Published in final edited form as:

Biotechnol Bioeng. 2010 October 15; 107(3): 403–412. doi:10.1002/bit.22844.

Systematizing the Generation of Missing Metabolic Knowledge

Jeffrey D. Orth¹ and Bernhard Ø. Palsson¹

¹ Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

Abstract

Genome-scale metabolic network reconstructions are built from all of the known metabolic reactions and genes in a target organism. However, since our knowledge of any organism is incomplete, these network reconstructions contain gaps. Reactions may be missing, resulting in dead-ends in pathways, while unknown gene products may catalyze known reactions. New computational methods that analyze data, such as growth phenotypes or gene essentiality, in the context of genome-scale metabolic networks, have been developed to predict these missing reactions or genes likely to fill these knowledge gaps. A growing number of experimental studies are appearing that address these computational predictions, leading to discovery of new metabolic capabilities in the target organism. Gap-filling methods can thus be used to improve metabolic network models while simultaneously leading to discovery of new metabolic gene functions.

Keywords

Gap-filling; gene annotation; GrowMatch; metabolic network reconstruction; SMILEY

Introduction

Genome-scale metabolic network reconstructions are biochemically, genetically, and genomically (BiGG) structured knowledge bases that seek to formally represent the known metabolic activities of an organism. These reconstructions are based on the annotated genome of a particular organism, and they are organism specific. Reconstructions have been assembled for the metabolic networks of *Escherichia coli* (Feist et al. 2007), *Saccharomyces cerevisiae* (Mo et al. 2009), human (Duarte et al. 2007), and about forty other organisms (Feist et al. 2009). Reconstructions are useful because they can be converted into constraint-based models, allowing practical calculations and simulations to be performed. In these models, the reactions in the reconstruction are mathematically represented by a stoichiometric matrix, in which every unique reaction is represented by a row and every metabolite is represented by a column. Minimum and maximum flux constraints can be imposed on each reaction, and the associated genes and proteins can be represented with a Boolean formalism (Reed et al. 2003). Computational analysis techniques such as flux balance analysis, which uses linear programming to identify optimal reaction flux distributions under different sets of constraints (Orth et al. 2010; Varma and Palsson 1993a; Varma and Palsson 1993b), can then be used to analyze models. Mathematical models based on genome-scale metabolic network reconstructions have many practical uses, as discussed in recent reviews (Feist and Palsson 2008; Oberhardt et al. 2009).

Even the most complete models are not perfect, however. They all contain gaps, or missing information. There are two types of missing information in metabolic network

Correspondence to: Bernhard Ø. Palsson¹ Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, Mail Code 0412, La Jolla, CA 92093-0412, USA. Tel: +1 858 534 5668; Fax: +1 858 822 3120; palsson@ucsd.edu.

reconstructions. The first type is gaps in a network, places where a reaction that consumes or produces a metabolite is missing, creating a dead-end. The second type is orphan reactions. These are reactions that are known to exist, but it is not yet known which gene or genes code for their enzymes. Both types of gaps are the result of our incomplete knowledge of the metabolism of an organism. In recent years, methods have been developed to predict which genes or reactions should be associated with the gaps in metabolic network reconstructions. These gap-filling methods have the potential to be very useful because they can improve the predictive capabilities of models by making them more realistic while simultaneously improving our knowledge of an organism by characterizing a previously unknown gene. Gap-filling is therefore a discovery tool as well as a model refinement tool. This review introduces the various computational methods that are used to fill gaps in metabolic networks. These methods are summarized in Table I and Figure 1, and examples of cases in which these methods led to new discoveries are listed in Table II. Purely experimental methods that lead to characterization of enzymes can also be used to improve models, but these will not be discussed in this review as many have been reviewed previously (Breitling et al. 2008).

Types of gaps and orphans

There are two fundamentally different types of missing information in metabolic network reconstructions (Figure 2). First, there are gaps; holes in the network where a reaction that should occur in the organism is absent. These gaps are manifest in network reconstructions as dead-end metabolites, which have a producing reaction but no consuming reaction. It is also possible for a metabolite to be consumed by a reaction but not produced by any reactions. These two types of dead-ends have been termed *root no-consumption metabolites* and *root no-production metabolites* (Kumar et al. 2007). When simulating steady-state reaction fluxes in these networks (as in flux balance analysis), the reactions producing or consuming dead-end metabolites can never carry flux under any conditions. These reactions are said to be “blocked,” and any reactions upstream or downstream from them will also be unable to carry flux under steady-state conditions. An extreme case of a dead-end metabolite is one that has no producing or consuming reactions, and thus forms an isolated island (Palsson 2009).

There are several reasons why there may be gaps in a metabolic network. First, it is possible that the actual biochemical network is missing an enzyme that is part of a completed pathway in related organisms. An example of this situation is in the O-antigen synthesis pathways of many *E. coli* K-12 strains, in which an IS5 insertion element has disrupted a rhamnosyltransferase gene, inactivating it (Liu and Reeves 1994; Rubires et al. 1997). These strains are therefore unable to produce a functioning rhamnosyltransferase enzyme, blocking the entire downstream pathway that produces O-antigen. In cases such as this, there really is a gap in the biochemical network, known as a *biological gap*. The *E. coli* K-12 MG1655 reconstruction iAF1260 (Feist et al. 2007) contains a biologically realistic gap, and this pathway is unable to carry flux. Another reason for gaps in network models is the scope of the models themselves. To date, most large-scale metabolic network models do not include other systems such as signaling or transcription and translation. Metabolites that are produced in metabolism but then enter these other systems may be left as gaps in models, even though their biological functions are known. These are *scope gaps*. An example of scope gaps is the tRNAs in iAF1260. This model contains tRNA charging reactions, but there are no consuming reactions for these tRNAs, even though it is well known how charged tRNAs are used in the process of translation. Finally, there may be a gap because it is not known what biochemical reaction produces or consumes a certain metabolite. In this case, the gap is not biologically realistic; it is the result of limited knowledge. These are called *knowledge gaps*, and to fill them, new biological discoveries must be made.

