Connecting parts with processes: SubtiWiki and SubtiPathways integrate gene and pathway annotation for Bacillus subtilis

Christoph R. Lammers,1† Lope A. Flórez,1† Arne G. Schmeisky,1
Sebastian F. Roppel,1 Ulrike Mäder,2 Leendert Hamoen3 and Jörg Stülke1

1Department of General Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Germany
2Interfaculty Institute for Genetics and Functional Genomics, Ernst-Moritz-Arndt-Universität Greifswald, Germany
3Centre for Bacterial Cell Biology, Institute for Cell and Molecular Biosciences, Newcastle University, Newcastle upon Tyne, UK

Bacillus subtilis is the model organism for a large group of Gram-positive bacteria, the Firmicutes. Several online databases have been established over time to manage its genetic and metabolic information, but they differ greatly in their rate of update and their focus on B. subtilis. Therefore, a European systems biology consortium called for an integrated solution that empowers its users to enrich online content. To meet this goal we created SubtiWiki and SubtiPathways, two complementary online tools for gene and pathway information on B. subtilis 168. SubtiWiki (http://subtiwiki.uni-goettingen.de/) is a scientific wiki for all genes of B. subtilis and their protein or RNA products. Each gene page contains a summary of the most important information; sections on the gene, its product and expression; sections concerning biological materials and laboratories; and a list of references. SubtiWiki has been seeded with key content and can be extended by any researcher after a simple registration, thus keeping it always up to date. As a complement, SubtiPathways (http://subtipathways.uni-goettingen.de/) is an online tool for navigation of the metabolism of B. subtilis and its regulation. Each SubtiPathways diagram presents a metabolic pathway with its participating enzymes, together with the regulatory mechanisms that act on their expression and activity, in an intuitive interface that is based on Google Maps. Together, SubtiWiki and SubtiPathways provide an integrated view of the processes that make up B. subtilis and its components, making it the most comprehensive web resource for B. subtilis researchers.

INTRODUCTION

Bacillus subtilis serves as the model for a large group of Gram-positive bacteria with a low G+C content in their genomic DNA: the Firmicutes. This group includes important pathogens such as Staphylococcus aureus, Listeria monocytogenes, Bacillus anthracis and Clostridium botulinum. The lactic acid bacteria, which are widely used in the dairy industry, important enzyme producers such as Bacillus licheniformis, and the insect pathogen Bacillus thuringiensis, which is used for crop protection, are also all members of the Firmicutes phylum. Finally, the mollicutes, such as Mycoplasma genitalium, are a phylogenetic branch of the Firmicutes that experienced substantial evolution leading to the smallest genomes that allow host-independent life.

†These authors contributed equally to this work.

The genome sequence of B. subtilis was first determined as a joint European and Japanese effort (Kunst et al., 1997). With the availability of the genome information and detailed experimental data on metabolic pathways and their players, the metabolism of B. subtilis is today quite well understood. There are models of the metabolic and regulatory pathways of B. subtilis available in the literature (Goelzer et al., 2008; Henry et al., 2009; Oh et al., 2007). Nonetheless, even today, more than 10 years after the publication of the original genome sequence, about 30% of the genes of B. subtilis have no defined functions. However, sporadic pieces of information have become available for many of the unknown genes. This information is mainly derived from genome-level analysis such as proteomic and transcriptomic studies as well as from global interaction screenings. This information may provide clues to the function of a certain gene (e.g. if the gene is expressed...
during sporulation, the function of the encoded protein will most likely be related to the sporulation process. Clearly, efficient data management is required to explore the function of the 1370 unknown genes and to obtain novel insights into the functions and molecular activities of those genes and gene products that are already under investigation.

To enable access to information on the genes and proteins of *B. subtilis*, the SubtiList database was created and subsequently integrated into GenoList, a suite of microbial genome databases (Lechat et al., 2008; Moszer et al., 1995, 2002). However, this database is not updated frequently enough to keep pace with the rate of ongoing research (Barbe et al., 2009).

In addition to GenoList, much of the desired information is provided in other centrally curated databases. Today, we use general databases such as Swiss-Prot or the collection of NCBI databases, databases with some focus such as GenoList or Prodoric (Grote et al., 2009) that are centred on a group of micro-organisms, and strongly specialized databases such as DBTBS (Sierro et al., 2008), which provides information on transcriptional regulation in *B. subtilis*. All these databases are very valuable tools, but they differ substantially in the frequency of updates and thus in the timeliness of the information they provide. Moreover, the episodic nature of the information that is available for many of the unknown genes makes it very difficult to store it in a traditional relational database. Therefore, even if published, this information is not easily accessible for the scientific community.

