HHS Public Access

Author manuscript

Nat Chem Biol. Author manuscript; available in PMC 2013 July 01.

Published in final edited form as:

Nat Chem Biol. 2012 October ; 8(10): 848–854. doi:10.1038/nchembio.1063.

Global probabilistic annotation of metabolic networks enables enzyme discovery

Germán Plata[1,2,*], **Tobias Fuhrer**[3,*], **Tzu-Lin Hsiao**[1,4,*], **Uwe Sauer**[3], and **Dennis Vitkup**[1,4,a]

[1]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY, U.S.A. [2]Integrated Program in Cellular, Molecular, Structural, and Genetic Studies, Columbia University, New York, NY, U.S.A. [3]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [4]Department of Biomedical Informatics, Columbia University, New York, NY, U.S.A.

Abstract

Annotation of organism-specific metabolic networks is one of the main challenges of systems biology. Importantly, due to inherent uncertainty of computational annotations, predictions of biochemical function need to be treated probabilistically. We present a global probabilistic approach to annotate genome-scale metabolic networks that integrates sequence homology and context-based correlations under a single principled framework. The developed method for Global Biochemical reconstruction Using Sampling (GLOBUS) not only provides annotation probabilities for each functional assignment, but also suggests likely alternative functions. GLOBUS is based on statistical Gibbs sampling of probable metabolic annotations and is able to make accurate functional assignments even in cases of remote sequence identity to known enzymes. We apply GLOBUS to genomes of *Bacillus subtilis* and *Staphylococcus aureus*, and validate the method predictions by experimentally demonstrating the 6-phosphogluconolactonase activity of *ykgB* and the role of the *sps* pathway for rhamnose biosynthesis in *B. subtilis*.

Introduction

Advances in DNA sequencing technologies and high-throughput experiments provide a unique opportunity to study cellular function at the systems level. The systems biology perspective seeks to understand how the interaction between multiple genomic components determines cellular physiology. Genome-scale metabolic networks serve as an important platform for such systems analyses and have been very successful in predicting various emergent properties of biological systems. They also have great potential for guiding metabolic engineering[1] and aiding drug target discovery[2]. Unfortunately, accurate manual

*equal contribution

[a]communication should be addressed to DV at dv2121@columbia.edu

Author contributions G.P., T-L.H and D.V. performed computational research and data analysis. T.F. performed experimental research and analysis. D.V. directed computational research. D.V. and U.S. directed experimental research. G.P., T.F., T-L.H and D.V. wrote the manuscript. All authors read and edited the manuscript.

Competing financial interests The authors declare no competing financial interests.

annotations of organism-specific metabolic networks are laborious and can take up to a year for a typical microbial genome. Efforts have been made to automate the reconstruction process, particularly the initial steps of genome annotation and network assembly[3-5].

The annotation process usually relies on sequence homology methods, in which the function of a metabolic gene is assigned based on sequence similarity to known enzymes[6]. Although homology methods have been successful overall, annotations established based solely on weak sequence identity are often unreliable due to frequent functional divergence between distant homologues. It was demonstrated that a sequence identity above 60% is usually required to accurately transfer a precise enzyme function, i.e. all four digits of an Enzyme Commission (EC) number[7]. Consequently, homology-based methods fail to assign functions to a substantial fraction of genes in completely sequenced genomes and have been known to produce multiple imprecise or incorrect annotations[8,9].

The metabolic network reconstruction for a given genome is usually performed based on a functional annotation of all metabolic genes. Functional databases such as BRENDA[10], GeneCards[11], KEGG[3], MetaCyc[12] or Swiss-Prot[13] are useful resources for establishing initial associations between metabolic genes and corresponding biochemical reactions. Draft metabolic models are typically reconstructed by assembling annotated biochemical reactions into a network. One disadvantage of this two-step approach is that genes are annotated individually rather than being considered together in a proper network context. Therefore, some successful computational approaches utilize pre-defined or manually curated metabolic pathways[5] and subsystems[14] to annotate network reactions. Naturally, the accuracy of such methods depends both on the quality of the initial annotation and the evolutionary conservation of reference pathways.

Context based methods such as phylogenetic profiles[15], protein fusions[16], gene co-expression[17], and chromosomal gene neighborhood[18] capture conserved functional relationships and often provide information complementary to sequence homology[19]. The effectiveness of these methods has been shown by determining members of protein complexes, functional modules, and molecular pathways[20,21]. Multiple studies have also demonstrated that context associations combined with local network structure can be used to identify genes responsible for orphan metabolic activities and to improve existing annotations of metabolic genes[22,23]. Therefore, it is natural to combine sequence homology and context functional descriptors using a unified probabilistic framework.

Although powerful probabilistic approaches, such as Bayesian and Boolean networks, have been applied to reconstruction of regulatory and signaling networks based on high-throughput data[24], global probabilistic methods to annotate metabolic networks have not been developed. Here, we present such a global probabilistic approach that integrates sequence homology and context associations to annotate genome-scale metabolic networks. The method for Global Biochemical reconstruction Using Sampling (GLOBUS) not only provides annotation probabilities for each gene and each metabolic activity, but also suggests possible alternative functions. We applied GLOBUS to the genomes of *Bacillus subtilis* and *Staphylococcus aureus*, evaluated the accuracy of the reconstructed networks,

and experimentally validated three *B. subtilis* predictions that have important functional consequences.

Results

Strategy of a global probabilistic reconstruction

The conceptual outline of GLOBUS is shown in Figure 1. First, we built a generic metabolic network containing all possible metabolic activities characterized in the Enzyme Commission (EC) system (http://www.chem.qmul.ac.uk/iubmb/enzyme/). Nodes of this EC network represent known enzymatic activities (Fig. 1a), and network edges are established by metabolites shared between the activities either as substrates or products[25]. The usage of the global EC network allowed us to consider gene function in a proper network context without predefining metabolic pathways. With the EC network as a scaffold, the global metabolic reconstruction for a given organism is equivalent to assigning metabolic genes to their correct network locations (Fig. 1b). In this way, organism-specific networks will occupy a subset of all possible locations (activities) in the global EC network.

