BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 3 2015, pages 373–381
doi:10.1093/bioinformatics/btu658

Systems biology

Advance Access publication October 5, 2014

Computing autocatalytic sets to unravel inconsistencies in metabolic network reconstructions

Ralf Schmidt[1,†], Silvio Waschina[1,2,†], Daniela Boettger-Schmidt[3], Christian Kost[2] and Christoph Kaleta[1,4,*]

[1]Research Group Theoretical Systems Biology, Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, 07743 Jena, [2]Department of Bioorganic Chemistry, Experimental Ecology and Evolution Research Group, Max Planck Institute for Chemical Ecology, 07745 Jena, [3]Department of Biomolecular Chemistry, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, 07745 Jena and [4]Department of Computational Biology, Institute for Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense M, Denmark

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Genome-scale metabolic network reconstructions have been established as a powerful tool for the prediction of cellular phenotypes and metabolic capabilities of organisms. In recent years, the number of network reconstructions has been constantly increasing, mostly because of the availability of novel (semi-)automated procedures, which enabled the reconstruction of metabolic models based on individual genomes and their annotation. The resulting models are widely used in numerous applications. However, the accuracy and predictive power of network reconstructions are commonly limited by inherent inconsistencies and gaps.

Results: Here we present a novel method to validate metabolic network reconstructions based on the concept of autocatalytic sets. Autocatalytic sets correspond to collections of metabolites that, besides enzymes and a growth medium, are required to produce all biomass components in a metabolic model. These autocatalytic sets are well-conserved across all domains of life, and their identification in specific genome-scale reconstructions allows us to draw conclusions about potential inconsistencies in these models. The method is capable of detecting inconsistencies, which are neglected by other gap-finding methods. We tested our method on the Model SEED, which is the largest repository for automatically generated genome-scale network reconstructions. In this way, we were able to identify a significant number of missing pathways in several of these reconstructions. Hence, the method we report represents a powerful tool to identify inconsistencies in large-scale metabolic networks.

Availability and implementation: The method is available as source code on http://users.minet.uni-jena.de/~m3kach/ASBIG/ASBIG.zip.

Contact: christoph.kaleta@uni-jena.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 17, 2014; revised on September 23, 2014; accepted on September 30, 2014

1 INTRODUCTION

In recent years, genome-scale metabolic network reconstructions have become an important tool in systems biology (Thiele and Palsson, 2010). They have the strong potential to combine distinct experimental data with bibliomic resources to generate a comprehensive knowledge base (Feist and Palsson, 2008; Wessely *et al.*, 2011). The resulting network reconstructions have been widely used to simulate metabolic processes or to explore the metabolic capabilities of various species (Kaleta *et al.*, 2011; Ruppin *et al.*, 2010). A prominent approach applying network reconstructions is constraint-based modeling, such as flux balance analysis (FBA) (Lewis *et al.*, 2012; Ruppin *et al.*, 2010). The use of metabolic models and associated methods has granted access to diverse scientific subjects, such as analysis of the bacterial metabolism (Yus *et al.*, 2009), the prediction of growth rates of *Escherichia coli* (Varma and Palsson, 1994), the comparison of growth rates between wild type and mutant strains of *E.coli* (Segrè *et al.*, 2002) and metabolic engineering (Wang *et al.*, 2006).

Until recently, the process of network reconstruction was time-consuming and necessarily required laborious curation effort (Durot *et al.*, 2009). To cope with the increasing amount of available data, automated methods became necessary to produce high-throughput reconstructions as well. Several approaches are available to address this issue (Barua and Reed, 2010; Jerby *et al.*, 2010; Li *et al.*, 2010). The Model SEED project successfully implemented one of these approaches and contains >190 metabolic reconstructions, which were primarily generated in an automated way (Henry *et al.*, 2010).

However, the automated process may neglect metabolic capabilities of the focal organism or include wrongly identified functionalities (Orth and Palsson, 2010). To minimize this bias, manual refinement and optimization are instrumental in the reconstruction process (Henry *et al.*, 2010; Thiele and Palsson, 2010), representing the most intricate elements in the workflow. To accelerate model curation, automated and semi-automated methods for the detection and correction of gaps were developed and integrated in the automated reconstruction process by the Model SEED (GapFind and GapFill) (Henry *et al.*, 2010; Satish Kumar *et al.*, 2007). The most common approach used to identify gaps in network reconstructions is constraint-based

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

modeling that relies on optimization-based algorithms. The predominant aim is to detect metabolites that cannot be produced at a steady state (Latendresse *et al.*, 2012; Ponce-de Leon *et al.*, 2013; Satish Kumar *et al.*, 2007).

Here we report ASBIG (Autocatalytic Set-Based Identification of Gaps), a method that detects incomplete parts of network reconstructions based on a novel approach: identifying elements (compounds) of catalytic cycles. ASBIG screens models for essential, self-replicating metabolites, which are pivotal elements of the underlying metabolism. A metabolite is considered to be pivotal if it is required for metabolism to proceed. The compounds are called 'self-replicating', or 'autocatalytic', as they are usually required for their own biosynthesis (Kun *et al.*, 2008). Hence, their production is inaccessible unless an initial amount of the compound is already available. Self-replicating metabolites are energy-currency cofactors, such as ATP and NAD. If these compounds are not present within the cell, even a rich nutritional medium is often insufficient for an organism to produce all of its biomass components. The definition of self-replicating metabolites originates from the work of Eigen and Schuster (1977), in which the concepts of catalytic cycles and autocatalysis were described on a molecular level. The search for autocatalytic compounds is conducted via the method of scope analysis (Handorf *et al.*, 2005).

The set of self-replicators in each network reconstruction is generally small (Kun *et al.*, 2008). Furthermore, autocatalytic sets of different organisms very likely contain similar well-conserved compounds. These observations can be used to survey any given metabolic network to detect inherent inconsistencies. If the number of self-replicating compounds is comparably large or unexpected elements and large macromolecules are included in the set, conclusions about possible inconsistencies can be drawn. Hence, gaps in the reconstructed network can lead to the presence of unexpected compounds in the set of self-replicators (owing to the impaired biosynthesis of this compound or its successors).

