Microbiology (1995), **141**, 261–268 Printed in Great Britain

SubtiList: a relational database for the *Bacillus subtilis* genome

Ivan Moszer, Philippe Glaser and Antoine Danchin

Author for correspondence: Antoine Danchin. Tel: +33 1 45 68 84 41. Fax: +33 1 45 68 89 48.
e-mail: adanchin@pasteur.fr

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

In the framework of the international collaborative project aiming to sequence the whole *Bacillus subtilis* chromosome, we have created a relational database for managing and analysing information associated with the molecular genetics of this bacterium: *SubtiList*. It allows recovery of non-redundant DNA sequences of the *B. subtilis* genome, as well as related information, i.e. genes, proteins, etc. A logical structure has been designed with appropriate links between the different objects, and a set of procedures has been implemented for data updating and management. The database is organized around a core constituted by all known contigs of *B. subtilis*, i.e. sets of non-redundant sequences created from original entries in the EMBL data library. A user-friendly interface has been developed to make the database easy to consult. Sequence analysis tools have been integrated into the database, such as a program for rapid similarity searching of protein data banks, and a powerful DNA pattern searching program. Thanks to the consistency of *SubtiList*, we have performed a codon usage analysis by Factorial Correspondence Analysis, and a study of the distribution of the isoelectric points of known proteins of *B. subtilis*. The *SubtiList* database is available through anonymous ftp (address 'ftp.pasteur.fr' or IP number 157.99.64.12, directory '/pub/GenomeDB/SubtiList').

Keywords: *Bacillus subtilis*, genome sequencing, database, codon usage, isoelectric point

Introduction

With the advent of efficient sequencing techniques and the emergence of large collaborative projects, it appears likely that several complete sequences of bacterial genomes will be known by the end of the century. To profit from this information, fast and easy software is required for quick recovery of knowledge and data associated with gene sequences. For many years, international organizations took charge of maintaining and diffusing DNA and protein sequence data. However, this kind of data storage is not well adapted to complete genome analysis for several reasons. (i) It does not allow easy extraction of aggregate information about genomic data for a given organism. (ii) Data collection by free submission generates discrepancies and redundancy. (iii) The information is organised in flat files that do not allow the establishment of logical relationships between data. (iv) Systematic sequencing programs supply large contiguous stretches of genomic sequences, which require a new kind of data handling. Therefore, it has become necessary to organize the large quantity of new sequences produced by systematic sequencing projects, as well as data obtained by piecemeal sequencing. This requires appropriate quality control by informed scientists and the construction of specialized data libraries.

The first effort which has been made in most cases is to extract data deposited in data banks, and to devise appropriate environment structures allowing the interested scientists to retrieve information rapidly. Several such data libraries for the *Escherichia coli* genome have been constructed. Since 1989, Kröger *et al.* (1993) have collected the *E. coli* sequences from the EMBL and GenBank data banks. A data library was similarly developed in Japan for managing *E. coli* sequences (Kunisawa *et al.*, 1990). More recently, K. Mori and others organized *E. coli* sequence data into a series of flat files, GenoBase (unpublished). However, this structure for data libraries does not provide the relational interface required for multicriteria searches. It is therefore necess-

Abbreviations: CdS, coding sequences; DBMS, database management system; FCA, factorial correspondence analysis; SASP, small acid-soluble spore proteins.

ary to use a structure where the different data types and records can be automatically linked. This requires the use of particular software: database management systems (DBMSs).

Rudd and co-workers have developed dedicated software for *E. coli* which contains both flat data files and programs for collecting, aligning, representing graphically and analysing sequences (Bouffard *et al.*, 1992; Rudd, 1993). At the same time, Médigue *et al.* (1993) developed a relational database for the *E. coli* genome. The data structure, which allows multicriteria searches, was exported to develop specialized databases for other genomes including *Saccharomyces cerevisiae* (Slonimski & Brouillet, 1993).

No such tool has been developed for *Bacillus subtilis*. The only sequence compilation has been a data library containing all published protein genes of *B. subtilis* in a series of flat files (Sharp *et al.*, 1990a). As part of the international collaborative program to determine the whole sequence of the *B. subtilis* genome (Kunst *et al.*, 1995), we have constructed a relational database, i.e. a logical, dynamic and evolving structure, named *SubtiList*. It is derived from the model originally constructed for the *E. coli* genome, the Colibri database (Médigue *et al.*, 1993). This structure allows the rational and efficient management of the data produced by sequencing laboratories, as well as information from international data libraries. Thus, all information on the *B. subtilis* genome is made available in a form suitable for retrieval and extraction of specific data for various purposes. Information is logically connected, independent of its physical organization, leading to verified, consistent, and non-redundant data, defined as a set of clean data. The exploration of this text will result in the discovery of rules that govern the organization of the genomic information of *B. subtilis*, and its operation, namely its consistency. The results should then be compared with those obtained from the genome of *E. coli*, using the Colibri database.