A gene assigned to its correct network location usually has at least remote sequence identity to enzymes known to catalyze the corresponding activity. In addition, a correctly assigned gene often has good context correlations with its network neighbors. As we demonstrated previously, the genes with high mutual context correlations tend to be located closer in metabolic networks[22]. For example, we show that the higher a context correlation between a pair of *S. cerevisiae* genes, the more likely that the genes are direct network neighbors (Supplementary Results, Supplementary Figure 1).

In GLOBUS we used sequence homology and context correlations to evaluate a given global assignment of multiple metabolic genes into a set of network locations using a Markov-like fitness function. The contribution of each gene to the fitness function depends on the sequence identity to the assigned location and the context correlations with the genes assigned to neighboring network positions. The overall GLOBUS fitness function $E(g_1, g_2, \ldots, g_n)$ (see Methods), which is calculated based on a given assignment of metabolic genes $(g_1, g_2, \ldots, g_n)$, consists of the following terms:

$$E(g_1, g_2, \ldots, g_n) = -b_{\text{homology}} f_{\text{homology}} - b_{\text{orthology}} f_{\text{orthology}} - b_{\text{context}} f_{\text{context}} - b_{\text{ECco-occurrence s}} f_{\text{ECco-occurrence s}} \quad (1)$$

where f_s are various homology-based and context-based functional descriptors, and b_s are corresponding positive coefficients representing weights of each descriptor in the fitness function. For homology descriptors we used two separate terms: 1.) the highest sequence identity to a Swiss-Prot[13] protein annotated to catalyze the corresponding activity in other species (annotations marked as based exclusively on computational methods were excluded), and 2.) a binary (0 or 1) descriptor indicating if a protein ortholog in another species is annotated to catalyze the activity. For context-based descriptors we used three types of gene-gene correlations: phylogenetic profiles (which quantify the co-occurrence of gene orthologs across species, see Methods), chromosomal gene clustering across sequenced genomes, and mRNA co-expression. For each context descriptor, we considered the maximum correlation Z-score (see Methods) between the gene under consideration and

genes assigned to neighboring network locations. In addition, we also considered a context term describing the co-occurrence across sequenced genomes of various metabolic activities according to annotations available in the KEGG database.

Using the described fitness function, the global probability for a particular assignment of multiple genes into their network locations is given by $P(g_1, g_2, \ldots, g_n)$ based on the relationship used in statistical physics and Markov Random Fields (MRF)[26]

$$P(g_1, g_2, \ldots, g_n) = \frac{1}{Z} \times e^{-E(g_1, g_2, \ldots, g_n)} \quad (2)$$

, where $E(g_1, g_2, \ldots, g_n)$ is the aforementioned fitness function, and Z is a normalizing partition function, which is necessary to insure that probabilities of all possible metabolic assignments sum to one. Using the defined probabilities we sampled from all possible assignments proportionally to their likelihood using Gibbs sampling[27]. Gibbs sampling is a version of Markov Chain Monte Carlo (MCMC)[28] and has been successfully used in many computational biology applications, such as finding transcription factor binding sites in a set of DNA sequences[29]. The efficiency of the Gibbs sampling in GLOBUS is due to the fact that although there is a combinatorially large number of possible metabolic assignments, the vast majority of them have very low probabilities. The Gibbs sampling allows to efficiently sample the most relevant global assignments according to their probabilities.

A step in a Gibbs chain was simulated by: 1.) selecting a random gene assigned to a particular network location, 2.) determining the probabilities for all possible locations of the selected gene, including the present location, and 3.) re-assigning the gene to a location according to the calculated probabilities (Fig. 1c,d). In the sampling we only considered the locations with at least remote sequence identity to the corresponding gene. In addition to possible locations in the network, a special out-of-the-network node was created, and in all Gibbs steps the move to the out-of-the-network node was also considered. The energy contribution to the fitness function for all genes located in the out-of-the-network node was the same. The energy in the out-of-the-network node is a parameter of the simulation (see below), it ensures that genes with little sequence identity or context correlation to any network location have a low probability of being assigned to an EC number. Importantly, we empirically established the absence of ergodicity problems in Gibbs sampling of microbial genomes. In other words, the annotation probabilities converged to essentially the same values for chains started from different random assignments; after about 20000 iterations the maximum probability difference across all genes was < 1%. Based on the convergent Gibbs chains we obtained the marginal probabilities for each metabolic assignment, consistent with the global fitness function.

Optimization of the fitness function parameters

The GLOBUS fitness function contains several important adjustable parameters b_s, that represent relative weights of several sequence and context correlations. The values of these parameters directly affect the sampling and the resulting gene annotation probabilities. To learn the parameters we applied a maximal likelihood approach using a well-annotated metabolic model of *S. cerevisiae* (iLL672[30]). Specifically, following the approach

commonly used in MRF[26], we optimized the fitness function parameters to maximally increase the product of the probabilities for correct gene assignments in the yeast network. Multiple simulated annealing[31] runs were used to the search the parameter space for maximal likelihood values. Importantly, in searching for the parameters over-fitting was not an issue as many hundreds of known metabolic annotations (485 yeast genes with EC numbers in the iLL672 model) dominate the number of optimized parameters (7 parameters in total). As a result of the maximum likelihood optimization, the yeast genes in their correct network locations had a geometric mean probability of 0.617, and an overall prediction accuracy of 80.5%, i.e. the overlap with the iLL672 model when genes were assigned to their most probable locations. Using more recent metabolic models of *S. cerevisiae* (iMM904[32]) or *B. subtilis* (iBsu1103[33]) for optimization resulted in similar parameter values and similar GLOBUS probabilities (Supplementary Fig. 2). Thus, we used the parameters optimized with the iLL672 model for GLOBUS metabolic annotations in other species.