The second category of missing information in a metabolic network is orphan reactions. These are biochemical reactions that are known to occur but are catalyzed by an unknown gene product. Reactions can be identified without genes through several different types of evidence, including biochemical assays of crude cell extracts, a known phenotype such as uptake or secretion of a particular substrate, or the implied presence of a reaction due to the presence of other genes in a conserved pathway. Even the most well studied organisms have many genes with unknown functions, and many of these genes may code for orphan reactions. *E. coli* K-12 MG1655, for example, still has 981 partially or fully uncharacterized genes according to the latest release of EcoCyc (version 13.6) (Keseler et al. 2009). It is also possible that some genes with currently known functions may also catalyze orphan reactions, so these genes must not be ignored when attempting to identify the gene associated with an orphan. Orphan reactions can be local, where the gene is unknown in one organism but known in at least one other. They can also be global orphan reactions, in which there are no known genes in any organism that code for a catalyzing enzyme. It has been determined that 30–40% of all known enzymatic activities are global orphans (Karp 2004; Lespinet and Labedan 2006b; Pouliot and Karp 2007). Global orphans are evenly distributed across the different types of known biochemical reactions (Chen and Vitkup 2007). Identifying the gene responsible for an orphan reaction in one organism increases the probability of identifying genes in other organisms, since it would provide a reference sequence for homology search methods. A database called ORENZA (ORphan ENZyme Activities) at <http://www.orenza.u-psud.fr/> lists currently known global orphan reactions (Lespinet and Labedan 2006a).

Methods for predicting gap-filling reactions

BNICE

When a gap exists in a biochemical network, there are potentially many sets of reactions that might fill the gap and restore connectivity. An algorithm called BNICE (Biochemical Network Integrated Computational Explorer) was developed to identify all biochemical reactions that could realistically link two metabolites (Hatzimanikatis et al. 2005). In this algorithm, molecules are represented as bond-electron matrices (BEMs) and reactions are represented by matrix addition. It is possible to model the activity of an enzyme catalyzed reaction occurring with alternative substrates by adding the same reaction matrix to different BEMs, resulting in different product BEMs from each one. The enzyme catalyzed reactions in the KEGG database (Kanehisa and Goto 2000; Kanehisa et al. 2006) were analyzed, and it was found that there are fewer than 250 known generalized enzyme reactions, each of which can be represented by a reaction matrix. These generalized enzyme reactions typically correspond with third level Enzyme Classification (EC) numbers (e.g. EC 1.1.1._), while the fourth level EC numbers indicate the particular substrates and products of a reaction. BNICE then iteratively searches for all possible pathways of chemical transformations from one metabolite to another, offering gap-filling hypotheses. The pathways for synthesis of the aromatic amino acids phenylalanine, tyrosine, and tryptophan were analyzed with BNICE restricted to use only the generalized enzymatic reactions already known to occur in these pathways. Several dozen alternative pathways were identified using these reaction mechanisms alone, some including metabolites not found in the KEGG or Chemical Abstracts Services databases. Thermodynamic analysis of the predicted pathways using the group contribution method showed that the currently known native pathways are the most favorable (Hatzimanikatis et al. 2005). Although gap-filling predictions generated by BNICE have not yet been experimentally verified, it remains a potentially useful tool.

GapFind and GapFill

The algorithms GapFind and GapFill were developed to attempt to minimize the total number of gaps in a metabolic network model (Kumar et al. 2007). GapFind is a mixed-integer linear programming (MILP) algorithm that can identify every gap in a network by identifying blocked metabolites, i.e. those that cannot be produced or consumed at steady-state under any conditions. Although it may be easy to identify some gaps by inspection of network or pathway maps, certain unusual model structures such as cycles can lead to non-intuitive sets of blocked metabolites, so GapFind can be very useful (Figure 3). In the *E. coli* model *iAF1260*, for example, the compounds in the ubiquinol-8 synthesis pathway are blocked even though the pathway appears to have a clear entry and exit. This is because ubiquinol-8 and ubiquinone-8 are interconverted in several reactions, but there are no reactions that consume either of these compounds, so no additional ubiquinol-8 can be produced at steady-state (Orth and Palsson 2010, in preparation). GapFill is another MILP algorithm, and its objective is to minimize the total number of gaps by reversing the directionality of existing reactions, adding new reactions (from the MetaCyc database (Caspi et al. 2008)), adding transport reactions for blocked metabolites, or adding intracellular transport reactions between compartments for multi-compartment models. GapFill attempts to reduce the number of gaps with the smallest number of model modifications possible. These algorithms were used to make gap-filling predictions on the single-compartment *E. coli* metabolic model *iJR904* (Reed et al. 2003) and the multi-compartment *S. cerevisiae* metabolic model *iND750* (Duarte et al. 2004). Although the hypotheses generated by GapFill were not experimentally verified, most were found to be in reasonable agreement with predicted free energy changes and database information.

