Combining multiple functional annotation tools increases completeness of metabolic annotation

Marc Griesemer[1,§], Jeffrey Kimbrel[1], Carol Zhou[2], Ali Navid[1], Patrik D'haeseleer[1,2]

[1] Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

[2] Global Security Computing Applications Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

[§] Corresponding author (griesemer1@llnl.gov)

Abstract

The dirty little secret behind genome-wide systems biology modeling efforts is that they are invariably based on very incomplete functional annotations. Annotated genomes typically contain 30-50% of genes with little or no functional annotation [1], severely limiting our knowledge of the "parts lists" that the organisms have at their disposal. In metabolic modeling, these incomplete annotations are often sufficient to derive a reasonably complete model of the core metabolism at least, typically consisting of well-studied (and thus well-annotated) metabolic pathways that are sufficient for growth in pure culture. However secondary metabolic pathways or pathways that are important for growth on unusual metabolites exchanged in complex microbial communities are often much less well understood, resulting in missing or lower confidence functional annotations in newly sequenced genomes. For example, one third of the EC database consists of "orphan enzymes" that have been described in the literature but for which no sequence data is available [1].

Individual metabolic annotation tools often return annotations for different subsets of genes, offering the potential to greatly increase the completeness of metabolic annotations by combining the outputs of multiple tools. Indeed, recent genome-scale modeling of Clostridium beijerinckii NCIMB 8052 demonstrated that the total number of genes and reactions included in the final curated model could be almost doubled by incorporating multiple annotation tools [2].

Here, we present preliminary results on a comprehensive reannotation of 27 bacterial Tier 1 and Tier 2 reference genomes from BioCyc[3], focusing on enzymes with EC numbers annotated by KEGG[4], RAST[5], EFICAz[6], and the Brenda enzyme database [7].

Methods

Reference Genomes

We utilized a total of 27 prokaryotic genomes of Metacyc Tier 1 & Tier 2 organisms. These genomes are from a range of phyla, including 15 Proteobacteria, 6 Firmicutes, 3 Actinobacteria, 2 Bacteroidetes and 1 Cyanobacteria. In addition to a range of phyla, these 27 organisms also display a range of lifestyles, including a human gut symbiont (*Bacteroides thetaiotaomicron* VPI-5482) and pathogens (e.g. *Mycobacterium tuberculosis*), an obligate insect endosymbiont (*Candidatus* Evansia muelleri), a bacterium with interesting metabolism (*Aurantimonas manganoxydans* SI85-9A1), and an important marine primary producer (*Synechococcus elongatus* PCC 7942). These organisms present various types of annotation challenges, such as identifying incomplete pathways in obligate intracellular organisms with small genomes, as well as the challenges of annotating incomplete, “draft” contigs compared to “finished” genomes.

Table 1. Reference genomes used in this study

Genome Name	Phylum	NCBI Accessions	Proteins
Mycobacterium tuberculosis CDC1551	Actinobacteria	AE000516	4189
Mycobacterium tuberculosis H37Rv	Actinobacteria	AL123456	4018
Streptomyces coelicolor A3(2)	Actinobacteria	NC_003888, NC_003903, NC_003904	8152
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes	AE015928, AY171301	4825
Candidatus Cardinium hertigii[b]	Bacteroidetes	HG422566, CBQZ010000001 - CBQZ010000011	739
Synechococcus elongatus PCC 7942	Cyanobacteria	CP000100, CP000101	2661
Listeria monocytogenes 10403S	Firmicutes	CP002002	2814
Bacillus subtilis 168	Firmicutes	AL009126	4185
Clostridium saccharoperbutylacetonicum ATCC 27021	Firmicutes	CP004121, CP004122	5821
Eubacterium rectale ATCC 33656	Firmicutes	CP001107	3626
Peptoclostridium difficile 630	Firmicutes	AM180355, AM180356	3809
Agrobacterium fabrum C58	Proteobacteria	AE008687, AE008688, AE008689, AE008690	5402
Aurantimonas manganoxydans SI85-9A1	Proteobacteria	AAPJ01000001-AAPJ01000035	3650
Caulobacter crescentus CB15	Proteobacteria	AE005673	3737
Caulobacter crescentus NA1000	Proteobacteria	CP001340	3885
Escherichia coli CFT073	Proteobacteria	AE014075	5379
Escherichia coli K-12 substr. W3110	Proteobacteria	NC_007779	4410

