Sequence-based Network Completion Reveals the Integrality of Missing Reactions in Metabolic Networks

Elias W. Krumholz and Igor G. L. Libourel
From the Department of Plant Biology and the Biotechnology Institute, University of Minnesota, Saint Paul, Minnesota 55108

Background: Genome-scale draft metabolic networks are incomplete, even for well studied organisms. Reactions are referred to as gaps. These gaps need to be filled to enable functional analysis, and gap-filling choices influence model predictions. To investigate whether functional networks existed where all gap-filling reactions were supported by sequence similarity to annotated enzymes, four draft networks were supplemented with all reactions from the Model SEED database for which minimal sequence similarity was found in their genomes. Quadratic programming revealed that the number of reactions that could participate in a gap-filling solution was vast: 3,270 in the case of *E. coli*, where 72% of the metabolites in the draft network could connect a gap-filling solution. Nonetheless, no network could be completed without the inclusion of orphaned enzymes, suggesting that parts of the biochemical integral to biomass precursor formation are uncharacterized. However, many gap-filling reactions were well determined, and the resulting networks showed improved prediction of gene essentiality compared with networks generated through canonical gap filling. In addition, gene essentiality predictions that were sensitive to poorly determined gap-filling reactions were of poor quality, suggesting that damage to the network structure resulting from the inclusion of erroneous gap-filling reactions may be predictable.

Metabolic network reconstructions are instrumental in aggregating metabolic knowledge about organisms (1–3). See “Glossary” for explanations of the technical terms pertinent in our field that we use in this article. Network reconstructions have steadily grown in size, reflecting increasingly comprehensive genome annotations (4–7). In addition, reconstructions have grown in complexity. Current reconstructions contain detailed gene-to-protein-to-reaction mappings, thermodynamic constraints, and, in some cases, signal transduction layers (8–10). The most sophisticated reconstructions have been extensively curated (6, 11, 12), but draft reconstructions are now mostly machine-generated or machine-assisted (7, 13–16). The Model SEED uses annotations from the “Rapid Annotation Using Subsystems Technology” (RAST) web service (17, 18) as part of a network reconstruction pipeline for prokaryotes (15). In addition to a starting point for curated reconstructions, draft metabolic networks facilitate interpretation of the metabolic capabilities of newly sequenced organisms or communities of organisms (7, 19, 20).

Metabolic networks are reconstructed in a bottom-up fashion from identified genes following genome annotation (21). Knowledge of metabolic pathways can guide gene annotation, as implemented by the Pathway Tools software (22). Similarly, RAST simultaneously annotates genes that are part of a metabolic subsystem (18), utilizing mutually corroborating information on genes involved in closely related metabolic processes. As a logical extension of the subsystem approach, network-wide mutually corroborating information may be used to guide reconstructions. An application of this concept is to require draft networks to be able to carry out the production of all essential cellular building blocks, collectively referred to as biomass, from a well defined medium source (23).

Metabolic networks resulting from assembling all reactions inferred from gene annotations (draft networks) are currently unable to describe the synthesis of all biomass components. Draft networks contain gaps, isolated reactions, and reactions that cannot carry flux under any circumstances (1). Although isolated or blocked reactions are easily identified (24), it is not obvious whether they result from under-annotation or over-annotation. Hence, an isolated reaction may need to be connected through additional reactions that were under-annotated in the draft network, or the isolated reaction resulted from a spurious annotation. Gaps in the network pose the opposite problem; although a network can be readily completed to enable production of all biomass components (25), the location...
of the actual missing reaction may be illusive. The appearance of gaps in metabolic networks is not exclusively the result of under-annotation. Incorrect reaction reversibility assignments (thermodynamic constraints) (26) or stoichiometric constraints resulting from dead-end metabolites may also prevent production of biomass components (25, 27). Last, part of the biochemistry of an organism may not have been associated with genes, or the biochemistry may be yet to be discovered. Adding reactions to fill these gaps is known as “gap filling” and has been the subject of considerable inquiry and has been reviewed in detail elsewhere (24).

Commonly, mixed integer linear programming (MILP) optimization is used to perform bottom-up gap-filling (27). In this approach, reactions are iteratively added until production of biomass becomes feasible, often while minimizing the number of reactions required (15, 25, 27). Several other optimization strategies have been reviewed here (28). In the case of the Model SEED, reactions are prioritized based on their nature. For instance, adding an internal reaction incurs a lower cost than adding a transporter. Bottom-up gap filling works well for well annotated genomes, but for networks that require extensive gap filling, a top-down approach is more robust (29). In the more recently developed top-down methods, all gap-filling reactions are added, followed by the successive preferential removal of unneeded gap-filling reactions with little or no sequence similarity in the genome of the organism for which the network is reconstructed (14, 29, 30). Prioritization of the removal of reactions without sequence similarity minimizes the inclusion of locally (enzymes with an associated sequence that is not present in the target genome) and globally (reactions without sequence association) orphaned reactions. Very recently, a bottom-up MILP approach also used sequence similarity as a likelihood metric for the existence of a gap-filling reaction in the target genome (31). Gap analysis itself has been used to identify knowledge gaps in human metabolism (32) and to leverage contextual information of networks to hypothesize gene function (33).

This work investigated the need for adding gap-filling reactions to draft networks, the extent for which sequence similarity to enzymes can be found for these reactions, and how orphaned enzymes influence gene essentiality predictions by metabolic networks. To assess the presence of sequence support for gap-filling solutions, new linear programming (LP)- and quadratic programming (QP)-based gap-filling problems were formulated that minimize the utilization of unsupported reactions. All gene sequences associated with the Model SEED gap-filling reaction database (11,858 reactions, received on April 20, 2012) (15) were queried against four prokaryotic genomes, and unique gap-filling solutions were retrieved that minimized the utilization of unsupported reactions. Unlike recently reported BLAST-weighted MILP-based work (31), the networks resulting from our approach outperformed networks gap-filled by the Model SEED (15).

