a coding density of about 85-90%. Another general feature of the sequenced microbial genomes is the finding that about 20-30% of their genes are unique, i.e. the databases do not contain related genes from other organisms (orthologs). The genome of P. horikoshii is an exception to this rule, with almost 60% of its putative protein sequences being unique. This is probably a result of the fact that this species is a hyper-thermophilic archaebacterium, growing optimally at temperatures of nearly 100°C (Kawarabayasi et al., 1998). The other thermophilic archaebacteria, the genomes of which have been determined, M. jannaschii, M. thermoautotrophicum, and A. fulgidus, have much lower optimal growth temperatures, of 85°C, 65°C, and 83°C, respectively. Apparently, with growth temperatures approaching 100°C, this poses unique constraints on protein architecture. Since the fraction of protein sequences without database matches in a genome have not significantly decreased with each new sequenced microorganism, this is likely to indicate that about one quarter of these organisms’ genes are probably specifying the unique character traits typical of that organism.

The genomes of H. influenzae, M. genitalium, M. tuberculosis, M. jannaschii, C. trachomatis, R. prowazekii, T. pallidum, A. aeolicus, H. pylori, M. thermoautotrophicum, A. fulgidus, and B. burgdorferii have been determined by a strategy called whole-genome random
sequencing (see below). The *B. subtilis* genome, however, like the genomes of *P. horikoshii*, *Synechocystis*, *S. cerevisiae*, *E. coli*, has been determined by a directed approach. All research groups involved in the *B. subtilis* genome-sequencing project were assigned a particular region from the genetic map of the chromosome (Anagnostopoulos et al., 1993). Cloning of the assigned region was carried out by various positional cloning methods, including plasmid walking, marker rescue, lambda bank screening, and (inversed) long-range PCR (Cheng et al., 1994).

**Whole-genome random sequencing**

The strategy of choice for the sequencing of an entire genome is presently whole-genome random sequencing. The basic principle behind this method is a modification of the Janus strategy (Burland et al., 1993) and involves the construction of two independent chromosomal DNA banks: a plasmid bank with relatively short inserts and a bank in a phage λ derivative with large inserts. Subsequently, high-throughput DNA sequencing of clones from primarily the first bank is performed, and this is supplemented by sequences obtained from the second, λ bank. The plasmid bank generates the main body of the sequence, while the λ bank is used primarily for controlling the physical integrity of the sequence and as a source of linking clones for the determination of physical gaps in the sequence after assembly of sequences obtained from the plasmid bank. This strategy is particularly powerful when the organism to be sequenced has not been extensively studied, and hence no genetic and physical maps of the chromosome are available at forehand. Most known genomes have been determined by this strategy (Fleischmann et al., 1995).

The essence of this approach consists of several, partially overlapping phases, summarised in Table I.2. In the first phase, BAL 31 nuclease-generated random genomic libraries are constructed in a high-copy number plasmid (e.g. pUC18) for *E. coli*, and in a phage lambda derivative (*e.g.* λ GEM-12 or λ DASH II). The plasmid shotgun bank contains relatively small DNA fragments in the 1.6 to 2.0 kb range, and the lambda genomic library contains fragments in the 15-20 kb range. These are verified for their random coverage of the genome.

The second phase consists of high-throughput DNA sequencing and assembly, primarily of clones from the plasmid library, and complemented by sequences from the lambda library. To estimate the amount of DNA that needs to be sequenced to yield the desired coverage of the genome, the following equation for the Poisson distribution can be used: $P_O = e^{-m}$, where $m$ is the sequence coverage, and $P_O$ is the probability that a base will not be sequenced. Thus, with *B. subtilis* as an example (genome size 4.2 Mb), after sequencing 4.2 Mb from random clones, $P_O = e^{-1} = 0.37$, which means that 37 percent of the genome is expected to be unsequenced. If $L$ is the genome length and $n$ is the number of random sequences generated, the total gap length is $Le^{-m}$, and the average gap size is $L/n$. So, again using *B. subtilis* as example, fivefold coverage (21,000 clones sequenced from both ends with an average of 500
bases), would yield a total gap length of 28,299 bases and an average gap size of about 100 bases.

In the third phase, contigs are ordered and the remaining sequence gaps are closed. This is done by primer walking, primarily from linking clones in the lambda bank.

