Systems biology in *Bacillus subtilis*: databases for gene function and software tools for pathway discovery

PhD Thesis

in partial fulfilment of the requirements
for the degree “Doctor rerum naturalium (Dr. rer. nat.)”
in the Molecular Biology Program
at the Georg August University Göttingen,
Faculty of Biology

submitted by

Lope Andrés Flórez Weidinger

born in

Bogotá, Colombia

2010

Members of the Thesis Committee:

Prof. Dr. Jörg Stülke (Supervisor and Reviewer)
Department for General Microbiology
Institute for Microbiology and Genetics
Georg-August-University of Göttingen
Göttingen

Prof. Dr. Burkhard Morgenstern (Reviewer)
Department of Bioinformatics
Institute for Microbiology and Genetics
Georg-August-University of Göttingen
Göttingen

Prof. Dr. Christian Griesinger
Department of NMR-based structural biology
Max Planck Institute for Biophysical Chemistry
Göttingen

Date of the oral examination: November 1st, 2010

I hereby declare that the PhD thesis entitled, “Systems Biology in *Bacillus subtilis:* databases for gene function and software tools for pathway discovery” has been written independently and with no other sources and aids than quoted.

Lope Andrés Flórez Weidinger

Acknowledgements

The last four years in Göttingen, especially the last three years as a graduate student at the Department of General Microbiology, have passed in the blink of an eye. For all the wonderful memories of this time I would like to express my gratitude to everyone involved.

First of all, I would like to thank Jörg, my supervisor, my guide, my mentor. Thanks for the amazing trust in me and for giving me always the very best opportunities to develop and become a better scientist. Thanks for your common sense and your feeling for things that work. Experiencing your way of doing things was perhaps the most important lesson in these three years. Thanks for the honesty, the encouragement, and the good chemistry between us.

I would like to thank Burkhard Morgenstern and Christian Griesinger. Thank you for your time and effort as members of the thesis committee. Thank you especially for all other occasions where we also met, either as your lab rotation student, the tutor of your lecture, or other opportunities.

I would also like to thank Leendert Hamoen, Ulrike Mäder, Jens Baumbach, Rasmus Steinkamp, and Lorena Zambrano for the helpful discussions, and their dedication to technical and design topics, which contributed to better results.

My acknowledgement goes also to the IMPRS of Molecular Biology at the University of Göttingen, the Georg-Lichtenberg Stipend of Lower Saxony, and the Stiftung der Deutschen Wirtschaft for their financial support.

I would like to give very a warm *thank you* to the “wet lab” members of the Department (a.k.a. everybody?). First of all, thanks to Fabian Commichau und Christina Herzberg, for patiently introducing me to all the molecular biology techniques... I admired you right from the beginning and still admire you now. I stopped pipetting soon in my graduate career, but you provided me with the best understanding of what it takes to produce results in the lab.

To the PhD generation before me (Birte, Claudine, Falk, Kalpana, Sebastian H.), thank you for introducing me to the lab culture! (And I don’t mean LB broth). From the very beginning you made me feel part of the group! To “my” PhD generation (Christoph W., Hinnerk, Katrin, Nico, Sebastian S.): as I could spend most time with you in the Department, I am glad to have met you closer, from the daily work, the seminars, but most of all from the myriad events out of the lab... thank you for your comments, for your help, your support. Thank you very much as well for the nice time together in conferences (Copenhagen, Jugendburg castle, San Diego, Tijuana ...)! To the “next PhD” generation (Denise, Fabian R., Frederik, Jens, Martin, Sebastian K., Tini), and the remaining members of the Department, thank you too for your friendliness, comments and good vibes!

In particular I would like to thank all the students that adventured into the “exotic world” of the dry lab. To Fips, Arne, and Repel (no, I won’t use the name “wiki-boys”... oops, I did already), not only a big thank you for all your dedication during the thesis (hope you enjoyed the tequila), but also for serving as guinea pigs for my supervision.

To my lab rotation students (Sinem, Koray, Cadu, Eline, Max): seeing the lab rotations from the other side was very enriching for me. Thanks a lot for choosing my project, it was a lot of fun to be with you in the office; it felt like a continuation from my free time.

The same goes to my two Bachelor students, Adrian and Raphael. Thank you for sharing my enthusiasm for programming and computer science!

To my friends in Göttingen: due to you, I will miss this place *tremendously*! I assure you that I will keep this time very fond in my memory. Hopefully we meet again and again in the future!

To my friends from Colombia, ¡muchísimas gracias por acompañarme en este camino! ¿Sí ven? Ya pronto soy “dotor” (y no de los de corbata :)). Tal vez no les parezca raro que esté acá, pero para mí es parte de un sueño que se hace realidad. ¡Realmente espero que este sueño continue con ustedes y que me sigan dando esa alegría y companía que tanto aprecio!

A mi gran y querida familia de Colombia... ¡muchas gracias por su apoyo! Das Gleiche gilt für Thomas, Monika und Wolfgang... vielen, vielen Dank!

Muito, muito obrigado pela sua ajuda, sua companhia, pelos cuidados que você tem me dedicado, e pelo carinho que você tem me dado. Em muitos momentos críticos, você me disse as palavras que eu mais precisava... estou muito feliz, deixa eu ser o companheiro de sua vida! Obrigado!!!

Y a Chepe, papá y mamá: aquí estoy, ya a punto de terminar mi carrera académica y empezando (ojalá) la vida laboral. Se creció el muchacho. Del “monito” que gateaba y se trapaba a muebles pasamos al que se trepa a estatuas de muchachas con gansos... Gracias, por llevarme a lo que soy hoy. Gracias...

Table of Contents

List of Abbreviations

ADP	Adenosine Diphosphate
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
ATP	Adenosine Triphosphate
B.	*Bacillus*
E.	*Escherichia*
FBA	Flux Balance Analysis
GWDG	Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen
kb	Kilo base pairs
KEGG	Kyoto Encyclopedia of Genes and Genomes
L.	*Listeria*
MILP	Mixed Integer Linear Programming
MOMA	Minimization Of Metabolic Adjustment
MW	Molecular weight
NADH	Nicotinamide adenine dinucleotide (reduced)
NCBI	National Center for Biotechnology Information
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
pI	Isoelectric point
ROOM	Regulatory On/Off Minimization
S.	*Staphylococcus*
SAT	(Boolean) satisfiability
SBGN	Systems Biology Graphical Notation
SBML	Systems Biology Markup Language
SPABBATS	Short PAthways Between a Basis And a Target Set of metabolites
URL	Uniform Resource Locator
XML	eXtensible Markup Language

List of Publications

Published before the thesis:

Herzberg, C., **Weidinger, L.A**., Dörrbecker, B., Hübner, S., Stülke, J., Commichau, F.M. (2007) SPINE: a method for the rapid detection and analysis of protein-protein interactions *in vivo*. *Proteomics* **7**: 4032-4035

Published during the thesis:

Flórez,L.A., Roppel,S.F., Schmeisky,A.G., Lammers,C.R., Stülke,J. (2009) A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki *Subti*Wiki. *Database (Oxford)* **2009**: bap012.

Lammers,C.R.*, **Flórez,L.A.***, Schmeisky,A.G., Roppel,S.F., Mäder,U., Hamoen,L., Stülke,J. (2010) Connecting parts with processes: *Subti*Wiki and *Subti*Pathways integrate gene and pathway annotation for *Bacillus subtilis*. *Microbiology* **156**: 849-859.

* These authors contributed equally to this work.

Manuscripts in preparation (and part of this thesis):

Flórez,L.A., Lammers,C.R., Michna,R., Stülke,J. (2010) CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling, and regulatory pathways. *Bioinformatics* (in revision).

Flórez,L.A., Gunka,K., Polania,R., Tholen,S., Stülke,J. (2010) SPABBATS: A pathway-discovery method based on Boolean satisfiability that facilitates the characterization of suppressor mutants. *BMC Systems Biology* (submitted).

Abstract

Systems biology studies the way in which interaction of different types of molecules results in complex behaviour at the cellular, organism, and higher levels. It approaches these questions using a combination of high-throughput experiments, mathematical and computational models, and information storage in specialized databases. This work has a special focus on the soil bacterium *Bacillus subtilis*. This bacterium became the model for the Gram-positive bacteria due to its natural competence, its simple differentiation program, and the ease of handling in the lab. It is also used in biotechnological processes for enzyme and vitamin production and serves as a model for several pathogenic Gram-positive bacteria, such as *Staphylococcus aureus* and *Bacillus anthracis*.

At the onset of this thesis, the information about the genes, proteins, and metabolic pathways of *B. subtilis* was spread out between several different databases and the scientific literature. In addition, the information present in most of the databases was of poor quality, or was not updated for several years. This hampered the development of systems biology models, as well as the work at the bench. To overcome this problem, *Subti*Wiki and *Subti*Pathways were created. *Subti*Wiki (http://subtiwiki.uni-goettingen.de) is a wiki on all genes and gene products of *B. subtilis*. Its resemblance to Wikipedia makes it easy to use and update, even by new users. Each page of the wiki is structured to find the most relevant information quickly, and is interconnected with other pages and external databases. *Subti*Pathways (http://subtipathways.uni-goettingen.de) is a collection of the main metabolic pathways of *B. subtilis*, together with the enzymes and their regulation. The website is composed of several pathway diagrams created with CellDesigner, a popular program from the systems biology community. These diagrams were enhanced with a navigation based on the interface of Google maps. The proteins and metabolites are linked to external websites, like *Subti*Wiki, the Protein Data Bank, and the PubChem database. To promote the use of the navigation features of *Subti*Pathways, the CellPublisher web server was created (http://cellpublisher.gobics.de). It allows any CellDesigner user to create online diagrams with the Google maps-based navigation features.

This work also addressed the analysis of genome-scale models of metabolism. These models are now available for several organisms (including *B. subtilis*) and consist of a list of all metabolic reactions, with the right stoichiometric coefficients, in a computer readable format. Several applications have been developed for these genome-scale models, in particular to study the intracellular fluxes that are possible under several environmental and genetic constraints. In this study, a computer program termed SPABBATS was developed, that uses these models to extract a list of alternative pathways connecting sets of input and output metabolites. SPABBATS was successfully put to the test in a lab experiment, involving the characterization of suppressor mutants of *B. subtilis*.

The software tools created here can readily be used for systems biology research in other organisms, as well as in related fields, like synthetic biology and metabolic engineering.

1. Introduction

1.1 The purpose of systems biology

Life is a complex phenomenon at many levels. First, life is complex because it requires understanding at many different scales (from the atomic and molecular scale of biophysics, biochemistry, and molecular biology, to the population's scale of ecology and evolutionary biology). Second, life is complex because at each of these scales (except the smallest ones) there are no "basic elements", like e.g. elementary particles in physics or chemical elements. Biomolecules (as well as cells, tissues, organisms, etc.) are the product of (random) evolution instead of being a direct consequence of physical laws and new biomolecules can arise anytime. It is known for instance, that in spite of the rich diversity present in the current proteins, the sequence space of all possible proteins has not yet been explored comprehensively by nature (Povolotskaya & Kondrashov, 2010).

To tackle this complexity, biology is divided into different disciplines, specialized on different scales (e.g. molecular biology to characterize phenomena at the molecular scale). In each of these disciplines, the individual elements are studied in great detail and classified into discrete classes and subclasses (Lehninger *et al.*, 2005). At the same time, links between the different scales are created, e.g. by determining the role of a specific protein in the formation of a tissue or organ (Affolter & Basler, 2007), or by recognizing that a specific gene is essential for the survival of the organism (Kobayashi *et al.*, 2003).

In many (perhaps most) cases, it is not possible to describe a biological phenomenon (i.e. phenotype) at a higher scale based only on the detailed understanding of a single element at a lower scale. Instead, the phenotype arises from a combined, synergistic action from many elements in the lower scale. Drawing an analogy from systems theory, a pattern (i.e. phenotype) at the system level *emerges* from the interactions of individual elements (Kitano, 2002).

The field of systems biology uses principles of systems theory to create links between molecular biology (as well as biochemistry) and higher scales. Systems biology is a rich field with increasing scientific awareness and importance (Marcus, 2008). At the same time, it is very diffuse, and a precise definition that captures all the different system biology efforts is elusive (just as the definition of *emergence* is elusive, see (Bedau & Humphreys, 2008)).

While a definition of systems biology is difficult to obtain, the goals and purpose of it are clearer, so it is common to define the field by its methods and goals (see Klipp *et al.* (2005), Palsson (2006), and Alon (2007)). The field aims to provide a mechanistic understanding of biological phenomena that involve a large collection of different molecules (i.e. a molecular system). In addition, it tries to mimic the behavior of a subsystem of molecules via computer models, to be able to predict values or properties that are difficult or expensive to measure.

1.2 Main approaches in systems biology

As mentioned above, systems biology tries to link cause and effect between interactions of biomolecules and events at higher biological scales. Clearly, this requires a deep understanding of the individual molecules involved, and a precise description (many times quantitative) of the known interactions.

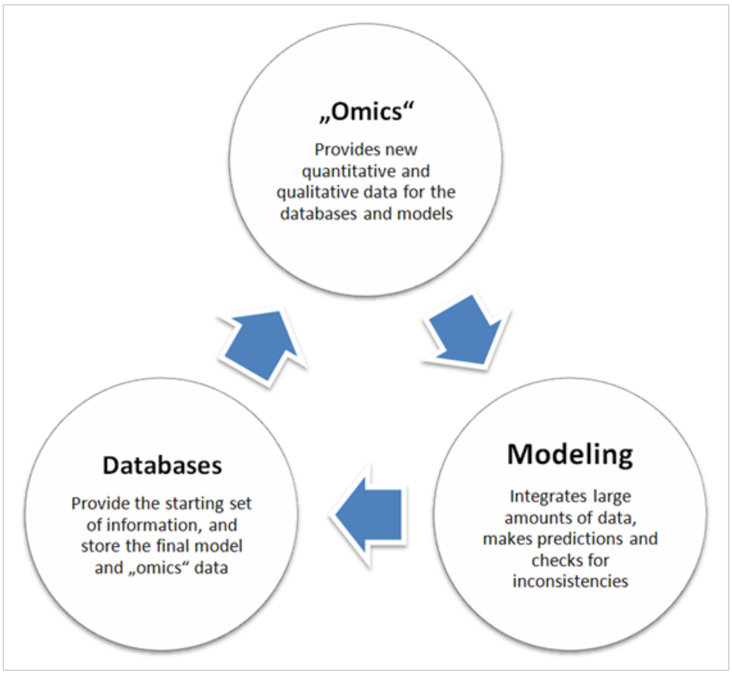

„Omics“
Provides new quantitative and qualitative data for the databases and models
Modeling
Integrates large amounts of data, makes predictions and checks for inconsistencies
Databases
Provide the starting set of information, and store the final model and „omics“ data

Figure 1.1: Main approaches in systems biology. High-throughput data generation, computational modelling, and the utilization of special purpose databases complement each other to generate new knowledge about biological systems.

For this reason, one of the most important approaches in systems biology is exhaustive data collection of specific classes of molecules or interactions (“omics”) (see Figure 1.1). Beside the genome sequence of the organism (genomics), there is an interest in determining and/or quantifying each mRNA (transcriptomics), protein (proteomics), lipid (lipidomics), metabolite (metabolomics), phosphorylation site (phosphoproteomics), protein-protein interaction (interactomics), etc., in the cell, at different conditions. One purpose of “omics” is to correlate the phenotype at the different conditions with the changes in concentration of the individual molecular species (i.e. a top-down approach, Bruggeman & Westerhoff, 2007). The second purpose is to provide qualitative (e.g. network topology) and quantitative (e.g. concentrations) data for other systems biology approaches.

The second approach is computational modeling (Kahlem & Birney, 2006). Again, this approach has at least two purposes. The first one is to integrate large amounts of data into a coherent representation. In this way, it is possible to find new correlations and structural elements present in the current data, such as e.g. novel pathways or functional protein complexes (see Battle *et al*., 2010). It also becomes possible to use tools from systems theory in a biological context (Alon, 2007). The second purpose is to predict the behavior of the system under conditions that have not been measured yet, but are of interest (Lewis *et al*., 2010).

The third approach is the creation of structured databases and other data repositories for the individual elements (i.e. molecules and interactions), as well as languages (written, visual, and computational) to describe biological processes. In most part, characterizing a single biomolecule is the task of biochemistry and molecular biology and many databases are designed for the purposes of these disciplines. Nonetheless, some of the databases have been repurposed and extended to bridge the gap between data collection and modeling. This is evidenced in three sample databases: SABIO-RK for proteins (Rojas *et al*., 2007), Reactome for a collection of reactions and pathways in humans (Matthews *et al.*, 2009), and Biomodels, a repository of quantitative biochemical models (Li *et al.*, 2010). All three databases have the specific purpose to serve as data feeders for quantitative models.

The three approaches complement each other (see Figure 1.1). Usually, the databases provide the foundation of current knowledge about the system. This knowledge is integrated in the form of a model. To calibrate and test the model, high-throughput experiments are conducted. The result of these experiments is contrasted with the expectations from the model. When the model and predictions don't match, the gaps in the knowledge are investigated. In the end, the data, the model, and the experimental conclusions are entered into the databases to start the cycle anew.

1.3 Mathematical description of metabolic fluxes operating at steady-state

Since the array of problems that can be tackled using systems biology approaches is very large (an overall review can be found in Klipp *et al*., 2005), for the purpose of this introduction a subset of problems serves for illustration: the problems that depend on an understanding of the whole set of metabolic reactions of an organism (sometimes termed the "reactome").

As with other areas of systems biology, metabolic analysis can be conducted at different levels of granularity. A coarse level would be to compare a bacterial species with a reference species based on the sequence identity of their enzymes and other genomic criteria. The aim is to infer if a certain metabolic function is present or absent in the former species (von Mering *et al*., 2003; Francke *et al.*, 2005). A very detailed level would be to have a quantitative, dynamic description of the changes in concentration of every metabolite in the cell under certain environmental conditions, a goal that is pursued by e.g. the E-cell project (Ishii *et al.*, 2004). In the spectrum between these two extremes lies the set of problems that depend only on the stoichiometry and topology of the metabolic network. This means that the fundamental information about the system is the stoichiometric equation of every metabolic reaction in the cell (usually with directionality in the case of thermodynamic restrictions), without the kinetics, and without the regulation of the enzymes.

In this case, it is convenient to describe the network in the form of a stoichiometric matrix (Palsson, 2006 and Figure 1.2), where each row of the matrix corresponds to one metabolite, each column corresponds to a reaction, and every value (i.e. metabolite-reaction pair) is an integer defining the stoichiometry of that metabolite in the corresponding reaction. The values are negative for substrates of a reaction, and positive for products. For reversible reactions, two columns with opposite signs are used instead of one.

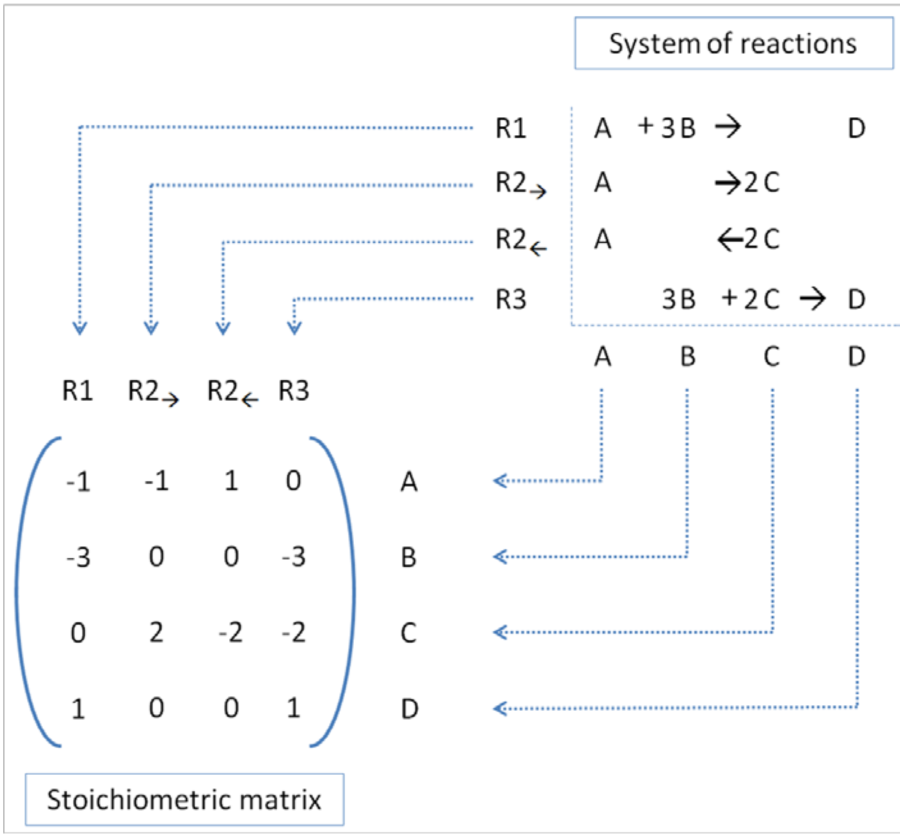


Figure 1.2: The stoichiometric matrix. A system of reactions (top right) can be described as a matrix, where each row corresponds to a metabolite and each column corresponds to a reaction. The values in the matrix correspond to the stoichiometry of each metabolite. These values are positive for products and negative for adducts.

Several mathematical properties of the stoichiometric matrix, or of a submatrix of it, can be interpreted in biological terms (for a thorough mathematical description, see Palsson, 2006, for the biological applications of this matrix see Feist & Palsson, 2008 and Oberhardt *et al.* 2009). A very common practice is to analyze the properties of the internal matrix, which results from only considering the metabolites that are present inside of the cell (i.e. excluding the transport reactions). The key property of the internal matrix is its null space. This vector space can be interpreted as the set of all metabolic fluxes that result in no net accumulation or utilization of internal metabolites (i.e. sets of metabolic fluxes that operate at steady-state).

The generation of computational tools that make use of this null space for their predictions is a thriving field. Flux balance analysis (FBA, Orth *et al.*, 2010) is considered the gold standard, due to its successful predictions of optimal growth rate and gene essentiality (Lewis *et al.*, 2010). Moreover, most applications are derivatives of this method. FBA defines a cellular objective (e.g. growth or ATP production) via a linear combination of fluxes (e.g. by adding all the fluxes required to duplicate a cell, or by adding all reactions that produce ATP). Afterwards, it uses linear optimization methods to find the optimal value for this cellular objective, under the constraint that the internal fluxes should be in steady-state (i.e. belong to the null space). The same calculation can be done with a stoichiometric matrix that lacks a column (which simulates a gene knock-out). If the predicted optimal growth rate of the mutant is zero, it is taken as an evidence of gene essentiality.

1.4 Pathway analysis in systems biology

In the past decade, genome-scale metabolic models of several organisms have been reconstructed. The improvement of a metabolic model is an iterative process that starts with all reactions described in the literature for that particular organism and then systematically tries to find gaps or contradictions in the knowledge. These gaps are then filled either by using comparative genomics, wet-lab experiments, or other methods. Additional experiments are then conducted to test the newly incorporated reactions and to find additional gaps in the knowledge, continuing the cycle (for a review see Durot *et al.*, 2009; for a step-by-step protocol to create a genome-scale model see Thiele & Palsson, 2010a).

The resulting stoichiometric matrix can be used to provide answers to physiological and evolutionary problems (a clear example of linking various biological scales through systems biology). A relatively abstract question that has been addressed is to enumerate all possible steady-state pathways inside of an organism, to assess its metabolic potential, and to compare it with other organisms. This question can be reformulated when considering that most steady-state pathways are the result of joining two or more simpler pathways. The alternative formulation of the problem becomes to find all “elementary pathways”, i.e. the minimal subset of pathways that can generate all possible pathways at steady-state (there are several alternative definitions for the term “elementary pathways”, see Llaneras & Picó (2010) for a review; here the term will mean the collection of all these alternative definitions).

Flux balance analysis and the related constraint based methods are not helpful in this setting. The reason is that these methods define a single objective that has to be met optimally and thus provide only one solution. Although it is possible to find more pathways by removing columns from the stoichiometric matrix, or adding additional constraints, this is not a systematic approach. For this reason, several algebraic methods operating on the null space of the inner metabolic matrix have been successfully developed (this is the case in “extreme pathways”, see Palsson, 2006, and “elementary flux modes”, see Pfeiffer *et al.*, 1999). Although the final number of pathways varies between the approaches (due to the slightly different definitions of “elementary pathway”), it is interesting to realize that in all methods the number of elementary pathways increases exponentially with network size, and is about half a million for a relatively small metabolic reconstruction of the metabolism of *E. coli*, which has 110 reactions and 89 metabolites (Klamt & Stelling, 2002). Due to the computational complexity of this problem, some approaches calculate the elementary pathways iteratively in ascending order of length, which is convenient in practical contexts (de Figueiredo *et al*, 2009).

There is an important difference in the definition of “pathway” in this context, when compared with the traditional biochemical definition. In biochemistry, a pathway is discovered step-by-step, in the interest of finding all intermediate metabolites and reactions between a basis substrate and a product (for recent examples, see Morinaga *et al.*, 2010 and Olszewski *et al.*, 2010). During the discovery process, some intermediates of a known pathway become substrate of other pathways, serving as link between different branches of metabolism. This gives the idea of pathways as linear (or circular) processes, connecting key metabolites.

In contrast, a “pathway” should be regarded as a network in the context of steady-state metabolic analysis. The main property of this network is that it can operate sustainably without the accumulation or degradation of intermediates. Under this definition, traditional biochemical pathways such as glycolysis and the citric acid cycle are no longer considered, since some of the intermediates (e.g. NADH) accumulate.

1.5 Applying pathway analysis to physiological problems: SPABBATS

The computation of all elementary pathways has been useful for solving several biological questions (Trinh *et al*., 2009). However, in some occasions the steady-state definition of a pathway can be too restrictive. In many situations, like in the example of glycolysis above, steady-state is not a constraint, since the intermediates of one pathway can be utilized by other pathways outside the scope of the question at hand. In these situations, the elementary pathways have difficulties in finding alternative (biochemical) pathways between sets of metabolites. These difficulties are further increased by the computational complexity of finding all steady-state pathways.

In this context, several heuristic methods have been introduced to relax the steady-state condition and still focus on the relevant pathways. One heuristic is to remove highly connected metabolites from the metabolic network. The assumption is that these metabolites serve as “currency” (e.g. ATP as energetic currency) between different pathways. Using this heuristic in combination with the steady-state condition, Beasley & Planes (2007) have reconstructed most of the traditional biochemical pathways using an optimization strategy.

Another heuristic uses established computational methods for graph analysis. These strategies regard the metabolic network as a directed graph, and they find the pathways by calculating paths connecting two metabolites in the graph. This heuristic works best in combination with the previous one, to prevent “shortcuts” in metabolism (e.g. connecting glycolysis and the citric acid cycle through ATP instead of pyruvate). A review of path-directed approaches (in comparison with the steady-state, so called “stoichiometric” methods) can be found in Planes & Beasley (2008), and a recent method that is based on this heuristic is present in Veeramani & Bader (2010).

In this work, we wanted to address a relevant biological question that cannot be answered by the aforementioned heuristics. Our problem formulation required us to find sustainable alternative pathways connecting glutamate with 2-oxoglutarate in the soil bacterium *Bacillus subtilis* (see Chapter 5) in a stoichiometrically balanced way. Glutamate is highly connected in the metabolic network of every organism, participating in at least 37 reactions in *B. subtilis* (Oh *et al.*, 2007). For this reason, it is mostly left out of the analysis in the previous heuristics. When glutamate is included, the path-directed methods find many pathways consisting of just one reaction. This is obvious considering that glutamate is the universal amino-group donor in the cell. Nonetheless this list does not provide new insight, since the only reaction that does not consume metabolic intermediates is the one catalyzed by the glutamate dehydrogenase, and in our setting, this was precisely the reaction we wanted to exclude.

For this reason, we required a method that would be able to have stoichiometric constraints (like the methods for “elementary pathways” and FBA), but where some metabolites could accumulate or be considered “currency metabolites”. At the beginning of the thesis, no such method was available. As a consequence, one of the aims of this work was to develop a new method for discovering alternative pathways in large metabolic networks. The method, named SPABBATS (=Short PAthways Between a Basis And a Target Set of metabolites), is described in Chapter 5. It uses Boolean satisfiability: a problem-solving strategy that is commonly used in electronic circuit verification and design (Velev & Bryant, 2003), but has never been used in a metabolic context before. During the course of the thesis, de Figueiredo *et al.* (2009) have developed a parallel approach based on an optimization strategy that can operate on the same constraints. The advantage of SPABBATS over their method is that it has been tested successfully in the context of a physiological problem (see Chapter 5). In addition, since Boolean satisfiability has provided performance benefits in comparison to analogous optimization methods in other areas (Graça *et al.*, 2007), SPABBATS might prove faster than the method of de Figueiredo *et al.* Nonetheless, a direct comparison between these two methods has not been made, as it was deemed to be out of the scope of this thesis project.