Escherichia coli B str. REL606	Proteobacteria	CP000819	4209
Escherichia coli K-12 substr. MG1655[a]	Proteobacteria	U00096	4140
Escherichia coli O157:H7 str. EDL933	Proteobacteria	AE005174, AF074613	5449
Candidatus Evansia muelleri[b]	Proteobacteria	LM655252	330
Helicobacter pylori 26695	Proteobacteria	CP003904	1594
Methylosinus trichosporium OB3b	Proteobacteria	NZ_ADVE02000001 - NZ_ADVE02000003	4344
Candidatus Portiera aleyrodidarum BT-QVLC[b]	Proteobacteria	CP003867	280
Shigella flexneri 2a str. 2457T	Proteobacteria	AE014073	4068
Vibrio cholerae O1 biovar El Tor str. N16961	Proteobacteria	AE003852, AE003853	3828

a: Tier 1 Pathway Genome Database (EcoCyc)

b: Endosymbiont with reduced genome

Annotation Tools

RAST (Rapid Annotation Subsystem Technology, [5]) is an open-source web server for genome annotation, using a "Highest Confidence First" assignment propagation strategy based on manually curated subsystems and subsystem-based protein families that automatically guarantees a high degree of assignment consistency. RAST returns an analysis of the genes and subsystems in each genome, as supported by comparative and other forms of evidence. We used the NMPDR website (rast.nmpdr.org) to generate genome-wide annotations for our 27 reference genomes.

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of genome and pathway databases for systems biology. We used the KAAS (KEGG Automatic Annotated Server, www.genome.jp/tools/kaas, [4]) to generate genome-wide annotations for our 27 reference genomes. KAAS assigns KEGG Orthology (KO) numbers using the bi-directional best hit method (BBH) against a set of default prokaryotic genomes in the KEGG database.

EFICAz [6] uses large scale inference to classify enzymes into functional families, combining 4 methods into a single approach without the need for structural information. Recognition of functionally discriminating residues (FDR) allows EFICAz to use a method called evolutionary footprinting.The latest EFICAz has high precision and recall ability: under test using sequence similarity of >40%, precision and recall were 0.88 and 0.85, while at sequence similarity of >60%, they were near unity. We used a local install of EFICAz2.5 to generate EC number predictions for our 27 reference genomes.

Brenda [7] is a publicly available enzyme database (containing 82,568 enzymes and 7.2 million enzyme sequences as of 2017) based on the literature and contains functional and molecular information such as nomenclature, enzyme structure, and reactions and specificity. We

annotated our 27 reference genomes based on a BLAST search at >60% sequence identity against a local copy of the Brenda database of enzyme reference sequences.

Results

In total, the four annotation tools produces 47,447 unique Gene-EC annotations ("gene X codes for an enzyme with EC number Y") across the 27 reference genomes, or an average of 1757 annotations per genome. The Venn diagram in Figure 1 illustrates the degree of overlap - and non-overlap - between the sets of annotations produced by the four tools.

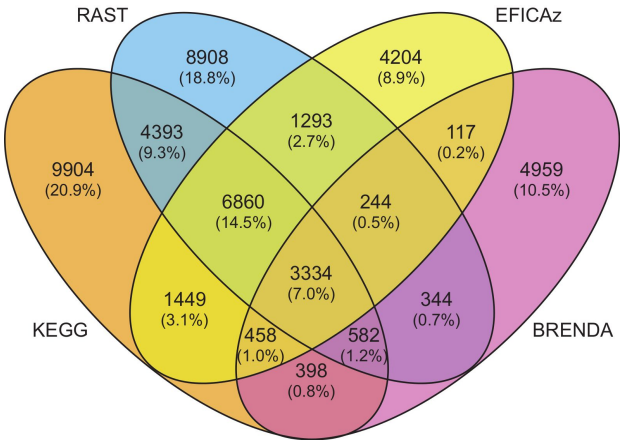

Figure 1. Large differences exist between the sets of Gene-EC annotations generated by the four annotation tools across the 27 reference genomes.

