FEBS Letters 430 (1998) 28–36 FEBS 20384

Minireview

The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis

Ivan Moszer*

Unité de Régulation de l'Expression Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Received 8 May 1998

Abstract The completion of the entire 4.2-Mb genome sequence of the Gram-positive bacterium *Bacillus subtilis* has been a milestone for biological studies on this model organism. This paper describes bioinformatics work related to this joint European and Japanese project: methods and strategies for gene annotation and detection of sequencing errors, using an integrated cooperative computer environment (Imagene); construction of a specialized database for data management and a WWW server for data retrieval (SubtiList); DNA sequence analysis, yielding striking results on oligonucleotide bias, repeated sequences, and codon usage, all landmarks of evolutionary events shaping the *B. subtilis* genome.

Key words: Complete genome; Sequence annotation; Genomic database; DNA sequence analysis; *Bacillus subtilis*

1. Introduction

Complete genome sequencing of microbial organisms has become almost routine. By May 1st, 1998, thirteen complete bacterial genomes were listed in the 'TIGR Microbial Database' (http://www.tigr.org/tdb/mdb/mdb.html), which provides regularly updated lists of genome sequencing projects (see also http://www-c.mcs.anl.gov/home/gaasterl/genomes.html). The analysis of genome sequence data should yield major insights into the molecular biology of the organism concerned. Genomic information not only provides the chemical definition of nucleic acids and proteins, it may also result in more in-depth knowledge of major processes such as replication, metabolism, gene expression and evolution. This requires that the overall properties of genes and associated regulatory networks can be deciphered, notably through computer analysis, and the combination of in silico results with data from in vivo and in vitro experiments.

Bacillus subtilis is a model Gram-positive bacterium, and its biochemistry, physiology and genetics have been intensely studied for forty years [1,2]. This aerobic, endospore-forming rod-shaped bacterium, commonly found in soil and in association with plants, has remarkable genetic and physiological features: the ability to differentiate into a resistant spore in unfavorable external media [3,4], and the aptitude for DNA-mediated transformation and homologous recombination [5], both coupled to complex networks of regulatory pathways and developmental checkpoints [6]. It has a compact genome with an intermediate G+C content, for which detailed genetic and physical maps are available [7,8]. Furthermore, *B. subtilis* is amenable to laboratory culture and reverse genetics for functional analysis, and it is an interesting model system for numerous industrial, medical, and ecological applications, notably because of its secretion ability.

Most of the complete microbial genome sequences available were determined by a single laboratory or institute. In contrast, as for the *Saccharomyces cerevisiae* genome [9], the *B. subtilis* genome sequencing project followed a network approach, involving 35 laboratories, from 1989 to 1998 [10–12]: 25 laboratories supported by the European Commission (Science and Biotechnology programs) [13,14], 7 laboratories funded by the Human Genome Project in Japan [15,16], one Korean laboratory, and two biotechnology companies. The completion of the entire sequence of the *B. subtilis* genome (reference strain 168) was officially announced during the Ninth International Conference on *Bacilli* in Lausanne on July 19th, 1997, and the sequence and annotations were released in international databanks on November 20th, 1997 (accession number AL009126) [17].

Bacillus subtilis has a single chromosome, 4 214 814 bp long [17]. Table 1 summarizes its main features. This minireview, rather than providing extensive information about the *B. subtilis* genome and proteome, is focused on bioinformatics activities aimed at the annotation, data handling and analysis of the DNA sequence, core material for genome maintenance and expression.

2. Annotation of the *B. subtilis* genome sequence: strategies, methods and results

2.1. An integrated computer environment for sequence annotation

An integrated cooperative computer environment, Imagene, was used for sequence annotation as part of the *B. subtilis* genome project [18,19]. This system provides an elaborate representation scheme of biological and methodological knowledge, and several graphical interfaces allowing one to run tasks in a cooperative way, and to interact with the system in order to verify, recombine or refute the information generated.

Imagene makes it possible to store and manipulate the sequences themselves, and the data resulting from the application of methods. Entities are internally represented in a uniform way through an object-oriented model. A data manager is used to edit and query the objects associated with the data. Imagene also provides an expandable set of sequence analysis methods, acting at three different levels: 'modules' are proce-

*Fax: +33 (0)1 45 68 89 48.
E-mail: moszer@pasteur.fr

PII: S0014-5793(98)00620-6

dures, either written in low-level programming languages or external executables; 'tasks' are ordered combinations of modules; and 'strategies' are ordered combinations of tasks. A graphical task manager allows the user to perform many operations, without having to program in a low-level language: to chain modules and tasks, to modify the characteristics of a task and re-execute part of a strategy at any time during the solving process, and to visualize successive versions of an execution. Finally, a generic cartographic interface gives a graphical representation of a set of objects (features) associated with a sequence. Advanced functions are available, such as synchronized scrolling of several maps with different scales and units (e.g. genetic and physical maps). Thus a common interface can be used to superimpose and compare the results of several independent analysis tasks, making the final biological evaluation more efficient. An overview of the architecture of Imagene and more details about its precise implementation are given in reference [19].

2.2. Criteria for the annotation of coding regions

Each group involved in the genome sequencing project chose its own strategies and methods so as to annotate as accurately as possible the chromosome region it had sequenced. The annotation of the whole genome sequence was then homogenized using Imagene.

