Journal of
Proteomics & Bioinformatics

Nguyen et al., J Proteomics Bioinform 2014, S9
DOI: 10.4172/0974-276X.S9-003

Research Article Open Access

Megafiller: A Retrofitted Protein Function Predictor for Filling Gaps in Metabolic Networks

Nam Ninh Nguyen[1]*, Wanwipa Vongsangnak[2,3]*, Bairong Shen[2], Phi-Vu Nguyen[1] and Hon Wai Leong[1]

[1]*Department of Computer Science, National University of Singapore, 117417, Singapore*
[2]*Center for Systems Biology, Soochow University, Suzhou 215006, China*
[3]*Department of Zoology, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand*

Abstract

Background: A bottleneck in investigating the cellular metabolism and physiology of organisms is the presence of metabolic gaps in the genome-scale metabolic networks. Metabolic gaps are reactions in the network that the corresponding genes have not yet been identified. Previous gap filling methods are generally based on identifying protein family in related organisms and then use this family to help for finding the target gene in a given genome. However, these methods fail when the protein family is not well-defined. There are therefore still many gaps in current metabolic networks. Here, we attempt to fill these gaps via an indirect approach by retrofitting protein function predictors and post-processing their results to identify the candidate genes.

Results: We developed a novel method for metabolic gap filling, called MeGaFiller that uses an ensemble of multiple retrofitted state-of-the-art protein function predictors. The ensemble scheme was adopted to boost the prediction performance. MeGaFiller can propose the candidate genes for 35% of the metabolic gaps in different metabolic networks (i.e. yeast, three filamentous fungi and bacterium). MeGaFiller can predict novel candidate up to hundreds genes for earlier annotated functions in the metabolic networks. MeGaFiller can also provide novel candidate genes for novel putative reactions throughout the metabolic networks.

Conclusions: MeGaFiller method demonstrates our first effort for filling metabolic gaps in the metabolic networks by retrofitted protein function predictors. It serves as a bioinformatics tool assisting for improved annotation through metabolic network reconstruction at a genome-scale.

Keywords: Ensemble scheme; Gap filling; MeGaFiller; Metabolic network; Retrofitted protein function predictors

Background

The metabolic network of a cell is the complete set of interconnected metabolic processes that determine the physiological and biochemical properties of the cell. In recent years, metabolic networks have enormously contributed to our understanding of metabolic genotype and phenotype relationship. This leads to important applications through systems biology and metabolic engineering. Recently, meta-proteome-scale metabolic network reconstruction has also emerged as a promising and challenging approach for investigating the metabolism of microbial communities [1].

In recent years, there has been an effort to reconstruct the genome-scale metabolic networks for hundreds species [2-7]. In principle, the reconstruction of metabolic networks is an iterative multi-stage process [8,9], which starts from gene annotation, and goes all the way to network development. Several sophisticated techniques have been developed for metabolic network reconstruction [10-15].

Metabolic Gaps and their Implications

However, most of the reconstructed networks remain incomplete, namely there are significant numbers of metabolic gaps [16,17]. A metabolic gap in a network for genome (G) is a metabolic reaction (R) (described by its EC number) that is present in the network. But the annotation through the network reconstruction methods have failed to find the corresponding gene in G that is responsible for that reaction R. We distinguish two types of metabolic gaps: local metabolic gap where the corresponding gene responsible for R can be found in other related organisms and global metabolic gap where the corresponding gene responsible for R has not been found in any known organism or have not been so annotated in any genome.

Metabolic gaps impede downstream biological analysis of these metabolic networks. For examples, in the reconstructed metabolic networks of yeast *Saccharomyces cerevisiae* [2], filamentous fungi *Aspergillus oryzae* [3], *Aspergillus nidulans* [4], and *Aspergillus niger* [5], and bacterium *Streptomyces coelicolor* [6], between 6% to 19% of the biochemical reactions are metabolic gaps.

Filling these metabolic gaps (and thus, enhancing these networks) is the most time-consuming task that may take years to complete since there is a lot of manual curation involved [16-18]. This can be a bottleneck for gaining high quality metabolic networks. Our work therefore proposes to fill those metabolic gaps by considering new algorithmic methods.

Current Metabolic Gap Filling Methods

Current direct methods for filling local metabolic gaps (i.e. genes that are un-annotated in the target organism, but have been found in

***Corresponding authors:** Nam Ninh Nguyen, Department of Computer Science, National University of Singapore, Singapore 117417, Tel: +6597623768; E-mail: nguyennn@comp.nus.edu.sg

Wanwipa Vongsangnak, Center for Systems Biology, Soochow University, Suzhou 215006, China, E-mail: wanwipa@suda.edu.cn

Received December 15, 2013; **Accepted** June 17, 2014; **Published** June 20, 2014

Citation: Nguyen NN, Vongsangnak W, Shen B, Nguyen PV, Leong HW (2014) Megafiller: A Retrofitted Protein Function Predictor for Filling Gaps in Metabolic Networks. J Proteomics Bioinform S9: 003. doi:10.4172/0974-276X.S9-003

other related organisms) are based on the profile of the protein family, such as GFAOP [3] and Reed et al. [19]. In these methods, the first step is to identify the specific protein family for the metabolic gap. The next step retrieves and/or builds a family profile of this protein from reference databases. Then this profile is used to search against the target organism (i.e. genome sequence) to detect candidate gene. Finally, candidate gene (if any) is manually validated by querying with annotation databases. These gap filling methods have been successfully used to fill some of the gaps in the metabolic networks of *Escherichia coli* [19] and *A. oryzae* [3]. Despite that, many gaps still remain in these networks.

The success (and failure) of these direct gap filling methods is highly dependent on the first step of identification of the protein family for the metabolic gaps. This step requires an expert to precisely interpret the proper family that performs the intended enzymatic function (described by an EC number) of the metabolic gap, specific up to substrate binding. However, this first step is challenging for two reasons. Firstly, the protein family is not well-defined. For example, in the Swiss-Prot database, there are 251 annotated genes encoding for the EC 2.1.1.64. Of these, 242 genes have no known specific protein family which are available.

Secondly, even if the protein family is found, it may not be particular enough to be helpful for finding the candidate gene with a specific function. Based on Swiss-Prot database, for example, there are a total of 673 annotated genes encoding for the EC 2.7.7.3 and all of them carry the Cytidylyltransferase (PF01467) domain. Meanwhile, there are the other 527 annotated genes encoding for the same domain, but show different functions. This indicates that the domain is too general and hence the generated protein family profile will not be specific enough to find good candidate gene for the EC 2.7.7.3. We noticed that a current *A. oryzae* metabolic network, 52 gaps (out of 61 gaps) do not have a specific protein family interpretation. Hence, these direct gap filling methods have failed in the beginning step as shown in example case of the *A. oryzae* metabolic network (*iWV1314*).

Overview of Our Approach

In this work, we aimed to develop MeGaFiller, an ensemble of indirect approach to overcome the difficulties of existing direct methods in cases of poorly characterized protein family. Our indirect approach leverages the following duality between gap filling and protein function prediction. In gap filling, we determine protein function (*f*), and we desire to search for candidate gene (*p*) in the genome sequence of the target organism with the protein function *f*. In protein function prediction, we determine a candidate gene *p*, and we desire to predict its protein functions *f*. Thus, gap filling and protein function prediction are dual problems of one another.

In theory, therefore, if we have a candidate gene *p* in the target organism (genome *G*) that performs a protein function *f*, then the "perfect" gap filler is able to find gene *p* in *G* given the protein function *f*, and the "perfect" protein function predictor is able to predict protein function *f* given the gene *p* in *G*. In reality, however both procedures are far from the "perfect". Current direct gap filling methods are not able to find the genes encoding for some protein functions, as evident from the gaps in current metabolic networks. Additionally, current protein function predictor may miss some protein function *f* due to very low scores from the prediction.

We propose to use the dual approach using protein function prediction methods to help find candidate gene *p* that have the predicted protein function *f*. We initially predict the functions of all the proteins in a target organism (genome *G*) using a protein function predictor. Then, we keep only the genes with predicted protein functions that are matched to those of the metabolic gaps. Once completed, the list of candidate genes is generated for the metabolic gaps.