It is clear that the metabolic annotations produced by these automated genome-wide annotation tools can differ drastically. Each tool produced on average between 23% (EFICAz) and 48% (Brenda) unique gene-EC annotations that were not predicted by any of the other tools. Overall, fewer than a quarter of all gene-EC annotations are agreed on by at least 3 tools.

The EC numbers on which the different tools most often agree across the 27 reference genomes tend to belong to well studied core metabolic pathways, such as glycolysis, amino acid and nucleotide biosynthesis, etc.

The EcoCyc database is an extensively hand-curated and continuously updated database summarizing all the experimentally determined enzymatic functions in *Escherichia coli* K-12 substr. MG1655, the single best studied model organism in the history of modern biology. We can use the EC numbers annotated in EcoCyc as a gold standard to evaluate how well the automated annotation tools are able to annotate the enzymes in E. coli K-12. Figure 2 shows how the set of gene-EC annotations generated by KEGG and RAST compare against EcoCyc.

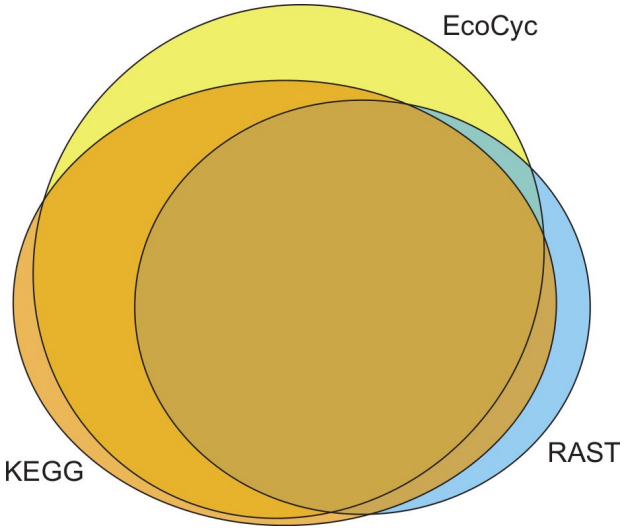

Figure 2. Gene-EC annotations produced by KEGG and RAST for *E. coli* K12, compared to the EcoCyc gold standard. The sets and intersections are drawn proportionally to the number of annotations in each.

Using EcoCyc as the Gold Standard for *E. coli* K12, the different tools achieve a Precision of 78% (BRENDA) to 92% (KEGG), and a Recall of 33% (BRENDA) to 85% (KEGG). Note that even on *E. coli* K12, which we expect to be a best-case situation, there are significant differences in the annotations produced by the different tools, and each tools is only able to cover a subset of the known enzymes in EcoCyc.

The annotation tools show much more agreement on E. coli than on more remote lineages such as Actinomycetes, Bacteroidetes, or Clostridia. For *E. Coli* K12, 60% of EC numbers are agreed on by 3 or more tools, while 28% EC numbers come from only a single tool. In contrast, for P. difficile 630, only 33% of EC numbers are agreed on by 3 or more tools, and 48% of EC numbers come from only a single tool.

The disjoint sets of annotations produced by the different tools provide us with an opportunity to trade off confidence in the annotations versus coverage. If higher confidence is required, we can focus solely on the subset of annotations that is agreed upon by multiple tools. Conversely, if the lack of genome coverage or metabolic network coverage is considered a problem, we can use the union of multiple tools to achieve a wider annotation.

Figure 3A shows the number of unique EC numbers annotated by each of the tools across the 27 reference genomes, which reflects the size of the metabolic network reconstruction (average 868 EC numbers per genome). Figure 3B shows the number of genes in all of the reference genomes annotated with one or more EC numbers by each of the tools, which reflects the overall genome annotation coverage (average 1361 genes per genome, or 34% of the genes).

