SubtiList: the reference database for the *Bacillus subtilis* genome

Ivan Moszer[1,*], Louis M. Jones[2], Sandrine Moreira[1], Cécilia Fabry[1] and Antoine Danchin[1,3]

[1]Unité de Génétique des Génomes Bactériens and [2]Logiciels et Banques de Données, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France and [3]Hong Kong University-Pasteur Research Centre, Dexter HC Man Building, 8, Sassoon Road, Pokfulam, Hong Kong, China

Received August 21, 2001; Accepted September 10, 2001

DDBJ/EMBL/GenBank accession no. AL009126

ABSTRACT

SubtiList is the reference database dedicated to the genome of *Bacillus subtilis* 168, the paradigm of Gram-positive endospore-forming bacteria. Developed in the framework of the *B.subtilis* genome project, SubtiList provides a curated dataset of DNA and protein sequences, combined with the relevant annotations and functional assignments. Information about gene functions and products is continuously updated by linking relevant bibliographic references. Recently, sequence corrections arising from both systematic verifications and submissions by individual scientists were included in the reference genome sequence. SubtiList is based on a generic relational data schema and a World Wide Web interface developed for the handling of bacterial genomes, called GenoList. The World Wide Web interface was designed to allow users to easily browse through genome data and retrieve information according to common biological queries. SubtiList also provides more elaborate tools, such as pattern searching, which are tightly connected to the overall browsing system. SubtiList is accessible at http://genolist.pasteur.fr/SubtiList/. Similar bacterial databases are accessible at http://genolist.pasteur.fr/.

INTRODUCTION

Complete genome sequences, in particular those of bacterial species, are crowding public databanks. Their analysis will yield major insights into the processes involved in the construction of self-replicating organisms. This initially requires the collection, the organization, and the actualization of data from heterogeneous sources (e.g. sequence, biochemical, physiological and genetic data). To fulfill these requirements, and to ensure an efficient access to the information, specialized databases have been developed. The SubtiList database was created within the framework of the *Bacillus subtilis* genome project (1), by collating and integrating various aspects of the genomic information from *B.subtilis* strain 168—the paradigm of Gram-positive endospore-forming bacteria. SubtiList provides a continuously updated dataset of DNA and protein sequences, linked to the relevant annotations and functional assignments. Taking advantage of a user-friendly, specialized interface on the World Wide Web, the SubtiList database constitutes the reference resource for the analysis of the genome of *B.subtilis*.

INTERNAL ORGANIZATION AND TECHNICAL FEATURES

SubtiList is a specialized database upgrading previous database releases started >10 years ago. The first versions of the Colibri (2) and SubtiList (3) databases were conceived when genome sequences were only partial. They were therefore constructed around the concept of the ‘contig’ (i.e. sets of non-redundant sequences gathered from databank entries, or large stretches of sequence produced by genome projects). Conversely, the current data structure used for the construction of SubtiList was designed to handle complete microbial genomes. The corresponding data model was developed into a general data frame—called GenoList—in such a way as to be generic. This makes it possible to set up similar databases for other bacterial species (see last section). GenoList databases are based on a relational data schema, organized so as to efficiently deal with the complexity of long continuous stretches of sequences. The internal organization of the data involves artificial subdivisions of the genome sequence. Associated genomic features are located relative to these chromosomal sections, which are then mapped to the whole chromosome. This leads to partial and efficient loading of the data whenever needed (query and update). Thus, sequence and feature updates do not require re-evaluation of the whole set of objects in the database, only of the chromosome fragment concerned. This artifice is completely transparent to the end-user of the database, as described below. Other features of the data model, albeit not discussed at length here, are mentioned in the following sections about the database content. The complete data schema is available at the SubtiList World Wide Web server.