ASBIG enriches the tool box of gap-finding procedures, as common flux calculating methods are susceptible to overlook gaps associated with autocatalytic cycles. One simple example is a cofactor that is required for the production of a biomass component but itself is not present in the biomass reaction. In such a case, it would be sufficient for constraint-based methods, such as FBA, if reactions that replenish the cofactor exist to produce biomass. A biosynthetic route of this essential cofactor would not be required making the detection of a lack of this route impossible for FBA. ASBIG uses scope analysis to determine such inconsistencies. Although this approach disregards the exact stoichiometric properties of metabolic reactions, our method is not susceptible to problems known to occur when using purely topology-based procedures (de Figueiredo *et al.*, 2008).

To outline the benefit of ASBIG, selected reconstructions from the following organisms were examined: *E.coli*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Chlamydomonas reinhardtii*, and *Zea mays* (Table 1). Using our method, we were able to identify missing pathways in all of these networks except for the *E.coli* reconstruction. Subsequently, the detected inconsistencies were resolved using information from the KEGG database (Kanehisa *et al.*, 2006) and other resources, like BioCyc, BsubCyc (Caspi *et al.*, 2010) or the Plant metabolic pathway database (Zhang *et al.*, 2010). Furthermore, all available reconstructions of the Model SEED project (Henry *et al.*, 2010) have been screened to demonstrate the applicability of the method on a large scale. As one substantial benefit of this screening, common inconsistencies widespread across numerous models were identified. Additionally, ASBIG detected missing reactions that were deleted manually from the model of *E.coli* for validation purposes, demonstrating the reliability and applicability of our method.

In summary, ASBIG is a reliable method for the detection of inconsistencies in any metabolic network reconstruction. It provides an efficient approach to validate metabolic models. By pinpointing discrepancies of network reconstructions, the method supports the improvement of the high number of incomplete models generated by high-throughput methods. Thus, ASBIG can significantly contribute to improve the quality of metabolic network reconstructions.

2 MATERIAL AND METHODS

2.1 Models

ASBIG uses a genome-scale reconstruction of metabolic networks as input. It was applied to investigate the following models: iJO1366 of *E.coli* K-12 MG1655, *i*BSU1103 of *B.subtilis* 168, AraGEM of *A.thaliana*, iRC1080 of *C.reinhardtii*, iMM904 of *S.cerevisiae* and DTMaize_C4GEM_45632 of *Z.mays*. Additionally, >190 models provided by the Model SEED project (Henry *et al.*, 2010) (as available in October 2012) were screened.

2.2 Identification of autocatalytic metabolic sets

The algorithm was implemented in JAVA. Genome-scale metabolic networks are processed in the standardized Systems Biology Markup Language using the JAVA package jigcell.sbml2 (Vass *et al.*, 2004).

2.3 Initial set of metabolites (seed set)

A predefined set of metabolites, named 'initial seed set', acts as the starting point for ASBIG. It combines the components of a nutritional medium with additional pivotal elements. The set of additional elements is composed of two subsets: metabolites identified as crucial by ASBIG (see Section 3.1) and autocatalytic compounds whose biosynthesis depends on their own presence (cf. Kun *et al.*, 2008). These were as follows: Adenosine triphosphate (ATP), nicotinamide adenine dinucleotide (NAD) and coenzyme A (CoA).

Hence, the composition of the initial seed set is assumed to include all essential components necessary for the growth of the organism.

2.4 Scope analysis

ASBIG uses the concept of scopes (Handorf *et al.*, 2005), which can be determined for a given sets of metabolites. The scope is a (large) set of metabolites and corresponds to all compounds that can, in principle, be produced from a (small) predefined set of metabolites. The concept of scope analysis relies on three aspects (Handorf *et al.*, 2005) (i) a reaction is considered to take place if all of its substrates (or products in the case of reversible reactions) have non-zero concentrations, (ii) products of one reaction are immediately considered as potential substrates of another one and (iii) starting with a small set of metabolites, iteratively increasing sets of reactions (and metabolites) are generated by screening the model for further reactions with a non-zero flux. The fundamental principle

Table 1. Closely investigated models

Organism	Model
E.coli K-12 MG1655	iJO1366 (Orth *et al.*, 2011)
B.subtilis 168	*i*BSU1103 (Henry *et al.*, 2009)
A.thaliana	AraGEM (de Oliveira Dal'Molin *et al.*, 2010b)
C.reinhardtii	iRC1080 (Chang *et al.*, 2011)
S.cerevisiae	iMM904 (Mo *et al.*, 2009)
Z.mays	DTMaize_C4GEM_45632 (de Oliveira Dal'Molin *et al.*, 2010a)

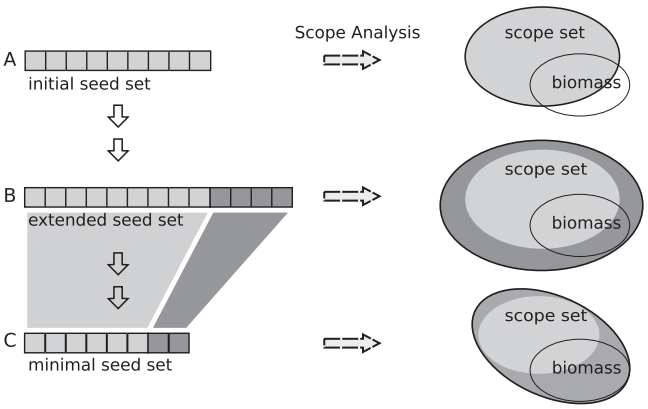

Fig. 1. Workflow of ASBIG. (**A**) Based on an initial seed set of metabolites (light gray squares), a scope set is calculated using scope analysis (this scope set might not include all biomass compounds). (**B**) The initial seed set is extended with add-on metabolites (dark gray squares) leading to a larger scope set (dark gray circle), until the corresponding scope set contains all biomass compounds. (**C**) The extended seed set is reduced to the smallest possible seed set that still contains all biomass components (final minimal seed set)

of ASBIG relies on the calculation of a scope set based on a given seed set of metabolites.

2.5 Biomass reaction

An essential feature among most network reconstructions is the biomass reaction, which is crucial for the ASBIG method. During a run of ASBIG, this reaction is used as a benchmark to evaluate generated metabolite sets. The biomass reaction constitutes a hypothetical reaction within the model that includes all metabolites necessary for the growth of the corresponding organism as substrate (biomass compounds). Usually, it is an essential element of FBA (Feist and Palsson, 2010). For ASBIG, it is reasonable to use this reaction as benchmark and compare metabolite sets based on their ability, to provide access to the biomass compounds. A seed set cannot be considered as adequate for the corresponding organisms until all biomass compounds are within the scope of the seed set.