SMILEY

With a constraint-based metabolic network model, flux balance analysis can be used to simulate the growth of microbial organisms on different substrates, including many different carbon and nitrogen sources. When an *in vivo* experiment shows that an organism can grow on a certain substrate but the model predicts that it cannot, the model is likely missing one or more of the reactions required to consume the substrate. An algorithm called SMILEY was created to predict which reactions are likely missing from a network when the model predicts no growth but experiment indicates growth (Reed et al. 2006). This algorithm was based on the OptStrain algorithm used in metabolic engineering (Pharkya et al. 2004). SMILEY uses MILP to attempt to identify a flux distribution that leads to growth on the substrate of interest while minimizing the total number of reactions added from a universal database of reactions (based on the KEGG database). It thus predicts which reactions should be added to the model parsimoniously, adding the smallest number of reactions necessary to reconcile *in silico* and *in vivo* growth predictions. SMILEY does not directly target or attempt to fill gaps in a model, but the reactions it adds could end up filling gaps. SMILEY was used to predict reactions missing from the *E. coli* model *iJR904*, and made several predictions that have been experimentally verified (see **New *E. coli* reactions discovered by SMILEY**) (Reed et al. 2006).

GrowMatch

The newest algorithm for filling gaps in metabolic networks is called GrowMatch. This algorithm uses experimentally determined gene essentiality data to identify incorrect model predictions (Kumar and Maranas 2009). If the model predicts growth but an experiment shows that the organism cannot grow with this gene knockout (i.e. it is an essential gene), there is a growth-no growth inconsistency (GNG). When the model predicts no growth but an experiment shows that growth is possible, there is a no growth-growth inconsistency (NGG). GrowMatch attempts to correct these two types of disagreements differently. In the case of a GNG mutant, the model has some extra capabilities that are not realistic and should

be removed or constrained. In cases of NGG inconsistency, the model is missing one or more reactions that allow growth *in vivo*. GrowMatch then uses the three single compartment methods from GapFill (reversing reaction directions, adding reactions from a database, or adding transport reactions) to attempt to correct the model. Corrections to GNG or NGG mutants made by GrowMatch can be either global or conditional. Global corrections resolve at least one inconsistency while creating no new inconsistencies, and conditional corrections resolve one inconsistency while creating inconsistencies in other mutant strains. GrowMatch was used to analyze the *E. coli* model *iAF1260* by comparing it to gene essentiality data from the Keio Collection (Baba et al. 2006). 72 GNG mutants and 38 NGG mutants were identified, and GrowMatch suggested corrections for 56 GNG and 13 NGG mutants (Kumar and Maranas 2009).

OMNI

Another algorithm related to SMILEY and GrowMatch that can also be used for gap-filling is called OMNI (Optimal Metabolic Network Identification) (Herrgård et al. 2006). This is another MILP based algorithm that compares *in silico* predictions to experimental measurements in an attempt to improve a constraint-based model. In this case, OMNI uses measured metabolic flux data, which is obtained through ¹³C labeling experiments (Fong et al. 2006). OMNI compares experimentally measured fluxes to fluxes predicted by flux balance analysis, and then attempts to minimize the total difference between measured and predicted fluxes by adding or removing reactions while maintaining a predicted growth rate above a defined minimum. It uses a matrix **F** that contains fixed reactions which cannot be deleted and a matrix **D** that contains reactions that may be deleted. To improve a model by predicting missing reactions, a library of candidate reactions such as the one used by SMILEY can be provided as **D** to the algorithm, which will then add reactions as needed to achieve a more realistic flux distribution. The OMNI algorithm has many potential uses other than gap-filling. Because it can remove reactions as well as add them, it can be used to remove reactions corresponding to poorly annotated genes if they do not match the experimental data. OMNI can also be used to identify alternative reaction mechanisms or bottleneck reactions in evolved strains or metabolically engineered strains.

Methods for predicting metabolic gene functions

PathoLogic Pathway Hole Filler

PathoLogic is a program for automatically constructing metabolic networks from annotated genomes. It uses EC numbers, Gene Ontology terms, or annotated gene names to map reactions to genes, and then assembles the reactions into pathways by comparing the reactions to the reference database MetaCyc and adding any missing reactions (Karp et al. 2002; Paley and Karp 2002). After performing these steps, however, many of the reactions in the new pathways may be orphans. Pathologic includes a Pathway Hole Filler (PHFiller) program that attempts to identify the genes associated with these reactions (Green and Karp 2004). The first step is to identify genes in other organisms that code for enzymes that catalyze an orphan reaction from the Swiss-Prot and PIR (Wu et al. 2003) databases. The sequences of these genes are then compared to the genome of the organism of interest using BLAST (Altschul et al. 1990), and genes with similar sequences are identified. Because PHFiller needs sequence data from other organisms, it cannot predict the genes associated with global orphan reactions. After identifying candidate genes, the hole filler program then uses a simple Bayesian network to calculate the probability that each gene actually encodes an enzyme that catalyzes the orphan activity. An extended PHFiller, called PHFiller-GC, which uses genome-context methods in addition to sequence comparisons to predict pathway genes, was released in 2007 (Green and Karp 2007). PHFiller-GC uses proteins that are known to occur in the same complex, proteins from the same operon, or known

transcription factor/regulatory target pairs to identify associated genes. Unlike the original PHFiller, PHFiller-GC is able to predict genes associated with orphan reactions, but it requires that at least one enzyme in the target pathway have a known gene association. Both PHFiller and PHFiller-GC were tested on known *E. coli* reactions with homologous genes in other organisms, and both identified the true gene in the top ten predictions well (85.1% for PHFiller and 91% for PHFiller-GC). For reactions associated with genes with no known homologs, PHFiller-GC identified the true gene in the top ten predictions 58.9% of the time.