### Table I.2. Stages of the whole-genome sequencing strategy (adapted from Fleischmann et al., 1995)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random small insert and large insert</td>
<td>Shear genomic DNA randomly to ~2 kb and 15 to 20 kb fragments, respectively</td>
</tr>
<tr>
<td>library construction</td>
<td></td>
</tr>
<tr>
<td>Library plating</td>
<td>Verify random nature of library and maximise random selection of small insert and large insert clones for template production</td>
</tr>
<tr>
<td>High-throughput DNA sequencing</td>
<td>Sequence sufficient number of fragments from both ends for about a 6× coverage of the entire sequence</td>
</tr>
<tr>
<td>Assembly</td>
<td>Assemble random sequence fragments and identify repeat regions</td>
</tr>
<tr>
<td>Gap closure</td>
<td></td>
</tr>
<tr>
<td>Physical gaps</td>
<td>Order all contigs (fingerprints, peptide links, λ clones, PCR) and identify templates for closure</td>
</tr>
<tr>
<td>Sequence gaps</td>
<td>Complete the genome sequence by primer walking</td>
</tr>
<tr>
<td>Editing</td>
<td>Inspect the sequence visually and resolve sequence ambiguities, including frameshifts</td>
</tr>
<tr>
<td>Annotation</td>
<td>Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions)</td>
</tr>
</tbody>
</table>

### Developments in sequencing techniques

The advent of genome sequencing projects has been entirely dependent on recent developments in sequencing techniques; a decade ago such efforts would not have been practically feasible undertakings. Therefore, a short overview will be given on developments from the Maxam & Gilbert (1977) chemical sequencing method to present-day, fully automated systems based on the dideoxy chain-termination –or enzymatic- method of Sanger et al (1977). Both methods produce nested sets of (radioactively) labelled polynucleotides, from 1 to 500 bases long, that begin at a fixed point and terminate at points that depend on the location of a particular base in the original DNA strand. The polynucleotides are then separated by polyacrylamide gel electrophoresis (PAGE), and the order of nucleotides in the original DNA can be read directly from an autoradiograph, or a fluorogram in the case of fluorescent labelling (Griffin & Griffin, 1993).

Cycle sequencing is an adaptation of the Sanger dideoxy method of sequencing. It has the advantages that reactions are simpler to set up, less template is required, and the quality and purity of the template are not as critical as in the standard procedure. In this method, using polymerase chain reaction (PCR) technology, a single primer is used to amplify the region to be sequenced in a linear manner using *Taq* DNA polymerase in the presence of deoxy-nucleotide triphosphates (dNTP’s) and a dideoxy-nucleotide triphosphate (ddNTP).
Automated DNA sequencing has become one of the major advancements in modern biotechnological research. It is based on the Sanger chain termination method of sequencing, but uses fluorescent instead of radioactive labelling techniques. The label is attached either to the sequencing primer, the nucleotides, or the dideoxy nucleotides. During electrophoresis of the DNA fragments in a PAA gel, the fluorescent label is excited by a laser beam and the emitted fluorescence is collected by detectors. The fluorescence signal that is generated is analysed by a computer. Two approaches of this basic concept are presently employed. In the first, one type of label is used for all four sequencing reactions (A, C, G and T), and the reactions are run in separate lanes of the sequencing gel. In the other approach, four different types of label, each having different fluorescence wavelengths, are used in the sequencing reactions and the products of the four reactions are run in one lane on a sequencing gel. Signal detection is then performed at the four different fluorescence wavelengths of the labels.

Today, sequencing systems exist that are fully automated through almost all stages of the process: from toothpicking colonies off a plate, culturing cells, extracting DNA, and sequencing of the template, to the computer-assisted assembly of the raw sequence data. Although the concept itself is not new (Drmanac et al., 1993), sequencing by hybridization (SBH) to an array of oligonucleotides has only recently become feasible with the advent of the DNA chip technology (Drmanac et al., 1998). This methodology is discussed in the paragraph dedicated to applications of the DNA chip technology (see below).

Finding structural features in a DNA sequence

Once a DNA sequence has been completed, the annotation phase begins. The aim of this phase, which consists of a logical order of analyses, is to identify as many as possible primary structural features within the DNA that has been sequenced. This includes identification of open reading frames (ORFs) and verification of codon usage, identification of the start (start codon + ribosomal binding site) and stop sites of ORFs, and analysis of possible terminator structures and promoters. All the analyses discussed below are performed with the aid of computer programs designed for that purpose.