Software

To obtain a clean dataset of all genomic data for *B. subtilis*, we implemented a relational database using commercial software, 4th Dimension® (4D, ACI), and a Macintosh computer. 4D is a relational DBMS which allows the design and construction of a complex database structure. It supplies a user-friendly interface to define 'files' made up of 'fields' of different types intended to describe data, as well as logical links or relationships between files. 4D also provides a procedure language which associates the capabilities of a fourth-generation language with those of Pascal or C language. Databases can be compiled and external procedures written in another language can be integrated, permitting a significant increase in running speed. Moreover, a user-friendly interface including layouts to display the data can be designed easily which make the database easy to consult. Finally, several modules are available for word processing and other office tasks within 4D, and for connecting 4D databases with other relational DBMSs. A more detailed description of the internal conception of 4D can be found in a previous article describing the Colibri database (Médigue *et al.*, 1993).

Constructing and updating *SubtiList*

All numerical results presented below were calculated using release 40 of the EMBL data library plus cumulative updates up to November 15, 1994.

The *SubtiList* database is organized around a core constituted by all known contigs of *B. subtilis*. These contigs were constructed by removing all redundancy from the sequences of *B. subtilis* available in public data banks. Firstly, we extracted entries for *B. subtilis* from the EMBL library (Rice *et al.*, 1993) and imported them into the 4D database. Data that was not from the paradigm strain 168 was eliminated. The database currently contains 546 EMBL entries. Secondly, we detected all the sequences completely included within other sequences. Sequences which are not entirely covered by another sequence entry are called 'Mothers'. The remaining EMBL sequences are termed 'duplicates'. Three hundred and sixty Mothers were identified. Thirdly, the contigs were constructed by detecting overlaps between Mothers and joining the corresponding sequences. Finding both Mothers and contigs was associated with a sequence alignment to identify nucleotide discrepancies in overlapping sequences. The differences may be due either to natural polymorphism or to mistakes in sequencing and data acquisition. The error/polymorphism rate for the *B. subtilis* sequences in public data banks was about 0·2%. From the 1654 kb of DNA sequence of *B. subtilis* in the EMBL data library, we constructed 241 contigs spanning 1221 kb, i.e. 29% of the 4188 kb chromosome. Thus, the redundancy level in public data banks for *B. subtilis* sequences is about 35%.

The localisation of protein-coding sequences (CdS) of RNA genes and of control signals was calculated from the Features Tables of the EMBL entries. We took into account, as far as possible, overlaps and checked possible mistakes when two entries from the same region differed. In many instances this required verification in original publications and could not be performed automatically. The description of each sequenced gene was generally extracted from the genetic map established by Anagnostopoulos *et al.* (1993). Other sources of data, such as the EMBL entries themselves and the associated bibliography, were also considered. We attempted to standardize the information that was not directly related to the sequence itself (e.g. keywords, features, comments). Finally, a procedure automatically generated records for each of the amino acid sequences translated from the CdS, and calculated data such as the predicted molecular mass and the estimated isoelectric point of the corresponding proteins. The CdS were systematically verified to eliminate possible mistakes and to check that no frameshift had been created during the construction of the contigs. In the near future, unique accession numbers will be attributed to each gene to establish cross-references with the protein data bank Swiss-Prot (Bairoch & Boeckmann,

