Using genome-context data to identify specific types of functional associations in pathway/genome databases

Michelle L. Green and Peter D. Karp*

Bioinformatics Research Group, SRI International, Menlo Park, CA 94025, USA

ABSTRACT

Background: Hundreds of genes lacking homology to any protein of known function are sequenced every day. Genome-context methods have proved useful in providing clues about functional annotations for many proteins. However, genome-context methods detect many biological types of functional associations, and do not identify which type of functional association they have found.

Results: We have developed two new genome-context-based algorithms. Algorithm 1 extends our previous algorithm for identifying missing enzymes in predicted metabolic pathways (pathway holes) to use genome-context features. The new algorithm has significantly improved scope because it can now be applied to pathway reactions to which sequence similarity methods cannot be applied due to an absence of known sequences for enzymes catalyzing the reaction in other organisms. The new method identifies at least one known enzyme in the top ten hits for 58% of EcoCyc reactions that lack enzyme sequences in other organisms. Surprisingly, the addition of genome-context features does not improve the accuracy of the algorithm when sequences for the enzyme do exist in other organisms. Algorithm 2 uses genome-context methods to predict three distinct types of functional relationships between pairs of proteins: pairs that occur in the same protein complex, the same pathway, or the same operon. This algorithm performs with varying degrees of accuracy on each type of relationship, and performs best in predicting pathway and protein complex relationships.

Contact: pkarp@ai.sri.com

1 BACKGROUND

The crusade to elucidate functional information for the superfluity of sequences with no homology to known sequences continues. Each day, approximately 400 new proteins of unknown function are sequenced. Most methods for the reconstruction of metabolic networks depend on the validity and completeness of the annotation of the set of genes predicted in an organism. Sequences of unknown function continue to present a growing challenge to accurate reconstruction of a sequenced organism’s metabolic potential.

The BioCyc collection of Pathway/Genome databases (PGDBs) includes predicted metabolic networks for more than 250 organisms (Karp et al., 2005). In a predicted metabolic pathway, despite the presence of enzymes known to catalyze one or more reactions in the pathway, ‘pathway holes’ may occur. A pathway hole, or missing reaction, is a reaction for which no enzyme has been identified in the genome annotation. Table 1 shows the fraction of pathway holes in several BioCyc organisms, ranging from 30% to 59% of the total number of small-molecule metabolic reactions in the database (excluding the manually curated database for Escherichia coli, EcoCyc).

We previously developed the Pathway Hole Filler (PHFiller) to identify and evaluate candidate enzymes that fill ‘pathway holes’ in an organism’s metabolic network (Green and Karp, 2004). The PHFiller algorithm searches the genome for sequences homologous to a set of known enzymes from other organisms. The known enzymes are really functional ‘analogs’ given that we do not require sequence homology among them. Homology data and pathway context are then used to evaluate pathway hole fillers, that is, enzymes, putatively catalyzing the missing reaction.

The 900 metabolic pathways in the MetaCyc database are used to predict pathways in an organism-specific PGDB based on the organism’s genome annotation. Of the almost 2531 reactions in MetaCyc pathways, nearly one-third (741/2531) lack isozyme sequences. That is, no sequence exists in any organism for that enzyme activity. We refer to these as ‘orphan’ enzyme activities (Karp, 2004; Roberts, 2004). So, for these 741 reactions, PHFiller cannot be used to search an organism’s genome for the enzymes that catalyze them. Between 13% and 52% of the pathway holes in the metabolic networks of computationally predicted BioCyc PGDBs are orphan activities. By incorporating genome-context data into the features used by PHFiller, we may be able to identify candidate enzymes for these reactions.