SEED: a subsystems approach

In response to the increasing number of fully sequenced microbial genomes, the Fellowship for Interpretation of Genomes (FIG) has launched the Project to Annotate 1000 Genomes. This large scale project is using a subsystems approach to annotate these genomes (Overbeek et al. 2005). Briefly, a subsystem is defined as a set of “functional roles” that act together to carry out a specific biological process such as a traditional biochemical pathway. Annotators focus on annotating one subsystem at a time in many different organisms using the SEED annotation environment (Overbeek et al. 2004), rather than the traditional approach of annotating one full genome at a time. The organization of genes in many different genomes into common subsystems can be beneficial in attempts to characterize genes and fill gaps. When one organism is missing a particular subsystem component (reaction) that most other organisms with the same subsystem have, it is an indication of missing content. Bioinformatics methods can then be used to analyze the genes composing the subsystems in these other organisms to identify which genes in the target organism likely encode the missing reactions (Osterman and Overbeek 2003). Analysis of chromosomal clustering has proven to be the most useful analysis method (Osterman 2006). Genes that encode enzymes in the same pathways are often clustered close together on the genome, so genes and their orthologs that are found near each other in multiple genomes are likely to have related functions (Dandekar et al. 1998; Lee and Sonnhammer 2003; Overbeek et al. 1999). This approach followed by experimental verification has already led to several new discoveries (see **Pathways discovered by SEED**).

ADOMETA

Since the first genomes were sequenced over a decade ago, numerous bioinformatics methods have been developed to attempt to identify the functions of genes. These methods include analysis of gene co-expression (Wu et al. 2002), phylogenetic profiles (Huynen and Bork 1998; Pellegrini et al. 1999), chromosomal clustering (Lee and Sonnhammer 2003; Overbeek et al. 1999), and protein fusions (Enright et al. 1999; Yanai et al. 2001). More recently, a bioinformatics framework called ADOMETA (ADoption of Orphan METabolic Activities) that combines these types of information has been developed (Chen and Vitkup 2006; Kharchenko et al. 2006; Kharchenko et al. 2004). The genes in an organism with unknown functions (and genes with known functions as well) can be compared to the genes in the local metabolic network surrounding an orphan reaction using different types of functional association evidence. Genes that are close to each other in a metabolic network are more likely to have similar co-expression profiles, phylogenetic profiles, and so forth. In ADOMETA, the genes in an organism are compared to the surrounding genes using different methods, and the different association scores are combined using the Adaboost method (Freund and Schapire 1997; Schapire 2002). The products of the highest scoring genes are the most likely to catalyze the orphan reaction.

The first type of functional association evidence used by ADOMETA is analysis of gene co-expression (Wu et al. 2002). When two genes have similar expression patterns, they are likely to have related functions because they may be used in the same functional pathway. ADOMETA uses the Spearman's rank correlation coefficient to compare co-expression of

genes using data from the Stanford Microarray Database (Sherlock et al. 2001). Phylogenetic profiles compare the occurrence of genes in different organisms (Huynen and Bork 1998; Pellegrini et al. 1999). A phylogenetic profile consists of a vector of 1's and 0's, with a 1 for every organism that contains an ortholog of a certain gene and a 0 for every organism that doesn't. Genes with related functions evolve together, so genes with the most similar phylogenetic profiles are therefore more likely to exist in the same pathways. ADOMETA uses a BLAST-based dataset of evolutionarily distinct genomes to determine phylogenetic profiles. Chromosomal clustering, which is used in the subsystems approach, can be a very powerful type of association evidence, especially in prokaryotic organisms (Lee and Sonnhammer 2003; Overbeek et al. 1999). ADOMETA uses a dataset of orthologous gene clusters based on 108 genomes from KEGG. Finally, protein fusion data can be used (Enright et al. 1999; Yanai et al. 2001). One gene in an organism may be homologous to two separate genes in other organisms, and it is likely that these two genes have similar functions. ADOMETA uses BLAST to identify genes containing homology to both candidate orphan filling genes and network adjacent genes. ADOMETA was tested by comparing its predictions to the known enzyme encoding genes for reactions in both *E. coli* and *S. cerevisiae* using the *iJR904* and *iFF708* (Forster et al. 2003) models, respectively. 60% of the correct *E. coli* genes were in the top ten predictions for their reactions, and 43% of the correct yeast genes were in the top ten predictions. ADOMETA is available as a web server at <http://vitkuplab.cu-genome.org/html/adometa/adometa.html> and can currently (as of June 2010) predict orphan associated genes for *E. coli*, *S. cerevisiae*, and *Bacillus subtilis*.

Examples of experimental verification of *in silico* gap-filling predictions

New *E. coli* reactions discovered by SMILEY

When the SMILEY algorithm was applied to the *E. coli* model *iJR904*, Biolog growth phenotype data (Bochner et al. 2001) was collected and 54 conditions were identified in which *E. coli* could grow *in vivo* but the *iJR904* model predicted that growth was not possible (Reed et al. 2006). SMILEY identified growth restoring reactions for 26 cases, with multiple solutions found for 15 cases. Several of these predictions were investigated experimentally. For the substrates propionate and 5-keto-D-gluconate, SMILEY predicted that only transport reactions needed to be added to the *in silico* model. Eight different genes tentatively annotated as transporters were tested by growing gene knockout strains on these two substrates. The $\Delta putP$ (b1015) strain had a reduced growth rate on propionate and the $\Delta idnT$ (b4265) strain had a reduced growth rate on 5-keto-D-gluconate. It was also determined by RT-PCR that expression of both of these genes in wild type *E. coli* increased during growth on these substrates, providing strong evidence that *putP* codes for a propionate transporter and *idnT* codes for a 5-keto-D-gluconate transporter.

