© 1998 Oxford University Press

The non-redundant *Bacillus subtilis* (NRSub) database: update 1998

Guy Perrière*, Manolo Gouy and Takashi Gojobori[1]

Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard, Lyon 1,
43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France and [1]Center for Information Biology,
National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan

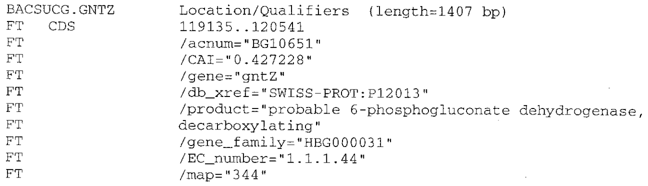
Received September 29, 1997; Accepted October 3, 1997

ABSTRACT

The non-redundant *Bacillus subtilis* database (NRSub) has been developed in the context of the sequencing project devoted to this bacterium. As this project has reached completion, the whole genome is now available as a single contig. Thanks to the ACNUC database management system and its associated retrieval system Query_win, each functional region of the genome can be accessed individually. Extra annotations have been added such as accession numbers for the genes, locations on the genetic map, codon adaptation index values, as well as cross-references with other collections. NRSub is distributed through anonymous FTP as a text file in EMBL format and as an ACNUC database. It is also possible to access NRSub through two dedicated World Wide Web servers located in France (http://acnuc.univ-lyon1.fr/nrsub/nrsub.html) and in Japan (http://ddbjs4h.genes.nig.ac.jp/).

INTRODUCTION

As the *Bacillus subtilis* genome sequencing project has now reached completion, the whole genome is now available as a single sequence contig. The sequencing project of this organism started 7 years ago and was completed by May 1997 (1–4). In the context of this project, the first public release of the non-redundant *B.subtilis* database (NRSub) was made available in July 1994 (5). Since this date, regular updates were provided. In January 1995, the NRSub database started to use sequence contigs assembled at Pasteur Institute for the SubtiList database (6). These contigs were built from the *B.subtilis* 168 (and derivatives) chromosomal sequences available from the EMBL (7), GenBank (8) and DDBJ (9) collections by removing all the redundancy. Sequences from strains other than 168 and plasmidic sequences were not considered. Many additional data, not available in EMBL/GenBank/DDBJ and SubtiList databases are provided by NRSub: (i) a measure of codon usage bias for protein genes, (ii) references to similar genes in all other bacterial species, (iii) systematic cross-references to other molecular biology databases such as SWISS-PROT (10) and ENZYME (11). Release 10 of NRSub (October 1997) contains a single sequence of 4 214 814 bp long corresponding to the whole *B.subtilis* genome. This contig allows to access 4100 protein genes, 30 rRNA and 88 tRNA genes.



```
BACSUCG.GNTZ        Location/Qualifiers  (length=1407 bp)
FT   CDS            119135..120541
FT                  /acnum="BG10651"
FT                  /CAI="0.427228"
FT                  /gene="gntZ"
FT                  /db_xref="SWISS-PROT:P12013"
FT                  /product="probable 6-phosphogluconate dehydrogenase,
FT                  decarboxylating"
FT                  /gene_family="HBG000031"
FT                  /EC_number="1.1.1.44"
FT                  /map="344"
```