There are additional reasons to pursue this dual approach. Firstly, there has been tremendous progress in the state-of-the-art protein function prediction methods and they have vastly improved recall and precision rates. Many enhancements have been developed for BLAST-based protein function predictors and these included Gotcha [20], PFP [21], and Blast2GO [22,23]. There have also been enhancements to protein function predictors that used Hidden Markov Model (HMM) profiles with improved prediction accuracy and these included ModEnzA [24] and EFICAz [25,26]. Hence, it is timely to leverage on these state-of-the-art protein function predictors to help find candidate genes for the "difficult-to-fill" metabolic gaps. Another crucial reason is that our dual approach does not require knowledge of the specific protein family and hence gets around the inherent challenge of identifying the proper family at the beginning. Thirdly, even in case where the protein family is not well-defined, there may still be some of individual protein list in the annotation databases that can infer function. Several protein function predictors can leverage on the protein list (via sequence similarity for BLAST-based methods or profile similarity for HMM based methods) to help in their function prediction. In this way, the protein list may help, indirectly, to predict the given metabolic function for achieving the candidate genes, without having the certain protein family.

To successfully use this dual approach, we identified good protein function predictors that were suitable for retrofitting for the purpose of filling metabolic gaps. We focused on protein function predictors that gave more predicted functions, even those with low scores. To deal with a large pool of predictions, we needed a retrofitting procedure to carefully filter and find the "difficult-to-fill" candidate genes for the metabolic gaps (Details given in the next section). Based on this, we retrofitted gap fillers based on the different state-of-the-art function predictors. After evaluating their individual effectiveness, we found that no one method dominates the rest and each of them has different strengths and weaknesses. General protein function predictors tend to give many more predicted functions (per protein) and thus, achieve higher coverage, but lower accuracy. In contrast, enzyme-specific function predictors give fewer predicted functions (per protein) and gain higher accuracy, but lower coverage.

To leverage on their different relative strengths, we developed novel MeGaFiller (Metabolic Gap Filler) method that uses a weighted ensemble of the individual retrofitted protein function predictors. We then optimized the relative weights of the individual protein function predictors within MeGaFiller. This optimization of MeGaFiller was then performed separately in each of five species, as our performance analysis showed that the optimal parameter setting is species-dependent. To the end, the optimized MeGaFiller was used to fill the gaps in five different metabolic networks. MeGaFiller showed effective in filling metabolic gaps remaining in these networks. It was also able to predict more candidate genes for existing reactions and novel putative reactions in these metabolic networks.

Methods

Our proposed method is called MeGaFiller, which carefully combined the prediction results of several individual gap fillers based on retrofitted protein function predictors. In the following, we first

explain how these individual gap fillers were designed, and then discuss how they were retrofitted into MeGaFiller.

In the dual approach for gap filling, we used protein function predictors to indirectly find candidate genes that have the predicted protein functions of the metabolic gaps. For a given metabolic network of a target genome *G*, and any chosen protein function predictor (*FP*), the general procedure is as follows: First, we used *FP* to predict the functions of all the proteins in a target genome. This produces, for each gene *p* in the target genome *G*, a list of predicted protein functions (given in EC numbers or GO terms). Usually, these predicted protein functions were ranked based on a variety of scores (e.g. confidence, significance) depending on the predictor used. Metabolic gaps were given by their EC numbers. Hence, the next step was to use EC2GO mapping [27] to map any predicted GO terms into EC numbers. Next, we matched these predicted protein functions with the gaps in the metabolic network. For each metabolic gap with EC number, we collected all the genes for which the protein function *f* that was predicted by *FP* and formed the list of candidate genes for the metabolic gaps.

We pointed out that the ranking/scores of candidate gene *p* for a given metabolic gap were produced by considering candidate gene *p* individually, and the ranking/scores were given relative to the other predicted protein functions for that candidate gene *p*. Hence, it does not make sense to directly compare the ranking/scores of genes in the candidate list. Hence, we may need to do some retrofitting by post-processing of the candidate list (for example, re-ranking by their confidence/significance scores) to produce the list of missing gene candidates.

With the recent advances and proliferation of protein function predictors, we selected suitable ones. For our purposes, we focused on state-of-the-art protein function predictors that are also relatively easy to retrofit for metabolic gap filling. We thus selected two general protein function predictors, PFP [21] and Blast2GO [22,23] that can predict general protein functions. We also selected an enzyme-specific function predictor, EFICAz [25,26]. We ran all the three protein function predictors using their default settings (see Appendix A for more details).

Next, we describe how we retrofitted them for gap filling and called them as PFP-GF, B2G-GF, and EFICAz-GF, respectively.

PFP-GF: Retrofitted PFP for gap filling

Among the general protein function predictors tested, PFP [21] gave the most predicted functions with the best coverage (but lower accuracy). Given an input as protein sequences, PFP uses PSI-BLAST on the Uniprot database to predict a list of GO terms, each of which comes with several scores (i.e. rank, raw score, P-value, and three different confidence scores). By default, PFP sorts GO terms by the "4-edge confidence score". This list contains many predicted protein functions (some may contain up to 500 GO terms), including many predicted protein functions with very low scores.

We retrofitted the protein function predictor PFP for gap filling as shown in Figure 1. After PFP was done, the output GO terms were mapped into EC numbers using EC2GO [27]. Then, for each EC number, we re-ranked the list of candidate genes based on the following criteria (in priority order): confidence score, raw score, and P-value. A top-rank cut-off to filter the candidate list was chosen to maximize the F_2 score. The best F_2 score for PFP-GF was obtained when setting this cut-off to keep only the top 5 candidates.

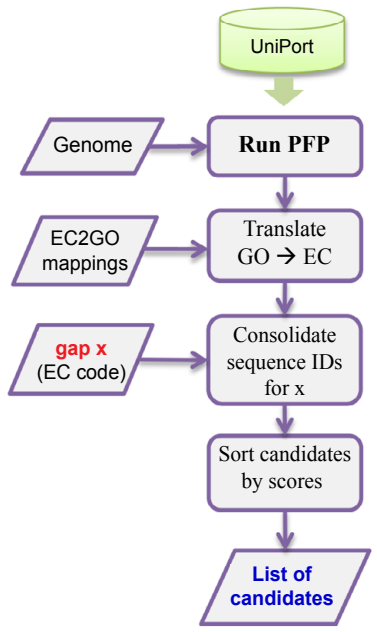

Figure 1: Retrofitting procedure for gap filler PFP-GF. The original PFP takes a genome sequence as an input and generates a list of predicted GO terms for each sequence. PFP-GF takes the PFP's output, the EC2GO mapping, and a metabolic gap described by an EC number. Then, PFP-GF generates the sorted list of candidate genes for the gaps.

B2G-GF: Retrofitted Blast2GO for gap filling

The other general protein function predictor, Blast2GO, is very simple to retrofit. Blast2GO produces EC numbers for their predicted functions. It produces relatively few predicted functions per proteins and they tend to have high accuracy. Thus, the retrofitting procedure consists of just matching predicted functions per proteins with the given set of EC numbers, and consolidating all the candidate genes for each of the given EC numbers.

EFICAz-GF: Retrofitted EFICAz for gap filling

EFICAz is an enzyme-specific function predictor. It predicts only enzyme-specific functions. It gives relatively few predicted functions and these have high accuracy. The predicted protein functions are also given by their EC numbers. Thus, like B2G-GF, retrofitting consists of just matching predicted functions per proteins with the given set of EC numbers, and consolidating all the candidate genes for each of the given EC numbers.

Our proposed method: MeGaFiller

To achieve better prediction results and higher confidence, our proposed method, MeGaFiller, considers an ensemble of these three individual gap fillers. For this integration, we need to handle the fact that the three individual gap fillers produce different types of prediction scores. Specifically, PFP-GF gives several scores, and, after our performance analysis, the confidence score was chosen as the score for PFP-GF. In contrast, B2G-GF ranking relies on the BLAST hit score, thus the best bit score of hits was chosen as the score of candidates for B2G-GF. Finally, EFICAz-GF does not produce any prediction score, and so all its predictions were equally weighted (i.e. all have score of 1.0).