In another interesting case, SMILEY predicted that *E. coli* is able to grow on thymidine when a transport reaction for thymine is added to the model. The model then predicts that thymidine is consumed, split into deoxyribose-1-phosphate and thymine by the thymidine phosphorylase reaction, and thymine is then excreted. HPLC analysis of *E. coli* growing on thymidine verified this mechanism by showing that thymine concentration in the media increased as thymidine concentration decreased. 29 gene knockout strains were screened for potential thymine transport activity, and only the $\Delta tolC$ (b3035) strain showed a reduced growth rate on thymidine, indicating that this gene may code for a thymine transporter. However, it was also found that $\Delta tolC$ has a reduced growth rate on other substrates as well and that it can still secrete thymine, so the exact role of TolC is still unclear.

Pathways discovered by SEED

Several new pathways have been discovered by using the subsystems approach to genome annotation. A previously unknown pathway in *E. coli* that degrades uracil and thymidine was recently discovered using experimental analysis of gene knockout phenotypes (Loh et al. 2006). The genes responsible for this pathway are all part of the *E. coli rut* operon. When this pathway was analyzed as a subsystem by SEED annotators, it was found that homologous genes exist in several other bacterial species (*Yersinia enterocolitica*, *Acinetobacter* sp. ADP1, *Pseudomonas syringae*, *Caulobacter crescentus*, and *Agrobacterium tumefaciens*), and that these genes also tend to cluster together in similar patterns (Osterman 2006). This provides strong evidence for the existence of this pathway in these organisms as well.

Comparative genomics methods in the context of subsystems were used to fill a gap in the metabolism of *Thermotoga maritima* (Rodionov et al. 2007). This organism, like other hyperthermophilic bacteria and archaea, produces di-*myo*-inositol 1,1'-phosphate (DIP) as an osmoprotecting metabolite. The four reaction pathway that produces DIP in many of these organisms has been partially characterized, and the genes encoding two of the reactions in *T. maritima*, *ips* (TM1419) and *imp* (TM1415), were already known. However, the genes encoding the enzymes for the other two steps had not been discovered in any organism. Chromosomal clustering and phylogenetic profiles were used in the SEED platform to predict which genes may be responsible for the two reactions. It was found that the genes *dipA* (TM1418a) and *dipB* (TM1418b) tend to cluster with *ips* and *imp* in several species of thermophilic bacteria and archaea. DipA was expressed in *E. coli* and purified, and was then confirmed to carry out its predicted CTP:inositol monophosphate cytidyltransferase reaction *in vitro*, and crude extracts of *E. coli* with *dipB* expressed were confirmed to produce DIP. The entire pathway was also cloned from *T. maritima* into *E. coli* and was confirmed to function as expected. Among the other organisms that were predicted to contain this pathway is *Aeropyrum pernix*, which was shown by ³¹P NMR to produce DIP.

E. coli succinate semialdehyde dehydrogenase gene discovery by ADOMETA

For over 25 years, it has been known that *E. coli* is able to carry out the NAD⁺-dependent succinate semialdehyde dehydrogenase (SSADH) reaction, but it was not known which gene encoded the catalyzing enzyme (Donnelly and Cooper 1981). This reaction converts the toxic intermediate succinate semialdehyde to succinate. Using the iJR904 *E. coli* model, ADOMETA was used to test all *E. coli* genes with unknown function for potential SSADH activity, and *yneI* (b1525) was identified as the top candidate (Fuhrer et al. 2007). A $\Delta yneI$ strain along with a $\Delta gabD$ (b2661) strain (missing the known gene encoding NADP⁺-dependent SSADH) were grown in M9 glycerol media with GABA as the only nitrogen source. Both strains grew to lower concentrations than wild type, presumably due to succinate semialdehyde toxicity. NAD⁺- and NADP⁺-dependent SSADH activities were assayed in crude cell extracts, and this activity was not detected in the $\Delta yneI$ strain but was detected in $\Delta gabD$ and wild type. Finally, the YneI and GabD proteins were expressed and purified, and YneI was found to accept both NAD⁺ and NADP⁺ as cofactors. All of this evidence confirmed the ADOMETA prediction that *yneI* encodes the NAD⁺-dependent SSADH of *E. coli*.

New reconstructions that incorporate gap-filling predictions

Several recent genome-scale metabolic network reconstructions have incorporated some of the gap-filling procedures outlined in this review. Both GapFill and GrowMatch were used in the assembly of the *Mycoplasma genitalium* metabolic reconstruction iPS189 (Suthers et al. 2009). The initial version of this model was assembled in an automated manner by comparing homologous genes in the *M. genitalium* genome to genes in other

reconstructions. This resulted in an initial model with 187 genes. This initial model, however, had many gaps, and 177 of the 179 reactions were blocked. GapFill was used to add a total of 86 reactions to the model, several of which were associated with genes based on sequence homology. GrowMatch was then used to further refine the model. The model gene essentiality predictions were 79% accurate before using GrowMatch, and were improved to an accuracy of 87% after. The final version of the model had 189 genes and 262 reactions. A new reconstruction of *Salmonella typhimurium* LT2, iMA945, was constructed using a similar strategy (AbuOun et al. 2009).

In another study, a reconstruction of *B. subtilis* 168, iBsu1103 was built using both SEED annotations and GrowMatch (Henry et al. 2009). First, an initial model was assembled based on the subsystems-based annotations in the SEED database. Reaction reversibilities were assigned based on thermodynamic calculations using the group contribution method, and then the model was refined with a modified version of GrowMatch. In this version of GrowMatch, the objective function has been modified so that the more plausible potential reactions (based on being in KEGG, having substrates with known structures, having an estimated ΔG° , and operating in the predicted thermodynamically favorable direction) are added before less plausible reactions. A gap-filling reconciliation step is also included that corrects the entire model by adding the minimum total number of new reactions predicted by GrowMatch while also minimizing the number of new false positives created.

