Genome-scale Reconstruction of Metabolic Network in Bacillus subtilis Based on High-throughput Phenotyping and Gene Essentiality Data*§

Received for publication, May 7, 2007. Published, JBC Papers in Press, June 15, 2007, DOI 10.1074/jbc.M703759200

You-Kwan Oh1†, Bernhard O. Palsson6, Sung M. Park7‡, Christophe H. Schilling4, and Radhakrishnan Mahadevan§3

From the 4Department of Bioengineering, University of California at San Diego, La Jolla, California 92093-0412 and 6Genomatica, Inc., San Diego, California 92121

In this report, a genome-scale reconstruction of Bacillus subtilis metabolism and its iterative development based on the combination of genomic, biochemical, and physiological information and high-throughput phenotyping experiments is presented. The initial reconstruction was converted into an in silico model and expanded in a four-step iterative fashion. First, network gap analysis was used to identify 48 missing reactions that are needed for growth but were not found in the genome annotation. Second, the computed growth rates under aerobic conditions were compared with high-throughput phenotypic screen data, and the initial in silico model could predict the outcomes qualitatively in 140 of 271 cases considered. Detailed analysis of the incorrect predictions resulted in the addition of 75 reactions to the initial reconstruction, and 200 of 271 cases were correctly computed. Third, in silico computations of the growth phenotypes of knock-out strains were found to be consistent with experimental observations in 720 of 766 cases evaluated. Fourth, the integrated analysis of the large-scale substrate utilization and gene essentiality data with the genome-scale metabolic model revealed the requirement of 80 specific enzymes (transport, 53; intracellular reactions, 27) that were not in the genome annotation. Subsequent sequence analysis resulted in the identification of genes that could be putatively assigned to 13 intracellular enzymes. The final reconstruction accounted for 844 open reading frames and consisted of 1020 metabolic reactions and 988 metabolites. Hence, the in silico model can be used to obtain experimentally verifiable hypothesis on the metabolic functions of various genes.

Bacillus subtilis has been the organism of choice for the production of several important industrial products, including antibiotics, enzymes, nucleosides, and vitamins. Several aspects of the biochemistry, genetics, and physiology of B. subtilis have been studied extensively making B. subtilis the best characterized prokaryote second only to Escherichia coli (1, 2). In addition various “omics” data sets such as transcriptomic (3), proteomic (4), and metabolomic (5) are available for B. subtilis. This wealth of experimental data enables the development of a genome-scale metabolic in silico model that can be used not only for quantitative interpretation and structured integration of such extensive data sets but also as a tool for hypothesis generation and engineering of B. subtilis metabolism.

To date, constraint-based reconstruction and analysis (COBRA)§ of cellular metabolism has been employed successfully to develop organism-specific genome-scale in silico models that have enabled numerous applications (6). Unlike other modeling strategies such as kinetic (7), stochastic (8), and cybernetic (9) methods, the COBRA approach does not attempt to compute precisely what a biochemical network does; rather, it seeks to distinguish between the network states that are achievable from those that are not, based on a detailed reconstruction of metabolism and incorporation of physiological parameters that are consistent with known experimental information. The COBRA approach is based on the successive imposition of governing physicochemical constraints on genome-scale reconstructions. The COBRA approach has been employed to generate genome-scale in silico models from each of the three major domains of the tree of life: archaea (i.e. Methanosarcina Barkeri (10)), bacteria (i.e. E. coli (11)), and eukarya (i.e. Saccharomyces cerevisiae (12)).

For B. subtilis, a stoichiometric model consisting of around 35 reactions in the central metabolism and the purine metabolism has been constructed previously (13, 14). The model has been used to obtain information on the intracellular flux distribution based on fractional carbon isotope labeling (15) and for the estimation of the P/O ratio (16) and the prediction of maximum theoretical yields of commercial biochemical products (14). Here we describe the reconstruction of a highly detailed and validated genome-scale metabolic network for B. subtilis, which reconciles established genomic and physiological data.

* This work was supported in part by the Department of Energy and by the Small Business Innovation Research grant program (DE-FG02-05ER84280), which enabled the high-throughput phenotyping integration studies and overall model development. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Some of the authors are stockholders of Genomatica, a company that develops and commercializes genome-scale metabolic models.

† Supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation. Present address: Bioenergy Research Center, Korea Institute of Energy Research, 71-2, Jang-dong, Yuseong-gu, Daejeon 305-343, Korea.

‡ Present address: Novozymes North America Inc., Franklington, NC 27525.

§ To whom correspondence should be addressed: Present address: Dept. of Chemical Engineering and Applied Chemistry, University of Toronto, Ontario M5S3E5, Canada. Tel.: 416-046-0096; Fax: 416-978-8605; E-mail: mahadevan@chem-eng.utoronto.ca.