The database is operated using two different DataBase Management Systems (DBMS): 4th Dimension® on a Macintosh microcomputer is used for the daily data handling (sequence corrections, literature updates and the like), whereas a copy of the database structure and data is run on a mainframe computer using the Sybase® SQL DBMS. The Sybase implementation is

*To whom correspondence should be addressed. Tel: +33 1 45 68 84 41; Fax: +33 1 45 68 89 48; Email: moszer@pasteur.fr used to generate dynamic HTML pages through a CGI script written in the C/C++ and Perl programming languages.

THE SubtiList WORLD WIDE WEB INTERFACE

In order to make the exploration and in-depth analysis of genome information easier, one needs appropriate ways to browse and query the corresponding data. The World Wide Web interface of SubtiList was built up with these specifications in mind: it is mainly oriented towards the typical questions asked by biologists. The front page is divided into three HTML frames. The left-hand frame makes the most common queries accessible, such as searches by gene name, chromosome location or keyword. This frame also gives access to more elaborate queries and data analysis (see below). A checkbox illustrates one of the biologist-oriented functionalities of the interface: it allows direct access to the chromosome region around a gene of interest. This is indeed one of the most frequent and typical requests, since a gene is usually best looked at in its genome environment. The second (upper right-hand) frame shows either a list of genes generated from various simple or complex queries, or a graphical representation of a chromosome region. This frame may also display forms allowing the user to launch more complex analyses, as well as their outcome (see below). The layout of the gene lists can be easily customized, e.g. adding or removing columns, or changing the sort order according to several data fields. The information thus displayed can be easily exported in various formats (including easy to parse tab-delimited files). When the gene list corresponds to a chromosome region, appropriate controls allow navigation around this location. The DNA sequence of the region can be exported, as well as the DNA or protein sequences of the corresponding genes. Finally, the third (bottom right-hand) frame presents detailed information about the gene of interest selected from the upper frame. Characteristics of these data are further detailed below. User-friendly functions available from this frame include the immediate access to the neighboring genome region, and the possibility to get the gene (protein) sequence together with the adjacent nucleotides—where regulatory and other signals are usually searched for.

The SubtiList World Wide Web server can also be used with a higher level of expertise: indeed, advanced functionalities allow one to extract intricate features. For instance, an extended search function is available to combine several search criteria and retrieve detailed information. A powerful pattern-searching program makes it possible to search for degenerate patterns in nucleic acid or protein sequences. The syntax used to describe the patterns allows one to define ambiguous or mandatory positions, the maximum number of mismatches allowed, and complex patterns made up of several small patterns separated by a variable number of letters. Moreover, the search can be restricted to a genome region and/or to areas potentially implicated in regulation processes (i.e. regions surrounding the start and stop codons of genes, intergenic regions, etc.). Results of the search—clickable positions of the occurrences found—are displayed with relevant information about neighboring genes, and gene lists can be generated from that point. These gene lists may be further customized as described in the previous paragraph. In the same way, classic sequence analysis tools (BLAST and FASTA sequence databank scanning programs) are available, and similarly integrated with the overall genomic resource.

SubtiList DATA CONTENT

The data contained in SubtiList originates mainly from the annotation process performed during the *B.subtilis* genome project (1). The database content is made up of three main parts. As described above, the genome sequence is artificially split out into shorter fragments. Genomic objects represent the physical locations on the chromosome of biologically relevant DNA regions. These include the protein- and RNA-coding genes, as well as a number of signals such as the ribosome binding sites, the Rho-independent transcription terminators, and some promoters and regulatory signals (see below). The methods used for creating this set of genomic objects, notably using the integrated and cooperative computer environment for sequence annotation Imagene (4), have already been extensively described (1,5).