To rationally combine these gap fillers, a generic weighted ensemble

scheme (similar to [28]) was adopted, as shown in Figure 2. Firstly, for each EC number (e), the scores for each candidate gene p for e were normalized using as follows:

$$S_N^i(e,p) = \frac{s^i(e,p)}{\max_k s^i(e,p_k)}$$

where $S_N^i(e,p)$ is the normalized score and $s^i(e,p)$ is the predicted score of pair (e, p) produced by gap filler (i). Secondly, we assigned weight (w_i) to each gap filler i as a measure of its prediction significance. These were also normalized, i.e. $\sum(w_i) = 1$.

The combined score, $C(e,p)$, for a pair of a candidate gene p and an EC number e, was given by a weighted summation over all individual gap filler, as follows:

$$C(e,p) = \sum_i w_i \cdot S_N^i(e,p)$$

To the end, the candidate gene list was filtered, keeping only candidates with the combined scores not below a given threshold (θ).

Data sources

In this research, we focused on using MeGaFiller to fill metabolic gaps in the reconstructed metabolic network of *A. oryzae iWV1314* [3]. However, we have also run MeGafiller on the other four networks, those

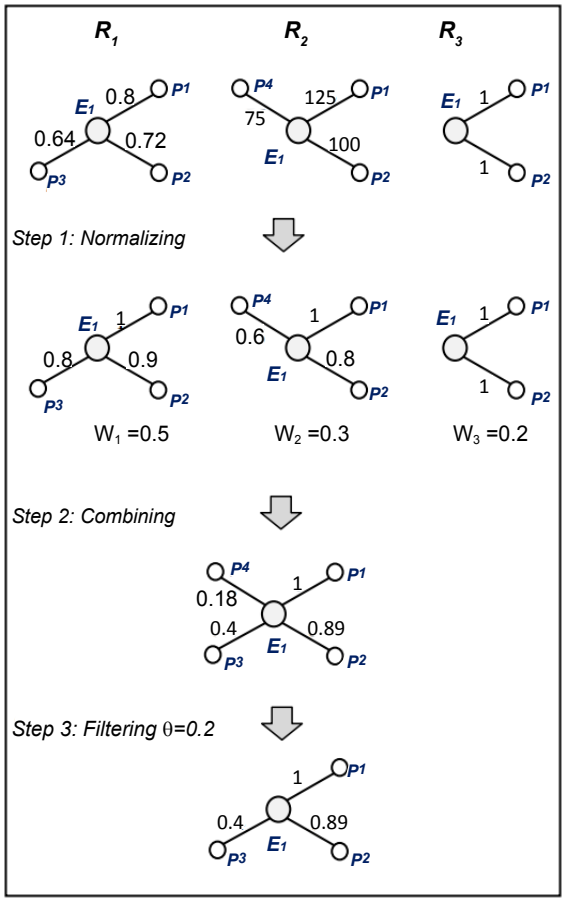

Figure 2: Ensemble scheme for MeGaFiller. A big circle denotes an EC number, while a small circle presents a candidate gene. Scores of predictions made by each predictor were normalized. Then, the predictions were combined by the weight w_i of each predictor. Finally, MeGaFiller applied a threshold θ to filter the final prediction list.

Network name	Species	Strain	Release	Reference	Genome Source
iIN800	*Saccharomyces cerevisiae*	S288c	2008	Nookaew et al. [2]	SGD[1]
iWV1314	*Aspergillus oryzae*	RIB40	2008	Vongsangnak et al. [3]	DOGAN[2]
iHD666	*Aspergillus nidulans*	FGSC A4	2008	David et al. [4]	AspGD[3]
iMA871	*Aspergillus niger*	ATCC1015	2008	Andersen et al. [5]	AspGD[3]
iIB711	*Streptomyces coelicolor*	A3(2)	2005	Borodina et al. [6]	*S. coelicolor* genome[4]

Metabolic networks were retrieved from BioMet Toolbox website (http://biomet-toolbox.org/). The corresponding genome sequence was downloaded from sources listed in last column.

[1]http://www.yeastgenome.org/
[2]http://www.bio.nite.go.jp/dogan/top
[3]http://www.aspgd.org/
[4]http://www.sanger.ac.uk/resources/downloads/bacteria/streptomyces-coelicolor.html

Table 1: Metabolic networks used in our study.

Network characteristics	*iIN800*	*iWV1314*	*iHD666*	*iMA871*	*iIB711*
Number of gene-EC number pairs	573	1,329	847	818	694
Number of unique EC numbers	481	711	455	482	407
Number of metabolic reaction without genes	83	65	29	112	72
Number of unique EC numbers without genes (number of gaps)	***52***	***61***	***28***	***90***	***68***
Percentage of metabolic gaps	11%	9%	6%	19%	17%

This table shows detailed content of the five metabolic networks. Unique EC numbers are accounted for the set of EC numbers appeared in the network. Number of metabolic gaps is the number of unique metabolic reactions that have no genes annotated. Percentage of gaps is calculated as ratio of gaps over unique EC numbers.

Table 2: Metabolic network contents.

for *S. cerevisiae iIN800* [2], *A. niger iMA871* [5], *A. nidulans iHD666* [4], as well as *S. coelicolor iIB711* [6]. These metabolic networks were obtained from BioMet Toolbox [29] website (http://biomet-toolbox.org/)

Each of these metabolic networks contains two datasets: the known dataset (N_K) and the metabolic gap dataset (N_G). N_K is a list of known pairs of metabolic reaction associated between EC numbers and gene. The N_K was used to tune the parameters of MeGaFiller and its component gap fillers. N_G is a list of metabolic gaps associated between EC numbers and gap reactions. From each metabolic network of the five species, we extracted N_K and N_G and the statistics are shown in the Tables 1 and 2.

Concerning on the genome sequences of these five species, we retrieved them from reference databases *S. cerevisiae* S288c-SGD (http://www.yeastgenome.org/) [30], *A. oryzae* RIB40-DOGAN (http://www.bio.nite.go.jp/dogan/top) [31], *A. niger* ATCC1015 and *A. nidulans* FGSC A4-AspGD (http://www.aspgd.org/) [32] and *S. coelicolor* (http://www.sanger.ac.uk/resources/downloads/bacteria/streptomyces-coelicolor.html) [33]. We used these genomes to extract the list of protein sequences to serve as input to MeGaFiller (and to the individual metabolic gap fillers).

Performance metrics

To tune the parameters of MeGaFiller and its component gap fillers, we ran them on the known datasets N_K of the five species and measured the following performance metrics: precision, recall, and F_2

score, calculated as follows: Suppose that a gap filler method proposes a candidate gene p for an input EC number e. Then the predicted pair (e, p) is said to be a true positive if (e, p) is in the N_K; otherwise it is called as a false positive. A pair (e, p) is called a false negative if it is in the N_K, but is not predicted by the gap filler. Let *TP, FP* and *FN* denote the number of true positive, false positive and false negative, respectively. Then the performance metrics are calculated as the following:

$recall=TP/(TP+FN)$

$precision=TP/(TP+FP)$

$F_2=5*precision*recall/(4*precision+recall)$

We used F_2 score to put more weight (double the importance) on the prediction coverage (recall), which is more important than the prediction precision for the purpose of finding missing gene candidates.

The aim of MeGaFiller is to find candidate genes to fill gaps that existing direct gap filling methods have already failed. So, MeGaFiller is biased towards higher true positives and finding new candidate genes, at the expense of an increase in the number of false positives. In other words, it is more important to be able to find the candidate genes while incorrect predictions may be ruled out later by manual curation with the help of additional independent data sources.

Parameter tuning for MeGaFiller

MeGaFiller has several parameters, the weights w_i for the component gap fillers and the threshold (θ). We tuned parameters using the known datasets as follows: Consider a species with genome G and reconstructed metabolic network N as $N_K(G)$. Let $R(G)$ is the list of metabolic reactions (given by their EC numbers) in $N_K(G)$. For any given values of the parameters w_i and θ, we ran MeGaFiller using the genome G and the $R(G)$ to predict pair (e, p), namely to predict gene p in G for the EC number e in $R(G)$. Let $P(w_i, \theta, G)$ denotes this list of predicted pair obtained by MeGaFiller with parameters w_i and θ. We then matched the predicted pair with $N_K(G)$ to compute $F_2(P(w_i, \theta, G) \mid N_K(G))$, the F_2 score for this parameter setting. These parameters are tuned by optimizing the F_2 score:

$$(w_i, \theta)_G = argmax\ \{F_2(P(w_i, \theta, G) \mid N_K(G))\}$$

All parameters were searched in the range (0,1) with step-size of 0.01.