2.2.1. Protein genes. Although searching for protein coding sequences (CDS) in prokaryotic genomes is often considered to be easy (high coding density, no intron), the detection of small genes and the accurate determination of translation initiation codons is not necessarily so simple. In *B. subtilis*, this latter difficulty was partly smoothed away, because ribosome binding sites (RBS) are highly similar to the 'consensus' sequence (complementary sequence of the 16S rRNA 3′-end: 5′-AAGGAGGTG-3′), probably due to the lack of S1 protein in *B. subtilis* ribosomes [20]. The presence of a canonical RBS sequence 5–15 bp upstream from a potential start codon was therefore strong evidence for there being a coding sequence in the adjacent downstream reading frame. Among the numerous methods which have been designed for searching for bacterial genes (e.g. nucleotide composition bias, codon usage, Fourier analysis), the GeneMark method [21], based on periodic Markov chain analysis, was the most commonly used procedure for efficiently predicting genes in *B. subtilis*. The Glimmer program [22], based on interpolated Markov models, has also been used recently, yielding similar results. Finally, sequence similarity searches against protein databanks, typically through BLASTX analysis [23], made it possible to identify potential CDSs. It should be noted that, in the final round of annotation of the *B. subtilis* genome, the minimal size of CDSs was not a decision criterion, so that small genes, a number of which had already been identified and their function precisely described, did not escape from this analysis.

Table 1
General features of the *B. subtilis* genome

Genome size	**4,214,814 bp** (*oriC*: 0 kb, *terC*: 2,017 kb)
European sequencing	2,677 kb (64%)
Japanese sequencing	1,368 kb (32%) (17 kb overlap)
Piecemeal sequencing	186 kb (4%)

Base composition (%)

	Genome	Leading strand	Protein genes	RNA genes	Intergenic regions	Prophages
A	28.2	29.4	29.9	25.0	31.1	30.8
C	21.8	20.1	20.3	23.5	18.9	19.2
G	21.7	23.5	24.1	31.4	18.5	18.1
T	28.3	27.0	25.7	20.1	31.5	31.9

=> whole genome: 43.5% G+C

Gene counts

Protein genes	4,100 (2,852 'y' genes, see similarities in Table 2)
rRNAs	10 operons (16S-23S-5S)
tRNAs	88 (12 operons, 8 single genes)
Other stable RNAs	3 (10S, 4.5S, RNase P)

Coding capacity (kb)

Protein genes	3,646.2 (86.5%)
Stable RNAs	53.8 (1.3%)
=> overall coding	3,700.0 (87.8%)
Intergenic regions	514.8 (12.2%)

Orientation of transcription (CDS)

	Clockwise	Leading strand
+	1,937 (47.2%)	3,030 (73.9%)
-	2,163 (52.8%)	1,070 (26.1%)

Lengths (bp)

	Mean	Min	Max
Protein genes	891	63	14,790
Intergenic regions	137	<0	2,191

=> density: 1 gene / 1.03 kb

Start and stop codons

Start		Stop	
ATG	3,185 (77.7%)	TAA	2,545 (62.1%)
TTG	513 (12.5%)	TGA	964 (23.5%)
GTG	387 (9.4%)	TAG	591 (14.4%)
CTG	8 (0.2%)		
ATT	6 (0.1%)		

Codon usage (CDS)

Class 1	3,375
Class 2 (high expression)	188
Class 3 (prophages)	537

Prophage(-like) regions

Name	Coordinates (bp)	Length (kb)
1	202,000-220,000	18
2	529,000-570,000	41
3	652,000-664,500	12.5
4	1,263,000-1,279,000	16
PBSX	1,320,000-1,348,000	28
5	1,879,000-1,900,000	21
6	2,046,000-2,078,000	32
SPß	2,151,274-2,285,689	134.416
skin	2,652,598-2,700,635	48.038
7	2,707,000-2,750,000	43

=> 9.3% of the genome

Non-strictly repeated DNA sequences longer than 100 bp (RNA, *srf*, *pks*, *pps* and prophage genes excluded)

Length (bp)	Coordinates (bp)	Genes
748	3,664,916 and 3,671,947	*gtaB* and *tagH*
410	4,101,919 and 4,102,329	*yxaL* and *yxaK*
237	4,188,553 and 4,190,346	*yyaO* and *yyaL*
174	526,346 and 526,520	*ydcI*
118	4,095,842 and 4,095,960	*yxbB* and *yxbC*

See positions of the 190 bp non-coding element in Figure 2

Four thousand one hundred (4100) protein genes were predicted in the *B. subtilis* genome using these criteria [17].

2.2.2. RNA genes. The ribosomal operons had already been sequenced at the start of the genome project. There are ten, mostly located very close to the origin of replication, each comprising three genes (16S, 23S, 5S). The situation is slightly different for tRNA genes: they are small and may exist as isolated genes as well as operons, so it was necessary to specifically search for these kinds of coding regions throughout the entire genome sequence. Two programs were used: tRNAscan, which relies on the detection of invariant or semi-invariant bases and potential base-pairing structures consistent with the cloverleaf secondary structure of typical tRNA sequences [24]; Palingol, which is a declarative programming language designed to describe RNA secondary structures, taking into account both sequence and structural criteria [25]. Eighty-eight (88) tRNA genes were identified in this way, clustered into 8 single genes plus 12 operons, 7 located between, adjacent to, or even within rRNA operons [17].