GLOBUS precision-recall performance

To understand the utility of GLOBUS for metabolic network annotations we applied it to the genomes of a gram-positive model bacterium, *B. subtilis*, and a medically important bacterium, *S. aureus*. The genomes of these bacteria contain 1244 (*B. subtilis*) and 854 (*S. aureus*) genes with at least remote sequence identity to known enzymes in other species. Several curated metabolic models are also available for these species: iYO844[34] and iBsu1103[33] for *B. subtilis* and iSB619[35] for *S. aureus.* The parameters optimized using the yeast model (see above) were used in Gibbs sampling of all possible metabolic assignments in the two bacteria. The GLOBUS annotation probabilities were generated and precision-recall curves calculated (Fig. 2a) based on comparison with the corresponding curated models. For comparison we also show in the figure the precision-recall curves calculated based only on sequence identity to enzymes in other species; similar results were obtained using either BLAST or PSI-BLAST[36] (Supplementary Fig. 3). The precision-recall calculations demonstrate that GLOBUS substantially outperforms homology in the areas of high recall and high precision.

Further analysis (Fig. 2b,c) demonstrates that the main source of the superior GLOBUS performance lies in more accurate annotations of genes with low sequence identity to known enzymes. In Figure 2b we show the recall (at 70% precision) for gene annotations in *B. subtilis* and *S. aureus* as a function of sequence identity to known enzymes. GLOBUS recovers significantly more correct assignments compared to homology (10%, $P < 4 \times 10^{-4}$ for *B. subtlis*, and 14%, $P < 5 \times 10^{-5}$ for *S. aureus*, χ^2 test), especially for cases with less than 40% sequence identity. In Figure 2c we show that at the same level of recall (90%) GLOBUS achieves significantly higher precision (9% and 11% more, $P < 8 \times 10^{-5}$ and $P < 5 \times 10^{-3}$). The difference in precision is again highest for genes with low sequence identity to known enzymes, which constitutes a substantial fraction of all potential metabolic genes (Supplementary Fig. 4).

To investigate the contribution of individual context correlations to the GLOBUS performance, we optimized the coefficients of the fitness function without each context descriptor. We then compared the precision and recall values for predictions using all

context correlations and predictions obtained without individual correlations (see Supplementary Fig. 5). This analysis showed that all correlations contribute to the method's accuracy and that – similar to the complete fitness function - the effects of the individual context correlations are most apparent for cases with lower sequence identity.

We investigated the potential utility of GLOBUS for refining existing metabolic models by comparing two curated models of *B. subtilis*[33,34] (older iYO844, newer iBsu1103) and two models of *S. cerevisiae*[30,32] (older iLL672, newer iMM904). Specifically, we considered all annotations with non-zero GLOBUS probabilities that were not included in the older metabolic models. We then subdivided these non-zero GLOBUS annotations into those that were included in the newer models and those that were not included in the newer models for each species. This analysis showed (see Supplementary Fig. 6) that for both species, and across different sequence identity bins, higher GLOBUS probabilities corresponded to higher likelihoods of being included in the newer metabolic models.

Specific metabolic predictions and biochemical validation

GLOBUS results indicate that in many cases context correlations provide crucial functional evidence determining correct annotations, especially when sequence identity is small. One example is the *B. subtilis* gene *hemD*, known to be responsible for the uroporphyrinogen-III synthase activity[37] (EC 4.2.1.75). The sequence identity of *hemD* to the closest Swiss-Prot sequence performing its correct function is only ~24%; however, GLOBUS assigned a high probability (P=0.86) to the correct EC number because of the excellent context associations with its neighboring enzymes at this location: the gene clustering Z-score (defined as the number of standard deviations from the mean based on all gene-gene context scores, see Methods) is 21.2, the co-expression Z-score is 5.64. Context correlations are also helpful in selecting between potential functions with comparable sequence identity. For instance, the *B. subtilis* 8-amino-7-oxononanoate synthase *bioF*[38] has ~39% sequence identity to both its correct function (EC 2.3.1.47) and to glycine C-acetyltransferase (EC 2.3.1.29). GLOBUS selected the correct assignment (P=0.64 vs. 0.02) despite the equivalent sequence identity due to high clustering and co-expression Z-scores (16.6 and 4.3, respectively) in the correct location compared to the alternative location (1.1 and 2.4).

In Table 1 (*B. subtilis*) and Supplementary Table 1 (*S. aureus*) we list GLOBUS predictions without experimental validation that have high annotation probabilities despite low sequence identity to enzymes responsible for corresponding functions in other species. The annotations in the tables are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing sequence identity distance to known enzymes. For each prediction in the table we also show the average Z-score for the three context correlations in the corresponding network location.

From the predictions listed in Table 1 we selected the genes *spsI*, *spsJ*, and *ykgB* for experimental validation. The first two genes were selected because they were predicted to catalyze the first two steps in a rhamnose biosynthesis pathway (Supplementary Fig. 7); the other two genes from the pathway (*spsK* and *spsL*, in Table 1) were also predicted by GLOBUS. Rhamnose is a main sugar component of the *B. subtilis* exosporium[39]. The *sps* genes are transcribed from a σ^K-controlled promoter at late stages of *B. subtilis* sporulation

when the outer components of the spore coat are being assembled[40]. The gene *ykgB* was selected because GLOBUS predicted (with probability P=0.51) that this gene catalyzes the long elusive 6-phosphogluconolactonase activity of the *B. subtilis* pentose phosphate (PP) pathway. Despite a central role of PP pathway in the *B. subtilis* metabolism, this enzymatic activity remains without experimental validation in this important model organism.