Manual curation of candidate genes for metabolic gaps

For new predictions made by MeGaFiller (and others methods) to fill the metabolic gaps in various networks, there is no ground truths. To evaluate, hence the reliability of these new predictions by MeGaFiller, we manually curated some of them to provide independent supporting evidences. For each predicted pair (e, p), we used the protein sequences of the candidate gene p to search against several annotation databases, such as NR (http://www.ncbi.nlm.nih.gov/refseq/), UniProt [34], Pfam [35], CDD [36] and KEGG [37] databases to look for significant hits with annotation of the function e. We also looked at the updated function annotation of the candidate p if such annotation exists in genome databases of the species. All the relevant supporting information for the candidate gene p was collated for further manual curation.

Results

We now present our extensive results which are organized as follows: We first show an evaluation of individual retrofitted gap fillers using the known datasets from the five species studied. We show how these results supported our use of an ensemble scheme in MeGaFiller. Next, we provide results on parameter tuning of MeGaFiller. We show that MeGaFiller (with default optimal parameters) outperformed the individual component gap fillers with the highest recall and F_2 score.

We then describe the main result of this study, namely using MeGaFiller for gap filling and the ability of MeGaFiller to fill critical gaps in the *A. oryzae* network. Besides, we then show comparison of our method, MeGaFiller with two other existing methods: GFAOP [3], a homology-based method and ADOMETA [38], a context-based method. Finally, we discuss how to use MeGaFiller to predict novel candidate genes for existing reactions and/or predict novel putative reactions in the metabolic network for a given species.

Evaluation of the individual retrofitted gap fillers

To evaluate the relative performance of the retrofitted gap fillers (PFP-GF, B2G-GF, EFICAz-GF), we ran each of them on the same known datasets of the different species. A completely uniform comparison was not possible since each method relies on its own (different) reference data that were retrieved at different timestamps. To do a fairer comparison, we restricted the comparison to only those reactions that found in all three methods. We also excluded those EC numbers that were too general and focused on those that were specified in all 4 digits. We noted that, across the 5 genomes, between 1.0-8.3% of the pairs, were excluded by this process (and overall of only 4.7%).

After this pre-processing, the known dataset for metabolic network *iIN800* contains 573 known gene-EC number pairs, while the known dataset for network *iWV1314* contains 1,329 known pairs. (The numbers of gene-EC number pairs in known datasets for *iMA871*, *iHD666* and *iIB711* are 818, 847 and 694 pairs, respectively).

Figure 3 shows the prediction results of PFP-GF, B2G-GF, and EFICAz-GF on the known datasets for *iIN800* and *iWV1314* datasets. For the well-studied yeast *iIN800* dataset, PFP-GF gave the largest number of true positive predictions, 520 out of 573. This was followed by EFICAz-GF with 445 true positive predictions while B2G-GF gave 402 true positives. However, PFP-GF also predicted more false positives (1,438) compared to B2G-GF (164) and EFICAz-GF (119).

For the relatively less well-studied filamentous fungus *iWV1314* dataset, PFP-GF and B2G-GF gave many more true positive results (833 and 844 out of 1,329) than EFICAz-GF (445). PFP-GF also had the highest false positives (2137), while the enzyme-specific EFICAz-

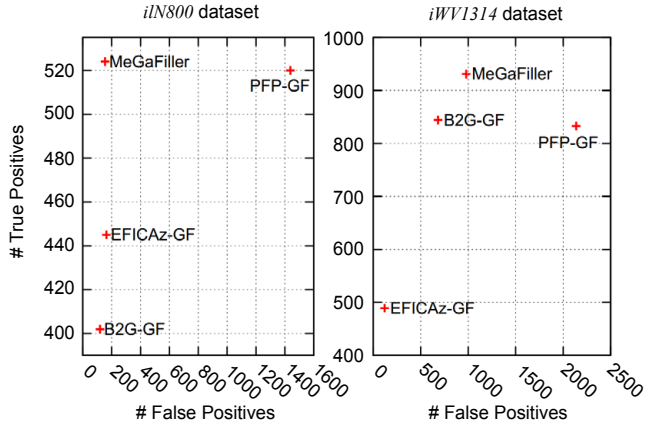

Figure 3: Performance evaluation of gap fillers. The dot plots show the number of true positives and false positives on iIN800 (*S. cerevisiae*) dataset (left), and *iWV1314* (*A. oryzae*) dataset (right) by component gap fillers and the MeGaFiller.

GF had the lowest (119), and B2G-GF was in between (682), but closer to EFICAz-GF. The results for the other three genomes (also less well-studied and annotated) were similar to that for *iWV1314* and are not shown in Figure 3.

We combined the results for all five genomes and observed that the retrofitted general protein function predictors gave higher average recall (PFP-GF: 67.5%, B2G-GF: 57.2%) compared to retrofitted enzyme-specific function predictors (EFICAz-GF: 46.4%). EFICAz-GF has higher precision (77.5%) than B2G-GF (62.4%) and PFP-GF (27.4%).

This can be explained by the fact that, enzyme-specific function predictors like EFICAz focuses on accurately classifying only enzymatic functions. Hence, they tend to make fewer predictions that are more accurate (higher precision) and they may lose in prediction coverage (lower recall), especially on datasets of less well-studied species. On the other hand, PFP, being a general protein function predictor gave more predictions than either EFICAz or Blast2GO, but suffered from lower precision.

Evidence to support an ensemble approach

We compared the actual predictions made by PFP-GF, B2G-GF, and EFICAz-GF for the *iIN800* and *iWV1314* datasets. The Venn diagram is shown in Figure 4. For each species, we observed that the number of true positive predictions in the 3-way common intersection is quite small, and each of gap filler produced its own unique true positive predictions. We also observed that PFP-GF gave the most unique true positive predictions, but with a high false positive rate. Thus, combining the prediction results of PFP-GF with that of B2G-GF or EFICAz-GF will increase the *precision*.

More generally, this suggests that we need to leverage on the results of all the gap fillers. Our method, MeGaFiller used a weighted ensemble scheme to combine the results of all three gap fillers to leverage on the high recall of PFP-GF and the high precision of the B2G-GF and EFICAz-GF.

Parameter tuning for MeGaFiller

To tune MeGaFiller, the known datasets of the five species for different parameter settings were used. The weights w_i for the component gap fillers and the threshold θ were searched for the parameter setting in order to obtain the highest F_2 score. As mentioned earlier, the parameters were searched in the range (0,1) with step-size of 0.01.

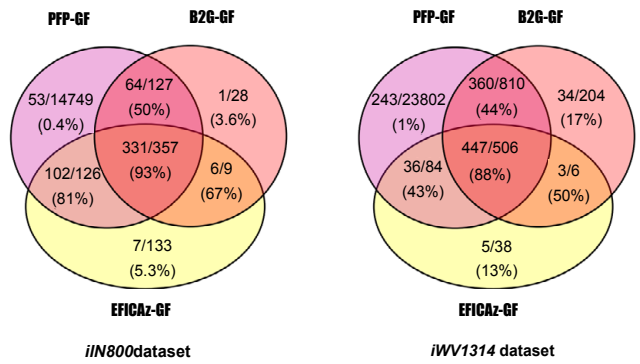

Figure 4: Intersection of predictions made by different gap fillers on testing datasets. Common predictions made by different methods and the corresponding correctly predicted portion for *iIN800* (left) and *iWV1314* (right) datasets. There are 357 predictions made by all gap fillers on *iIN800* dataset, in which 331 predictions are correct (93%). None of methods can cover all possible correct predictions.

Network dataset	w_1 (PFP-GF)	w_2 (B2G-GF)	w_3 (EFICAz-GF)	Θ
iIN800	0.60	0.10	0.30	0.63
iWV1314	0.43	0.47	0.10	0.40
iHD666	0.65	0.20	0.15	0.63
iMA871	0.40	0.50	0.10	0.40
iIB711	0.13	0.30	0.57	0.30

Values of parameters (w_i, θ) optimized by maximizing F_2 score on the known dataset N_K for each network. These weights and thresholds were used for gap filling on corresponding metabolic network.

Table 3: Optimal parameters for MeGaFiller.