First, ORF searches are performed, and the ORFs that are found are subjected to several filtering procedures that are meant to identify the ones that are likely to constitute protein coding regions. One should keep in mind that all these filtering procedures are imperfect, and also tend to reject several ORFs that are in fact real genes (Bains, 1992). ORFs are first selected on the basis of their length. Usually, a cut-off value of 50 to 100 codons is employed, implying that smaller genes will be missed. Genes encoding RNA species (rRNA, tRNA) are identified by sequence similarity.

The second step in gene annotation is the identification of translational start and stop signals. A typical bacterial translational start site consists of a Shine-Dalgarno (SD) sequence, also called Ribosomal Binding Site (RBS), followed within 4-10 basepairs by one of the start codons ATG, TTG, or GTG. The RBS is recognised by the 16S ribosomal RNA, and should be (partially) complementary to its 3’ end. In B. subtilis, the RBS should preferably contain
part of the sequence 5'-AGAAAGGAGGTGATC-3'. Although ATG (78%), and to a lesser
extend TTG (13%) and GTG (9%) are the main start codons in *B. subtilis*, ATT and CTG
have also been identified as the start codon in a small number of genes (Kunst *et al.*, 1997). An
ORF ends with any of the three stop codons TAA, TGA, or TAG.

Subsequently, the codon usage of the ORFs is compared with the average codon usage
of the organism from which the DNA sequence was obtained. In *B. subtilis*, genes can be
separated into three classes according to codon usage (Kunst *et al.*, 1997). Class I comprises
the majority of the *B. subtilis* genes (82%), including the genes for sporulation. Class II
includes genes that are highly expressed during exponential growth (4.6%), such as those
encoding the transcription and translation machineries, stress proteins, and core intermediary
metabolism. Class III genes (13%) are mainly of unknown function and these genes are
enriched in A + T residues. They are mostly located in or associated with (remnants of)
bacteriophages, transposons, or functions related to the cell envelope. Codon usage tables for
many organisms are obtainable via the www (at URL: http://www.dna.affrc.go.jp/~nakamura/CUTG.html).

The fourth step is the identification of structures involved in transcription termination. A
(rho-independent) terminator is a sequence element that can form a stem-loop structure; it
consists of a short inverted repeat (the stem) separated by a few bases (the loop). Identification
of terminators yields insight into the possible transcriptional organisation of genes in operon
structures.

A further option is the search for promoters for transcription initiation. This is, however,
not an option that is easily performed, since bacterial promoters are too variable in sequence
and spacing for unambiguous identification by simple search-strings. This might still be feasible
if bacteria would not have possessed multiple RNA polymerase sigma factors with different
sequence specificities for their binding. For *E. coli* however, a neural network was trained to
recognise $\sigma^A$-specific promoter sequences.

Finally, possible functions for ORFs can be deduced by performing homology analyses.
When an amino acid sequence displays a high level of similarity to a sequence with known
function from another organism, it is very likely that the putative gene from the organism of
interest performs the same, or a similar function. When no full-length homology to known
protein sequences is observed, identification of functional domains or motifs can be useful in
determining at least part of the function of a given gene. Such domains include, for instance,
ATPase domains characteristic of ABC transporters, helix-turn-helix domains of DNA-binding
domains, signal sequences typical of secreted proteins, transmembrane helices, but also smaller
sequence elements, called motifs, like the Walker A ATP-binding motif. However, this kind of
analysis already resides in the realm of functional genomics, and therefore will be discussed
below.
I.3. Functional genomics

Introduction

The recent explosion in available sequence data has induced the rapid development of a new field of science, now termed genomics. Concomitant with the advent of this research area, the vocabulary of biological terms concerning the classification of proteins has expanded as well. Therefore, a brief overview of this new terminology will be presented here. Like organisms, proteins can be classified on the basis of their relationships. Homologous protein sequences can either be orthologs or paralogs. Orthologs are homologous sequences with a common ancestor, separated by speciation events, which perform the same role in different species. Paralogs are homologous proteins resulting from gene duplications within a species. Normally, orthologs retain the same function in the course of evolution, whereas paralogs may evolve new functions, which may or may not be related to the original one (Tatusov et al., 1997). Thus, paralogs usually perform similar functions, but not necessarily the same (Henikoff et al., 1997). When in different organisms the same function is performed by nonorthologous proteins, this is called a nonorthologous displacement (Mushegian & Koonin, 1996). In the context of protein homologies, the term orphan is also used, to describe a gene product that does not show any similarity to protein sequences from other organisms, or shows similarity only to proteins of unknown function (Dujon, 1998). The complete set of protein sequences encoded by a genome is called the proteome, and the complete set of (possible) RNA transcripts from a genome is called the transcriptome.