The three proteins selected for experimental validation were over-expressed in *E. coli* and purified by His-Tag affinity and anion exchange chromatography. The correct identity of the purified proteins was confirmed by in-gel tryptic digestion and subsequent peptide analysis using mass spectrometry (Supplementary Dataset 1). In vitro enzymatic assays for SpsI and SpsJ were performed using a published method[41]. Predicted SpsI substrates (dTTP and α-D-glucose-1-phosphate, Fig. 3a) were observed in negative ionization mode high-precision mass-spectra profiles at 259.022 m/z and 480.981 m/z (**M**-H^+) respectively. Intensities of both dTTP and α-D-glucose-1-phosphate decreased only when SpsI was present in the assays, indicating that the enzyme uses these compounds as substrates (Supplementary Fig. 8). In addition, the predicted reaction product (dTDP-glucose) accumulated at 563.068 m/z (**M**-H^+) only in the presence of SpsI (Fig. 3b,c). The product of SpsJ (dTDP-4-dehydro-6-deoxy-glucose) was observed at 545.058 m/z (**M**-H^+) only in the presence of both SpsI and SpsJ (Fig. 3b,d), suggesting that SpsJ indeed converts dTDP-glucose into dTDP-4-dehydro-6-deoxy-glucose (Fig. 3a). Product accumulation, as well as substrate consumption, exhibited a clear dependence on the protein concentrations within a wide range around the estimated *in vivo* concentration of glucose-1-phosphate thymidylyltransferase (~1 μM for RfbA in *Escherichia coli*[42]).

Similarly to SpsI/SpsJ, the YkgB activity (Fig. 4a) was followed by observing the 6-phospho-gluconolactone degradation with online flow injection into a high-precision mass-spectrometer operating in the negative ionization mode. The intensity at the mass of 257.007 m/z (**M**-H^+), corresponding to 6-phospho-gluconolactone, decreased with rates faster than the rate of spontaneous background hydrolysis only when YkgB was present in the assays (Fig. 4b). The 6-phospho-gluconolactone degradation rate also exhibited a clear dependence on the protein concentration (Fig. 4c) within a wide range around the estimated *in vivo* 6-phosphogluconolactonase concentration (~1.5 μM for YbhE in *Escherichia coli*[42]). Similarly, the production rate of 6-phosphogluconic acid was consistently higher than the background when YkgB was present in the assays (Supplementary Fig. 9). Interestingly, available expression and proteomic data show that the *ykgB* gene is transcribed during several environmental conditions[43,44], such as heat and phenol stress. This suggests that YkgB - similar to lactonases in other species[45] - is likely to play a role in removing toxic byproducts of the PP pathway.

Discussion

Due to inherent uncertainty of computational annotations, predictions of biochemical function need to be treated probabilistically. Currently, most publicly available biochemical databases do not provide quantitative probabilities or confidence measures for existing annotations. This makes it hard for the users of these valuable resources to distinguish between confident assignments and mere guesses. As the application and impact of genome-

scale metabolic networks rapidly expands[1], a probabilistic treatment of annotations is essential. The GLOBUS approach, which is based on statistical sampling of possible biochemical assignments, provides a principled framework for such global probabilistic annotations. The method assigns annotation probabilities to each gene, as well as suggests likely alternative functions.

We demonstrate that context correlations can significantly improve the accuracy of biochemical predictions, especially when annotations are based on distant sequence identity. Over half of potential metabolic genes, even in such well-studied model organisms as *S. cerevisiae* and *B. subtilis*, have remote sequence identity (<40%) to known enzymes (Supplementary Fig. 4). Application of GLOBUS to less-studied organisms should be straightforward, as context-based correlations, excluding gene co-expression, are calculated directly from genome sequences; the reduction in the overall accuracy due to the co-expression term is relatively small (<1%). The precision of other context correlations should only improve with the rapid growth of fully sequenced genomes.

Probabilistic predictions generated by GLOBUS can be directly used to annotate sequences and genomes. GLOBUS annotations can be also used by various gap identification and gap filling approaches[22,23,46,47] to produce simulation-ready flux balanced networks. In addition, recent advances in metabolomics, proteomics, and fluxomics offer complementary opportunities to expand and refine biochemical annotations and network reconstructions [48]. The flexibility of the GLOBUS framework makes it easy to integrate metabolomics and proteomics data. For example, as genes are moved through the network to sample possible assignments, available data for corresponding proteins and metabolites can be included in the global fitness function. Additional functional descriptors, for example based on protein structure and information about protein localization, can be also considered in the framework. Such probabilistic integration of diverse biochemical data will be crucial for exploiting the ongoing avalanche of genomic sequencing.

Methods

Construction of the generic EC network

In the construction of the EC (Enzyme Commission) network we considered 3284 EC numbers (http://www.chem.qmul.ac.uk/iubmb/enzyme/) responsible for biochemical activities involving small compounds as substrates and products; activities such as “RNA polymerase” or “protein kinase” were excluded. In the global EC network, nodes represent EC numbers connected by edges representing metabolites shared between reactions. Following a common procedure [25], linkages through the top 40 most highly connected metabolites and cofactors were not considered (Supplementary Table 2).

Identification of potential metabolic genes and their functions

The program BLAST[36] (with E-value cutoff of $5*10^{-2}$) was used for homology searches against enzymes in Swiss-Prot[13], excluding sequences that were: 1) from genomes of closely related species (species in the same taxonomic genus) or 2) likely annotated based exclusively on computational methods, i.e., annotations containing words *probable*, *by*

similarity, *hypothetical*, *like*, or *putative*. Although many remaining annotations in Swiss-Prot are also derived using computational methods, they are usually curated, ensuring that the misannotation rate in this database is relatively low [8,9].

To account for multi-functional enzymes, when non-overlapping regions of a query gene could be mapped to different enzymatic functions - indicating domains responsible for distinct metabolic activities - the mapped regions of the query gene were allowed to be assigned independently to different network locations.

The functional descriptors in the GLOBUS fitness function

Detailed description of the energy function and related calculations are given in Supplementary Methods. Denoting by *n* the total number of considered metabolic genes, the components of the fitness function used in GLOBUS are as follows:

Sequence homology. $f_{homology}$—As the sequence identity descriptor we used the logarithm of the conditional probability that the gene performs the assigned metabolic function, given the highest sequence identity to a Swiss-Prot [13] protein annotated to catalyze the corresponding activity:

$$f_{\text{homology}}=\sum_{i=1}^{n} log\, P\,(\text{gene performs target function}|\text{highest sequence identity to annotated Swiss}-\text{Prot protein}) \quad (3)$$

Orthology. $f_{orthology}$—An additional binary descriptor related to sequence homology was the likely gene orthology to a gene from another species annotated with the target activity. For each gene, the orthology term was either 1, if at least one possible ortholog was annotated in Swiss-Prot to perform the target activity, or 0, if no orthologs with the target activity could be identified.