2.6 Expanded seed set

Given the initial seed set, not all biomass compounds are necessarily part of the computed scope. Hence, the initial seed set is insufficient and needs

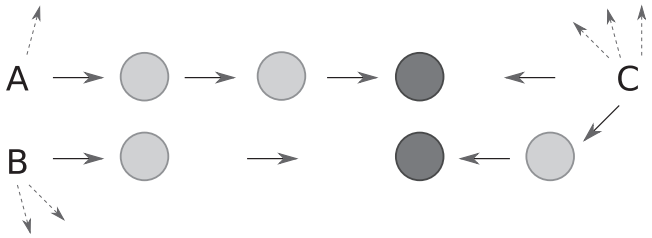

Fig. 2. Scenario to illustrate the necessity of a random approach to minimize the extended seed set. A and B represent initial metabolites, whereas C is an add-on compound. Light gray circles symbolize intermediates of the depicted pathways, and the dark gray circles represent biomass compounds. Dashed arrows hint further reaction routes. A and B together lead to the same biomass components as the availability of C. If, for example, A is deleted in the second phase of ASBIG, B alone lacks the potential to maintain the producibility of the two biomass compounds. Consequently, C would emerge in the final minimal seed set

to be expanded to generate a scope comprising all biomass components. Iteratively, metabolites, named '*add-on metabolites*', are added using a greedy approach. In each iteration, each metabolite that is not within the previously calculated scope is considered and inserted into the extended seed set on a trial basis. Finally, the compound, which results in the biggest increase of scope size represents the add-on metabolite of this iteration and remains in the extended seed set (independent of its impact on the biomass function). To this end, a scope analysis has to be performed for each metabolite, which was not part of the previously computed scope. If two metabolites yield an identical increase in scope size, the smaller metabolite in terms of the number of carbon atoms is chosen. Subsequently, a scope analysis is conducted with the temporary expanded seed set. The expansion process stops when the current expanded seed set results in a scope set that comprises all biomass compounds (Fig. 1 A and B).

2.7 Determining the minimal seed set

Most likely, not every metabolite of the expanded seed set contributes significantly to biomass production. For this reason, we aimed to determine a minimal seed set in terms of size during the second phase of ASBIG (Fig. 1C). To minimize the size of the expanded seed set, each element is removed on trial, and the scope of the remaining seed set is computed. If all biomass compounds are still within the scope, the element remains removed, otherwise it is reinserted. Potential dependencies between compounds of the expanded seed set impede a deterministic strategy to remove elements from the expanded seed set. Alternatively, a strategy accomplishing randomized removal of seed compounds is applied—with respect to one constraint: Elements of the initial seed set (depicted as light gray squares in Fig. 1) were prioritized and were only examined once all other add-on metabolites (dark gray squares in Fig. 1) have been tested (random removal with priorities). Unconstrained randomized removal can be critical, as the following example illustrates (Fig. 2): Three metabolites, A, B, and C, are included in the expanded seed set. A and B are initially given compounds, whereas C was appended as add-on metabolite. C, self-evidently, enlarges the scope size. However, the enlargement may not affect the biomass production (Fig. 2A and B together unlock the same biomass compounds as the presence of C) In other words, C constitutes an add-on metabolite, because of its impact on the scope size, but C is not necessary for biomass production. However, in case of an unconstrained random approach to calculate the minimal seed set, it is two times more likely that C remains in the minimal seed set: if and only if C is deleted first, A and B remain in the minimal seed set. In the two remaining cases of primary deletion of A or B, C will

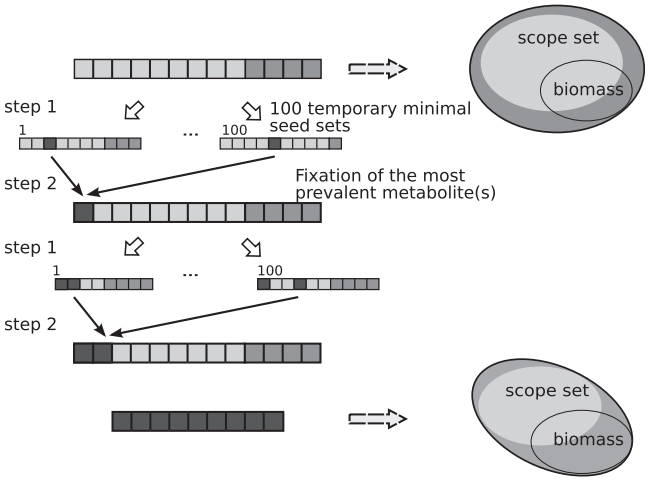

Fig. 3. Visualization of the minimal seed set generation. Beginning with the extended seed set, the method computes 100 (potentially different) minimal seed sets with the constrained random procedure as described. Out of these, the most prevalent compound(s) (depicted in black) are marked as inherent part of the emerging minimal seed set and are not considered in following iterations. Within these, the procedure is repeated until the set of fixed elements is capable of producing a scope that comprises all biomass components (bottom part)

be part of the minimal seed set to ensure the accessibility of the relevant biomass compounds. This would be in contrast to the basic principle of ASBIG to retain the initial metabolites. To minimize this bias, all components of the initial seed set are prioritized.

This illustrates the problem of dependencies between different compounds of the seed set, which, of course, can occur between any subset of metabolites.

To account for such properties of the expanded seed set, the actual implementation comprises two iteratively repeated steps: During the first step, a distinct number of runs of random removal with priorities is performed to generate a pool of potential minimal seed sets (possibly, all potential minimal seed sets are different). For the analyses presented here, 100 runs were performed in one cycle. In the second step, the temporary minimal seed sets are evaluated and the metabolite(s) with the maximal number of occurrences is (are) marked as definite element(s) of the final minimal seed set (marked in black in step 2 of Fig. 3). Subsequently, fixed metabolites are not removed on trial anymore in the following iterations. Instead, they constitute an invariant part of every temporary minimal seed set once they have been fixed. In the next iteration, only the remaining compounds of the expanded seed set are removed on trial. The procedure is repeated until the set of fixed metabolites is capable of generating a scope that includes all biomass compounds.