**Experimental Procedures**

*Metabolic Networks, Biochemistry Database, and Gene Annotations—*Metabolic networks for *Streptococcus pneumoniae*, *Bacillus subtilis*, *Escherichia coli* MG1655, and *Acinetobacter baylyi* ADP1 were downloaded from the Model SEED web site on May 3, 2013 (15) along with medium conditions and biomass formulations. The Model SEED gap-filling biochemistry database, experimental gene essentiality results, and associated medium formulations were kindly provided by Chris Henry (Argonne National Laboratory, Lemont, IL). Gene annotations for 891 prokaryotic species were downloaded from the RAST sapling server (34), totaling 690,445 genes encoding 7,218 functional roles. The biochemistry database maps genes to reactions through the use of functional roles and enzyme complexes made up of functional roles (15). Table 1 includes a summary of the size of the downloaded Model SEED database and draft metabolic networks.

**Identification of Functional Roles**—For each functional role in the biochemistry database, a BLAST amino acid database was generated using all protein sequences associated with that particular role. The complete genomes of target organisms were queried against each functional role BLAST database using BLASTX (35) with the BLOSUM62 scoring matrix (36). The $E$-values for the best BLAST high-scoring segment pairs from each functional role database query were used to weight biochemical reactions. $E$-values were chosen because they are comparable between different calls against distinct functional role databases and correct for multiple comparisons by penalizing the score by both the length of the enzyme database and the length of the target genome (37). Only the lowest $E$-value was recorded. To adjust the weights for each enzyme complex independently, duplicate reactions were created so that each complex had an independent mapping with a reaction. Reactions were weighted with the geometric mean of the $E$-values for the constituent roles of an enzyme complex. This treats the $E$-values as probabilities in determining the support for the existence of an enzyme, which is here defined as enzyme sequence support (ESS). Reactions with an ESS of less than $1.0E^{-240}$ were set to the value of $1.0E^{-240}$. Reactions were weighted by the logarithm of the ESS values of the associated enzymes,

$$W_R = \ln(E_R) - \ln(E_{\text{min}}) \quad (\text{Eq. 1})$$

where $W_R$ is the weight for a reaction, $E_R$ is the ESS for a reaction, and $E_{\text{min}}$ is the minimum $E$-value. This formulation results in small weights for well supported reactions relative to unsupported reactions while constraining the weights to a smaller numerical range, which improved the numerical stability of the LP and QP solver software.

**Gene Essentiality and Metabolite Production**—Flux balance analysis (FBA) (38) was used to check for the existence of a synthesis route for individual biomass components. A gene was classified as computationally essential if removing the reaction(s) uniquely associated with an enzyme complex resulted in a network that could not carry flux greater than $1.0E^{-6}$ to biomass. Similarly, an individual metabolite was classified as producible if a flux solution could be found that carried a flux $>1.0E^{-6}$ of the tested metabolite through an export reaction that was temporarily added for testing purposes (25).

**Gap-filling Algorithm**—The BLAST-weighted LP gap-filling algorithm was formulated as follows,
**Integrity of Orphaned Enzymes in Metabolic Networks**

**Model SEED database and model summary**

Metabolic networks produced by the Model SEED are subsets of the complete Model SEED gap-filling database. Relationships from gene to functional role to enzyme complex to reaction are encoded as a gene to reaction relationship in draft metabolic networks, thus removing the enzyme complex abstraction from the model. This compact encoding of relationships allows gene knockouts to be quickly translated into reaction knockouts in draft networks. NA, not available.

| Model SEED gap-filling database | S. pneumoniae | B. subtilis | E. coli K-12 MG1655 | A. baylyi ADP1 |
|-------------------------------|--------------|------------|-------------------|--------------|
| No. of functional roles       | 7,218        | 1,496      | 2,606             | 3,658        |
| No. of unique reactions       | 10,516       | 880        | 1,537             | 1,638        |
| No. of metabolites            | 7,732        | 848        | 1,280             | 1,278        |
| No. of enzyme complexes (equal to no. of reactions, including duplicated reactions) | 11,858 | NA | NA | NA |
| No. of genes                  | 690,445      | 480        | 952               | 1067         |

\[
\min f(v) = \mathbf{w}^T \mathbf{v} 
\]

(Eq. 2)

\[
\mathbf{Sv} = 0 \text{ such that: } 0 \leq v \leq v_{\text{max}} 
\]

[Eq. 3]

\[
v_{\text{bio}} = 1e - 3 
\]

where \( \mathbf{w} \) is a column vector of weights (Equation 1), and \( \mathbf{v} \) is a column vector of reaction fluxes including separate terms for forward and reverse reactions. The stoichiometric matrix (\( \mathbf{S} \)) relates reactions to metabolites through stoichiometric coefficients. A negative value in the stoichiometric matrix specifies a metabolite that is consumed by a reaction, and a positive value describes the production of a metabolite by a reaction. \( \mathbf{S} \) has dimensions \( m \) (metabolites) by \( 2n \) (\( n \) reactions, \( 2n \) for both directions). The constraint \( \mathbf{Sv} = 0 \) enforces that all metabolites have a net balance of production and consumption, known as a mass balance constraint. \( v_{\text{max}} \) is a vector of upper bounds on reaction fluxes, \( v_{\text{bio}} \) is the required flux through the biomass reaction. Similarly, a weighted QP gap-filling algorithm was formulated as follows,

\[
\min f(v) = \mathbf{v}^T \mathbf{W} \mathbf{v}
\]

(Eq. 4)

\[
\mathbf{Sv} = 0 \text{ such that: } v_{\text{min}} \leq v \leq v_{\text{max}} 
\]

[Eq. 5]

\[
v_{\text{bio}} = 1e - 3
\]

where \( \mathbf{W} \) is a diagonal matrix of the weights. Other terms were identical to the LP formulation, except that reactions did not need to be divided into forward and reverse directions and were constrained directly using two vectors: \( v_{\text{min}} \) and \( v_{\text{max}} \). The QP formulation results in fluxes that minimize the sum of weighted squared fluxes, which effectively distributes fluxes across available biomass routes inversely proportional to the weights and number of reactions of a given route.