Gene-gene context correlations. $f_{context}$—In GLOBUS we used the context correlations (phylogenetic profiles, chromosomal clustering, mRNA co-expression) by: 1.) transforming them into Z-scores[49] (number of standard deviations from the mean) using the distribution of correlations for all pairs of metabolic genes, and 2.) estimating the conditional probability that two genes are direct network neighbors, given their context association Z-score. The corresponding conditional probabilities were derived using the iLL672 yeast metabolic model (Supplementary Fig. 1a-c). In the GLOBUS fitness function, for each assigned gene we considered the maximum log probability among all network neighbors of the gene:

$$f_{\text{context}}=\sum_{i=1}^{n} max\,(logP\,(\text{two genes are network neighbors}|\text{context correlation Z}-\text{score between the genes}))$$

EC co-occurrences. $f_{ECco\text{-}ocurrence}$—This descriptor measures the correlation between the occurrences of different metabolic activities (EC numbers) across sequenced species without considering specific genes assigned to the activities. In the GLOBUS fitness

function for each assigned gene we considered the EC co-occurrence descriptor equal to the average correlation between the EC activity of the assigned gene and the EC activities for all its network neighbors. The EC co-occurrence term provides information additional to that available from direct sequence homology. The most relevant information about homology usually comes from annotated enzymes with the highest sequence identity to a protein under consideration. On the other hand, the EC co-occurrence reflects common presence and absence of metabolic activities across multiple KEGG genomes. Thus, this term quantifies tendencies of closely related activities to be filled together.

Experimental validation of biochemical predictions

Different amounts of purified SpsI or SpsJ were incubated at 37 °C in 1 mL of 10 mM potassium phosphate buffer pH 7.4, 2.5 mM $MgCl_2$, 1 mM glucose-1-phosphate (Sigma-Aldrich, >= 97% purity), 1 mM dTTP (Sigma-Aldrich, >= 96% purity) and 1 U pyrophosphatase[41]. The enzyme reaction samples were assayed after 4 hours by flow-injection into a time of flight mass spectrometer (6520 Series QTOF, Agilent Technologies) operated in the negative ionization mode. High-precision mass spectra were recorded from 50-1000 m/z and analyzed as described previously[50]. Acquired masses were deviating less than 0.001 atomic mass units (amu) from the reference masses 259.022, 480.982, 545.058, and 563.068 for α-D-glucose-1-phosphate, dTTP, dTDP-glucose, and dTDP-4-dehydro-6-deoxy-glucose, respectively.

Purified YkgB was assayed in 1 mL 5 mM potassium phosphate buffer pH 7, 2.5 mM $MgCl_2$, and freshly prepared 6-phospho-gluconolactone. The lactone was prepared freshly from 6-phospho-gluconic acid (Sigma-Aldrich, >= 90% purity) by lyophilization, and its degradation due to the YkgB activity was followed by direct online flow-injection into a time of flight mass spectrometer as described above. Acquired masses were deviating less than 0.001 atomic mass units (amu) from the reference masses 257.007 and 275.017 for 6-phospho-gluconolactone and 6-phosphogluconic acid, respectively.

A detailed description of the cloning, purification and protein identification procedure is given in the Supplementary Methods.

Computational requirements and statistical analysis

The calculations were performed using the 3GHz Intel Xeon quad core processor with 256MB of RAM memory. GLOBUS run times depend both on the number of iterations and the number of genes considered for a given species. For the *S. cerevisiae*, *S. aureus*, and *B. subtilis* genomes, 10,000 iterations over all genes took about 10 minutes. The run time increased linearly with the number of iterations and number of genes. 20,000-50,000 iterations (20-50 minutes) were required to achieve 1% convergence of annotation probabilities, i.e. so that there were no gene assignments different in their annotation probabilities by more than 1% between different runs. Pre-computed GLOBUS predictions for 10 bacterial species of medical interest can be found at: http://vitkuplab.c2b2.columbia.edu/globus/index.html

P-values used to compare the precision-recall performances for GLOBUS and sequence identity were calculated using the one-tailed χ^2 test, N = 332 to 717 annotations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Ruth Hüttenhain from the Aebersold lab at IMSB Zurich for technical assistance and measuring the peptide samples. We thank Patrick Eichenberger for providing mutant strains. This work was supported in part by NIH grant GM079759 to D.V. and National Centers for Biomedical Computing (MAGNet) grant U54CA121852 to Columbia University.

References

1. Oberhardt MA, Palsson BO, Papin JA. Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 2009; 5:320. [PubMed: 19888215]
2. Almaas E, Oltvai ZN, Barabasi AL. The activity reaction core and plasticity of metabolic networks. PLoS Comput. Biol. 2005; 1:e68. [PubMed: 16362071]
3. Kanehisa M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008; 36:D480–4. [PubMed: 18077471]
4. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B. Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics. 2006; 7:296. [PubMed: 16772023]
5. Karp PD, et al. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief. Bioinform. 2010; 11:40–79. [PubMed: 19955237]
6. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat. Rev. Mol. Cell Biol. 2007; 8:995–1005. [PubMed: 18037900]
7. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol. 2003; 333:863–82. [PubMed: 14568541]
8. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. 2009; 5:e1000605. [PubMed: 20011109]
9. Hsiao TL, Revelles O, Chen L, Sauer U, Vitkup D. Automatic policing of biochemical annotations using genomic correlations. Nat. Chem. Biol. 2010; 6:34–40. [PubMed: 19935659]
10. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res. 2009; 37:D588–92. [PubMed: 18984617]
11. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics. 1998; 14:656–64. [PubMed: 9789091]
12. Caspi R, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2008; 36:D623–31. [PubMed: 17965431]
13. Boeckmann B, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003; 31:365–70. [PubMed: 12520024]
14. Overbeek R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33:5691–702. [PubMed: 16214803]
15. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA. 1999; 96:4285–8. [PubMed: 10200254]