To obtain more information about the reasons why a specific metabolite occurs in the seed set, it can be beneficial to repeat the workflow using a slightly different approach (‘*producibility constraint*’). The difference to the default procedure is an additional constraint during seed set expansion restricting the possibility of any considered compound to become part of the expanded seed set. More precisely, each metabolite outside the previously computed scope is considered as putative add-on metabolite only if its corresponding scope contains a reaction that ensures the production of the metabolite.

The described change in the workflow is not part of the default preferences. It is left to the user whether an additional run with the altered approach is required.

2.8 Final minimal seed set

ASBIG computes a final minimal seed set, whose scope set includes all biomass compounds. The minimal seed set consists of two major parts: initially given metabolites and add-on metabolites. The incorporated add-on metabolites can be divided into two major classes: (i) metabolites described in Section 3.1, which are elementary because of the structure of the model and not of principal interest, and (ii) metabolites, which show an autocatalytic behavior and embody a pivotal part of the metabolism. These compounds constitute the crucial result of ASBIG, as their presence in the minimal seed set indicates potential gaps within the metabolic network reconstruction.

3 RESULTS AND DISCUSSION

We developed a new method to detect inconsistencies in metabolic network reconstructions. The key step in this method is the identification of metabolites, which are necessary to ‘unlock’ specific metabolic pathways, that is, to make them accessible. Such metabolites usually represent cofactors that are required, besides the substrate of a pathway, to produce the corresponding products. As outlined above, although a wide range of metabolites of a biochemical network can usually be produced from a small set of precursor metabolites, an additional set of cofactors such as ATP and NAD is required to make most of metabolism accessible. The central goal of our approach is to identify these cofactors, which we call autocatalytic compounds, for a specific network and compare them with the usually well-conserved set of autocatalytic compounds of other metabolic models. If uncommon autocatalytic compounds are identified, this can usually be regarded as an indicator of missing metabolic functionality in a network reconstruction.

Figure 1 depicts the methodical procedure. First, the reconstruction of interest is examined via scope analysis with a predefined initial seed set (always containing the same metabolites regardless of the metabolic model investigated). This initial seed set is incrementally expanded by additional compounds until all biomass components are included in the computed metabolic scope (leading to an expanded metabolic seed set, Fig. 1B). Each compound added, named add-on metabolite, unlocks certain pathways in the model and leads to a larger metabolic scope. However, not every add-on metabolite contributes to the biomass production. Hence, in the last step of the ASBIG analysis, the expanded seed set is minimized to the minimal set of compounds required for biomass production (Fig. 1C).

In summary, for every metabolic reconstruction, a minimal set of metabolites is computed. This minimal set allows the production of all biomass components and consists of initially given metabolites and a subset of add-on metabolites (see 2.8). As prerequisite for each model investigation, it is assumed that those parts of the model that are necessary for the biomass production can be unlocked with the initially given metabolites. However, this assumption was not true for most investigated metabolic reconstructions and add-on metabolites had to be included. Each add-on metabolite has the capacity to provide access to priorly blocked parts of metabolism, and thus, uncovers putative gaps within a metabolic network reconstruction.

3.1 Commonly identified inconsistencies in multiple reconstructions

Initial runs of ASBIG indicated the presence of common autocatalytic compounds in all (or at least the majority of) the investigated metabolic models. This observation was a consequence of similar features among different models. Often, protein-derived reaction components, like the cofactor thioredoxin or the cofactor-carrying acyl carrier protein (ACP), were implemented as discrete compounds. However, biosynthesis of ribosomal protein is not included in most metabolic network reconstructions. As a result, these protein-derived compounds were present in the minimal seed set. Occasionally, metabolites that require protein-derived compounds for their biosynthesis appeared in the minimal seed set. Their assignment to the corresponding protein-derived compound involved additional effort, which could be avoided by integrating the protein-derived reaction components in the initial seed set.

Another metabolite identified by ASBIG in several network reconstructions was dihydrolipoamide. In these models, dihydrolipoamide was part of small isolated reaction cycles, excluding its own *de novo* biosynthesis. Hence, dihydrolipoamide and other commonly identified metabolites could be regarded as autocatalytic, although the lack of their biosynthetic pathways represented a common flaw of the corresponding metabolic networks. Consequently, the compounds were added to the initial seed set for subsequent analyses (see supplementary information for a complete list of initial seed set compounds).

3.2 Application to selected reconstructions

3.2.1 Escherichia coli The metabolism of *E.coli* is well investigated and the associated metabolic reconstruction is one of the best-curated models (Orth *et al.*, 2011). Accordingly, no peculiarities or gaps were detected by ASBIG.

3.2.2 Arabidopsis thaliana ASBIG identified several inconsistencies in the *A.thaliana* model. The presence of cytosolic nicotinamide adenine dinucleotide phosphate (NADP) in the minimal seed set unraveled the lack of a NAD phosphorylation reaction in the cytosol. Similarly, plastidic NADP turned out to be autocatalytic because of its metabolic connection to plastidic hexadecanoic acid (palmitic acid), which was also found in the minimal seed set.

Plant hexadecanoic acid production involves the NADH-dependent reduction of *trans*-2-hexadecenoyl-ACP to hexadecanoyl-ACP. However, the insertion of plastidic NAD, ACP and CoA did not result in the biosynthesis of hexadecanoic acid. Instead, plastidic NADP that is required in its reduced form (NADPH) during the reduction of oxohexadecanoyl-ACP to hydroxypalmitoyl-ACP was present in the minimal seed set. Thus, plastidic CoA, ACP, NADH and NADPH are crucial factors to enable fatty acid biosynthesis in the *A.thaliana* model. However, the requirement of NADPH in addition to NADH was unexpected and had to be resolved through detailed manual network analysis.

The requirement of both plastidic and cytosolic NADPH suggested a gap in the phosphorylation process of NAD. The *A.thaliana* genome contains two genes coding for NAD kinases, which are annotated as NADK1 and NADK2 (Turner *et al.*, 2004). The two corresponding enzymes were found to be localized in the cytosol and plastid stoma, respectively (Waller *et al.*, 2010). The reactions catalyzed by theses enzymes (EC 2.7.1.23) were absent in the metabolic reconstruction of *A.thaliana*, and NADPH lost its autocatalytic property by adding the reaction EC 2.7.1.23 to the cytosol and to the plastidic stoma compartment. Consequently, plastidic and cytosolic NADPH was not detected in the minimal seed set anymore.