The second and major section of the database comprises information related to the functional assignments of protein- and RNA-coding genes. This information is distributed into several tables and fields of the database, such as the ‘Function’ and ‘Product’ fields of the ‘Gene’ table. A ‘Description’ field is also used for every gene to provide a short summary of the annotation, which is used, for example, in gene list formats. The sources of this information are multiple: original annotation by authors in earlier World Wide DNA Data Library (WWDDL, also known as DDBJ/EMBL/GenBank) entries, information extracted from the *B.subtilis* genetic map (6). However, the source that is now the most informative is the direct browsing of the literature relevant to *B.subtilis*. This time-consuming curation is nevertheless the most exhaustive and accurate way to get detailed and up-to-date information about *B.subtilis* gene functions and products. Automatic selection of the relevant references is made in a very broad fashion, and manual refinement is then performed to keep the most relevant references only (also including references not closely related to *B.subtilis*, yet providing useful information). Each reference is then linked to one or several genes, combined with one or several relevance values for each gene/reference pair that indicates why this particular reference was selected. In parallel, information is manually extracted from the abstracts and/or full papers to update the description fields of the related genes. Other important information about gene annotation includes a dictionary of synonymous names—useful for tracking history of changes of gene names—and the functional category each gene is assigned to (note that a gene may belong to several categories, which are browsable through a hierarchical representation). Functional assignments of the genes of unknown function (‘y’ genes, see below) are mainly based on sequence similarity data. Since this information may evolve more rapidly than it is manually updated, Smith and Waterman scanning reports of a non-redundant protein databank—run on a multi-processor Paracel GeneMatcher (7)—are linked to every *B.subtilis* protein gene, and updated on a regular basis (at least once a month). Finally, biochemical information, e.g. calculated molecular mass, estimated isoelectric point, and results of more elaborate data analysis, e.g. gene classification based on codon usage (8), are available too.

Other information connected to genes is present in the database. In particular, thanks to the assignment of unique accession numbers to every gene, several cross-references to other data collections have been established. Links to SWISS-PROT

B.subtilis protein entries are of particular importance since they provide access to this reliable resource of annotated protein information (9), notably in the framework of the new HAMAP project. Other cross-references include links to some *B.subtilis*-dedicated data collections, such as the proteomic resource Sub2D (10), the phenotypic resource Micado (11) and the Japanese functional analysis resource JAFAN (12). Recently, documentary information has been integrated: illustrations extracted from the new edition of the *Bacillus* 'bible' (13), which summarize complex regulatory systems or cellular processes, are dynamically linked to the relevant genes.

Finally, the third section of the database is related to regulation data. Dedicated tables and fields of the data schema have been created for storing this information, which originates from several sources. First, data are extracted from the current gene annotations and the relevant literature. The recent integration of the DBTBS (database of *B.subtilis* promoters and transcription factors), kindly provided by K. Nakai and co-workers (14), dramatically increases the amount of information on transcription signals and regulators. σ^A site predictions are provided by Jarmer *et al.* (15), associated with the compilation of experimental σ^A signals performed by Helmann (16). Finally, a renewed annotation of Rho-independent transcription terminators has recently been performed, using the Termit program (C.Thermes, unpublished).

LATEST DATA UPDATE (RELEASE R16.1)

Data release R16.1 of SubtiList was frozen on May 2001. Major modifications were included in this update as compared to the previous one (data release R15.1, June 1999). First, following a systematic scanning of the whole genome sequence and features using appropriate procedures developed for that purpose (17), sequence corrections were performed and integrated into the database. This crucial step of quality assessment in the genome sequencing program was performed after completion of the sequence: this was especially required because of the collaborative nature of the program (up to 34 laboratories involved) (1). Due to the expected unevenness of the sequence quality, regions containing putative errors were targeted through BLASTX and GeneMark analysis and re-sequenced from direct PCR amplification products. New sequences (up to 6% of the *B.subtilis* genome) were then compared to the original ones and re-annotated, in order to decide whether they should be incorporated into the reference sequence. A number of other sequence corrections communicated by individual users were considered as well. On the whole, 288 sequence corrections (often clustered in chromosome regions) were included in the *B.subtilis* genome sequence, significantly improving its overall quality. Note that the integration of these sequence modifications (and subsequent feature updates) was facilitated by the split structure described above.

Sequence corrections, mainly frameshift errors, led in most cases to the modification of gene boundaries in the region of the sequence update. In a few cases, more dramatic changes, such as merging several genes or adding/removing genes, were required. Furthermore, additional corrections of gene locations resulted from a revised analysis of the annotations (e.g. re-assessment of some start codons), independently of the sequence corrections. About 180 genes in total were consequently updated (Table 1).