In the past 5 years or so, genome-context methods have been used to hypothesize functionally related networks of proteins without reliance on homology to sequences of known function. Genome-context methods infer that two genes A and B are functionally related based on evidence from patterns conserved across many genomes. Two example genome-context methods are the conserved chromosomal proximity method (Dandekar et al., 1998; Overbeek et al., 1998; Pellegrini et al., 2001; Yanai et al., 2002), which infers that genes A and B are functionally related if orthologs of A and B across many organisms are nearby on the chromosome (also called gene neighbors), and the phylogenetic profile method (Gaasterland and Ragan, 1998; Pellegrini et al., 1999), which infers that genes A and B are functionally related if orthologs of A and B have similar patterns of presence and absence across many genomes. Most of these methods are capable of identifying networks of interacting proteins, but either do not incorporate multiple sources of data [5–16] or do not provide a
Table 1. Fraction of pathway holes across multiple PGDBs

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of Holes per reaction</th>
<th>PGDB tier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>0.03</td>
<td>1</td>
</tr>
<tr>
<td>Caulobacter crescentus</td>
<td>0.34</td>
<td>2</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>0.36</td>
<td>2</td>
</tr>
<tr>
<td>Archaeoglobus fulgidus</td>
<td>0.56</td>
<td>3</td>
</tr>
<tr>
<td>Thermosynechococcus elongatus</td>
<td>0.55</td>
<td>3</td>
</tr>
<tr>
<td>Bradyrhizobium japonicum</td>
<td>0.42</td>
<td>3</td>
</tr>
<tr>
<td>Salmonella enterica</td>
<td>0.30</td>
<td>3</td>
</tr>
<tr>
<td>Pyrococcus furiosus</td>
<td>0.59</td>
<td>3</td>
</tr>
</tbody>
</table>

The fraction of reactions that are pathway holes in a sample of BioCyc organism-specific PGDBs.

Several researchers have utilized multiple non-homology-based evidence types to identify functionally related enzymes. Only two have integrated these data in a rigorous manner to provide an overall assessment of the evidence (Kharchenko et al., 2006; von Mering et al., 2005), and only one has applied these results specifically to the problem of finding missing enzymes in a metabolic network. Kharchenko et al. (2006) have developed a method that uses multiple types of evidence, including three genome-context methods and protein interaction (for yeast) and gene expression data, to identify enzymes that fill holes in a metabolic network. The authors include all non-metabolic genes in E.coli and yeast as candidate enzymes to fill holes in those organisms’ networks. They use the rank of the true enzyme in the prioritized list of candidates to indicate the performance of the method and successfully identify 60% of known enzymes within the top ten candidates identified for the missing reaction in the E.coli metabolic network.

In this work, we have extended our PHFiller algorithm to use genome-context data to find missing enzymes to fill pathway holes. We achieved results slightly better than those of previous studies, with a more limited set of features used by our classifier. In addition, rather than considering the entire network as an equivalent set of reactions from which to select data for training and validation, our work examines separately the group of reactions for which functional analogs are available and the group of orphan reaction activities for which the PHFiller-GC method would be most beneficial.

In addition to using genome-context methods to identify and evaluate candidate enzymes to fill pathway holes in metabolic pathways, we also present a method for identifying putative relationships between proteins based on multiple types of distinct functional associations. These associations include:

1. proteins that appear in the same complex,
2. protein pairs whose genes appear in the same operon, and
3. protein pairs where protein A regulates transcription of the gene encoding protein B.

The Functional Association Finder (FAFinder) algorithm applies a Bayesian classifier to predict the probability that a pair of proteins is functionally related by a specific type of association. We validate the FAFinder method against literature-derived data from EcoCyc (Keseler et al., 2005) and computational data from a ‘Tier 2’ PGDB.

2 RESULTS AND DISCUSSION

2.1 PHFiller-GC: using genome-context data to fill pathway holes

2.1.1 Subset of known reactions used for training and validation
PHFiller-GC depends on the presence of at least one known enzyme in a pathway to identify and evaluate candidate enzymes to fill pathway holes—the existing enzyme(s) within the pathway provide the genome context that is used to search for other enzymes in the pathway. For validation studies, we required at least two reactions with identified enzymes in the pathway so that data from the enzyme catalyzing reaction one can be used to identify candidates for reaction two and vice versa, based on the rules described in the Implementation and Methods section.