2.2.3. Prediction of potential operons. Operons are a major feature of gene organization in bacteria because they play an

Table 2
Functional classification of *B. subtilis* protein coding genes

1	**Cell envelope and cellular processes**			**867 (21%)**
	1.1	Cell wall		93
	1.2	Transport/binding proteins and lipoproteins		381
	1.3	Sensors (signal transduction)		38
	1.4	Membrane bioenergetics (electron transport chain and ATP synthase)		78
	1.5	Mobility and chemotaxis		55
	1.6	Protein secretion		18
	1.7	Cell division		21
	1.8	Sporulation		139
	1.9	Germination		23
	1.10	Transformation/competence		21
2	**Intermediary metabolism**			**742 (18%)**
	2.1	Metabolism of carbohydrates and related molecules		261
		2.1.1	Specific pathways	214
		2.1.2	Main glycolytic pathways	28
		2.1.3	TCA cycle	19
	2.2	Metabolism of amino acids and related molecules		205
	2.3	Metabolism of nucleotides and nucleic acids		83
	2.4	Metabolism of lipids		77
	2.5	Metabolism of coenzymes and prosthetic groups		99
	2.6	Metabolism of phosphate		9
	2.7	Metabolism of sulfur		8
3	**Information pathways**			**482 (12%)**
	3.1	DNA replication		22
	3.2	DNA restriction/modification and repair		39
	3.3	DNA recombination		17
	3.4	DNA packaging and segregation		10
	3.5	RNA synthesis		244
		3.5.1	Initiation	19
		3.5.2	Regulation	213
		3.5.3	Elongation	8
		3.5.4	Termination	4
	3.6	RNA modification		19
	3.7	Protein synthesis		96
		3.7.1	Ribosomal proteins	56
		3.7.2	Aminoacyl-tRNA synthetases	25
		3.7.3	Initiation	6
		3.7.4	Elongation	6
		3.7.5	Termination	3
	3.8	Protein modification		27
	3.9	Protein folding		8
4	**Other functions**			**289 (7%)**
	4.1	Adaptation to atypical conditions		72
	4.2	Detoxification		68
	4.3	Antibiotic production		30
	4.4	Phage-related functions		83
	4.5	Transposon and IS		10
	4.6	Miscellaneous		26
5	**Similar to unknown proteins**			**667 (16%)**
	5.1	From *B. subtilis*		177
	5.2	From other organisms		490
6	**No similarity**			**1053 (26%)**

essential role in the coordination of gene expression. The complexity of the very nature of promoters and the large number of associated regulatory proteins currently prevent efficient searches for these biological signals (consensus sequences alone do not contain enough information). The existence of multiple sigma factors in *B. subtilis* (at least 14) made this problem even more difficult. In contrast, Rho-independent transcription terminators were accurately predicted, using the Petrin program, which is based on the ratio between the free energy of formation and the length of the hairpin structure, plotted against the number of bases in the adjacent U-run (calculated with a decreasing weight positional function) [26]. About 1300 putative operons were predicted in this way (probably a slight underestimate), with a mean size of three genes per operon [17].

Overall 88% of the *B. subtilis* genome was predicted to be either translated into protein or transcribed into stable structural RNA (Table 1). A few unexpectedly long regions were found to have no detectable feature [17]. These 'grey-holes' should therefore be priority targets for searching for new DNA signals, or unknown structural RNAs.

2.3. Annotation of protein similarities

The function of about 1250 *B. subtilis* genes was already known, at least via a phenotypic observation or as part of a known operon: these genes were assigned names according to their function. The other genes were given names beginning with the letter '*y*', indicating that their function was unknown. Typically, one wanted to know whether the proteins encoded by these genes were similar to other proteins available in sequence databanks. The raw output of similarity search programs, useful for the end-user interested in a given protein, is available through the SubtiList database (see Section 3). Independently, semi-automatic parsing of this output was performed for more in-depth analysis on a large set of proteins. The BLAST2P program [23] has been run for all *B. subtilis* predicted protein sequences against a databank comprising entries from SWISS-PROT, which has high quality annotations, and TREMBL, which is exhaustive and has identical flat-file format [27]. Output was filtered using automatic combinations of thresholds of BLAST2P *p*-value probability and percentage identity between the query and databank sequences. The results were then processed to obtain further information from SWISS-PROT (features, references). Finally, this information was integrated into a relational database, facilitating more efficient browsing of the results to manually assess every decision.

All *B. subtilis* protein sequences were also compared with themselves, using another statistical method [28]. This yielded many classes of paralogs, large ones (77 genes) up to gene doublets, whereas half of the genome was made up of genes encoding proteins with no apparent paralog. In addition, several protein duplications were detected in genes located very close to each other, up to an entire operon which was duplicated 3 kb away from the original, with 80% amino acid identity for the resulting proteins [17].

2.4. Functional classification of protein gene products

The functional classification devised for *B. subtilis* proteins (Table 2) was based on the one hand on the distinction between the machinery of the chemical reactions taking place in the cell (part 2 of the classification) and the information that generates this machinery (part 3 of the classification), and on the other hand on the fundamental notion of compartmentalization that prevails in the living world (part 1 of the classification) [14,17]. As for most of the other microorganism genomes sequenced to date, a significant fraction of the proteins did not match with any other proteins, or only matched with proteins of unknown function, thus giving no clues in the hunt for their putative functions. More information on the classification of predicted gene products is available in reference [17].