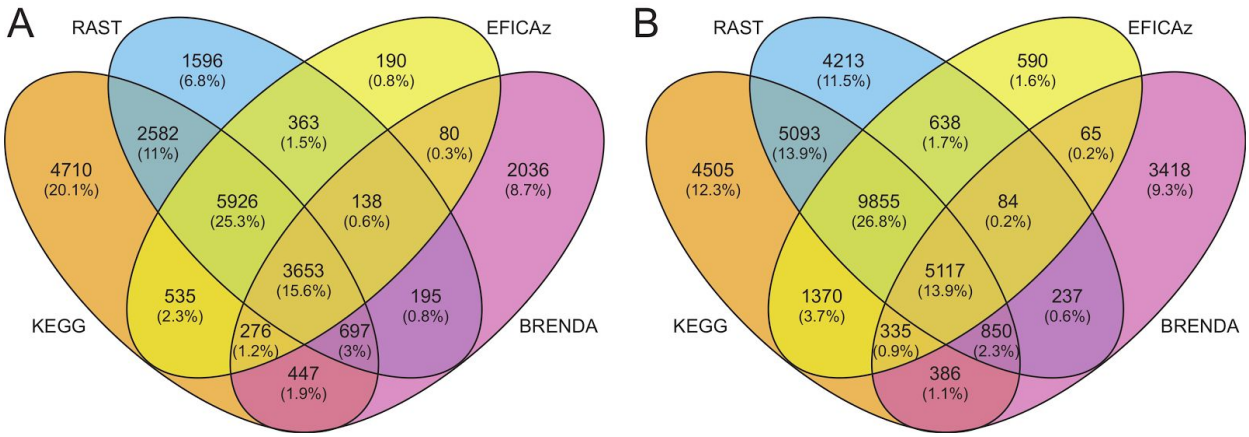

Figure 3. Combining Annotation Tools Improves Genome annotation and Metabolic Network Coverage. A: Total unique EC numbers annotated. B: Total genes annotated with EC numbers

Note that each tool adds a significant number of reactions to the metabolic network model, and each tool significantly contributes to the number of genes covered with metabolic annotations.

Discussion

RAST and KEGG are the most widely used tools for metabolic network reconstruction. However, they do not necessarily produce identical annotations. In our analysis, KEGG produced the best Precision and Recall on E. coli K12. In general, KEGG produces a larger number of unique EC numbers, which could indicate more over-prediction, or more comprehensive pathway coverage. Note that both also generate many reactions without official EC numbers.

EFICAz produces the least number of unique EC numbers, but can be used in combination with RAST or KEGG to highlight high confidence annotations. EFICAz also produced 3-digit EC number annotations which may be used for hole filling.

BLAST against the Brenda database of reference enzymes produced the smallest number of annotations, but a high fraction of unique EC numbers. Of the top 10 unique EC numbers produced by this method, only one is also covered by RAST and KEGG, two of the EC numbers have been deprecated, and six are EC numbers that have been assigned in 2000 or newer, and may not have been incorporated into the predictions by the other annotation tools yet.

Recommendations

1. Do not just use a single annotation tool unless you are only interested in core metabolism.
2. Trade off confidence versus coverage by looking at the intersection or union of multiple annotation tools.

3. EFICAz can be used to identify higher confidence annotations, or partial EC numbers for hole filling
4. BLASTing against a database of reference sequences such as the Brenda database is generally an inefficient method for annotating enzymes, but may be useful to cover more recently assigned EC numbers not yet included by other tools.
5. More tool development is needed to allow merging of annotations beyond simple EC numbers.

Acknowledgements

This work was supported by the Department of Energy through the Genome Sciences Program as part of the LLNL Biofuels SFA (contract SCW1060). Work at LLNL was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

References

1. Hanson AD, Pribat A, Waller JC, de Crécy-Lagard V. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. Biochem J. 2009 Dec 14;425(1):1-11.
2. Milne CB, Eddy JA, Raju R, Ardekani S, Kim PJ, Senger RS, Jin YS, Blaschek HP, Price ND. Metabolic network reconstruction and genome-scale model of butanol-producing strain Clostridium beijerinckii NCIMB 8052. BMC Syst Biol. 2011 Aug 16;5:130.
3. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016 Jan 4;44(D1):D471-80.
4. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W182-5.
5. Overbeek R, Olson R, Pusch G D, Olsen GJ, Davis JJ, DiszT, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014 Jan;42(Database issue):D206-14.
6. Kumar N, Skolnick J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. Bioinformatics. 2012 Oct 15;28(20):2687-8.
7. Schomburg I, Chang A, Placzek S, Söhngen C, Rother M, Lang M, Munaretto C, Ulas S, Stelzer M, Grote A, Scheer M, Schomburg D. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 2013 Jan;41(Database issue):D764-72.