16. Yanai I, Derti A, DeLisi C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA. 2001; 98:7940–5. [PubMed: 11438739]
17. Wu LF, et al. Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat. Genet. 2002; 31:255–65. [PubMed: 12089522]
18. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA. 1999; 96:2896–901. [PubMed: 10077608]
19. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature. 2000; 405:823–6. [PubMed: 10866208]
20. Korbel JO, Jensen LJ, von Mering C, Bork P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat. Biotechnol. 2004; 22:911–7. [PubMed: 15229555]
21. von Mering C, et al. Genome evolution reveals biochemical networks and functional modules. Proc. Natl. Acad. Sci. USA. 2003; 100:15428–33. [PubMed: 14673105]
22. Chen L, Vitkup D. Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol. 2006; 7:R17. [PubMed: 16507154]
23. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics. 2006; 7:177. [PubMed: 16571130]
24. Price ND, Shmulevich I. Biochemical and statistical network models for systems biology. Curr. Opin. Biotechnol. 2007; 18:365–70. [PubMed: 17681779]
25. Kharchenko P, Church GM, Vitkup D. Expression dynamics of a cellular metabolic network. Mol. Syst. Biol. 2005; 1:0016. 2005. [PubMed: 16729051]
26. Li, SZ. Markov random field modeling in image analysis. Springer; Tokyo: 2001. p. xixp. 323
27. Casella G, George EI. Explaining the Gibbs sampler. Am. Stat. 1992; 46:167–174.
28. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57:97–109.
29. Lawrence CE, et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science. 1993; 262:208–14. [PubMed: 8211139]
30. Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome Res. 2005; 15:1421–30. [PubMed: 16204195]
31. Kirkpatrick S, Gelatt CD Jr. Vecchi MP. Optimization by Simulated Annealing. Science. 1983; 220:671–680. [PubMed: 17813860]
32. Mo ML, Palsson BO, Herrgard MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst. Biol. 2009; 3:37. [PubMed: 19321003]
33. Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. Genome Biol. 2009; 10:R69. [PubMed: 19555510]
34. Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. J. Biol. Chem. 2007; 282:28791–9. [PubMed: 17573341]
35. Becker SA, Palsson BO. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. BMC Microbiol. 2005; 5:8. [PubMed: 15752426]
36. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–402. [PubMed: 9254694]
37. Stamford NP, Capretta A, Battersby AR. Expression, purification and characterisation of the product from the Bacillus subtilis hemD gene, uroporphyrinogen III synthase. Eur. J. Biochem. 1995; 231:236–41. [PubMed: 7628476]
38. Bower S, et al. Cloning, sequencing, and characterization of the Bacillus subtilis biotin biosynthetic operon. J. Bacteriol. 1996; 178:4122–30. [PubMed: 8763940]
39. Faille C, et al. Morphology and physico-chemical properties of Bacillus spores surrounded or not with an exosporium: consequences on their ability to adhere to stainless steel. Int. J. Food Microbiol. 2010; 143:125–35. [PubMed: 20739077]

40. Eichenberger P, et al. The program of gene transcription for a single differentiating cell type during sporulation in Bacillus subtilis. PLoS Biol. 2004; 2:e328. [PubMed: 15383836]
41. Timmons SC, Mosher RH, Knowles SA, Jakeman DL. Exploiting nucleotidylyltransferases to prepare sugar nucleotides. Org. Lett. 2007; 9:857–60. [PubMed: 17286408]
42. Ishihama Y, et al. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics. 2008; 9:102. [PubMed: 18304323]
43. Hecker M, Reder A, Fuchs S, Pagels M, Engelmann S. Physiological proteomics and stress/starvation responses in Bacillus subtilis and Staphylococcus aureus. Res. Microbiol. 2009; 160:245–58. [PubMed: 19403106]
44. Tam le T, et al. Proteome signatures for stress and starvation in Bacillus subtilis as revealed by a 2-D gel image color coding approach. Proteomics. 2006; 6:4565–85. [PubMed: 16847875]
45. Galperin MY, Moroz OV, Wilson KS, Murzin AG. House cleaning, a part of good housekeeping. Mol. Microbiol. 2006; 59:5–19. [PubMed: 16359314]
46. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics. 2007; 8:212. [PubMed: 17584497]
47. Henry CS, et al. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat. Biotechnol. 2010; 28:977–82. [PubMed: 20802497]
48. Breitling R, Vitkup D, Barrett MP. New surveyor tools for charting microbial metabolic maps. Nat. Rev. Microbiol. 2008; 6:156–61. [PubMed: 18026122]
49. Faith JJ, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007; 5:e8. [PubMed: 17214507]
50. Fuhrer T, Heer D, Begemann B, Zamboni N. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. Anal. Chem. 2011; 83:7074–80. [PubMed: 21830798]