1.6 *Bacillus subtilis* as a model organism

The previous section provided an example of a systems biology tool used for solving physiological problems involving the soil bacterium *B. subtilis*. This rod-shaped bacterium belongs to the phylum of the Firmicutes, a group of Gram-positive bacteria with low GC content. *B. subtilis* usually grows in chains, but under stress conditions is able to build heat-resistance spores, as well as single motile cells (Graumann, 2007).

B. subtilis has become the best studied Gram-positive model organism for a number of reasons. The first reason is the relative ease of handling in wet lab experiments. The growth rate of the organism is relatively fast (it is comparable to that of *E. coli*). It also possesses a natural competence system and the machinery for homologous recombination. This makes it easy to manipulate genetically.

A second reason for its success as model organism is the suitability of this bacterium for industrial and biomedical research. *Bacillus* species are widely used in biotechnological enterprises (see Schallmey *et al.* (2004) for a review), due to the efficient secretion system of these bacteria. Secretion is crucial for the large-scale production of engineered proteins (e.g. enzymes of detergents). Moreover, the genes for many biotechnologically relevant pathways are already encoded in the genome (e.g. for the production of biotin and riboflavin). Strains of the genus *Lactobacillus*, a close relative of *B. subtilis*, are an important ingredient of diary products.

Moreover, in contrast to other related Gram-positive bacteria – like e.g. *Staphylococcus aureus*, *Listeria monocytogenes*, *Bacillus anthracis*, and *Clostridium botulinum* – *Bacillus subtilis* is apathogenic and generally regarded as safe. This safety is additionally warranted by the tryptophan auxotrophy of the widely used *B. subtilis* laboratory strain 168.

A third reason for its success as model organism is the long history of research on this organism, with the first descriptions dating from the 19th century. This has resulted in detailed knowledge of almost every aspect of its physiology, which in turn has encouraged more labs around the world to test new “omics” methods on it (Rasmussen *et al*., 2009). Moreover, in comparison to e.g. the community of *E. coli* researchers, the *Bacillus* community has remained relatively cohesive through time and the main groups doing basic research on this organism have been collaborating for many years.

1.7 Databases about *B. subtilis* and *Subti*Wiki

For the reasons outlined above, *B. subtilis* has also been a good model organism for systems biology. The first global project undertaken by the *Bacillus* community was the sequencing of the entire genome (Kunst *et al*., 1997), one of the first genomes to become available in the dawn of the genomics era.

This ambitious project was possible only through the extensive collaboration between many labs across Europe and Japan. The data that was produced by this consortium was stored and administered by the Institut Pasteur in France. Bioinformaticians at that institute created a very successful relational database for accessing the genome data, called SubtiList (Moszer, 1998).

SubtiList was updated for the last time in 2001 (according to the website). Since then, the focus of the creators of SubtiList changed, and was targeted mainly to comparative genomics. Based on the data structures of SubtiList (and related databases for other model organisms), a new database called GenoList was created (Lechat *et al*., 2008). GenoList incorporated additional tools for comparative genomics and data from several different genome projects. A fundamental change when compared to SubtiList was the user interface. The new focus simplified comparative analyses between species, but introduced some complicated additional steps in the interface for the everyday tasks performed by a wet lab *Bacillus* scientist.

At the onset of this thesis project, the lack of update of SubtiList, and the difficulty of using GenoList in the lab, proved to be a serious hurdle for collaboration and communication between several European *Bacillus* groups. These groups were engaged in several systems biology efforts that were bringing together multiple “omics” strategies with computational modeling. The main concern was the lack of a central repository, with up-to-date knowledge on *B. subtilis*, and with the possibility to be easily expanded by the different members of the groups.

For this reason, one of the aims of this thesis project was to initiate a collaborative database for *B. subtilis*. The user interface of this database had to be easy to use and reminiscent of the one of SubtiList. In contrast to *E. coli*, there is no bioinformatics group dedicated mainly to the curation of a central database on *Bacillus* (like EcoCyc (Keseler *et al.*, 2009)). For this reason, an additional constraint was to have a platform that is easy to maintain and that permits the expansion with freely available (instead of proprietary) software products.

The result is *Subti*Wiki, a wiki on all genes and proteins of *B. subtilis*. A scientific wiki fulfills all the previous criteria. Chapter 2 outlines the design principles behind *Subti*Wiki, in the interest of allowing other scientific communities to initiate similar efforts. Other scientific wikis existed before *Subti*Wiki, like e.g. EcoliWiki (for all aspects of *E. coli* research (Hu *et al*., 2008)), and WikiPathways (for collaborative curation of biochemical pathways (Pico *et al*., 2008)). Nonetheless, *Subti*Wiki responds to very specific requirements on the research on *B. subtilis* (detailed in Chapter 3) and is thus a qualitatively novel use of an existing platform.

Since its creation, the user base of *Subti*Wiki has increased steadily and it has now become the world-wide reference on gene annotation for this model organism. In the context of the systems biology approach, it provides extensive details on each individual gene and gene product, including the main interactions with other genes. Through its design, *Subti*Wiki aids in the everyday tasks of molecular biology research. In addition, it helps to make sense out of patterns obtained through “omics” techniques, thus linking molecular biology and systems biology.

1.8 The Systems Biology Graphical Notation

*Subti*Wiki provides a close-up view of the entire gene set of *B. subtilis*. Nonetheless, this close-up view is a disadvantage for the purpose of understanding phenomena involving several genes at the same time.

As mentioned above, systems biology addresses these challenges through modeling. A prerequisite for creating a good model is to have a very detailed representation of what is already known about the organism in terms of the interactions of the individual genes and gene products.

For metabolism, this representation could be a stoichiometric matrix. The advantage of this representation is that it already provides the basis for modeling several metabolic events (see previous sections). Nonetheless, the stoichiometric matrix does not contain information about how the individual enzymes are regulated at the genetic and biochemical levels. It is possible to introduce matrix formalisms that incorporate regulatory constraints (e.g. Gianchandani *et al.*, 2006). Nonetheless, most lab researchers are not familiar with the matrix formalism. For this reason, this formalism is inappropriate for laying the foundation of previous knowledge in collaboration between modelers and wet lab scientists.

The general problem is representing biological knowledge in a way that is intuitive for biologists, and at the same time precise and unambiguous for modeling. One way could be to introduce a standard vocabulary for biological processes and formulate the knowledge about the pathways following strict rules. An effort in this direction is on course: the Systems Biology Ontology (SBO) initiative (Le Novère, 2006). Nonetheless, for large pathways a textual description becomes cumbersome and does not provide a quick overview on the most important facts.

An alternative approach is to introduce a standard graphical language, like the one used in engineering for electric circuits. This idea has resulted in the Systems Biology Graphical Notation (SBGN, Le Novère *et al.*, 2009). The notation lays out the rules to draw the different entities (genes, RNAs, proteins, metabolites …) as well as how to connect them with arrows that have unambiguous meaning (i.e. there is an arrow for transcription, one for translation, one for state transition, etc. as well as for activation, catalysis, and others). Moreover, powerful and intuitive software tools have been developed to draw these diagrams. One of these tools is CellDesigner (Funahashi *et al*., 2008), a popular program that is freely available and widely used.

1.9 *Subti*Pathways and CellPublisher

At the onset of this thesis, pathway information about *B. subtilis* was dispersed in several books (Sonenshein *et al*., 1993; Sonenshein *et al*., 2002) and papers. For this reason, starting a modeling initiative was a frustrating process that involved a slow literature research. The main metabolism databases like e.g. KEGG (Kanehisa *et al*., 2010) and BioCyc (Caspi *et al.*, 2010) contained information about metabolic pathways, but in many cases these were derived from other model organisms (primarily *E. coli*) and for this reason were not very reliable. Moreover, they did not include the gene regulation of the individual enzymes.

This deficiency motivated us to formulate the *Subti*Pathways project: to make a comprehensive representation of all major pathways in *B. subtilis*, together with their regulation, in a form that would be useful for lab scientists doing molecular biology research, as well as for modelers working on systems biology. The initiation of this task was simplified considerably with the publication of a reconstruction of metabolic and regulatory pathways in *B. subtilis* by Goelzer *et al.*

(2008). This reconstruction was based on a deep literature review, and accounts for the central metabolic pathways (and their regulation) in the form of diagrams.

With additional research on the literature, it was possible to collect information on several other relevant pathways present in *B. subtilis* (see Chapter 3), leading to a comprehensive set of pathways. The next step was to repurpose this information in a format that would be most useful for lab scientists and modelers alike. In this context CellDesigner and the Systems Biology Graphical Notation provided a big opportunity. All the collected information was laid out in the form of CellDesigner diagrams.

However, the CellDesigner files by themselves did not fulfill the purpose of making the information useful for a wide audience. Although CellDesigner is a popular program in the systems biology community, it was not so widespread among *Bacillus* researchers. This meant that the diagrams had to be visible to anyone without the need to install further software. In addition, we reasoned that the usefulness of these diagrams would be greatly increased, if we could provide a way to see more information about the individual elements by clicking on them. This connection between the bird’s-eye and close-up view in biochemical pathways is very useful in systems biology, since it allows to understand the properties of the individual molecular species in a broader context. It therefore became necessary to develop a way to repurpose CellDesigner diagrams for the Internet so that they could be visible by anyone. At the same time, the navigation of the diagrams had to be enhanced, to make it easy to find individual elements and to connect them to outer sources of information, like *Subti*Wiki. Available software packages, like e.g. WikiPathways (Pico *et al.*, 2008), did not comply with the Systems Biology Graphical Notation and the navigation of the resulting pathways was a serious limiting factor.

The solution to this problem was to use the Google maps Application Programming Interface (API). This freely available JavaScript library permits the representation of complex graphical data in a form that is very intuitive to navigate. Moreover, through the use of info windows (see Chapters 3 and 4) it is possible to attach further information on specific graphical elements in the map.

In this work, the first step was to write software to convert the diagrams of *B. subtilis* created in CellDesigner to online Google maps, linked to the chemical database PubChem, the Protein Data Bank (Berman *et al*., 2007), and *Subti*Wiki. Chapter 3 presents the result and puts it in context with the work done in *Subti*Wiki.

Afterwards, we considered that the same visualization technique would be useful for other systems biology initiatives operating with the same constraints. For instance, an online tool created independently by Kono *et al.* (2009), called Pathway Projector, also uses the Google maps API to display pathway information; in this case the pathways present in the KEGG database (and without the Systems Biology Graphical Notation). We deemed that the use of the Systems Biology Graphical

Notation in combination with a custom upload of pathways by anyone would provide a powerful medium for communicating pathway knowledge.

To allow any researcher to create Google maps-based online diagrams starting with CellDesigner, the CellPublisher web server was created (Chapter 4). A similar software package was developed recently by Matsuoka *et al.* (2010) and is called Payao. It presents several similarities with CellPublisher, like the possibility to upload a CellDesigner file and add more information regarding the individual species. Nonetheless, the purpose of Payao is to create a discussion platform about diagrams and the user interface does not have the navigation features of Google maps. This makes it less suitable for presentation of finished pathways, interconnected to external resources.

1.10 Aims of this work

In the previous sections, the projects of this thesis have been introduced in the context of research in their related fields. As a summary, the aims of the projects will now be presented in their relationship to each other and their common goal.

The primary aim of this work is to create useful software tools for the analysis of the metabolism of *B. subtilis*. The tools should serve as reference for molecular biologists working with *Bacillus* and related organisms, and aid them in their daily research. At the same time, the tools should have a clear focus on systems biology, and introduce new ways to analyze the metabolism of model organisms.

*Subti*Wiki is an online collaborative database that contains up-to-date knowledge on *B. subtilis* and can be expanded by any member of the *Bacillus* community. It contributes to the general goal by serving as reference for researchers, as well as by facilitating everyday tasks in the lab. Pathway information is stored in the *Subti*Pathways database. The pathways contained in this database can be navigated easily due to the Google maps interface. *Subti*Pathways and *Subti*Wiki are interconnected with each other and with external resources, facilitating new discoveries. Together, they contribute to the larger goal by facilitating the creation of models for the metabolism of *B. subtilis*. This purpose is shared with CellPublisher that aims to facilitate the creation of interactive diagrams, and the communication between scientists. Finally, the aim of SPABBATS is to make the genome-scale models of *B. subtilis* (and other model organisms) understandable. For this purpose, it extracts pathways that satisfy specific metabolic criteria from a complex metabolic network encoded in a stoichiometric matrix.

Together, the four resources created in this thesis contribute to the better understanding of *B. subtilis* and promote new discoveries based on a better use of the available information on this organism.

2. A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki *Subti*Wiki

The results described in this chapter were published in:

Flórez,L.A., Roppel,S.F., Schmeisky,A.G., Lammers,C.R., Stülke,J. (2009) A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki *Subti*Wiki. *Database (Oxford)* **2009**: bap012.

Authors' contributions:

The structure and purpose of the wiki were designed by LAF and JS. LAF created the starting pages for each gene, and the programs to interface with the wiki. SFR, AGS, and CRL added further information to the wiki under the supervision of LAF. LAF did the figures and wrote the manuscript with JS. All authors read and approved the final manuscript.

2.1 Abstract

Bacillus subtilis is the model organism for Gram-positive bacteria, with a large amount of publications on all aspects of its biology. To facilitate genome annotation and the collection of comprehensive information on *B. subtilis*, we created *Subti*Wiki as a community-oriented annotation tool for information retrieval and continuous maintenance. The wiki is focussed on the needs and requirements of scientists doing experimental work. This has implications for the design of the interface and for the layout of the individual pages. The pages can be accessed primarily by the gene designations. All pages have a similar flexible structure and provide links to related gene pages in *Subti*Wiki or to information in the World Wide Web. Each page gives comprehensive information on the gene, the encoded protein or RNA, as well as information related to the current investigation of the gene/protein. The wiki has been seeded with information from key publications and from the most relevant general and *B. subtilis*-specific databases. We think that *Subti*Wiki might serve as an example for other scientific wikis that are devoted to the genes and proteins of one organism.

Database URL: The wiki can be accessed at **http://subtiwiki.uni-goettingen.de/**

2.2 Introduction

With the completion of more and more genome sequences, their accurate annotation has become an important matter. Usually, the initial annotation is done automatically, and is subsequently improved by manual curation. All major model organisms that are subject to extensive investigation have been sequenced and annotated in the early phase of the genomic age. However, once a genome sequence and the corresponding annotation have been published, there is decreasing support for and interest in keeping the annotation information up-to-date. Since the work in the “traditional” molecular and cell biology labs goes on, new information is continuously being generated but not included in the annotation. A good example for this problem is the small RNA SR1 of the bacterium *Bacillus subtilis* that was originally described in 2005, but that is not annotated even in the most recent publication of the *B. subtilis* genome (Licht *et al*., 2005; Barbe *et al*., 2009). Since experimental work focuses on a few model organisms, this problem is of specific urgency for these organisms. At the same time, the lack of complete up-to-date annotation information may prevent the lab researchers from getting important new insights because the relevant information is not easily accessible from primary literature. The problem of outdated annotation is even aggravated by the fact, that annotation for one organism is usually controlled by one institution that might change its focus and thus be unable to guarantee updated annotation in the long term.

A way to overcome these problems might be to establish an annotation based on the wiki concept. This concept offers several advantages: First, each interested scientist can easily contribute any information to the existing annotation and make it thus more useful. The result is that novel information can be added immediately upon its generation and its inclusion does not depend on the availability of a usually unknown curator. Thus, the task of annotation is distributed among a complete scientific community. Second, a wiki makes it very easy to retrieve complete sets of object-oriented information, which are represented as a wiki page. Moreover, the information provided can by be enriched by internal and external links to pages of the wiki and the internet, respectively. These links establish different classes of connections that make the interrelatedness of all processes of life visible and tractable. Third, a wiki is a liberal way to manage shared information. Alternative opinions can be exchanged and presented as such without somebody who has the power to decide what the truth is. Instead, each user can make its own judgement and assess the validity of opposing statements based on the evidence provided but also on the own additional knowledge.

Compared to classical relational databases, a wiki has some similarities but there are also fundamental differences. Both provide the user with the requested information. However, whereas the structure of a relational database is very rigid, a wiki can be very flexible. In principle, each page could have an individual structure that is adjusted to the information to be presented on the page. This may cause problems if one wants to extract an identical set of information from each page of the wiki. However, since the wiki is object-centred, it can be especially successful for the retrieval of information on individual objects such as genes or proteins. In contrast, relational databases outcompete wikis for the retrieval of cross-sectional information. The more flexible structure of wiki pages allows the presentation of information to a level of detail that is unprecedented in relational databases. The simple structure of the wiki pages and the inherent user-friendliness make the wiki an easy-to-access and easy-to-contribute marketplace of information.

These advantages of the wiki concept resulted in a large number of different kinds of scientific wikis that have been established in the past few years resulting in the new “discipline” of “wikiomics” (Waldrop, 2008). Wikis have been set up for different biological purposes such as ArrayWiki for the annotation of microarray experiments (Stokes *et al*., 2008), Proteopedia for protein structures (Hodis *et al*., 2008), and the model bacterium *Escherichia coli* (Hu *et al*., 2008). In addition to these more specialized wikis, there are general wikis devoted to all genes and proteins as well as to metabolic pathways (Hoffmann, 2008; Pico *et al*., 2008; Mons *et al*., 2008). The wiki concept has been suggested to be of specific value for genome re-annotation due to the challenges mentioned above (Salzberg, 2007).

We are interested in the Gram-positive model bacterium *Bacillus subtilis* (Stülke and Hillen, 2000; Commichau *et al*., 2009). These bacteria are of great practical importance because they are used in biotechnology for the production of vitamins and enzymes (Schallmey *et al*., 2004). Moreover, *B. subtilis* undergoes a simple differentiation program and is the model to understand many important pathogens such as *Bacillus anthracis*, *Staphylococcus aureus* and *Listeria monocytogenes*. Therefore, *B. subtilis* has attracted substantial research interest during the past decades that has made this bacterium the best-studied in addition to *E. coli* (Sonenshein *et al*., 2002). The genome sequence of *B.*

subtilis has been published in 1997 (Kunst *et al*., 1997) and the publicly available annotation has not been updated from 2001 to 2009 (Barbe *et al*., 2009). In an attempt to facilitate continuous genome annotation, we have set up a wiki devoted to the genes and proteins of *B. subtilis*. In this wiki, designated *Subti*Wiki, information is centred on the genes and the corresponding proteins (or RNAs) of *B. subtilis*. The wiki provides information on mutant phenotypes, gene expression and regulation, to the functions, modifications, interactions and localizations of proteins. Moreover, *Subti*Wiki provides links to databases specialized in gene expression, genome organisation, protein structures and enzyme activities. Finally, the wiki provides information on biological materials, specialists as well as links to relevant publications.

2.3 Description of the wiki

The central objects of *Subti*Wiki are the genes, proteins, and functional RNAs of *B. subtilis*. Thus, most pages of the wiki are devoted to a specific gene and its corresponding product(s). The central position of the genes is indicated by a search box on the start page of *Subti*Wiki, which can be used to enter the name of the gene of interest, to get access to detailed information on this gene. Moreover, information can be retrieved by text search through all pages of the wiki. Both the gene pages and the main page provide links to other categories of pages such as pages for the labs that work with *B. subtilis*, or pages for important plasmids and methods (see below).

Gene names as identifiers

There are two principal options to get access to gene-specific pages. One would be to use genetic gene designations, whereas the alternative is the use of gene identifiers derived from genomic projects. The latter option is preferable for organisms in which only a small part of the genes had been studied before, and where annotations are therefore not yet stably established in the scientific literature. In contrast, *B. subtilis* has been the object of substantial investigation since the middle of the last century, and this interest is going on, and has become even more intensive, with the availability of the genome sequence (Sonenshein *et al*., 2002). The use of classical gene designations has a long-standing tradition in the work with *B. subtilis*, and it is safe to predict that each *Bacillus* researcher knows the designations of at least one hundred genes together with the corresponding products and functions. With such a strong tradition, and with the needs of the scientific *Bacillus* community in mind, we decided to build the pages on the gene names. This brings of course the problem of instability of certain designations and of synonyms. Based on a collection of these synonyms that is available in our group, we ensured that all common designation of a gene guide the user to the same gene page via redirects. This is the case for 559 genes that represent about 12% of all *B. subtilis* genes. With the ongoing research and the identification of new gene functions and the introduction of novel mnemonic designations, more redirects are likely to be required.

However, gene identifiers are the most stable way to refer to specific genes and proteins since they do not change with the accumulation of novel information. The standard gene identifiers for *B. subtilis* genes are the identifiers provided by the annotation team at the Institut Pasteur (Barbe *et al*., 2009; Kunst *et al*., 1997). These identifiers are mapped to the gene designations used in *Subti*Wiki and the Uniprot identifiers in an Excel table. This table is available on the front page of *Subti*Wiki.

Semi-structured pages balance intuitive orientation with the ease of contributing

The structure of the pages of scientific wikis may be very different. On one end of the scale are WikiGenes and Proteopedia, which use text descriptions. This may make it difficult to find the requested information on a page since every page may present the content differently. On the other end, the information in EcoliWiki is entered in tables with a rigid structure. Such a structure is very helpful to achieve consistency in the presented information and in the design of the pages. At the same time it poses two problems: First, for some data there may be no obviously appropriate table. Second, and perhaps even more serious, such a structure might discourage the casual user from contributing information. However, the wiki is aimed specifically at such users that generate new experimental information and add it to the wiki without training in the curation of a wiki.

For *Subti*Wiki, we decided to strike a balance between the two extreme strategies outlined above. The most critical information that is required very often is presented in a table at the top of each page (see Fig. 1). All additional information is provided as text under preset headlines. These headlines are listed in a Table of contents on the top of each page next to the table with the key information. This general outline is derived from a template page that was used to generate all the individual gene pages. While the structure of the table is quite rigid, all other headlines can be easily adapted, irrelevant headlines can be deleted and new headlines be added. With these possibilities, each page can be adapted to the specific requirements of the gene and its products although the general layout of all pages remains still very similar. This makes it very easy for the user to go directly to the set of information he/ she is interested in. Moreover, this way of arranging the pages is very advantageous for the addition of information: The common general headlines facilitate the automatic entry of information using scripts, but it is also very easy for the user to edit the information since an “edit” button is present next to each headline, and new contents can be added intuitively. We are confident that our page layout will lower the barrier for the casual user.

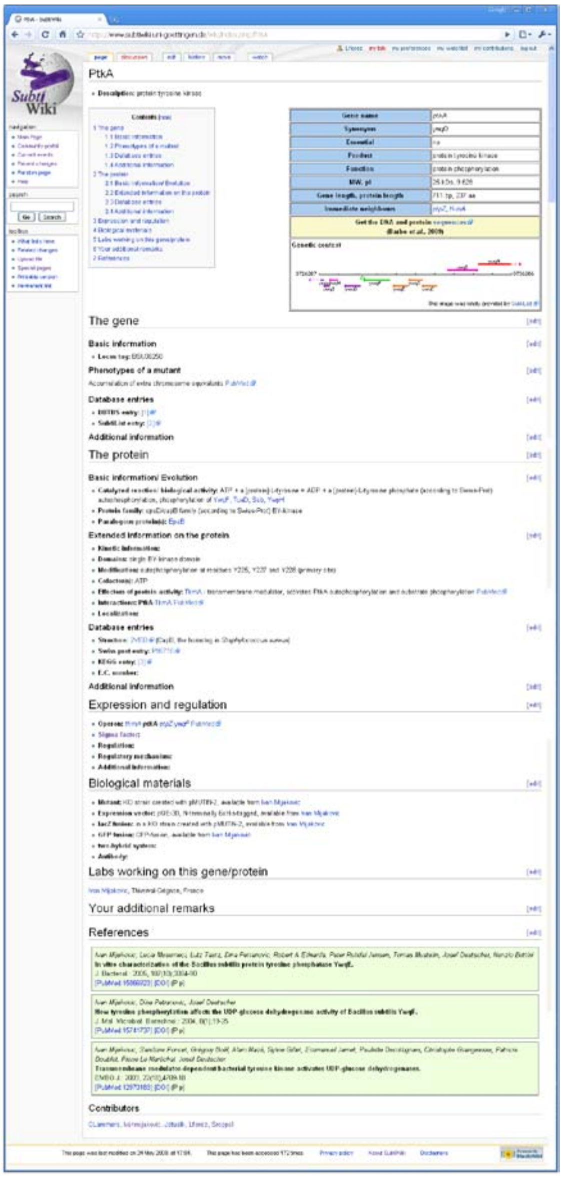
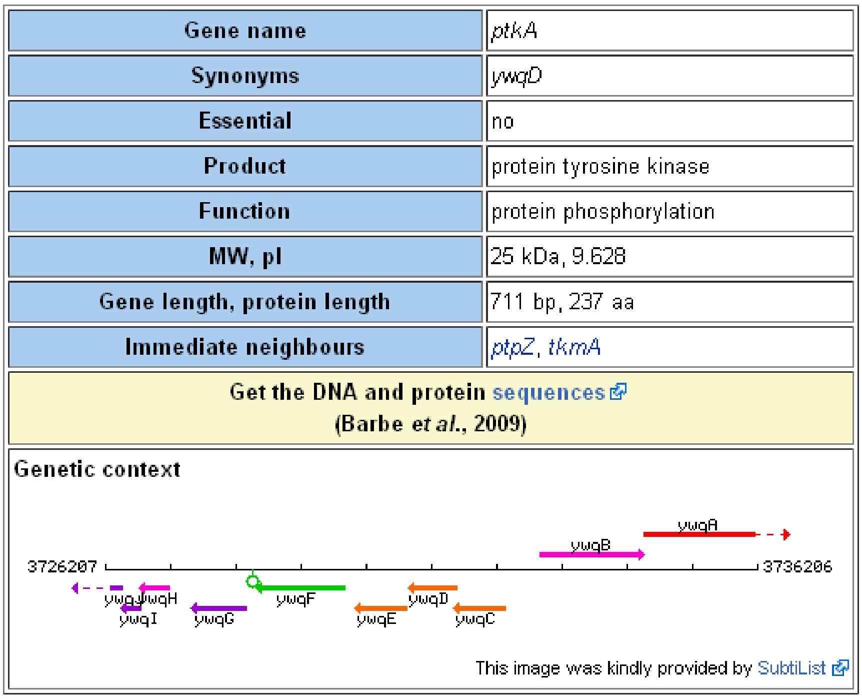


Figure 1. The layout of gene pages in *Subti*Wiki. The pages adhere to the design used in Wikipedia. At the top, the user finds a clickable table of contents of the detailed information. Next to it there is another table with the most important information on a gene/ protein and a scheme of the genomic context (see Fig. 2 for details). These tables are then followed by detailed information on the gene, the protein or RNA and gene expression/ regulation (see Fig. 3 for details). The next sections cover biological materials related to the gene/ protein, the labs working on the gene or protein, and provide space for additional remarks, which do not seem to fit elsewhere on the page. Finally, references and information on the contributors are listed (see Fig. 4 for details).

Features of the pages for the individual genes

As mentioned above, at the top of each gene page there is the table of contents for the detailed information provided in the bottom part of the page as well as a table with the most important information on a gene (see Fig. 2). This table contains information on gene designations and synonyms, the gene product and its function, whether the gene is essential or not, quantitative information important for the experimental work (gene and protein length, molecular weight and isoelectric point of the protein), and the gene context (the neighbouring genes and a figure showing the context). Moreover, this table provides a link to the DNA and amino acid sequences entry in the EMBL Nucleotide Sequence Database (Kulikova *et al*., 2007). In contrast to the detailed information for the genes, which will remain work in progress as long as the research on *B. subtilis* continues, the table with the key information has been completed for all genes (for the source of information, see below).