**Software**—LP and QP problems were solved with CPLEX\textsuperscript{TM} (IBM, Armonk, NY). LP problems were solved using the dual simplex solver to minimize constraint violations. Custom Matlab\textsuperscript{TM} (Mathworks, Natick, MA) and Python scripts were used for the preparation of matrices and databases. Productivity of metabolites from medium components was tested by FBA using the COBRA Toolbox (39).

BLASTX comparisons for the four target genomes were run on a commodity quad core Intel i3 desktop computer, taking roughly 8 total h to complete. Further processing of the BLASTX output using custom Python scripts required approximately 4 h of computation time.

**Results**

*Metabolic Networks Require Gap Filling*—Draft metabolic networks for *S. pneumoniae*, *B. subtilis*, *E. coli* MG1655, and *A. baylyi* ADP1 were downloaded from the Model SEED. Corresponding experimental gene essentiality results of genome-scale single gene knock-out libraries (40–43) along with mappings to the model genes were provided by the Model SEED upon request in October 2011. The four organisms were selected because the associated gene knock-out libraries were generated through full-length gene deletion methods rather than transposon insertion methods. Transposon knockouts can display complex gene knockdown behavior that complicates the interpretation of gene essentiality predictions (44). The gap-filling reactions added by the Model SEED were stripped from the downloaded models. In addition, all genes not directly associated with metabolic reactions were not evaluated to limit gene essentiality evaluations to precursor metabolism only.

The draft networks now contained gaps resulting from under-annotation of the genome, incorrect reaction reversibility constraints resulting from inaccurate Gibbs free energy estimates, and stoichiometric constraints caused by dead-end metabolites. Consequently, there were three approaches to gap-fill metabolic networks by addressing each of the three causes. To test whether removal of thermodynamic constraints alone could enable biomass production, all reactions were made reversible. The existence of a route to biomass was tested using FBA by maximizing flux through the biomass reaction. No such route existed for any of the four networks, demonstrating that removal of thermodynamic constraints alone was insufficient to gap-fill the tested networks. Removal of stoichiometric constraints caused by dead-end metabolites by allowing all metabolites to leave the network was also insufficient. Furthermore, a combination of relaxing thermodynamic constraints and allowing metabolites to leave the network (all reactions in the metabolic network were made reversible, and all metabolites could leave the network) also did not result in feasible biomass production in the networks. Hence, the addition of reactions to all tested metabolic networks was necessary.

**Network Completion Requires Reactions with no ESS**—For all further gap-filling approaches, the reactions from the Model SEED gap-filling database (11,858 reactions; Table 1) were used as candidates for gap filling (Fig. 1). The database included a subset of curated transport reactions and had been thermodynamically constrained using the group contribution method (45). The sequences of the RAST-annotated genes of all organisms in the Model SEED database were extracted, and a
sequence database was generated for each gene. Using the RAST mapping between genes, functional roles, enzyme complexes, and reactions, an organism-specific weight (Equation 1) was calculated for each reaction of the gap-filling database (Fig. 1; see “Experimental Procedures” for details).

To test whether networks could be completed by restricting incorporation of reactions with a predefined level of support, reactions were divided into three tiers: highly supported reactions (ESS of 1.0E−240), significantly supported reactions (30) (ESS ≤ 1.0E −10), and unsupported reactions (ESS > 1.0E −10). For each tier, all reactions were added to the base models, and FBA was used to evaluate whether biomass could be produced. This revealed that no networks could be completed with only highly supported reactions or even significantly supported reactions (Table 2). This suggested that the tested networks required locally orphaned enzymes (no similarity to known enzymes in the organism) or globally orphaned enzymes (no known sequence) to produce all biomass components. Hence, orphaned metabolic functionality was integral to the core of metabolic networks and included reactions essential to biomass production. However, after releasing all thermodynamic constraints, including those in the gap-filling reactions and stoichiometric constraints caused by dead-end metabolites, networks containing only significantly supported reactions were able to produce all biomass components. The role of thermodynamic constraints in network completion was investigated in detail and will be reported in a specialist journal.

Non-producible Biomass Metabolites Are Distributed across Metabolism—The four tested organisms were unable to produce a significant portion of their biomass components, in each case spanning multiple classes of metabolites (Figs. 2 and 3). Some of the biomass components in S. pneumoniae, B. subtilis, and A. baylyi that were not producible in the base models were producible with the models that were augmented with the strongly supported reactions only. This suggested that the original networks were under-annotated. This was especially true for S. pneumoniae, which could not produce half of its biomass metabolites (39 of 79), yet 33 metabolites could be produced solely using significantly supported reactions (Fig. 3 and Table 2). Only 6–8 biomass metabolites could not be produced with only supported reactions in each organism. The ability of the augmented networks to produce many more biomass components than the base models, even if only the highly supported reactions were used, suggests that there was sufficient potential for ESS values to guide gap-filling solutions.

