Nucleic Acids Research, 2005, Vol. 33, No. 13 ***4035–4039***
doi:10.1093/nar/gki711

Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers

M. L. Green* and P. D. Karp

Bioinformatics Research Group, Artificial Intelligence Center, SRI International, Menlo Park, CA 94025, USA

Received April 11, 2005; Revised and Accepted June 29, 2005

ABSTRACT

We report on a new type of systematic annotation error in genome and pathway databases that results from the misinterpretation of partial Enzyme Commission (EC) numbers such as '1.1.1.-'. This error results in the assignment of genes annotated with a partial EC number to many or all biochemical reactions that are annotated with the same partial EC number. That inference is faulty because of the ambiguous nature of partial EC numbers. We have observed this type of error in multiple databases, including KEGG, VIMSS and IMG, all of which assign genes to KEGG pathways. The *Escherichia coli* subset of the KEGG database exhibits this error for 6.8% of its gene-reaction assignments. For example, KEGG contains 17 reactions that are annotated with EC 1.1.1.-. A group of three *E.coli* genes, b1580 [putative dehydrogenase, NAD(P)-binding, starvation-sensing protein], b3787 (UDP-*N*-acetyl-D-mannosaminuronic acid dehydrogenase) and b0207 (2,5-diketo-D-gluconate reductase B), is assigned to 15 of those reactions, despite experimental evidence indicating different single functions for two of the three genes. Furthermore, the databases (DBs) are internally inconsistent in that the description of gene functions for genes with partial EC numbers is inconsistent with the activities implied by reactions to which the genes were assigned. We infer that these inconsistencies result from the processing used to match gene products to reactions within KEGG's metabolic pathways. These errors affect scientists who use these DBs as online encyclopedias and they affect bioinformaticists who use these DBs to train and validate newly developed algorithms.

INTRODUCTION

The Enzyme Commission (EC) system (http://www.chem.qmul.ac.uk/iubmb/enzyme/) is perhaps the earliest, and one of the most widely used, examples of a hierarchical controlled vocabulary in biology. Enzymologists recognized early on that great confusion can result from the use of uncontrolled terminology in a complex technical domain. The use of multiple synonyms for one enzyme, or of the same name for different enzymes, can result in chaos in the biomedical literature. These problems are even more severe in biological databases (DBs) and computational applications, because computers lack the biochemical knowledge with which scientists can sometimes disambiguate confusing terminology. However, misuse of a controlled vocabulary in a bioinformatics DB is worse than having no controlled vocabulary at all, because users of the DB will assume that they can reliably compute with terms of the vocabulary.

This article demonstrates a systematic set of errors within several pathway DBs involving what we call partially qualified EC numbers (partial EC numbers), such as '2.1.1.-'. These DBs include many cases where a gene annotated with a partial EC number, which, by definition, does not denote a specific reaction, is inaccurately assigned to a set of reactions that are all annotated by the same partial EC number. These errors can have serious repercussions for users of the DBs. For example, since DBs such as KEGG are used for training and validating algorithms in bioinformatics research (1–8), algorithms trained on DBs containing these errors will have learned their rules and parameters from faulty training data.

We infer in the case of KEGG that these errors result from the assumption within the processing underlying KEGG that, if two genes are assigned to the same KEGG Orthology (KO) group (http://www.genome.jp/kegg/ko.html), their products therefore catalyze all reactions that are assigned the partial EC number associated with the KO group. When the KO entry involves a diverse group of reactions with the same partial EC number, this approach can lead to the erroneous assignment of multiple genes to a set of reactions inconsistent with their existing annotations.

We have observed examples of the same type of error within the VIMSS (http://vimss.lbl.gov/) and IMG (http://img.jgi.doe.gov/v1.0/main.cgi) DBs.