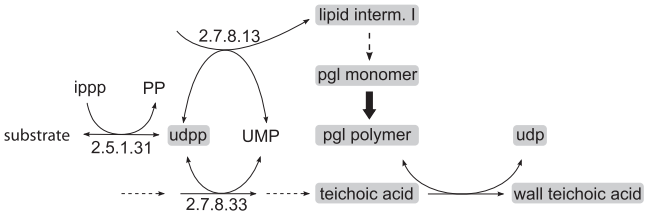

Fig. 4. Schematic representation of parts of the PGL and the teichoic acid biosynthetic pathways in the *B.subtilis* model. The conflicting minimal seed substrate (italics) is geranylgeranyl PP in case of the model and, in contrast, *trans,trans*-farnesyl PP according to information from KEGG and BsubCyc. Further, addition of the reaction depicted by the bold arrow would omit the need for PGL as crucial compound. Key intermediates of the pathway are highlighted in gray; the remaining metabolites complete the reactions. Each reaction is labeled with the corresponding EC number (if available). Dashed arrows mark several reaction steps. Abbreviation: IPPP—isopentenyl diphosphate

3.2.3 Bacillus subtilis Applying ASBIG to *i*BSU1103, the metabolic network model of *B.subtilis*, revealed two pathways with deficient or no accessibility: cell wall biosynthesis and thiamine biosynthesis.

Although the relevant genes for thiamine biosynthesis in *B.subtilis* are known (Lawhorn *et al.*, 2004), the corresponding pathway was not present in *i*BSU1103. As a result, thiamine was an inevitable component of the minimal seed set. Adding the thiamine biosynthetic pathway resulted in a gain of functionality for the model and the elimination of thiamine from the minimal seed set.

In addition to thiamine, geranylgeranyl diphosphate (PP) and glycerol teichoic acid were part of the final minimal seed set. In the model, geranylgeranyl PP is the direct precursor of undecaprenyl PP (UDPP), which is an intermediate in the biosynthesis of the cell wall components peptidoglycan (PGL) and teichoic acid (Fig. 4, EC 2.7.8.13 and EC 2.7.8.33) (Swoboda *et al.*, 2010). Hence, geranylgeranyl PP and glycerol teichoic acid were required in the minimal seed set for the cell wall biosynthesis.

The corresponding pathways in the model were examined to resolve the need for the two compounds. The teichoic acid biosynthetic pathway was implemented in *i*BSU1103. However, teichoic acid synthesis remained locked in a scope analysis with the initial seed set owing to the unavailability of UDPP. As mentioned above, geranylgeranyl PP is assigned as the direct precursor of UDPP in the *i*BSU1103 model. In contrast, the database BsubCyc indicates *trans,trans*-farnesyl PP as the precursor of UDPP (BsubCyc: di-*trans*,poly-*cis*-undecaprenyl phosphate biosynthesis). Furthermore, the gene *uppS* in *B.subtillis* encodes an UDPP synthase (Inaoka and Ochi,

2012), which requires *trans,trans*-farnesyl PP to produce UDPP (EC 2.5.1.31, Fig. 4). Apparently, the reaction EC 2.5.1.31 of the model included an incorrect substrate assignment leading to the substrate incorporation into the minimal seed set, a conflicting minimal seed metabolite. The exchange of geranylgeranyl PP with *trans,trans*-farnesyl PP in the reaction EC 2.5.1.31 resolved the need for the conflicting minimal seed metabolite and geranylgeranyl PP was eliminated from the minimal seed set. The identification of conflicting minimal seed metabolites demonstrates the ability of ASBIG to detect wrongly implemented reactants.

Despite the modification of reaction EC 2.5.1.31, teichoic acid remained in the minimal seed set because of its ability to provide the PGL polymer to the metabolic scope (the backward reaction of EC. 3.6.3.40 removes teichoic acid from PGL, Fig. 4). To avoid this property of teichoic acid, a reaction, which transforms multiple PGL monomers to a polymer, was implemented in the model (Fig. 4, light gray dashed arrow). As a result, no polymers had to be provided as self-replicators anymore, and teichoic acid was eliminated from the final minimal seed set. To sum up, the application of ASBIG substantially improved the integrity of the *B.subtilis* model *i*BSU1103 leading to a more comprehensive representation of the *in vivo* metabolism. Furthermore, the investigation of *i*BSU1103 illustrates central characteristics of the method: computed add-on metabolites not necessarily indicate the corresponding pathway or specific location of an inconsistency, and the results generated by ASBIG can suggest the inclusion of further functionality for the model.

3.2.4 Chlamydomonas reinhardtii The minimal seed set for iRC1080, the metabolic model of *C.reinhardtii*, contained add-on metabolites that indicated the lack of functionalities in some biosynthetic pathways. To access all components of the biomass reaction, thiamine, glutathione, chorismate, magnesium-protoporphyrin IX 13-monomethyl ester and plastidic plastoquinone expanded the initial seed set.

The biosynthesis for thiamine was not implemented in the model, likely because critical parts of the pathway still require experimental confirmation (Moulin *et al.*, 2013). However, the basic metabolic routes of *de novo* thiamine biosynthesis in *C.reinhardtii* are known, and the corresponding reactions could be added to the model. After the implementation, thiamine lost its autocatalytic property.

The presence of glutathione in the minimal seed set could be linked to the biosynthetic pathway of cysteine. Assimilatory sulfate reduction in *C.reinhardtii* depends on glutathione and leads to the generation of hydrogen sulfide. Subsequently, hydrogen sulfide can be incorporated into *O*-acetyl-L-serine to form cysteine. Hence, it was assumed that hydrogen sulfide, glutathione or cysteine were required to be part of the minimal seed set. Three runs of ASBIG, each with one out of the three components in the initial seed set, confirmed this assumption (and resulted, for hydrogen sulfide and cysteine, in the replacement of glutathione in the minimal seed set).

It was further possible to identify an impaired biosynthesis of chlorophyllide a explaining the presence of protoporphyrin IX 13-monomethyl ester in the minimal seed set. *S*-adenosyl-L-methionine (SAM) was one of the substrates to produce the protoporphyrin ester. Hence, another run of ASBIG was performed with SAM included in the initial seed set. This led to the removal of protoporphyrin IX 13-monomethyl ester, implicating the relevance of SAM.