Databases on metabolism such as KEGG or BioCYC are usually focused on the presentation of the metabolic pathways. It would, however, be desirable to use databases that link information on metabolic pathways and their regulation to the expression of the genes that encode the enzymes of the pathway. This problem could be solved by using diagrams that are generated using the Systems Biology Graphical Notation as used by the CellDesigner software (Kitano et al., 2005; Le Novère et al., 2009).

METHODS

Wiki implementation. *SubtiWiki* runs on a MediaWiki installation hosted and maintained at the Gesellschaft für Wissenschaftliche Datenverarbeitung (GWDG) in Göttingen, Germany. The functionality of the wiki is enhanced by three third-party extensions. The reCAPTCHA extension (http://www.mediawiki.org/wiki/Extension:ReCAPTCHA; von Ahn et al., 2008) prevents the creation of user accounts using scripts and serves as a primary step to prevent malicious modification of the wiki. The ContributionCredits extension (http://www.mediawiki.org/wiki/Extension:ContributionCredits) gives credit to the users that add content to the wiki. It creates a list at the bottom of each page with the names of all contributing authors. The third extension is PubMed (http://www.mediawiki.org/wiki/Extension:Pubmed) and is used to present the relevant references for each gene. It uses the web service of the NCBI to extract the full citation of an article based on e.g. the PubMed identifier (PMID). It then formats this information in the structure that a certain community requires. Thus, *SubtiWiki* provides information on the enzymes and metabolites. The WikiPathways as a complement to all genes and proteins, such as WikiGenes or WikiProteins, provides a complementary visual presentation of metabolic and regulatory pathways in *B. subtilis* as well as links to detailed information on the enzymes and metabolites. The diagrams were created with CellDesigner so as to allow their integration into systems biological applications. Together, *SubtiWiki* and *SubtiPathways* are a comprehensive and up-to-date source of information related to all aspects of research on *B. subtilis*.

Seeding of the wiki. *SubtiWiki* was seeded as described previously (Florez et al., 2009). Briefly, a template page was created with the skeleton structure of all the gene pages. Then a Python (http://www.python.org) script was used to create a copy of this template for every gene of *B. subtilis*. The gene names, basic biochemical
information (gene and protein lengths, pl and molecular mass of the protein) as well as the genetic context were extracted from SubtiList and inserted into SubtiWiki via scripts. Additional pages were created semi-automatically to include the new gene annotations in the recently published genome sequence (Barbe et al., 2009). Redirects from old gene designations to their current new designations, as well as the reannotation of the description, function and product of each gene, are based on an extensive recompilation of literature from the years 2003 to the present. Further Python scripts added additional information from various sources such as DBTBS (Sierro et al., 2008), GenoList (Lechat et al., 2008), MPIDB (Goll et al., 2008) and SwissProt (Uniprot Consortium, 2009), as well as selected publications. These publications were chosen due to their genome-scale approach (Florez et al., 2009). Finally, each gene page was linked to the relevant entries in EMBL-Bank (gene and protein sequences; Kulikova et al., 2007), KEGG (gene pages in this database; Kanehisa et al., 2008) and PDB (structure of the proteins; Berman et al., 2007).

Creation of the metabolic and regulatory diagrams. The diagrams were created manually using CellDesigner (Kitano et al., 2005). They are based on a previously published metabolic and genetic reconstruction (Goelzer et al., 2008). This reconstruction was then significantly expanded and updated based on the KEGG database and an extensive literature research. The PubChem identifier was collected for the metabolic compounds. In addition, the Swiss-Prot/Uniprot identifier was collected for each protein.

Development of a navigation interface. The online navigation interface for the diagrams was implemented using version 2 of the Google Maps Application programming interface (API), which is freely available at http://code.google.com/apis/maps/. The first step was to export the diagrams from CellDesigner to the PNG file format. This image was then processed using Python scripts to create tiles for all zoom levels of the metabolic and regulatory map. The images were then uploaded to the server of the GWGD. Separately, a file was created with markers for all proteins and metabolites based on the specification in the Google Maps API and uploaded to the same server. To show the structure of proteins in the marker’s pop-up window, a copy of the pdb file (downloaded from http://www.pdb.org/) was uploaded to the GWGD server and JavaScript code to load the Jmol applet (http://www.jmol.org/) was added to the marker’s file. The position and caption of all the markers were extracted from the Systems Biology Markup Language (SBML) file of CellDesigner using Python. Finally, an HTML file with embedded JavaScript was created for each diagram. These files contain the code that connects the information on the GWGD servers with the software functionality in the Google servers, thus providing the dynamic interface.