EcoCyc 10.0 (Keseler et al., 2005) includes 206 metabolic pathways comprising 663 unique reactions. Of the 206 EcoCyc pathways, 169 are contiguous (i.e. at least two reactions act in series with the first reaction converting A to B and the second reaction converting B to C) and include at least one sequenced enzyme. Of these 169 pathways, 132 include two or more reactions with a sequenced enzyme. From these 132, we compiled 557 unique reactions, from which we remove 124 reactions catalyzed by enzymes catalyzing multiple reactions in the same pathway. Thus, our final dataset comprises 433 unique reactions, 507 unique enzymes (75 reactions are catalyzed by multiple enzymes or a complex), and a total of 547 enzyme-reaction pairs (33 enzymes catalyze multiple reactions). The same selection procedures were also applied to the CauloCyc PGDB for C.crescentus.

2.1.2 Validation using known reactions in EcoCyc
Table 2 summarizes each of the experiments completed using data from EcoCyc. The features available to PHFiller-GC included phylogenetic profiles (PP), conserved gene neighbors (GN), gene fusions (RS), gene clusters (GC) and gene-reaction adjacency (AD). For each combination of features investigated (homology and/or genome context), we report the fraction of true hits appearing in the top 10 candidates identified for each reaction. To evaluate the ability of our method to identify the correct enzyme catalyzing a known reaction, we performed 5-fold or 10-fold cross-validation on data taken from EcoCyc reactions as described in the Methods and Implementation section. Our method includes the identification of candidate enzymes for each pathway hole; if the true enzyme is not identified by our method, the miss counts against the performance of the method (see ‘Methods and Implementation’).

To evaluate the performance of the method without homology data, we first performed a cross-validation study using our entire EcoCyc reaction dataset, excluding homology
data from the identification and evaluation of candidate enzymes.

We compared the fraction of known enzymes identified in the top $N$ candidates for each of the individual features, and for the complete model combining all features. Figure 1A displays the fraction of known enzyme-reaction pairs identified in the top $N$ candidates, while Figure 1B displays the fraction of reactions for which one or more known enzymes are identified in the top $N$ candidates. In other words, Figure 1A describes how often the method identifies all known enzymes catalyzing a reaction, while Figure 1B describes how often the method identifies at least one of the known enzymes catalyzing a reaction.

Both ways of counting the fraction of known enzymes in the top $N$ candidates show similar trends in the decreasing order of performance of each of the individual genome-context methods. The conserved gene neighbors method provides the most effective discovery of known enzymes, followed by the gene cluster method, gene-reaction adjacency (the candidate is adjacent in the genome to the gene catalyzing an adjacent reaction), gene fusions and finally phylogenetic profiles. The full model (including all five nodes) provides the most accurate predictions. The performance of the gene neighbors method alone is comparable to the performance of the five methods combined.

For each validation experiment performed for EcoCyc and CauloCyc, we present the percent of known enzyme-reaction pairs in the top ten, the number of known enzymes in the top ten, and the number of known enzymes identified as the top-scoring candidate. The ‘% pairs in top 10’ is the number of reactions used in the validation. ‘Known homologs?’ is true if the known enzymes that catalyze the reactions used in that experiment have known homologs in other organisms (i.e. if the original PHFiller could be used to identify the reactions’ catalysts). For instance, the 111 EcoCyc reactions in the fourth row under ‘EcoCyc’ have no known homologous sequences available in other organisms; hence, candidates for filling these 111 pathway holes cannot be identified by homology.