The abbreviations used are: COBRA, constraint-based reconstruction and analysis; GPR, gene-protein-reaction association; ORF, open reading frame; KEGG, Kyoto Encyclopedia Genes and Genomes.
with high-throughput phenotyping data generated herein. The process used for the reconstruction illustrates the iterative nature of model development. Furthermore it demonstrates how the model can be used to highlight conflicts and discrepancies between various experimental data sets, and design experiments to resolve any gaps in our understanding of cellular metabolism, thereby accelerating biological discovery.

**EXPERIMENTAL PROCEDURES**

*Metabolic Network Reconstruction*—The reconstruction procedure followed the approach outlined in Edwards and Palsson (17). The reconstruction was carried out in SimPheny (Genomatica, Inc., San Diego, CA), a platform for development of cellular metabolic models. Lists of the genes, proteins, and reactions, corresponding literature information, definitions of the metabolite abbreviations, and a list of the exchange fluxes are described in the supplemental material (*Bacillus subtilis Model Excel file*). All the gene-protein-reaction (GPR) associations are provided in a GPR Associations zip file as JPEG images, which can be obtained from the authors, and a detailed document describing how to interpret these images is available elsewhere (11).

*Biomass Composition*—An equation for biomass formation was developed to account for the drain of precursors and building blocks into biomass. Biomass synthesis was incorporated as a linear combination of six macromolecular components (protein, DNA, RNA, lipid, lipoteichoic acid, and cell wall components) along with ions and metabolites, which were considered to account for the overall biomass composition. Lipoteichoic acid is a polymer of glycerolphosphate 24–33 units long, linked by phosphodiester bonds with a 3-gentibiosylglycerolide as the membrane anchor, and protrudes through the cell wall (18). The detailed calculation of the biomass composition is provided in the supplemental *Biomass Composition Word file*.

*In Silico Computations*—The metabolic capabilities of the *B. subtilis* network were calculated by using flux balance analysis and linear optimization (17). For growth simulations, biomass synthesis (see above) was selected as the objective function to be maximized, and optimization was solved using Linear Programming techniques in the SimPheny platform. In addition to the metabolic reactions, reversible exchange reactions for all external metabolites were also included in the simulations to allow external metabolites to cross the system boundary. All flux values during simulation are in millimoles/g cell/h. For simulation of aerobic growth on minimal media, the following flux values during simulation are in millimoles/g cell/h. For simulation of growth on complex Luria-Bertani (LB) medium, the chemical composition of LB medium was approximated based on typical analysis of the yeast extract and Tryptone provided by the manufacturers (see supplemental *Complex Medium Composition Word file*). The threshold of specific growth rate for cell viability was assumed to be 10⁻⁸ h⁻¹. All simulation conditions are detailed in the supplemental *Simulation Conditions Excel file*.

*Biolog Phenotyping Experiments*—*B. subtilis* 168 (ATCC 23857) was obtained from the American Type Culture Collection (ATCC). Biolog’s Phenotype MicroArray™ technology (19) was used for the phenotypic analysis of *B. subtilis*. It permits assays of 190 carbon (PM1- and PM2A-Microplates), 95 nitrogen (PM3B-Microplate), and 59 phosphorus and 35 sulfur source (PM4A-Microplate) utilizations at once. A defined medium containing 25 mm sodium pyruvate, 25 mm glucose, 5.0 mm NaCl, 2.0 mm Na₃PO₄, 0.25 mm Na₂SO₄, 100 mm NaCl, 30 mm triethanolamine HCl (pH 7.1), 0.05 mm MgCl₂, 1.0 mm KCl, 1.0 mm FeCl₃, and 0.01% tetraciazol violet was used for the Phenotype Microarrays™ tests. The (PM) plates contained various carbon, nitrogen, phosphorus, or sulfur sources, which are omitted from the defined medium. The microplates were incubated at 37 °C, and the absorbances were measured both at 570 and 750 nm after 24- and 48-h incubations, respectively, resulting in a total of 8 measurements. The detailed experimental conditions were described elsewhere (20). The colorimetric assay was considered as positive when the absorbance corresponding to the reduced dye (indicating substrate utilization) was 1.2 times higher than that of negative control. This threshold was chosen based on the calculation of standard deviations of the absorbance of the reduced dye in the negative control wells (S.D. was ~19%). The confidence levels for the data were determined based on the number of positive reactions out of the 8 measurements: high-confidence (growth, 8; no growth, none), medium-confidence (growth, 6 or 7; no growth, 1 or 2), and low-confidence (3 to 5). Low-confidence data were used for *in silico* analysis only if corresponding experimental data were available from the literature. Lists of the compounds, confidence level on data, predicted growth rates, and corresponding literature information are detailed in the supplemental Table S1.