Although some readers may consider it inappropriate for us to publish these errors, rather than to simply report them to the

*To whom correspondence should be addressed. Tel: +1 650 859 5669; Fax: +1 650 859 3735; Email: green@ai.sri.com

DB authors, the fact that the same errors have been made by three different groups suggests that partial EC numbers are not well understood in the bioinformatics community. Therefore, we consider it important to raise awareness around this issue among both DB providers (current and future) and DB users.

Overview of the EC system

The EC system is a hierarchical controlled vocabulary that assigns unique combinations of four numbers to different enzyme activities. For example, the EC term 2.1.1.1 corresponds to the enzyme activity nicotinamide *N*-methyltransferase, in which nicotinamide and *S*-adenosyl-L-methionine are converted to *S*-adenosyl-homocysteine and 1-methylnicotinamide. All enzymes with that activity, regardless of the source organism, are assigned that EC number to indicate their catalytic function. A multifunctional enzyme is assigned multiple EC numbers corresponding to each of the reactions that it catalyzes. For example, the enzyme that is the product of the *Escherichia coli* trpC gene catalyzes two reactions whose EC numbers are 5.3.1.24 and 4.1.1.48.

The EC system is a hierarchical classification of reactions according to several criteria that include the nature of the chemical transformation they accomplish, and the chemical classes of their substrates. For example, the term 2.1.1.1 is in class 2 (transferase reactions, which are of the form $XY + Z = X + YZ$, that is, the Y group has been transferred from X to Z), and is in subclass 2.1 (transferase reactions that transfer 1-carbon groups), and is in sub-subclass 2.1.1 (transferase reactions in which the 1-carbon group is a methyl group).

The last '1' in 2.1.1.1 is simply a sequence number with no meaning as part of the classification system. That is, reaction 2.1.1.1 is the first reaction in class 2.1.1 that was assigned an EC number by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB).

What is a partial EC number and why are they used? Partial EC numbers look like EC numbers except the last number is replaced by a dash, e.g. 2.1.1.-. Partial EC numbers are used with two different intended meanings.

Meaning 1. Consider a scientist who is performing sequence analysis of a newly discovered gene, who finds that the gene shows equally close sequence similarity to several different methyltransferases. The scientist might choose to annotate the gene with the function 'methyltransferase (EC 2.1.1.-)' as shorthand for saying 'I feel confident inferring that the enzyme has a methyltransferase activity, but I do not feel confident inferring exactly which methyltransferase activity'.

Meaning 2. Consider an experimentalist from Stanford who has just characterized a novel methyltransferase that catalyzes the reaction A + B = C + D, has sequenced the gene, and has deposited the sequence in a public sequence DB. During the deposition process the experimentalist is prompted to enter an EC number. However, EC numbers can be assigned only by the NC-IUBMB and no other group is authorized to officially assign EC numbers. Because of their rigorous review process, it takes several months before the NC-IUBMB can assign an official EC number to a newly discovered enzyme. The experimentalist chooses not to wait, and when prompted for the EC number enters 2.1.1.- as a way of saying 'I know this enzyme is a methyltransferase, and even though I know the exact reaction catalyzed by the enzyme, because I do not yet know the sequence number it will be assigned within 2.1.1, I will omit the last digit in the EC number and enter 2.1.1.-'. Alternatively, the experimentalist might leave the EC number field completely blank, and the number 2.1.1.- might be assigned by the DB curation staff.

Consider an experimentalist at UC Berkeley who has just characterized another novel methyltransferase, and one that catalyzes a different methyltransferase reaction than that discovered by the Stanford team, say E + F = G + H. When the Berkeley team deposits the sequence of the enzyme, the team might also assign it the partial EC number 2.1.1.- because, again, the team is simply stating that the enzyme is in the class of methyltransferases, because the sequence number is unknown.

We now see the key issue. The different enzymes that catalyze different reactions within the same class can be assigned the same partial EC number. But the fact that they are assigned the same partial EC number does not mean that they have the same activities.