Gene name	*ptkA*
Synonyms	*ywqD*
Essential	no
Product	protein tyrosine kinase
Function	protein phosphorylation
MW, pI	25 kDa, 9.628
Gene length, protein length	711 bp, 237 aa
Immediate neighbours	*ptpZ*, *tkmA*

Get the DNA and protein sequences
(Barbe *et al*., 2009)

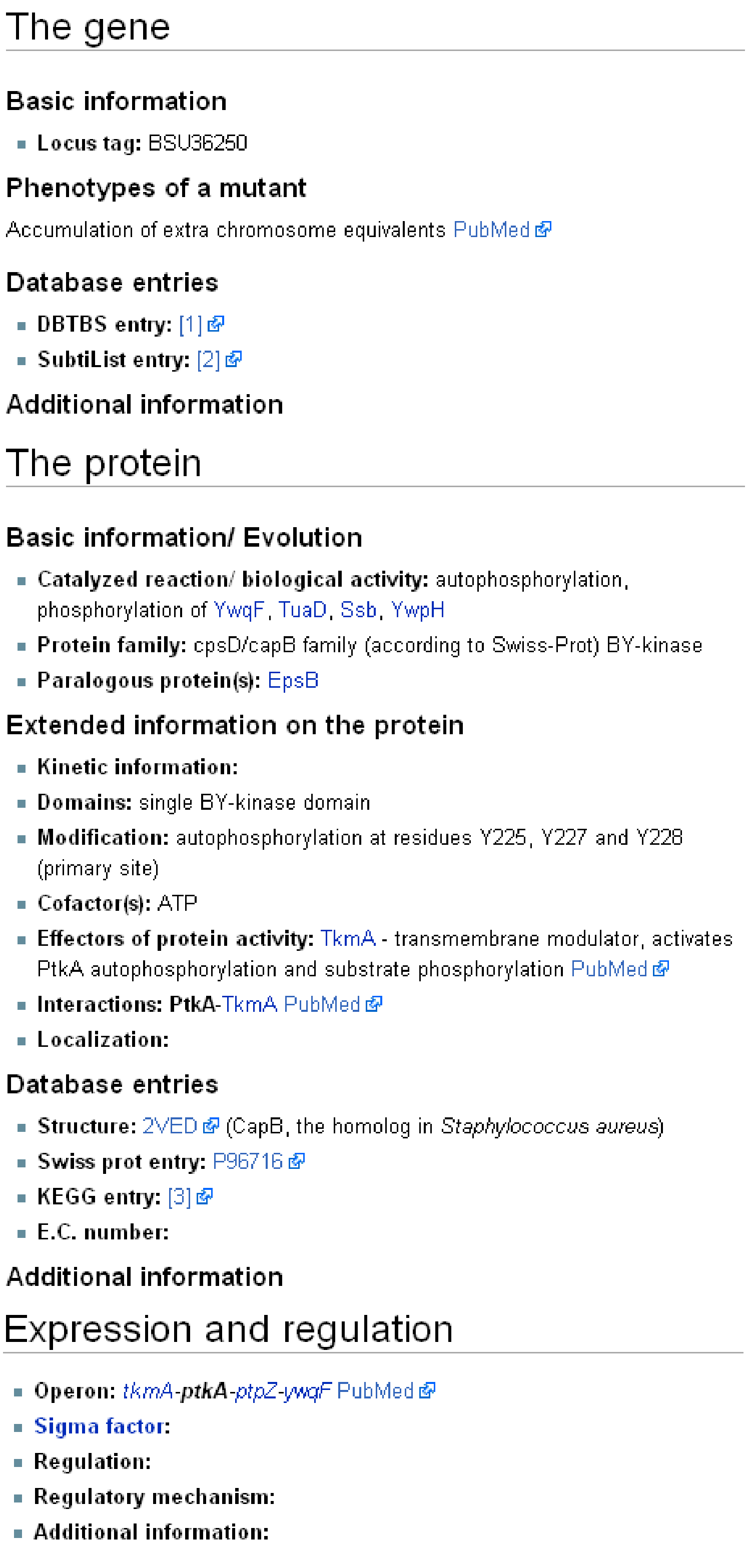
Genetic context

This image was kindly provided by SubtiList

Figure 2. Key information on any gene/ protein. Each page contains a table with information that is most often accessed by experimental biologists. The table provides information on synonyms, states whether a gene is essential or not and gives details on the gene product and its function(s). Moreover, the table provides “technical data” such as lengths of the gene and the corresponding protein as well as the molecular weight and the isoelectric point of the protein. Next, the table contains information on the genomic context (neighbouring genes and a figure with of the 10 kb region). Finally, there is a link that gives access to the DNA and protein sequences.

The detailed information is provided in seven categories that belong either to the molecular description of the gene and its product or that are related to the ongoing research on the gene. For the molecular description (see Fig. 3), there is first a section with information on the gene such as the locus tag, the phenotype(s) of a mutant, and links to gene-centred databases. The second section deals with the properties of the encoded protein. The first sub-section covers the biological activity of the protein and evolutionary aspects such as the protein family and paralogous proteins encoded in the genome of *B. subtilis*. The second sub-section provides detailed information on the protein such as kinetic information, the domain structure, modifications, cofactors and effectors of the biological activity, interactions and the localization of the protein. Finally, the third part of this section contains links to protein-centred databases that cover protein structures and protein activities. The third section providing molecular information is devoted to gene expression and regulation. Here, the operon structures and sigma factors are listed. Moreover, this section provides information on gene regulation and the corresponding regulators and their regulatory mechanisms. To facilitate the research on *B. subtilis*, the second part of the pages provides information on biological materials and their availability (mutants, expression vectors, GFP fusions, antibodies, etc.), on the labs that work on a gene/ protein and the key references on the gene or protein (see Fig. 4). For all information that does not easily fit into the provided frame, there is a section for additional information. Moreover, such a sub-section is also present in each of the three parts that cover the molecular biology. There are some genes that encode RNAs rather than proteins. For such genes, the page content was adapted accordingly, and the section on the protein was replaced by a section on the RNA.

An important general feature of all gene-specific pages is the availability of external and internal links. External links direct the user to databases or to the publications that describe the data that are presented in the wiki. Internal links relate each gene/ protein with all other genes or proteins with which it interacts in one or the other way (physical interaction, regulation, or co-localization on the genome). Moreover, the part of the page devoted to the research on a gene provides internal links to pages on resources like plasmids and experimental approaches as well as to pages with information on the labs that study a gene or protein. These features are intended to facilitate the collaboration among the *Bacillus* labs.



The gene

Basic information

- **Locus tag:** BSU36250

Phenotypes of a mutant

Accumulation of extra chromosome equivalents PubMed

Database entries

- **DBTBS entry:** [1]
- **SubtiList entry:** [2]

Additional information

The protein

Basic information/ Evolution

- **Catalyzed reaction/ biological activity:** autophosphorylation, phosphorylation of YwqF, TuaD, Ssb, YwpH
- **Protein family:** cpsD/capB family (according to Swiss-Prot) BY-kinase
- **Paralogous protein(s):** EpsB

Extended information on the protein

- **Kinetic information:**
- **Domains:** single BY-kinase domain
- **Modification:** autophosphorylation at residues Y225, Y227 and Y228 (primary site)
- **Cofactor(s):** ATP
- **Effectors of protein activity:** TkmA - transmembrane modulator, activates PtkA autophosphorylation and substrate phosphorylation PubMed
- **Interactions: PtkA-**TkmA PubMed
- **Localization:**

Database entries

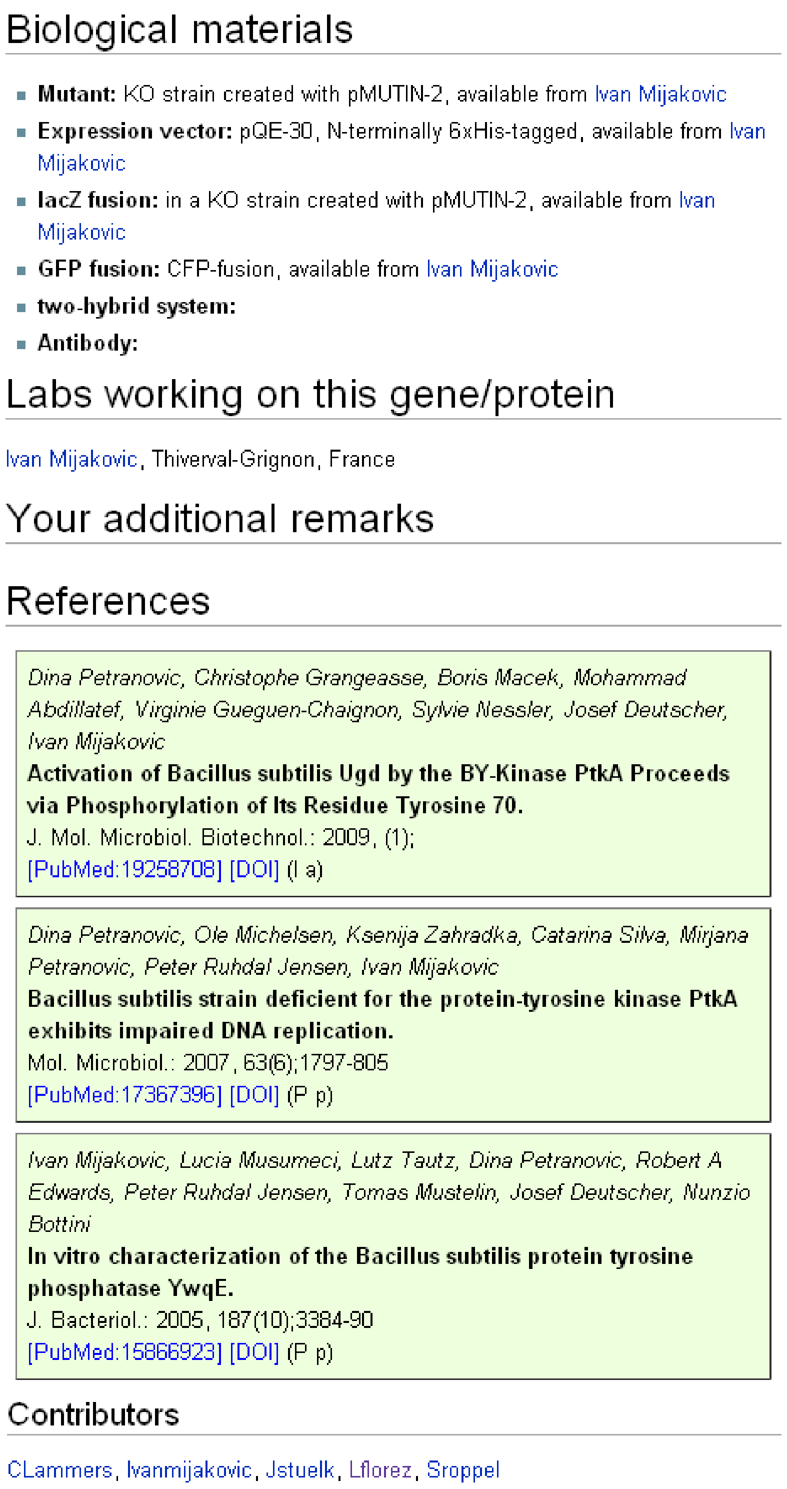
- **Structure:** 2VED (CapB, the homolog in *Staphylococcus aureus*)
- **Swiss prot entry:** P96716
- **KEGG entry:** [3]
- **E.C. number:**

Additional information

Expression and regulation

- **Operon:** *tkmA*-***ptkA***-*ptpZ*-*ywqF* PubMed
- **Sigma factor:**
- **Regulation:**
- **Regulatory mechanism:**
- **Additional information:**

Figure 3. Biological information on the gene/ protein. The pages provide a frame for entering detailed information on any gene or protein. The first section is devoted to the gene and provides information such as the locus tag, phenotypes of mutants and gene-related database entries. The second section covers information on the protein such as the biological activity, the membership in a protein family, and the presence of paralogous proteins in *B. subtilis*. Moreover, features such as the domain structure, modifications, cofactors and effectors of the biological activity, interaction partners and the protein localization are available. Again, this section ends with protein centred-databases. The third section describes gene expression and regulation. Here, the user finds the operon structure, information on the sigma factor(s) and regulatory mechanisms.



Biological materials

- **Mutant:** KO strain created with pMUTIN-2, available from Ivan Mijakovic
- **Expression vector:** pQE-30, N-terminally 6xHis-tagged, available from Ivan Mijakovic
- **lacZ fusion:** in a KO strain created with pMUTIN-2, available from Ivan Mijakovic
- **GFP fusion:** CFP-fusion, available from Ivan Mijakovic
- **two-hybrid system:**
- **Antibody:**

Labs working on this gene/protein

Ivan Mijakovic, Thiverval-Grignon, France

Your additional remarks

References

Dina Petranovic, Christophe Grangeasse, Boris Macek, Mohammad Abdillatef, Virginie Gueguen-Chaignon, Sylvie Nessler, Josef Deutscher, Ivan Mijakovic
Activation of Bacillus subtilis Ugd by the BY-Kinase PtkA Proceeds via Phosphorylation of Its Residue Tyrosine 70.
J. Mol. Microbiol. Biotechnol.: 2009, (1);
[PubMed:19258708] [DOI] (I a)

Dina Petranovic, Ole Michelsen, Ksenija Zahradka, Catarina Silva, Mirjana Petranovic, Peter Ruhdal Jensen, Ivan Mijakovic
Bacillus subtilis strain deficient for the protein-tyrosine kinase PtkA exhibits impaired DNA replication.
Mol. Microbiol.: 2007, 63(6);1797-805
[PubMed:17367396] [DOI] (P p)

Ivan Mijakovic, Lucia Musumeci, Lutz Tautz, Dina Petranovic, Robert A Edwards, Peter Ruhdal Jensen, Tomas Mustelin, Josef Deutscher, Nunzio Bottini
In vitro characterization of the Bacillus subtilis protein tyrosine phosphatase YwqE.
J. Bacteriol.: 2005, 187(10);3384-90
[PubMed:15866923] [DOI] (P p)

Contributors

CLammers, Ivanmijakovic, Jstuelk, Lflorez, Sroppel

Figure 4. Information on the research on the gene/ protein. The first section of this part gives information on biological materials such as mutants, reporter fusions, expression systems or antibodies. The second section shows the labs that study the gene/ protein, and there is a section for additional remarks that do not seem to be appropriate at any other position of the wiki. The last section covers the references. At the very bottom of each page, the contributors to this page are shown. These entries are generated automatically.

2.4 Implementation of *Subti*Wiki

The content management system

MediaWiki was chosen as the software platform for the wiki (www.mediawiki.org). This interface is identical to that used by Wikipedia, and thus most users can be expected to be immediately familiar with the way of interacting with the system. In addition, MediaWiki allows the use of extensions. Due to the popularity of MediaWiki, many extensions are already available for immediate use. For *Subti*Wiki, we use three extensions: First, ContributionCredits (http://www.mediawiki.org/wiki/Extension:ContributionCredits) allows giving any contributor a credit for his work by placing his username at the bottom of each page. This is very important for two reasons: On the one hand, the user is acknowledged for each contribution, but on the other hand, this protects the wiki from potential anonymous spam. Second, reCAPTCHA (http://www.mediawiki.org/wiki/Extension:ReCAPTCHA) is used to prevent anonymous automated registration by malicious scripts. This is achieved by requesting the entry of two words upon registration. These words are easily recognized by any person, but they cannot be processed by computer programs (von Ahn *et al.*, 2008). Third, the extension Pubmed (http://www.mediawiki.org/wiki/Extension:Pubmed) serves to fetch literature citations from PubMed entries.

User access and restrictions

All pages of *Subti*Wiki are freely accessible without prior registration. However, the contribution of information is only possible for registered users who are logged in. Our policy with respect to registrations is based on two conflicting aims: On the one hand, we wish to invite users to contribute to the wiki rather than to discourage them by complicated procedures. On the other hand, the reliability of the information provided by *Subti*Wiki is of crucial importance. A mandatory but liberal registration policy seemed to be the best way to balance these two aims. Thus, registration is simple and does not require the approval of another user. As mentioned above, automated spam registrations are prevented by the reCAPTCHA extension. Once a registered user has logged into the system, any input is possible. With our system of giving credits to all contributors, there is another level of security since nobody is able to modify any pages without being revealed. Moreover, we do not expect vandalism to be a major problem for specialized scientific wikis.

Sources of information

The information provided in *Subti*Wiki is derived from three principal types of sources. First, we used information from general databases (Table 1). Second, and most importantly, we derived information from databases that are specifically devoted to *B. subtilis*, and finally, the scientific literature was an important source of information (Table 2). Each page contains information derived from each of these sources because the different databases serve specialized purposes. We did not only extract information from these sources, but we did also provide links to them whenever possible and appropriate. This allows the user to profit from the specific strengths of the individual information sources.

The pages were generated using SubtiList, the semi-official database on *B. subtilis* (Moszer *et al*., 2002). Moreover, information on gene products and functions, key quantitative properties of the genes (size) and proteins (size, molecular weight, isoelectric point) as well as the figure showing the genetic context were derived from SubtiList. The gene nomenclature used in SubtiList served as the basis for *Subti*Wiki. Gene designations that have been replaced since the last update in 2001 are still recognized with the SubtiList designation due to page redirects (see above). Information on protein families, the biological activities of proteins and their localization was, in addition to specific references, derived from SwissProt. The DBTBS database on transcriptional regulation in *B. subtilis* was used to extract information on transcription units, sigma factors, and transcription regulators. The other databases provided important specialized information for many genes (see Table 1).

Table 2 lists the papers that were of special importance for filling *Subti*Wiki with information. For this purpose, we selected publications that followed genome-wide strategies and that are therefore relevant to a large number of genes.

The data processing workflow

The databases and publications that served as source of information are very heterogeneous in the way of presenting information and in the use of gene designations. Data processing from each source consisted of five steps, (i) the acquisition of a holistic data set, (ii) mapping of the data to the gene designations in *Subti*Wiki, (iii) re-formatting the information to wiki markup language, (iv) batch upload of the information, and (v) manual curation of the uploaded information. For all these steps, special purpose Python (www.python.org) scripts were developed.

Data were derived from either manually curated lists or from lists that were provided in the databases or publications. From these lists, the identifiers were parsed. If the identifiers were gene names, they were checked and, if necessary, converted to the actual annotation in *Subti*Wiki. For this purpose, a list of gene designations and synonyms (including redirects) was used. If accession numbers were used as identifiers we used tables that mapped the accession numbers to the *Subti*Wiki designations. The information was then re-formatted. The resulting entries conformed to the formatting style of MediaWiki and were enriched by internal and external links. For the upload, the Python scripts added the new information taking care not to replace any prior information. Since any automated process is prone to errors, some entries (about 1% of the entries) were randomly sampled to verify the correctness of the upload. Occasional systematic errors were corrected by new scripts or manually, if they were relevant only for a small number of pages.

Table 1: Databases used to seed *Subti*Wiki

Database, Reference	Extracted information, URL
General databases	
EMBL-bank (Kulikova *et al*., 2007)	Link to the nucleotide and protein sequences. EC Numbers of enzymes. http://www.ebi.ac.uk/embl/
KEGG (Kanehisa *et al*., 2008)	Link to the corresponding gene entry. http://www.genome.jp/kegg/
MPIDB (Goll *et al*., 2008)	Protein-protein interactions. http://www.jcvi.org/mpidb
PDB (Berman *et al*., 2007)	Link to the molecular structures. http://www.pdb.org/
PubMed	Literature citations. http://www.ncbi.nlm.nih.gov/pubmed
Swiss-Prot/UniProt (Uniprot Consortium, 2009)	Link to the protein entry. Protein family, localization, and catalyzed reaction. http://www.uniprot.org/
Databases specific for *B. subtilis*	
DBTBS (Sierro *et al*., 2008)	Link to the corresponding operon entry. Operon structure and regulation. http://dbtbs.hgc.jp/
SubtiList (Moszer *et al*., 2002) GenoList (Lechat *et al*., 2008)	Link to the corresponding gene entry. Name, length, product, function, and genomic context of the genes. Molecular weight, isoelectric point, and length of the protein. http://genolist.pasteur.fr/

Table 2: Key publications used to seed *Subti*Wiki

Extracted information	References
Essential genes	Kobayashi *et al*., 2003; Hunt *et al*., 2006; Thomaides *et al*., 2007
Gene regulation	Blencke *et al*., 2003; Mäder *et al*., 2002; Höper *et al*., 2005; Petersohn *et al*., 2001; Molle *et al*., 2003
Localization	Meile *et al*., 2006; Hahne *et al*., 2008; Voigt *et al*., 2009
Protein phosphorylation	Lévine *et al*., 2006; Macek *et al*., 2007; Eymann *et al*., 2007
Groups of related genes/ proteins	Fabret *et al*., 1999; Reizer *et al*., 1999; Quentin *et al*., 1999; Saito *et al*., 2009

2.5 Considerations for the attraction of a wide and active audience

The ultimate goal of research on any organism is to get a comprehensive understanding of its biology. A main part of this research is the elucidation of the functions of all genes, RNAs and proteins, their regulation, localization, and interactions. By making such information available and easily accessible, *Subti*Wiki serves the scientific community that works on *B. subtilis* and closely related bacteria such as the pathogens *S. aureus*, *L. monocytogenes* or *B. anthracis*. *Subti*Wiki helps these communities to keep up with the research progress and provides insights into novel links between genes or proteins that might otherwise have escaped the attention of the busy scientist.

One could also imagine concentrating curation efforts on a small number of “large” wikis instead of creating novel, more specialized wikis. Arguments in favour of the “central solution” are the possibility of using the same set of tools and to provide a uniform interface for comparative genomics approaches. Nonetheless, there are also arguments against this approach: First of all, each model organism has its specific traits. The wiki should be tailored to the needs of the community that studies these traits, and the “one size fits all” solution might be a strait jacket. More importantly, as the use of scientific wikis is in an early stage, we are still in the process of identifying the best concepts that define a successful wiki. A healthy competition between alternative solutions will allow us to discern what works best.

To make *Subti*Wiki a vivid platform of information exchange, it needs the attention and the commitment of the *Bacillus* scientific community. With the information already provided for all genes (see Fig. 2), we are confident that any researcher looking at it will find it useful and we count on the willingness of the colleagues to share their expertise and to improve *Subti*Wiki by contributing. We trust that we provide already so much in terms of information content and simplicity of the interface that the hurdle is set quite low for any other potential contributor. This might be the solution for the problem that many scientific wikis experience, *i. e.* the difficulty to attract new contributors (Welch & Welch, 2009).

In principle, there are two ways for finding information, the “ants’ perspective” and the “birds perspective”. Wikis provide information usually in the ants’ perspective. However, *Subti*Wiki is complemented by a presentation of metabolic and regulatory pathways in *B. subtilis* (*i. e.* *Subti*Pathways). This presentation provides an approach to information on *B. subtilis* metabolism from the birds eyes view. It allows visual navigation in the pathways and links the genes and enzymes to *Subti*Wiki. On the other hand, *Subti*Wiki provides links to *Subti*Pathways to visualise all genes/ RNAs/ proteins in the context of their cellular functions. The CellDesigner (Kitano *et al*., 2005) source files of *Subti*Pathways (in Systems biology markup language) are available upon request.

At this early stage of using wikis in biocuration, we feel that *Subti*Wiki might prove to be useful for many scientists who are considering the creation of a scientific wiki for their own purpose.

3. Connecting parts with processes: *Subti*Wiki and *Subti*Pathways integrate gene and pathway annotation for *Bacillus subtilis*

The results described in this chapter were published in:

Lammers,C.R.*, Flórez,L.A.*, Schmeisky,A.G., Roppel,S.F., Mäder,U., Hamoen,L., Stülke,J. (2010) Connecting parts with processes: *Subti*Wiki and *Subti*Pathways integrate gene and pathway annotation for *Bacillus subtilis*. *Microbiology* **156**: 849-859.

*These authors contributed equally to this work.

Authors' contributions:

LAF, UM, LH and JS initiated the *Subti*Wiki project. LAF created the wiki starting pages, as well as programs to add content. SFR added information on *Subti*Wiki with computer programs under the supervision of LAF. LAF and JS planned the *Subti*Pathways project. CRL created the programs for *Subti*Pathways, under the supervision of LAF. AGS added content to *Subti*Wiki and created the CellDesigner diagrams. LAF did the figures and wrote the manuscript with JS. All authors read and approved the final manuscript.

3.1 Abstract

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 strongly 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 *Subti*Wiki and *Subti*Pathways, two complementary online tools for gene and pathway information on *B. subtilis* 168. *Subti*Wiki (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 labs; and a list of references. *Subti*Wiki 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, *Subti*Pathways (http://subtipathways.uni-goettingen.de/) is an online tool for navigation of the metabolism of *B. subtilis* and its regulation. Each *Subti*Pathways 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, *Subti*Wiki and *Subti*Pathways provide an integrated view of the processes that make up *B. subtilis* and its components, making it the most comprehensive resource for *B. subtilis* researchers in the web.

3.2 Introduction

B. 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 comprises important pathogens such as *Staphylococcus aureus*, *Listeria monocytogenes*, *Bacillus anthracis*, and *Clostridium botulinum*. Likewise, the lactic acid bacteria that are widely used in dairy industry, important enzyme producers such as *Bacillus licheniformis*, and the insect pathogen *Bacillus thuringiensis* that is used for crop protection are 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.

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, 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 ten 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 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 in the functions and molecular activities of those genes and gene products that are already under investigation.

For the access of 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 SwissProt or the collection of NCBI databases, databases with some focus such as GenoList or Prodoric (Grote *et al*., 2009) that are centered on a group of microorganisms, and strongly specialized databases such as DBTBS (Sierro *et al*., 2008) that provides information on transcription 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).

In view of the above, there is a need to complement the centrally curated databases with interfaces that provide up-to-date reliable information in a rapidly accessible manner. In addition, the researcher needs a more dynamic and flexible interface than those provided by the non-specialist central databases.

The limitations of centrally curated databases have led to a re-interpretation of the way a knowledge repository should be managed. The critical aspect is to empower the user with the option of adding new knowledge. A technological platform that effectively serves this purpose is a wiki (Hu *et al*., 2008). A wiki can be designed to provide exactly that information in the structure that a certain community needs. Thus, wikis are being set up for different scientific purposes resulting in the new field of “wikiomics” (Waldrop, 2008). There are more general wikis devoted to all genes and proteins such as WikiGenes or WikiProteins as well as to metabolic pathways such as WikiPathways (Hoffmann, 2008; Mons *et al*., 2008; Pico *et al*., 2008). In addition, there are more specialized wikis such as ArrayWiki for the annotation of transcriptome studies (Stokes *et al*., 2008), EcoliWiki for the information on *E. coli* (Hu *et al*., 2008), or Proteopedia for the collection of protein structures (Hodis *et al*., 2008).

We are interested in the regulation of carbon and nitrogen metabolism in *B. subtilis*. When performing experiments at the genome level, we experienced the need to have a comprehensive and up-to-date source of information on the genes and proteins of *B. subtilis*. This opinion was shared within a European consortium that studies systems biology of *B. subtilis* and therefore, we developed *Subti*Wiki and *Subti*Pathways as information tools on *B. subtilis* that are designed with the requirements of the experimental scientist in mind. *Subti*Wiki provides information on all genes, functional RNAs and proteins of *B. subtilis subsp. subtilis* 168 and is based on the recently published sequence (Barbe *et al*., 2009; GenBank acc. no. AL009126). For each gene, the basic information is available, and the scientific community can extend this information, thus keeping it always up-to-date. A central feature of *Subti*Wiki is extensive internal links that make the interconnections between different genes/proteins directly visible and accessible. *Subti*Pathways 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, *Subti*Wiki and *Subti*Pathways are a comprehensive and up-to-date source of information related to all aspects of the research on *B. subtilis*.

3.3 Materials and Methods

Wiki implementation

*Subti*Wiki 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 citation and presents it on the page.

Seeding of the wiki

*Subti*Wiki was seeded as described previously (Flórez *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, pI and MW of the protein) as well as the genetic context were extracted from SubtiList and inserted in *Subti*Wiki 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 re-annotation 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 (Flórez *et al*., 2009). Finally, each gene page was linked to the relevant entries in the EMBL-Bank (gene and protein sequences, Kulikova *et al*., 2009), 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 SwissProt/Uniprot identifier was collected for each protein.

Development of a navigation interface

The online navigation interface for the diagrams was implemented using the version 2 of the Google Maps Application programming interface (API) that is freely available on 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 GWDG. 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 GWDG 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 GWDG servers with the software functionality in the Google servers, thus providing the dynamic interface.

3.4 Results and Discussion

The concept of *Subti*Wiki and *Subti*Pathways

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

*Subti*Wiki is centered 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 in 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 *Subti*Wiki, as 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 with any other gene of *B. subtilis*, whereas the external links guide the user to relevant databases, structural information, or evidence as links to publications.

To facilitate the understanding of metabolism and its regulation, we have created a companion site, *Subti*Pathways. 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 *Subti*Wiki and the NCBI database PubChem, respectively. If there is a structure for a protein available, then this structure and a link to the corresponding Proteopedia (Hodis *et al.*, 2008) page appear in a pop-up window. On the other hand, the *Subti*Wiki pages for all 750 proteins that appear in *Subti*Pathways provide links to the corresponding pathway diagram(s).

The current state of *Subti*Wiki is just intended as a starting point. For the future expansion and to keep it an updated data source *Subti*Wiki 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, edition by a user requires prior registration.

The start page

The central element of the *Subti*Wiki 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) received new names in the past few years after the last revision of the SubtiList database (Moszer *et al*., 2002). Of these newly baptised 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 *Subti*Wiki, both the old and new designations will lead the user to the same page with the most recent designation.

The start page of *Subti*Pathways (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 *Subti*Wiki or *Subti*Pathways and resources for the *Bacillus* community. The user not so familiar with the site may find guided tours and a tutorial that explains how to add new information in *Subti*Wiki helpful. The resources for the community include links to *Bacillus* labs, open positions, 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 *Subti*Wiki, SubtiList and GenoList with the NCBI locus tags, UniProt identifiers and brief functional information.

The last feature on the start pages of *Subti*Wiki and *Subti*Pathways are links to Wikipedia-type "main pages". The main page for *Subti*Wiki provides links to some example gene pages as well as to a collection of additional wiki pages that enrich the contents of *Subti*Wiki but that are not centered on individual genes or proteins (see below). The *Subti*Pathways main page again allows selecting a pathway of interest and gives information on the color code that was used throughout *Subti*Pathways.

Gene designations

Since the central entry point of *Subti*Wiki is a search box and the main element 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 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 *Escherichia 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 the ongoing research, 355 of these "y" genes were functionally analyzed and these genes got a new designation.