**RESULTS**

*Metabolic Reconstruction*—The overall iterative model development procedure of genome-scale metabolic reconstruction for *B. subtilis* was based on established reconstruction methods augmented with high-throughput substrate utilization experiments and large-scale gene essentiality data sets (Fig. 1). First, a *B. subtilis* metabolic network (*modelv1* in Fig. 1) was reconstructed based on the annotated genome sequence (21) and available biochemical data (1, 2) containing 940 reactions. The metabolic network was expanded in size to 1015 reactions (*modelv2* in Fig. 1) based on high-throughput Biolog phenotyping data (see “Experimental Procedures”) and network-based gap analyses. *Metabolic modelv2* computed the qualitative outcome of substrate utilization data with 74% agreement (200 in 271 cases). Finally, *metabolic modelv2* was compared with a large-scale gene essentiality study (22), and this analysis led to the addition of five metabolic reactions to *modelv2*. This expanded model could...
The abundance of various classes of transporters in the ORFs are associated with the transport of metabolites indi-
ing and gene-essentiality analyses. For example, Biolog pheno-
tion, but 53 were based on high-throughput Biolog phenotyp-
subtilis}

The functional classification of the 844 ORFs included in the reconstruction are available as supplemental informa-
detailed in Fig. 2. The completed reconstruction (metabolic
that agree with the experimental data.

Characteristics of the Reconstructed Network—The charac-
teristics of the reconstructed B. subtilis metabolic network are
detailed in Fig. 2. The completed reconstruction (metabolic
model) accounted for 844 open reading frames (ORFs) and
consisted of 1020 reactions and 988 metabolites (Fig. 2A).
The details of the list of genes, reactions, and the GPR associations
in the reconstruction are available as supplemental informa-
tion. The functional classification of the 844 ORFs included in
the reconstruction is summarized in Fig. 2B. The majority of
the ORFs are associated with the transport of metabolites indi-
cating the abundance of various classes of transporters in B.
subtilis. 191 of 244 transporters included in the reconstruction
were identified based on genomic and biochemical informa-
tion, but 53 were based on high-throughput Biolog phenotyp-
ing and gene-essentiality analyses. For example, Biolog pheno-
typing indicated that D-malic acid, L-arabitol, glyoxylic acid, or

dulcitol could be utilized by B. subtilis as a carbon source, and
its corresponding transporter was incorporated into the bio-
chemical-reaction network (see supplemental Tables S2 and S3
for lists of transporters).

The basic capabilities of the in silico model to predict quan-
titatively aerobic growth on various carbon sources were deter-
mained (Fig. 2C). A growth demand function was formulated
based on biomass composition detailing the required metabo-
lites in the appropriate ratios (see supplemental Biomass Com-
position Word file). This demand function was used as the
to compute the experimental growth rates of B. subtilis. In B. subtilis,
four different terminal oxidases are known to exist, which
catalyze the transfer of electrons to O2 with different energy-coupling
efficiencies (16, 24) (Fig. 2D). Three oxidases accept electrons from

---

**Figure 1.** Reconstruction of genome-scale metabolic networks of B. subtilis and iterative model devel-

optment guided by high-throughput substrate utilization experiments and large-scale gene essentiality
data. The B. subtilis network was reconstructed based on genome sequence and available biochemical infor-
mation. Then, the network was expanded with high-throughput phenotyping data from Biolog and finally
verified against large-scale gene deletion studies. The reconstructed network may be used to update, validate,
and generate hypotheses about the phenotypes of B. subtilis. The accuracy of the model predictions that agree
with the experimental data.

---

**Genome-scale Metabolic Model of B. subtilis**

The demand function was used as the objective function for flux balance analysis analysis. The growth rates for
different substrate uptake rates were computed based on minimal
media (see “Experimental Procedures” for simulation conditions)
and compared with the experimental results of aerobic chemostat
cultivation (13, 23). Under these conditions, the model used the caa3
terminal oxidase activity as part of the optimal flux distribution and
resulted in higher in silico growth rates than experimental results (Fig.
2, C and D). However, when the flux through caa3 oxidase reaction was
restricted to zero under the same conditions to reflect the experimen-
tally observed repression of caa3 oxidase in the presence of glucose
(see below), the in silico model could correctly compute the exper-
imental growth rates of B. subtilis. In B. subtilis, four different terminal
oxidases are known to exist, which catalyze the transfer of electrons to
O2 with different energy-coupling efficiencies (16, 24) (Fig. 2D). Three
oxidases accept electrons from

---

**Analysis of High-throughput Substrate Utilization—The in silico computations and high-throughput phenotyping data
from Biolog’s Phenotyping MicroArray™ technology (Biolog
Hayward, Inc., CA (19)) were compared (Fig. 3). 271 of 379
substrates tested (190 for carbon, 95 for nitrogen, 59 for phos-
phorus, and 35 for sulfur sources) were identified as data with
sufficient confidence (see “Experimental Procedures” for deter-
mining the confidence) and were analyzed for consistency using
the B. subtilis model. In the case of nitrogen and sulfur
sources, obtaining high confidence data was difficult due to the
high background of negative control, and 44% (42 in 95) and
49% (17 in 35) of the compounds tested could not be used for in silico
analyses, respectively. A list of substrates, confidence lev-
els of data, model refinements, predicted growth rates, and literature information for corresponding substrates, are described in supplemental Table S1. Growth on substrates was simulated by fixing its specific uptake rate at 5 mmol/g cell/h under aerobic conditions based on minimal media (see “Experimental Procedures” for simulation conditions). However, it should be noticed that the assumption of utilization rate of substrate might affect growth prediction (see below). The network (metabolic model in Fig. 1), which was developed only based on genomic and biochemical information, could predict qualitatively the outcomes of Biolog data with 52% accuracy (140 in 271), but, after network expansion and metabolic gap analysis, overall prediction accuracy (metabolic model in Fig. 1) was considerably improved to 74% (200 in 271) (Fig. 3A).