Two biomass components, acyl carrier protein and peptidoglycan polymers, could not be produced by any of the organisms. Spermidine and thiamine pyrophosphate (TPP) could also not be produced by any organism but were imported from the media by S. pneumoniae (Fig. 2). The Gram-positive bacteria S. pneumoniae and B. subtilis required the cell wall precursor glycerol teichoic acid. Acyl carrier protein, peptidoglycan polymers, calomide, and glycerol teichoic acid could be produced in isolation with significantly supported reactions if the biomass reaction was replaced by independent export reactions for all biomass components. However, acyl carrier protein, peptidoglycan polymers, and calomide biosynthesis were not required for total biomass production in the models because their precursors were regenerated by the biomass equation itself. All other non-producible metabolites are discussed in detail below.

Note that not all biomass components were necessarily essential; for instance, spermidine is part of the canonical E. coli biomass equation but may not be essential (5, 46). In some cases, genetic evidence may support the classification of a metabolite as essential if a sole pathway synthesizes the metabolite; riboflavin is one such example. In the biosynthesis of riboflavin in E. coli, all genes associated with riboflavin biosynthesis

---

**TABLE 2**

**Biomass components that require gap filling**

To investigate which biomass metabolites required gap-filling, FBA was used to maximize the export of each individual biomass component, given the stoichiometric and thermodynamic constraints imposed on the network. An exchange reaction was added for each biomass component that was tested, and FBA was used to maximize flux through each component exchange reaction in turn. Metabolites that could not be exported at a flux greater than a numerical cut-off of 1.0E−6 were considered non-producible. Production of the individual biomass components was attempted using gap-filling reaction sets with three different levels of support as well as the base models with no gap-filling reactions.

| Biomass metabolites producible with: | S. pneumoniae | B. subtilis | E. coli K-12 MG1655 | A. baylyi ADP1 |
|--------------------------------------|---------------|-------------|---------------------|---------------|
| Base model                           | 39            | 57          | 60                  | 47            |
| Strong support (ESS = 1E−240)        | 41            | 74          | 60                  | 51            |
| Some support (ESS < 1E−10)           | 71            | 76          | 66                  | 61            |
| All reactions in gap-filling database | 79            | 83          | 73                  | 67            |
Integrality of Orphaned Enzymes in Metabolic Networks

were experimentally essential (40), suggesting that riboflavin is indeed an essential biomass metabolite. It was surprising that the complete synthesis of riboflavin required unsupported reactions, despite the final steps in riboflavin synthesis being present in the metabolic network (also see below).

**BLAST-weighted Gap Filling**—A weighted LP problem was formulated to incorporate reactions into gap-filling solutions depending on their ESS. Each reaction in the gap-filling database was weighted inversely proportionally to the associated ESS scores (see “Experimental Procedures”). The improbability (approximated by 1 E-value) of a sequence similarity score occurring by chance was treated as the level of support for an enzyme activity existing in the network. This simplification was vulnerable to detecting false positives caused by strong similarity to a short sequence or domain only, but assessment at network level immunizes this approach to most effects of false positives. Hence, for an incorrect pathway to be included, all reactions in a pathway would have to be false positives. Treating support for a reaction as a probability, the support for a pathway was expected to scale with the product of the support of the underlying ESS values. Sequence support for the existence of a pathway of n reactions may then be described as the product of the ESS values for the individual reactions: \( II_n \). To avoid penalities against existing reaction annotations, reactions already included in the draft network received a weight of zero. The linear and quadratic programming objectives were minimized while requiring a set flux through the biomass reaction (see “Experimental Procedures”). Utilization of LP and QP made gap filling very fast. On a typical desktop computer, solutions were retrieved in seconds, compared with minutes for MILP.

**Quadratic Programming Reveals the Gap-filling Solution Space**—The QP formulation of the weighted gap-filling algorithm minimized the squared sum of weighted reaction fluxes. Squaring the weighted reaction fluxes limits large fluxes because penalties increase quadratically with flux. Conversely, small fluxes are penalized lightly, even if the associated weights are high. This results in the distribution of flux through alternative gap-filling solutions inversely proportional to the combined weights of reactions in a given pathway (Fig. 4). The number of reactions used in a QP gap-filling solution can thus provide a lower bound estimate on the number of reactions that can participate in gap-filling solutions. QP revealed that several thousand reactions could participate in gap-filling reactions for each organism (Fig. 5). Importantly, QP is not guaranteed to identify all potential gap-filling routes. Combinations of irreversible reactions and reaction weights can lead to hidden gap-filling reactions (Fig. 4). Removing reversibility constraints and adding random reaction weights allowed for an extreme estimate of gap-filling solutions in the gap-filling database. It was revealed that even more reactions could potentially participate in gap filling. For *E. coli*, this extreme QP solution included 7,337 reactions, almost double the number of reactions in the constrained QP solution.

The LP solutions were necessarily always contained in the QP solution for a given set of reversibility constraints and reaction weights. LP solutions that were based on uniform weights were mostly, but not always, contained in the QP solution (Figs. 4 and 5). The solutions of the Model SEED were more frequently outside the QP solution space (Fig. 5). The uniformly weighted LP solutions contained the lowest number of gap-filling reactions for the four tested networks and were probably the minimal reaction solutions in most cases. Strictly, the uniformly weighted LP solution is a minimal flux solution, making it imaginable that an alternative LP solution with fewer reac-
tions, but that carry more flux, may exist. In contrast, the weighted LP solutions often contained high flux reactions, but only if such reactions were associated with very low weights (Fig. 3). The weighted LP solution always contained substantially more reactions than either the uniformly weighted LP or the Model SEED gap-filling solutions, suggesting that a strong enough ESS signal existed to significantly influence gap filling. LP and Model SEED gap-filling solutions often shared several reactions, indicating that some biomass components may only be made producible in a limited number of ways, but no solutions were close to identical (Figs. 5 and 6).