RESULTS AND DISCUSSION

The concept of SubtiWiki and SubtiPathways

The key idea in establishing SubtiWiki was to provide the community with an easily accessible tool that gives an overview of the most relevant information on each gene and protein of B. subtilis and the possibility of decentralized input of data by the scientific community.

SubtiWiki is centred on the genes and proteins of B. subtilis, with a dedicated page for each gene/protein. All pages have the same principal design to facilitate orientation. Moreover, we decided to use a rather simple design that allows easy modification and addition of novel information even by those users who are at the beginner level in the use of wikis. Each page provides the most essential information on the gene/protein at the top in a small table, and detailed information can be found in the lower part of the pages. This information is basically divided into three parts: (i) molecular biology of the gene/protein, (ii) the research on the gene or protein, and (iii) the references (see below). An important feature of SubtiWiki, as for most other wikis, is the extensive use of internal and external links. The internal links provide a quick impression of the relation between one gene/protein and any other gene of B. subtilis, whereas the external links guide the user to relevant databases, structural information, or evidence such as links to publications.

To facilitate the understanding of metabolism and its regulation, we have created a companion site, SubtiPathways. This site provides a graphical presentation of major metabolic pathways and regulatory events that control these pathways in an intuitive way that is based on the Google Maps software. The diagrams allow zooming in and out and use drag-and-drop navigation. Moreover, the proteins and metabolites are clickable and provide links to SubtiWiki and the NCBI database PubChem, respectively. If a structure for a protein is available, then this structure and a link to the corresponding Proteopedia (Hodis et al., 2008) page appear in a pop-up window. Conversely, the SubtiWiki pages for all 750 proteins that appear in SubtiPathways provide links to the corresponding pathway diagram(s).

The current state of SubtiWiki is intended only as a starting point. For future expansion and to keep it as an updated data source SubtiWiki requires the input of the members of the scientific community. To facilitate this, we have made the modification and data entry process very simple. To maintain some control of the data, editing by a user requires prior registration.

The start page

The central element of the SubtiWiki start page (http://subtiwiki.uni-goettingen.de) is a Google-type search box that allows immediate access to the individual gene pages upon entering the gene designation (by using the ‘Go’ button). Moreover, all pages can be searched for the occurrence of any gene name or any other term by using the ‘Search’ button. When searching for a gene or protein, the user may be uncertain about the correct designation since nearly 10% of all genes (357 genes) have received new names in the past few years after the last revision of the SubtiList database (Moszer et al., 2002). Of these newly baptized genes, there are even 71 gene designations supported by published evidence that are not covered in the recently released annotation of the B. subtilis genome (Barbe et al., 2009). When interrogating SubtiWiki, both the old and new designations will lead the user to the same page with the most recent designation.
The start page of SubtiPathways (http://subtipathways.uni-goettingen.de) is organized around a drop-down menu that allows the selection of the pathway of interest.

The lower part of the start pages provides materials that are related to SubtiWiki or SubtiPathways and resources for the Bacillus community. Guided tours and a tutorial that explains how to add new information to SubtiWiki are provided for users unfamiliar with the site. The resources for the community include links to Bacillus laboratories, open positions, and links to other important web pages on Bacillus and databases. In addition, the user can download an Excel file that maps the different gene designations in SubtiWiki, SubtiList and GenoList with the NCBI locus tags, UniProt identifiers and brief functional information.

The last feature on the start pages of SubtiWiki and SubtiPathways is links to Wikipedia-type 'main pages'. The main page for SubtiWiki provides links to some sample gene pages as well as to a collection of additional wiki pages that enrich the contents of SubtiWiki but that are not centred on individual genes or proteins (see below). The SubtiPathways main page again allows selection of a pathway of interest and gives information on the colour coding that is used throughout SubtiPathways.

Gene designations

Since the central entry point of SubtiWiki is a search box and the main elements of the wiki are gene-specific pages, the gene designations are of crucial importance. This is even more the case since using the same designations facilitates communication and information exchange in the scientific community. Traditionally, the designation of B. subtilis genes has three sources. (i) Many genes got their designations when the corresponding phenotypes of mutants or properties of the encoded proteins were studied. (ii) During the initial genome project, many genes got the designations of their E. coli counterparts. (iii) A large set of genes, especially those for which no function was known, got a 'y' designation (Kunst et al., 1997). With ongoing research, 355 of these 'y' genes were functionally analysed and these genes got a new designation.