However, 14 disagreements (7 for carbon, 4 for nitrogen, and 3 for sulfur sources) were observed (Fig. 3B), and out of these 14, 8 cases were compared with experimental growth data available in the literature by Medline search for the corresponding substrates (Fig. 3C).

Biolog phenotyping result and Mortl et al. (29) indicated together that glycine could not be utilized by B. subtilis as a sole carbon source, but the model showed growth, although the predicted growth was low at 0.036 h^{-1} for a specific uptake rate of 5 mmol/g cell/h. However, it should be noticed that for a lower specific uptake rate of 1 mmol/g cell/h no growth was predicted. Glycine is converted to glyoxylate, hydrogen peroxide, and ammonia by glycine oxidase (30). Hydrogen peroxide is toxic, and its production during glycine fermentations at 25 mM (Biolog phenotyping) and 20 mM (29) might also inhibit cell growth of B. subtilis. For other cases, relatively higher growth rates were predicted and accordingly presented in supplemental Table S1 without further computations. Direct comparison of present Biolog results and growth data from literature is difficult, because fermentation conditions, including medium composition, can affect significantly cell growth. In addition, in many reports the experimental conditions have been poorly defined. Nevertheless, for the remaining cases, in silico predictions generally agreed with experimental data from the literature but not with the phenotype micro-arrays. For example, B. subtilis does not have a glyoxylate shunt, a pathway that allows acetate to be used as a sole carbon source (31). Bryan et al. (32) reported that no growth was observed for B. subtilis with acetate as a sole carbon source. The metabolic network could produce NADH and ATP (namely aerobic respiration), but not biomass from acetate. In the high-throughput phenotype data, all of eight measurements for acetate indicated positive growth compared with that of the negative control (see supplemental Table S1). Biolog phenotyping system is based on colorimetric assay utilizing chemical reduction (purple color formation) of tetrazolium dye by NADH as a reporter system (19). However, NADH is an important electron carrier both in respiration and...
cell growth events, and cell respiration can occur independent of cell growth.

The growth predictions of the in silico model for the Biolog substrates evaluated as low-confidence data were also compared with the data available in the literature by Medline search for the corresponding substrates. For 18 out of the 108 low-confidence cases, corresponding experimental growth data were available from biochemical literature, and the model could correctly predict the utilization of these substrates in 94% accuracy (17 of 18 cases) (Fig. 3D). Incorrect prediction for D-serine was further analyzed. Fisher and Debarobuille (33) indicated no growth with D-serine as a sole nitrogen source, although the growth condition was not defined, the model simulations suggested the growth was possible. In the biochemical network, dsdA gene was assumed to encode D-serine deaminase (EC 4.3.1.18) based on a sequence similarity, which is necessary for utilizing D-serine as a sole nitrogen source (2). Therefore, the functionality of dsdA gene product or toxic effect of D-serine on cell growth (34) could be the subject of further experimental investigations.

The examples of glycine and serine utilization illustrate the capability of the COBRA approach to pinpoint potential gaps in the understanding of metabolism and to guide experimental design.

As a result of the high-throughput substrate utilization experiments, the B. subtilis network was expanded in size from 940 to 1015 unique reactions through the use of network gap analysis to identify additional reactions that would reconcile the model with the experimental data. Network gap analysis first finds metabolites that are either consumed only or produced only, and therefore, violate the material balance on the metabolites. Then the candidate reactions that could fill the gap are identified by analyzing the online pathway data base such as Kyoto Encyclopedia Genes and Genomes (KEGG) and/or biochemical literature, and metabolic pathway(s) on the basis of physiological information such as Biolog phenotyping, could be completed (6). This analysis resulted in the identification of 75 reactions in model v2 (4 for amino acids, 16 for carbohydrates, 1 for cell wall, 3 for nucleotides, 2 for phosphate, and 49 for transporters; see supplemental Table S2 for details). Note that these reactions do not have genetic or biochemical evidence but would be required to sustain growth on corresponding substrates. The protein sequences from other organisms where the protein associations with the reaction are characterized were used in BLASTP searches (39) of the B. subtilis genome SubtilList data base (40) to identify candidate ORFs that may have the corresponding functions. The conserved domain(s) and sequence of ORF product identified were also compared against source protein in an NCBI data base (www.ncbi.nlm.nih.gov) and using a ClustalW program (41) at European Bioinformatics Institute, respectively. These analyses...
TABLE 1
Hypothetical annotation based on high-throughput substrate utilization data, network gaps analysis, sequence similarity, and conserved domain and ClustalW analyses