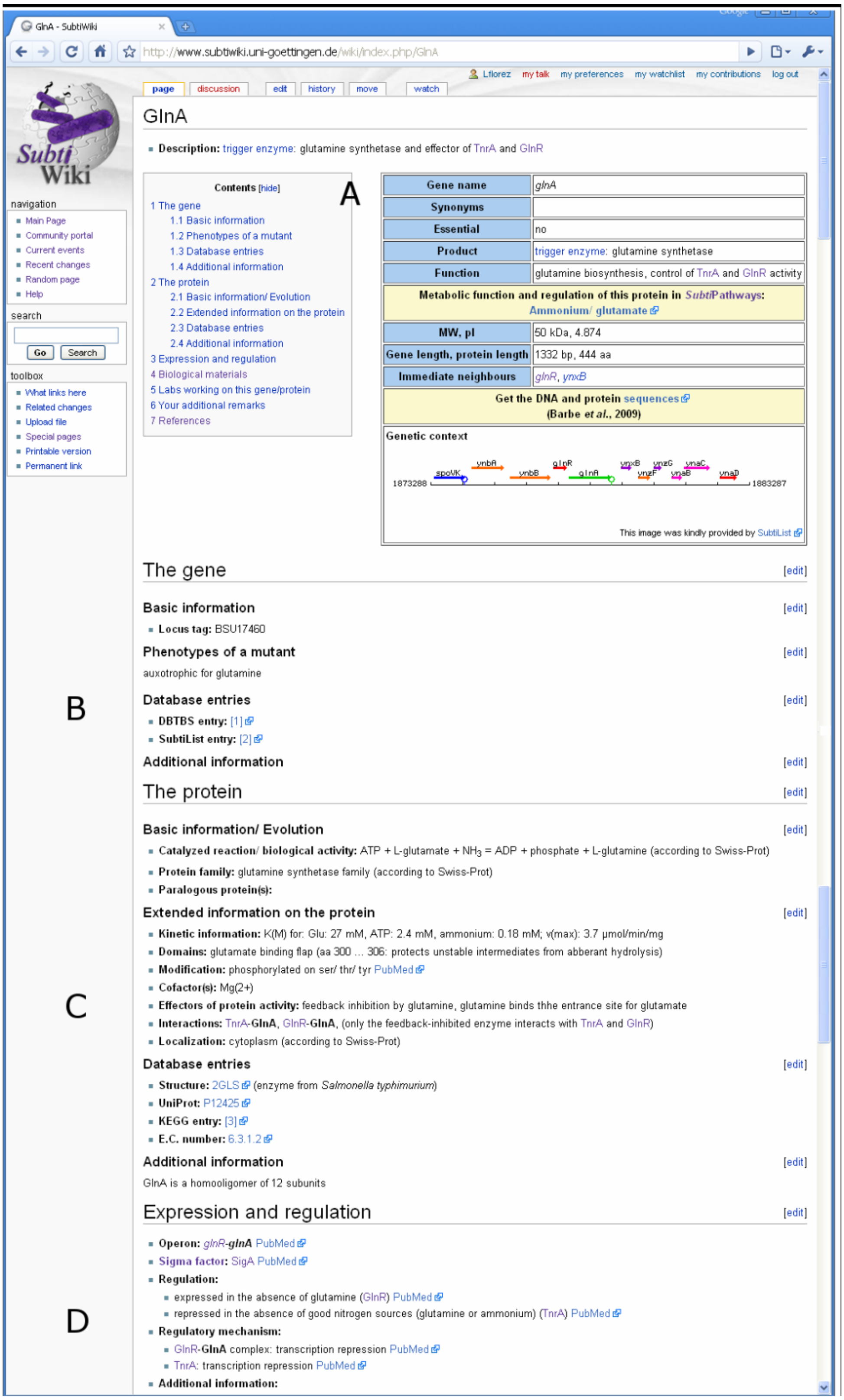
The nomenclature used in *Subti*Wiki is based on that of SubtiList. This means, each SubtiList gene name will lead to a gene page, even if the gene has been renamed in the meantime. In those cases, the user will be directly guided to the page with the new name, and the old name will be indicated as a 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 re-named by the corresponding experts. It is important to note, that there are 162 new designations that are not based on published evidence in GenoList (Barbe *et al.*, 2009). For these genes, *Subti*Wiki prefers to use the classic SubtiList designations to which the scientific community is used and which do already appear in many publications.

Features of the gene pages

*Subti*Wiki 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 *Subti*Pathways pages, numerical data (molecular weight and isoelectric points for proteins, gene and protein lengths). Moreover, the table lists the neighbors on the chromosome, 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-centered 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. Moreover, this section provides links to databases related to proteins and metabolism including (among others) protein structure databases and E. C. numbers. The third section contains information on gene expression and regulation such as operon structures, sigma factors, regulators and regulatory mechanisms.



GlnA

- **Description:** trigger enzyme: glutamine synthetase and effector of TnrA and GlnR

Gene name	*glnA*
Synonyms	
Essential	no
Product	trigger enzyme: glutamine synthetase
Function	glutamine biosynthesis, control of TnrA and GlnR activity
Metabolic function and regulation of this protein in *Subti*Pathways: **Ammonium/ glutamate**	
MW, pI	50 kDa, 4.874
Gene length, protein length	1332 bp, 444 aa
Immediate neighbours	*glnR*, *ynxB*
Get the DNA and protein sequences **(Barbe *et al.*, 2009)**	
Genetic context	

This image was kindly provided by SubtiList

The gene [edit]

Basic information [edit]

- **Locus tag:** BSU17460

Phenotypes of a mutant [edit]

auxotrophic for glutamine

Database entries [edit]

- **DBTBS entry:** [1]
- **SubtiList entry:** [2]

Additional information [edit]

The protein [edit]

Basic information/ Evolution [edit]

- **Catalyzed reaction/ biological activity:** ATP + L-glutamate + NH_3 = ADP + phosphate + L-glutamine (according to Swiss-Prot)
- **Protein family:** glutamine synthetase family (according to Swiss-Prot)
- **Paralogous protein(s):**

Extended information on the protein [edit]

- **Kinetic information:** K(M) for: Glu: 27 mM, ATP: 2.4 mM, ammonium: 0.18 mM; v(max): 3.7 µmol/min/mg
- **Domains:** glutamate binding flap (aa 300 ... 306: protects unstable intermediates from abberant hydrolysis)
- **Modification:** phosphorylated on ser/ thr/ tyr PubMed
- **Cofactor(s):** Mg(2+)
- **Effectors of protein activity:** feedback inhibition by glutamine, glutamine binds thhe entrance site for glutamate
- **Interactions:** TnrA-**GlnA**, GlnR-**GlnA**, (only the feedback-inhibited enzyme interacts with TnrA and GlnR)
- **Localization:** cytoplasm (according to Swiss-Prot)

Database entries [edit]

- **Structure:** 2GLS (enzyme from *Salmonella typhimurium*)
- **UniProt:** P12425
- **KEGG entry:** [3]
- **E.C. number:** 6.3.1.2

Additional information [edit]

GlnA is a homooligomer of 12 subunits

Expression and regulation [edit]

- **Operon:** *glnR*-***glnA*** PubMed
- **Sigma factor:** SigA PubMed
- **Regulation:**
  - expressed in the absence of glutamine (GlnR) PubMed
  - repressed in the absence of good nitrogen sources (glutamine or ammonium) (TnrA) PubMed
- **Regulatory mechanism:**
  - GlnR-**GlnA** complex: transcription repression PubMed
  - TnrA: transcription repression PubMed
- **Additional information:**

Figure 1. A sample gene page in *Subti*Wiki. All gene pages have a common layout with the following elements: A) a summary table with the most important information, B) a section on the gene with the mutant's phenotype and references to gene centered databases, C) a section on the gene product (usually a protein) with biochemical characterization and links to pertinent databases, and D) a section on regulation with factors that modulate expression. In addition, each page has a section to foster collaboration in the *Bacillus* community, and references (see Figure 2).

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 *Subti*Wiki 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.

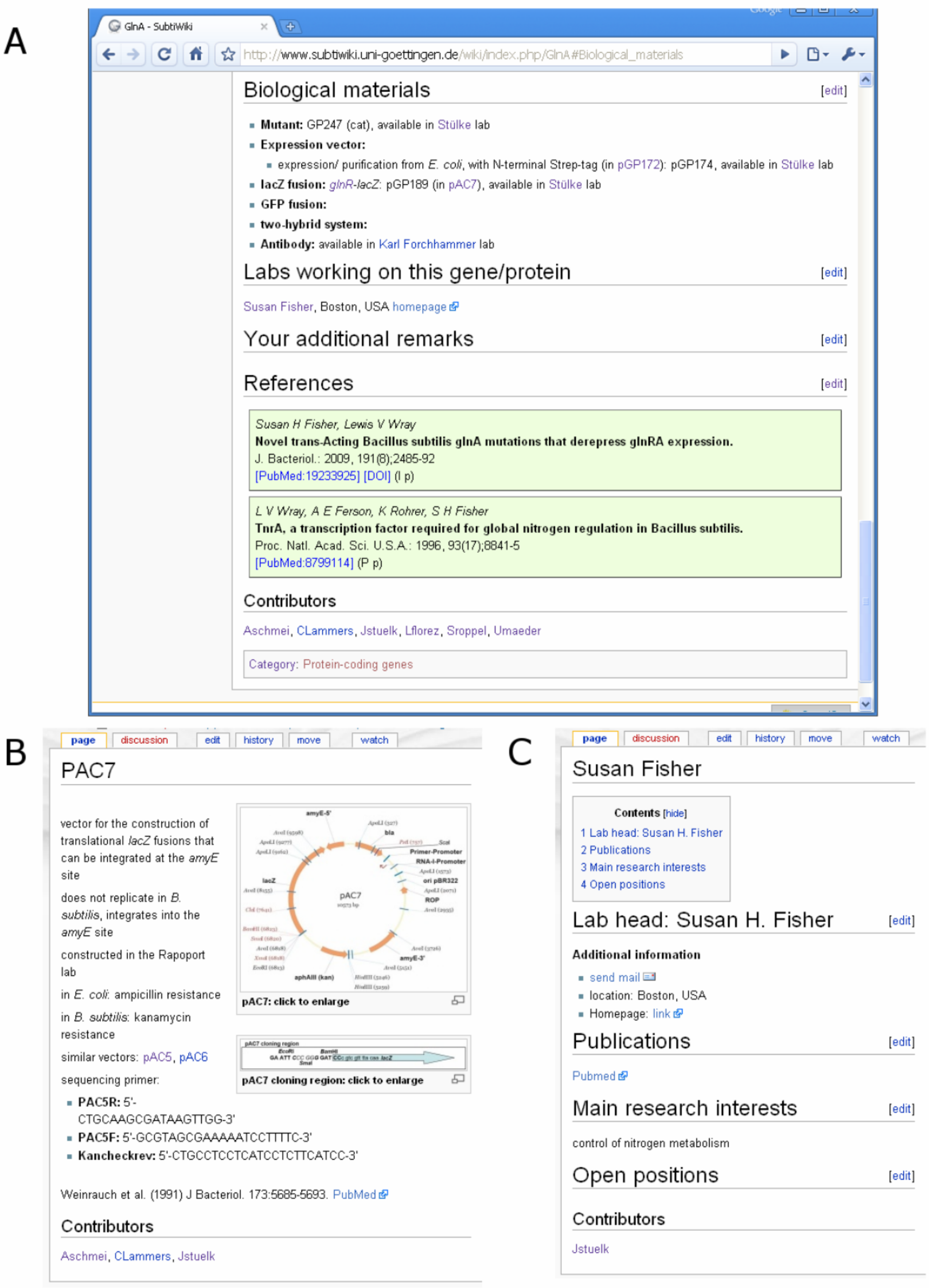
Other pages that enrich the content

Apart from the gene pages, there are additional wiki pages for plasmids, methods and *Bacillus* labs (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 *Subti*Wiki was the possibility of simple editions of each page by any qualified user. After a short registration procedure (the link can be found on top of each page), the user can login, and the user 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 *Subti*Wiki. 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 *Subti*Wiki 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.



Biological materials [edit]

- **Mutant:** GP247 (cat), available in Stülke lab
- **Expression vector:**
  - expression/ purification from *E. coli*, with N-terminal Strep-tag (in pGP172): pGP174, available in Stülke lab
- **lacZ fusion:** *glnR-lacZ*: pGP189 (in pAC7), available in Stülke lab
- **GFP fusion:**
- **two-hybrid system:**
- **Antibody:** available in Karl Forchhammer lab

Labs working on this gene/protein [edit]

Susan Fisher, Boston, USA homepage

Your additional remarks [edit]

References [edit]

Susan H Fisher, Lewis V Wray
Novel trans-Acting Bacillus subtilis glnA mutations that derepress glnRA expression.
J. Bacteriol.: 2009, 191(8);2485-92
[PubMed:19233925] [DOI] (I p)

L V Wray, A E Ferson, K Rohrer, S H Fisher
TnrA, a transcription factor required for global nitrogen regulation in Bacillus subtilis.
Proc. Natl. Acad. Sci. U.S.A.: 1996, 93(17);8841-5
[PubMed:8799114] (P p)

Contributors

Aschmei, CLammers, Jstuelk, Lflorez, Sroppel, Umaeder

Category: Protein-coding genes

PAC7

vector for the construction of translational *lacZ* fusions that can be integrated at the *amyE* site

does not replicate in *B. subtilis*, integrates into the *amyE* site

constructed in the Rapoport lab

in *E. coli*: ampicillin resistance

in *B. subtilis*: kanamycin resistance

similar vectors: pAC5, pAC6

sequencing primer:

- **PAC5R:** 5'-CTGCAAGCGATAAGTTGG-3'
- **PAC5F:** 5'-GCGTAGCGAAAAATCCTTTTC-3'
- **Kancheckrev:** 5'-CTGCCTCCTCATCCTCTTCATCC-3'

pAC7: click to enlarge

pAC7 cloning region: click to enlarge

Weinrauch et al. (1991) J Bacteriol. 173:5685-5693. PubMed

Contributors

Aschmei, CLammers, Jstuelk

Susan Fisher

Contents [hide]
1 Lab head: Susan H. Fisher
2 Publications
3 Main research interests
4 Open positions

Lab head: Susan H. Fisher [edit]

Additional information

- send mail
- location: Boston, USA
- Homepage: link

Publications [edit]

Pubmed

Main research interests [edit]

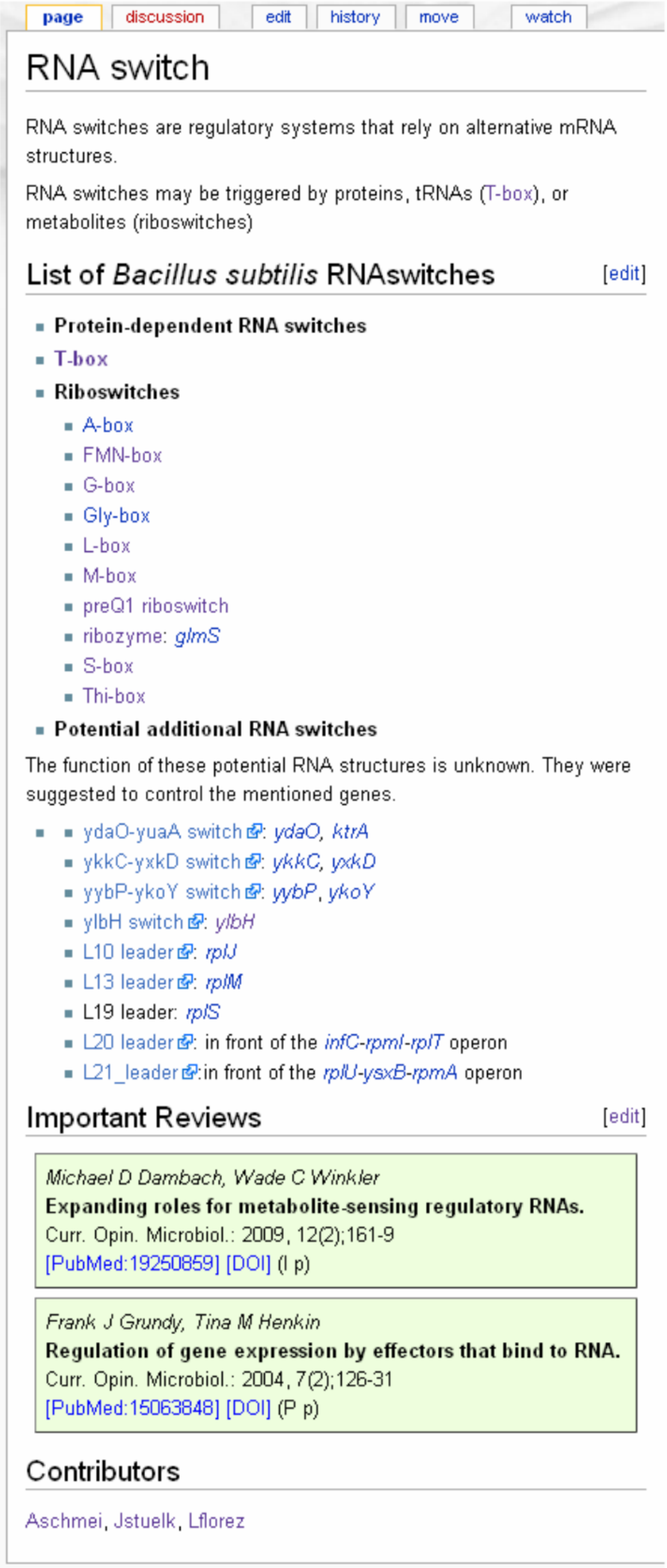
control of nitrogen metabolism

Open positions [edit]

Contributors

Jstuelk

Figure 2. Information on current research in *Subti*Wiki. After the section on the molecular biology of the gene/gene product (see Figure 1), each *Subti*Wiki page contains a section on biological materials available in the community, a list of researchers working on the gene, and a list of publications (A). The former two sections contain links to special purpose material and methods (B) and researcher (C) pages, respectively. Note that, some publications have temporarily been left out for the purpose of this figure.



RNA switch

RNA switches are regulatory systems that rely on alternative mRNA structures.

RNA switches may be triggered by proteins, tRNAs (T-box), or metabolites (riboswitches)

List of *Bacillus subtilis* RNAswitches [edit]

- **Protein-dependent RNA switches**
- **T-box**
- **Riboswitches**
  - A-box
  - FMN-box
  - G-box
  - Gly-box
  - L-box
  - M-box
  - preQ1 riboswitch
  - ribozyme: *glmS*
  - S-box
  - Thi-box
- **Potential additional RNA switches**

The function of these potential RNA structures is unknown. They were suggested to control the mentioned genes.

-
  - ydaO-yuaA switch: *ydaO, ktrA*
  - ykkC-yxkD switch: *ykkC, yxkD*
  - yybP-ykoY switch: *yybP*, *ykoY*
  - ylbH switch: *ylbH*
  - L10 leader: *rplJ*
  - L13 leader: *rplM*
  - L19 leader: *rplS*
  - L20 leader: in front of the *infC-rpml-rplT* operon
  - L21_leader:in front of the *rplU-ysxB-rpmA* operon

Important Reviews [edit]

Michael D Dambach, Wade C Winkler
Expanding roles for metabolite-sensing regulatory RNAs.
Curr. Opin. Microbiol.: 2009, 12(2);161-9
[PubMed:19250859] [DOI] (I p)

Frank J Grundy, Tina M Henkin
Regulation of gene expression by effectors that bind to RNA.
Curr. Opin. Microbiol.: 2004, 7(2);126-31
[PubMed:15063848] [DOI] (P p)

Contributors

Aschmei, Jstuelk, Lflorez

Figure 3. Page on RNA switches in *Subti*Wiki. The main page of *Subti*Wiki links to wiki pages that categorize the content. Shown here is one of these pages, centered on the RNA switches present in *B. subtilis*. Some links (e.g. "T-box") lead to more specific wiki pages on the particular type of riboswitch, with a list of genes containing this regulatory element. Note that, some publications have temporarily been left out for the purpose of this figure.

Features of *Subti*Pathways

*Subti*Pathways is aimed at a comprehensible presentation of metabolic pathways and their regulation in *B. subtilis*. As a starting point for the generation of *Subti*Pathways, 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 modelers 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 color 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.

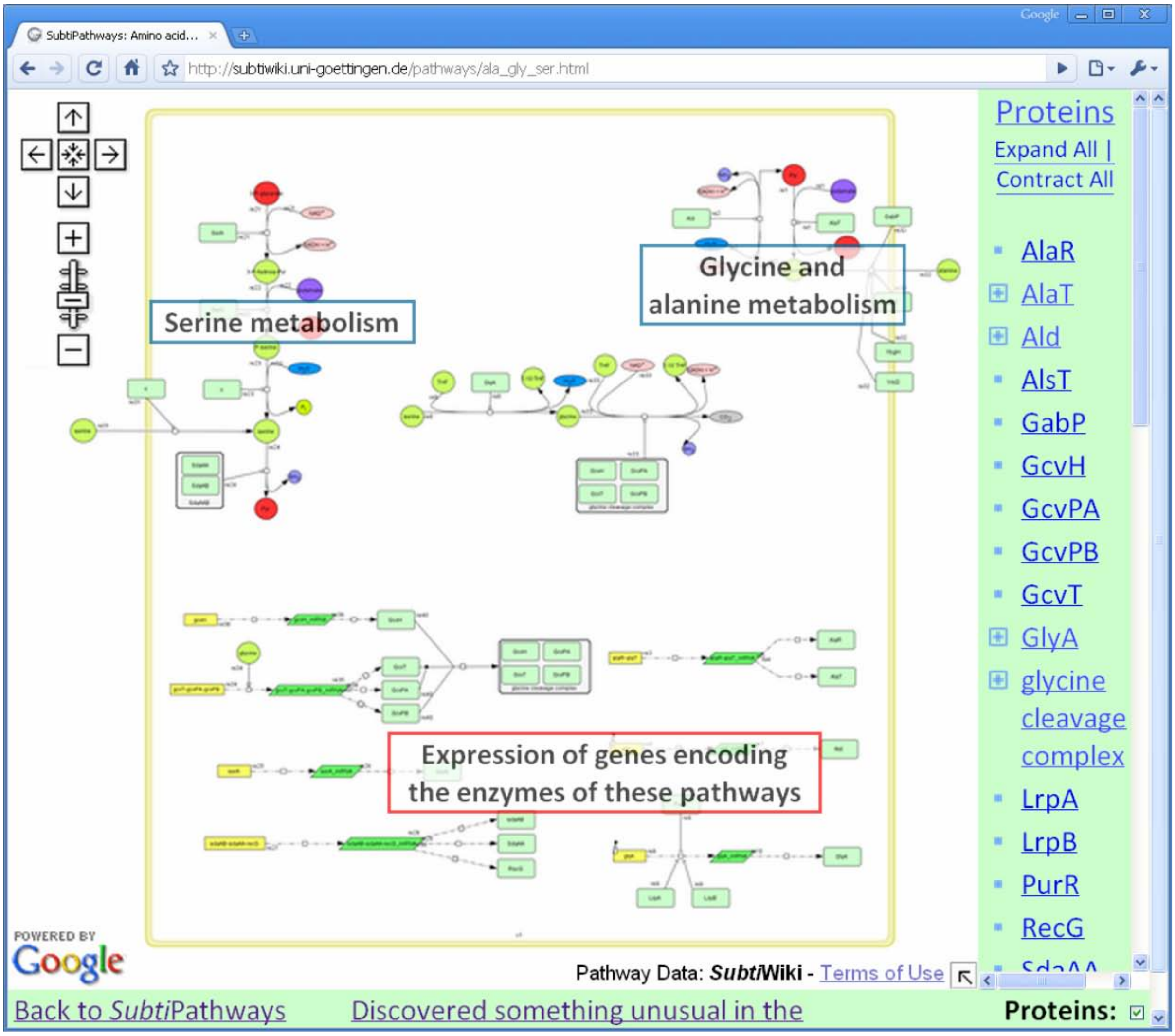

SubtiPathways: Amino acid...
http://subtiwiki.uni-goettingen.de/pathways/ala_gly_ser.html
Serine metabolism
Glycine and alanine metabolism
Expression of genes encoding the enzymes of these pathways
Proteins
Expand All |
Contract All
AlaR
AlaT
Ald
AlsT
GabP
GcvH
GcvPA
GcvPB
GcvT
GlyA
glycine cleavage complex
LrpA
LrpB
PurR
RecG
SdaAA
POWERED BY Google
Pathway Data: *Subti*Wiki - Terms of Use
Back to *Subti*Pathways
Discovered something unusual in the
Proteins: ☑

Figure 4. Screenshot of the *Subti*Pathways interface. *Subti*Pathways is composed of several diagrams on the metabolism of *B. subtilis* and its regulation. The page shown here corresponds to the metabolism of glycine and serine. At lower levels of magnification, the different areas of the diagram are marked with labels. The right part of the page contains direct links to the metabolites and proteins in the diagram. The diagram can be navigated as in Google Maps (for example, zooming and drag-and-drop navigation).

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 responsible enzymes. 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 enzymatic 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 labeled with a clickable icon that provides additional information in a pop-up window. For the proteins, the window contains a short description (taken from *Subti*Wiki) and a link to the corresponding *Subti*Wiki 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 *Subti*Wiki 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 *Subti*Pathways diagrams is indicated in a tab.

It is important to note that the main aim of *Subti*Pathways is to allow easy and intuitive navigation in the rather complex metabolic and regulatory networks of *B. subtilis*. As *Subti*Wiki is focused on gene products, the proteins are also central in *Subti*Pathways. These two features distinguish *Subti*Pathways from other initiatives centered 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*., 2008; Pico *et al*., 2008).

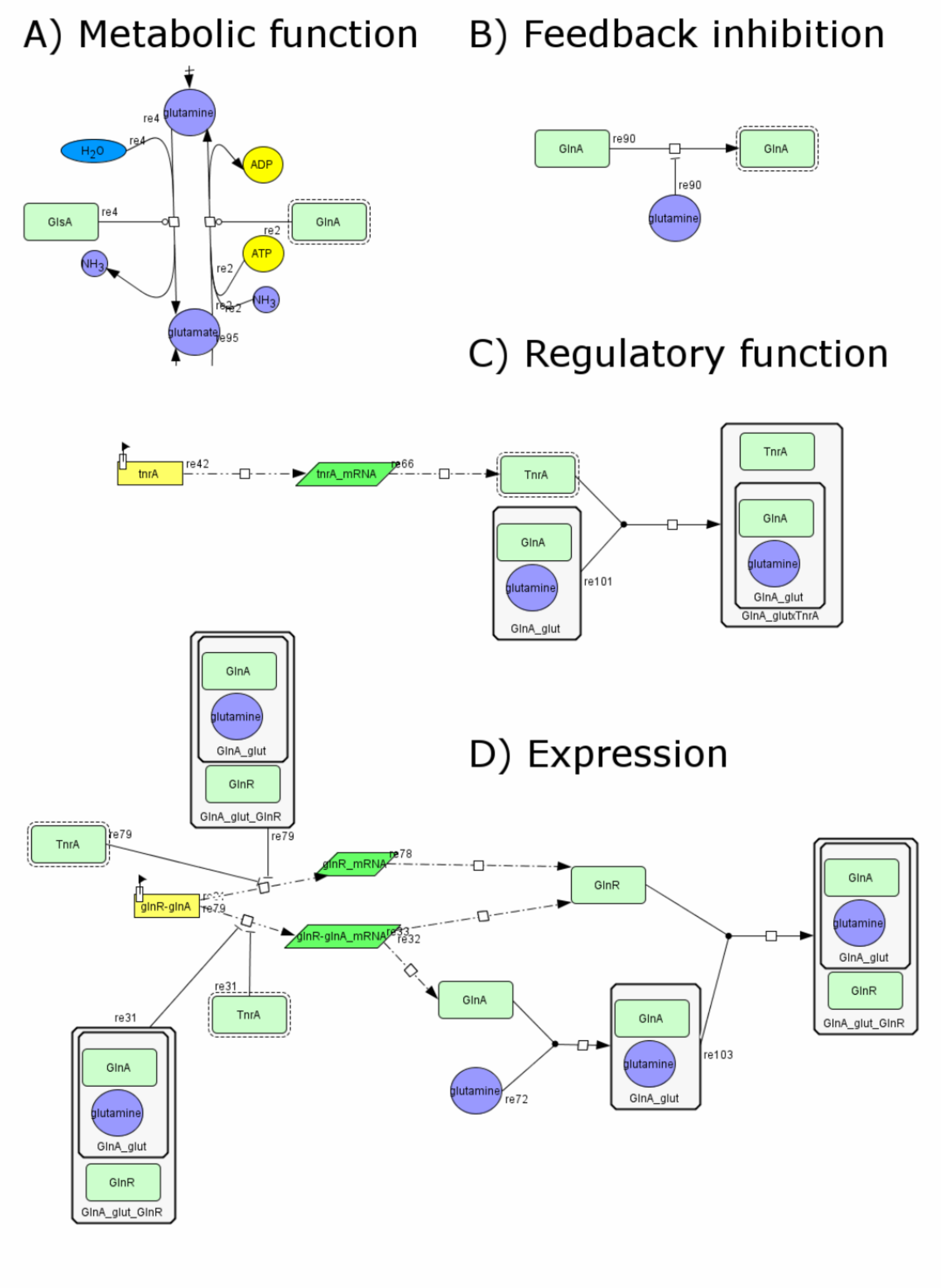
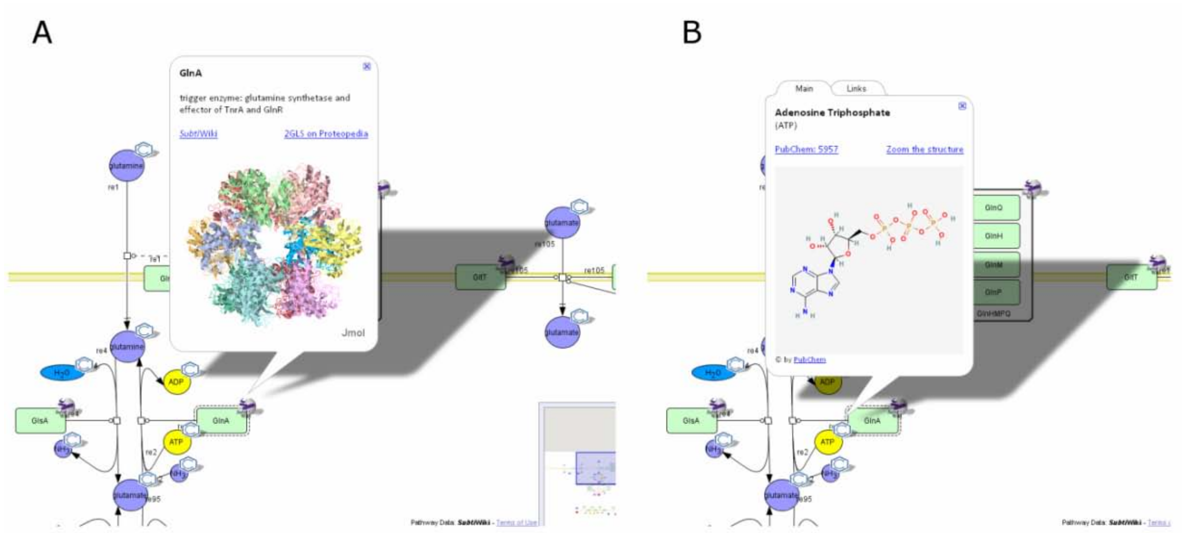


Figure 5. The different aspects of a gene/ protein in *Subti*Pathways. The diagrams of *Subti*Pathways cover a broad spectrum of metabolic and regulatory events, including A) enzymatic reactions, B) feedback mechanisms, C) regulation of protein activity, and D) regulation of gene expression. It also represents active and inactive states where required (via dotted lines around the protein).