Conclusions

There are now a number of different computational methods for identifying candidate gap-filling reactions and candidate genes to associate with orphan reactions. As more high-throughput data becomes publicly available in databases in the future, the effectiveness of these methods should only increase. Already, experimental validation of gap-filling predictions by SMILEY, SEED, and ADOMETA has demonstrated the utility of these methods. The other methods have not yet been fully tested, but their potential usefulness is immense. All of these methods are applicable to models of any organism, provided the right type of experimental data is available.

As new genomes are sequenced and annotated, new metabolic network reconstructions will be assembled for different organisms. As these new reconstructions are built, we can expect use of gap-filling methods to become a standard part of the reconstruction process, just as gap-filling was used in the recent reconstructions of *M. genitalium*, *Salmonella*, and *B. subtilis*. New gap-filling methods will also likely be developed in the next few years, and the combination and reconciliation of predictions from different methods such as SMILEY and GrowMatch could be used to generate even more plausible hypotheses. Another potential area for improvement of gap-filling methods is the inclusion of thermodynamic information. Several studies have already shown that the inclusion of thermodynamic constraints can improve the accuracy of constraint-based model predictions (Beard et al. 2004; Fleming et al. 2009; Fleming et al. 2010; Henry et al. 2006). So far, only studies using BNICE and the modified GrowMatch used to improve the iBsu1103 *B. subtilis* model have considered thermodynamic constraints when making gap-filling predictions. Although consideration of these additional constraints should improve the accuracy of gap-filling predictions in the future, many constraint-based thermodynamics methods are extremely computationally intensive, limiting their applicability at the present. Gap-filling methods have not been used very extensively in the past, but with the increasing availability of high-throughput data for many organisms, these methods are likely to lead to many new discoveries in the future.

Acknowledgments

This work was supported by National Institutes of Health grant number R01 GM057089.

References

- The Universal Protein Resource (UniProt). Nucleic Acids Res. 2010; 38(Database issue):D142–8. [PubMed: 19843607]
- AbuOun M, Suthers PF, Jones GI, Carter BR, Saunders MP, Maranas CD, Woodward MJ, Anjum MF. Genome scale reconstruction of a *Salmonella* metabolic model: comparison of similarity and differences with a commensal *Escherichia coli* strain. J Biol Chem. 2009; 284(43):29480–8. [PubMed: 19690172]
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215(3):403–10. [PubMed: 2231712]
- Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006; 2:2006.0008.
- Beard DA, Babson E, Curtis E, Qian H. Thermodynamic constraints for biochemical networks. J Theor Biol. 2004; 228(3):327–33. [PubMed: 15135031]
- Bochner BR, Gadzinski P, Panomitros E. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Res. 2001; 11(7):1246–55. [PubMed: 11435407]
- Breitling R, Vitkup D, Barrett MP. New surveyor tools for charting microbial metabolic maps. Nat Rev Microbiol. 2008; 6(2):156–61. [PubMed: 18026122]
- Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2008; 36(Database issue):D623–31. [PubMed: 17965431]
- Chen L, Vitkup D. Distribution of orphan metabolic activities. Trends Biotechnol. 2007; 25(8):343–8. [PubMed: 17580095]
- Chen LF, Vitkup D. Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biology. 2006; 7(2)
- Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 1998; 23(9):324–8. [PubMed: 9787636]
- Donnelly MI, Cooper RA. Two succinic semialdehyde dehydrogenases are induced when *Escherichia coli* K-12 is grown on gamma-aminobutyrate. J Bacteriol. 1981; 145(3):1425–7. [PubMed: 7009588]
- Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A. 2007; 104(6):1777–82. [PubMed: 17267599]
- Duarte NC, Herrgård MJ, Palsson BØ. Reconstruction and Validation of *Saccharomyces cerevisiae* iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. Genome Res. 2004; 14(7): 1298–309. [PubMed: 15197165]
- Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999; 402(6757):86–90. [PubMed: 10573422]
- Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007; 3(121)
- Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. Nat Rev Microbiol. 2009; 7(2):129–43. [PubMed: 19116616]
- Feist AM, Palsson BØ. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. Nat Biotech. 2008; 26(6):659–667.
- Fleming RM, Thiele I, Nasheuer HP. Quantitative assignment of reaction directionality in constraint-based models of metabolism: application to *Escherichia coli*. Biophys Chem. 2009; 145(2–3):47–56. [PubMed: 19783351]