The sheer size of the QP solution made it clear that many different gap-filling solutions existed (Fig. 6). 78% of the metabolites of the original network were used in the quadratic gap-filling solution of S. pneumoniae (Fig. 6), suggesting that there were no obvious metabolites that should serve as connecting metabolites to gap-filling reactions. With the size and level of connections in mind, it was surprising that gap-filling solutions shared reactions at all. Further investigation revealed that the shared reactions were often associated with high ESS values and sometimes represented a sole gap-filling solution to a subset of biomass components (Table 2 and supplemental Tables S1 and S2). The high ESS values of the shared reactions indicated that, in reality, alternative pathways to these biomass components may exist and highlighted the possible existence of missing biochemistry in the gap-filling database.

Comparison of Computational and Experimental Gene Essentiality—To investigate the quality of the gap-filling solutions, gene essentiality predictions from the gap-filled networks were compared with experimental data of full-length single
gene knock-out libraries. Gene deletions were simulated by removing all reactions that required a given gene. Gene essentiality was then predicted from the feasibility of biomass production from the specified media using FBA. Gene deletions that resulted in networks that could no longer produce biomass were considered computationally essential.

Networks that were gap-filled with weighted gap filling (referred to as BLAST LP) predicted gene knock-out outcomes better than networks filled by the Model SEED or uniformly weighted gap filling (Table 3). Weighted gap filling outperformed the alternatives methods both at the essential and non-essential gene predictions. The improved performance for both essential gene and nonessential gene predictions was striking because the weighted gap filling added significantly more reactions to networks, yet this did not result in fewer true essential gene predictions (except for in B. subtilis). More importantly, it suggested that the ESS signal was strong enough to enhance gap filling of draft metabolic networks, although all of the solutions included reactions associated with maximum ESS values. All genes associated with supported reactions in the BLAST LP gap-filling solutions for the E. coli network were consistent with RAST annotations. However, the additional functionalities associated with AceE and SucB (supplemental Table S1) were not supported by the literature (47) and were probably incorrect.

A Subset of Knock-out Predictions Are Sensitive to Weight Changes—Two sensitivity analyses were performed to investigate the robustness of the network support (NS) for the weighted gap-filling solutions and the influence of the gap-filling solutions on gene essentiality predictions. NS is here defined as how well a gap-filling reaction selection is determined by the entire network. NS for a reaction is calculated from the gap-filling penalty function increase after exclusion of that reaction from the gap-filling database. To test which gene essentiality predictions were sensitive to any gap-filling solution, 100 Monte Carlo gap-filling simulations of the E. coli network with randomly shuffled weights were calculated. This resulted in 97 genes with alternating essentiality prediction (supplemental Table S2), suggesting that a significant portion (9.1%) of gene predictions were sensitive to gap filling.

In a second sensitivity analysis, weights were shifted by a small random amount from the sequence-derived weights to test sensitivity to variations in the ESS calculation. Only 18 genes changed predicted essentiality status over 100 runs (supplemental Table S2). This suggested that the sequence-based gap-filling approach was fairly robust to variations in the BLAST sequence comparisons. The number of genes changing essentiality prediction was fairly insensitive to the magnitude of the added noise, ranging from a normal distribution centered at zero, with an S.D. from 0.1 to 10 units (reaction weights scaled between 0 and 553), indicating that most essentiality predictions were well determined by the gap-filling approach.

The BLAST LP optimal gap-filling solution utilized 15 reactions, eight of which were present within all shifted weight Monte Carlo gap-filling solutions. The additional reactions varied substantially over the Monte Carlo runs, but the number of reactions that were featured at least once in the gap-filling solutions was insensitive to the magnitude of the noise (supplemental Table S2). Of the eight reactions that were always retrieved, four had minimal ESS values, and four had maximum ESS values, including one mandatory reaction for which no alternative existed. Removal of any of the eight reactions that were always included resulted in solutions with substantially higher objectives, indicating strong NS and explaining their consistent inclusion. 17 of the 18 genes with variable essentiality predictions were essential in the noiseless solution. Remarkably, 15 of these 17 computationally essential predictions were wrong, which was on par with random predictions, considering that only 10% of genes were experimentally essential. Note that overall, 70% of the experimentally essential genes were correctly predicted as essential. However, of the computationally essential genes, only 44.6% were experimentally essential. Disregarding the 18 genes with alternating essentiality calls improved the latter statistic to 48.3%.

 Combined, these results indicated that the ESS signal was strong enough to determine >80% of otherwise variable essentiality predictions. The seven gap-filling reactions for which the inclusion was sensitive to reaction weights determined the 18 gene essentiality calls that were of very poor quality (supplemental Table S2). Therefore, the implied presence of orphaned enzymes in all networks did not nullify the ability to find meaningful gap-filling solutions, but the poorly determined reactions significantly deteriorated a subset of essentiality calls.

Analysis of Gap-filling Reactions with High ESS Values—A subset of the metabolites that could not be produced by significantly supported reactions still required unsupported reactions for production after breaking the biomass equation into independent export reactions. These unsupported metabolites
were investigated in more detail by using the BLAST LP algorithm on the gap-filled networks to calculate flux utilization of the gap-filling reactions for unsupported metabolites (Fig. 7).

The two reactions required by *E. coli* for riboflavin, FAD, and TPP synthesis had strong NS values and were therefore always included in shifted weight sensitivity analysis. The remaining reaction associated with spermidine synthesis was included in 71 of 100 solutions. This suggested that these reactions were strongly determined by the gap-filling approach. In contrast, only the reaction associated with riboflavin and FAD was always included in the shuffled weight sensitivity analysis because no alternative reaction was present in the gap-filling database. The reactions for TPP and spermidine synthesis were only included in 4 and 24 cases of 100 in the shuffled weight solutions, respectively. This suggests that despite the lack of ESS support for these reactions, NS strongly determined gap-filling reactions among many other poor alternatives.