2.5. Detection of sequencing errors

Several strategies have been developed using Imagene to detect sequencing errors in the *B. subtilis* genome sequence (C. Médigue, personal communication). One method made use of the results of a BLAST2X scan [23] performed on the entire genome sequence against a non-redundant protein databank. Frameshift errors were shown as two protein similarity predictions for the same databank entry, but corresponding to two adjacent *B. subtilis* CDSs in different reading frames. However, this procedure gave no clues if no significant similarity was detectable. Therefore, a second strategy was used, based on GeneMark predictions [21] associated with RBS-CDS predictions (see above): if the boundaries of a predicted RBS-CDS and the corresponding GeneMark curve did not coincide, then it was assumed that a frameshift error generated such an atypical figure. Finally, the strong prediction of a Rho-independent transcription terminator in an intergenic region, possibly too far downstream from a gene, may indicate a sequencing error shortening this gene, or possibly a small gene missing in the annotations.

After a first screening of the complete genome sequence using these various strategies, about 500 fragments have been re-sequenced, 50% of which were found to be identical to the original, whereas the others contained insertions-deletions or substitutions (C. Médigue, personal communication). Interestingly, several 'possible' frameshift errors, called 'authentic frameshifts', were shown to occur on the chromosome, indicating that the putative genes were either non-functional or subject to regulation processes such as programmed translational frameshifts [17]. New methods should also be devised to detect sequencing errors other than those involving frameshifts.

3. Whole genome data handling: the relational database and WWW server SubtiList

3.1. A relational database structure for complete bacterial genomes

Sequence data alone do not provide all the information required to interpret the messages encoded in the genome sequence. Other kinds of knowledge (e.g. biochemical, physiological and genetic), usually available from various sources, are of fundamental importance, and should thus be organized in combination with sequence data. The specialized SubtiList database was constructed to fulfill these requirements for the *B. subtilis* genome [29]. This database uses a relational data scheme, in which entities to be represented are grouped together into tables, linked by logical relationships. Thus, complex connections can be made between different types of data, and well designed internal structures facilitate complex

queries. The first versions of SubtiList, which has been available since 1994, were based on the concept of contigs, sets of non-redundant sequences gathered from *B. subtilis* databank entries, or large stretches of sequence produced by the genome project. This acted as a natural partition of the genome, facilitating its manipulation and visualization. A new relational structure has been set up for the complete chromosome, which internally parcels out the genome so as to implement more efficient procedures for query, management and update. All genomic features are located relative to these chromosomal sub-divisions, which are in turn mapped to the whole chromosome. Thus sequence and feature updates do not require reevaluation of the whole set of objects in the database, only of the chromosome fragment concerned. This artificial division of the chromosome is not visible to the user, who benefits from an adapted interface for browsing and querying the genome data (see below). Procedures have been written to facilitate the use of this new structure for genome projects currently underway (i.e. with split contigs).

3.2. An adapted WWW interface for bacterial genomes

SubtiList is accessible on the Internet through a WWW server at the URL http://www.pasteur.fr/Bio/SubtiList.html, where it has been implemented on a UNIX system using the Sybase relational database management system. The SubtiList interface enables the user to browse the *B. subtilis* genome according to various criteria, and to perform multicriteria requests in order to retrieve specific information. This interface involves the division of the screen into three frames (Fig. 1): frame 1 contains the controls required to perform common queries at any time (gene name, chromosome region, keyword, sequence analysis, etc.); frame 2 presents a list of genes generated from a user query, or a graphical representation of the chromosome region selected, as applicable; frame 3 provides more detail about one particular gene selected from the previous frame, such as mapping information, functional classification, or links with other databanks. Hyper-text links thus facilitate the retrieval of relevant information about *B. subtilis* gene and protein sequences, notably from the EMBL (EBI server), SWISS-PROT and ENZYME databanks (ExPASy