Composite proteins consisting of multiple modules, the functional building blocks of proteins, are called chimeras. Modules can contain one or multiple motifs. These smallest units in protein classification are recognised as highly conserved similar regions (amino acids) in alignments, and these often represent an enzyme’s active site residues. They can be as small as, for example, the zinc-finger DNA binding motif C2H2. By virtue of forming an independently folded structure, this motif is the signature of a module. Motifs can either reflect common ancestry or convergence from different origins (e.g. the Walker A ATP-binding motif).

Major challenges in genomics are the systematic analysis of genome function, regulation, and evolution. Several methodologies have been designed to address these goals, separately or in concert, with single genes or genome-wide, and some of these will be reviewed below.

Analysis of (single) genes: mutant construction & phenotype screening
When the genome of an organism has been annotated and the putative genes are known, the real work in a genomic study has just begun. To uncover the function of all (orphan) genes in a genome, a systematic high-throughput approach for mutant construction and phenotype screening is essential. For the genomes of *B. subtilis*, yeast, and *C. elegans*, programs are now underway to generate strain collections in which all genes—especially the orphan genes—are mutated one-by-one in individual mutant strains. Subsequently, the mutants are systematically screened for a phenotype (Dujon, 1998). In this paragraph, only the projects on the functional analysis in *B. subtilis* and yeast will be discussed.

In order to study gene function in *B. subtilis* in a systematic way, a consortium of research groups in the European Union and Japan was set up for this purpose. The goal of this program was to construct mutants of all genes with unknown function, and to assign functions to these genes by systematic and high-throughput phenotype screening. Proteins without any significant similarity to proteins in the public databases, or with similarity to uncharacterised proteins only, are considered unknown in this project. An integrational vector, pMUTin2, was constructed, that can be used for the construction of insertional mutants. Concomitant with the formation of an insertional mutation, a transcriptional fusion of the promoter of the gene of interest to the *lacZ* reporter gene is generated, such that possible downstream genes in operon are placed under the control of the IPTG-inducible $P_{spac}$ promoter (Vagner et al., 1998). Each participant in the consortium is responsible for the construction of an assigned number of mutant strains and the subsequent analysis of these strains with respect to growth and *lacZ* expression in rich and minimal medium. Also, each mutant strain has to be screened for phenotypic effects with respect to a number of characteristics, as listed in Table I.3.

When mutant strains show a phenotype concerning any of these characteristics, these are subsequently analysed in more detail by members of the consortium which are specialised in that particular area of research. All data are entered in a central database called Micado (at http://locus.jouy.inra.fr/cgi-bin/genmic/madbase_home.pl), which for a limited period of time is accessible only to members of the consortium. By the end of 1998, 1035 mutant strains had been deposited in the *B. subtilis* mutant collection.

### Table I.3. Areas of phenotype screening in *Bacillus subtilis* (from Gas et al., 1998)

<table>
<thead>
<tr>
<th>Small and inorganic molecules</th>
<th>Stress and stationary phase</th>
<th>Cell structure and motility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon metabolism</td>
<td>Stress analysis</td>
<td>Cell envelope</td>
</tr>
<tr>
<td>Nitrogen and sulphate</td>
<td>Stationary phase</td>
<td>Motility</td>
</tr>
<tr>
<td>Macromolecules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA</td>
<td>Cell cycle</td>
<td></td>
</tr>
<tr>
<td>RNA</td>
<td>Competence</td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>Sporulation</td>
<td>Germination</td>
</tr>
</tbody>
</table>
The yeast *S. cerevisiae* is being subjected to a similar systematic functional analysis program by a consortium of 134 research groups called EUROFAN (European Functional Analysis Network). In this program, orphan ORFs are completely deleted and replaced by a kanamycin marker cassette. The disruptants, when viable in the haploid state, are then subjected to a first level analysis that includes growth on three basic laboratory media at three different temperatures, mating, and sporulation (Dujon, 1998).

Genomics at the transcriptome level: DNA chip technology

The DNA chip technology provides one of the most powerful genetic research tools that have been developed in the past few years. The technique can be used for many purposes, among which are: genome-wide parallel gene expression studies, identification of regulons and their regulatory motifs, gene discovery studies, polymorphism screening, mapping of DNA clones, and DNA sequencing by hybridization (Ramsay, 1998; Drmanac *et al.*, 1998). Another method for analysis of the transcriptome is Serial Analysis of Gene Expression (SAGE; Velculescu *et al.*, 1995). In this method, very short cDNA sequences from a complete mRNA population are concatamerisated in a cloning vector and subsequently sequenced, providing insight in which genes are transcribed and how abundantly. This technique will not be further discussed here.