Two reactions implicated by NS were supported by circumstantial evidence. The gap-filling export reaction for TPP may be through spontaneous diffusion due to the chemical properties of 4-hydroxy-benzyl alcohol, a byproduct of TPP synthesis. Riboflavin and FAD required a reaction for which no alternative existed in the Model SEED biochemistry database. This reaction has been hypothesized in the literature, and only recently, a gene in *E. coli* has been associated with the activity (48).

**Discussion**

Draft metabolic networks of four species were investigated for the ability to produce a complete set of biomass metabolites. The observed inability of networks to produce biomass, even after the removal of thermodynamic and stoichiometric constraints caused by dead-end metabolites, necessitated the addition of gap-filling reactions. Although each network could be readily filled using the Model SEED biochemistry database, no networks could be filled solely with reactions that were supported by sequence similarity to known enzymes. The need for orphaned enzymes implied that all metabolic networks were missing essential biochemistry annotations. Possibly, these reactions are of unknown biochemistry, suggesting fundamental gaps in our biochemistry knowledge for even the best-studied organisms. This realization suggests that our biochemistry knowledge or inclusion of this knowledge in the database, rather than the quality of machine annotations, is limiting our ability to further improve automated network reconstructions. Note that given the very small flux requirement through unsupported reactions (Fig. 3), it is conceivable that some of the orphaned activities may be attributed to secondary catalytic activity of promiscuous enzymes.

The presence of orphaned enzymes in gap-filling solutions and the very large size of the solution spaces, made evident by the quadratic programming, prompted the question of how robust the gap-filling solutions were in response to noise and to what extent gene essentiality predictions were influenced by gap-filling solutions. One hundred repeated gap-filling runs using randomly shuffled weights for the *E. coli* network showed that a substantial number of reactions could be part of the gap-filling solution, which resulted in many alternate gene essentiality assignments (supplemental Table S2). However, in response to noise added to the correct weights, a much smaller subset of genes showed alternating gene essentiality. This suggested that many gene essentiality predictions sensitive to the gap-filling solutions were strongly determined by the sequence-derived weights. Additionally, eight gap-filling reactions were always present in gap-filling solutions, suggesting that they were strongly determined by NS. Interestingly, the essentiality of the group of genes sensitive to gap filling was predicted very poorly, which suggested that the fallout of the partially arbitrary gap-filling process due to a simplified relationship between E-values and ESS as well as the addition of orphaned enzymes may be limited to a small subset of gene essentiality predictions.

LP- and QP-based gap-filling algorithms generated fast and meaningful gap-filling solutions. LP optimization resulted in gap-filled networks that better predicted gene essentiality compared to networks that were filled with existing gap-filling technology. The large majority of gap-filling reactions was supported by sequence similarity and had often been identified by RAST, yet these reactions had not been included in the Model SEED draft models. The fairly insignificant computational time to establish ESS values (2 h/organism on a quad core Intel i3 desktop computer) should be well worth the effort, although the network quality improvement may be modest. This is par-

### Table: Unsupported metabolite gap-filling reactions

| Unsupported metabolite | Unsupported gap-filling reaction | Affected organisms |
|------------------------|----------------------------------|--------------------|
| Riboflavin (RIBF)      | rxn05039                         | E. coli, A. baylyi |
|                        | 5-Amino-6-5-phosphoribitylaminouracil + H2O |                     |
|                        | -> Phosphate + 4-1-D-Ribitylamino-5-aminouracil + H+ |                     |
| The enzyme 5-amino-6-(5-phospho-D-ribitylamino)uracil phosphatase was recently discovered in *E. coli* (48) and was therefore not included in the Model SEED biochemistry database. |
| FAD                    | rxn05039                         | E. coli, A. baylyi |
|                        | 5-Amino-6-5-phosphoribitylaminouracil + H2O |                     |
|                        | -> Phosphate + 4-1-D-Ribitylamino-5-aminouracil + H+ |                     |
| The same gap-filling reaction identified for riboflavin was also selected for FAD. Riboflavin is a precursor to FAD synthesis, and this result suggests that only the precursor metabolism was missing. |
| Spermidine             | rxn00125                         | E. coli, A. baylyi, B. subtilis |
|                        | H2O + S-Adenosyl-L-methionine    |                     |
|                        | <= L-Homoserine + H+ + MTA       |                     |
| The activity of adenosylmethionine hydrolase has been observed in soil bacteria (49), but no associated gene was cataloged in the Model SEED, KEGG (50) (accessed May 7, 2015), or Biocyc (51) (accessed May 7, 2015) databases. |
| Spermidine             | rxn13085                         | A. baylyi           |
|                        | Putrescine + S-Adenosyl-L-methionine |                     |
|                        | <= CO2 + MTA + Spermidine        |                     |
| The *A. baylyi* model was missing the spermidine synthesis step. The *A. baylyi* genome had no significant similarity to genes in the Model SEED database that could fill this role. Spermidine synthesis may occur through a mechanism that does not have homology with known synthesis pathways. Alternatively spermidine may be incorrectly included in the biomass equation of *A. baylyi*. |
| TPP                    | rxn12376                         | E. coli, A. baylyi, B. subtilis |
|                        | 4-Hydroxy-benzylalcohol          |                     |
|                        | <= 4-Hydroxy-benzylalcohol       |                     |
| 4-Hydroxy-benzyl alcohol transport does not contain a gene association in the Model SEED, KEGG, or Biocyc database. However, the ChemSpider entry for 4-Hydroxy-benzyl alcohol (http://www.chemspider.com/Chemical-Structure.122.html, accessed 14:52, May 7, 2015) lists a very high water solubility and large octane/water partition coefficient, which suggests that it may readily passively traverse membranes. |
particularly true for the inclusion of BLAST LP in network reconstruction pipelines.