<table>
<thead>
<tr>
<th>Features</th>
<th># rxns</th>
<th>Known homologs?</th>
<th>% pairs in top 10</th>
<th>% enzymes in top 10</th>
<th>% enzymes in top 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EcoCyc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homology</td>
<td>322</td>
<td>Yes</td>
<td>80.4</td>
<td>85.1</td>
<td>82.3</td>
</tr>
<tr>
<td>Homology + genome context</td>
<td>322</td>
<td>Yes</td>
<td>84.8</td>
<td>91.3</td>
<td>82.9</td>
</tr>
<tr>
<td>Genome context</td>
<td>322</td>
<td>Yes</td>
<td>42.6</td>
<td>49.3</td>
<td>25.3</td>
</tr>
<tr>
<td>Genome context</td>
<td>111</td>
<td>No</td>
<td>53.9</td>
<td>58.9</td>
<td>25.0</td>
</tr>
<tr>
<td><strong>CauloCyc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genome context</td>
<td>274</td>
<td>Yes</td>
<td>44.6</td>
<td>54.0</td>
<td>24.5</td>
</tr>
</tbody>
</table>

For each validation experiment performed for EcoCyc and CauloCyc, we present the percent of known enzyme-reaction pairs in the top ten, the number of known enzymes in the top ten, and the number of known enzymes identified as the top-scoring candidate. The ‘% rxns’ is the number of reactions used in the validation. ‘Known homologs?’ is true if the known enzymes that catalyze the reactions used in that experiment have known homologs in other organisms (i.e. if the original PHFiller could be used to identify the reactions’ catalysts). For instance, the 111 EcoCyc reactions in the fourth row under ‘EcoCyc’ have no known homologous sequences available in other organisms; hence, candidates for filling these 111 pathway holes cannot be identified by homology.

**Fig. 1.** Validation against 433 EcoCyc reactions. **A:** The fraction of known EcoCyc enzyme-reaction pairs (i.e. known enzymes) identified in the top $N$ candidates as found without homology data. **B:** The fraction of EcoCyc reactions for which one or more true enzymes are identified in the top $N$ candidates as found without homology data. The ‘AD’ data points include candidates identified not only by homology, but also gene-reaction coadjacency and genes occurring in the same direction as another pathway gene. Each of the remaining models utilizes, in addition, a single genome-context method as noted, except ‘full’ which includes gene-reaction co-adjacency (AD), gene neighbors (GN), gene clusters (GC), gene fusions (RS) and phylogenetic profiles (PP).
that describe the relationship between a pathway hole and candidate enzyme and the remaining reactions and known enzymes in the pathway. For example, gene-reaction adjacency is one pathway context feature. For 322 of the 433 EcoCyc reactions included in our dataset, homologs of isozyme sequences from other organisms were identified when we used BLAST to query the *E.coli* proteome. We used these 322 reactions (and 409 known enzymes catalyzing them) to determine if the addition of genome context to homology data improved our chances of identifying the known enzyme(s) catalyzing each reaction.

We found a slight advantage (91% versus 85.1%) in the addition of genome-context data in the search for pathway hole fillers for reactions where the true hit could be identified by homology. Figure 2 shows the fraction of true hits appearing in the top ten candidates for the EcoCyc reactions with homology data. The ‘homology’ data points include only candidates identified by homology and evaluated using the nodes included in the original PHFiller incarnation (Green and Karp, 2004) (i.e. average rank in BLAST output, best E-value, average % query aligned, number of query sequences matching hit, candidate in a potential pathway operon and gene-reaction adjacency).

The real benefit in adding genome-context data comes in the identification of pathway hole fillers for missing reactions that lack any known sequenced enzymes, that is, orphan reactions. PHFiller-GC identified one or more known enzymes in the top ten candidates for 65 of the 111 EcoCyc reactions for which no homology-based hits were available. While these reactions are catalyzed by known enzymes in *E.coli*, either no enzymes have been identified in other organisms or BLAST searches against the *E.coli* genome revealed no homologs to the known enzymes from. Searching with genome-context data alone can identify the true enzyme for 58.9% of the 111 EcoCyc reactions (Fig. 3).

Figure 3 also includes a measure of performance to compare our results to previous studies using different methods of identifying candidates for validation studies. Candidate lists for previous studies (Kharchenko et al., 2004; Kharchenko et al., 2006) were composed of all non-metabolic proteins in *E.coli* plus one known catalyst of the reaction. To compare our results to these studies, we removed all metabolic proteins from the list of candidates generated by our method, except the known enzyme, leaving a list of all non-metabolic candidates plus the known catalysts of the reaction. In Figure 3, the curve labeled ‘GN (non-metabolic)’ reflects the fraction of reactions for which our method identifies the known enzyme in the top ten candidates when all metabolic candidates other than the known enzyme are removed from our candidate list. With this additional restriction, PHFiller-GC identifies the known enzyme in the top ten candidates for 61.6% of reactions.