| Carbon substrate (KEGG no.) | Gene<sup>a</sup> | Suggested annotation<sup>c</sup> | EC number | BLAST search<sup>d</sup> | Published annotation |
|-----------------------------|------------------|----------------------------------|------------|--------------------------|----------------------|
| D-Malate (C00497)           | yccA             | D-Malate dehydrogenase           | 1.1.1.83   | 2E-95                    | Pseudomonas putida (Q51945) |
| L-Arabinol (C00532)         | yggN             | Aldose reductase                 | 1.1.1.21   | 7E-49                    | Rattus norvegicus (AAA40721) |
| Glycolic acid (C00160) and glyoxylic acid (C00048) | ytbE | Tartronate semialdehyde reductase<sup>e</sup> | 1.1.1.21   | 6E-48                    | Plant metabolite dehydrogenase (21, 22) |
| Glycolic acid (C00160) and glyoxylic acid (C00048) | ykwC | Tartronate semialdehyde reductase<sup>e</sup> | 1.1.1.60   | 4E-49                    | Plant metabolite dehydrogenase (21, 22) |
| Glycolic acid (C00160) and glyoxylic acid (C00048) | yfg<sup>f</sup> | Tartronate semialdehyde reductase<sup>e</sup> | 4.1.1.47   | 6E-25                    | 3-Hydroxyisobutyrate dehydrogenase (21, 22) |
| Glycolic acid (C00160) and glyoxylic acid (C00048) | ydal<sup>f</sup> | Tartronate semialdehyde reductase<sup>e</sup> | 4.1.1.60   | 4E-50                    | Pyruvate oxidase (21, 22) and pyruvate dehydrogenase-related protein (42) |
| Dulcitol (C01697)           | yjmD             | Galactitol 1-phosphate 5-dehydrogenase | 1.1.1.251  | 3E-30                    | Sorbitol dehydrogenase (21) and alcohol dehydrogenase (22) |

<sup>a</sup> From the KEGG compound database.<br>
<sup>b</sup> Gene names are based on the Subtilist database.<br>
<sup>c</sup> Suggested annotation arose from analyzing the network gaps in modelv2 in Fig. 1.<br>
<sup>d</sup> E-values were obtained at the protein sequence level based on BLAST 2 alignment of sequences obtained from Subtilist and NCBI databases. The conserved domain(s) and sequence similarity were also compared against source protein in a NCBI database and using a ClustalW program at EBI, respectively (see supplemental Figs. S1 and S2).<br>
<sup>e</sup> YggN protein was expressed in E. coli and showed glyoxal reductase (EC 1.1.1.78) activities for glyoxal, methylglyoxal, glyoxylate, glyceraldehyde, and glyceraldehyde 3-phosphate (43). However, it should be noticed that aldose reductase generally have a broad substrate specificity, including arbutin, glyoxal, methylglyoxal, and glyceraldehyde (BRENDA database).<br>
<sup>f</sup> The tartronate semialdehyde reductases are a structurally and mechanistically related family of enzymes that includes 3-hydroxyisobutyrate dehydrogenase, 6-phosphogluconate dehydrogenase, and a novel D-phenylserine dehydrogenase (44).<br>

TABLE 2
Impact of single-gene deletions on growth in B. subtilis

For 766 of 844 metabolic genes included in the model, experimental gene deletion data were available from the Bacillus subtilis Genome data base (22), and the deletions of these genes were analyzed by simulating their growth on rich medium.

| Functional classification<sup>a</sup> | (+/+ in vivo) | (+/+ in silico) | (-/- in silico) | (-/-) | Accuracy<sup>b</sup> |
|--------------------------------------|--------------|----------------|-----------------|-------|----------------------|
| Amino acids and related molecules (129) | 118          | 9              | 0               | 2     | 98                   |
| Carbohydrates and related molecules (146) | 133          | 6              | 0               | 7     | 95                   |
| Cell wall (37)         | 13           | 14             | 9               | 1     | 73                   |
| Coenzymes and prosthetic group (67) | 53           | 12             | 0               | 2     | 97                   |
| Lipids (55)          | 30           | 16             | 8               | 1     | 84                   |
| Membrane bioenergetics (43) | 41           | 0              | 0               | 2     | 95                   |
| Nucleotides and nucleic acids (66) | 53           | 5              | 1               | 7     | 88                   |
| Other functions (13) | 12           | 0              | 0               | 1     | 92                   |
| Phosphate and sulfur (7) | 6            | 1              | 0               | 0     | 100                  |
| Transport (203)       | 198          | 0              | 0               | 5     | 98                   |
| Total (766)          | 657          | 63             | 18              | 28    | 94                   |