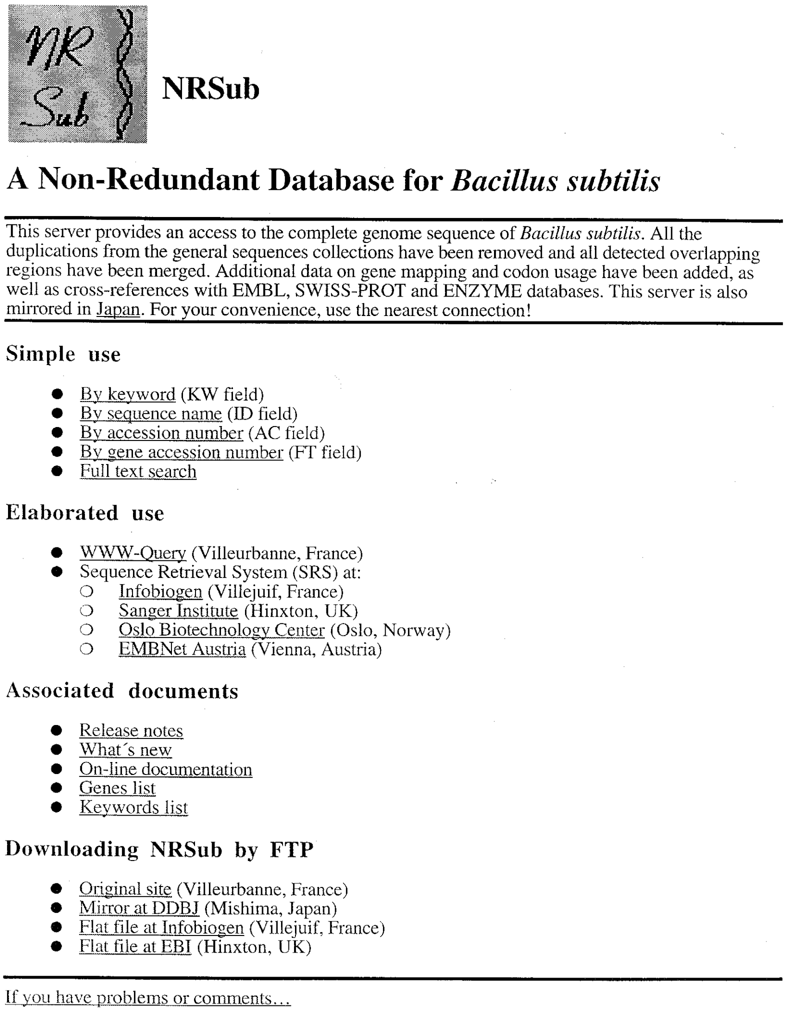
Figure 1. Structure of the FT field for a protein gene in NRSub. Three additional, non-standard qualifiers have been defined in a way to introduce specific informations: '/acnum', for the accession number, '/CAI', for the codon adaptation index value, and '/gene_family' for the accession number of the corresponding gene family, if any.

DATABASE CONTENT

The main difference between the present release of NRSub and the previous ones is that the contents of the KW, RA, RT, RL and RX fields of the EMBL/GenBank/DDBJ entries are no longer merged in the contig corresponding to the whole genome. Indeed, the length of the resulting fields would have been extremely large and their usefulness questionable. On the other hand, we have still performed our usual corrections on the information provided by the general sequence databases. These corrections took place mainly in the FT field of the unique entry (named BACSUCG) now integrated in NRSub (Fig. 1). For each protein or structural RNA gene we have added an accession number under the '/acnum' qualifier. Thanks to this introduction of gene accession numbers it was possible to cross-reference SWISS-PROT with NRSub. This reference can be found in the DR field of *B.subtilis* SWISS-PROT entries on the ExPASy server (12). When the feature was a Coding DNA Sequence (CDS) we have added a Codon Adaptation Index (CAI) (13) value under the '/CAI' qualifier. These CAI values were established with the table we have previously published (5).

When information on locus name was available we have added a '/gene' qualifier for the CDS or the structural RNA described. The gene names used were taken from the recently published computerized genetic map of *B.subtilis* (14) or from the original publications. In some cases we have corrected the names when they were inconsistent with the official nomenclature (15). To name Open Reading Frames (ORFs) we have used a naming convention similar to the one employed for *Escherichia coli* (16): as there is no gene name in *B.subtilis* that begins with y, we have

*To whom correspondence should be addressed. Tel: +33 4 72 44 62 96; Fax: +33 4 78 89 27 19; Email: perriere@biomserv.univ-lyon1.fr



NRSub

A Non-Redundant Database for *Bacillus subtilis*

This server provides an access to the complete genome sequence of *Bacillus subtilis*. All the duplications from the general sequences collections have been removed and all detected overlapping regions have been merged. Additional data on gene mapping and codon usage have been added, as well as cross-references with EMBL, SWISS-PROT and ENZYME databases. This server is also mirrored in Japan. For your convenience, use the nearest connection!

Simple use

- By keyword (KW field)
- By sequence name (ID field)
- By accession number (AC field)
- By gene accession number (FT field)
- Full text search

Elaborated use

- WWW-Query (Villeurbanne, France)
- Sequence Retrieval System (SRS) at:
  - Infobiogen (Villejuif, France)
  - Sanger Institute (Hinxton, UK)
  - Oslo Biotechnology Center (Oslo, Norway)
  - EMBNet Austria (Vienna, Austria)

Associated documents

- Release notes
- What´s new
- On-line documentation
- Genes list
- Keywords list

Downloading NRSub by FTP

- Original site (Villeurbanne, France)
- Mirror at DDBJ (Mishima, Japan)
- Flat file at Infobiogen (Villejuif, France)
- Flat file at EBI (Hinxton, UK)

If you have problems or comments…

Figure 2. Home page of the NRSub World Wide Web server.

used this letter as the first letter in naming each ORF. The second letter was assigned according to the group in charge of sequencing the region of the chromosome in which the ORF was located. In the case of an unmapped gene the letter z was used. The third letter was freely chosen by the group sequencing the region. Preferably, each different letter corresponding to a different operon. In the case of genes obtained by piecemeal sequencing the letters x or y were used. The fourth and last letter was always a capital letter, freely chosen by the group sequencing the region. Preferably genes in the same operon were ordered using the alphabet from A to Z.