The essential character for both plastidic SAM and cytosolic chorismate could be explained with a missing transport reaction between plastid and cytosol or missing biosynthetic pathways in one of the compartments. A putative transport enzyme could overcome this lack in both cases, even though no such transporters have been identified in *C.reinhardtii* yet. However, an enzyme to import SAM into the chloroplast was reported for *A.thaliana* (Bouvier *et al.*, 2006). Therefore, the findings of ASBIG highlight the need for further investigations of the metabolic capabilities of *C.reinhardtii*.

The putative plastoquinone biosynthetic pathway for *C.reinhardtii* is available in the KEGG database (KEGG: ubiquinone and other terpenoid-quinone biosynthesis). As it was not integrated in the investigated *C.reinhardtii* metabolic model, plastoquinone possessed autocatalytic property. Despite the lack of experimental evidence, the pathway was added to the model. First, the network reconstruction had to be expanded by two metabolites previously not implemented: nonaprenyl PP and 2-methyl-6-solanyl-1,4-benzoquinol. Nonaprenyl PP is the product of the reaction of geranylgeranyl PP with five molecules of isopentenyl PP (catalyzed by EC 2.5.1.85). Subsequently, nonaprenyl PP reacts with homogentisate to form 2-methyl-6-solanyl-1,4-benzoquinol (Merchant *et al.*, 2007), which is further converted to plastoquinone in a SAM-dependent reaction. The enzyme involved in the plastoquinone formation is not yet characterized, however, it is assumed that *VTE3* codes for the corresponding enzyme (Merchant *et al.*, 2007). By adding the aforementioned metabolic reactions to the *C.reinhardtii* model, plastoquinone lost its autocatalytic property and was eliminated from the minimal seed set.

The results for the examined network reconstruction of *C.reinhardtii*, iRC1080, show the versatile capabilities of ASBIG, for example, to identify inconsistencies that require additional experimental investigation as shown for the case of SAM or chorismate.

3.2.5 Saccharomyces cerevisiae In the *S.cerevisiae* model iMM904, dolichol was identified as a crucial add-on metabolite of the minimal seed set, implying an impaired biosynthesis of the compound. A survey of the metabolic model confirmed the absence of the corresponding pathway, although the biosynthesis of dolichol in *S.cerevisiae* has been previously described (Grabińska and Palamarczyk, 2002). To implement the metabolic route in the network reconstruction, several additions were required: (i) the condensation of farnesyl PP with 13 units of isopentenyl pyrophosphate to form polyprenyl PP (EC 2.5.1.87), (ii) the formation of polyprenol from polyprenyl PP and water (both polyprenyl PP and polyprenol had to be implemented as novel compounds in the model) and (iii) the conversion of polyprenol to dolichol. Manual expansion of the network reconstruction of *S.cerevisiae* provided access to the novel biosynthetic pathway of dolichol, thus eliminating the compound from the minimal seed set.

Another unexpected metabolite in the minimal seed set was NADP located in the endoplasmatic reticulum (ER). This compound remained in the minimal seed set because of

the lack of information on a potential pathway providing NADP in the ER.

Even though dolichol is the only gap-indicating add-on metabolite, the identification of NADP as an autocatalytic compound suggests general inconsistencies of the model.

3.2.6 Zea mays Application of ASBIG to the *Z.mays* model DTMaize_C4GEM_45632 revealed four add-on metabolites in the final minimal seed set that required further investigation: thiamine, NADP, undecaprenyl phosphate (UDP), and 4-methylthio-2-oxobutanoate (all localized in the cytosol).

Although maize is capable of thiamine biosynthesis, the pathway has not yet been fully characterized (Belanger *et al.*, 1995), and thus, is not entirely represented within the metabolic network model. As a result, the biosynthetic route of thiamine was blocked and thiamine had to be incorporated in the minimal seed set. With the future availability of novel information, the model can be improved by additionally including this specific biosynthetic pathway.

The absence of a cytosolic NAD kinase necessitated NADP to be an element of the minimal seed set. However, transcriptome analysis and automated annotation suggested the existence of a functional NAD kinase in *Z.mays* (UniProt ID: B6TAB2_MAIIZE) (Alexandrov *et al.*, 2009). Therefore, the reaction (EC 2.7.1.23) was added leading to the elimination of NADP from the minimal seed set.

Subsequently, the crucial character of UDP was examined and a missing reaction to phosphorylate uridine 5′-phosphate (UMP) to UDP in UTP biosynthetic pathway was identified. As for the NAD kinase, transcriptome data and electronic annotation (UniProt ID: B6T904_MAIZE) (Alexandrov *et al.*, 2009) suggested the presence of a functional UMP kinase. Hence, the detected gap could be resolved by implementing the additional reaction, thus eliminating UDP from the minimal seed set.

The application of the more restrictive 'producibility constraint' during the seed set expansion (see Section 2) allowed to link the autocatalytic metabolite 4-methylthio-2-oxobutanoate with the biosynthetic pathway of methionine. 4-Methylthio-2-oxobutanoate was part of the minimal seed set owing to its ability to bypass a blocked canonical methionine biosynthetic pathway (KEGG: map00270). It was assumed that a limited availability of cysteine, an essential intermediate in the methionine biosynthesis, caused the blocked methionine production (see also cysteine biosynthesis in *C.reinhardtii*). By adding cysteine to the initial seed set, folate instead of 4-methylthio-2-oxobutanoate was represented as an essential add-on metabolite of the minimal seed set. Evidently, the canonical methionine biosynthetic pathway was inaccessible unless cysteine and folate (or precursors/intermediates of the corresponding pathways) were part of the minimal seed set. A derivative of folate, the methyl group donor methyltetrahydropteroyltri-L-glutamate, is vital for the last step of the methionine biosynthesis (Hanson and Gregory, 2011). However, the biosynthesis of folate and its derivatives was impaired in the model, likely because the *in vivo* pathway is not yet fully understood (Hesse *et al.*, 2004), leading to the inevitable role of folate. Comparable with thiamine, the implementation of the biosynthetic pathway in the model depends on additional experiments to elucidate folate biosynthesis in *Z.mays*.