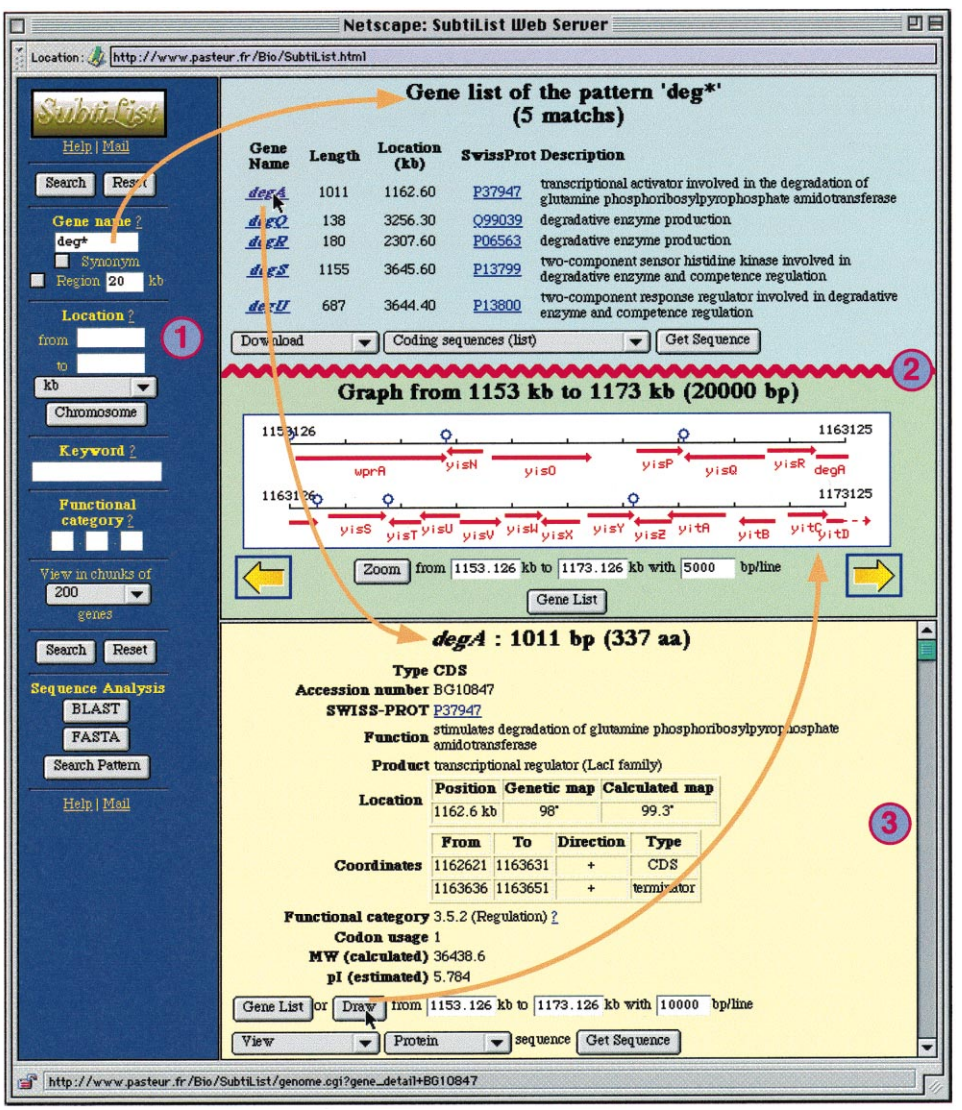

Fig. 1. Screen from the SubtiList WWW server. Circled numbers refer to the three frames described in the text. Possible user actions are indicated by orange arrows. The red wavy line indicates that this figure has been edited, so as to show both lists of genes (blue background) and graphical representations (green background), which cannot be displayed simultaneously by the real interface.

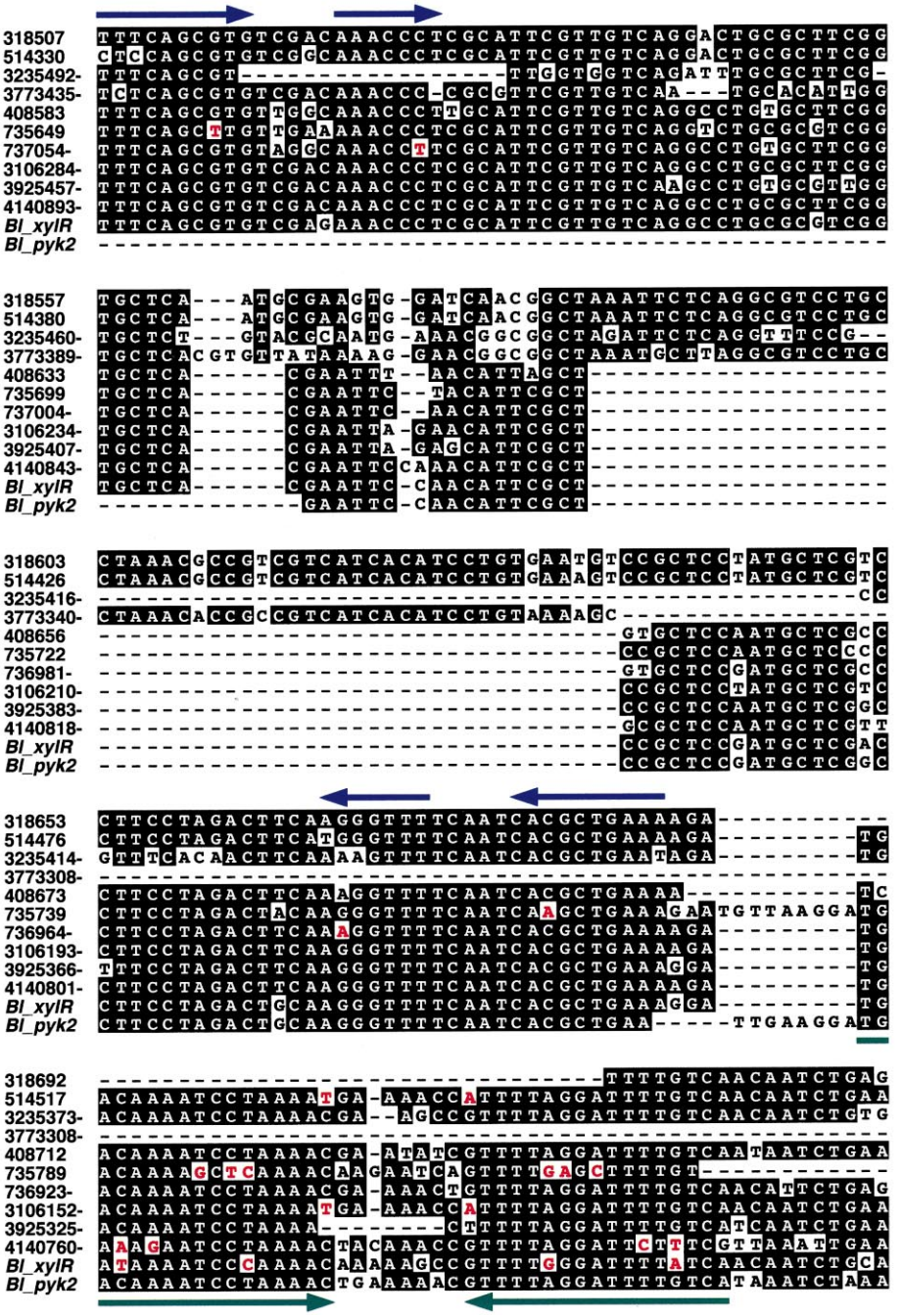

Fig. 2. Multiple alignment of the sequences of the non-coding repeated element present ten times in *B. subtilis* and twice in *B. licheniformis*. Sequences are labeled according to their position in the *B. subtilis* chromosome (followed by '-' when found on the complementary strand), or their name preceded by '*Bl*' in *B. licheniformis* (accession numbers Z80222 and D31955). Possible secondary structures and compensatory mutations are indicated.