We found that the optimal parameter setting was species dependent. So, we optimized the parameters for each species (dataset) separately as shown in Table 3. These were used as default settings for MeGaFiller on corresponding dataset. As can be seen, the *iIB711* dataset (the last one) gave different results from the other 4 datasets. So, we discuss results for the first 4 datasets. The optimal weights w_i for PFP-GF and B2G-GF were high while the optimal weights for EFICAz-GF were the lowest (except for the *iIB711* dataset). The optimal threshold θ ranged from 0.4 to 0.63. For each dataset, the optimal θ was quite close to the highest weight w_i. This suggests that the score of the highest weight component gap filler (usually PFP-GF) must be very close to 1 or it is made by at least two component gap fillers (e.g. Recall that EFICAz-GF scores are all 1.0, while PFG-GF scores range from 0.1 to 1). This result is consistent with our earlier stated objective of (a) leveraging the high recall of PFP-GF and (b) increasing the precision by having it "confirmed" by the other component gap filler.

Evaluation of MeGaFiller on known datasets

After parameter tuning, the optimized parameter settings were then used as default for MeGaFiller on the given dataset. The prediction results of MeGaFiller on known datasets for *iIN800* and *iWV1314* are shown in Figure 3. For both species, MeGaFiller produced more true positives than any one of its component gap fillers. In addition, the number false positive has improved drastically as compared to PFP-GF.

For the yeast *iIN800* dataset, the number of true positives from MeGaFiller was 524, compared to 520 for PFP-GF, 445 for B2G-GF and 402 for EFICAz-GF. Additionally, the number of false positive was 154, drastically lower than 1,438 for PFP-GF, and comparable to those of EFICAz-GF (164) and B2G-GF (119). For the filamentous fungus *iWV1314* dataset, the number of true positives for MeGaFiller was 931, which was much better than 833 for PFP-GF, 844 for B2G-GF and 489 for EFICAz-GF. While the number of false positive went down to 979, compared with 2,137 for PFP-GF, 682 for EFICAz-GF and 119 for B2G-GF.

Figure 5 shows the F_2 score of MeGaFiller and the individual gap fillers, (PFP-GF, B2G-GF, EFICAz-GF) on the known datasets of all five species. Clearly, MeGaFiller was consistently better than all of its component gap fillers over all the five datasets. In particular, we noted that the *iMA871* dataset, MeGaFiller had F_2 score of almost 60%, even when all the component gap filler had F_2 score below 50%. Summing over all the five species studied, MeGaFiller achieved average F_2 score of 68%, with average recall of 73% and average precision of 55%.

Comparing MeGaFiller with other variants of ensemble scheme

To further analyse the contribution of the weighted ensemble

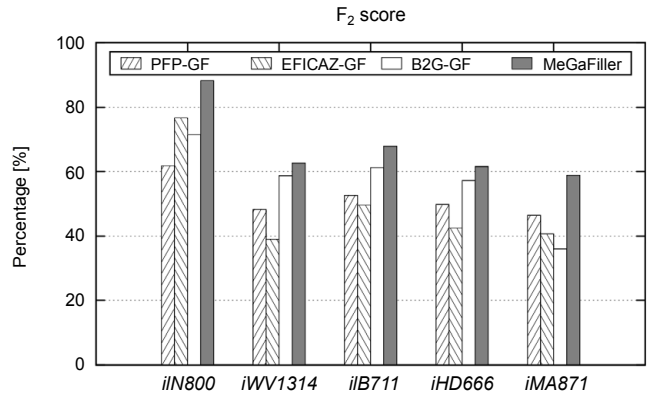

Figure 5: Relative performance of different gap fillers and MeGaFiller. F_2 score is shown for each method on 5 network datasets. MeGaFiller achieved the highest value, which outperformed all components.

Network dataset	MeGaFiller	Non-weighted	Common-2
iIN800	**88.27**	86.71	86.40
iWV1314	**62.67**	60.48	59.86
iHD666	**61.55**	58.77	58.39
iMA871	**58.86**	54.53	53.64
iIB711	**67.91**	67.14	66.91

F_2 score [%] of our weighted ensemble (MeGaFiller) is always better than that of non-weighted version and Common-2 voted version. The non-weighted version was run the same as weighted version, except that the weights were fixed equally. The common voted version was run by taking only predictions that were made by at least 2 component gap fillers (ignoring both weights and scores). The largest value for each row is shown in bold. The non-weighted version slightly performed better than the Common-2 voted version.

Table 4: Performance of different variants of our ensemble gap fillers.

Network dataset	*iIN800*	*iWV1314*	*iHD666*	*iMA871*	*iIB711*
Number of metabolic gaps	52	61	28	89	68
Number of metabolic gaps putatively filled	***15***	***12***	***17***	***23***	***25***
Number of putative candidates	***25***	***15***	***68***	***61***	***64***
Number of percentage metabolic gap filled	29%	20%	61%	26%	37%

For *A. oryzae* metabolic network (*iWV1314*), MeGaFiller predicted one or more candidate genes for 12 (out of 61 (20%)) metabolic gaps (and a total of 15 candidates).

Table 5: Number of putatively filled metabolic gaps for the five metabolic networks.

scheme used in MeGaFiller, we compared with two other ensemble variants. The first is a Non-Weighted ensemble in which the weights are equal (in this case, w_i =1/3 for each of the 3 components). The second, called Common-2 ensemble, uses simple voting and keeps only predictions made by at least 2 component gap fillers (this version ignores all the weights and the scores in MeGaFiller).

We repeated the evaluation on these two ensemble variants and compared them with MeGaFiller. Table 4 shows the F_2 score of the three ensemble variants. As expected, the three variants have very similar F_2 scores. Table 4 also shows that MeGaFiller (the weighted version) achieved the highest F_2 score over all five datasets. The non-weighted ensemble variant performed only slightly better than the Common-2 ensemble variant. We suggest that if no parameter tuning can be done, then the Common-2 ensemble variant based on simple voting may also be a good strategy.

Effectiveness of MeGaFiller in Filling Gaps in Metabolic Networks

We evaluated the effectiveness of MeGaFiller in filling the metabolic gaps in the metabolic gap dataset N_G for the five metabolic networks. For each metabolic gap from the N_G, MeGaFiller produced a list of candidate genes, if any, from the target genome that perform the function of the metabolic gap. We measured the number of gaps filled and the total number of candidate genes predicted for these gaps.

Table 5 shows the results obtained by MeGaFiller. For each network, Table 5 shows the number of metabolic gaps, the number of gaps filled (putatively with at least one candidate gene), the total number of candidate genes predicted for these gaps, and the percentage of gaps putatively filled by MeGaFiller. For the *iIN800* network, MeGaFiller putatively filled 15 out of 52 gaps (29%), and for *iWV1314* network, it putatively filled 12 out of 61 gaps (20%). It obtained even better results for two of the less well-studied species which were 37% for *iIB711* network and 61% for *iHD666* network. On average, MeGaFiller putatively filled 35% of the metabolic gaps in the five networks. In following description, we highlight some results from MeGaFiller.

Example 1: Consider the Fumarylacetoacetate hydrolase reaction (EC 3.7.1.2) in Phenylalanine, tyrosine, and tryptophan biosynthesis pathway in the *S. cerevisiae iIN800* network. This is currently a metabolic gap in the *S. cerevisiae iIN800* network. MeGaFiller predicted the candidate gene identifier YNL186C in *S. cerevisiae* for this reaction (EC 3.7.1.2). Through our manual curation, we found that YNL186C matches to the Pfam FAA_hydrolase family (PF01557) with an e-value of 1.1e-49, and significantly matches with InterPro entry IPR002529 (Fumarylacetoacetase, EC 3.7.1.2). Hence, there is direct evidence to support this candidate gene, even though it is currently still unknown function in the SGD database.

Example 2: In *A. nidulans iHD666* network, the reaction EC 3.1.3.3 (Phosphoserine phosphatase) is a metabolic gap. MeGaFiller predicted AN10593 is a candidate gene for this reaction. Currently, in AspGD database, this gene is annotated as uncharacterized function. In fact, this candidate gene hits to HAD family in Pfam database with significant e-value of 2.6e-16. This family involves phosphoserine phosphatase activity. Besides, sequence similarity searching against Swiss-Prot database also gives supporting evidence for this candidate gene.