Another inconsistency in the last step of the methionine biosynthesis was identified: According to literature, the triglutamate of methyltetrahydrofolate (MeTHF), methyltetrahydropteroyltri-L-glutamate, acts as the substrate for the methionine synthase (EC 2.1.1.14) (Eckermann *et al.*, 2000; Eichel *et al.*, 1995). In the *Z.mays* model, this enzyme used 'pure' MeTHF as methyl donor (Fig. 4), indicating a conflicting minimal seed metabolite. Additionally, another methionine-producing reaction (EC 2.4.2.7), which involved methyltetrahydropteroyltri-L-glutamate as substrate, was available. However, the latter reaction was misassigned, as the denoted EC number characterizes an adenine phosphoribosyltransferase (EC 2.4.2.7). Thus, methyltetrahydropteroyltri-L-glutamate was implemented as substrate for the methionine synthase (EC 2.1.1.14), and the redundant misassigned reaction (EC 2.4.2.7) was deleted. As the *Z.mays* model lacked the reactions necessary for the transfer of the single carbon to tetrahydropteroyltri-L-glutamate, the triglutamate of MeTHF replaced folate (precursor of MeTHF) in the final minimal seed set. Although the exact mechanism of the methyl group transfer is not completely characterized for plants, the basic pathway composed of two reactions is known (Ravanel *et al.*, 1998). Hence, the pathway was implemented in the model to connect methyltetrahydropteroyltri-L-glutamate with the folate metabolism. Subsequently, folate and cysteine were part of the minimal seed set again and enabled methionine biosynthesis.

Each of the four identified add-on metabolites indicates substantial restrictions in the *Z.mays* model DTMaize_C4GEM_45632. Even though not each issue could be resolved, the results of ASBIG provide valuable indications to the *Z.mays* model for improvements.

3.3 Application to automatically reconstructed networks

To further scrutinize the ability of ASBIG to improve the consistency of metabolic network reconstructions, we analyzed 193 genome-scale metabolic networks, which have been reconstructed automatically based on the organisms' genome sequences and genomic annotations (Henry *et al.*, 2010). Among all models, 234 different metabolites were identified as crucial elements in the final minimal seed set after adding them during scope analysis. This list of compounds suggested that reactions were possibly missing in the biosynthetic pathways linked to these metabolites (Supplementary Table S1). A median of 6 additional compounds per model, necessary to facilitate biomass production, was identified (Supplementary Fig. S2). In 69% of all models, PGL polymer (n–1 subunits) was identified as an inevitable compound. Peptidoglycan polymers are found in the membranes of bacteria and occur with a varying number of subunits (Vollmer and Holtje, 2004). The elongation and shortening of this polymer are included in the models, but not the *de novo* biosynthesis. Furthermore, in multiple models, which were analyzed by ASBIG, external spermidine (47%), the glutamate-accepting tRNA (26%), thiamine (24%), external alanylhistidine (22%) and glycyl-L-asparagine (21%) were frequently identified as compounds that could not be synthesized. The recurrence of these compounds might be owing to missing knowledge of

possible biosynthetic routes or owing to properties of the reconstruction process (all 193 automated reconstructions analyzed here were generated using the same procedure) (Henry *et al.*, 2010). On the other hand, 194 of the 243 compounds were identified in <5% of all automated reconstructions analyzed, suggesting model-specific gaps or errors in the reconstruction or actual auxotrophies of the organisms.

Taken together, our analysis of a wide range of different automatically reconstructed metabolic networks reveals that several metabolites have an autocatalytic property or are inevitable for biomass production. The identification of these autocatalytic metabolites using ASBIG can improve the quality of the reconstruction by suggesting that the *de novo* biosynthetic pathways of the identified metabolites are either incomplete or missing.

3.4 Validation of ASBIG on auxotrophic mutants of *E.coli*

As mentioned above, no gaps were identified in the *E.coli* model iJO1366. Using this network, we knocked out reactions, recalculated the minimal seed set and compared the changes with reported experimental phenotypes of single-gene deletion mutants of *E.coli* K12 derivatives (Bertels *et al.*, 2012). Using this approach, we were able to validate ASBIG in terms of the method's potential to predict gene essentiality as well as altered nutritional requirements of mutants compared with the wild type strain. We deleted 10 reactions, one at the time, each one in the biosynthetic pathways of a specific amino acid. The deleted reactions correspond to gene deletions, which are known to cause a specific auxotrophy for the focal amino acid of the mutant strain (Bertels *et al.*, 2012). For each of the mutated networks, ASBIG identified an add-on metabolite, which was necessary to ensure biomass production. The add-on metabolite was the final product (the amino acid) of the affected pathway, an intermediate within the pathway or a derivative of the focal amino acid (for an overview see Supplementary Table S2). Hence, the output of ASBIG reflects the auxotrophy of the mutant strain impressively demonstrating the capability of ASBIG to detect gaps within primary metabolic pathways.

4 CONCLUSION

In conclusion, ASBIG is an efficient method for detecting inconsistencies in existing genome-scale metabolic network reconstructions. The method facilitates network validation and automated gap detection in primary metabolism contributing considerably to the quality improvement of metabolic models. ASBIG combines the conceptual approaches of autocatalytic metabolites and scope analysis, thereby allowing to test network reconstructions for their integrity. For each investigated metabolic model, a minimal seed set of metabolites, which provides access to all fundamental metabolic pathways, is computed. Putative flaws of the model can be deduced directly from identified minimal seed set.

The benefit of ASBIG was demonstrated by investigating different metabolic models. In the closely examined models, numerous kinds of errors could be identified including i) missing reactions, ii) missing pathways or iii) gaps caused by insufficient knowledge of metabolic processes. Furthermore, not only single gaps in individual models were identified, but also common flaws simultaneously present in several network reconstructions, could be detected in a large screen of 190 network reconstructions.

In contrast to the commonly applied gap-finding approach (i.e. deriving putative gaps from the calculated flux distributions), ASBIG is a simple method and detects other types of inconsistencies. Hence, the application of ASBIG contributes significantly to model refinement and validation representing a complementary approach to existing gap-finding methods.

Funding: This work was supported by the Jena School for Microbial Communication (JSMC) to S.W. and the Volkswagen Foundation to C.Ko.

Conflict of interest: none declared.

REFERENCES

Alexandrov,N.N. *et al.* (2009) Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol. Biol.*, **69**, 179–194.

Barua,D. and Reed,J.L. (2010) An automated phenotype-driven approach (geneforce) for refining metabolic and regulatory models. *PLoS Comput. Biol.*, **6**, e1000970.