server), but also from sites dedicated to particular aspects of the biology of *B. subtilis* (sporulation, 2D protein data). Several sequence analysis tools are also accessible via SubtiList for specific studies of the genome of *B. subtilis*. For example, programs for rapid protein sequence comparisons (BLAST(2), FASTA), and DNA or protein pattern searches, running against *B. subtilis* sequence data, can be used. DNA pattern searches may be restricted to specific genomic regions, such as areas surrounding start or stop codons, or intergenic regions, making possible the more accurate exploration of signals important for the regulation of gene expression. Other information about the molecular biology of *B. subtilis* is available on the Internet, including the BSORF-DB (http://bacillus.genome.ad.jp/BSORF-DB.html), Micado (http://locus.jouy.inra.fr/cgi-bin/genmic/madbase_home.pl) [30] and NRSub (http://acnuc.univ-lyon1.fr/nrsub/nrsub.html) [31] WWW sites (a more complete list is available in SubtiList).

4. Initial exploration of the DNA information content

4.1. Base composition and oligonucleotide bias

The simplest way to consider a genome is to look at its base

composition: the *B. subtilis* genome sequence is 43.5% G+C-rich on average. However, significantly heterogeneous regions have been identified on the chromosome, which contain a higher proportion of either G+C nucleotides (notably rDNA regions) or A+T nucleotides (see below for a possible interpretation of these regions) [17]. A large imbalance in the purine and pyrimidine composition of the replication leading and lagging strands has been observed [32]. This made it possible to directly identify the origin and terminus of replication, which could also be identified by plots of the frequency along the chromosome of some dinucleotides (AG, GA, CT, TC), showing dramatic variations around these two regions [17,32], or by measuring the G−C/G+C ratio [33]. These results were correlated with 74% of the protein genes (which contain a higher proportion of A+G nucleotides than the genome as a whole) being transcribed in the same direction as movement of the replication fork (Table 1) [17].

These types of analysis led to studies of longer oligonucleotides, their frequency as well as spatial distribution, in order to identify contrast words, i.e. short lengths of sequence that are significantly under- or over-represented, indicating possible selection constraints [34,35]. One of the main ideas was to build models of the chromosome, and to compare the actual number of times a word appeared in the genome sequence with the expected count derived from a given model. Theoretical models were frequently used, such as Markov models, in which the frequency of an oligonucleotide is evaluated by considering smaller words within the word studied [36]. Results should also be tested using different sets of data: this makes it possible to isolate the various selective constraints acting on the genome, such as the separation of the sequence into leading and lagging strands to study replication signals, or the separation of sequences into genes and intergenic regions to study signals related to transcription and translation.

Some interesting results have already been obtained. In *B. subtilis*, as in several other bacteria for which complete genome sequences are available, the total number of biased words 2–8 nucleotides in length was much larger than would be expected by chance, and there were more biased words of size 7 than of any other size; this latter observation was thought to be due to interactions with DNA or RNA polymerases [32]. Furthermore, palindromes have been found to be under-represented, probably due to the role of restriction sites related to restriction-modification (R-M) systems [37,38]. However, two further observations were quite surprising: all restriction sites seemed to be under-represented in bacterial genomes (i.e. not only the ones recognized by endogenous R-M systems), and palindromes were avoided to a smaller extent in prophage-like regions of the *B. subtilis* genome (see below), although the converse was expected as a condition for successful phage transduction [32]. These results, in contrast to the widely recognized role of R-M systems in protecting against foreign DNA, were tentatively interpreted as being due to the ease with which R-M systems can be horizontally transferred, causing a general avoidance of all possible palindromes in bacterial genomes [32].

More complex and realistic models of the chromosome were also built, intended to integrate pre-existing biological knowledge so as to discover new signals: simulated sequences in which the average length of the genes, dinucleotide frequency, codon usage, and dipeptide frequency were conserved [39]. Studies on the distribution of tetranucleotides using such models shed light on probable alterations in the pattern of translation of a number of genes of *B. subtilis* [40]. Indeed, it has been found that pairs of AGCT motifs separated by a multiple of three nucleotides were significantly over-represented. The overall frequency of AGCT motifs in *B. subtilis* was twice that in *Escherichia coli*, but a similar, uneven distribution was found in both organisms; this was true only for the AGCT motif, not for the other 255 tetranucleotides. Remarkably, regions containing many such AGCT pairs were frequently found inside genes, often those coding for the same function in *E. coli* and *B. subtilis*, or corresponding to functions specific to each organism. Moreover, the three reading frames likely to contain these AGCT motifs were significantly unevenly represented, whereas there was no bias in the polypeptide sequence, indicating that selection pressure probably operated on mRNA during translation. Several other lines of evidence, including the function of the proteins concerned, the prediction of their three-dimensional structure, and mRNA potential pseudoknot structures, strongly suggested that AGCT motifs may be used for ribosomal frameshifting or hopping, resulting in a single mRNA molecule being translated into several different proteins [40].