Table 1. Summary of updates in SubtiList data release R16.1 (May 2001)

Type of modification	Number of genes
Genomic sequence changed	
Location updated (start and/or stop codons)	67
Substitutions	3
Internal compensating frameshift	2
Two genes merged into one single gene	18 (→9)
Three genes merged into one single gene	3 (→1)
One gene split out into two genes	3 (→6)
New genes added in the annotations	5
Genes deleted from the annotations	2
Genomic sequence unchanged	
Location updated (start and/or stop codons)	71
New genes added in the annotations	8
Genes deleted from the annotations	6
Gene name changed	
'y' → not-'y'	181
Not-'y' → not-'y'	54
Not-'y' → 'y'	4
Description updated	~800

Finally, updates of gene functions were performed. They were based both on the consultation of the recent literature related to *B.subtilis* (see above) and on personal communications. This resulted in a number of changes of gene names, especially genes whose name began with the letter 'y' (meaning that their function was unknown), which were renamed with significant names pertinent to their newly identified function. Altogether, 520 new references were imported and linked to the relevant genes, 239 genes were renamed, amongst them 181 'unknown' genes, and about 800 genes were updated—regarding any feature, including description fields (figures compared to data release R15.1) (Table 1).

The new version of the *B.subtilis* genome sequence and annotations has been submitted to the WWDDL (accession no. AL009126). Flat files are available at the SubtiList World Wide Web server (genome and gene/protein sequences, annotations in various formats). Furthermore, in order to help users track data changes, history files are provided in two different formats: easy to read and clickable HTML pages on the one hand, easy to parse and strictly formatted files (identifier-value line syntax) on the other hand, that can be used by automatic procedures to duplicate data changes in other datasets (e.g. results of complex analysis involving genome coordinates).

FUTURE PROSPECTS

The SubtiList database is accessible at http://genolist.pasteur.fr/SubtiList/. A mirror site has been set up in Beijing (http://genolist.mirror.edu.cn/SubtiList/). The generic data model and engine that SubtiList is based on, GenoList, makes it possible to set up similar databases for other bacterial species. Indeed, five other databases are already available through World Wide Web servers comparable to SubtiList; most of them also benefited from specialized curation analogous to that

Table 2. GenoList-based bacterial genomic databases

Name	Species	Collaboration	Reference
Colibri	*Escherichia coli* K-12	K.E.Rudd (University of Miami)	(18)
TubercuList	*Mycobacterium tuberculosis* H37Rv	S.T.Cole (Institut Pasteur)	(19)
PyloriGene	*Helicobacter pylori* 26695 and J99	A.Labigne, H.de Reuse (Institut Pasteur), P.Legrain, Y.Chemama (Hybrigenics)	I.G.Boneca *et al.*, manuscript in preparation.
MypuList	*Mycoplasma pulmonis* UAB CTIP	A.Blanchard, I.Chambaud (INRA)	(20)
Leproma	*Mycobacterium leprae* TN	S.T.Cole (Institut Pasteur)	(21)

World Wide Web servers are accessible at the respective URLs http://genolist.pasteur.fr/<database_name>.

performed for *B.subtilis* (Table 2). New functionalities for multi-genome integration are being developed in GenoList, as exemplified by the multi-strain features of PyloriGene, and a larger number of microbial genomes will be integrated into the same data structure in the near future.

In parallel, a new section of SubtiList is being developed to provide an adequate framework for collecting, querying and analyzing transcriptome data, especially those generated as part of the 'BACELL Network' program funded by the European Commission (http://www.ncl.ac.uk/bacellnet/, contract QLG2 CT9901455). This new database section complies with international standards about expression data storage that are being defined world wide, including the list of minimum information about microarray experiments, defined by the international Microarray Gene Expression Database (MGED) group.