GlnA

trigger enzyme: glutamine synthetase and effector of TnrA and GlnR

SubtiWiki 2GLS on Proteopedia

Jmol

Main Links

Adenosine Triphosphate
(ATP)

PubChem: 5957 Zoom the structure

© by PubChem

Figure 6. Pop-up windows for proteins and metabolites in *Subti*Pathways. At the highest level of magnification the diagrams present clickable logos with additional information for proteins and metabolites. A) A pop-up for a protein that contains a link to *Subti*Wiki, to the structures centered wiki Proteopedia, and a 3D molecular structure if available. Not shown here is a tab that appears on top when the protein is present in more than one diagram. B) A pop-up for a metabolite that contains a link to the PubChem database, a representation of the structure, and an extra tab (top) with links to other diagrams that contain this metabolite.

Perspectives

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

To unfold its full potential, *Subti*Wiki 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.

*Subti*Wiki 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 *Subti*Wiki may certainly be relevant to those who study other organisms. Second, the layout of *Subti*Wiki might be a useful basis for a wiki database for other micro-organisms; the source code is freely available on request.

4. CellPublisher: a web platform for the intuitive visualization and sharing of metabolic, signalling, and regulatory pathways

Authors' contributions:

LAF and JS planned the project. CRL provided the starting programs. LAF adapted these programs and programmed the web server and offline versions of CellPublisher. RM implemented additional features under the supervision of LAF. LAF and JS wrote the manuscript. LAF created the supplementary material and figures. All authors read and approved the final manuscript.

4.1 Abstract

Systems biology relies increasingly on collaborations between several groups with different expertise. Therefore, the systems biology community is adopting standards that allow effective communication of concepts, as well as transmission and processing of pathway information. The Systems Biology Graphical Notation (SBGN) is a graphical language for biological pathways that has both a biological as well as a computational meaning. The program CellDesigner allows the codification of biological phenomena in an SBGN compliant form. CellPublisher is a web server that allows the conversion of CellDesigner files to web-based navigatable diagrams based on the user interface of Google maps. Thus, CellPublisher complements CellDesigner by facilitating the understanding of complex diagrams and by providing the possibility to share any CellDesigner diagram online with collaborators and get their feedback. Due to the intuitive interface of the online diagrams, CellPublisher serves as a basis for discovery of novel properties of the modelled networks.

Availability. The freely available web server and the documentation can be accessed at: http://cellpublisher.gobics.de/. The source code and the offline version for Microsoft Windows are freely available at http://sourceforge.net/projects/cellpublisher/.

4.2 Introduction

The detailed understanding of biological pathways and the design of novel molecular pathways with engineered properties lies at the core of systems biology. To meet these goals, two complementary approaches are combined in modeling: simulation and graphical representation. While the former aims to mimic the behaviour of the pathway *in silico*, the latter aims to make the pathway properties and structure understandable to humans.

The pace of discovery in systems biology has been greatly accelerated by the adoption of standards by the community (e.g. SBML). The Systems Biology Graphical Notation (SBGN, Le Novère *et al.*, 2009) is likely to become the standard for graphical representation. Adequate software tools like CellDesigner (Funahashi *et al.*, 2008) are freely available and actively used.

While several online databases such as KEGG, BioCyc, Reactome, Panther, Pathway projector, or *Subti*Pathways exist for visualizing centrally curated pathways (Bauer-Mehren *et al.*, 2009, Kono *et al.*, 2009), only few of them are adopting a standard graphical language (e.g. Panther and *Subti*Pathways; Lammers *et al.*, 2010; Mi *et al.*, 2010). Moreover, as novel pathways are constantly discovered or designed, there is a specific need for presenting custom pathways and receiving feedback from other scientists.

So far, authoring diagrams for online sharing is possible using the web servers BioPP, Payao and WikiPathways (Viswanathan et al., 2007; Matsuoka et al., 2010; Kelder et al., 2009). However, these servers do not allow the generation of easily navigatable diagrams which are required for the evaluation of pathway maps by the non-specialist busy lab scientist. With the advent of Asynchronous JavaScript and XML (AJAX) technologies in the web, which have given rise to applications like Google maps, the possibilities for navigation of graphical information have been greatly improved.

To complement the powerful authoring capabilities of CellDesigner with an improved visualization interface and to overcome the limitations of the existing pathway presentation tools, we have developed CellPublisher, a web server for the intuitive visualization and sharing and discussion of metabolic, signalling, and regulatory pathways based on the interface of Google maps.

4.3 The web server

User input

In the design of the input screen we have focused on making the submission procedure as fast, simple and intuitive as possible. Apart from entering the author and title of the pathway, the user is asked to upload the CellDesigner file and an image of it. The image can be exported easily from CellDesigner and can be modified with other programs before uploading (see below). Moreover, the user has the option to enter an e-mail address, to get notified when the conversion is finished.

Upon submission, the user is directed to an admin page with links to download a local copy of the interactive pathway, and to the online version of the diagram. Both versions are basically identical, but the local copy can be modified further and shared on any web server. In addition, the admin page includes a link to delete the pathway, together with all user input. The link to the admin page can be bookmarked. It is also accessible from the e-mail sent to the user.

The same functionality of the web server is also available as a stand-alone, “offline” version (see Supplementary material).

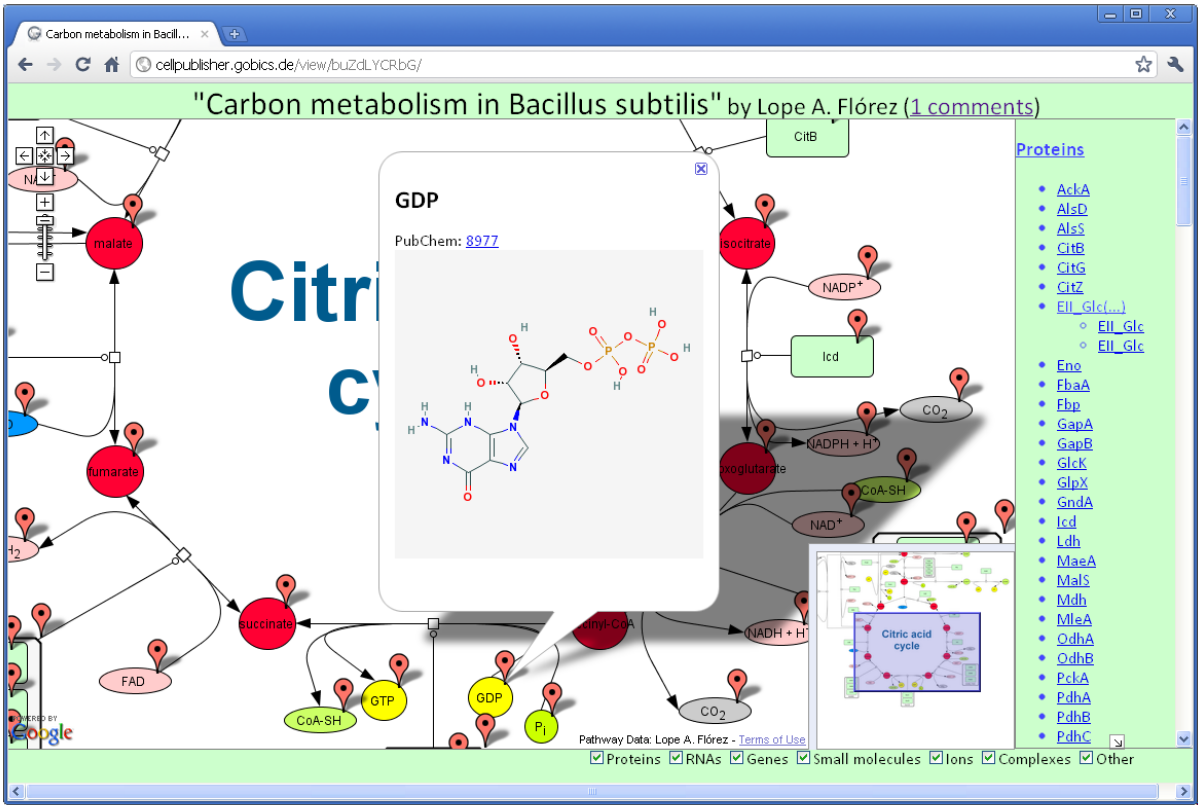
Features of the computed diagram

In the pathway navigation interface (Fig. 1.) the majority of the screen is devoted to the interactive pathway map. It can be zoomed in and out with the mouse, and navigated as is known from Google maps. A bird's eye view on the lower right provides orientation on higher magnifications. At the highest magnification, all the species include a clickable marker that opens a pop-up window with additional information on the species. The content of the window is directly taken from the notes entered in CellDesigner.

For the PDB, PubChem Compound, PubMed, and Uniprot databases, an ID entered in the CellDesigner notes will automatically be linked to the external resource. Moreover, the molecular structure of species linked to PubChem Compound will be incorporated in the info window (Fig 1). Likewise, under Mozilla Firefox, PDB references will also include a Jmol applet with the 3D structure.

The map is framed on the right with an ordered list of all species in it. Clicking on a link in this list will open the pop-up window of the corresponding species in the map. This is useful to explore the various places where a species participates in the diagram.

A forum for each pathway can be accessed through the link on the top right of the page. In this forum, every viewer can give feedback to all other viewers.



Carbon metabolism in Bacil... ×

cellpublisher.gobics.de/view/buZdLYCRbG/

"Carbon metabolism in Bacillus subtilis" by Lope A. Flórez (1 comments)

GDP

PubChem: 8977

Proteins

- AckA
- AlsD
- AlsS
- CitB
- CitG
- CitZ
- EII_Glc(...)
  - EII_Glc
  - EII_Glc
- Eno
- FbaA
- Fbp
- GapA
- GapB
- GlcK
- GlpX
- GndA
- Icd
- Ldh
- MaeA
- MalS
- Mdh
- MleA
- OdhA
- OdhB
- PckA
- PdhA
- PdhB
- PdhC

☑Proteins ☑RNAs ☑Genes ☑Small molecules ☑Ions ☑Complexes ☑Other

Pathway Data: Lope A. Flórez - Terms of Use

Figure 1. Screenshot of an online diagram. The CellDesigner-based representation can be navigated like a Google map. Clickable markers appear on the species at the highest magnification, which contain the information in the CellDesigner notes and external links for selected databases.

Sharing of the diagrams

The main purpose of CellPublisher is to provide an easier way to navigate SBGN compliant diagrams and share them with a wider audience. We envision at least two scenarios where the use of CellPublisher is of particular convenience: publishing and discussion among colleagues and collaborators.

For instance, a publication can cite the link to the online diagram and it will be immediately accessible to everyone without the need of installing programs; the use of any modern web browser is sufficient. The species notes can be enriched with references to other external sources, such as online databases or web pages with extended information. As novel information is discovered, the diagrams can be enriched with comments from the community.

Moreover, the enhanced navigation makes it easier to analyze and discuss a diagram. For instance, it is easier to upload a diagram with need of discussion to CellPublisher, and send the URL to the collaborators, instead of attaching the full CellDesigner file.

As mentioned above, the image of the diagram can be processed with additional programs (e.g. Adobe Photoshop®) before uploading it to CellPublisher. While preparing a discussion, it could be useful to mark in some way uncertain areas of the diagram. These marks will be visible in the online diagram, something that is not possible in current CellDesigner versions.

These features can be especially useful in the early phases of discussion, where an agreement on the pathway connectivity is necessary. In contrast, Payao (Matsuoka *et al*., 2010) could be most useful in later phases, involving the meta-annotation of the individual components by curation experts.

An additional level of customization can be achieved with the local version, which can be obtained from the admin page or the offline version. The look-and-feel and the interactivity can be adapted to specific needs and the resulting pathway can be uploaded to any server. This customization of CellPublisher diagrams is exemplified by *Subti*Pathways, a collection of diagrams covering metabolism and gene regulation in the bacterium *Bacillus subtilis* (http://subtipathways.uni-goettingen.de; Lammers *et al*., 2010).

4.4 Conclusion

As the complexity of the studied pathways increases, the encoded diagrams themselves will increase in size and complexity. This requires software that makes the diagrams understandable in spite of their size. In addition, sharing pathway information with collaborators in a graphical way becomes increasingly important, as well as receiving comments by the community.

CellPublisher is perfectly suited for this task due to its combination of intuitive visualization, the possibility to process custom-made diagrams, the forum for each pathway, and free availability via a web browser.

5. SPABBATS: A pathway-discovery method based on Boolean satisfiability that facilitates the characterization of suppressor mutants

Authors' contributions:

LAF, KG, and JS planned the project. LAF and RP conceived and implemented the SPABBATS algorithm. KG and ST provided the experimental validation of the method. LAF, KG, and JS wrote the manuscript and created the table and figures. All authors read and approved the final manuscript.

5.1 Abstract

Introduction. Several computational methods exist to suggest rational genetic interventions that improve the productivity of industrial strains. Nonetheless, these methods are less effective to predict possible genetic responses of the strain after the intervention. This problem requires a better understanding of potential alternative metabolic and regulatory pathways able to counteract the targeted intervention.

Results. Here we present SPABBATS, an algorithm based on Boolean satisfiability (SAT) that computes alternative metabolic pathways between input and output species in a reconstructed network. The pathways can be constructed iteratively in order of increasing complexity. SPABBATS allows the accumulation of intermediates in the pathways, which permits discovering pathways missed by most traditional pathway analysis methods. In addition, we provide a proof of concept experiment for the validity of the algorithm. We deleted the genes for the glutamate dehydrogenases of the Gram-positive bacterium *Bacillus subtilis* and isolated suppressor mutant strains able to grow on glutamate as single carbon source. Our SAT approach proposed candidate alternative pathways which were decisive to pinpoint the exact mutation of the suppressor strain.

Conclusions. SPABBATS is the first application of SAT techniques to metabolic problems. It is particularly useful for the characterization of metabolic suppressor mutants and can be used in a synthetic biology setting to design new pathways with specific input-output requirements.

5.2 Introduction

A holistic understanding of cellular metabolism is central to systems biology and metabolic engineering: In order to amplify the flux through production pathways in industrial strains we have to understand how the metabolic network responds to our interventions.

Several methods can suggest rational interventions that may lead to favourable industrial phenotypes (see Feist & Palsson (2008) for a review). Their goal is to optimize the distribution of metabolic fluxes towards the product of interest, either directly (e.g. FBA, MOMA or ROOM) or indirectly by coupling it to another characteristic (e.g. OptKnock) that facilitates further strain improvements via mutation and screening.

While these methods can predict a final flux distribution, they do not predict the range of genetic and metabolic responses of the organism after the targeted mutation. At the same time, it would be highly desirable to have tools that may predict these responses, since they can suggest ways to generate more stable strains, or accelerate the adaptation to an intended optimal flux. The challenge of the question is the need to understand why particular genetic responses make sense in an evolutionary setting. Thus, the ultimate question is: Which parallel pathways - that were not active previously - result in an adaptive advantage under the screening conditions?

Pathway analysis has received increased attention due to the reconstruction of genome scale metabolic networks for many organisms. These methods can be divided into two categories: stoichiometric and path oriented (see Planes & Beasley (2008) for a review). The first approach generates all pathways that conform to the pseudo-steady-state assumption for internal metabolites. However, it presents two problems: the number of predicted pathways is in the order of millions for genome scale models, making the approach totally intractable for the question at hand (Klamt & Stelling, 2002). Its second shortcoming is the constraint imposed by the pseudo-steady-state assumption for internal metabolites. This assumption may rule out feasible pathways or (in case we include a large number of "freely available" metabolites) result again in a combinatorial explosion of pathways. The alternative approach - path oriented pathway reconstructions - is advantageous since it usually generates a small (and thus tractable) set of possible pathways. This is due to the choice of starting and ending metabolites and heuristics on the characteristics of the "optimal" pathway. However, the path-oriented approach may result in unrealistic pathways that consume internal metabolites not present in sufficient quantities inside the cell.

What is needed is an algorithm that reconstructs stoichiometrically balanced pathways in increasing order of complexity, with relaxed mass-balance constraints in comparison to the traditional pseudo-steady-state restriction.

A solution based on mixed-integer linear programming (MILP) has been suggested by de Figueiredo *et al.*, 2009, but it has not been used in an evolutionary context so far. Here we describe the use of Boolean satisfiability (SAT, Claessen *et al*., 2009) for the reconstruction of alternative pathways in metabolic networks. Given a set of basal metabolites (that are considered freely available) and a set of target metabolites (whose concentration *must* increase), our SAT method constructs the shortest pathway between the basal and target sets (SPABBATS) of metabolites that is stoichiometrically balanced, while allowing the concentration of the intermediate metabolites to increase, if needed. The constraints are more relaxed than the ones for e.g. flux balance analysis, thus retaining the metabolically significant pathways. Using the algorithm iteratively, we obtain a prioritized list of pathways, whose elements can be tested individually by common molecular biology techniques.

To demonstrate the power of this concept, we applied the SPABBATS algorithm to a complex physiological problem, which is a result of an evolutionary experiment. We have elucidated a novel pathway of glutamate degradation present in the metabolic network of *B. subtilis* that had been decryptified upon inactivation of the normal glutamate catabolic genes. By using our SAT approach, we proposed four different new pathways that could be present in the mutant to utilize glutamate as single carbon source. These predictions were experimentally tested and revealed that one of these pathways was indeed active in the mutant strain and that this novel “suppressor” pathway is required and sufficient for glutamate utilization. This proves that the results of our approach correspond to valid metabolic alternatives for living cells.

5.3 Materials and Methods

Algorithm for finding short pathways between a basis and a target set of metabolites (SPABBATS)

Our approach draws inspiration from flux-balance analysis (FBA Orth *et al.*, 2010) in the sense that it searches the flux space of a metabolic network for fluxes that comply with a set of stoichiometric constraints. The major difference to FBA lies in the optimality criterion; in FBA the value to optimize is the target flux. In our case we change from optimization to satisfiability: we search for a flux that satisfies all the constraints, including a maximum number of allowed reactions.

Another important difference, that is a consequence of satisfiability approach, is that we use two variables for each flux instead of one. The first variable is a positive integer, which is a relative measure of the contribution of that particular flux to the total pathway. The second variable is Boolean and defines whether or not the particular flux takes part in the solution.

As in FBA, we define **S** as the stoichiometric matrix of the network with *n* reactions and *m* compounds. Reversible reactions are split into two unidirectional reactions. We divide the set of compounds into three disjoint sets:

- **B** is the set of basis compounds that are considered freely available, either because they are provided in the medium, or because they are “currency metabolites”, whose concentration is buffered by the whole system (e.g. ATP, ADP, NADH, etc.)
- **T** is the set of target compounds, the ones constrained to be produced in the pathways of interest
- **I** is the set containing all other compounds, that can be intermediates of the resulting pathway

We use different constraints for each of these sets. The compounds in the set **B** are left unconstrained. For each compound in the set **T**, we write a constraint in the form:

$$\sum_{i=1}^{n} s_{ij} a_i b_i > 0 \quad , \qquad (1)$$

where s_{ij} is the stoichiometric coefficient of compound *j* in reaction *i*, and a_{i} and b_{i} are the integer and Boolean valued variables of reaction *i*, respectively. These constraints mean that in the solution pathway the overall flux to these metabolites should be positive.

For the compounds in the set **I** we use a constraint similar to (1), with the difference that we use a “greater than or equal to” ($\geq$) sign. In FBA, an equality sign is used here, to constraint the fluxes to the steady-state space. We purposely do not constrain the pathway to the steady-state space, since the candidate solutions to the problem will not be the only pathway active in the cell and the intermediates that are accumulated in our pathway can be used by other pathways operating in parallel in the system. We require the total flux to these compounds to be non-negative, since the supposition is that they are not present in sufficiently high amounts to allow sustained growth on their consumption.

Next, we add constraints that limit the directionality of reversible reactions. This is done with constraints in the form:

$$b_i + b_j < 2 , \qquad (2)$$

where b_i and b_j are the Boolean variables of two reactions that together characterize a reversible reaction. These constraints mean that no two directions of a reversible reaction can appear in the final pathway at the same time.

Last, we add a constraint for the total length of the solution. This constraint is:

$$\sum_{i=1}^{n} b_i \leq k , \qquad (3)$$

where *k* is a positive integer value that determines the maximum number of reactions that can appear in the pathway. This constraint does not immediately find the best solution, but it puts successively stricter upper-bounds to the maximum number of reactions that are allowed. Thus, it is able to find the shortest solution after some iterations by choosing successively smaller numbers for *k*.

The constraints for the compounds in **T** and **I** are not linear, since each term in the sum is composed of two variables instead of one. For this reason, a linear optimization strategy cannot be used directly. This limitation is not present when we use the SAT-solver HySAT (Fränzle *et al.*, 2007). It is able to find assignments to the variables that satisfy all the constraints in the system, even when these are non-linear. It is also able to detect if no such assignment exists.

If the shortest solution has been found, the best sub-optimal solution can be found by adding an additional constraint in the form:

$$\sum_{i \in K} b_i < k_{op} , \qquad (4)$$

where k_op is the number of reactions in the shortest solution and *K* is the set of indices for the reactions in the shortest solution. In other words, we constrain the sum of all the Boolean variables of the optimal solution to be less than k_op, thus leaving out the shortest solution from the solution space. By iterating this process with the Boolean variables of the sub-optimal pathway, we can find solutions with successively higher number of reactions.

The particular implementation of this algorithm for the problem mentioned in the Results section is as follows: we used the genome-scale reconstruction of *B. subtilis* (Oh *et al.*, 2007). We removed the biomass "reaction"; it is useful for FBA, since it describes the target flux to cellular growth, but is meaningless in our context. In addition, we removed the reaction "glutamate dehydrogenase" (R_GLUDxi) to simulate the conditions of the strain GP717. We also scaled the noninteger stoichiometric coefficients of the model to integer values (and divided by the greatest common denominator). In our case, the set **B** contained the metabolites ATP, ADP, NAD^+, NADH, FAD, $FADH_2$, H_20, H^+, NH_4^+, and glutamate. These "currency metabolites" were chosen due to their participation in most catabolic pathways in the cell. The set **T** contained just 2-oxoglutarate. The remaining compounds were assigned to the set **I**. We set the interval for the a_i to [1, 1000]. The calculations were done using an Intel Core2 Duo processor at 2.66GHz, with 3.25GB of RAM. The first pathway (the one involving leucine as intermediate) was found after 28 seconds. All other pathways took less than 8 minutes each to calculate.

Bacterial strains and growth conditions

All *B. subtilis* strains used in this work are derived from the laboratory wild type strain 168. They are listed in Table 1. *E. coli* DH5α (Sambrook *et al.*, 1989) was used for cloning experiments. *B. subtilis* was grown in C minimal medium containing ammonium as basic source of nitrogen (Wacker *et al.*, 2003). Glutamate and/ or glucose were added as carbon source as indicated. The medium was supplemented with auxotrophic requirements (at 50 mg/l). *E. coli* was grown in LB medium and transformants were selected on plates containing ampicillin (100 µg/ml). LB, SP and CSE plates were prepared by the addition of 17 g Bacto agar/l (Difco) to LB, SP or CSE medium, respectively.

Table 1 – *B. subtilis* strains used in this study

Strain	Genotype	Source or Reference
GP28	trpC2 ΔgudB::cat rocG::Tn10 spc amyE::(gltA-lacZ aphA3)	(Commichau *et al.*, 2007b)
GP717	trpC2 ΔgudB::cat rocG::Tn10 spc amyE::(gltA-lacZ aphA3) gltB1 ansR-C107A	(Commichau *et al.*, 2008)
GP811	trpC2 ΔgudB::cat rocG::Tn10 spc amyE::(gltA-lacZ aphA3) ΔansR::tet	see Materials and methods
GP1154	trpC2 ΔgudB::cat rocG::Tn10 spc amyE::(gltA-lacZ aphA3) gltB1 ansR-C107A ΔansAB::ermC	see Materials and methods

DNA manipulation and transformation

Transformation of *E. coli* and plasmid DNA extraction were performed using standard procedures (Sambrook *et al.*, 1989). Restriction enzymes, T4 DNA ligase and DNA polymerases were used as recommended by the manufacturers. DNA fragments were purified from agarose gels using the Nucleospin Extract kit (Macherey and Nagel, Germany). Phusion DNA polymerase was used for the polymerase chain reaction as recommended by the manufacturer. All primer sequences are provided as supplementary material (Table S1). DNA sequences were determined using the dideoxy chain termination method (Sambrook *et al*., 1989). All plasmid inserts derived from PCR products were verified by DNA sequencing. Chromosomal DNA of *B. subtilis* was isolated as described (Kunst & Rapoport, 1995).

E. coli transformants were selected on LB plates containing ampicillin (100 µg/ml). *B. subtilis* was transformed with plasmid DNA or PCR products according to the two-step protocol described previously (Kunst & Rapoport, 1995). Transformants were selected on SP plates containing tetracyclin (Tet 10 µg/ml), or erythromycin plus lincomycin (Em 2 µg/ml and Lin 25 µg/ml).

Plasmid and mutant strain construction

To express a plasmid-borne *ansR* gene in *B. subtilis*, we constructed plasmid pGP873. For this purpose the *ansR* gene was amplified with the primers KG18 and KG19 using chromosomal DNA of *B. subtilis* as a template. The PCR product was digested with *Bam*HI and *Sal*I and cloned into the overexpression vector pBQ200 (Martin-Verstraete *et al*., 1994).

Deletion of the *ansAB* and *ansR* genes was achieved by transformation with PCR products constructed using oligonucleotides to amplify DNA fragments flanking the target genes and an intervening erythromycin and tetracyclin resistance cassettes from plasmids pDG647 and pDG1514, respectively, (Guérout-Fleury *et al*., 1995) as described previously (Wach, 1996). The PCR products were used to transform GP717 and GP28 for the deletion of the *ansAB* and *ansR*, respectively.

Reverse transcription-real-time quantitative PCR

For RNA isolation, the cells were grown to an OD_{600} of 0.5 – 0.8 and harvested. Preparation of total RNA was carried out as described previously (Ludwig *et al*., 2001). cDNAs were synthesized using the One-Step RT-PCR kit (BioRad) as described (Rietkötter *et al*., 2008). Real time quantitative PCR was carried out on the iCycler instrument (BioRad) following the manufacturer's recommended protocol by using the primers KG26/KG27 for the *ansA* gene, KG38/KG39 for the *ald* gene and KG40/KG41 for the *bcd* gene, respectively. Their recommended data analysis procedure was also used. The *rpsE* and *rpsJ* genes encoding constitutively expressed ribosomal proteins were used as internal controls and were amplified with the primers *rpsE*-RT-*fwd/ rpsE*-RT-*rev* and *rpsJ*-RT-*fwd/ rpsJ*-RT-*rev,* respectively. The expression ratios were calculated as fold changes as described (Rietkötter *et al*., 2008). RT-PCR experiments were performed in duplicate.