- Fleming RM, Thiele I, Provan G, Nasheuer HP. Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *J Theor Biol.* 2010; 264(3):683–92. [PubMed: 20230840]
- Fong SS, Nanchen A, Palsson BØ, Sauer U. Latent pathway activation and increased pathway capacity enable *Escherichia coli* adaptation to loss of key metabolic enzymes. *J Biol Chem.* 2006; 281(12):8024–33. [PubMed: 16319065]
- Forster J, Famili I, Fu PC, Palsson BØ, Nielsen J. Genome-Scale Reconstruction of the *Saccharomyces cerevisiae* Metabolic Network. *Genome Research.* 2003; 13(2):244–53. [PubMed: 12566402]
- Freund Y, Schapire R. A decision-theoretic generalization of online learning and an application to boosting. *J Computer and System Sci.* 1997; 55(1):119–139.
- Fuhrer T, Chen L, Sauer U, Vitkup D. Computational prediction and experimental verification of the gene encoding the NAD⁺/NADP⁺-dependent succinate semialdehyde dehydrogenase in *Escherichia coli*. *J Bacteriol.* 2007; 189(22):8073–8078. [PubMed: 17873044]
- Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics.* 2004; 5:76. [PubMed: 15189570]
- Green ML, Karp PD. Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics.* 2007; 23(13):i205–11. [PubMed: 17646298]
- Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. *Bioinformatics.* 2005; 21(8):1603–9. [PubMed: 15613400]
- Henry CS, Jankowski MD, Broadbelt LJ, Hatzimanikatis V. Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys J.* 2006; 90(4):1453–61. [PubMed: 16299075]
- Henry CS, Zinner JF, Cohoon MP, Stevens RL. iBsu1103: a new genome-scale metabolic model of *Bacillus subtilis* based on SEED annotations. *Genome Biol.* 2009; 10(6):R69. [PubMed: 19555510]
- Herrgård MJ, Fong SS, Palsson BØ. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol.* 2006; 2(7):e72. [PubMed: 16839195]
- Huynen MA, Bork P. Measuring genome evolution. *Proc Natl Acad Sci U S A.* 1998; 95(11):5849–56. [PubMed: 9600883]
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28(1):27–30. [PubMed: 10592173]
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006; 34(Database issue):D354–7. [PubMed: 16381885]
- Karp PD. Call for an enzyme genomics initiative. *Genome Biol.* 2004; 5(8):401. [PubMed: 15287973]
- Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics.* 2002; 18(Suppl 1):S225–32. [PubMed: 12169551]
- Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, Paulsen IT, et al. EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.* 2009; 37(Database issue):D464–70. [PubMed: 18974181]
- Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics.* 2006; 7(177)
- Kharchenko P, Vitkup D, Church GM. Filling gaps in a metabolic network using expression information. *Bioinformatics.* 2004; 20(Suppl 1):I178–I185. [PubMed: 15262797]
- Kumar VS, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics.* 2007; 8:212. [PubMed: 17584497]
- Kumar VS, Maranas CD. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. *PLoS Comput Biol.* 2009; 5(3):e1000308. [PubMed: 19282964]
- Lee JM, Sonnhammer EL. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 2003; 13(5):875–82. [PubMed: 12695325]
- Lespinet O, Labedan B. ORENZA: a web resource for studying ORphan ENZyme activities. *BMC Bioinformatics.* 2006a; 7:436. [PubMed: 17026747]
- Lespinet O, Labedan B. Orphan enzymes could be an unexplored reservoir of new drug targets. *Drug Discov Today.* 2006b; 11(7–8):300–5. [PubMed: 16580971]

- Liu D, Reeves PR. *Escherichia coli* K12 regains its O antigen. Microbiology. 1994; 140 (Pt 1):49–57. [PubMed: 7512872]
- Loh KD, Gyaneshwar P, Markenscoff Papadimitriou E, Fong R, Kim KS, Parales R, Zhou Z, Inwood W, Kustu S. A previously undescribed pathway for pyrimidine catabolism. Proc Natl Acad Sci U S A. 2006; 103(13):5114–9. [PubMed: 16540542]
- Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Syst Biol. 2009; 3:37. [PubMed: 19321003]
- Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. Mol Syst Biol. 2009; 5:320. [PubMed: 19888215]
- Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010; 28(3):245–8. [PubMed: 20212490]
- Osterman A. A hidden metabolic pathway exposed. Proc Natl Acad Sci U S A. 2006; 103(15):5637–8. [PubMed: 16595627]
- Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol. 2003; 7(2):238–51. [PubMed: 12714058]
- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crecy-Lagard V, Diaz N, Disz T, Edwards R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33(17):5691–702. [PubMed: 16214803]
- Overbeek R, Disz T, Stevens R. The SEED: a peer-to-peer environment for genome annotation. Commun ACM. 2004; 47(11):46–51.
- Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999; 96(6):2896–901. [PubMed: 10077608]
- Paley SM, Karp PD. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. Bioinformatics. 2002; 18(5):715–24. [PubMed: 12050068]
- Palsson BØ. Metabolic systems biology. FEBS Lett. 2009; 583(24):3900–4. [PubMed: 19769971]
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96(8):4285–8. [PubMed: 10200254]
- Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of microbial production systems. Genome Res. 2004; 14(11):2367–76. [PubMed: 15520298]
- Pouliot Y, Karp PD. A survey of orphan enzyme activities. BMC Bioinformatics. 2007; 8:244. [PubMed: 17623104]
- Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BØ. Systems approach to refining genome annotation. Proc Natl Acad Sci U S A. 2006; 103(46):17480–4. [PubMed: 17088549]
- Reed JL, Vo TD, Schilling CH, Palsson BØ. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biology. 2003; 4(9):R54.1–R54.12. [PubMed: 12952533]
- Rodionov DA, Kurnasov OV, Stec B, Wang Y, Roberts MF, Osterman AL. Genomic identification and in vitro reconstitution of a complete biosynthetic pathway for the osmolyte di-myo-inositol-phosphate. Proc Natl Acad Sci U S A. 2007; 104(11):4279–84. [PubMed: 17360515]
- Rubires X, Saigi F, Pique N, Climent N, Merino S, Alberti S, Tomas JM, Regue M. A gene (wbbL) from *Serratia marcescens* N28b (O4) complements the rfb-50 mutation of *Escherichia coli* K-12 derivatives. J Bacteriol. 1997; 179(23):7581–6. [PubMed: 9393727]
- Schapire R. The boosting approach to machine learning: An overview. MSRI Workshop on Nonlinear Estimation and Classification. 2002
- Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, et al. The Stanford Microarray Database. Nucleic Acids Res. 2001; 29(1):152–5. [PubMed: 11125075]
- Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, Maranas CD. A genome-scale metabolic reconstruction of *Mycoplasma genitalium*, iPS189. PLoS Comput Biol. 2009; 5(2):e1000285. [PubMed: 19214212]