Example 3: In *A. niger iMA871* network, the reaction EC 4.2.3.5 (Chorismate synthase) is a metabolic gap. MeGaFiller found the gene ID 54235 as a candidate in *A. niger* genome. This candidate matches well with the chorismate synthase domain with e-value of 6.5e-130 in Pfam database.

Example 4: In *S. coelicolor iIB711* network, the reaction EC 4.1.1.36 (Phosphopantothenoylcysteine decarboxylase) is a metabolic gap. MeGaFiller predicted SCO1477 is a candidate gene. This candidate is annotated as putative flavoprotein homologue in Uniprot database. However, it matches well with DFP (with e-value of 1.6e-69) and Flavoprotein (with e-value of 8.6e-34) domains in Pfam database. Furthermore, KEGG database also confirms EC 4.1.1.36 activity for this candidate gene.

In addition, we further analysed which component gap fillers predicted the most gaps. As expected, within MeGaFiller, PFP-GF always produces the most number of candidate genes predictions. For examples, for the *S. coelicolor iIB711* dataset, PFP-GF predicted 63 out

of the 64 candidate genes from MeGaFiller, while B2G-GF predicted 60 candidate genes and EFICAz-GF predicted 9 candidate genes. Overall, within MeGaFiller, the general gap fillers (PFP-GF and B2G-GF) always contribute more predictions than the enzyme-specific gap filler (EFICAz-GF).

Filling Critical Gaps in the Metabolic Network of *A. oryzae*

A more detailed analysis of the metabolic network *iWV1314* for *A. oryzae* shows that there are 61 metabolic gaps (EC numbers) involved in 65 reactions that are spread over 37 metabolic pathways. To judge whether a gap is critical, we manually examined its reference pathway map given by KEGG database. A reaction in a pathway is called critical if it is the only reaction that consumes/produces a metabolite that is specific to the pathway. In other words, without that transformation, this pathway will not be connected, and hence the reaction is critical.

There are 28 critical gaps in the *A. oryzae iWV1314* network, and 33 non-critical gaps. Significantly, MeGaFiller predicted 12 candidate genes for 10 of these critical gaps. This means that MeGaFiller filled the gaps that are most likely to improve the connectivity of metabolic networks.

Example 5: Consider the Pantetheine-phosphate adenylyltransferase (PPAT) reaction (EC 2.7.7.3) which is currently a metabolic gap in the *A. oryzae iWV1314* network. This reaction is a critical gap in the Coenzyme A and pantothenate biosynthesis pathway (Figure 6) as it is the only transformation that produces Dephospho-CoA, which is the substrate to produce Coenzyme A. Without this reaction, the pathway is not functional and the Coenzyme A cannot be synthesized by this pathway.

This PPAT gap reaction by EC 2.7.7.3 was predicted by MeGaFiller

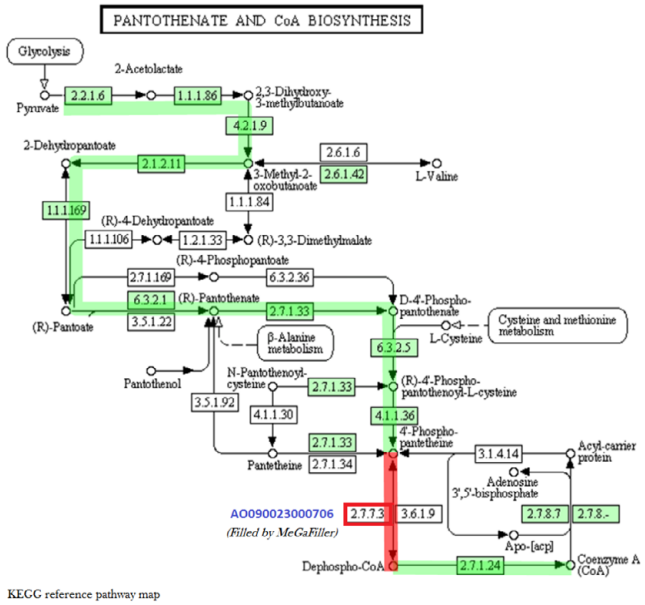

Figure 6: Filling gap for Pantothenate and CoA biosynthesis pathway in *A. oryzae*. The picture was modified from KEGG Pantothenate and CoA biosynthesis pathway (aor00770). Green-filled boxes are reactions with already identified genes in *A. oryzae*. White boxes are reactions without genes identified in *A. oryzae*. The EC 2.7.7.3 reaction (thick red-border box) is the "bottle-neck" for producing Dephospho-CoA, the substrate metabolite for CoA synthesis. MeGaFiller predicted AO090023000706 is the protein that catalyses for this reaction in *A. oryzae*.

(both PFP-GF and B2G-GF) to be catalysed by the candidate gene with identifier AO090023000706 in the *A. oryzae* genome. But, currently this gene shows un-annotated function in the DOGAN database.

However, we found strong supporting evidence for this prediction. Firstly, the protein matches with PPAT_CoAS (Phosphopantetheine adenylyltransferase domain) in CDD database with an e-value of 1.48e-36. It is also matches to Pfam's cytidylyltransferase domain (which is more general than PPAT) with an e-value of 6.1e-05. Another matching CDD entry is PRK01170, which is provisionally annotated as phosphopantetheine adenylyltransferase/unknown domain fusion protein. In addition, the corresponding ortholog in yeast (assigned by Ortho-MCL database) is the gene YGR277C, which is annotated as a PPAT by SGD database. With these strongly supporting evidences, we believed that AO090023000706 is the missing gene for the reaction with EC 2.7.7.3. With the assignment of the gene AO090023000706 to this reaction, the pathway becomes complete function. The full list of novel candidate genes for the metabolic gaps in *A. oryzae iWV1314* network filled by MeGaFiller can be found in additional file 1.

Comparison of MeGaFiller and GFAOP for gap filling

We compared MeGaFiller with an existing homology-based direct gap filling method, GFAOP [3]. A direct comparison with GFAOP was not possible since the first step of GFAOP with identifying the protein family given the EC number requires expert domain knowledge which is not easy to automate in software. Instead, we compared the predictions of MeGaFiller against the set of metabolic gaps in *A. oryzae* that were previously filled by the GFAOP method. While this comparison is not ideal, it is a reasonably close approximation. GFAOP used older datasets than MeGaFiller, but GFAOP have domain expert input while MeGaFiller does not.

For this comparison, we first extracted the set of metabolic gaps from the *A. oryzae iWV1314* metabolic network that were previously filled by GFAOP, together with the set of gene-EC number pairs predicted by GFAOP. This set represents the "difficult-to-fill" metabolic gaps that had remained in the network before applying GFAOP, but were then successfully filled by GFAOP. We called this the recently-filled gaps dataset WV-RFG and it contains 162 gene-EC number pairs.

We used MeGaFiller to fill these metabolic gaps in WV-RFG. MeGaFiller managed to predict 169 gene-EC number pairs. These predictions were compared with WV-RFG (the results obtained by GFAOP in [3]). MeGaFiller also predicted 102 (or 63%) of the 162 gene-EC number pairs filled by GFAOP. We note that this is a reasonably good performance by MeGaFiller since GFAOP uses domain expert input while MeGaFiller does not.

We then analysed the other 67 pairs predicted by MeGaFiller that were not in the WV-RFG. These predictions are either false predictions or additional gene-EC number pairs that were missed by GFAOP earlier. After our manual curation, we found that 38 (out of 67) pairs have strong supporting homology evidences over multiple annotation databases (e.g. CDD, Pfam, and UniProt databases), see additional file 2. Thus, it is likely that these 38 pairs predicted by MeGaFiller are actually additional gene-EC number pairs for the *A. oryzae* metabolic network, but they were missed by GFAOP. In the following, we give two illustrative examples.

Example 6: Consider the Endo-1,4-beta-xylanase reaction (EC 3.2.1.8) which is a metabolic gap in WV-RFG and is in the Polysaccharide metabolism. GFAOP predicted 6 gene-EC number pairs. For this reaction (EC 3.2.1.8), MeGaFiller gave a total of 8 gene-EC number

pairs (including the 6 pairs predicted by GFAOP). The two additional pairs involve candidate genes with identifier AO090026000103 and AO090103000141 in *A. oryzae*. But, both these genes are currently un-annotated in the *A. oryzae* genome.