4.2. Repeated sequences

Bacillus subtilis is naturally competent for DNA-mediated transformation, but neither insertion sequences (IS) nor transposons have been identified in the genome of strain 168 [17]. Was this genome sequence completely devoid of large repeats? A statistical model was used such that, taking into account mononucleotide frequencies, finding one word longer than 25 bp strictly repeated twice in a 4.2-Mb random sequence had a probability of 10^{-3} [41]. Unexpectedly, a large number of such repeats was identified in the *B. subtilis* genome sequence, even when rRNA and tRNA genes were excluded from the analysis. Some were readily accounted for (highly repetitive proteins such as polyketide synthases, prophage-like regions, paralogous genes), whereas others, based on their size and distribution, may be landmarks of an evolutionary strategy by which *B. subtilis* acquires new genetic information, through a mechanism independent of IS or extensive homologous recombination (E. Rocha, personal communication). Some longer regions (ca. 100–400 bp) were also found to be repeated two or three times, with successive copies being contiguous (Table 1) [17].

Finally, the presence of other repeats with more copies was investigated, possibly with a number of differences between the copies. One such element has been found, which was repeated ten times in the genome sequence, always in non-coding regions [17,42]. This 190-bp long sequence has several striking features. Its distribution was not completely random, as all repeats were located on either side of the origin of replication, spanning only one half of the chromosome. Furthermore, all copies but one were in the same orientation as the direction of movement of the replication fork. This element had several inverted repeats in its internal structure, which were often conserved by compensatory mutations in various copies. Multiple alignment of the ten repeats showed that they could be classified into two subfamilies, with an insert in three of the copies (Fig. 2). Finally, a similar sequence has been found in a closely related bacterium, *Bacillus licheniformis* [43]. It has been suggested that these sequences are remnants of an ancient IS [44], are involved in chromo-

some replication or partitioning, or are landmarks of a structural RNA molecule [17].

4.3. Codon usage

It is generally thought that there is bias in codon usage due to the correlation between tRNA availability in the cell and the level of gene expression required under various growth conditions [45]. Codon usage has been studied in *B. subtilis* by statistical data analysis methods: multivariate factorial correspondence analysis (FCA) [46] and automatic classification [47]. This involves considering protein coding sequences (CDSs) as a cloud of 4100 points in a 61-dimensional space, each dimension corresponding to the frequency with which a given codon is used relative to synonymous codons for the same amino acid. Using the χ^2 distance between each pair of genes, FCA projects the cloud of points into two-dimensional space with a minimum loss of information, giving maximum scattering. This calculation is followed by a general clustering method, defining classes of CDSs close to one another, with no a priori knowledge of their nature and number. *Bacillus subtilis* genes were partitioned into three well defined classes in this way (Table 1). This classification was probably significant because these classes were also clearly distinguished by their biological properties [17]. Class 1 contained most of the genes, with the exception of genes from classes 2 and 3. Class 2 contained genes that were constitutively highly expressed during exponential growth, such as those encoding components of the transcription and translation machineries, the core of intermediary metabolism, or stress proteins [29,48]. Class 3 contained a very high proportion of genes of unknown function, often clustered into small groups on the chromosome, corresponding to the A+T-rich islands detected along the chromosome (see above). Genes from class 3 generally corresponded to functions found in, or associated with, bacteriophages, as well as functions related to the cell envelope. Together with known bacteriophages or bacteriophage-like elements (SPβ, PBSX and the *skin* element), these possible cryptic prophages brought to at least 10 the number of such elements, covering more than 9% of the genome of *B. subtilis* 168 (Table 1) [17].

5. Conclusion

Correction of the DNA sequence and annotation of the complete genome sequence of *B. subtilis* is still an ongoing task. A new systematic functional analysis program funded by the European Commission was initiated in 1996, to determine the function of about half of the genes the role of which is still unknown [12]. A similar project is underway in Japan. Preliminary results obtained from the analysis of the *B. subtilis* genome sequence have already revealed that investigation at a first level of the information content of a genome is rewarding. The recent explosion of new data requires tight integration between in silico analysis and experimental biological approaches. Whatever the future of these interactions, genomics has already changed our way of conceiving biology.

Acknowledgements: This work was supported by the European Commission (contract BIO4-CT96-0655), the *BACillus* Industrial Platform (BACIP), the GIP GREG (Groupement de Recherche et d'Étude sur les Génomes 2 93 G404 00 71201 21), and the Institut Pasteur. I thank A. Danchin, P. Glaser, and F. Kunst for their critical comments and suggestions on the manuscript, and C. Médigue and E. Rocha for sharing unpublished results.