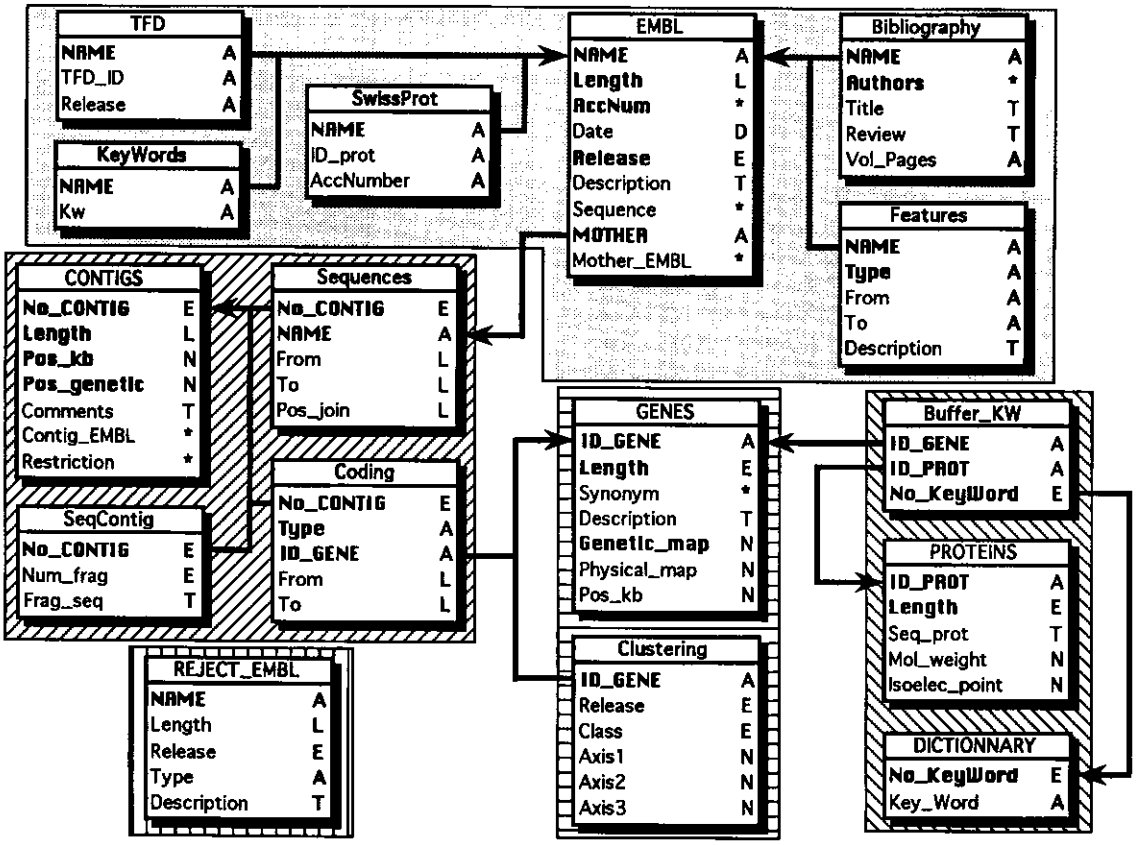

Fig. 1. Schematic structure of the *SubtiList* database. Each box is a 4D file associated with its fields. Each field is followed by its type with a one-letter code (A, string; T, text; E, integer; L, long integer; N, real number; D, date; *, 4D sub-file). Bold fields are indexed. Arrows indicate logical links between files (links of type $N \rightarrow 1$). ▨, EMBL section; ▨, contigs section; ▤, genes section; ▧, proteins section; ▥, rejected EMBL section.

1993). The number of protein-coding gene sequences currently recorded in *SubtiList* is 997. The function of 365 remains unknown (mainly genes identified by systematic sequencing).

When *SubtiList* is being updated, a series of procedures scans the pre-existing contigs with new EMBL entries and manages the incorporation of these sequences into the database. Four cases are considered: the new sequence (i) is a duplicate of an older EMBL sequence, (ii) is included within a contig, (iii) overlaps one or several contigs, or (iv) makes up a new contig. Subsequently, another set of procedures identifies new or modified contigs to semi-automatically bring data up-to-date. However, a fast and completely automated procedure of updating cannot be reasonably considered because decisions have to be made about conflicting information. *SubtiList* is updated once a month by this procedure. In contrast, the software will evolve to respond to new needs.

This procedure of constructing and updating the database is a significant improvement on the procedure used by Colibri. The contigs of Colibri were detected using the physical restriction map established by Kohara *et al.* (1987), i.e. with a resolution of about 500 bp. For this reason, overlaps could not be detected exhaustively, nor could discrepancies between overlapping sequences be revealed. In contrast, our method for detecting contigs, using FASTN and dynamic alignment (see below), gives a resolution at the nucleotide level. Furthermore, different versions of contig sequences can be constructed, for more precise analysis of the corresponding region.

Logical structure of *SubtiList*

Conceiving the database requires firstly that the various kinds of data making up the totality of the information are well identified, and secondly that the relationships between these objects are well defined. Pertinent data are mainly sequences, genes, control regions identified on the sequences, proteins deduced from the DNA sequences of genes and bibliography related to these objects. Each piece of data has to be annotated. This information may come from a variety of sources. The structure of the database has to integrate all this information and allow the user to recover whatever he needs easily. A well-conceived database will also make modifications of the database easy.