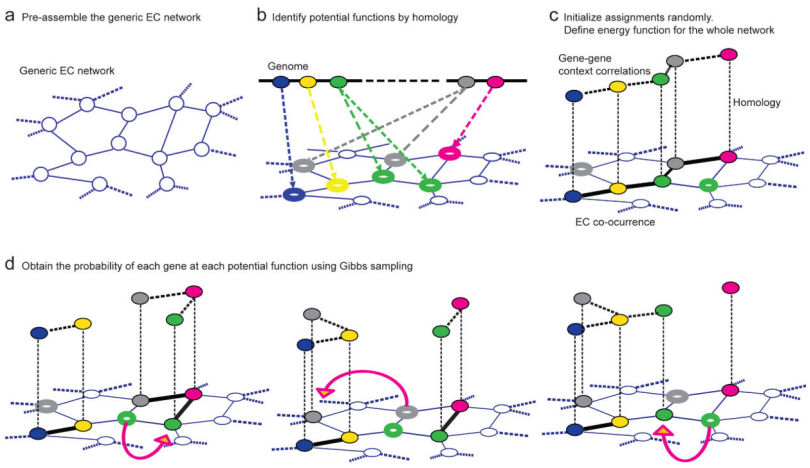

Figure 1. Overview of the GLOBUS method

(a) A generic Enzyme Commission (EC) network, where nodes represent all known biochemical activities and edges indicate metabolites shared between activities. **(b)** For a genome of interest, the potential network locations of each gene are assigned based on sequence homology to known enzymes. **(c)** Each gene is initially assigned randomly to one of its possible locations. A fitness function is defined such that assignments to locations with high sequence identity and good context correlations with neighboring genes correspond to higher values of the fitness function (higher probability). **(d)** Gibbs sampling is used to sample all possible assignments of genes to their candidate network locations. At each step of a Gibbs chain a random gene is selected and re-assigned to one of its possible locations (arrows). The marginal probabilities for assigning every gene to each candidate network location are derived from converged Gibbs chains.

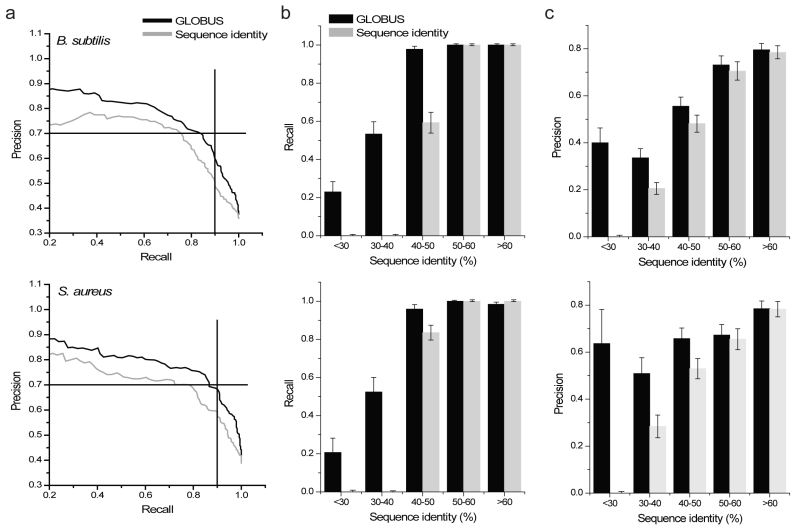

Figure 2. GLOBUS precision-recall performance

Using available metabolic models (iBsu1103[33] for *B. subtilis* and iSB619[35] for *S. aureus*) we compared predictions by GLOBUS to predictions made using sequence homology; predictions for *B. subtilis* are on the top, and predictions for *S. aureus* are on the bottom. **(a)** Precision–recall curves for GLOBUS (black lines) were calculated by ranking genes using assignment probabilities. Precision-recall curves for homology (red lines) were calculated by ranking genes using sequence identity. **(b)** Recall of known metabolic genes (at 70% precision) as a function of sequence identity to the closest enzymes from other species with the annotated functions. **(c)** Prediction precision (at 90% recall) for known metabolic genes as a function of sequence identity to the closest enzymes from other species with the annotated functions. In the figure error bars represent the S.E.M,

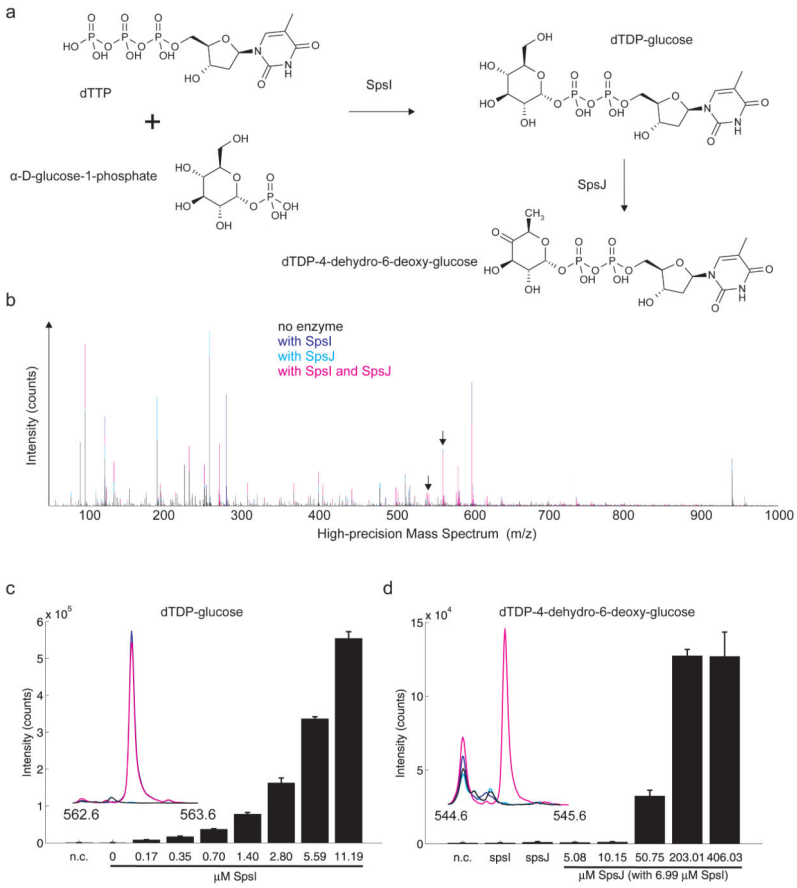

Figure 3. In vitro biochemical assays used to characterize activities of SpsI and SpsJ using high-precision mass spectrometry

(a) Reaction diagram. **(b)** Mass spectrum plot showing intensities for masses corresponding to the products dTDP-glucose and dTDP-4-dehydro-6-deoxy-glucose of the reactions catalyzed by SpsI and SpsJ (black arrows, detailed in panel c). Observed masses deviated by less than 0.001 atomic mass units (amu) from the corresponding reference masses. Spectra were recorded from two independent assays. **(c, d)** Bar plots show dependency of dTDP-glucose and dTDP-4-dehydro-6-deoxy-glucose accumulation on protein concentration of SpsI and SpsJ, respectively. As negative control (n.c.), the protein free filtrate of 6.99 μM spsI or 203.01 μM SpsJ solution was used. Error bars represent standard deviations calculated using two independent assays.