- Varma A, Palsson BØ. Metabolic capabilities of *Escherichia coli*: I. Synthesis of biosynthetic precursors and cofactors. *Journal of Theoretical Biology*. 1993a; 165(4):477–502. [PubMed: 21322280]
- Varma A, Palsson BØ. Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *Journal of Theoretical Biology*. 1993b; 165(4):503–522.
- Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, et al. The Protein Information Resource. *Nucleic Acids Res*. 2003; 31(1):345–7. [PubMed: 12520019]
- Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat Genet*. 2002; 31(3):255–65. [PubMed: 12089522]
- Yanai I, Derti A, DeLisi C. Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc Natl Acad Sci U S A*. 2001; 98(14): 7940–5. [PubMed: 11438739]

Figure 1.
Overview of the different algorithms used for predicting gap-filling reactions and orphan-filling genes.

Figure 2.

Examples of a gap (A) and an orphan reaction (B) in metabolic networks. Gaps occur when the reaction that consumes or produces a particular metabolite is completely unknown. Orphans occur when a particular reaction is known to occur, but it is not known which gene encodes the catalyzing enzyme. In this figure, metabolites are represented by orange circles and reactions by blue arrows connecting them. Gene-protein-reaction interactions are shown for each reaction.

Figure 3.

An example of how a cycle in a metabolic network can lead to a non-obvious gap. Reactions 2–5 form a cycle with reaction 1 as the only input or output. At steady-state, reaction 1 is unable to carry flux without violating mass balance constraints because of this cycle. The four reactions in the cycle are free to carry any flux (as long as all reactions carry the same flux), but metabolite A is blocked even though it has both producing and consuming reactions. Algorithms such as GapFind are useful for identifying these types of gaps.

Table ISummary of *in silico* gap-filling and orphan-filling methods.

Method	Gaps or orphans?	Data needed (other than the original metabolic network)	References
BNICE	gaps	a set of generalized enzyme reactions	(Hatzimanikatis et al. 2005)
GapFill	gaps	database of potential reactions (e.g. MetaCyc)	(Kumar et al. 2007)
SMILEY	gaps	growth phenotype data (e.g. Biolog), database of potential reactions (e.g. KEGG)	(Reed et al. 2006)
GrowMatch	gaps	gene essentiality data, database of potential reactions (e.g. MetaCyc)	(Kumar and Maranas 2009)
OMNI	gaps	metabolic flux data, database of potential reactions (e.g. KEGG)	(Herrgard et al. 2006)
PHFiller-GC	orphans	annotated gene/protein sequences from other organisms (e.g. Swiss-Prot, PIR), genomic-context data	(Green and Karp 2004; Green and Karp 2007)
SEED	orphans	annotated genome sequences from other organisms (the SEED)	(Osterman 2006; Osterman and Overbeek 2003)
ADOMETA	orphans	Co-expression data (e.g. SMR), annotated genome sequences from other organisms (e.g. GenBank), sets of gene clusters (e.g. KEGG)	(Chen and Vitkup 2006; Kharchenko et al. 2006; Kharchenko et al. 2004)

Table II

Summary of applications of gap-filling and orphan-filling methods and new discoveries made.

Discovery/application	Gap-filling method used	Validation methods used	References
<i>putP</i> as <i>E. coli</i> propionate transporter	SMILEY	gene knockout phenotype, upregulation of gene detected (RT-PCR)	(Reed <i>et al.</i> , 2006)
<i>idnT</i> as <i>E. coli</i> 5-keto-D-gluconate transporter	SMILEY	gene knockout phenotype, upregulation of gene detected (RT-PCR)	(Reed <i>et al.</i> , 2006)
<i>tolC</i> as a possible <i>E. coli</i> thymine transporter (not verified)	SMILEY	measured thymine excretion (HPLC), gene knockout phenotype	(Reed <i>et al.</i> , 2006)
<i>dctA</i> , <i>yeaU</i> , <i>yeaT</i> as D-malate uptake genes	SMILEY	gene knockout phenotypes, upregulation of genes detected (RT-PCR), assay of purified YeaU enzyme, ChIP analysis of YeaT	(Reed <i>et al.</i> , 2006)
pyrimidine catabolism pathway in <i>Y. enterocolitica</i> , <i>Acinetobacter sp. AP1</i> , <i>P. syringae</i> , <i>C. crescentus</i> , and <i>A. tumefaciens</i>	SEED	gene knockout phenotypes in <i>E. coli</i> , detection of pathway products in <i>E. coli</i> (GC/MS)	(Loh <i>et al.</i> , 2006; Osterman, 2006)
<i>dipA</i> and <i>dipB</i> in DIP synthesis pathway of <i>T. maritima</i>	SEED	assay of purified DipA enzyme, assay of DipB and entire pathway in cell extracts	(Rodionov <i>et al.</i> , 2007)
<i>yneI</i> as <i>E. coli</i> succinate semialdehyde dehydrogenase	ADOMETA	gene knockout phenotypes, assay of purified YneI enzyme	(Fuhrer <i>et al.</i> , 2007)
86 new reactions for iPS189 <i>M. genitalium</i> reconstruction	GapFill and GrowMatch	improved accuracy of model gene essentiality predictions	(Suthers <i>et al.</i> , 2009)
refinement of iMA945 <i>Salmonella</i> reconstruction	GapFill and GrowMatch	improved accuracy of model growth phenotype predictions	(AbuOun <i>et al.</i> , 2009)
32 new reactions for iBsu1103 <i>B. subtilis</i> reconstruction	modified GrowMatch	improved accuracy of model gene essentiality and growth phenotype predictions	(Henry <i>et al.</i> , 2009)