The *SubtiList* database is divided into four main sections (Fig. 1). The first section stores all the information contained in the EMBL data library concerning

B. subtilis 168 or 168-derived strains. These data are stored unmodified so that the information can be retrieved exactly as entered in the EMBL data bank. However, they are organized to allow a more convenient method of consultation than is possible in EMBL. The core of this section is the 4D file [EMBL]. It contains information including the name of the entry, accession numbers, the length of the sequence and the sequence itself. Several 4D files are linked to the [EMBL] file to store data such as bibliographic references or the Features Tables. As an annex to this first section, the 4D file [REJECT_EMBL] stores data concerning non-168 entries, plasmids or RNA sequences.

The second section stores information related to the contigs. The central 4D file [CONTIGS] contains fields describing various data such as the length or the mapping information, while the sequence itself is stored in a linked 4D file. This section also includes relationships between the contigs and the original sequences from the EMBL data library (4D file [Sequences]), as well as annotation of biologically significant features identified on the contigs (4D file [Coding]). These features are extracted from the Features Tables of the EMBL entries, but may also be computed independently as new data appear, whether from experimental work or from computer analysis.

The information related to the sequenced genes of *B. subtilis* is contained in the third section. This is mainly made up of a single 4D file [GENES], which is linked to corresponding features on the contigs thanks to the [Coding] file. The [GENES] file allows description of basic information concerning genes: length, function, accession numbers, genetic mapping, etc. As in the case of the Colibri database, two fields are present to store physical mapping data for the genes. However, in the case of *SubtiList*, the only physical map available comprises only 98 sites in the whole chromosome (Itaya & Tanaka, 1991). Thus, these fields cannot be filled in for the moment, due to the lack of a detailed restriction map of the chromosome of *B. subtilis*.

Finally, the last section of the database stores data relevant to proteins (such as molecular mass, isoelectric point) and the sequence itself in the 4D file [PROTEINS]. The files [PROTEINS] and [GENES], and a third file containing keywords, [DICTIONARY], are linked together via a buffer file, [Buffer_KW]. The keywords are extracted from the EMBL and SWISS-PROT data banks, and from the genetic map of *B. subtilis* (Anagnostopoulos *et al.*, 1993).

Once this structure is established, it is a relatively easy task to add or modify 4D fields and files as the demand for information increases. In particular, this database is also intended to store experimental and computer-aided results on DNA sequences, genes and proteins, so that existing correlation between data can be highlighted quickly and readily.

Using *SubtiList*

The *SubtiList* database allows easy retrieval of any information concerning the *B. subtilis* genome. To make the consultation of the database as simple as possible, we developed a user-friendly interface using the 4D interface generator. A set of layouts presenting data on the screen has been designed. They contain a combination of data from different 4D files and allow the user to navigate between the diverse sections of the database described above. Links between data are automatically generated, and logical connections can be established between a set of genes and the bibliography associated with it, for instance, without knowing how data are physically organized in the database. Basic manipulations of the data, such as sorting, exporting or printing are possible via a button or menu interface. A powerful multicriteria-search tool has been developed, taking advantage of the query module of 4D. Moreover, one can obtain a graphical representation of each contig, and perform zooms or get information by clicking with the mouse on a gene name or an EMBL entry. As an example, Fig. 2 shows the layout of a contig record on which one can identify EMBL entries that constitute this contig, as well as genes annotated on this sequence. We are currently developing a completely graphic interface which will allow the user to perform a series of zooms from the whole 'circular' chromosome, and to access data by clicking with the mouse on the corresponding object.

Moreover, several sequence analysis tools have been directly integrated into the database. They are therefore accessible from the user-interface and can be applied directly to the data in the database, avoiding tedious manipulations of exporting and managing ASCII text files. These tools are written in C language and are called from the 4D database as external procedures, so they can operate very quickly. Simple programs have been implemented, such as translation of a nucleic acid sequence, calculation of a restriction map and calculation of the molecular mass and isoelectric point of a protein. In the present version, we have added two more complex tools.

A FASTP program, extracted from the Smarties package for Macintosh (A. Viari, unpublished), is used to scan a protein data bank with a given protein sequence of the *SubtiList* database for rapid similarity searching (Lipman & Pearson, 1985). Because of the numerous parameters needed by the program, an interface has been developed to choose the protein data bank to scan, the distance matrix between amino acids and the various search options.