The nomenclature used in SubtiWiki is based on that of SubtiList. This means that each SubtiList gene name will lead to a gene page, even if the gene has been renamed in the meantime. In those cases, the user will be directly guided to the page with the new name, and the old name will be indicated as a synonym.

For the scientific community it is important that genetic nomenclature remains stable or changes only if new evidence becomes available in the literature. Therefore, we decided to use novel designations only for those genes that have been validly renamed by the corresponding experts. It is important to note that there are 162 new designations in GenoList that are not based on published evidence (Barbe et al., 2009). For these genes, SubtiWiki prefers to use the classic SubtiList designations to which the scientific community is used and which have already appeared in many publications.

Features of the gene pages

SubtiWiki contains individual pages for each of the 4394 protein- or RNA-coding genes of B. subtilis. As mentioned above, all these pages have a similar structure (see Fig. 1). At the very top of the page, there is a short description of the function of the gene product, followed by the table of contents of the page and another table that provides the most important information on the gene and its product. This information includes potential synonyms, functional essentiality, the gene product and its function, links to corresponding SubtiPathways pages, numerical data (molecular mass and isoelectric points for proteins, gene and protein lengths). The table also lists the neighbours on the chromosome, includes a map showing the chromosomal arrangement, and provides a link to the DNA and protein sequences in the EMBL-Bank database (Kulikova et al., 2007).

The second part of the pages provides information on the gene and the gene product (see Fig. 1). The first section is devoted to the gene itself. It lists the phenotype of mutants and provides links to gene-centred databases. The second section describes the gene product, i.e. the protein or the RNA. After some basic information on the biological function of the gene product and homologous proteins, biochemical details such as kinetic parameters, modifications, cofactors, interaction partners and the localization are listed. This section also provides links to databases related to proteins and metabolism including (among others) protein structure databases and EC numbers. The third section contains information on gene expression and regulation such as operon structures, sigma factors, regulators and regulatory mechanisms.

The third part of each page provides information that is related to the research on the gene/protein (see Fig. 2A). This part is intended to support the collaboration within the Bacillus scientific community. One section is devoted to biological materials available in the scientific community such as mutants or expression vectors. For common vectors or methods, there are links to dedicated wiki pages.
(see Fig. 2B). The second section names the experts who study this particular gene or protein, and provides links to a specific SubtiWiki page for that expert (see Fig. 2C) as well as to his/her own homepage. Finally, the last section shows major references and provides direct links to the corresponding PubMed entries.
Other pages that enrich the content

Apart from the gene pages, there are additional wiki pages for plasmids, methods and Bacillus laboratories (see above, Fig. 2). Moreover, there are pages that list specific categories of genes, proteins or RNAs. These lists are accessible from the main page or from the gene pages if appropriate. These pages are of special importance for newly emerging fields such as small RNAs or riboswitches (see Fig. 3). Nonetheless, we decided to keep the focus on the genes and their products. Thus, we refrained from introducing a large number of extra categories of pages. This ensures that the relevant information is available in a concentrated form on the genes pages, and the limited number of additional pages prevents dilution of the information on a large number of pages that might become difficult to find.

Modification of the wiki

A successful wiki requires the contributions of a large community. Therefore, a major concern in the design of SubtiWiki was the possibility of simple editing of each page by any qualified user. After a short registration procedure (the link can be found at the top of each page), the user can log in, and will then be able to add new information. However, registration and logging in is not required for ‘passive’ access to the content. Upon logging in, ‘edit’ links appear on the top of each page and next to each section of SubtiWiki. On the edit screen, there is a toolbar for formatting the entries, and there is some information how to handle PubMed links in the main body of the page and in the references section. Detailed information on the registration and modification procedures can be found in the tutorial on the SubtiWiki start page.

The history of modifications can be accessed for each page, and erroneous or unintentional modifications can be undone at any stage. Moreover, on the ‘history’ page, it can be seen who made which modification. Contributors are also listed at the bottom of each page. This gives credit to the contributors.