Through our manual curation, we found that both these candidates AO090026000103 and AO090103000141 matches (with e-value 2.3e-44 and 2.5e-44, respectively) to Glycoside hydrolase (Glyco_hydro_11) domain in Pfam (PF00457) database. The family 11 of this domain comprises enzymes with only one known activity of xylanase (EC 3.2.1.8).

The previous method GFAOP missed both candidates. We found that there are two protein families (PF00457 and PF00331) that can perform this metabolic activity. We conjecture that GFAOP probably took only one family (PF00331) as its input and hence missed these predictions.

Example 7: Consider the *A. oryzae* gene with identifier AO090009000675, which is currently annotated in DOGAN database as a putative sugar kinase, and assigned in the *A. oryzae iWV1314* network as a NADH kinase (EC 2.7.1.86, in NAD and NADP Conversion pathway). MeGaFiller predicted (by all component gap fillers) this gene for another metabolic reaction NAD(+) kinase (EC 2.7.1.23). The Pfam homology confirms that it is a member of NAD(+) kinase family (PF01513) with an e-value (1.3e-47). This kinase family (PF01513) includes both EC 2.7.1.23 and EC 2.7.1.86 enzymatic functions. GFAOP found one of these, namely the EC 2.7.1.86 enzymatic function, while MeGaFiller found both of them. In this instance, MeGaFiller filled a gap and at the same time, found an additional metabolic reaction for an existing enzyme.

This example shows that MeGaFiller can also supplement protein function predictions, and in this case, it improves the network for the NAD and NADP Conversion co-factor pathway. The full list of 67 gene-EC number pairs predicted by MeGaFiller, together with their homology evidences can be found in additional file 2.

Explaining why GFAOP missed the remaining 61 metabolic gaps: We analyzed why the GFAOP missed the 61 metabolic gaps in the *iWV1314* network. To fill a gap (given by EC number), GFAOP firstly needs to find specific protein family for the given EC number. Examining the 61 EC numbers of the gaps in the *iWV1314* dataset, we found that only 9 EC numbers (of the 61) have corresponding Pfam protein family translation. However, GFAOP could not find any protein encoding in the genome of *A. oryzae* that matches with these 9 protein families. The remaining 52 EC numbers cannot map into specific Pfam protein family (as already explained in Background section). Some of these are too general, some are mapped to many Pfam families, and some are grouped under complicated Pfam sub-domain structures. Thus, there is no single specific protein family that can be used by GFAOP for those gaps. In contrast, MeGaFiller was able to fill 12 of these 52 gaps in this category, as explained earlier.

Comparison of MeGaFiller with ADOMETA

We also carried out a comparison of MeGaFiller with ADOMETA [38] which is a context-based method for gap filling. ADOMETA leverages gene association data [38-40] and can be used to predict new gaps as well as filling existing and predicting gaps. For comparison of MeGaFiller with ADOMETA, the published results of ADOMETA were used. The dataset used in ADOMETA was the metabolic network *iFF708* of yeast *S. cerevisiae* from a year 2003. This dataset has 513 genes, 386 EC numbers, and 541 pairs. It was reported that during self-testing for ADOMETA with this *iFF708* dataset and combined with gene association data, achieved 60% recall based on their top-50 predicted candidates [38]. It is noted that the precision of ADOMETA was not reported.

We ran MeGaFiller on the same iFF708 dataset, and achieved a recall of 87% with a precision of 77%. These are significantly better results, in both recall and precision. Of course, this is not a completely fair comparison-part of the improvement could be due to the more up-to-date reference information used by the component protein function predictors. However, we believe that homology evidence (where they exist) is stronger than association evidence in predicting these candidate genes. By relying on homology evidence to make its prediction, we believe that the candidate genes predicted by MeGaFiller are more reliable.

This result also suggests that one reason MeGaFiller worked well for less-characterized genomes is that the homology reference for them, in other existing genomes, was richly available and these could help MeGaFiller and other homology based methods like GFAOP to find the correct candidate genes.

Using MeGaFiller to make putative enhancement of metabolic networks

While MeGaFiller was designed primarily to fill metabolic gaps, we can also use it as a method to make putative enhancement to current metabolic networks. This enhancement is in the form of (a) putative candidate genes for existing reactions in the network, and (b) novel putative reactions for the current metabolic network. Here, we give some results.

Novel candidate genes: To do this for any target species, we ran MeGaFiller on the genome of the target species using the list of all EC numbers from the metabolic network of the species. We then filtered out the predictions that were already found in the networks. The remaining predictions contain novel candidate genes for existing reactions in the network.

We ran this for all the five networks and the results are shown in Table 6. For the *A. oryzae iWV1314* network, MeGaFiller predicted 587 novel candidate genes (for 215 EC numbers). The numbers of novel candidate genes for *S. cerevisiae iIN800*, *A. nidulans iHD666*, *A. niger iMA871* and *S. coelicolor iIB711* networks were 231, 384, 289 and 280, respectively (see Additional file 3). These predicted candidate genes need to be further curated and validated, but they give a valuable supplement of candidate genes for enhancement of these metabolic networks.

Novel putative metabolic reactions: To use MeGaFiller to predict novel metabolic reaction for current networks, we first retrieved all the EC numbers in the reference metabolic pathway (with identifier ec01100) from KEGG database. In all, there were 1,464 EC numbers relevant to metabolism. We filtered all EC numbers that were already found in the *A. oryzae* network (*iWV1314*). The remaining 753 EC numbers were used as input for MeGaFiller. Of these, MeGaFiller

Network dataset	*iIN800*	*WV1314*	*iHD666*	*MA871*	*iIB711*
Number of genes in current network	707	1346	674	831	711
Number of novel candidate genes	***231***	***587***	***384***	***289***	***280***

The number of novel candidate genes predicted by MeGaFiller for the five metabolic networks. These candidates need to be further curated, but they represent big potential enhancement in the gene coverage of these metabolic networks.

Table 6: Novel candidate genes predicted by MeGaFiller for the five metabolic networks.

predicted 369 candidate genes for 119 new EC numbers with corresponding novel putative metabolic reactions for *A. oryzae*. These novel putative reactions also need to be further curated and validated. Potentially they can further enhance the metabolic network.

Discussion

Metabolic gaps that exist after network reconstruction are usually "difficult-to-fill", since earlier gap filling methods have already failed to find them. While previous homology methods for gap filling that are based on protein family profile have been successfully used to enhance the reconstructed networks [3,19], they may fail if the protein family is poorly defined. Our approach based on retrofitting protein function prediction has indeed overcome the issue, since it does not require the concrete protein family. In this case, any individual protein in reference databases could help. The fact that MeGaFiller was able to fill 12 gaps (missed by previous method GFAOP for *A. oryzae* network), which have no specific protein family interpretation, has confirmed our idea. Furthermore, MeGaFiller is able to predict many additional candidate genes for existing reactions, as well as novel putative metabolic reactions throughout the metabolic network.

Conclusions

In this work, we demonstrated that retrofitting state-of-the-art protein function predictors can help to find candidates for "difficult-to-fill" local metabolic gaps that missed by previous direct gap filling methods. We implemented and tested an ensemble MeGaFiller method, which rationally combined three retrofitted gap fillers. We also performed gap filling and manual curation for *A. oryzae* network, and putative enhancement for the other four reconstructed metabolic networks.

Re-validation on filled gaps in *A. oryzae* network and manual inspection showed that our method was able to reliably propose more candidate genes that were missed by previous methods. There were strong supporting evidences found for these candidate genes in *A. oryzae* metabolic network, which suggests that our methodology is reliable. Thus, our method can serve as an effective bioinformatics tool for filling metabolic gaps and enhancing reconstructed metabolic networks.

We gave results on the use of MeGaFiller in tackling the related problem of detection and filling of gaps in metabolic network [14,15]. We also gave results on using MeGaFiller to predict novel candidate genes for existing reactions and novel putative reactions for the reconstructed metabolic network. This suggests that MeGaFiller may also be a powerful tool for investigating metabolism from metagenomics data, possibly with augmentation of context-based (association data) used in [38-42].