The principle of DNA chip technology is the following. An ordered array of oligonucleotides or DNA fragments, for instance representing all genes from a genome, is synthesised and immobilised on a solid base. Subsequently, this array is exposed to a fluorescently labelled probe, or a set of differently labelled probes, and fluorescence is monitored at each position of the array. Two basic variants of chip manufactory exist today: delivery or synthesis. In the delivery variant, prefabricated oligonucleotides, or PCR-generated DNA fragments, are immobilised in an ordered array on a solid surface. In the synthesis variant, an array of oligonucleotides is synthesised *in situ* through a combination of photolithography and oligonucleotide chemistry (Ramsay, 1998; Marshall & Hodgson, 1998; Schena *et al.*, 1998).

One important application of this technology is genome-wide simultaneous gene expression monitoring. Here, a chip containing oligonucleotides or complementary DNA fragments, each uniquely complementary to a gene from the genome under investigation, is used. The DNA on this chip is then hybridised with total mRNA from a culture grown in a certain medium or under specific conditions. This analysis yields insights in the transcriptional program used under the particular growth condition, and it can also serve as a reference transcriptome. Subsequently, the same experiment can be performed with mRNA isolated from a culture that is subjected to different growth conditions or medium components. When the transcriptional programs from these two experiments are compared, this identifies genes that are specifically induced or repressed under the conditions tested. The sensitivity of the method, as tested with human, yeast, and bacterial cells, is such that between 0.1 to 5 mRNA copies per cell can be detected (Lockhart *et al.*, 1996; Wodicka *et al.*, 1997; de Saizieu *et al.*, 1998). The
DNA chip technology has been used successfully in several instances. In budding yeast it has been applied to identify genes involved in the transcriptional program of sporulation and, concomitantly, regulator-binding consensus sequences upstream of these genes (Chu et al., 1998). Roth and co-workers (1998) have shown in yeast, using the galactose response, heat shock, and mating type as examples, that alignment of upstream sequences of genes that are expressed in concert readily reveals the consensus regulatory motif of the respective regulon. Also in yeast, genes involved in growth in rich and minimal media were identified (Wodicka et al., 1997), as well as the temporal program involved in the metabolic shift from fermentation to respiration (DeRisi et al., 1997). In this latter paper, another application of the method, the identification of target genes of transcriptional regulators, was also presented. In this case, targets of regulators (TUP1 and YAP1) involved in the shift from fermentation to respiration were identified.

To investigate DNA sequence differences (polymorphisms), two approaches can be used. In the first approach, sequencing by hybridization (SBH), an array is used consisting of a complete set of noncomplementary oligonucleotides. With seven-mers, for instance, 8192 oligonucleotides are necessary in the array. Hybridization of an unknown DNA fragment to this array yields its sequence. Using this method, a number of polymorphisms in a 1.1 kb DNA fragment carrying the human the p53 gene was determined (Drmanac et al., 1998). In the other approach, an array of oligonucleotides designed to match specific sequences is used. Sequence polymorphisms in human mitochondrial DNA were identified successfully with this approach (Chee et al., 1996).

**Genomics on the proteome level: 2D gelelectrophoresis and the two-hybrid system**

Two methods exist to perform genome analysis at the protein level; these are two-dimensional (2D) gel electrophoresis and the yeast two-hybrid system.