This work demonstrated that orphaned enzymes were integral to essential metabolic functions and that a fully supported and functionally complete metabolic network could not be assembled even with the extensive compilation of enzymes and biochemistry from RAST and the Model SEED. Nonetheless, sequence similarity-driven gap filling improved the quality of the networks and identified deficiencies in our biochemistry knowledge. The large set of significantly supported gap-filling reactions in all gap-filling solutions showed the potential for network-based identification of candidate gene annotations. Nonetheless, truly realistic models will probably require further expansion of the Model SEED biochemistry database or the discovery of not yet observed metabolic reactions and their gene associations.

Glossary

Metabolic Network—Relational structure of interconnected chemical reactions that constitute metabolism.

Functional Role—Activity that an individual protein performs as part of an enzyme complex.

Model SEED—Online automated metabolic reconstruction service for prokaryotes.

Locally Orphaned Enzyme—Enzyme activity that is observed or predicted to occur in an organism but for which no gene has been discovered in that organism.

Globally Orphaned Enzyme—Enzyme activity that is observed or predicted to occur in biology but with which no gene has been associated.

Stoichiometric Matrix—Matrix representing the stoichiometry of the metabolites in the reactions of a metabolic network.

Metabolic Network Reconstruction—Systematic representation of enzyme-catalyzed reactions that occur in an organism.

Genome-scale Metabolic Model—Comprehensive metabolic model that includes reactions for all enzymatic genes.

Draft Metabolic Network—A network assembled from gene annotations, often created algorithmically.

Gene Essentiality—Binary property of a gene that is specific to a given environmental condition. A gene is essential if the model requires that gene to produce biomass.

Dead-end Metabolite—Metabolite that is either not consumed or produced by any reaction in a metabolic network.

Isolated Reaction—Reaction with at least one product AND substrate that do not participate in another reaction.

Blocked Reaction—Reaction with at least one product OR substrate that does not participate in another reaction.

Gap—Reaction(s) missing from a network reconstruction that result(s) in isolated or blocked reactions.

Gap-filling Solution—Set of reactions that, upon addition to a network, allow synthesis of a set of target metabolites from a defined nutrient source.

Top-down Gap Filling—Addition of the full reaction database to a draft network followed by pruning of reactions until a minimal gap-filling solution is obtained.

Bottom-up Gap Filling—Iterative addition of reactions to a metabolic network until a gap-filling solution is obtained.

Optimization-based Gap Filling—Gap-filling approach that aims to minimize or maximize a gap-filling objective through numerical optimization.

Linear Programming (LP)—Maximizing or minimizing of a (multivariate) linear equation given specified linear constraints and parameter bounds.

Quadratic Programming (QP)—Maximizing or minimizing of a (multivariate) quadratic equation given specified linear constraints and parameter bounds.

Mixed Integer Linear Programming (MILP)—Linear programming method where some parameters must be integers.

Weighted Linear Programming—Linear programming with weight coefficients for each linear equation.

Weighted Quadratic Programming—Quadratic programming with weight coefficients for each quadratic equation.

Flux Balance Analysis (FBA)—Linear programming method that calculates a flux solution optimized for an inferred metabolic function, such as maximum biomass yield.

Enzyme Sequence Support (ESS)—Measure of similarity between known enzymes that catalyze a reaction and the best matched sequence region in the target genome.

Network Support (NS)—Network-wide measure of support for the inclusion of a gap-filling reaction. NS for a reaction is determined from the gap-filling penalty function increase in the absence of that reaction.

Author Contributions—E. W. K. and I. G. L. L. designed and performed the experiments, analyzed the data, and wrote the manuscript.

Acknowledgments—We gratefully acknowledge Chris Henry for facilitating the use of the Model SEED gap-filling database and assembled gene essentiality data. We are grateful for the insightful feedback from the reviewers.

Note Added in Proof—The current Fig. 7 originally appeared as Table 4 in the version of the article published as a Paper in Press on June 3, 2015.

References

1. Thiele, I., and Palsson, B. Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat. Protoc. 5, 93–121
2. Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D., and Palsson, B. Ø. (2012) Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. PLoS One 7, e5635
3. Oberhardt, M. A., Palsson, B. Ø., and Papin, J. A. (2009) Applications of genome-scale metabolic reconstructions. Mol. Syst. Biol. 5, 320
4. Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biol. 4, R54
5. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbent, L. J., Hatzimanikatis, V., and Palsson, B. O. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3, 121
6. Orth, J. D., and Palsson, B. Ø. (2012) Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions. BMC Syst. Biol. 6, 30
7. Henry, C. S., Overbeek, R., Xia, F., Best, A. A., Glass, E., Gilbert, J., Larsen, P., Edwards, R., Disz, T., Meyer, F., Vonstein, V., Dejongh, M., Bartels, D., Desai, N., D’Souza, M., Devo, S., Keegan, K. P., Olson, R., Wilke, A.,...
Integrity of Orphaned Enzymes in Metabolic Networks

Wilkening, J., and Stevens, R. L. (2011) Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochim. Biophys. Acta.* 1810, 967–977.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr., Assad-Garcia, N., Glass, J. L., and Covert, M. W. (2012) A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401.

Reed, J. L. (2012) Shrinking the metabolic solution space using experimental datasets. *PLoS Comput. Biol.* 8, e1002662.

Berestovsky, N., Zhou, W., Nagrath, D., and Nakhleh, L. (2013) Modeling integrated cellular machinery using hybrid Petri-Boolean networks. *PLoS Comput. Biol.* 9, e1003306.

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Ho, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bölting, C., Bordel, S., Chavalí, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Huka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, I. J., Jury, N., Keating, S., Noskow, I., Le Noé, N., Malys, N., Mazin, A., Papin, J. A., Price, N. D., Selkow, E. S., Sigurdsson, M. I., Simeonis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H. M., Weichert, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Koll, D. B., Mendes, P., and Palsson, B. Ø. (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425.