2.1.3 Application of PHFiller-GC to a Tier 2 PGDB Our results for the application of this method to EcoCyc are encouraging; however, most of the PGDBs being curated are what we call ‘Tier 2’ PGDBs. That is, the PGDB has been created using the PathoLogic program to infer pathways based on the organism’s genome annotation and the PGDB has then undergone additional review and annotation to add experimentally verified data (such as pathways) relevant to the organism. To evaluate the method on computationally generated databases with less manual curation than EcoCyc, we applied our method to the CauloCyc PGDB.

Like our analysis for EcoCyc, we computed for CauloCyc the fraction of reactions with at least one known enzyme (those enzymes assigned to reactions in the PGDB by PathoLogic based on the genome’s annotation) identified in the top ten candidates for the reaction. Figure 4 shows the number of reactions with hits identified at each rank for the full model and each of the individual features. The full model achieves the best performance, identifying a known enzyme in the top ten candidates for 54% of CauloCyc reactions. The gene-neighbors method performed nearly as well as the full model and the remaining individual genome-context methods fell far behind. In general, gene fusions and gene clusters are more efficient at
identifying pathway hole fillers than phylogenetic profiles and gene-reaction adjacency.

2.2 General functional association prediction compared to prediction of distinct types of functional association

A metabolic pathway can be thought of as a set of functional associations among the genes that code for enzymes within that pathway. Many researchers have used this definition of functional association to identify and validate novel genome-context methods, especially those genome-context methods that were used in our development of PHFiller-GC, but, participation in the same metabolic pathway is only one type of functional association that might exist between a pair of genes.

We investigated the ability to detect additional distinct functional associations among genes using genome-context data and gene coexpression profiles. These additional functional associations include:

1. proteins that appear in the same complex,
2. protein pairs whose genes appear in the same operon, and
3. protein pairs where protein A regulates transcription of the gene encoding protein B.

Our predictor integrates phylogenetic profiles (PP), conserved gene neighbors (GN), gene fusions (RS), gene clusters (GC), (excluded for identifying ‘same-operon’ pairs), the Spearmann rank correlation between coexpression profiles (CO) and gene-reaction adjacency (AD) to identify candidate pairs and applies a Bayesian classifier to evaluate the probability that two genes are functionally associated by one or more of the above criteria.

2.2.1 Overlap among types of functional associations

Many known pathway enzymes are protein complexes and many known complexes are encoded together in the same operon. If the categories of functional association described above overlap significantly, there would be no need to develop individual predictors for each. Thus, we examined the extent to which functionally associated pairs in each category overlap with those in the other categories.

Table 3 shows that of the 4055 unique pairs that appear in the same pathway in EcoCyc, only 239 of those pairs also appear in the same protein complex and 510 appear in the same operon. Similarly, of the 2858 pairs that appear in the same complex, only 729 appear in the same operon. Given the divergence in the sets of proteins related by each type of functional association, it would not be surprising if predicting each type required a different predictive model.

We explored this hypothesis using data from the EcoCyc database.

2.2.2 Validation results

The EcoCyc PGDB includes extensive data curated from the literature on protein complexes, transporters, known operons and the gene regulatory network of E.coli. These data can be used to easily identify known protein or gene pairs related by each of the relationships listed earlier. The predictor was evaluated and trained using protein pairs from EcoCyc known to be associated by one of these functional association criteria. We also trained and validated a predictor that combined all of the individual types of associations into a single predictor. We performed cross-validation studies in EcoCyc to determine the predictive value of the algorithm for identifying known functionally associated protein pairs and to determine which features (e.g. coexpression data and conserved gene neighbors) were most useful in identifying each type of functional association. We compared the area under the precision-recall curve (AUC-PR) to identify the model with the best predictive performance over the entire recall range. Table 4 shows the results from each type of functional association. For comparison, the AUC-PR for PHFiller is 0.90 and its precision at 50% recall is 94%. AUC-PR curves for each model comparison are available as Supplementary Data.