A powerful pattern searching program has also been integrated into the database. This can search for degenerate patterns in nucleic acid sequences. The grammar used to describe the pattern is very simple, allowing indication of ambiguous positions, mandatory positions, repetition of a position, etc. A maximum number of mismatches allowed can also be defined. Moreover, complex patterns, made up of several small patterns separated by a variable number of base pairs, can be used. Finally, the search may be restricted in different ways: it can be performed on the whole database or on a given set of contigs, and it can be limited to the areas of the contigs potentially implicated in regulation processes, i.e. regions surrounding the start of genes.

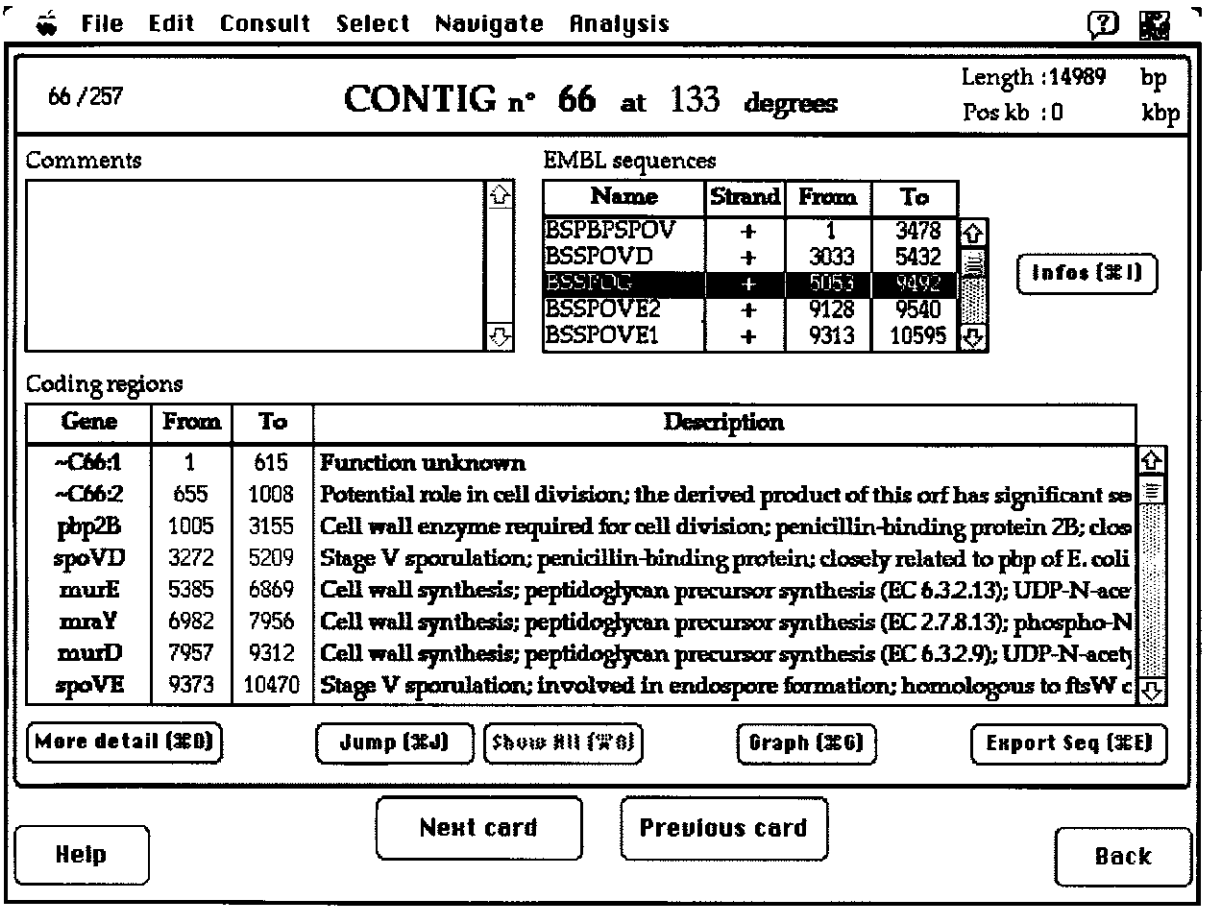

Fig. 2. Screen capture of a layout presenting a contig record, with data from several 4D files: [CONTIGS], [EMBL] linked by [Sequences], [GENES] linked by [Coding].