Belanger,F.C. *et al.* (1995) Evidence for the thiamine biosynthetic pathway in higher-plant plastids and its developmental regulation. *Plant Mol. Biol.*, **29**, 809–821.

Bertels,F. *et al.* (2012) Design and characterization of auxotrophy-based amino acid biosensors. *PLoS One*, **7**, e41349.

Bouvier,F. *et al.* (2006) *Arabidopsis* SAMT1 defines a plastid transporter regulating plastid biogenesis and plant development. *Plant Cell*, **18**, 3088–3105.

Caspi,R. *et al.* (2010) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–D479.

Chang,R.L. *et al.* (2011) Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Mol. Syst. Biol.*, **7**, 518.

de Figueiredo,L.F. *et al.* (2008) Can sugars be produced from fatty acids? A test case for pathway analysis tools. *Bioinformatics*, **24**, 2615–2621.

de Oliveira Dal'Molin,C.G. *et al.* (2010a) C4gem, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiol.*, **154**, 1871–1885.

de Oliveira Dal'Molin,C.G. *et al.* (2010b) Aragem, a genome-scale reconstruction of the primary metabolic network in *Arabidopsis*. *Plant Physiol.*, **152**, 579–589.

Durot,M. *et al.* (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.*, **33**, 164–190.

Eckermann,C. *et al.* (2000) Plant methionine synthase: New insights into properties and expression. *Biol. Chem.*, **381**, 695–703.

Eichel,J. *et al.* (1995) Vitamin-B12-independent methionine synthase from a higher plant (*Catharanthus roseus*), molecular characterization, regulation, heterologous expression, and enzyme properties. *Eur. J. Biochem.*, **230**, 1053–1058.

Eigen,M. and Schuster,P. (1977) The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, **64**, 541–565.

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

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

Grabińska,K. and Palamarczyk,G. (2002) Dolichol biosynthesis in the yeast *Saccharomyces cerevisiae*: an insight into the regulatory role of farnesyl diphosphate synthase. *FEMS Yeast Res.*, **2**, 259–265.

Handorf,T. *et al.* (2005) Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.*, **61**, 498–512.

Hanson,A.D. and Gregory,J.F. III (2011) Folate biosynthesis, turnover, and transport in plants. *Annu. Rev. Plant Biol.*, **62**, 105–125.

Henry,C.S. *et al.* (2009) iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on seed annotations. *Genome Biol.*, **10**, R69.

Henry,C.S. *et al.* (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.

Hesse,H. *et al.* (2004) Current understanding of the regulation of methionine biosynthesis in plants. *J. Exp. Bot.*, **55**, 1799–1808.

Inaoka,T. and Ochi,K. (2012) Undecaprenyl pyrophosphate involvement in susceptibility of *Bacillus subtilis* to rare earth elements. *J. Bacteriol.*, **194**, 5632–5637.

Jerby,L. *et al.* (2010) Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism. *Mol. Syst. Biol.*, **6**, 401.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Kaleta,C. *et al.* (2011) *In silico* evidence for gluconeogenesis from fatty acids in humans. *PLoS Comput. Biol.*, **7**, e1002116.

Kun,A. *et al.* (2008) Computational identification of obligatorily autocatalytic replicators embedded in metabolic networks. *Genome Biol.*, **9**, R51.

Latendresse,M. *et al.* (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics*, **28**, 388–396.

Lawhorn,B.G. *et al.* (2004) Biosynthesis of the thiamin pyrimidine: the reconstitution of a remarkable rearrangement reaction. *Org. Biomol. Chem.*, **2**, 2538–2546.

Lewis,N.E. *et al.* (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of *in silico* methods. *Nat. Rev. Microbiol.*, **10**, 291–305.

Li,P. *et al.* (2010) Systematic integration of experimental data and models in systems biology. *BMC Bioinformatics*, **11**, 582.

Merchant,S.S. *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.

Mo,M.L. *et al.* (2009) Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.*, **3**, 37.

Moulin,M. *et al.* (2013) Analysis of *Chlamydomonas* thiamin metabolism *in vivo* reveals riboswitch plasticity. *Proc. Natl Acad. Sci. USA*, **110**, 14622–14627.

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

Orth,J.D. *et al.* (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol.*, **7**, 535.

Ponce-de Leon,M. *et al.* (2013) Solving gap metabolites and blocked reactions in genome-scale models: application to the metabolic network of *Blattabacterium cuenoti*. *BMC Syst. Biol.*, **7**, 114.

Ravanel,S. *et al.* (1998) The specific features of methionine biosynthesis and metabolism in plants. *Proc. Natl Acad. Sci. USA*, **95**, 7805–7812.

Ruppin,E. *et al.* (2010) Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Curr. Opin. Biotechnol.*, **21**, 502–510.

Satish Kumar,V. *et al.* (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, **8**, 212.

Segrè,D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA*, **99**, 15112–15117.

Swoboda,J.G. *et al.* (2010) Wall teichoic acid function, biosynthesis, and inhibition. *Chembiochem*, **11**, 35–45.

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

Turner,W.L. *et al.* (2004) Cloning and characterization of two NAD kinases from *Arabidopsis*. Identification of a calmodulin binding isoform *Plant. Physiol.*, **135**, 1243–1255.

Varma,A. and Palsson,B.Ø. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.*, **60**, 3724–3731.

Vass,M. *et al.* (2004) The jigcell model builder and run manager. *Bioinformatics*, **20**, 3680–3681.

Vollmer,W. and Holtje,J.V. (2004) The architecture of the murein (peptidoglycan) in Gram-negative bacteria: vertical scaffold or horizontal layer(s)? *J. Bacteriol.*, **186**, 5978–5987.

Waller,J.C. *et al.* (2010) Subcellular and tissue localization of NAD kinases from *Arabidopsis*: Compartmentalization of de novo NADP biosynthesis. *Planta*, **231**, 305–317.

Wang,Q. *et al.* (2006) Genome-scale *in silico* aided metabolic analysis and flux comparisons of *Escherichia coli* to improve succinate production. *Appl. Microbiol. Biotechnol.*, **73**, 887–894.

Wessely,F. *et al.* (2011) Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. *Mol. Syst. Biol.*, **7**, 515.

Yus,E. *et al.* (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science*, **326**, 1263–1268.

Zhang,P. *et al.* (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant .Physiol.*, **153**, 1479–1491.