5.4 Results

Isolation of a mutation that allows a bypass of the glutamate dehydrogenase for the utilization of glutamate

Glutamate is the most abundant metabolite in a bacterial cell. Although its exact concentration in *B. subtilis* is unknown, it is known to account for about 40% of the internal metabolite pool of an *Escherichia coli* cell (Yuan *et al*., 2009). Glutamate serves as an osmotic regulator (Whatmore *et al*., 1990), as well as universal amino group donor in anabolism thus linking carbon and nitrogen metabolism (Commichau *et al*., 2006). In *B. subtilis*, at least 37 reactions make use of glutamate as cofactor for transamination (Oh *et al*., 2007).

The key reactions of glutamate biosynthesis and degradation in *B. subtilis* are summarized in Figure 1. 2-oxoglutarate, an intermediate of the citric acid cycle, is aminated by the glutamate synthase, encoded by the *gltA* and *gltB* genes. Glutamate degradation to 2-oxoglutarate requires the glutamate dehydrogenase RocG. Additionally, the laboratory strain *B. subtilis* 168 harbours a cryptic gene, *gudB*, coding for an inactive glutamate dehydrogenase. This gene is readily decryptified in *rocG* mutants (Belitsky & Sonenshein, 1998; Commichau *et al*., 2008). In addition, RocG controls the expression of the *gltAB* operon and therefore prevents glutamate biosynthesis in the presence of arginine (Commichau *et al*., 2007a; Belitsky & Sonenshein, 2004).

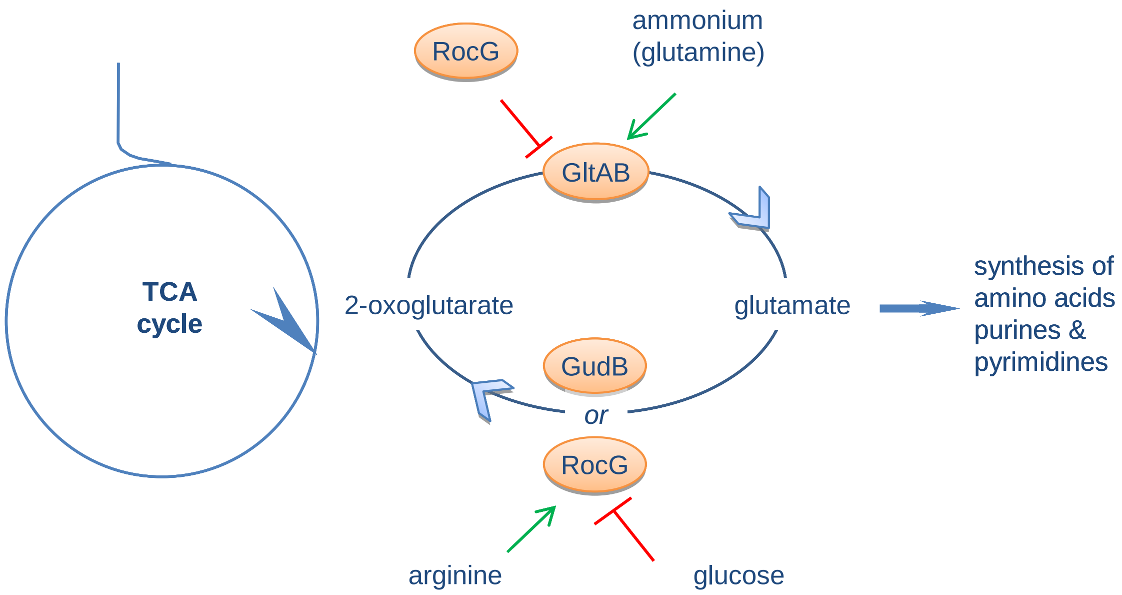


Figure 1. Key reactions for glutamate biosynthesis and degradation in *Bacillus subtilis*. Glutamate is the universal amino group donor in all living cells and in that way links the carbon and nitrogen metabolisms. In *B. subtilis* the synthesis of glutamate depends on the glutamate synthase GltAB. In addition, the genome encodes two glutamate dehydrogenases, RocG and GudB, although the latter is inactive in the laboratory *B. subtilis* strain 168 (see text). The synthesis and degradation of glutamate are tightly regulated in response to the availability of carbon and nitrogen sources.

Inactivation of both the *rocG* and the *gudB* gene results in loss of any glutamate dehydrogenase activity and concomitant inability of the bacteria to utilize glutamate (Belitsky & Sonenshein, 1998; Commichau *et al*., 2008). The *rocG gudB* double mutant strain GP28 grows poorly on SP medium (an amino acid-rich medium) due to the accumulation of degradation products of arginine metabolism (Commichau *et al*., 2007b). However, cultivation of GP28 on SP plates eventually resulted in the isolation of a mutant (GP717) that carries a mutation inactivating the *gltB* gene, encoding a subunit of the glutamate synthase (Commichau *et al*., 2008). This *gltB1* mutation leads to glutamate auxotrophy and might therefore prevent the accumulation of intermediates of arginine degradation. A careful analysis of the mutant strain revealed that it had acquired the ability to utilize glutamate as the only source of carbon and energy. This might have resulted from a re-activation of the *rocG* or *gudB* genes or from the establishment of a novel pathway for glutamate utilization. We tested therefore the *rocG* and *gudB* alleles by PCR analysis. Both the transposon insertion in *rocG* and the replacement of the *gudB* gene by a chloramphenicol resistance gene were identical to the parent strain GP28. Clearly, a new pathway of glutamate degradation was activated in this suppressor mutant that was not active in the wild type and *rocG gudB* mutant cells.

Development of a pathway-finding algorithm

The most reasonable hypothesis to explain the suppression was that the mutation had activated a redundant pathway that is inactive in the wild type strain in a medium with glutamate as single carbon source. Since glutamate is a highly abundant metabolite and is involved as a substrate in 20 reactions in *B. subtilis,* it was not obvious which mutation could have lead to glutamate utilization proficiency in *B. subtilis* GP717.

To address this problem by use of the power of bioinformatics, we developed an approach that harnesses the strengths of Boolean satisfiability (SAT) to find valid pathways (see Materials and Methods). It is able to find short pathways between a basis and a target set (SPABBATS) of metabolites that can operate in a sustained way. It is convenient for its focus on short pathways and the fact that it can calculate pathways that comply with the steady-state constraint. It also allows the relaxation of this constraint, by allowing some metabolites to accumulate if necessary.

The first four pathways suggested by our algorithm are presented in Figure 2. In each case, the first step is a transamination reaction that leads to the production of 2-oxoglutarate. The substrate for transamination is then replenished via the remaining reaction(s) of the pathway. The first pathway (Fig. 2A) involves transamination to form alanine and subsequent oxidative deamination of alanine by the alanine dehydrogenase Ald resulting in the net formation of 2-oxoglutarate. The next two pathways (Fig. 2B) are very similar and involve enzymes of branched amino acid metabolism. In the transamination step, both pathways use the transaminases YbgE and YwaA. The branched chain amino acid dehydrogenase Bcd is then used for the oxidative deamination of the transamination products valine or leucine. Again, the net result of this pathway is the production of 2-oxoglutarate from glutamate. The last pathway (Fig. 2C) requires four steps, (i) the reaction of the aspartate aminotransferase AspB, (ii) the deamination of asparate to fumarate by the aspartase AnsB, (iii) the fumarase reaction (CitG) of the citric acid cycle, and finally (iv) the oxidation of malate by the malate dehydrogenase Mdh. As described for the other pathways, this reaction sequence results in the net formation of 2-oxoglutarate from glutamate. Since the original mutant GP28 did not grow with glutamate as the single carbon source, it is obviously not able to use any of these proposed pathways suggesting that they were activated by a suppressor mutation in GP717.

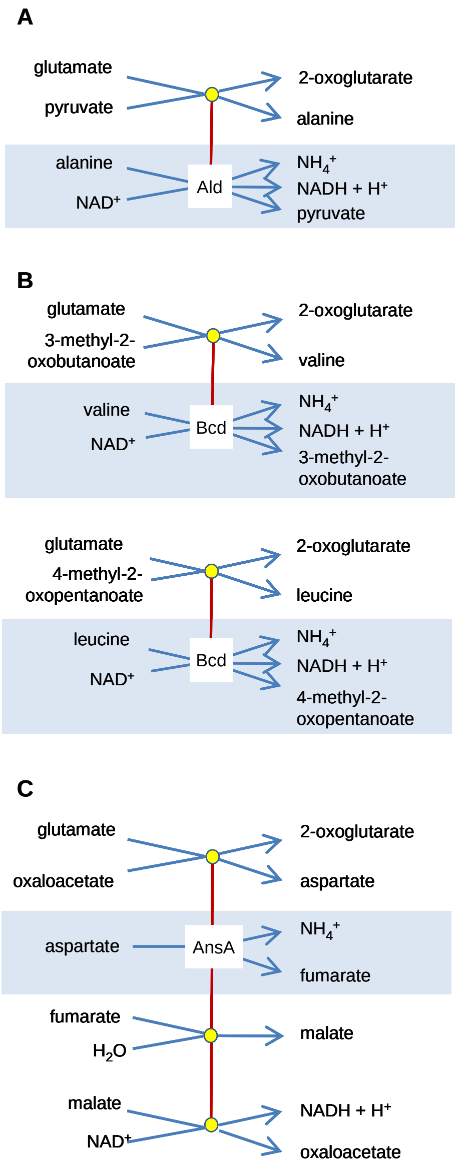


Figure 2. Predictions of alternative pathways for glutamate utilization based on SAT techniques. A *B. subtilis* strain (GP28) was constructed that lacks the glutamate dehydrogenases. An evolutionary adaptation resulted in a strain (GP717) that acquired the capacity to grow on glutamate as single carbon source. Using a SAT based search algorithm (see Materials and Methods) we predicted four alternative pathways that could be activate in the GP717. The genes involved in the highlighted reactions were analyzed further (see text).

Experimental validation of the predictions

Our experiments were performed in minimal medium suggesting that the activity of transaminases was not limiting. Similarly, the two enzymes of the citric acid cycle (CitG and Mdh) are constitutively expressed (Feavers *et al*., 1998; Jin & Sonenshein, 1994; Blencke *et al*., 2003). Thus, the mutation may have affected the expression of one of the deaminases Ald, Bcd or AnsB. This hypothesis was tested by reverse transcription-real-time quantitative PCR. As shown in Figure 3, the levels of *ald* and *bcd* mRNA are comparable for the original mutant GP28 and the suppressor strain GP717. In contrast, a strong increase of the expression of the *ansAB* operon encoding the asparaginase and aspartase was observed for the suppressor mutant that was able to utilize glutamate. This observation suggests that it is the high-level expression of AnsB that allows glutamate utilization in GP717.

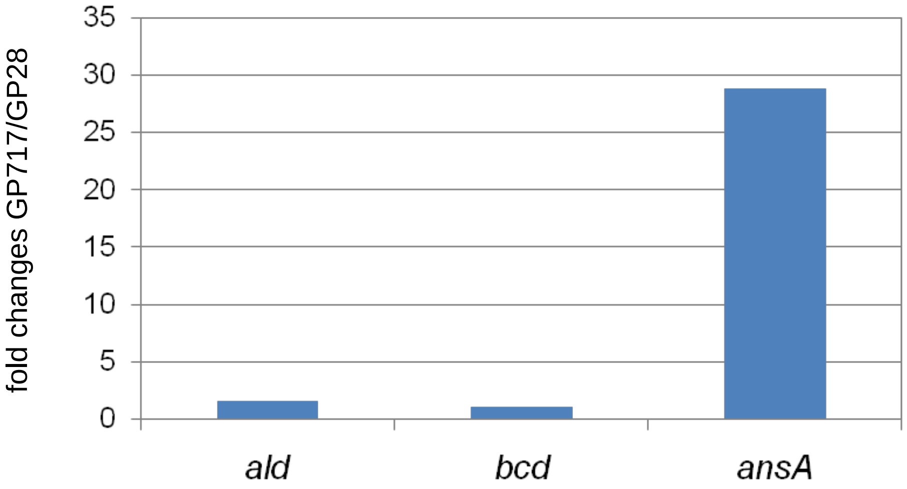


Figure 3. Comparison of gene expression patterns between mutant and parental strains, based on the predictions of the SPABBATS algorithm for pathway analysis. The predictions of the SPABBATS algorithm (see Figure 2) were further characterised by transcription analysis. The expression of the *ald* and *bcd* genes remains constant between the mutant (GP717) and parental (GP28) strains, suggesting that these genes are not involved in the newly activated catabolic pathway. In contrast, the expression of the *ansAB* operon is strongly increased in the mutant. This hints to a gain of function in the mutant strain that was analyzed further.

The involvement of the aspartase AnsB in the novel glutamate utilization pathway was verified by analysing the effect of a deletion of the *ansAB* operon. Growth of the original strain GP28, the suppressor mutant GP717 and its isogenic Δ*ansAB* mutant derivative GP1154 in minimal medium with glutamate or with glutamate and glucose was recorded. As shown in Fig. 4, all three strains were able to grow with glutamate and glucose. In contrast, the deletion of the *ansAB* operon reverted the capability of the suppressor strain of using glutamate as the single carbon source, and the Δ*ansAB* mutant GP1154 was unable to grow with glutamate as was the original strain GP28. This finding strongly supports the idea that the activity of the aspartase AnsB is the reason for the ability of the suppressor strain GP717 to utilize glutamate.

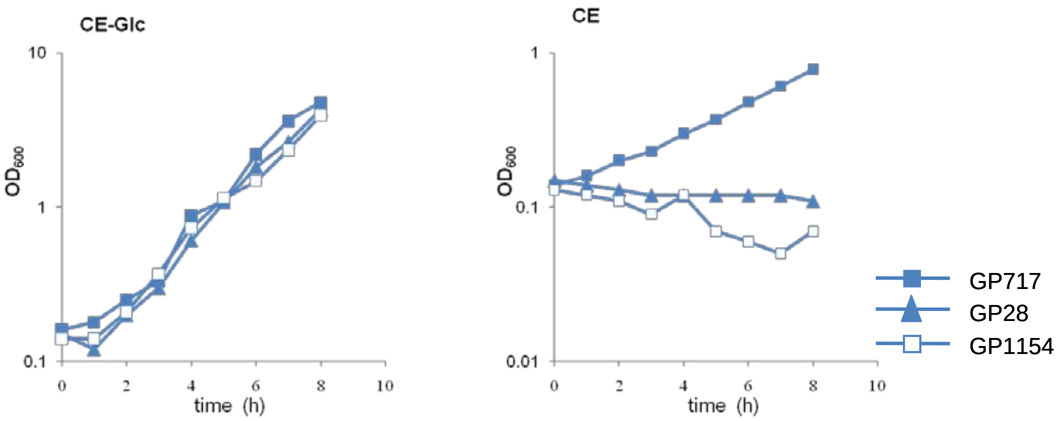


Figure 4. Requirement of the aspartase gene in the alternative pathway for glutamate utilization. The SPABBATS algorithm (see Fig. 2) and the transcription analysis (see Fig. 3) suggested that the overexpression of the asparaginase and aspartase genes (*ansAB*) is the cause for the metabolic gain of function of the mutant strain GP717. To prove this, the *ansAB* operon was deleted in the GP717 strain. The resulting strain GP1154 lost the capacity to utilize glutamate as single carbon source. This strongly indicates that the induction of the aspartase gene is required and sufficient for the newly activated catabolic pathway. CE = Minimal medium containing 0.5% glutamate, CE-Glc = CE medium with an addition 0.5% glucose.

The *ansAB* operon is induced in the presence of asparagine due to inactivation of the AnsR repressor (Sun & Setlow, 1991; Sun & Setlow, 1993; Fisher & Wray, 2002). A comparative analysis of *ansAB* expression revealed about 30-fold induction by asparagine in GP28, whereas the expression levels were unaffected by the availability of asparagine in the suppressor mutant GP717 (data not shown). The observed induction in the wild type strain is good agreement with previous reports. The loss of regulation in GP717 and the high expression of the operon as compared to GP28 suggest constitutive *ansAB* expression that might be the result of an inactivation of the *ansR* repressor gene.

To test the hypothesis that inactivation of the AnsR repressor allowed glutamate utilization by GP717, we performed two tests: First, we deleted the *ansR* gene of the parental strain GP28 and tested the ability of the resulting strain GP811 to grow with glutamate as the single carbon source. Unlike GP28, this strain GP811 (Δ*ansR*) grew in CE minimal medium. Thus, inactivation of the *ansR* gene is sufficient to open a new pathway for glutamate catabolism. In a complementary approach, we complemented *B. subtilis* GP717 with a plasmid-borne copy of the *ansR* gene (present on pGP873) and tested the ability of the transformants to use glutamate. While the control strain (GP717 transformed with the empty vector pBQ200) grew well on CE medium, expression of AnsR from the plasmid completely blocked growth in this medium, *i. e.* the utilization of glutamate. This result confirms that a mutation in the *ansR* gene must be present in GP717 and that it is this mutation, which confers the bacteria with the ability to utilize glutamate via the new aspartase pathway.

To identify the mutation in *ansR*, we sequenced the *ansR* alleles of the parental strain GP28 and the glutamate-utilizing suppressor mutant GP717. While the wild type allele of *ansR* was present in GP28, a C-to-A substitution at position 107 of the *ansR* open reading frame was found in GP717.

This mutation changes codon 36 from UCA (Ser) to UAA (stop) and results in premature translation termination and the formation of an incomplete and non-functional AnsR repressor protein.

Taken together, these experiments confirmed that the metabolic pathway predicted by the SPABBATS algorithm corresponds to a valid metabolic state of the *rocG gudB ansR* mutant strain GP717.

5.5 Discussion

Comparison of SPABBATS with other methods for metabolic analysis

Flux balance analysis (Orth *et al*., 2010) and the majority of methods derived from it are based on constraining the admissible intracellular flux space to steady-state and choosing an adequate optimality criterion to calculate intracellular fluxes. Commonly used optimization criteria are biomass production and the maximization of energy output.

Although these methods predict the essentiality of genes with high accuracy (Oh *et al*., 2007), they are less suited for the characterization of alternative metabolic pathways in viable mutants. On the one hand, by restricting the admissible intracellular flux to steady-state, they discard pathways where a by-product accumulates. Nonetheless, the cell is still viable if this by-product is consumed by other pathways in the cell, not directly related to the process that is studied. SPABBATS solves this problem by allowing a larger flux-space, where intermediate products can accumulate, if necessary.

On the other hand, the optimality criterion can be artificial. For instance, maximizing cellular growth might lead to a theoretical maximum growth rate, or a flux distribution that is as close to the wild-type flux as possible, but it is hard to argue that the regulatory network of the strain is directed to the same target. The pathways discovered by SPABBATS are a structural property of the network and do not depend on an extrinsic optimality criterion (beyond the number of reactions of the resulting pathway). For this reason, the resulting pathways can be interpreted objectively.

Other methods for structural decomposition (e.g. extreme pathways and elementary flux modes, see Planes & Beasley (2008) for a review) rely on the same steady-state restriction of FBA related methods and for this reason share some of their disadvantages. Moreover, SPABBATS does not require the calculation of all possible pathways. Instead, it can be used iteratively to calculate pathways of increasing length, which results in a dramatic improvement in performance for finding relevant pathways in large networks.

An advantage over the method of de Figueiredo *et al*. (2009) is that we do not make use of an optimization framework, but select for satisfiability instead. Similar problems in other areas of computational biology (e.g. Graça *et al*., 2007) show a performance improvement of SAT methods over traditional mixed-integer linear programming methods.

Future perspectives

So far, our analysis of networks using SAT has been restricted to metabolic networks. Nonetheless, since SAT is especially suited for problems that involve Boolean constraints, it is possible to expand the analysis to regulatory networks. For *B. subtilis*, this implies the reconstruction of the metabolic network together with its regulatory complement. This reconstruction is in progress (Goelzer *et al*., 2008; Lammers *et al.*, 2010).

In parallel, we envision the development of novel SAT solvers that are optimized for the solution of metabolic constraints. This will result in the adoption of SAT based methods for metabolic engineering as well as for the design of synthetic circuits that are able to perform computations in the same way as their silicon-made counterparts (Lou *et al*., 2010).

5.6 Conclusions

In this contribution we have shown the use of SAT techniques to discover alternative pathways that connect sets of starting and target species. In addition, we provided a proof of concept for the applicability of the algorithm. We started with a complex physiological problem in *B. subtilis*: the need to characterize a suppressor mutation that allowed growth on glutamate without glutamate dehydrogenases. SPABBATS predicted four potential pathways for glutamate utilization that were decisive to suggest target genes for experimentation. These experiments confirmed the validity of the SPABBATS' prediction, closing the cycle between modelling and wet lab experimentation.

SPABBATS relies on Boolean satisfiability (SAT) to construct the metabolic pathways. SAT has been used for the determination of haplotypes from sequenced genotypes (Graça *et al*., 2007), the analysis of genome biology networks (Chin *et al*., 2008), the understanding of myogenic differentiation (Piran *et al*., 2009), and the characterization of steady states of regulatory circuits (Tiwari *et al*., 2007; de Jong & Page, 2008). Here we report the first application of SAT techniques to metabolic problems.

The SPABBATS algorithm was applied here to a specific problem, the analysis of glutamate metabolism in *B. subtilis*. However, the solution strategies are applicable to a broad spectrum of metabolic problems. For instance, SPABBATS can be particularly useful in the characterization of suppressor mutants. Moreover, SPABBATS can also be useful in synthetic biology. Although used here to find pathways in a reconstruction of the metabolism of *B. subtilis*, it is also possible to use a database of enzymes as the starting model. In this way, it can be used to construct synthetic pathways that satisfy specific input-output and mass-balance requirements.

6. Discussion

6.1 Complementary approaches to study the metabolism of *Bacillus subtilis*

In this work, three different methods were developed to better understand the metabolism of the model bacterium *B. subtilis*. *Subti*Wiki is focused on the genes and gene products of *B. subtilis* and provides a close-up view on the main players (i.e. the enzymes) and their regulation. Genes are connected to each other via the internal links in the wiki. This is well suited to see the relation between two genes, but does not allow seeing the big picture. *Subti*Pathways connects the main players graphically through their interactions in metabolism and regulation. In this way, it complements the individual and paired views offered by *Subti*Wiki. Since these two resources are interlinked, it is easy to go from the close-up to the birds-eye view. Nonetheless, *Subti*Pathways only contains the main pathways. Alternative pathways that are present in the metabolic network, but are not used in wild-type cells due to regulation of the enzymes, are not shown, in the interest of making simpler representations. This gap is closed by SPABBATS, which can find any pathway connecting pairs of metabolites. As shown in Chapter 5, this can elucidate the hidden metabolic potential of *B. subtilis.*

As mentioned in the Introduction, systems biology combines different approaches to understand phenomena involving a large number of different molecules. It also combines different groups with specific expertise. Although the tools developed here are helpful for all three approaches ("omics", modeling, and databases), the three tools have been created with a specific user group in mind: the molecular biologist in the wet-lab. The wet scientist focuses most of the time on individual molecules. As Chapter 2 makes clear, the design of the wiki pages provides a fast user experience in this context. Moreover, the wet scientist requires a deep understanding of the position of the individual proteins on the overall metabolic context. Instead of treating the system as a "black-box" or an abstract data structure, this user usually has a very good biochemical background and is trained in seeing the relationship between events in metabolism. This explains the focus on pathways in *Subti*Pathways and SPABBATS. Pathways allow putting individual enzymes into a broader metabolic context.

This way to approach metabolism differs from the one followed by modelers. Clearly, every model has a different purpose, but in general their aim is to extract the key features of the system (e.g. connectivity, stoichiometry, kinetics) in order to predict or describe systemic properties (e.g. structural patterns, optimal flux distributions, or metabolic control parameters). In other words, the focus of the modeler is on abstraction rather than on detail (a good example of this can be found in Alon (2007)). Nonetheless, in order to create the models, the biological knowledge first needs to be converted to a computer readable form. In this context, *Subti*Wiki and *Subti*Pathways are very valuable resources, since the information in them is structured consistently (in contrast to the scientific literature) and already uses the data representations of modelers, like the Systems Biology Graphical Notation.

In broad terms, the different expertise groups are specialized on specific systems biology approaches. For instance, the molecular biologists and biochemists are experts for molecular and “omics” techniques, and modelers are specialized in putting their knowledge into a useful computational representation. The third approach, the development of databases and standards, is the specific task of curators. The goal of curation is to structure the available knowledge in a way that makes it useful for new approaches, outside the field where it was generated (e.g. to allow comparisons between organisms). To achieve this, the data is stored in specific formats using established pipelines. In addition, it is interconnected to other specialized resources, like databases and web servers. Entering information in a consistent way requires training by specialists, for example at the European Bioinformatics Institute.

*Subti*Wiki can significantly aid in the development of external databases by expert curators. First of all, it contains up-to-date knowledge. For instance, it contains information on some RNA genes and small peptides that are not present in the current genome annotation at the NCBI (Barbe *et al*., 2009). In addition, most of the information entered is linked to the corresponding scientific literature. Going through literature is the most time consuming step in manual curation (Winnenburg *et al*., 2008). Finally, *Subti*Wiki is already interconnected with several external resources (see Chapters 2 and 3).

Due to its concept, *Subti*Wiki as a whole does not meet all standards of curated databases. For instance, since every user can contribute, there are no established pipelines for data entry and the quality control depends on the engagement of other users (see below). Moreover, the data is entered in text form in a flexible way. This flexibility can become an obstacle to extract information via computer programs (a common use of established databases). In addition, the text entered is not in the form of controlled vocabularies (i.e. ontologies). Standard ontologies are a very important step to compare information gathered from different organisms (Ashburner *et al*., 2000).

Nevertheless, although the whole wiki is not subject to the same standards as the ones of other databases, some parts of it are more structured than others, like e.g. the table with most important information. This opens the possibility for introducing standard curation methods to parts of the content on the page. In addition, the fact that anyone can update information opens a door for a collaborative curation (see below).

6.2 Visualization of complex data sets in the context of cellular metabolism

Apart from its focus on *B. subtilis*, this work introduced visualization strategies for complex metabolic information that are applicable to any organism. Visualization can be defined in a broad context, as a way to extract meaningful information from complex data. SPABBATS would fall into this definition, since it is able to extract single pathways from the whole metabolic network of an organism.

However, visualization is usually defined in a more restricted sense, implying the *graphical* representation of information in an insightful way. Due to the complex nature of biology, good visualization tools are crucial for making sense out of the large amounts of data. For this reason, developing software for data visualization is an active field, and a myriad of tools have been created in the areas of genomics (Nielsen *et al*., 2010), sequence analysis and phylogenetics (Procter *et al*., 2010), images of cells and organisms (Walter *et al*., 2010), molecular structures (O'Donoghue *et al*., 2010b), and systems biology (Gehlenborg *et al*., 2010).

Saraiya *et al*. (2005a) and Kono *et al*. (2009) have analyzed the requirements of biologists when using pathway-centric software. Many of the requirements (e.g. the capacity to construct new pathways, meaningful node representations, the possibility to add references to the nodes, etc.) are already present in CellDesigner. Nonetheless, CellDesigner misses other requirements, like the possibility to collaborate on pathways, interconnectivity between pathways, the possibility to get an overview on the whole process (e.g. through a good zoomable user interface), and the possibility to link data from other biological scales (e.g. molecular structures).

In this context, the interface underlying *Subti*Pathways and CellPublisher is a step forward in the development of intuitive visualization software, since these tools comply with these missing requirements: CellPublisher can enhance the collaboration between groups (see Chapter 4); *Subti*Pathways diagrams are connected through the info windows; the Google maps API, common to both tools, allows zooming in and out and selectively showing information on higher levels that is not present on more detailed levels; and finally, the info windows permit to show information on lower biological scales. The open-source nature of CellPublisher allows re-using these features to complement other software tools. Future work could focus on implementing some of the missing requirements, like e.g. the capacity to map high-throughput data on the pathways. Moreover, it would be useful to explore broader requirements, like the ones outlined by O’Donoghue *et al*. (2010a).

Apart from the requirements for systems biology software, it is also useful to think about the principles underlying simple and intuitive computer visualizations. In an information age, visualization of complex data is relevant in a variety of contexts, not only in biology. Some of the principles may be transferred between fields. For instance, innovative visualizations of data have been developed in the context of public health (Rosling 2007, see also http://www.gapminder.org/) or journalism (McCandless 2010, see also http://www.informationisbeautiful.net/). Maeda (2006, see also http://lawsofsimplicity.com/) has formulated general principles for simple design. Some of these principles can be found in CellPublisher. Two examples are hiding complexity through nested designs (e.g. by showing detailed information in info windows instead of the diagram), and to base the interfaces on something the users already know (e.g. Google maps for CellPublisher, or Wikipedia for *Subti*Wiki). Moreover, systems biology can profit from other scientific disciplines focused specifically on this area, like the field of human-computer interactions in computer science. Some advances in this synergy have already been made (Saraiya *et al*., 2005b).