Despite all the care taken in the construction of SubtiList, mistakes in the data and defects in the program may still remain. Information and comments about gene names and functions, as well as problems and possible improvements concerning the World Wide Web server engine, are welcomed.

ACKNOWLEDGEMENTS

Initial development stages of SubtiList were performed in the framework of the European *B.subtilis* genome project (European Commission Biotechnology program contracts BIO2-CT93-0272, BIO2-CT94-2011, BIO4-CT96-0655), coordinated by F. Kunst and supported by the *BACillus* Industrial Platform (BACIP). Early developments of the Colibri and SubtiList databases are owing to the work of P. Glaser, A. Hénaut, C. Médigue and A. Viari. We thank G. Fang for setting up the Beijing mirror, I. Martin-Verstraete for fruitful discussions, and M. Pupin for her contribution to the GenoList project. We acknowledge helpful messages received from the *Bacillus* international community and the fruitful collaboration with A. Bairoch and the SWISS-PROT team.

REFERENCES

1. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessières,P., Bolotin,A., Borchert,S. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.
2. Médigue,C., Viari,A., Hénaut,A. and Danchin,A. (1993) Colibri: a functional data base for the *Escherichia coli* genome. *Microbiol. Rev.*, **57**, 623–654.
3. Moszer,I., Glaser,P. and Danchin,A. (1995) SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology*, **141**, 261–268.
4. Médigue,C., Rechenmann,F., Danchin,A. and Viari,A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15**, 2–15.
5. Moszer,I. (1998) The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett.*, **430**, 28–36.
6. Biaudet,V., Samson,F., Anagnostopoulos,C., Ehrlich,S.D. and Bessières,P. (1996) Computerized genetic map of *Bacillus subtilis*. *Microbiology*, **142**, 2669–2729.
7. Horton,B. (1998) Sequence cyberbiology. *Nature*, **393**, 603–606.
8. Moszer,I., Rocha,E.P. and Danchin,A. (1999) Codon usage and lateral gene transfer in *Bacillus subtilis*. *Curr. Opin. Microbiol.*, **2**, 524–528.
9. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
10. Bernhardt,J., Buttner,K., Scharf,C. and Hecker,M. (1999) Dual channel imaging of two-dimensional electropherograms in *Bacillus subtilis*. *Electrophoresis*, **20**, 2225–2240.
11. Biaudet,V., Samson,F. and Bessières,P. (1997) Micado: a network-oriented database for microbial genomes. *Comput. Appl. Biosci.*, **13**, 431–438.
12. Ogasawara,N. (2000) Systematic function analysis of *Bacillus subtilis* genes. *Res. Microbiol.*, **151**, 129–134.
13. Sonenshein,A.L., Hoch,J.A. and Losick,R. (2001) *Bacillus subtilis and Its Closest Relatives: From Genes to Cells*. American Society for Microbiology, Washington, DC.
14. Ishii,T., Yoshida,K., Terai,G., Fujita,Y. and Nakai,K. (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res.*, **29**, 278–280.
15. Jarmer,H., Larsen,T.S., Krogh,A., Saxild,H.H., Brunak,S. and Knudsen,S. (2001) σ^A recognition sites in the *Bacillus subtilis* genome. *Microbiology*, **147**, 2417–2424.
16. Helmann,J.D. (1995) Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res.*, **23**, 2351–2360.
17. Médigue,C., Rose,M., Viari,A. and Danchin,A. (1999) Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence. *Genome Res.*, **9**, 1116–1127.
18. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.
19. Cole,S.T. (1999) Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett.*, **452**, 7–10.
20. Chambaud,I., Heilig,R., Ferris,S., Barbe,V., Samson,D., Galisson,F., Moszer,I., Dybvig,K., Wroblewski,H., Viari,A., Rocha,E.P.C. and Blanchard,A. (2001) The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.*, **29**, 2145–2153.
21. Jones,L.M., Moszer,I. and Cole,S.T. (2001) Leproma: a *Mycobacterium leprae* genome browser. *Lepr. Rev.*, in press.