Algorithms

Methodologically, we used two well-known programs to construct the contigs. Finding Mother–duplicate relations, and detection of similarities between a given sequence and the contigs of the database, were performed with a FASTN procedure (Wilbur & Lipman, 1983). This was followed by an alignment of the overlapping regions, using a dynamic programming algorithm (Needleman & Wunsch, 1970; Sellers, 1974). Although these algorithms could have been implemented directly within the database using the programming language integrated in 4th Dimension, this solution would have been much too slow to be of practical use. Thus, we have implemented these algorithms in C language, and optimized them to make them as fast as possible on a Macintosh computer.

Similarly, the FASTP module implements, in C, a variant of the original algorithm of Lipman & Pearson (1985). In the latter, substitution matrices used to represent equivalencies between amino acids are applied in the last step of scoring the diagonals resulting from the word-search. Here, we also took these matrices into account in the first step, i.e. searching for words common to different sequences. This permits significant improvement in the sensitivity of the program. For the pattern searching module, we used a consensus matrix approach implemented, in C, in a bit comparison manner so as to be as fast as possible.

Exploiting *SubtiList*

SubtiList makes available any sort of information about the genome of *B. subtilis* for analysis of different aspects of the organization of a bacterial genome. We describe two examples of such applications.

Thanks to the consistency of the database, we could readily extract genes encoding proteins without any duplication. We analysed the codon usage of *B. subtilis* using Factorial Correspondence Analysis (FCA). FCA is a data analysis method which finds the best two-dimensional representation of a cloud of points, whatever the original number of dimensions of the cloud is, using the χ^2 distance between each point (Hill, 1974). A codon usage table for *B. subtilis* can be deduced from the gene sequences, where each line is one gene and each column is the frequency, for this gene, of one codon among the

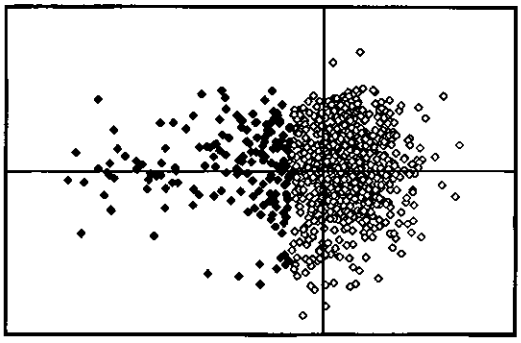

Fig. 3. Codon usage analysis by FCA. Representation of the two first axes (18% of total inertia). One diamond represents one gene. Filled symbols are the genes of the highly expressed genes class as calculated by the classification algorithm. It is worth noting that the frontier between classes may change as the number of points increases.

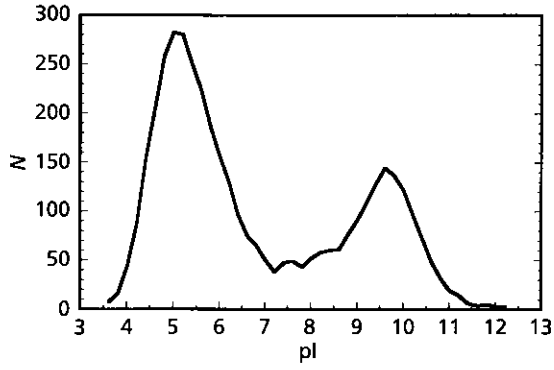

Fig. 4. Distribution of the isoelectric points (pI) of the proteins. *N* is the number of proteins in a window of 1 unit of pI sliding with a pitch of 0·2.

synonymous codons for the corresponding amino acid. Thus, the genes can be represented as a cloud of points in a 61-dimensional space (the codons). FCA spreads out this cloud along its largest axis, to project with a minimum loss of information these points in a human-readable way, i.e. a two-dimensional space. Thus, in the resulting plot, two points corresponding to genes having a similar codon usage appear as neighbours (the converse is not necessarily true). A classification method (Diday, 1971; Delorme & Hénaut, 1988) can be used to set up an adequate clustering of the whole genes according to this criterion. In the case of *E. coli*, three classes of genes were identified and interpreted according to the biological properties of the genes constituting them (Médigue *et al.*, 1991). In contrast, in *B. subtilis*, only two well-separated classes were apparent (Fig. 3). The smallest one comprises genes implicated in the core machinery of transcription and translation, genes of central metabolism and genes encoding small acid-soluble spore proteins (SASP), i.e. highly expressed genes. The large class includes all the other, and thus the majority of the genes. Interestingly, Shields & Sharp (1987) have already performed a similar analysis using far fewer genes (56 genes) and Sharp *et al.* (1990b) have studied this topic using 221 genes. They reported that genes for ribosomal proteins and SASP were the most biased in synonymous codon usage and were separated from the rest of the genes. Although we now know nearly 1000 *B. subtilis* genes, the general organization of these genes according to their codon usage is still the same. This difference to *E. coli* may be due to the fact that *B. subtilis*, unlike *E. coli*, is an organism naturally competent for DNA transformation and recombination. As a consequence, it tends to homogenize all DNA entering the cell, and does not evolve towards a codon strategy for horizontal transfer (i.e. an *E. coli*-third class equivalent). These results have been recorded in the database.