2D gel electrophoresis is a method that separates proteins in two dimensions. In the first dimension, proteins are separated on the basis of their charge, or isoelectric point, while they are separated on the basis of their mass in the second dimension. This is -in theory- a powerful tool for genome analysis at the protein level. It is, for instance, possible to study changes in protein levels in response to changes in growth conditions, or the effects of mutations in regulators. Progress in this field is, however, hindered by two limitations of the technique. First, only a limited number of protein spots is visible on a 2D gel (around 500 for B. subtilis and 1500 for yeast). Second, the number of identified protein spots is only slowly accumulating, since identification has to be done through time-consuming methods such as microsequencing and mass spectroscopy. In B. subtilis, attempts have been made to construct such a map, or 2D protein index (Schmid et al., 1997). Also, characterisation of protein changes in response to heat shock, salt- and ethanol stress, is well underway (Bernhardt et al., 1997). Proteome analysis through 2D gel electrophoresis is also employed in the yeast functional genomics program (Dujon, 1998).
The two-hybrid system is a genetic method that uses transcriptional activity as a measure of protein-protein interaction. It relies on the modular nature, characteristic of many site-specific transcriptional activators, which consist of a DNA-binding domain and a transcriptional activator domain. The DNA-binding domain targets the activator domain to the specific genes that will be expressed, and the activation domain contacts other proteins of the transcriptional machinery to enable transcription to occur. The two-hybrid system is based on the observation that the two domains of the activator, which is the yeast Gal4 protein in most applications, need not be covalently linked and can be brought together by the interaction of any two proteins. It requires that two hybrids are constructed: a DNA-binding domain fused to some protein, X, and a transcription activation domain fused to some protein Y. These two hybrids are expressed in a cell containing a reporter gene with an upstream consensus sequence for the DNA-binding domain. If the proteins X and Y interact, they create a functional activator, the activity of which can be detected through the expression of the reporter gene (Phizicky & Fields, 1995). In yeast, surveys of protein-protein interactions are already in progress, either using pre-mRNA splicing factors as initial bait (Dujon, 1998), or genome-wide, investigating all 18 million pair-wise combinations of the approximately 6,000 predicted proteins of yeast (Hieter & Boguski, 1997).

I.4. In silico genomics

The availability of complete genome sequences makes it possible to investigate the evolutionary relationships between genomes and gene families within a genome. Several aspects can be addressed with in silico genomics, at different levels of hierarchy: at the DNA level, the protein level, and level of the organism. A number of examples will be presented in the following paragraph.

In silico DNA analysis

When investigating genome features at the DNA level, searches for evidence of non-randomness of (oligo)nucleotides (words) in a DNA sequence can be made. This non-randomness is then analysed in a realistic model of the genome. The rationale for this is that words that are over- or under-represented in a sequence, in contrast to a model, may indicate a phenomenon of positive or negative selection (Rocha et al., 1998). For example, when investigating oligonucleotide frequencies in a genome, a period of three will be observed. This period is known to be related to the codon size of three, and in a realistic model, therefore, this period of three should be incorporated in the analysis. The same applies to frequencies of nucleotides (the G+C content) and all other known sequence elements that result in non-randomness of the nucleotide composition.

Analysis of codon usage in E. coli and B. subtilis has revealed the existence of three classes of genes, also distinguishable by their biological properties. The majority of the genes fall into Class I, representing genes that are expressed rarely and/or genes expressed at a low level. Class II consists of genes expressed continuously at a high level during exponential
growth. Class III consists of genes corresponding to surface elements of the cell, genes from mobile elements, as well as genes resulting in a high fidelity of DNA replication. It was suggested that class III genes are inherited through horizontal transfer (Médigue et al., 1991; Kunst et al., 1997).

Rocha et al. (1998) have observed biased word usage in the genome of *B. subtilis* with words of up to eight letters long. Biases in trinucleotide frequencies were mainly coupled to codon usage. Biased words of size seven are probably related to interaction with RNA or DNA polymerase, and avoidance of palindromic sequences would be the result of avoidance of restriction sites, since restriction/modification systems are easily horizontally transferred. In *E. coli*, codon usage of genes for major structural components of the outer membrane, porins and lipopolysaccharides, indicate that these genes might be obtained through horizontal gene transfer (Guerdoux-Jamet et al., 1997).

The distribution of the methylation motif GATC in the *E. coli* genome, which is known to be also involved in the long-patch mismatch repair system, revealed regions which are abnormally rich in this motif. These GATC clusters were mainly located within genes involved in the shift from anaerobic to aerobic growth. The GATC clusters probably represent some kind of transcriptional regulation process accompanying the shift of the bacterium from the host environment (high temperature, lack of oxygen, high osmolarity) to the external medium (low temperature, presence of oxygen, low osmolarity). A possible mechanism was presented which includes the fact that methylation lowers the $T_m$ of the DNA helix drastically. Also, a strong bias was observed against GATC motif pairs separated by 6 bases. This under-representation could be explained, since CAP binding sites (TGTGATCTAGATCACA) and FNR binding sites contain this double motif (Hénaut et al., 1996).