<sup>a</sup> See supplemental Table S4 for the complete list of genes analyzed.<br>
<sup>b</sup> See Fig. 3 for + and – descriptions.
These metabolic genes (pssA, psd, yfiX, ugtP, gtaB, dltABCD, ggaA, ggaB, and tagE) encode enzymes catalyzing the generation reaction of lipids and cell wall components, which are not essential for cell growth under normal conditions (22, 45–48). However, because the biomass composition defined in silico required the synthesis of these reactions, the corresponding genes were identified as lethal deletions. The remaining six cases (ndk, bkdAA, bkdAB, bkdB, ipdV, and yubB) appear to be due to metabolic gaps in the present biochemical network and/or the availability of external sources of the biomass precursors from rich media (22, 49). This could be subject of further experimental investigation. For the 28 false-positive disagreements (supplemental Table S6), 15 cases could be explained by accumulation of toxic compounds induced by gene deletions (tpiA and yueK), disruption of other functions such as pH homeostasis maintenance (mrxABCD), and formylation of methionyl-tRNA (fint), conditional essentiality (pgk), or regulation effects (racE, fbaA, murAA, yumC, and trxB) (22, 35–37), most of which are not yet predictable by the standard flux balance approach. metK, ytaG, and acpS products are related to syntheses of sioenzyme A, and acylcarrier protein, respectively, which might be necessary for cell viability. The remaining 11 false-positive cases were related to syntheses of siroheme, coenzyme A, and acyl-

The in silico metabolites of the metabolic network were also compared with experimental metabolomic data for B. subtilis grown in a moderately rich medium (5). 350 of 1692 compounds separated have been positively identified using standards (see supplemental Table 1 of Ref. 5) and were analyzed for consistency using the model. The biochemical network could only account for 160 (46%) of the 350 compounds identified (supplemental Table S7). This appears to be due to metabolic gaps in the present model. 270 (77%) of the 350 compounds could be referenced by the online KEGG compound database and/or Chemical Abstracts Service number, but the detailed metabolisms except for the in silico metabolites of the network have not been reported. The remaining 613 unique in silico metabolites could be the subject of further metabolomic investigations.

DISCUSSION

Herein, we have developed a genome-scale metabolic model of B. subtilis, a Gram-positive organism that is of both scientific and commercial importance (1, 2). The model-development process is iterative by nature and moves from the reconstruction of a genome-scale metabolic network to a fully functional in silico model that can be used to compute phenotypes. First, a metabolic reconstruction was developed that reconciles the genome sequence and annotation together with the known biochemistry and physiology of the organism. This reconstructed network was validated for consistency against newly generated high-throughput substrate phenotyping data and existing large-scale gene essentiality data. The model was subjected to iterative refinement based on the identified inconsistencies, leading to additional reactions incorporated in the network and other modifications to the model content. The net result was a biochemically and genetically detailed in silico model that consists of 1020 metabolic reactions that are catalyzed by 844 ORFs (~20% of the genes in B. subtilis). For B. subtilis, a stoichiometric model consisting of around 35 reactions in the central metabolism and the purine metabolism has been constructed previously (13, 14). However, with the genome-scale model, it was possible to compute not only quantitatively aerobic growth of B. subtilis, but also qualitatively its phenotypic behaviors for substrate utilization in 74% agreement (200 in 271 cases) and the outcome of single-gene deletions with 94% agreement (720 in 766 cases). Hence, the in silico model could represent a framework wherein a variety of heterogeneous data types are integrated and analyzed in a structured format (38) and be subject to further investigations (see below).

The computation of cellular functions from reconstructed networks allows for study of the relationship between environmental and genetic factors and their integrated function to orchestrate cellular phenotypes. The model development has also generated several experimentally verifiable hypotheses that could provide insight into the functioning of the large-scale metabolic network. An example of hypotheses generated in this model is the set of new annotation for metabolic genes based on network-based gap analysis and high-throughput phenotyping. Biolog phenotyping indicates that B. subtilis can consume various carbon substrates such as d-malic acid, l-arabitol, glycolic acid, glyoxylic acid, and dulcitol for cell growth. In Table 1, the reactions that are missing from these pathways and corresponding genes are identified based on network-gap and sequence analyses. Such reactions and genes should be validated with genetic and biochemical experiments to confirm their role in cellular metabolism. The second example is depicted in Table S5 in which the in silico deletions of the six metabolic genes (ndk, bkdAA, bkdAB, bkdB, ipdV, and yubB) are predicted to be essential; however, the growth of the corresponding mutants in vivo does not seem to be affected. This discrepancy indicates that there are possibilities of other genes in the genome that replace those specific reactions or the metabolites necessary for cell growth might be imported from the external source of rich media. Subsequently, we hope to see new experimental data generated that will support or refute these hypothesis, which will add further confidence to the network reconstruction or lead to content modifications. The model will continue to be iteratively refined further based on comparison with additional experimental data (Fig. 1). In this way the model can serve as an up-to-date representation of the cumulative knowledge of B. subtilis metabolic capabilities.

In addition to improving the model, this analysis presented
Genome-scale Metabolic Model of B. subtilis

in silico model could predict correctly the cellular respiration event in B. subtilis utilizing acetate as a sole carbon source. Biolog’s Phenotype MicroArray™ technology is based on NADH as a reporter system (19), but NADH is an important electron carrier both in respiration and cell growth events. The network could also predict the utilization in B. subtilis with an accuracy of 94% (17 of 18), for the substrates evaluated as low-confidence data in Biolog phenotyping. The ability to tightly integrate computational modeling approaches with high-throughput experimental data will be important in our ongoing quest to mathematically model and simulate complex biological functions, which represents a fundamental goal of systems biology.