A second insight into the general organization of the *B. subtilis* genome is provided by the analysis of the distribution of the isoelectric points (pI) of known proteins. As shown in Fig. 4, there is a strong bias against proteins which should have an isoelectric point around 7·5 (±1). This may indicate the intracellular pH of *B. subtilis*: proteins showing the corresponding pI would be subject to a considerable instability of their net charge. Furthermore, proteins exhibiting such a pI may be excessively sensitive to the intracellular proton content and implicated in proton-driven mechanisms. The 26 proteins of known function found between pI 7 and pI 8 include: (i) several excreted proteins, i.e. insensitive to the intracellular pH; (ii) two proteins whose function is related to protons: the γ subunit of ATP synthase, and one putative response regulator of a two-component system, disruption of which leads to a protonophore resistance phenotype. Twenty-seven proteins of unknown function are also found in the pI range 7–8.

Other similar analyses, notably about the length of the proteins, and global analysis about general signals (promoters, translation initiation signals, etc.) are being carried out and will be published elsewhere.

Conclusion

SubtiList is a database which permits recovery of non-redundant DNA sequences of the *B. subtilis* genome, as well as related information. A set of procedures has been implemented for data management, with the aim that the user will not have to understand the underlying structure of the database. The data structure has been designed with appropriate links between the different objects in the database, to allow immediate access to the variety of information associated with the molecular genetics of *B. subtilis*. This database is implemented on a low cost computer and does not require any specialist computing knowledge. This is the first effort to give an overall view of the knowledge associated to this bacterial genome. Improvements in the general procedure of updating the data are currently applied to the Colibri database.

A recent trend in the field of databases dedicated to the management of complete genomes is the use of advanced

computing techniques, allowing scientists to handle and analyse the very large volume of data generated by large-scale sequencing programs, and the knowledge which is associated. The ideal software should be able not only to manage the data and the results of analysis, but also to help the user choose the adapted methods for a given task in linking, more or less automatically, the procedures aiming at solving a global analysis task. Environments are being constructed which allow modelling and manipulation of descriptive knowledge generated by a genome sequencing program, to help the user solve his sequence analysis problems through task decomposition and method selection, and finally to display and manage the set of newly created objects. These systems are therefore meant to integrate descriptive knowledge on the entities involved (such as genes, promoters, maps, etc.) together with methodological knowledge on a large and extensible set of analysis methods. Such object-oriented knowledge bases have been developed for *E. coli* by Shin *et al.* (1992) and by Perrière & Gautier (1993). We are presently constructing a more evolved system for the general case of a bacterial genome, and in particular *B. subtilis*.

The *SubtiList* database is available through anonymous ftp. It is embedded in a runtime version of the 4D software so that it is a stand-alone application which does not require 4th Dimension to run. Users connected to the Internet network can download the program from the Institut Pasteur ftp server: address 'ftp.pasteur.fr' or IP number 157.99.64.12, login as 'anonymous' with your e-mail address as the password, change to the '/pub/GenomeDB/SubtiList' directory and get the 'README' file (session from the system prompt: type 'ftp ftp.pasteur.fr', 'anonymous', '<e-mail address>', 'cd /pub/GenomeDB/SubtiList', 'get README'). The 'README' file explains how to recover the database on Macintosh and how to start the software. An on-line help is available as you navigate in the database and a complete documentation is being written. Users who do not have access to the Internet may send three high-density (1·4 Mb) Macintosh-formatted floppy disks to: Secrétariat de Mr A. Danchin at the address given on p. 261. However, the use of the network is strongly encouraged.

We emphasize that in many cases, the consultation of articles has been necessary to elucidate conflicts or errors in the data. However, despite all the care taken, it is likely that some mistakes still remain. Any comments on the data of *SubtiList*, as well as any new information such as gene mapping data, are welcome and should be sent together with bug reports by e-mail or by regular mail to Professor Antoine Danchin at the address cited on p. 261.