**In silico protein analysis**

At the protein level, paralogous and orthologous relationships can be investigated. Although the definitions of orthologs and paralogs are clear, in practice several problems arise. First, what is the level of amino acid similarity that makes two proteins paralogous to each other? Related to this problem is the question whether paralogs should display similarity over the entire length of their sequence. This is especially problematic in the case of chimeras, composite proteins consisting of multiple domains. It is well known that two separate proteins from one organism can be conserved in another bacterium, where they may be translationally fused to form one protein (Henikoff et al., 1997). Although these proteins obviously have a paralogous relationship from a functional point of view, they will not be designated as such in most studies. Also, proteins can be categorised as paralogs on the basis of their amino acid similarity while they are not functional paralogs. Even identical proteins can, conceivably, be functionally non-paralogous because of differences in temporal and/or spatial expression patterns. A last remark should be made in this respect. Protein sequences from different organisms, though similar, may not be orthologs. This is the case when multiple paralogs are present in one organisms’ genome as possible candidates for being the ortholog to a protein.
from another organism. In general, though this is not a very reliable rule when investigating species that are phylogenetically very divergent, the pair of sequences displaying the highest similarity to each other are considered orthologs (Tatusov et al., 1997). In genomes, the size of gene families is related to genome size (Huynen & Nimwegen, 1998). In a genome-wide analysis of orthologous and paralogous relationships, Tatusov and co-workers (1997) have identified a total of 720 clusters of orthologous groups (COGs), in which 37% of all the genes from the seven analysed genomes could be classified.

A similar analysis, although on the level of protein structure instead of protein sequence, was performed by Gerstein & Hegyi (1998). These researchers observed that all known protein structures could be classified into a very limited number, currently about 350, of folding patterns. Probably, all proteins occurring in nature are composed of no more than around 1000 folds.

Wallin & Von Heijne (1998) have investigated the occurrence of transmembrane segments in fourteen completed genomes. Twenty to thirty percent of all ORFs were predicted to encode membrane proteins, and this number increases linearly with greater genome sizes. The widespread assumption that organisms have a preference for membrane proteins with 6 or 12 transmembrane segments, was invalidated to some extent, since this preference was observed in some genomes (E. coli, B. subtilis, A. fulgidus, H. influenzae, H. pylori, M. genitalium and Synechocystis), but not in others (S. cerevisiae, C. elegans, H. sapiens, M. thermoautotrophicum, M. jannaschii, C. acetobutylicum).

**In silico analysis of organisms**

At the organism level, several aspects can be studied. For instance, Rosa & Labedan, (1998) have analysed the evolutionary relationship between the bacteria E. coli and H. influenzae and deduced a putative last common ancestor. The approach used here was an exhaustive analysis of homologous proteins encoded by genes in both genomes through comparison of the evolutionary distances between orthologs and paralogs. Significant similarities were observed between 1,345 H. influenzae proteins and 3,058 E. coli proteins, many of them belonging to families of various sizes. In other studies, the minimal gene complement for self-sufficient cellular life was deduced from the genomes of two representatives of ancient bacterial lineages, the bacterium M. genitalium and the Gram-negative bacterium H. influenzae (Mushegian & Koonin, 1996; Koonin & Mushegian, 1996). Only 240 orthologous proteins were found to be encoded by both genomes, and these were assumed to be probably essential for cellular function. This set was complemented with nonorthologous displacements, and removing apparent functional redundancies and parasite-specific genes narrowed the resulting set further down. This yielded a minimal gene complement of only 256 genes. These analyses can be facilitated by a visual representation method called differential genome display, which enables the rapid identification of special features of organisms, such as virulence factors (Huynen, 1997).
An important goal in this type of analyses is, of course, to resolve the ancestry of all life. Attempts in this direction have also been made (Mushegian & Koonin, 1996; Koonin & Mushegian, 1996). Since the minimal gene complement does not contain eukaryal or archaeal homologs of the key proteins of bacterial DNA replication, it seems likely that the last common ancestor of the three primary kingdoms had an RNA genome. Furthermore, as archaea have significant deviations in the enzymology of the upstream reactions of glycolysis, the ancestor may have had a metabolism based on trioses and pentoses as energy source through the conversion to glyceraldehyde 3-phosphate.

I.5. Outline of this thesis

The central question that will be addressed in this thesis is what the scientific spin-off is that can obtained from the vast body of information that genome sequencing projects generate. The \textit{B. subtilis} genomic sequence, part of which was determined in our group, is used as an example to illustrate this general theme.