Although there is significant progress on transcriptional and phenomic analyses with metabolic models, the integrated analysis of other high-throughput data types and/or simultaneous reconciliation of multiple data sets are in the early stages of development (6). In this study, the Biolog phenotyping dataset could be used effectively for the metabolic reconstruction of B. subtilis, and furthermore the latter could enhance the quality and utility of the former (see above). However, the in silico model could only predict 46% (160 in 350) of the experimental metabolomic compounds reported from the literature (5). At the present time, the annotation of the B. subtilis genome is incomplete, and ~40% of its ORFs do not have a functional assignment (2). Thus the genome-scale model developed here might lack some metabolic capabilities that B. subtilis possesses. Thus, the combined genome-wide experiments and comparisons with the model predictions will be an integral part in the further development of in silico B. subtilis strain and represents a key step in the application of systems biology to industrial biotechnology.

Acknowledgments—We thank I. Famili, J. Reed, A. Feist, R. Thakar, J. Trawick, I. Thiele, M. Ziman (Biolog Inc.), and J. Woodcock for technical assistance and valuable discussions.

REFERENCES

1. Sonenshein, A. L., Hoch, J. A., and Losick, R. (1993) Bacillus subtilis and Other Gram-positive Bacteria. Biochemistry, Physiology, and Molecular Genetics. American Society of Microbiology, Washington, D. C.
2. Sonenshein, A. L., Hoch, J. A., and Losick, R. (2002) Bacillus subtilis and Its Closest Relatives. From Genes to Cells, American Society of Microbiology, Washington, D. C.
3. Auger, S., Danchin, A., and Martin-Verastraete, I. (2002) J. Bacteriol. 184, 5179–5186
4. Eymann, C., Dreisbach, A., Albrecht, D., Bernhardt, J., Becher, D., Gentner, S., Tam, L. T., Buttner, K., Buurman, G., Scharf, C., Venz, S., Volker, U., and Hecker, M. (2004) Proteomics 4, 2849–2876
5. Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M., and Nishitani, T. (2003) J. Proteome Res. 2, 488–494
6. Reed, J. L., Famili, I., Thiele, I., and Palsson, B. O. (2006) Nat. Rev. Genetics 7, 130–141
7. Loew, L. M., and Schaff, J. C. (2001) Trends Biotechnol. 19, 401–406
8. Arkin, A. P., Ross, J., and McAdams, H. H. (1998) Genetics 149, 1633–1648
9. Varner, J., and Ramkrishna, D. (1999) Curr. Opin. Biotechnol. 10, 146–150
10. Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F. J., and Ideker, T. (2006) Mol. Syst. Biol. 2, 2006.0004
11. Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003) Genome Biol. 4, R54
12. Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004) Genome Res. 14, 1298–1309
13. Dauner, M., and Sauer, U. (2001) Biotechnol. Bioeng. 76, 132–143
14. Sauer, U., Cameron, D. C., and Bailey, I. E. (1998) Biotechnol. Bioeng. 59, 227–238
15. Sauer, U., Hatzimanikatis, V., Bailey, J. E., Hoch, J. A., and Losick, R., eds pp. 21–41, American Society of Microbiology, Washington, D. C.
16. Bochner, B. R., Gadzinski, P., and Panomrrotos, E. (2001) Genome Res. 11, 1246–1255
17. Zhou, L., Lei, X.-H., Bochner, B. R., and Wanner, B. L. (2003) J. Bacteriol. 185, 4956–4972
18. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Aloni, G., Azavedo, V., Bertero, M. G., Bessières, P., Bolotin, A., Borchert, S., Borris, R., Bourrier, L., Brans, A., Braun, M., Briggnell, S. C., Bron, S., Broillett, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codd, J. J., Connerton, I. F., Danchin, A., et al. (1997) Nature 390, 249–256
19. Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., Asai, K., Ashikaga, S., Aymeric, S., Bessieres, P., Boland, F., Bringnelli, S. C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L. C., Danchin, A., Debarbouille, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S. K., Dreesen, O., Errington, J., Fillingter, S., Foster, S. J., Fujita, Y., Galizzi, A., Gardan, R., Eschevin, C., Fukushima, T., Haga, K., Harwood, C. R., Hecker, M., Hosoya, D., Hullo, M. F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Inamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Mauel, C., Meima, R., Mellado, R. P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O’Reilly, M., O’Rourke, M., Pragai, Z., Pooley, H. M., Rapoport, G., Rawlings, J. P., Rivola, L. A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H. H., Scanlan, E., Schumann, W., Seegers, J. F., Sekiguchi, J., Sekowska, A., Seror, S. I., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H. B., Vagner, V., van Dijl, M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yoda, K., Yoshikawa, H., Zuber, U., and Ogasawara, N. (2003) Proc. Natl. Acad. Sci. U. S. A. 100, 4678–4683
20. Dauner, M., Sonderegger, M., Hochuli, M., Szyppers, T., Wuthrich, K., Hohmann, H.-P., Sauer, U., and Bailey, I. E. (2002) Appl. Environ. Microbiol. 68, 1760–1771
21. von Wachenfeldt, C., and Hederstedt, L. (2002) in Bacillus subtilis and Its Closest Relatives. From Genes to Cells (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 163–179, American Society of Microbiology, Washington, D.C.
22. Santana, M., Kunst, F., Hulio, M. F., Rapoport, G., Danchin, A., and Glaser, F. (1992) J. Bacteriol. 176, 6851–6862
23. Liu, X., and Taber, H. W. (1998) J. Bacteriol. 180, 6154–6173
24. Burns, K. E., Xiang, Y., Kinsland, C. L., McLafferty, F. W., and Begley, T. P. (2005) J. Am. Chem. Soc. 127, 3682–3683
25. Winstedt, L., Yoshida, K.-I., Fujita, Y., and von Wachenfeldt, C. (1998) J. Bacteriol. 180, 6571–6580
26. Mortl, M., Diederichs, K., Welte, W., Molla, G., Motteran, L., Andriolo, G., Pinole, M. S., and Pollegioni, L. (2004) J. Bacteriol. 186, 2971–2977
27. Molla, G., Motteran, L., Job, V., Pinole, M. S., and Pollegioni, L. (2003) Eur. J. Biochem. 270, 1474–1478
28. Sonenshein, A. L. (2002) in Bacillus subtilis and Its Closest Relatives. From Genes to Cells (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 151–162, American Society of Microbiology, Washington, D. C.
29. Bryan, E. M., Beall, B. W., and Moran, C. P., Jr. (1996) J. Bacteriol. 178, 4778–4786
33. Fisher, S. H., and Debarbouille, M. (2002) in *Bacillus subtilis and Its Closest Relatives. From Genes to Cells* (Sonenshein, A. L., Hoch, J. A., and Losick, R., eds) pp. 181–191, American Society of Microbiology, Washington, D. C.