Heavner, B. D., Smallbone, K., Barker, B., Mendes, P., and Walker, L. P. (2012) Yeast 5: an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Mol. Syst. Biol.* 6, 55.

Dreyfuss, J. M., Zucker, J. D., Hood, H. M., Ocasio, L. R., Sachs, M. S., and Galagan, J. E. (2013) Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM. *PLoS Comput. Biol.* 9, e1003126.

Vitkin, E., and Shlomi, T. (2012) MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biol.* 13, R111.

Henry, C. S., Delongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982.

Feng, X., Xu, Y., Chen, Y., and Tang, Y. J. (2012) MicrobesFlux: a web platform for drafting metabolic models from the KEGG database. *Mol. Syst. Biol.* 6, 94.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. I., Diz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* 42, D206–D214.

Aziz, R. K., Bartels, D., Best, A. A., Delongh, M., Diz, T., Edwards, R. A., Formsmma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. I., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Pacciazz, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008) The RAST Server: rapid annotations using subsystems technology. *GBMC Genomics* 9, 75.

Konwar, K. M., Hanson, N. W., Pagé, A. P., and Hallam, S. J. (2013) MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. * BMC Bioinformatics* 14, 202.

Meyer, F., Paarmann, D., D’Souza, M., Olson, R., Glass, E. M., Kubal, M., Pacciazent, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. A. (2008) The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes. *Mol. Syst. Biol.* 4, 386.

Reed, J. L., and Palsson, B. Ø. (2003) Thirty years of building constraint-based in silico models of *Escherichia coli* J. Bacteriol. 185, 2692–2699.

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.* 11, 40–79.

Feist, A. M., and Palsson, B. Ø. (2010) The biomass objective function. *Curr. Opin. Microbiol.* 13, 344–349.

Orth, J. D., and Palsson, B. Ø. (2010) Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.* 107, 403–412.

Satish Kumar, V., Dasika, M. S., and Maranas, C. D. (2007) Optimization based automated curatin of metabolic reconstructions. *BMC Bioinformatics* 8, 212.

Fleming, R. M. T., Thiele, I., and Nasheuer, H. P. (2009) Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to Escherichia coli. *Biophys. Chem.* 145, 47–56.

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., and Palsson, B. Ø. (2006) Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 17480–17484.

Zomorodi, A. R., Suthers, P. F., Ranganathan, S., and Maranas, C. D. (2012) Mathematical optimization applications in metabolic networks. *Metab. Eng.* 14, 672–686.

Krumholz, E. W., Yang, H., Weisenhorn, P., Henry, C. S., and Libourel, I. G. L. (2012) Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *J. Exp. Bot.* 63, 2353–2362.

Christian, N., May, P., Kempa, S., Handorf, T., and Ebenböh, O. (2009) An integrative approach towards completing genome-scale metabolic networks. *Mol. Biosyst.* 5, 1889–1903.

Bolival, B., Jr., Assad-Garcia, N., Glass, J. I., and Covert, M. W. (2012) A meta-analysis of genome-scale metabolic models of cyanobacteria networks. *J. Bacteriol.* 194, 1889–1903.

Orth, J. D., Palsson, B. Ø., and Thiele, I. (2011) The human metabolic reconstruction Recon 1 directs hypotheses of novel human metabolic functions. *Mol. Syst. Biol.* 5, 155.

Kol, D. B., Mendes, P., and Palsson, B. Ø. (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* 31, 419–425.
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., Rawlins, J. P., Rivas, L. A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H. H., Scanlan, E., Schumann, W., Seegers, J. F. M. L., Sekiguchi, J., Sekowska, A., Séror, S. I., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H. B., Vagner, V., van Dijl, J. M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U., and Ogasawara, N. (2003) Essential Bacillus subtilis genes. Proc. Natl. Acad. Sci. U.S.A. 100, 4678–4683

43. Thanassi, J. A., Hartman-Neumann, S. L., Dougherty, T. J., Dougherty, B. A., and Pucci, M. J. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in Streptococcus pneumoniae. Nucleic Acids Res. 30, 3152–3162

44. Yang, H., Krumholz, E. W., Brutinel, E. D., Palani, N. P., Sadowsky, M. J., Odlyzko, A. M., Grahnich, J. A., and Libourel, I. G. L. (2014) Genome-scale metabolic network validation of Shewanella oneidensis using transposon insertion frequency analysis. PLoS Comput. Biol. 10, e1003848

45. Jankowski, M. D., Henry, C. S., Broadbelt, L. J., and Hatzeimanakis, V. (2008) Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys. J. 95, 1487–1499

46. Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S. A., Palsson, B. Ø., and Agarwalla, S. (2006) Experimental and computational assessment of conditionally essential genes in Escherichia coli. J. Bacteriol. 188, 8259–8257

47. Bi, H., Bai, Y., Cai, T., Zhuang, Y., Liang, X., Zhang, X., Liu, T., and Ma, Y. (2013) Engineered short branched-chain acyl-CoA synthesis in E. coli and acylation of chloramphenicol to branched-chain derivatives. Appl. Microbiol. Biotechnol. 97, 10339–10348

48. Haase, I., Sarge, S., Illarionov, B., Laudert, D., Hohmann, H. P., Bacher, A., and Fischer, M. (2013) Enzymes from the haloacid dehalogenase (HAD) superfamily catalyse the elusive dephosphorylation step of riboflavin biosynthesis. ChemBioChem 14, 2272–2275

49. Mazelis, M., Levin, B., and Mallinson, N. (1965) Decomposition of methyl methionine sulfonium salts by a bacterial enzyme. Biochim. Biophys. Acta 105, 106–114

50. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–D205

51. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Kesseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, I. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 42, D459–D471