In chapter one, an overview will be given of strategies and methods that are employed in genome sequencing projects. The central question here is: what is the value of this giant investment of time, money, and human recourses in these seemingly boring and certainly tedious projects. An outline will be presented on how \textit{in silico} analyses can be performed on raw sequence data that are no more than a silent string of letters at the start, but notwithstanding this, define the life of its owner. These analyses can be carried out to address several questions. What can be said about the function of a putative gene, solely by analysing the similarity of its deduced amino acid sequence to other proteins and investigating the presence of motifs, localisation signals and other features? What is its relationship to other proteins in the same and other genomes? What are the implications with respect to insights in the evolution of species, protein paralog and ortholog families, and the biological relevance of these?

In chapter two, the cloning, sequence, and annotation of the \textit{B. subtilis} DNA region that was determined in our group, is presented. This is the chromosomal region located between the \textit{prkA} and \textit{addAB} markers, a DNA fragment of 171,812 bps. This region was obtained through various cloning methods, including lambda bank screens, plasmid rescue, and (inversed) long-range PCR. The \textit{in silico} DNA analyses in this chapter are restricted to gene annotation, terminator searches, and searches for homologs in public databases. Because large differences were observed with respect to existing data, a correction of the original physical and genetic maps of the \textit{prkA} to \textit{addAB} region, originally published by Itaya & Tanaka (1991) and Anagnostopoulos \textit{et al.} (1993), will also be presented.

Chapter three deals with the results of further analyses on the protein sequences encoded by genes in the \textit{prkA} to \textit{addAB} region. These include: paralog frequency analysis, paralog positional analysis, analysis of compartmentalisation signals (membrane-spanning domains, signal sequences, and lipomodification signals), and the identification of several dysfunctional genes, including a remnant of a gene for anaerobic coproporphyrinogen III oxidase (\textit{hemN}-like
gene), and a dysfunctional ABC-type transporter. The chapters II and III have been published, albeit in a somewhat different form (Noback et al., 1996; Noback et al., 1998).

In chapter four, the complete genome sequence of the *B. subtilis* genome is presented. This chapter has already been published (Kunst et al., 1998).

Chapter five deals with the positional analysis of amino acid frequencies in fifteen known genomes. Comparison of average amino acid frequencies in the entire proteome with the average frequencies in the N- and C-terminal amino acids of these proteomes has revealed biases of many amino acids. Further investigation of the deduced (average) properties of the proteomes with respect to charge and hydrophobicity, showed that all proteomes display similar differences between the N- and C-termini with respect to these parameters. This could reflect that the N- and C-termini of proteins are usually located at the surface of proteins, and might be involved in the proper translocation of the nascent proteins through the ribosome.

In chapter six, the identification of the *gtaC* gene is presented. An ORF was found in the *prkA-addAB* region, *yhxB*, that displayed high similarity to phosphogluco- and phosphomanno-mutases. It was demonstrated that this ORF corresponds to the genetic marker *gtaC*, and is responsible for the glucosylation of cell wall teichoic acid and phage susceptibility in *B. subtilis*.

In chapter seven, functional analyses of the ubiquitous *hit* gene are presented. This ORF has highly conserved orthologs in (probably) all living organisms, including the organism with the smallest known genome, *M. genitalium*. Surprisingly though, the *hit* gene was not essential in *B. subtilis*, as shown by the viability of a strain with an insertional mutation of this gene. The insertional mutation of the *hit* gene induced a sensitive phenotype in heat-shock treatment. The Hit protein was found to have a twofold biochemical activity on ADP, the relative amounts of products of the Hit-mediated reaction being dependent on the pH. Hit hydrolyses ADP to AMP and Pi, and also acts as phosphotransferase in the reaction 2 ADP → ATP + AMP. The isolation of a transcriptional regulator of *hit*, the -also ubiquitous- *yabJ* gene, is also presented.

In chapter eight, the characterisation of a new forespore-specific gene of *B. subtilis*, *yhcN* is presented. Based on conserved amino acid sequence elements, specific for small acid soluble proteins (SASP; KLEVADE) and membrane anchored lipoproteins (LMTGC), this ORF was considered a likely candidate for involvement in spore formation. This assumption was shown to be correct; *yhcN* expression is dependent on the forespore-specific sigma factor \( \sigma^G \), it is expressed at a very high level in the forespore and is located in the inner spore membrane. *YhcN* mutant spores show a phenotype of slower outgrowth than wild-type spores. This work has been published (Bagyan et al., 1997).

Chapter nine is a typical example of paralog research. The *B. subtilis* genome encodes four closely related Type I signal peptidases, responsible for the removal of signal peptides from secretory precursor proteins, with overlapping substrate specificities and differing expression patterns. This work has been published (Tjalsma et al., 1997).
References