34. Bloom, F. R., and Mcfall, E. (1975) *J. Bacteriol.* **121**, 1078–1084

35. Fraenkel, D. G. (1996) in *Escherichia coli and Salmonella. Cellular and Molecular Biology* (Neidhardt, F. C., ed) pp. 189–198, American Society of Microbiology, Washington, D. C.

36. Nakano, M. M., Zhu, Y., Haga, K., Yoshikawa, H., Sonenshein, A. L., and Zuber, P. (1999) *J. Bacteriol.* **181**, 7087–7097

37. You, C., Lu, H., Sekowska, A., Fang, G., Wang, Y., Gilles, A.-M., and Danchin, A. (2005) *BMC Microbiol.* **5**, 57

38. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004) *Nature* **429**, 92–96

39. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402

40. Moszer, I., Jones, L. M., Moreira, S., Fabry, C., and Danchin, A. (2002) *Nucleic Acids Res.* **30**, 62–65

41. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680

42. Petersohn, A., Brigulla, M., Haas, S., Hoheisel, J. D., Volker, U., and Becker, M. (2001) *J. Bacteriol.* **183**, 5617–5631

43. Sakai, A., Katayama, K., Katsuragi, T., and Tani, Y. (2001) *J. Biosci. Bioeng.* **91**, 147–152

44. Njau, R. K., Herndon, C. A., and Hawes, J. W. (2000) *J. Biol. Chem.* **275**, 38780–38786

45. Matsumoto, K., Okada, M., Horikoshi, Y., Matsuzaki, H., Kishi, T., Itaya, M., and Shibuya, I. (1998) *J. Bacteriol.* **180**, 100–106

46. Nishibori, A., Kusaka, J., Haru, H., Umeda, M., and Matsumoto, K. (2005) *J. Bacteriol.* **187**, 2163–2174

47. Perego, M., Glaser, P., Minutello, A., Strauch, M. A., Leopold, K., and Fischer, W. (1995) *J. Biol. Chem.* **270**, 15598–15606

48. Wecke, J., Perego, M., and Fischer, W. (1996) *Microb. Drug Resist.* **2**, 123–129

49. Willecke, K., and Pardee, A. B. (1971) *J. Biol. Chem.* **246**, 5264–5272

50. van der Ploeg, J. R., Cummings, N. J., Leisinger, T., and Connerton, I. F. (1998) *Microbiology* **144**, 2555–2561

51. Chaudhuri, A., Mishra, A. K., and Nanda, G. (1982) *Folia Microbiol.* **27**, 73–75

52. Yoshida, K.-I., Yamamoto, Y., Omae, K., Yamamoto, M., and Fujita, Y. (2002) *J. Bacteriol.* **184**, 983–991

53. Sekowska, A., and Danchin, A. (2002) *BMC Microbiol.* **2**, 8