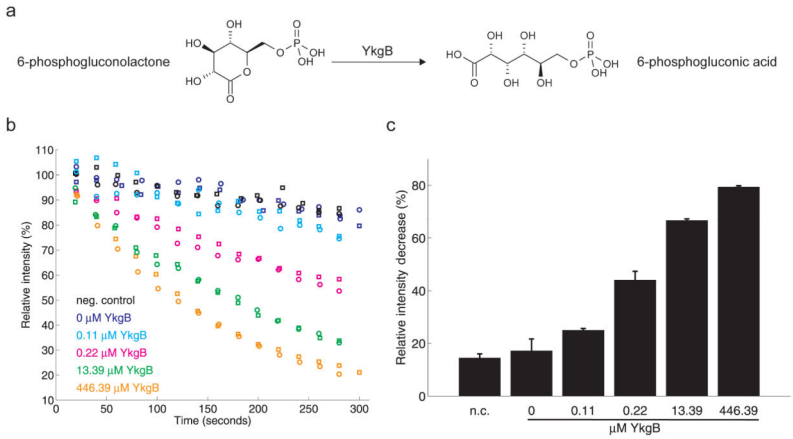

Figure 4. In vitro biochemical assays used to characterize the 6-phospho-gluconolactonase activity of YkgB

(a) Reaction diagram for 6-phospho-gluconolactonase. **(b)** Time courses of lactone degradation at different YkgB concentrations were recorded by direct flow injection analysis. Different symbols represent replicate assays. **(c)** Relative intensity increase from initial to final lactone intensities as a function of YkgB concentration. As negative control (n.c.), the protein-free filtrate of 223.2 μM YkgB solution was used. Error bars represent standard deviations calculated using two independent assays.

Table 1

Prediction of gene function in *B. subtilis*

In the table we show predictions without experimental validation that have GLOBUS-assigned probabilities above 0.5 and protein sequence identity to known enzymes below 50%. The first three activities in the table were experimentally validated in this study. The remaining annotations in the table are ordered by averaging the prediction ranks sorted by decreasing annotation probability and the prediction ranks sorted by decreasing sequence identity distance to known enzymes. The last column shows the average Z-score of phylogenetic correlations, gene clustering and gene co-expression when all sequences are assigned to their most probable locations. The Z-score for each type of data was calculated using the maximum context correlation between a gene and its immediate network neighbors (see Methods).

Gene	EC number	Enzyme name	Probability	Identity (%)	Average Context Z-score
spsI	2.7.7.24	glucose-1-phosphate thymidylyltransferase	0.93	44.4	11.6
spsJ	4.2.1.46	dTDP-glucose-4,6-dehydratase	0.97	48	12.0
ykgB	3.1.1.31	6-phosphogluconolactonase	0.51	30.4	2.6
murF	6.3.2.10	UDP-N-acetylmuramoyl-tripeptide-D-alanyl-D-alanine ligase	0.98	32.8	9.0
spsL	5.1.3.13	dTDP-4-dehydrorhamnose-3,5-epimerase	0.95	33.1	8.4
ycgM	1.5.99.8	proline dehydrogenase	0.76	25.6	3.6
yfnG	4.2.1.45	CDP-glucose-4,6-dehydratase	0.76	27.5	11.0
birA	6.3.4.15	biotin-[acetyl-CoA-carboxylase] ligase	0.77	31.7	2.3
gcvPB	1.4.4.2	glycine dehydrogenase (decarboxylating)	0.97	41.5	12.3
yloI	4.1.1.36	phosphopantothenoylcysteine decarboxylase	0.99	44.5	2.6
fruK	2.7.1.56	1-phosphofructokinase	0.88	40.4	10.9
spsK	1.1.1.133	dTDP-4-dehydrorhamnose reductase	0.87	39.6	8.4
murB	1.1.1.158	UDP-N-acetylmuramate dehydrogenase	0.97	43	5.2
folK	2.7.6.3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine diphosphokinase	0.99	45.3	8.0
sul	2.5.1.15	dihydropteroate synthase	0.99	47	8.2
yitJ	2.1.1.13	methionine synthase	0.54	30.6	2.1
ybbF	2.7.1.69	protein-Npi-phosphohistidine-sugar phosphotransferase	0.85	40.5	11.3
yloI	6.3.2.5	phosphopantothenate-cysteine ligase	0.97	44.5	2.9
pheA	4.2.1.51	prephenate dehydratase	0.69	36.1	6.7
purK	4.1.1.21	phosphoribosylaminoimidazole carboxylase	0.89	43.5	13.3
ysnA	3.6.1.15	nucleoside-triphosphatase	0.56	33.3	7.7
ywbC	4.4.1.5	lactoylglutathione lyase	0.6	35.2	3.6
pucE	1.2.3.14	abscisic-aldehyde oxidase	0.62	35.8	1.0
ydhR	2.7.1.4	fructokinase	0.77	41.5	5.3
yfnH	2.7.7.33	glucose-1-phosphate cytidylyltransferase	0.88	43.2	11.0
ybbD	3.2.1.52	beta-N-acetylhexosaminidase	0.52	33.1	3.1
yngE	6.4.1.4	methylcrotonoyl-CoA carboxylase	0.64	36.2	8.6

Gene	EC number	Enzyme name	Probability	Identity (%)	Average Context Z-score
kbl	2.3.1.29	glycine C-acetyltransferase	0.97	49	9.4
tenI	2.5.1.3	thiamine-phosphate diphosphorylase	0.7	40.6	6.6
pabB	4.1.3.27	anthranilate synthase	0.74	42.8	8.6