References

[1] Harwood, C.R. (1992) Trends Biotechnol. 10, 247–256.
[2] Sonenshein, A.L., Hoch, J.A. and Losick, R. (1993) *Bacillus subtilis* and other Gram-positive Bacteria: Biochemistry, Physiology, and Molecular Genetics, American Society for Microbiology, Washington, DC.
[3] Stragier, P. and Losick, R. (1996) Annu. Rev. Genet. 30, 297–341.
[4] Moir, A., Kemp, E.H., Robinson, C. and Corfe, B.M. (1994) J. Appl. Bacteriol. 76, 9S–16S.
[5] Dubnau, D. (1991) Microbiol. Rev. 55, 395–424.
[6] Grossman, A.D. (1995) Annu. Rev. Genet. 29, 477–508.
[7] Biaudet, V., Samson, F., Anagnostopoulos, C., Ehrlich, S.D. and Bessières, P. (1996) Microbiology 142, 2669–2729.
[8] Itaya, M. and Tanaka, T. (1991) J. Mol. Biol. 220, 631–648.
[9] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Science 274, 546–567.
[10] Kunst, F. and Devine, K.M. (1991) Res. Microbiol. 142, 905–912.
[11] Devine, K.M. (1995) Trends Biotechnol. 13, 210–216.
[12] Harwood, C.R. and Wipat, A. (1996) FEBS Lett. 389, 84–87.
[13] Kunst, F., Vassarotti, A. and Danchin, A. (1995) Microbiology 141, 249–255.
[14] Moszer, I., Kunst, F. and Danchin, A. (1996) Microbiology 142, 2987–2991.
[15] Ogasawara, N., Fujita, Y., Kobayashi, Y., Sadaie, Y., Tanaka, T., Takahashi, H., Yamane, K. and Yoshikawa, H. (1995) Microbiology 141, 257–259.
[16] Ogasawara, N. and Yoshikawa, H. (1996) Microbiology 142, 2993–2994.
[17] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.-K., Codani, J.-J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Seror, S.J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanaka, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H. and Danchin, A. (1997) Nature 390, 249–256.
[18] Médigue, C., Moszer, I., Viari, A. and Danchin, A. (1995) Gene 165, GC37–GC51.
[19] Médigue, C., Rechenmann, F., Danchin, A. and Viari, A. (1998) Bioinformatics, submitted.
[20] Muralikrishna, P. and Suryanarayana, T. (1985) Biochem. Int. 11, 691–699.

[21] Borodovsky, M. and McIninch, J.D. (1993) Comput. Chem. 17, 123–133.
[22] Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Nucleic Acids Res. 26, 544–548.
[23] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[24] Fichant, G.A. and Burks, C. (1991) J. Mol. Biol. 220, 659–671.
[25] Billoud, B., Kontic, M. and Viari, A. (1996) Nucleic Acids Res. 24, 1395–1403.
[26] d'Aubenton Carafa, Y., Brody, E. and Thermes, C. (1990) J. Mol. Biol. 216, 835–858.
[27] Bairoch, A. and Apweiler, R. (1998) Nucleic Acids Res. 26, 38–42.
[28] Comet, J.-P., Aude, J.-C., Glémet, E., Risler, J.-L., Hénaut, A. and Codani, J.-J. (1998) Comput. Chem., submitted.
[29] Moszer, I., Glaser, P. and Danchin, A. (1995) Microbiology 141, 261–268.
[30] Biaudet, V., Samson, F. and Bessières, P. (1997) Comput. Appl. Biosci. 13, 431–438.
[31] Perrière, G., Gouy, M. and Gojobori, T. (1998) Nucleic Acids Res. 26, 60–62.
[32] Rocha, E.P.C., Viari, A. and Danchin, A. (1998) Nucleic Acids Res. 26, 2971–2980.
[33] Lobry, J.R. (1996) Mol. Biol. Evol. 13, 660–665.
[34] Burge, C., Campbell, A.M. and Karlin, S. (1992) Proc. Natl. Acad. Sci. USA 89, 1358–1362.
[35] Karlin, S. and Cardon, L.R. (1994) Annu. Rev. Microbiol. 48, 619–654.
[36] Schbath, S., Prum, B. and de Turkheim, E.J. (1995) J. Comput. Biol. 2, 417–437.
[37] Karlin, S., Burge, C. and Campbell, A.M. (1992) Nucleic Acids Res. 20, 1363–1370.
[38] Gelfand, M.S. and Koonin, E.V. (1997) Nucleic Acids Res. 25, 2430–2439.
[39] Hénaut, A. and Danchin, A. (1996) in: *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology, Vol. 2, (Neidhardt, F., Curtiss, R. III, Ingraham, J., Lin, E., Brooks Low, K., Magasanik, B., Reznikoff, W., Riley, M., Schaechter, M. and Umbarger, H., Eds.) pp. 2047–2066, American Society for Microbiology, Washington, DC.
[40] Hénaut, A., Lisacek, F., Nitschké, P., Moszer, I. and Danchin, A. (1998) Electrophoresis 19, 515–527.
[41] Karlin, S. and Ost, F. (1985) in: Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. 1 (LeCam, L.M. and Olshen, R.A., Eds.), pp. 225–243, Wadsworth, Belmont.
[42] Popham, D.L. and Setlow, P. (1994) J. Bacteriol. 176, 7197–7205.
[43] Presecan, E., Moszer, I., Boursier, L., Cruz Ramos, H., de la Fuente, V., Hullo, M.-F., Lelong, C., Schleich, S., Sekowska, A., Song, B.H., Villani, G., Kunst, F., Danchin, A. and Glaser, P. (1997) Microbiology 143, 3313–3328.
[44] Kasahara, Y., Nakai, S. and Ogasawara, N. (1997) DNA Res. 4, 155–159.
[45] Berg, O.G. and Kurland, C.G. (1997) J. Mol. Biol. 270, 544–550.
[46] Hill, M.O. (1974) Appl. Statist. 23, 340–353.
[47] Delorme, M.-O. and Hénaut, A. (1988) Comput. Appl. Biosci. 4, 453–458.
[48] Shields, D.C. and Sharp, P.M. (1987) Nucleic Acids Res. 15, 8023–8040.