Features of SubtiPathways

SubtiPathways is aimed at a comprehensible presentation of metabolic pathways and their regulation in B. subtilis. As a starting point for the generation of SubtiPathways, we created a set of CellDesigner diagrams for B. subtilis metabolic pathways. The diagrams provide an interface for systems biology, as they link information relevant for modellers and bench biologists (Fig. 4). Each diagram consists of two parts, i.e. the metabolic pathways (upper part) and the expression and regulation of the corresponding genes (lower part). If appropriate, metabolic regulation at the level of enzyme activity is also included. We decided to use a colour code for the metabolites to indicate central intermediates of carbon (red) and nitrogen (blue)
metabolism. In all diagrams, the pathways either start or end with these central metabolites.

The diagrams cover as many aspects of metabolism and regulation as possible (Fig. 5). The metabolic pathways (Fig. 5A) show the names of intermediates, products and cofactors as well as the designations of the enzymes responsible. If relevant, the regulation of protein activities by interactions with activating or inhibiting metabolites or proteins or by covalent modification is shown. Accordingly, active and inactive proteins or protein complexes can be distinguished in the diagrams. This applies to the regulation of enzymic activities (see Fig. 5B for an example) and for the control of regulatory proteins (see Fig. 5C). Finally, the expression of the relevant genes and operons is shown in the lower part of each diagram (see Fig. 5D). Starting with the genes in yellow, the resulting mRNAs and proteins are shown. The regulators that control expression of the gene are depicted in their active form, as well as the type of control (positive or negative).

The CellDesigner diagrams were used to generate interactive web pages. These pages were generated using Google Maps software to allow easy navigation on the pages. A list of proteins and metabolites on the right margin allows immediate access to any occurrence of a protein or metabolite in the diagram. Moreover, at the highest zoom levels there are labels that indicate what is shown in the corresponding part of the diagram.

At the highest magnification, all proteins and metabolites are labelled with a clickable icon that provides additional information.
information in a pop-up window. For the proteins, the window contains a short description (taken from SubtiWiki) and a link to the corresponding SubtiWiki page. If available, the structure of the protein and a link to the corresponding Proteopedia page (Hodis et al., 2008) are provided (see Fig. 6A). Moreover, if a protein occurs in multiple diagrams, this is shown in a special tab at the top of the pop-up window. For protein complexes, the pop-up window contains the descriptions and SubtiWiki links for all components of the complex. For metabolites, the pop-up window shows the systematic name of the intermediate (derived from PubChem), a link to the specific PubChem page, and the chemical structure of the compound (Fig. 6B). As for the proteins, occurrence in multiple SubtiPathways diagrams is indicated in a tab.

**Fig. 5.** The different aspects of a gene/protein in SubtiPathways. The diagrams of SubtiPathways cover a broad spectrum of metabolic and regulatory events, including (A) enzymic reactions, (B) feedback mechanisms, (C) regulation of protein activity and (D) regulation of gene expression. They also represent active and inactive states where required (via dotted lines around the protein).
It is important to note that the main aim of SubtiPathways is to allow easy and intuitive navigation in the rather complex metabolic and regulatory networks of B. subtilis. As SubtiWiki is focused on gene products, the proteins are also central in SubtiPathways. These two features distinguish SubtiPathways from other initiatives centred on metabolic pathways such as KEGG, BioCyc or Wikipathways that have the metabolites in the centre or are dedicated to automated processing of pathway information (Kanehisa et al., 2008; Karp et al., 2005; Pico et al., 2008).

Perspectives

So far, we have started to seed SubtiWiki with some basal information. With its extensive internal and external links and the interconnection to SubtiPathways, SubtiWiki is today the most comprehensive source of information on B. subtilis on the web. We are confident that SubtiWiki is already useful for the Bacillus researcher in the lab.

To unfold its full potential, SubtiWiki depends on the collective effort of the B. subtilis scientific community to continuously enter new discoveries and other information and resources they find useful for their daily experimental work.

SubtiWiki might also be useful beyond the B. subtilis scientific community. First, since B. subtilis is a model organism for low-GC Gram-positive bacteria (the Firmicutes), the information provided in SubtiWiki may well be relevant to those who study other organisms.

Second, the layout of SubtiWiki might be a useful basis for a wiki database for other micro-organisms; the source code is freely available on request.

ACKNOWLEDGEMENTS

We are grateful to Lorena Zambrano for help with the creation of the SubtiWiki logo. This work was supported by the Federal Ministry of Education [Research SYMSO network (PtJ-BIO/0313978D)] and the Fonds der Chemischen Industrie to J. S. L. A. F. was supported by the International Molecular Biology Program of the University of Göttingen and the Studienförderwerk Klaus Murmann der Stiftung der Deutschen Wirtschaft.

REFERENCES


Edited by: J. M. van Dijl

http://mic.sgmjournals.org