As expected, since gene clusters and conserved gene neighbors primarily depend on the colocation of genes in the genome, these methods perform best on identifying pairs that occur in the same operon. Also, since many of these methods were trained and validated using pairs of proteins participating in the same pathway, it is not surprising that they also perform...
Table 4. AUC-PR for functional association prediction methods

<table>
<thead>
<tr>
<th>Functional association</th>
<th>Best performer</th>
<th>AUC-PR</th>
<th>Precision @ 50% recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>AD + GC</td>
<td>0.52</td>
<td>54%</td>
</tr>
<tr>
<td>Same pathway</td>
<td>AD + GC</td>
<td>0.57</td>
<td>66%</td>
</tr>
<tr>
<td>Same complex</td>
<td>AD + GC or full</td>
<td>0.32</td>
<td>37%</td>
</tr>
<tr>
<td>Same operon</td>
<td>AD + GN + CO</td>
<td>0.68</td>
<td>80%</td>
</tr>
<tr>
<td>Regulator</td>
<td>RS</td>
<td>0.03</td>
<td>3.5%</td>
</tr>
</tbody>
</table>

The best-performing model for each prediction method shown in the table was identified by comparing the area under the AUC-PR for each model. The ‘All’ functional association includes pairs related by one or more of the specific relationships (same pathway, complex, operon or regulation pair). AUC-PR curves are available as Supplementary Data.

well in identifying same-pathway pairs and perform poorly in identifying regulatory relationships.

3 CONCLUSIONS

We have developed and validated a method for filling orphan pathway holes using a pathway context-guided process for candidate identification and extending the PHFiller program to use genome-context data. In some organisms, these pathway holes comprise a large fraction of the total number of missing reactions across the set of metabolic pathways inferred by Pathologic. Our evaluation has shown that adding genome-context data on top of homology data when the data are available provides a slight improvement in performance over homology data alone. But, when used to identify enzymes for reactions lacking isozyme sequences in other organisms, we can identify the correct enzyme in the top ten candidates 58% of the time, thus increasing the scope of the PHFiller program.

In applying our method to CauloCyc, we have established that the method is generally applicable to computationally predicted PGDBs. Unlike validation against EcoCyc, where the differences in performance of the individual genome-context methods were only minor, validation using data from CauloCyc revealed more drastic differences in the ability of each method to identify known enzymes. Known enzymes in CauloCyc are based solely on the annotation of the C. crescentus genome. Hence, the differences may reflect noise in the validation because of errors in the original annotation.

The overall performance in FAFinder’s prediction of each type of functional association is relatively low compared to the performance of PHFiller. However, like the PHFiller-GC method, our goal in applying these methods is to narrow the field of candidates for experimental investigation. In addition, these results indicate that although genome-context methods can be applied effectively to identify pathway-, complex- and operon-based functional associations, they seem to be entirely ineffective for identifying regulatory relationships.

4 IMPLEMENTATION AND METHODS

The PHFiller-GC program is an add-on to the Pathway Tools software and is implemented in ANSI Common Lisp. The genome-context data for each organism was stored in a MySQL database.

4.1 Datasets used

Our investigations used data from several organism-specific BioCyc databases. BioCyc is a collection of databases where each database describes one organism; for example, EcoCyc describes E.coli. EcoCyc is a manually curated DB, whereas the metabolic network in the CauloCyc PGDB for C. crescentus was predicted computationally based on the MetaCyc pathway DB. CauloCyc is a ‘Tier 2’ database. After creation, Tier 2 databases undergo limited manual curation for the addition of specific pathways, known and predicted complexes, and transporters from the literature for a given organism (Paley and Karp, 2002). For our validation and evaluation studies, we used version 10.0 of each organism’s PGDB.