Acknowledgements

We are grateful to Dr Alain Viari for providing the FASTP program, for technical help in writing C programs and integrating them into 4D, and for fruitful discussions about managing analysis. We also thank Dr Claudine Médigue for supplying the original Colibri database structure, and Dr Alain Hénaut for his helpful advice in the practice of FCA. This work was supported by special grants from the 3e Section de l'Ecole Pratique des Hautes Etudes, from the Direction de la Recherche et des Etudes Doctorales (DRED), from the Centre National de la Recherche Scientifique (CNRS – GDR 1029 'Informatique et Génomes') and from the Groupement de Recherches et d'Etudes sur les Génomes (GREG). I. M. was a recipient of a fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche.

References

Anagnostopoulos, C., Piggot, P. J. & Hoch, J. A. (1993). The genetic map of *Bacillus subtilis*. In *Bacillus subtilis and Other Gram-positive Bacteria: Biochemistry, Physiology and Molecular Genetics*, pp. 425–461. Edited by A. L. Sonenshein, J. A. Hoch & R. Losick. Washington, DC: American Society for Microbiology.

Bairoch, A. & Boeckmann, B. (1993). The SWISS-PROT protein sequence data bank, recent developments. *Nucleic Acids Res* **21**, 3093 3096.

Bouffard, G., Ostell, J. & Rudd, K. E. (1992). GeneScape: a relational database of *Escherichia coli* genomic map data for Macintosh computers. *Comput Appl Biosci* **8**, 563–567.

Delorme, M. O. & Hénaut, A. (1988). Merging of distance matrices and classification by dynamic clustering. *Comput Appl Biosci* **4**, 453–458.

Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. *Rev Stat Appl* **19**, 19- 33.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Appl Stat* **23**, 340–353.

Itaya, M. & Tanaka, T. (1991). Complete physical map of the *Bacillus subtilis* 168 chromosome constructed by a gene-directed mutagenesis method. *J Mol Biol* **220**, 631–648.

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

Kröger, M., Wahl, R. & Rice, P. (1993). Compilation of DNA sequences of *Escherichia coli* (update 1993). *Nucleic Acids Res* **21**, 2973–3000.

Kunisawa, T., Nakamura, M., Watanabe, H., Otsuka, J., Tsugita, A., Yeh, L. S., George, D. G. & Barker, W. C. (1990). *Escherichia coli* K12 genomic database. *Protein Sequences & Data Anal* **3**, 157- 162.

Kunst, F., Vassarotti, A. & Danchin, A. (1995). Organization of the European *Bacillus subtilis* genome sequencing project. *Microbiology* **141**, 249–255.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. & Danchin, A. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222**, 851–856.

Médigue, C., Viari, A., Hénaut, A. & Danchin, A. (1993). Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol Rev* **57**, 623–654.

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443–453.

Perrière, G. & Gautier, C. (1993). ColiGene: object-centered representation for the study of *E. coli* gene expressivity by sequence analysis. *Biochimie* **75**, 415–422.

Rice, C. M., Fuchs, R., Higgins, D. G., Stoehr, P. J. & Cameron, G. N. (1993). The EMBL data library. *Nucleic Acids Res* **21**, 2967–2971.

Rudd, K. E. (1993). Maps, genes, sequences, and computers: an *Escherichia coli* case study. *ASM News* **59**, 335–341.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J Appl Math* **26**, 787–793.

Sharp, P. M., Higgins, D. G., Shields, D. C. & Devine, K. M. (1990a). Protein-coding genes: DNA sequence database and codon usage. In *Molecular Biological Methods for Bacillus*, pp. 557–569. Edited by C. R. Harwood & S. M. Cutting. Chichester: John Wiley and Sons.

Sharp, P. M., Higgins, D. G., Shields, D. C., Devine, K. M. & Hoch, J. A. (1990b). *Bacillus subtilis* gene sequences. In *Genetics and Biotechnology of Bacilli*, pp. 89–98. Edited by M. M. Zukowski, A. T. Ganesan & J. A. Hoch. San Diego: Academic Press.

Shields, D. C. & Sharp, P. M. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* **15**, 8023–8040.

Shin, D. G., Lee, C., Zhang, J., Rudd, K. E. & Berg, C. M. (1992). Redesigning, implementing and integrating *Escherichia coli* genome software tools with an object-oriented database system. *Comput Appl Biosci* **8**, 227–238.

Slonimski, P. P. & Brouillet, S. (1993). A data-base of chromosome III of *Saccharomyces cerevisiae*. *Yeast* **9**, 941–1029.

Wilbur, W. J. & Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc Natl Acad Sci USA* **80**, 726–730.

Received 13 July 1994; revised 16 September 1994; accepted 4 October 1994.