6.3 Capabilities and limitations of the tools for pathway analysis

The previous sections have provided examples of various contexts where the tools of this thesis are useful, but it is equally important to understand their limitations in related areas. On the one hand, this hints to opportunities for improvement. On the other hand, in some contexts these limitations can become very useful to obtain new knowledge in an indirect way.

The core strength of SPABBATS is the way it finds a set of reactions (i.e. a pathway) that satisfies certain restrictions on the use of metabolites (see Introduction and Chapter 5). Restrictions on reactions (e.g. gene knock-outs) can be added explicitly by setting the corresponding reaction variable to zero. Nonetheless, it is not possible to restrict the reactions indirectly via regulatory mechanisms. Although this can be an advantage in some situations (see Chapter 5), the method would need to be expanded in the case that regulatory interactions are an essential part of the problem. Some frameworks (Covert *et al.*, 2001; Gianchandani *et al.*, 2006; Shlomi *et al*., 2007; Covert *et al.*, 2008; Kotte *et al.*, 2010; and others) incorporate both regulatory and metabolic constraints. Interestingly, in many cases they use a Boolean framework to describe the metabolic network of the cell. Since Boolean satisfiability is an especially active field in computer science and electronic engineering, an incorporation of these constraints into the SPABBATS framework seems reasonable and feasible.

Another limitation of SPABBATS is that it depends on an exhaustive network reconstruction. For instance, when a mutation changes the specificity of an enzyme, new pathways become possible, since the metabolites are connected to each other in new ways. These new pathways would not be discovered by SPABBATS, unless the new specificity of the enzyme is also included in the original model. The same is true if the organism possesses an enzymatic function that has not been characterized yet, and is thus not included in the model. On the other hand, when this limitation is understood, it can be used as an advantage. As mentioned in the Introduction, systems biology proceeds in cycles between experimentation and modeling. When a model provides no explanation for an observed phenomenon, it is an indication that no matter how complex the model already is, it still misses interesting new data. This allows research to focus on the relevant aspects, as shown for instance by Nakahigashi *et al*. (2009).

*Subti*Pathways has two main advantages over SPABBATS. The first one is the simple visualization of pathways (see above). The second advantage is that *Subti*Pathways includes the regulatory mechanisms for all the enzymes involved (as far as they are understood). Nonetheless, there is an important limitation inherent in the use of CellDesigner. The software depends on a view of metabolism where each enzyme works independently from the others. However, it is becoming increasingly clear that enzymes assemble into macromolecular complexes in *B. subtilis* (Commichau *et al*., 2009) and other organisms (Persson & Johansson, 1989; Michels *et al*., 2006). The role of these interactions is still not fully understood. Several cases of metabolic channeling have been reported, especially for eukaryotes (Malaisse *et al*., 2004). In other cases, the complexes modulate the activity of the whole pathway they participate in (Haanstra *et al*., 2008).

Although CellDesigner offers the possibility to draw molecular complexes, it is not possible to show the pair-wise interactions of the proteins involved. However, these pair-wise interactions might be important determinants of substrate channeling or might affect the functionality of the whole complex. Moreover, these interactions might be transient. Representing several different states of the complex depending on the presence or absence of some proteins will lead to very complex diagrams that are ambiguous in their meaning.

This indicates the need to use other software tools to model enzymatic complexes, besides CellDesigner. A good starting point could be to use software specialized on the representation of molecular networks (e.g. Cytoscape (Shannon *et al*., 2003), see Gehlenborg *et al*. (2010) for a review of other possibilities). A caveat is that the software will represent the different interactions, but not the events that these interactions can trigger. For that reason, the second step could be to use more abstract levels of representation in combination with the previous tools. The Systems Biology Graphical Notation (SBGN) suggests a possible alternative. CellDesigner is based on the “Process Description” language of SBGN. However, SBGN also defines two other languages: the “Entity Relationship” and “Activity Flow” languages (Le Novère *et al*., 2009). Since the “Entity Relationship” language is specialized in showing the way in which an entity (e.g. a phosphatase) affects the state of another entity (e.g. a phosphorylated protein), it could prove useful to indicate the way in which one enzyme in the complex affects the activity of an interaction partner, or the whole complex.

Combining the different visualizations and representations in a clever way is a challenging task, but is likely to be the best overall description of the intricate processes involving macromolecular complexes. This challenge is present throughout systems biology, e.g. at the interface between proteomics and structural biology (Morris *et al*., 2010).

6.4 Considerations for scientific platforms based on user generated content

Two products of this thesis, namely *Subti*Wiki and CellPublisher, allow the users to upload new information and make it available to other groups over the Internet. The purpose of uploading differs between the web sites: in the case of *Subti*Wiki, users update the available knowledge to benefit the whole scientific community working on *B. subtilis* and related bacteria. In contrast, CellPublisher is more appropriate as a communication tool between collaborating labs based on a visual language (see above). Nonetheless, CellPublisher can also be used to create biochemical diagrams for a big audience. In this way it serves as a basis to allow community contribution of pathways to e.g. *Subti*Pathways.

Several curation teams of scientific databases have been interested in interacting more actively with the users in the process of knowledge update. However, the development of scientific wikis is a relatively new trend (Waldrop, 2008). Chapter 2 lists some scientific wikis that appeared before *Subti*Wiki, and new wikis have been created since then (Brohée *et al*., 2010a; Stehr *et al*., 2010; and others). Moreover, some databases have introduced a combined approach, where a wiki complements other modules that follow standard curation procedures (Legeai *et al*., 2010 and Katz *et al*., 2010). In addition, apart from allowing users to participate in data upload and update, some initiatives encourage participants to do computational tasks. Two examples of this are finding stellar dust among millions of images or predicting protein structures my manipulating structural models (Hand, 2010; Cooper *et al*., 2010).

Some drawbacks of the wiki model for scientific databases have been discussed by Arita (2009) and Welch & Welch (2009). Arita points out a technical problem of wikis: since every page can be modified independently of the others, user contribution can lead to data inconsistency between pages. He suggests introducing in-line searches: an extension to the wiki program that connects the pages to a relational database in the background (Arita & Suwa, 2008). In contrast, Welch & Welch mention that wiki users are reluctant to share their data in spite of the availability of technical solutions for online collaboration. They advocate a method for giving users recognition for data input, in a way that benefits their scientific careers.

The aforementioned critics are rightly addressed and complement each other: in-line searches might be too complicated for new users and might discourage new contributions. At the same time, data inconsistency deters the wider audience (which in turn results, in the long term, in lower participation). Nonetheless, under some conditions the benefits of the wiki can neutralize its drawbacks. First of all, there should be a clear need for the resource in the community. For instance, the *Bacillus* community came to depend on SubtiList for everyday research. A tool with the functionality of SubtiList, but with updated information, was strongly needed. It is unlikely that *Subti*Wiki would have been acknowledged if data curation of *B. subtilis* would meet the same standards as e.g. EcoCyc, the encyclopedia on the genes of *E. coli* (Keseler *et al*., 2009).

In principle, the previous condition could be met with a new relational database instead of a wiki. For this reason, the second condition is that the community identifies the wiki as the natural place to present recently published data to the whole community. For instance, GenBank (more precisely the International Nucleotide Sequence Database Collaboration (Benson *et al*., 2010)) is universally acknowledged as the standard repository for sequence data. As *Subti*Wiki is more widely used, it should become the standard repository for other types of information on *B. subtilis*, like e.g. for biological materials, data on transcriptomics, or on protein-protein interactions. The open nature of a wiki allows the groups to upload information directly, or contact the right partners in the community to do so. This would be more complex in the case of a resource managed by a central organization and would likely lead to several different databases, with almost no connection between each other.

The third condition is related to incentives. In many cases, making published information more accessible to other groups is already a big incentive, since this can result in more citations or new collaborations. This is especially the case in a cohesive community open for collaboration, like the one for *B. subtilis*. Nonetheless, as Welch & Welch (2009) point out, other incentives need to come from the community, and in the long term, from journal editors. For instance, a reviewer can request the upload of data to the wiki before accepting a paper. Similar mechanisms are already in course for sequence data, as journal editors expect that authors upload it into GenBank.

In the mid-term a curation team can be incorporated to enhance the data consistency and quality. Moreover, this would be especially helpful if similar initiatives are already on course, e.g. with other organisms. This could lead to better comparisons between the organisms and also to re-use of software and ideas.

Although maintaining data quality by curation is highly desirable, some compromise is required to maintain the usability of the resource and its open nature. It is not realistic to request every user to adhere to strict data entering guidelines. Instead, it should be possible for curators to easily correct the data entered by other participants. One way to do this is by splitting the wiki into two different sections. One section retains all the openness of standard wiki implementations and the other section creates structured annotations based on the previous entry. This could be especially helpful for information that is almost never updated, like the position of a gene in the chromosome. Gene Wiki provides a good example for this practice (Huss *et al.*, 2009). A second option is the use of extensions, as proposed by Arita & Suwa (2008) and Brohée *et al.* (2010b). Well designed, simple extensions can help intermediate and advanced users to correct entered data in a systematic way. Finally, a jamboree can be a good way to coordinate efforts between curators and several groups working on *B. subtilis*. Thiele & Palsson (2010b) present some examples of successful jamborees for metabolic reconstructions.

6.5 Integration in the broader context of systems biology and related fields

The ultimate goal of systems biology is to have a complete understanding of the interplay of all cellular processes. This goal is still distant, even for model organisms such as *E. coli* and *B. subtilis*. First of all, many gene functions remain unknown, so they cannot be easily incorporated into models. Second, the modeling paradigms are under constant change, as is evidenced e.g. by the increasing importance given to protein-protein interactions as opposed to isolated enzymes (see above). Third, to fully understand all cellular mechanisms, it will be necessary to frame them in the context of the evolution of the organism in its natural environment. However, it is not clear how to model these environments (with the exception of nutrient composition in the medium), let alone reproduce them all accurately in the lab. Several new challenges will become evident as we gain more understanding.

In spite of the long way ahead, it is important to develop a roadmap towards this ambitious goal. Understanding the function of each gene will require painstaking molecular biology research (i.e. a bottom-up approach) in combination with pattern discovery through high-throughput methods (i.e. a top-down approach). For both approaches, *Subti*Wiki becomes a central resource, due to its balance of detail (important for the former approach) and up-to-date gene descriptions (relevant for the latter approach). Moreover, the combination of *Subti*Pathways with genome-scale reconstructions of the metabolism of *B. subtilis* (Oh *et al*., 2007; Henry *et al*., 2009) is the most extensive model available of the metabolism of *B. subtilis* and its regulation. As mentioned above, finding out something that cannot be explained by the model (or openly contradicts it) is a direct way to follow the right leads to new knowledge. As this happens, CellPublisher provides a forum to propose alternative explanations (as long as they can be represented using CellDesigner). In this sense, the products of this thesis act as successive steps towards the major goal.

However, the decisive test for our understanding of biological systems lies in our capacity to combine the individual parts in new ways to produce useful products or an engineered behavior. Two emerging disciplines, metabolic engineering and synthetic biology, make use of the concepts gained from systems biology to design biological system with useful new properties.

For instance, metabolic engineering searches for purposeful genetic interventions that significantly improve the efficiency of biotechnological strains (Bailey, 1991). It differs from genetic engineering by its focus on the whole system, as opposed to individual molecules. It also differs from traditional strain improvement methods, like directed evolution, by the fact that each mutation introduced is fully characterized *a priori*. The strains are improved using two complementary strategies (Clomburg & Gonzalez, 2010). First, the metabolic fluxes of the cell are reoriented towards the desired product. Second, the metabolic network is modified to allow the efficient utilization of better (usually cheaper) substrates. Both strategies require mathematical models of metabolism that suggest meaningful interventions. Several methods derived from flux balance analysis have been devised for this purpose and in many cases have made valuable suggestions (see Kim *et al*., 2008, for a review of these methods). SPABBATS could complement them in areas where they are not very effective. For instance, many biotechnological processes take place in the stationary growth phase of the organisms. At the same time, the majority of the current methods are directed towards optimal growth, which is in conflict with this situation. Moreover, a better understanding of the regulatory mechanisms in the strains (a goal targeted by *Subt*iPathways in *B. subtilis*) will be necessary to increase not only the yield, but also the titer and productivity of the respective bioprocesses.

Metabolic engineering also profits from synthetic biology, which can be defined as the design of novel biological systems with defined behavior using engineering principles (like modularity, component testing, standards, etc., see Purnick & Weiss (2009) for a review). Based on these principles, strains can be improved by combining different optimized enzymes into novel pathways not present in the organism. Other applications of synthetic biology have been reviewed by Khalil &

Collins (2010). If a database of enzymes is created, SPABBATS can be used to extract pathways from this database with defined input-output relationships. In addition, since many circuits designed in synthetic biology are a combination of common biological elements, their first design could be performed with CellDesigner. In this context CellPublisher could provide a platform for storing these designs and to share them with collaborators, or the whole synthetic biology community.

Summary and Conclusions

Systems biology is a thriving field that aims to bridge the gap between the detailed knowledge about individual molecular species and the phenomena occurring at the cellular, organism, and higher biological levels. It draws inspiration from general systems theory, where the behaviour of a system (e.g. a cell) is explained by means of interactions of the individual components (i.e. the molecules it is composed of). To achieve this goal, three main approaches converge in systems biology: high-throughput experiments (including genomics, proteomics, transcriptomics, metabolomics, and others); mathematical and computational modelling; and the development of databases and data annotation standards.

These combined approaches have been especially useful to study model organisms. The abundant previous knowledge about these organisms is usually codified in scientific databases. The content of these databases serves as the basis for creating large-scale models. These models guide the generation of new data with high-throughput techniques and serve as a frame to correlate different types of data (e.g. the transcriptome and the proteome). In turn, these correlations provide new knowledge or challenge previous assumptions about the biological system, guiding future experiments.

The focus of this thesis was to assist the progress of systems biology for the model organism *Bacillus subtilis*. This Gram-positive soil bacterium is, together with *Escherichia coli*, the best understood organism in terms of its physiology, genetics, and molecular biology. It is possible to conduct precise genetic modifications of this organism, due to its natural competence and homologous recombination capacity. Under harsh environmental conditions, it can follow various cellular differentiation programs, like sporulation and biofilm formation. The study of these mechanisms provides concepts for the analysis of cellular differentiation in higher organisms. *B. subtilis* is also relevant for biotechnological processes, where it is used to produce engineered proteins and useful compounds, like e.g. some vitamins. Last but not least, it is used as a model for pathogenic Gram-positive bacteria, like *Staphylococcus aureus*, *Clostridium botulinum*, *Bacillus anthracis*, and others.

This thesis addresses several challenges faced by the *Bacillus* community in systems biology research. The first challenge is to obtain up-to-date information about the function of each gene and gene product in this organism. At the onset of this thesis, this information was spread out between several unconnected databases and the scientific literature. The most important database for the organism appeared as a result of the first sequencing efforts, but was not updated further after completion of the sequencing project. To provide an up-to-date database for all genes and gene products of *B. subtilis*, and to allow every member of the community to participate in the maintenance of knowledge on the organism, *Subti*Wiki was created (http://subtiwiki.uni-goettingen.de). It is a wiki based on the same platform as Wikipedia, adapted to suit the needs of molecular biologists and modelers alike. Every gene of *B. subtilis* has a structured page in this wiki, with the most relevant biochemical information on top (e.g. the gene name, its function, the length of the gene and protein, etc.), followed by detailed information on the gene, the protein, its expression, biological materials available in the community, and an updated list of references to scientific literature. In addition, the pages contain links to other pages in the wiki and to external databases. *Subti*Wiki is now used as reference for several systems biology efforts focused on *B. subtilis*.

The second challenge addressed in this thesis was the visualization of the metabolic and regulatory networks present in *B. subtilis*. The most popular online resources for the visualization of metabolic pathways, the KEGG and BioCyc databases, did not contain an accurate representation of the metabolism of this organism. Moreover, they do not contain the information about regulatory networks. A publication that reconstructed the main pathways of *B. subtilis*, together with their regulation, was available at the beginning of the thesis. This motivated an expansion of the published reconstruction based on further literature sources, and the creation of an online tool to visualize this information in a useful way.

The result is *Subti*Pathways (http://subtipathways.uni-goettingen.de), an online platform to visualize the metabolic and regulatory pathways of *B. subtilis*. The pathways of the organism were drawn using a popular program for systems biology called CellDesigner. This program uses different shapes for each class of biomolecules (e.g. DNA, RNA, proteins, metabolites, etc.) and different arrows for each process (e.g. state transition, activation, catalysis, etc.). CellDesigner adheres to standards set by the systems biology community on the representation of biological processes. The diagrams were then converted to an online representation, based on the interface of Google maps. In addition, clickable icons were added to each of the shapes at higher magnification. These icons display info windows that provide a link to metabolic databases, as well as to *Subti*Wiki and the Protein Data Bank.

The interface underlying *Subti*Pathways allows a simple navigation in complex diagrams and the interconnection with external data sources. To make this interface accessible to other groups working with CellDesigner, CellPublisher (http://cellpublisher.gobics.de) was created. This web server allows any user to upload a CellDesigner diagram and obtain an online representation of it, supported on the same Google maps-based navigation underlying *Subti*Pathways. CellPublisher is especially useful for the communication of pathway information between collaborating groups.

The fourth resource created in this thesis, SPABBATS, is focused on extracting useful information of genome-scale computational models of the metabolism of *B. subtilis*. These models are now available for a number of organisms and consist of a list of all metabolic reactions, with the precise stoichiometry, in a computer readable format. The models are used primarily to investigate the possible metabolic fluxes inside of a cell under steady-state conditions and different environmental and genetic constraints. SPABBATS is an algorithm that uses these models as a basis to discover alternative pathways connecting sets of metabolites. In particular, SPABBATS is especially suited to discover ways to bypass a specific metabolic reaction. We used it to discover alternative ways to catabolise glutamate in a strain that lacks the main catabolic enzyme: the glutamate dehydrogenase. In contrast to most existing methods for pathway discovery, SPABBATS predicted several pathways, in increasing order of length, which can be used to sustainably convert glutamate to 2-oxo-glutarate. The predictions of SPABBATS were tested in the lab and served to characterize a suppressor mutant. This mutant had acquired the capacity to catabolise glutamate, in spite of the fact that the two glutamate dehydrogenases of *B. subtilis* were knocked out in this strain.

In combination, these tools allow the Bacillus community to formulate more informed systems biology questions about the organism, as well as to make sense of genome-scale models and experimental techniques. Moreover, the methods developed in this thesis can easily be adapted to the needs of the communities of other model organisms. They can also be used in contexts where systems biology interacts with other areas, such as synthetic biology and metabolic engineering.

References

Affolter,M., Basler,K. (2007) The Decapentaplegic morphogen gradient: from pattern formation to growth regulation. *Nat Rev Genet.* **8**: 663-674

Alon,U. (2007) *An Introduction to Systems Biology: Design Principles of Biological Circuits.* Chapman & Hall/CRC Mathematical & Computational Biology, London.

Arita,M., Suwa,K. (2008) Search extension transforms Wiki into a relational system: a case for flavonoid metabolite database. *BioData Min*. **1**: 7.

Arita,M. (2009) A pitfall of wiki solution for biological databases. *Brief Bioinform.* **10:** 295-286.

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* **25**: 25-29.

Bailey,J.E. (1991) Toward a science of metabolic engineering. *Science* **252**: 1668-1675.

Barbe,V., Cruveiller,S., Kunst,F., Lenoble,P., Meurice,G., Sekowska,A. *et al*. (2009) From a consortium sequence to a unified sequence: The *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* **155**: 1758-1775.

Battle,A., Jonikas,M.C., Walter,P., Weissman,J.S., Koller,D. (2010) Automated identification of pathways from quantitative genetic interaction data. *Mol Syst Biol.* **6**: 379.

Bauer-Mehren,A., Furlong,L.I., Sanz,F. (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol*. **5**: 290.

Beasley,J.E., Planes,F.J. (2007) Recovering metabolic pathways via optimization. *Bioinformatics* **23**: 92-98.

Bedau,M.A., Humphreys,P. (2008) *Emergence: Contemporary readings in philosophy and science.* A Bradford book, The MIT Press, Cambridge, Massachusetts.

Belitsky,B.R., Sonenshein,A.L. (1998) Role and regulation of *Bacillus subtilis* glutamate dehydrogenase genes. *J Bacteriol* **180**: 6298-6305.

Belitsky,B.R., Sonenshein,A.L. (2004) Modulation of activity of *Bacillus subtilis* regulatory proteins GltC and TnrA by glutamate dehydrogenase. *J Bacteriol* **186**: 3399-3407.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.* **38**: D46-D51.

Berman,H., Henrick,K., Nakamura,H., Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35**: D301-D303.

Blencke,H.-M., Homuth,G., Ludwig,H., Mäder,U., Hecker,M., Stülke,J. *et al*. (2003) Transcriptional profiling of gene expression in response to glucose in *Bacillus subtilis*: regulation of the central metabolic pathways. *Metab. Engn.* **5**: 133-149.

Brohée,S., Barriot,R., Moreau,Y., André,B. (2010a) YTPdb: A wiki database of yeast membrane transporters. *Biochim Biophys Acta*. **1798**: 1908-1912.

Brohée,S., Barriot,R., Moreau,Y. (2010b) Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases. *Bioinformatics* **26**: 2210-2211.

Bruggeman,F.J., Westerhoff,H.V. (2007) The nature of systems biology. *Trends Microbiol.* **15**: 45-50.

Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. **38**: D473-D479.

Chin,G.Jr., Chavarria,D.G., Nakamura,G.C., Sofia,H.J. (2008) BioGraphE: high-performance bionetwork analysis using the Biological Graph Environment. *BMC Bioinformatics* **9**: S6.

Claessen,K., Een,N., Sheeran,M., Sörensson,N., Voronov,A., Åkesson,K. (2009) SAT-solving in practice, with a tutorial example from supervisory control. *Discrete Event Dyn Syst* **19**: 495-524.

Clomburg,J.M., Gonzalez,R. (2010) Biofuel production in *Escherichia coli*: the role of metabolic engineering and synthetic biology. *Appl Microbiol Biotechnol.* **86**: 419-434.

Commichau,F.M., Forchhammer,K., Stülke,J. (2006) Regulatory links between carbon and nitrogen metabolism. *Curr Opin Microbiol* **9**: 167-172.

Commichau,F.M., Herzberg,C., Tripal,P., Valerius,O., Stülke,J. (2007a) A regulatory protein-protein interaction governs glutamate biosynthesis in *Bacillus subtilis*: the glutamate dehydrogenase RocG moonlights in controlling the transcription factor GltC. *Mol Microbiol* **65**: 642-654.

Commichau,F.M., Wacker,I., Schleider,J., Blencke,H.M., Reif,I., Tripal,P., Stülke,J. (2007b) Characterization of *Bacillus subtilis* mutants with carbon source-independent glutamate biosynthesis. *J Mol Microbiol Biotechnol* **12**: 106-113.

Commichau,F.M., Gunka,K., Landmann,J.J., Stülke,J. (2008) Glutamate metabolism in *Bacillus subtilis*: gene expression and enzyme activities evolved to avoid futile cycles and to allow rapid responses to perturbations of the system. *J Bacteriol* **190**: 3557-3564.

Commichau,F.M., Rothe,F.M., Herzberg,C., Wagner,E., Hellwig,D., Lehnik-Habrink,M. *et al.* (2009) Novel activities of glycolytic enzymes in *Bacillus subtilis*: Interactions with essential proteins involved in mRNA processing. *Mol. Cell. Proteomics* **8**: 1350-1360.

Cooper,S., Khatib,F., Treuille,A., Barbero,J., Lee,J., Beenen,M. *et al*. (2010) Predicting protein structures with a multiplayer online game. *Nature* **466**: 756-760.

Covert,M.W., Schilling,C.H., Palsson,B.Ø. (2001) Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* **7**: 73-88.

Covert,M.W., Xiao,N., Chen,T.J., Karr,J.R. (2008) Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24**: 2044-2050.

de Figueiredo,L.F., Podhorski,A., Rubio,A., Kaleta,C., Beasley,J.E., Schuster,S., Planes,F.J. (2009) Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **25**: 3158-3165.

de Jong,H., Page,M. (2008) Search for steady states of piecewise-linear differential equation models of genetic regulatory networks. *IEEE/ACM Trans Comput Biol Bioinform* **5**: 208-222.

Durot,M., Bourguignon,P.Y., Schachter,V. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev*. **33**: 164-190.

Eymann,C., Becher,D., Bernhardt,J., Gronau,K., Klutzny,A., Hecker,M. (2007) Dynamics of protein phosphorylation on Ser/ Thr/ Tyr in *Bacillus subtilis*. *Proteomics* **7**: 3509-3526.

Fabret,C., Feher,V.A., Hoch,J.A. (1999) Two-component signal transduction in *Bacillus subtilis*: How one organism sees its world. *J. Bacteriol*. **181**: 1975-1983.

Feavers,I.M., Price,V., Moir,A. (1998) The regulation of the fumarase (*citG*) gene of *Bacillus subtilis* 168. *Mol Gen Genet* **211**: 465-471.

Feist,A.M., Palsson,B.Ø. (2008) The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol.* **26**: 659-667.

Fisher,S.H., Wray,L.V. Jr. (2002) *Bacillus subtilis* 168 contains two differentially regulated genes encoding L-asparaginase. *J Bacteriol.* **184**: 2148-2154.

Flórez,L.A., Roppel,S.F., Schmeisky,A.G., Lammers,C.R., Stülke,J. (2009) A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki *Subti*Wiki. *Database (Oxford)* **2009**: bap012.

Francke,C., Siezen,R.J., Teusink,B. (2005) Reconstructing the metabolic network of a bacterium from its genome. *Trends Microbiol.* **13**: 550-558.

Fränzle,M., Herde,C., Teige,T. (2007) Efficient solving of large non-linear arithmetic constraint systems with complex Boolean structure. *Journal on Satisfiability* **1**: 209-236.

Funahashi,A., Matsuoka,Y., Jouraku,A., Morohashi,M., Kikuchi,N., Kitano,H. (2008) CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE* **96**: 1254-1265.

Gehlenborg,N., O'Donoghue,S.I., Baliga,N.S., Goesmann,A., Hibbs,M.A., Kitano,H. *et al.* (2010) Visualization of omics data for systems biology. *Nat Methods* **7**: S56-S68.

Gianchandani,E.P., Papin,J.A., Price,N.D., Joyce,A.R., Palsson,B.Ø. (2006) Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comput Biol* **2**: e101.

Goelzer,A., Bekkal Brikci,F., Martin-Verstraete,I., Noirot,P., Bessières,P., Aymerich,S., Fromion,V. (2008) Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst Biol* **2**: 20.

Goll,J., Rajagopala,S.V., Shiau,S.C., Wu,H., Lamb,B.T., Uetz,P. (2008) MPIDB: the microbial protein interaction database. *Bioinformatics.* **24**: 1743-1744.

Graça,A., Marques-Silva,J., Lynce,I., Oliveira,A.L. (2007) Efficient haplotype inference with pseudo-Boolean optimization. In *Algebraic Biology* Springer Verlag Berlin/Heidelberg, pp 125-139

Graumann,P. (2007) Bacillus*: cellular and molecular biology*. Caister Academic Press, Wymondham, Norfolk, pp xiii-xv

Grote,A., Klein,J., Retter,I., Haddad,I., Behling,S., Bunk,B. *et al.* (2009) PRODORIC (release 2009): a database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res* **37**: D61-D65.

Guérout-Fleury,A.M., Shazand,K., Frandsen,N., Stragier,P. (1995) Antibiotic-resistance cassettes for *Bacillus subtilis*. *Gene* **167**: 335-336.

Haanstra,J.R., van Tuijl,A., Kessler,P., Reijnders,W., Michels,P.A., Westerhoff,H.V. *et al.* **(**2008**)** Compartmentation prevents a lethal turbo-explosion of glycolysis in trypanosomes. *Proc Natl Acad Sci U S A.* **105**: 17718-177123.

Hahne,H., Wolff,S., Hecker,M., Becher,D. (2008) From complementarity to comprehensiveness – targeting the membrane proteome of growing *Bacillus subtilis* by divergent approaches. *Proteomics* **8**: 4123-4136.

Hand,E. (2010) Citizen science: People power. *Nature* **466**: 685-687.

Henry,C.S., Zinner,J.F., Cohoon,M.P., Stevens,R.L. (2009) iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biol* **10**: R69.

Hodis,E., Prilusky,J., Martz,E., Silman,I., Moult,J., Sussman,J.L. (2008) *Proteopedia* – a scientific “wiki” bridging the rift between three-dimensional structure and function of biomacromolecules. *Genome Biol* **9**: R121.

Hoffmann,R. (2008) A wiki for the life sciences where authorship matters. *Nat Genet* **40**: 1047-1051.

Höper,D., Völker,U., Hecker,M. (2005) Comprehensive characterization of the contribution of individual SigB-dependent general stress genes to stress resistance of *Bacillus subtilis*. *J. Bacteriol*. **187**: 2810-2826.

Hu,J.C., Aramayo,R., Bolser,D., Conway,T., Elsik,C.G., Gribskov,M. *et al*. (2008) The emerging world of wikis. *Science* **320**: 1289-1290.

Hunt,A., Rawlins,J.P., Thomaides,H.B., Errington,J. (2006) Functional analysis of 11 putative essential genes in *Bacillus subtilis*. *Microbiology* **152**: 2895-2907.