We next discuss some future work in this area: the first is to integrate more tools into MeGaFiller. Prediction power of MeGaFiller comes from its component tools, thus, if the retrofitted tools are capable of predicting globally missing genes, the integrated metabolic gap filler will have that capability as well. Recently, there are a number of annotation tools that make use of association data (network-based function annotation) such as GeneMANIA [43], and FS-weight [28]. These tools and data allow inferring functional associations between genes/proteins without sequence similarity reference. Thus, retrofitting and integrating gap fillers based on these tools will give MeGaFiller the capability of finding globally missing genes.

Recently, problem of finding globally missing metabolic gaps (orphan enzymes [17]) has been emerged [16,17]. Currently, there are 6,320 enzymatic reactions are known in Enzyme database (December 2013). However, only 4,534 enzymes (72%) have been annotated (as stored in Swiss-Prot database, December 2013). This means that, about 28% of the known metabolic activities remain orphan. Several attempts have been made to tackle this problem, using context association data [38-42], integration of genomics, interactomics [44,45], and metagenomics data [1,42]. Nonetheless, this missing metabolic knowledge [16,17,46] still remains as a challenging problem.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

NNN: implemented the MeGaFiller, analyzed the results, performed gap filling and candidate validation, and drafted the manuscript. WV: introduced the research problem, helped in gap filling and candidate validation, drafted and revised the manuscript, and designed the evaluation study. SB: helped in discussion and design of the study. PVN: helped with implementation of MeGaFiller, and helped in analyzing the results and drafting the manuscript. LHW: supervised the research, designed the study, helped in result analysis and extensively revised the manuscript.

All the authors read and approved the final manuscript.

Acknowledgements

Nam Ninh Nguyen and Hon Wai Leong were supported in part by the National University of Singapore under ARF grant R252-000-461-112. Wanwipa Vongsangnak was supported by National Natural Science Foundation of China (NSFC) (Grant No. 31200989 and No. K124810312) and Soochow University (a starting Grant No. Q410700111). The authors want to thank Dr Ket Fah Chong and Dr Sriganesh Srihari for helpful discussions, Professor Limsoon Wong for suggesting the use of function prediction methods. The publication fee was supported by National Natural Science Foundation of China (NSFC) (Grant No. 31200989).

Appendix A

Parameters for running the protein function predictors

PFP: no specific parameter was required. No threshold was applied. The PFP software and relevant data (August, 2008) was obtained from its authors [21].

Blast2GO: The web-service at (http://www.blast2go.com/b2glaunch) was run in January 2013 with reference database chosen as NR; other parameters were set by default (Blast with the best hits: top 20, E-value for annotation hit filter: 1e-6, Annotation cut-off: 55, GO weight: 5).

EFICAz: It was run with CHIEFc (Conservation-controlled HMM Iterative procedure for Enzyme Family classification) option. The data was retrieved for EFICAz (http://cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html) at January, 2011.

References

1. Seifert J, Herbst FA, Halkjaer Nielsen P, Planes FJ, Jehmlich N, et al. (2013) Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities. Proteomics 13: 2786-2804.

2. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, et al. (2008) The genome-scale metabolic model *iIN800* of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. BMC Syst Biol 2: 71.

3. Vongsangnak W, Olsen P, Hansen K, Krogsgaard S, Nielsen J (2008) Improved annotation through genome-scale metabolic modeling of *Aspergillus oryzae*. BMC Genomics 9: 245.

4. David H, Ozçelik IS, Hofmann G, Nielsen J (2008) Analysis of *Aspergillus nidulans* metabolism at the genome-scale. BMC Genomics 9: 163.

5. Andersen MR, Nielsen ML, Nielsen J (2008) Metabolic model integration of the

bibliome, genome, metabolome and reactome of *Aspergillus niger*. Mol Syst Biol 4: 178.

6. Borodina I, Krabben P, Nielsen J (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. Genome Res 15: 820-829.
7. Vongsangnak W, Ruenwai R, Tang X, Hu X, Zhang H, et al. (2013) Genome-scale analysis of the metabolic networks of oleaginous *Zygomycete fungi*. Gene 521: 180-190.
8. DeJongh M, Formsma K, Boillot P, Gould J, Rycenga M, et al. (2007) Toward the automated generation of genome-scale metabolic networks in the SEED. BMC Bioinformatics 8: 139.
9. Pitkänen E, Rantanen A, Rousu J, Ukkonen E (2008) A computational method for reconstructing gapless metabolic networks. Communications in Computer and Information Science 13: 288-302.
10. Notebaart RA, van Enckevort FH, Francke C, Siezen RJ, Teusink B (2006) Accelerating the reconstruction of genome-scale metabolic networks. BMC Bioinformatics 7: 296.
11. Pitkänen E, Rousu J, Ukkonen E (2010) Computational methods for metabolic reconstruction. Curr Opin Biotechnol 21: 70-77.
12. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Brief Bioinform 11: 40-79.
13. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, et al. (2013) The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. PLoS Comput Biol 9: e1002980.
14. Brooks JP, Burns WP, Fong SS, Gowen CM, Roberts SB (2012) Gap detection for genome-scale constraint-based models. Adv Bioinformatics.
15. Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics 8: 212.
16. Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol 7: 238-251.
17. Hanson AD, Pribat A, Waller JC, de Crécy-Lagard V (2009) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list-and how to find it. Biochem J 425: 1-11.
18. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5: 93-121.
19. Reed JL, Vo TD, Schilling CH, Palsson BO (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4: R54.
20. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics 5: 178.
21. Hawkins T, Luban S, Kihara D (2006) Enhanced automated function prediction using distantly related sequences and contextual association by PFP. Protein Sci 15: 1550-1556.
22. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676.
23. Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008: 619832.
24. Desai DK, Nandi S, Srivastava PK, Lynn AM (2011) ModEnzA: Accurate Identification of Metabolic Enzymes Using Function Specific Profile HMMs with Optimised Discrimination Threshold and Modified Emission Probabilities. Adv Bioinformatics.
25. Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 32: 6226-6239.
26. Arakaki AK, Huang Y, Skolnick J (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics 10: 107.
27. Dimmer EC, Huntley RP, Alam-Faruque Y, Sawford T, O'Donovan C, et al. (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Res 40: D565-570.
28. Chua HN, Sung WK, Wong L (2007) An efficient strategy for extensive integration of diverse biological data for protein function prediction. Bioinformatics 23: 3364-3373.
29. Cvijovic M, Olivares-Hernández R, Agren R, Dahr N, Vongsangnak W, et al. (2010) BioMet Toolbox: genome-wide analysis of metabolism. Nucleic Acids Res 38: W144-149.
30. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, et al. (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40: D700-705.
31. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, et al. (2005) Genome sequencing and analysis of *Aspergillus oryzae*. Nature 438: 1157-1161.
32. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, et al. (2014) The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res 42: D705-710.
33. Bentley SD, Chater KF, Cerdeño-Tárraga AM, Challis GL, Thomson NR, et al. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141-147.
34. UniProt Consortium1 (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71-75.
35. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-301.
36. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, et al. (2013) CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res 41: D348-352.
37. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40: D109-114.
38. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM (2006) Identifying metabolic enzymes with multiple types of association evidence. BMC Bioinformatics 7: 177.
39. Kharchenko P, Vitkup D, Church GM (2004) Filling gaps in a metabolic network using expression information. Bioinformatics 20: i178-185.
40. Chen L, Vitkup D (2006) Predicting genes for orphan metabolic activities using phylogenetic profiles. Genome Biol 7: R17.
41. Yamanishi Y, Mihara H, Osaki M, Muramatsu H, Esaki N, et al. (2007) Prediction of missing enzyme genes in a bacterial metabolic network. Reconstruction of the lysine-degradation pathway of *Pseudomonas aeruginosa*. FEBS J 274: 2262-2273.
42. Yamada T, Waller AS, Raes J, Zelezniak A, Perchat N, et al. (2012) Prediction and identification of sequences coding for orphan enzymes using genomic and metagenomic neighbours. Mol Syst Biol 8: 581.
43. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38: W214-220.
44. Smith AA, Belda E, Viari A, Medigue C, Vallenet D (2012) The CanOE strategy: integrating genomic and metabolic contexts across multiple prokaryote genomes to find candidate genes for orphan enzymes. PLoS Comput Biol 8: e1002540.
45. Janga SC, Díaz-Mejía JJ, Moreno-Hagelsieb G (2011) Network-based function prediction and interactomics: the case for metabolic enzymes. Metab Eng 13: 1-10.
46. Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng 107: 403-412.