4.2 Genome-context relationships and data source

We used data from Prolinks to identify gene pairs related by the gene neighbors, gene clusters, gene fusion, or phylogenetic profile methods (Green, 2006). Each method uses a different metric to compute functional relatedness as described by Bowers et al. (2004) and summarized in (Green and Karp, 2006).

Datasets for all organisms were retrieved from the Prolinks download page (http://mysql5.mbi.ucla.edu/public/). After download, gene identifiers were translated to identifiers used in each PGDB using a local instance of BioWarehouse (Pouliot et al., 2005).

Co-expression profiles for the functional association predictor were downloaded from GEO (Barrett et al., 2005) and the Stanford Microarray Database (Ball et al., 2005).

4.3 Identification of candidate proteins for PHFiller-GC

The original PHFiller program uses isozyme sequences that catalyze the pathway hole reaction in other organisms to search the target genome for candidate proteins. As shown previously, an average of about 20% of the pathway holes in a PGDB have no known isozyme sequences, that is, they are orphan enzymes. PHFiller-GC uses genome-context methods to identify and evaluate candidate proteins from a PGDB to fill these pathway holes. Rather than querying all proteins in the PGDB or constructing a list of non-metabolic proteins plus the enzyme known to catalyze each reaction as earlier researchers have done, we build a list of candidate proteins using several criteria, as follows:

(1) Include all proteins that are in the same direction with another gene catalyzing a reaction in the pathway.

(2) Include all proteins that are functionally associated (by any one of the genome-context methods used) with any protein catalyzing any other reaction in the pathway.

(3) Exclude all proteins that catalyze another reaction in the pathway.

Since most genome-context methods were developed and validated based on protein pairs that appear together in the same pathway, these other known pathway enzymes will disproportionately increase the number of false positive candidates identified for each reaction.

For method validation, other groups have used constructed candidate lists including, for instance, all non-metabolic proteins plus the true enzyme. Rather than specifically adding the true enzyme(s) to this list and then assessing if our method can identify it, we rely on the first two criteria to identify all candidates. In other words, if the true
enzyme is not included by one of our first two criteria, our method will not be able to identify it.

4.4 Evaluation of candidate proteins for PHFiller-GC
We applied the same method used by the original PHFiller program to compute the probability that each candidate catalyzes the desired reaction (Green and Karp, 2004). Briefly, for each evidence node, we computed probability distributions for the set of true hits, P(evidence|true hit), and the set of false hits, P(evidence|false hit). Given these training distributions, we then applied a native Bayesian classifier to compute the probability that the enzyme is a true hit given the evidence, \( P(\text{true hit}|\text{evidence}) \).

4.5 Measures of performance for evaluation of PHFiller-GC
We considered two different measures in our evaluation of the method. We determined the fraction of true hits found in the top \( N \) candidates identified for each reaction when all candidates are ranked by the computed probability. Since many BioCyc reactions are catalyzed by multiple enzymes, either heteromultimeric complexes or multiple isozymes, we assessed accuracy by counting either the fraction of all true hits found, or the fraction of reactions for which a true hit was found (in the top ten candidates). For example, imagine that reaction \( R \) is catalyzed by enzymes A, B and C, and these three enzymes are the three top-scoring hits for reaction \( R \) in the order \( A > B > C \). If we count all hits for the reaction, our performance appears worse than in reality; only one-third of this reaction’s enzymes appear in the first position in the list of candidates, two-thirds appear in the top two and all appear in the top three. However, by counting only best hits, our performance appears to be enhanced; the top-scoring candidate for reaction \( R \), enzyme A, is one of the enzymes known to catalyze the reaction; enzymes B and C are not considered.

ACKNOWLEDGEMENTS
We would like to thank the referees for their careful review of the manuscript and thoughtful suggestions for improvement. This work was supported in part by grant No. DE-FG03-01ER63219 from the Department of Energy. This financial support does not constitute an endorsement of the views expressed herein.

Conflict of Interest: none declared.

REFERENCES