Huss,J.W., Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., *et al.* (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.* **38**:D633-D639.

Ishii,N., Robert,M., Nakayama,Y., Kanai,A., Tomita,M. (2004) Toward large-scale modeling of the microbial cell for computer simulation. *J Biotechnol.* **113**: 281-294.

Jin,S., Sonenshein,A.L. (1994) Transcriptional regulation of *Bacillus subtilis* citrate synthase genes. *J Bacteriol* **176**: 4680-4690.

Kahlem,P., Birney,E. (2006) Dry work in a wet world: computation in systems biology. *Mol Syst Biol*. **2**: 40.

Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**: D480-D484.

Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M., Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38**: D355-D360.

Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahrén,D. *et al*. (2005). Expansion of the BioCyc collection of pathway/ genome databases to 160. *Nucleic Acids Res* **33**: 6083-6089.

Katz,P.S., Calin-Jageman,R., Dhawan,A., Frederick,C., Guo,S., Dissanayaka,R. *et al*. (2010) NeuronBank: A Tool for Cataloging Neuronal Circuitry. *Front Syst Neurosci*. **19**: 4-9.

Kelder,T., Pico,A.R., Hanspers,K., van Iersel,M.P., Evelo,C., Conklin,B.R. (2009) Mining biological pathways using WikiPathways web services. *PLoS One* **4**: e6447.

Keseler,I.M., Bonavides-Martínez,C., Collado-Vides,J., Gama-Castro,S., Gunsalus,R.P., Johnson,D.A. *et al*. (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res*. **37**: D464-D470.

Khalil,A.S., Collins,J.J. (2010) Synthetic biology: applications come of age. *Nat Rev Genet.* **11**: 367-379.

Kim,H.U., Kim,T.Y., Lee,SY. (2008) Metabolic flux analysis and metabolic engineering of microorganisms. *Mol Biosyst.* **4**: 113-120.

Kitano,H. (2002) Systems biology: a brief overview. *Science* **295**: 1662-1664.

Kitano,H., Funahashi,A., Matsuoka,Y., Oda,K. (2005) Using process diagrams for the graphical representation of biological networks. *Nat Biotechnol* **23**: 961-966.

Klamt,S., Stelling,J. (2002) Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep.* **29**: 233-236.

Klipp,E., Herwig,R., Kowald,A., Wierling,C., Lehrach,H. (2005) *Systems Biology in Practice.* WILEY-VCH Verlay GmbH & Co. KGaA, Weinheim.

Kobayashi,K., Ehrlich,S.D., Albertini,A., Amati,G., Andersen,K.K., Arnaud,M. *et al*. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* **100**: 4678-4683.

Kono,N., Arakawa,K., Ogawa,R., Kido,N., Oshita,K., Ikegami,K. *et al*. (2009) Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One* **4**: e7710.

Kotte,O., Zaugg,J.B., Heinemann,M. (2010) Bacterial adaptation through distributed sensing of metabolic fluxes. *Mol Syst Biol* **6**: 355.

Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A. *et al*. (2007) EMBL Nucleotide database in 2006. *Nucleic Acids Res* **35**: D16-D20.

Kunst,F., Rapoport,G. (1995) Salt stress is an environmental signal affecting degradative enzyme. *J Bacteriol.* **177**: 2403-2407.

Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V. *et al*. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**: 249-256.

Lammers,C.R., Flórez,L.A., Schmeisky,A.G., Roppel,S.F., Mäder,U., Hamoen,L., Stülke,J. (2010) Connecting parts with processes: *Subti*Wiki and *Subti*Pathways integrate gene and pathway annotation for *Bacillus subtilis*. *Microbiology* **156**: 849-859.

Le Novère,N. (2006) Model storage, exchange and integration. *BMC Neurosci.* **7**: S11.

Le Novère,N., Hucka,M., Mi,H., Moodie,S., Schreiber,F., Sorokin,A. *et al.* (2009) The systems biology graphical notation. *Nat. Biotechnol.* **27**: 735-741.

Lechat,P., Hummel,L., Rousseau,S., Moszer,I. (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucleic Acids Res.* **36**: D469-D474.

Legeai,F., Shigenobu,S., Gauthier,J.P., Colbourne,J., Rispe,C., Collin,O. *et al*. (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol Biol.* **19**: 5-12.

Lehninger, A.L., Nelson,D.L., Cox,M.M. (2006) *Lehninger: Principles of Biochemistry.*W.H. Freeman, N.Y.

Lévine,A., Vannier,F., Absalon,C., Kuhn,L., Jackson,P., Scrivener,E. *et al.* (2006) Analysis of the dynamic *Bacillus subtilis* Ser/ Thr/ Tyr phosphoproteome implicated in a wide variety of cellular processes. *Proteomics* **6**: 2157-2173.

Lewis,N.E., Hixson,K.K., Conrad,T.M., Lerman,J.A., Charusanti,P., Polpitiya,A.D. *et al.* (2010) Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol.* **6**: 390.

Li,C., Donizelli,M., Rodriguez,N., Dharuri,H., Endler,L., Chelliah,V. *et al.* (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol*. **4**: 92.

Licht,A., Preis,S. Brantl,S. (2005) Implication of CcpN in the regulation of a novel untranslated RNA (SR1) in *Bacillus subtilis*. *Mol. Microbiol* **58**: 189-206.

Llaneras,F., Picó,J. (2010) Which metabolic pathways generate and characterize the flux space? A comparison among elementary modes, extreme pathways and minimal generators. *J Biomed Biotechnol*. **2010**: 753904.

Lou,C., Liu,X., Ni,M., Huang,Y., Huang,Q., Huang,L. *et al.* (2010) Synthesizing a novel genetic sequential logic circuit: a push-on push-off switch. *Mol Syst Biol* **6**: 350.

Ludwig,H., Homuth,G., Schmalisch,M., Dyka,F.M., Hecker,M., Stülke,J. (2001) Transcription of glycolytic genes and operons in *Bacillus subtilis*: Evidence for the presence of multiple levels of control of the *gapA* operon. *Mol Microbiol* **41**: 409-422.

Macek,B., Mijakovic,I., Olsen,J.V., Gnad,F., Kumar,C., Jensen,P.R., Mann,M. (2007) The serine/ threonine/ tyrosine phosphoproteome of the model bacterium *Bacillus subtilis. Mol. Cell. Proteomics* **6**: 697-707.

Mäder,U., Homuth,G., Scharf,C., Büttner,K., Bode,R., Hecker,M. (2002) Transcriptome and proteome analysis of *Bacillus subtilis* gene expression modulated by amino acid availability. *J. Bacteriol*. **184**: 4288-4295.

Maeda, J. (2006) *Simplicity: Design, Technology, Business, Life.* MIT Press, Cambridge, MA

Malaisse,W.J., Zhang,Y., Sener,A. (2004) Enzyme-to-enzyme channeling in the early steps of glycolysis in rat pancreatic islets. *Endocrine.* **24**: 105-109.

Marcus,F. (2008) *Bioinformatics and Systems Biology: Collaborative Research and Resources.* Springer-Verlag Berlin Heidelberg.

Martin-Verstraete,I., Débarbouillé,M., Klier,A., Rapoport,G. (1994) Interaction of wild-type truncated LevR of *Bacillus subtilis* with the upstream activating sequence of the levanase operon. *J Mol Biol* **241**: 178-192.

Matsuoka,Y., Ghosh,S., Kikuchi,N., Kitano,H. (2010) Payao: a community platform for SBML pathway model curation. *Bioinformatics*. **26**: 1381-1383.

Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B. *et al*. (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**: D619-D622.

McCandless, D. (2009) *Information is Beautiful: The Information Atlas*. Collins, San Francisco, CA

Meile,J.C., Wu,L.J., Ehrlich,S.D., Errington,J., Noirot,P. (2006) Systematic localisation of proteins fused to the green fluorescent protein in *Bacillus subtilis*: identification of new proteins at the DNA replication factory. *Proteomics* **6**: 2135-2146.

Mi,H., Dong,Q., Muruganujan,A., Gaudet,P., Lewis,S., Thomas,P.D. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res*. **38**: D204-D210.

Michels,P.A., Bringaud,F., Herman,M., Hannaert,V. (2006) Metabolic functions of glycosomes in trypanosomatids. *Biochim Biophys Acta*. **1763**: 1463-1477.

Molle,V., Nakaura,Y., Shivers,R.P., Yamaguchi,H., Losick,R., Fujita,Y., Sonenshein,A.L. (2003) Additional targets of the *Bacillus subtilis* global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis. *J. Bacteriol*. **185**: 1911-1922.

Mons,B., Ashburner,M., Chichester,C., van Mulligen,E., Weeber,M., den Dunnen,J. *et al.* (2008) Calling on a million minds for community annotation in WikiProteins. *Genome Biol.* **9**: R89.

Morinaga,T., Ashida,H., Yoshida,K. (2010) Identification of two scyllo-inositol dehydrogenases in *Bacillus subtilis*. *Microbiology* **156**: 1538-1546.

Morris,J.H., Meng,E.C., Ferrin,T.E. (2010) Computational tools for the interactive exploration of proteomic and structural data. *Mol Cell Proteomics* **9**: 1703-1715.

Moszer,I., Glaser,P., Danchin,A. (1995) SubtiList: a relational database for the *Bacillus subtilis* genome. *Microbiology* **141**: 261-268.

Moszer,I. (1998) The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett.* **430**: 28-36.

Moszer,I., Jones,L.M., Moreira,S., Fabry C, Danchin A. (2002) SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res*. **30**: 62-65.

Nakahigashi,K., Toya,Y., Ishii,N., Soga,T., Hasegawa,M., Watanabe,H. *et al.* (2009) Systematic phenome analysis of *Escherichia coli* multiple-knockout mutants reveals hidden reactions in central carbon metabolism. *Mol Syst Biol.* **5**: 306.

Nielsen,C.B., Cantor,M., Dubchak,I., Gordon,D., Wang,T. (2010) Visualizing genomes: techniques and challenges. *Nat Methods* **7**: S5-S15.

Oberhardt,M.A., Palsson,B.Ø., Papin,J.A. (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Biol.* **5**: 320.

O'Donoghue,S.I., Gavin,A.C., Gehlenborg,N., Goodsell,D.S., Hériché,J.K., Nielsen,C.B. *et al*. (2010a) Visualizing biological data-now and in the future. *Nat Methods* **7**: S2-S4.

O'Donoghue,S.I., Goodsell,D.S., Frangakis,A.S., Jossinet,F., Laskowski,R.A., Nilges,M. *et al*. (2010b) Visualization of macromolecular structures. *Nat Methods* **7**: S42-S55.

Oh,Y.K., Palsson,B.Ø., Park,S.M., Schilling,C.H., Mahadevan,R. (2007) Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* **282**: 28791-28799.

Olszewski,K.L., Mather,M.W., Morrisey,J.M., Garcia,B.A., Vaidya,A.B., Rabinowitz,J.D., Llinás,M. (2010) Branched tricarboxylic acid metabolism in *Plasmodium falciparum*. *Nature* **466**: 774-778.

Orth,J.D., Thiele,I., Palsson,B.Ø. (2010) What is flux balance analysis? *Nat Biotechnol* **28**: 245-248.

Palsson,B.O. (2006) *Systems Biology: Properties of Reconstructed Networks.* Cambridge University Press, Cambridge.

Persson,L.O., Johansson,G. (1989) Studies of protein-protein interaction using countercurrent distribution in aqueous two-phase systems: partition behaviour of five glycolytic enzymes from crude baker's yeast extract. *Arch. Biochem. Biophys.* **276**: 227–231.

Petersohn,A., Brigulla,M., Haas,S., Hoheisel,J.D., Völker,U., Hecker,M. (2001) Global analysis of the general stress response of *Bacillus subtilis*. *J. Bacteriol.* **183**: 5617-5631.

Pfeiffer,T., Sánchez-Valdenebro,I., Nuño,J.C., Montero,F., Schuster,S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics* **15**: 251-257.

Pico,A.R., Kelder,T., van Iersel,M.P., Hanspers,K., Conklin,B.R., Evelo,C. (2008) WikiPathways: pathway editing for the people. *PLoS Biol.* **6**: e184.

Piran,R., Halperin,E., Guttmann-Raviv,N., Keinan,E., Reshef,R. (2009) Algorithm of myogenic differentiation in higher-order organisms. *Development* **136**: 3831-3840.

Planes,F.J., Beasley,J.E. (2008) A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief Bioinform* **9**: 422-436.

Povolotskaya,I.S., Kondrashov,F.A. (2010) Sequence space and the ongoing expansion of the protein universe. *Nature* **465**: 922-926.

Procter,J.B., Thompson,J., Letunic,I., Creevey,C., Jossinet,F., Barton,G.J. (2010) Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods* **7**: S16-S25.

Purnick,P.E., Weiss,R. (2009) The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol.* **10**: 410-422.

Quentin,Y., Fichant,G., Denizot,F. (1999) Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *J. Mol. Biol*. **287**: 467-484.

Rasmussen,S., Nielsen,H.B., Jarmer,H. (2009) The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol Microbiol.* **73**: 1043-1057.

Reizer,J., Bachem,S., Reizer,A., Arnaud,M., Saier,M.H.Jr., Stülke,J. (1999) Novel phosphotransferase system genes revealed by genome analysis – the complete complement of PTS proteins encoded within the genome of *Bacillus subtilis*. *Microbiology* **145**: 3419-3429.

Rietkötter,E., Hoyer,D., Mascher,T. (2008) Bacitracin sensing in *Bacillus subtilis*. *Mol Microbiol* **68**: 768-785.

Rojas,I., Golebiewski,M., Kania,R., Krebs,O., Mir,S., Weidemann,A., Wittig,U. (2007) Storing and annotating of kinetic data. *In Silico Biol*. **7**: S37-S44.

Rosling,H.A. (2007) Visual technology unveils the beauty of statistics and swaps policy from dissemination to access. *Statistical Journal of the IAOS*. **24**: 103-104

Saito,S., Kakeshita,H., Nakamura,K. (2009) Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene* **428**: 2-8.

Salzberg,S.L. (2007) Genome re-annotation: a wiki solution? *Genome Biol*. **8**: 102.

Sambrook,J., Fritsch,E.F., Maniatis,T. (1989) *Molecular cloning: a laboratory manual*, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Saraiya,P., North,C., Duca,K. (2005a) Visualizing biological pathways: requirements analysis, systems evaluation and research agenda. *Inf. Vis.* **2005**: 1–15.

Saraiya,P., North,C., Duca,K. (2005b) An Insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Vis. Comput. Graph*. **11**: 443-456.

Schallmey,M., Singh,A., Ward,O.P. (2004) Developments in the use of *Bacillus* species for industrial production. *Can. J. Microbiol*. **50**: 1-17.

Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**: 2498-2504.

Shlomi,T., Eisenberg,Y., Sharan,R., Ruppin,E. (2007) A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol Syst Biol* **3**: 101.

Sierro,N., Makita,Y., de Hoon,M., Nakai,K. (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.* **36**: D93-D96.

Sonenshein,A.L., Hoch,J.A., Losick,R. (1993) *Bacillus subtilis and other Gram positive bacteria: Biochemistry, Physiology, and Molecular Genetics*. ASM press, Washington, D.C.

Sonenshein,A.L., Hoch,J.A., Losick,R. (2002) *Bacillus subtilis and its closest relatives: from genes to cells.* ASM press, Washington, D.C.

Stehr,H., Duarte,J.M., Lappe,M., Bhak,J., Bolser,D.M. (2010) PDBWiki: added value through community annotation of the Protein Data Bank. *Database (Oxford)* **2010**: baq009.

Stokes,T.H., Torrance,J.T., Li,H. and Wang,M.D. (2008) ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics* **9**: S18.

Stülke,J., Hillen,W. (2000) Regulation of carbon catabolism in *Bacillus* species. *Annu. Rev. Microbiol.* **54**: 849-880.

Sun,D.X., Setlow,P. (1991) Cloning, nucleotide sequence, and expression of the *Bacillus subtilis ans* operon, which codes for L-asparaginase and L-aspartase. *J Bacteriol* **173**: 3831-3845.

Sun,D., Setlow,P. (1993) Cloning and nucleotide sequence of the *Bacillus subtilis ansR* gene, which encodes a repressor for the *ans* operon coding for L-asparaginase and L-aspartase. *J Bacteriol* **175**: 2501-2506

Thiele,I., Palsson, B.Ø. (2010a) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* **5**, 93-121.

Thiele,I., Palsson,B.Ø. (2010b) Reconstruction annotation jamborees: a community approach to systems biology. *Mol Syst Biol.* **6**: 361.

Thomaides,H.B., Davison,E.J., Burston,L., Johnson,H., Brown,D.R., Hunt,A.C. *et al*. (2007) Essential bacterial functions encoded by gene pairs. *J. Bacteriol.* **189**: 591-602.

Tiwari,A., Talcott,C., Knapp,M., Lincoln,P., Laderoute,K. (2007) Analyzing Pathways Using SAT-Based Approaches. In *Algebraic Biology* Springer Verlag Berlin/Heidelberg, pp 155-169.

Trinh,C.T., Wlaschin,A., Srienc,F. (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. *Appl Microbiol Biotechnol.* **81**: 813-826.

Uniprot Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* **37**: D169-D174.

Veeramani,B., Bader,J.S. (2010) Predicting functional associations from metabolism using bi-partite network algorithms. *BMC Syst Biol.* **4**: 95.

Velev,M.N., Bryant,R.E. (2003) Effective use of Boolean satisfiability procedures in the formal verification of superscalar and VLIW microprocessors. *J Symbolic Comput* **35**: 73-106

Viswanathan,G.A., Nudelman,G., Patil,S., Sealfon,S.C. (2007) BioPP: a tool for web-publication of biological networks. *BMC Bioinformatics* **8**: 168.

Voigt,B., Antelmann,H., Albrecht,D., Ehrenreich,A., Maurer,K.H., Evers,S. *et al*. (2009) Cell physiology and protein secretion of *Bacillus licheniformis* compared to *Bacillus subtilis*. J. *Mol. Microbiol. Biotechnol.* **16**: 53-68.

von Ahn,L., Maurer,B., McMillen,C., Abraham,D., Blum,M. (2008) reCAPTCHA: human-based character recognition via Web security measures. *Science* **321**: 1465-1468.

von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A., Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A.* **100**: 15428-15433.

Wach,A. (1996) PCR-synthesis of marker cassettes with long flanking homology regions for gene disruptions in *Saccharomyces cerevisiae*. *Yeast* **12**: 259-265.

Wacker,I., Ludwig,H., Reif,I., Blencke,H.M., Detsch,C., Stülke,J. (2003) The regulatory link between carbon and nitrogen metabolism in *Bacillus subtilis*: regulation of the *gltAB* operon by the catabolite control protein CcpA. *Microbiology* **149**: 3001-3009.

Waldrop,M. (2008) Big data: Wikiomics. *Nature* **455**: 22-25.

Walter,T., Shattuck,D.W., Baldock,R., Bastin,M.E., Carpenter,A.E., Duce,S. *et al*. (2010) Visualization of image data from cells to organisms. *Nat Methods* **7**: S26-S41.

Welch, R., Welch, L. (2009) If you build it, they might come. *Nat. Rev. Microbiol.* **7**: 90.

Whatmore,A.M., Chudek,J.A., Reed,R.H. (1990) The effects of osmotic upshock on the intracellular solute pools of *Bacillus subtilis*. *J Gen Microbiol* **136**: 2527-2535.

Winnenburg,R., Wächter,T., Plake,C., Doms,A., Schroeder,M. (2008) Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform.* **9**: 466-478.

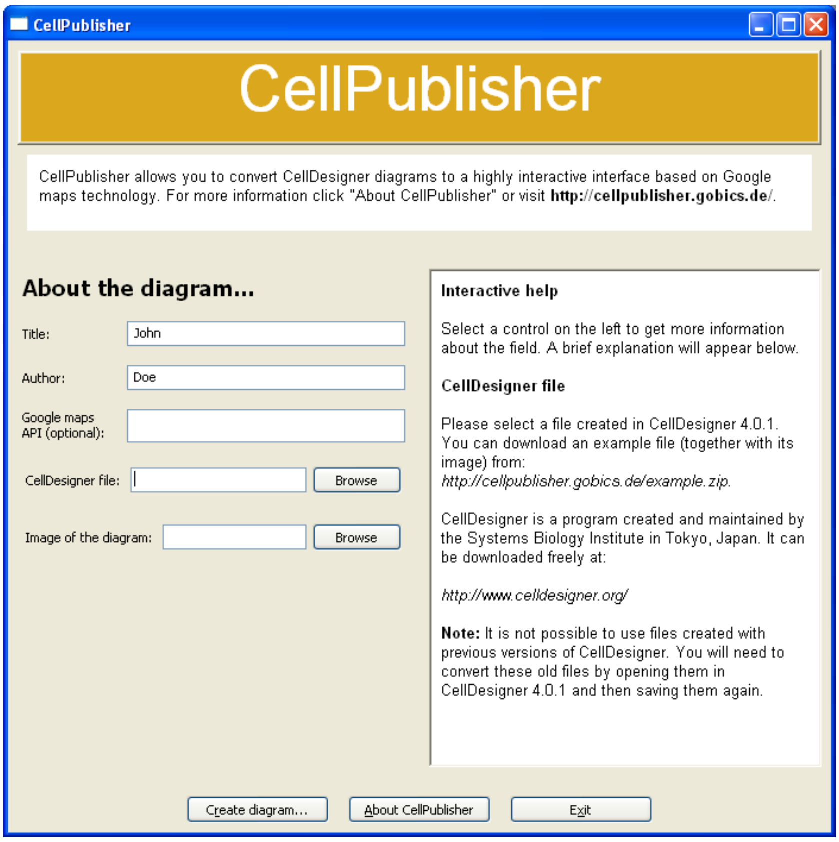
Yuan,J., Doucette,C.D., Fowler,W.U., Feng,X.J., Piazza,M., Rabitz,H.A. *et al*. (2009) Metabolomics-driven quantitative analysis of ammonia assimilation in *E. coli*. *Mol Syst Biol* **5**: 302.

Supplementary Material

The offline version of CellPublisher

Summary

The offline version of CellPublisher is a desktop application that can be installed on any computer running Windows. It is available for free at: http://sourceforge.net/projects/cellpublisher/. Once installed, there is no need to download extra software to make it work. The installed program requires the same input files as the web-based version: a CellDesigner file, and an image of it. These files are converted on the local computer and the results are stored in a local directory chosen by the user. These results can then be uploaded to any non-profit web server. Since all the computation and storage occur in the local computer, there is no data sent to the CellPublisher website. This can be important for confidential data.



CellPublisher allows you to convert CellDesigner diagrams to a highly interactive interface based on Google maps technology. For more information click "About CellPublisher" or visit **http://cellpublisher.gobics.de/**.

About the diagram...

Title: John

Author: Doe

Google maps API (optional):

CellDesigner file:

Image of the diagram:

Interactive help

Select a control on the left to get more information about the field. A brief explanation will appear below.

CellDesigner file

Please select a file created in CellDesigner 4.0.1. You can download an example file (together with its image) from: *http://cellpublisher.gobics.de/example.zip.*

CellDesigner is a program created and maintained by the Systems Biology Institute in Tokyo, Japan. It can be downloaded freely at:

http://www.celldesigner.org/

Note: It is not possible to use files created with previous versions of CellDesigner. You will need to convert these old files by opening them in CellDesigner 4.0.1 and then saving them again.

Screenshot: The offline version has the same input as the web server, except for the field for the Google maps API key. See text for a description.

Availability of the offline version

The installer for Microsoft Windows can be downloaded from http://sourceforge.net/projects/cellpublisher/. Once downloaded, the program is installed in the same way as any other software, with the possibility to create a shortcut on the desktop and the start menu.

The installer already includes all necessary libraries, so no extra software needs to be installed prior to use.

At the moment of this writing, there are no plans to create versions of the software for Mac, Linux, or other operating systems. Nonetheless, all the code is open-source and can be browsed and downloaded from Sourceforge.net.

Purpose of the offline version

CellPublisher aims to facilitate the navigation and sharing of CellDesigner diagrams. To achieve these goals, the online version requires the user to upload the files to the CellPublisher web server and returns a link that is open to everybody. In some cases, there might be confidentiality restrictions on the diagram that do not allow the submission of it via an open web server. Moreover, although the resulting links are codified and almost impossible to guess by outsiders, sometimes a more fine-grained control of access is desirable, for instance by requiring user registration.

The offline version offers more privacy and flexibility. First, the user does not interact at all with the CellPublisher web server. Second, the output of this version of CellPublisher can be modified to suit particular requirements. For instance, the look-and-feel of the webpage can be changed, or different images can be used for the different zoom levels. Some examples are shown on the User's Guide in the CellPublisher website.

Basic functionality

The main reason to use the offline version of CellPublisher is to host the pathway in another web server, like e.g. a department's home page or a website where registration is required. The files created with this version can be copied directly to the web server, without the need to install any additional framework (e.g. an Apache installation should be sufficient). The "User's guide" on the CellPublisher website shows step-by-step instructions on how to use the program.

The user interface of the offline version contains almost the same fields as the online version (see screenshot) with the exception of a text field for the Google maps API key. This key is a short text that indicates agreement with the conditions to use Google software on the web server. It can be obtained for free on a per-website basis on this link: http://code.google.com/apis/maps/signup.html

Further differences between versions

The diagrams resulting from the offline version do not include 3D structures from the Protein Data Bank. In addition, the viewers of the final diagram cannot add comments, as is the case in the online website.

Table S1: Primer sequences used in the experiments

Name	Sequence
KG12	5’ CCTATCACCTCAAATGGTTCGCTGGATGGCCAGCCGCTGAGTGAAG
KG13	5’ CCGAGCGCCTACGAGGAATTTGTATCGCCGAGAAGGTCAGCTGTATATTGAAGC
KG14	5’ ACCTCGTAAATGCTCATGTCTTCGCC
KG15	5’ CCGGAAGTCATTCTAGAGCTTGAGGA
KG18	5’AAAGGATCCCAGCTCAAGGTGAAAAAGGAGCGGAA
KG19	5’TTTGTCGACTCATTAACTCAGTTCCTCCTGTACTTTTCTTTTTGTG
KG25	5’ TTGAAGGGGAAAATGGGCTG
KG26	5’ CTATTTCCACCCAGTATTCAGG
KG28	5’ ATGGCTTGGACCCGTTATTGGGG
KG29	5’ CCTATCACCTCAAATGGTTCGCTGGAGCCAGCCCATTTTCCCCTTC
KG30	5’ CCGAGCGCCTACGAGGAATTTGTATCGCGGCGCTGATCATCTTGTTGATG
KG31	5’ AAGTCGGCACAACGCCTCCGG
KG38	5’ CCGTGTCGCATTAACACC
KG39	5’ ACCTGCTTCGGATCAGCA
KG40	5’ TAAACCTTGGCGGCGGAA
KG41	5‘ CCATATCCTCGACCGTTG
rpsJ-RT-fwd	5` GAAACGGCAAAACGTTCTGG
rpsJ-RT-rev	5` GTGTTGGGTTCACAATGTCG
rpsE-RT-fwd	5`GCGTCGTATTGACCCAAGC
rpsE-RT-rev	5` TACCAGTACCGAATCCTACG
mls-fwd (kan)	5`CAGCGAACCATTTGAGGTGATAGGGATCCTTTAACTCTGGCAACCCTC
mls-rev (kan)	5`CGATACAAATTCCTCGTAGGCGCTCGGGCCGACTGCGCAAAAGACATAATCG
Tc fwd1 (kan)	5`CAGCGAACCATTTGAGGTGATAGGGCTTATCAACGTAGTAAGCGTGG
Tc rev (kan)	5`CGATACAAATTCCTCGTAGGCGCTCGGGAACTCTCTCCCAAAGTTGATCCC

Curriculum vitae

Full name	Lope Andrés Flórez Weidinger
Date of Birth	29 May 1982
Place of Birth	Bogotá, Colombia
Nationalities	Colombian – German

Education

1986 – 2001	Colegio Andino – Deutsche Schule, Bogotá, Colombia
2001 – 2006	Bachelor in Biology with minor in Mathematics Universidad de los Andes, Bogotá Colombia Title of the Bachelor thesis: “Bioinfórmate: a virtual learning environment for bioinformatics” (http://bioinformate.uniandes.edu.co). Website in Spanish. Supervisors: Carlos Jaramillo, MSc., Silvia Restrepo, PhD, Rafael García, MSc
2006 – 2007	IMPRS MSc Program in Molecular Biology at the University of Göttingen, Germany Direct admission to the PhD program without Master thesis based on excellent academic results
2007 – 2010	IMPRS PhD Program in Molecular Biology at the University of Göttingen, Germany Title of the dissertation: “Systems biology in *Bacillus subtilis*: databases for gene function and software tools for pathway discovery” Supervisor: Prof. Dr. Jörg Stülke

Scholarships

2001	“Alberto Magno” stipend of the Universidad de los Andes, Colombia
2006-2007	Stipend of the International Max Planck Research School, Germany
2007-2008	“Georg-Lichtenberg” stipend of the State of Lower Saxony, Germany
2009-2010	Stipend of the Stiftung der Deutschen Wirtschaft, Germany