If alternative names were available for a given gene they were indicated under the '/alt_name' qualifier. In the case of CDS we have added a '/db_xref' qualifier for the cross-reference pointing to the corresponding entry in SWISS-PROT. We have rewritten or completed the '/product' qualifier using data from SWISS-PROT. In the case of ORFs, we used 'hypothetical protein' for the product associated with these putative genes. When an encoded protein was an enzyme, we have added its EC number in the '/EC_number' qualifier. The EC numbers used were taken from ENZYME. For each protein or structural RNA gene we have added its location on the genetic map of *B.subtilis* 168 in the '/map' qualifier. Finally, when the gene was known to belong to a family defined in the Hobacgen database (manuscript in preparation), we added the accession number of this family in a '/gene_family' qualifier.

Some elements of the original EMBL/GenBank/DDBJ features were discarded. Thus we have kept only the descriptions of signal, mature and leader peptides, CDS, tRNA, rRNA, –35 and –10 regions, promoters and terminators. Some mistakes were corrected, consisting mainly of signal, mature or leader peptides wrongly annotated as CDS, frameshifts resulting from bad start points, CDS shortened due to bad end points and in features described in the original publications but not inserted in the tables.

DATABASE USE

As a way to spread the information widely we distribute primarily NRSub as a text file in EMBL format. This format is recognized by many sequence analysis packages and retrieval systems, and is the standard for all the European Bioinformatic Institute (EBI) databases (7). To make access to NRSub easier, we have indexed this text file with the ACNUC sequence database management system (17). This system allows for the indexing of all collections in EMBL/SWISS-PROT, GenBank/DDBJ or NBRF/PIR (18) formats. Under ACNUC, each protein or structural RNA gene can be seen as an independant sequence and keywords can be attached to these subsequences. For instance the contents of the '/gene', '/product' and '/EC_number' qualifiers are inserted as keywords associated to the corresponding subsequences in the feature table.

The graphical retrieval system provided with ACNUC is called Query_win. This program is written in C and uses the Vibrant library which is a part of the toolbox distributed by the National Center for Biotechnology Information (NCBI) (19). Binaries of Query_win are available for UNIX workstations (Sun, DEC Alpha, IBM RS/6000, HP/UX, Silicon Graphics), and for all microcomputers under MacOS 7.1 (or higher) and Windows 95 (or higher) operating systems. Under Query_win, interrogations can be made on sequence names, accession numbers, keywords, bibliographic references, dates of insertion in the bank, etc. In the case of keywords, a graphical browser allows to parse their tree structure in such a way as to find the ones matching a template given by the user. Query_win allows the extraction of selected sequences or parts of them into data files, and offers the possibility to translate CDS. It also integrates a tool for string search in sequences annotations. Each kind of feature can be accessed and extracted as well as its flanking regions. Different formats are available for extracting sequences including FASTA, GCG and EMBL. Query_win integrates an on-line help mode that the user can invoke at any time during the session. General help describes main ACNUC concepts, and detailed help describes each command and its options.

A World Wide Web (WWW) server allowing access to NRSub has been set up in France (http://acnuc.univ-lyon1.fr/nrsub/nrsub.html) and a mirror has been installed in Japan (http://ddbjs4h.genes.nig.ac.jp/). The home page of the server gives access to entry points allowing one to make simple or complex queries (Fig. 2). Simple queries are made by keyword, sequence name, accession number, gene accession number and full text search. More sophisticated access is possible through WWW-Query, a WWW interface for the different ACNUC databases installed on our server (20). WWW-Query includes the main features of Query_win and permits the selection of sequences using different criteria linked by logical operators. The server also gives access to various documents such as release notes, a history of the database, an on-line documentation, a list of keywords and a list of all the protein genes accessible in NRSub. This list includes the accession numbers of these genes in NRSub and the accession numbers of their corresponding proteins in SWISS-PROT. Pointers to other servers related to *B.subtilis* are also listed.

Finally, NRSub has been indexed for use with the Sequence Retrieval System (SRS) (21) on four different European sites: Infobiogen (Villejuif, France), the Sanger Institute (Hinxton, UK), the Oslo Biotechnology Center (Oslo, Norway), and the Austrian EMBNet Center (Vienna, Austria). Links toward the SRS page of these servers are available on the home page of NRSub.

AVAILABILITY

The NRSub text file, as well as the ACNUC index tables, the sources and the binaries of Query_win are available through two anonymous FTP servers: one in France (ftp://biom3.univ-lyon1.fr/pub/nrsub) and one in Japan (ftp://ftp.nig.ac.jp/pub/db/nrsub). The sources of Query_win include a C library allowing one to interface user-developed applications with any ACNUC database. The text file is mirrored at the EBI (ftp://ftp.ebi.ac.uk/pub/databases/nrsub) and Infobiogen (ftp://ftp.infobiogen.fr/pub/db/nrsub) FTP servers. Any questions and comments related to NRSub can be sent by Email to the corresponding author (perriere@biomserv.univ-lyon1.fr).

PERSPECTIVES

We want to continue our effort on annotations improvement, firstly by adding data provided by comparative genomics (similarities between protein sequences and protein structure from other bacteria, organization of genes into families, etc.). Introduction of data such as CDS orientation on the chromosome or accession numbers for all genomic fragments of biological interest (ribosome binding sites, promoters, operators and terminators) are also considered.

Then, as the complete genome of *B.subtilis* is now available, we have planned to replace NRSub by a database using the same format and integrating all complete bacterial genomes. This database, named NRBact for non-redundant bacterial database, is already accessible from our WWW-Query server. At this point, NRBact contains the complete genomes of *B.subtilis*, *E.coli*, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Methanococcus jannaschii*, *Cynechocystis* PCC 6803 and *Helicobacter pylori*. All these genomes are integrated as a single contig and not as a set of overlapping entries as in EMBL/GenBank/DDBJ.

REFERENCES

1 Kunst,F. and Devine,K. (1991) *Res. Microbiol.*, **142**, 905–912.
2 Kunst,F., Vassarotti,A. and Danchin,A. (1995) *Microbiology*, **141**, 249–255.
3 Devine,K.M. (1995) *Trends Biotechnol.*, **13**, 210–216.
4 Harwood,C.R. and Wipat,A. (1996) *FEBS Lett.*, **389**, 84–87.
5 Perrière,G., Gouy,M. and Gojobori,T. (1994) *Nucleic Acids Res.*, **22**, 5525–5529.
6 Moszer,I., Glaser,P. and Danchin,A. (1995) *Microbiology*, **141**, 261–268.
7 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res.*, **25**, 7–13. [See also this issue *Nucleic Acids Res.* (1998), **26**, 8–15.]
8 Benson,D.A., Boguski,M.S., Lipman,D.J. and Ostell,J. (1997) *Nucleic Acids Res.*, **25**, 1–6. [See also this issue *Nucleic Acids Res.* (1998), **26**, 1–7.]
9 Tateno,Y. and Gojobori,T. (1997) *Nucleic Acids Res.*, **25**, 14–17. [See also this issue *Nucleic Acids Res.* (1998), **26**, 16–20.]
10 Bairoch,A. and Apweiler,R. (1997) *Nucleic Acids Res.*, **25**, 31–36. [See also this issue *Nucleic Acids Res.* (1998), **26**, 38–42.]
11 Bairoch,A. (1996) *Nucleic Acids Res.*, **24**, 221–222.
12 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.
13 Sharp,P.M. and Li,W.-H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
14 Biaudet,V., Samson,F., Anagnostopoulos,C., Ehrlich,S.D. and Bessières,P. (1996) *Microbiology*, **142**, 2669–2729.
15 Demerec,M., Adelberg,E.A., Clark,A.J. and Hartman,P.E. (1966) *Genetics*, **54**, 61–76.
16 Rudd,K.E. (1993) *ASM News*, **59**, 335–341.
17 Gouy,M., Gautier,C., Attimonelli,M., Lanave,C. and di Paola,G. (1985) *Comput. Applic. Biosci.*, **1**, 167–172.
18 George,D.G., Dodson,R.J., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Sidman,K.E., Srinivasarao,G.Y., Yeh,L.-S.L., *et al*. (1997) *Nucleic Acids Res.*, **25**, 24–27. [See also this issue *Nucleic Acids Res.* (1998), **26**, 27–32.]
19 NCBI (1993) NCBI Software Development Toolkit, Version 1.8. National Center for Biotechnology Information, National Library of Medecine, Bethesda, MD.
20 Perrière,G. and Thioulouse,J. (1996) *Comput. Applic. Biosci.*, **12**, 63–69.
21 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol.*, **266**, 114–128.