METHODS

To demonstrate the systematic errors resulting from the assignment of genes to multiple reactions with the same partial EC number, we investigated several cases of partial EC numbers from the KEGG, VIMSS and IMG DBs. We investigated examples drawn from *E.coli*, *Homo sapiens* and *Caulobacter crescentus*, but VIMSS and IMG are microbial DBs only. Note that our examples are drawn from websites that are actively updated and, therefore, may have changed since the time of this analysis.

We retrieved data from KEGG via the web interface to the GENES DB and from version 0.4 of the KGML organism-specific metabolic pathway datasets. At the time of our analyses, the online versions of the KEGG *E.coli* and *C.crescentus* Gene catalogs were last updated on January 19, 2005 and the most recent update to the *H.sapiens* Gene catalog was January 25, 2005. Using the KEGG text search interface to the *E.coli*, the *C.crescentus* and the *H.sapiens* 'Gene catalogs' all accessible via the KEGG2 web page (http://www.genome.jp/kegg/kegg2.html) we searched for specific partial EC numbers to find all KEGG maps including a reaction assigned that EC number. For example, querying KEGG with 2.7.2.- reveals that three *E.coli* KEGG pathways include a reaction with this partial EC number.

We searched for cases where reactions with partial EC numbers appeared in multiple KEGG pathway maps, with the same set of genes catalyzing multiple different reactions. That is, cases where some set of genes A are all assigned a partial EC number X, and genes in the set A appear in multiple pathways catalyzing those reactions. When we say 'gene G is assigned to a reaction R whose partial EC number is X', we mean that the KEGG DB contains an assertion that the product of G catalyzes reaction R, and that a user viewing the pathway map can access gene G and all other genes assigned to reaction R by clicking the corresponding reaction symbol (its EC number), and that the KGML representation of the pathway map

includes gene G. Our definition is consistent with the available descriptions of KEGG's annotation procedure. During the internal reannotation of the GENES entry for a gene, a K number is assigned by KEGG. That K number is the identifier for a KO entry that enables matching of multiple genes to all reactions associated with that KO entry in the KEGG metabolic pathways (5).

Our survey also considered the KEGG gene name and gene product description, which according to the KEGG authors remain unchanged from the original genome annotation during the KEGG reannotation process (5).

Our evaluation of gene-reaction assignments is based on two criteria. When possible, we utilize the experimental literature as determined by the EcoCyc and UniProt annotations of each gene as our gold standard data. In cases where such information is not available, we judge gene-reaction assignments based on the number of reactions assigned to the gene. Specifically, if a gene is assigned to more than five reactions, we reviewed the list of reactions to determine the likelihood that the assignments were correct based on the gene annotation and the diversity of the reaction substrates.

To retrieve data from the BioCyc DBs for our examples, we employed Lisp queries against version 8.6 of the BioCyc DBs. BioCyc is a collection of DBs where each DB describes one organism, for example, EcoCyc describes *E.coli*. EcoCyc (9) is a manually curated DB, whereas the metabolic networks in the other BioCyc DBs used in this study were predicted computationally (10). After the initial prediction, these databases were subjected to very limited curation to add known species-specific pathways and other information.

Our analyses of the VIMSS (http://vimss.lbl.gov) (website version as on March 25, 2005) and IMG (http://img.jgi.doe.gov/v1.0/main.cgi) (website version as on March 25, 2005) DBs were less systematic. We checked whether some of the same multiple-assignment errors present in KEGG were also present in these DBs. In each case we checked, they were. Note that VIMSS and IMG both use KEGG pathway maps in their sites, but the procedures by which they assign genes to reactions in these maps are unclear (we are unable to find publications about these DBs, and the website documentation does not specify these procedures in detail). However, we infer that VIMSS and IMG both perform their own gene-reaction assignments because in the examples we checked, VIMSS and IMG often inferred gene-reaction assignments for these partial EC numbers for additional genes beyond those present in KEGG.

RESULTS

We considered examples where a set of genes, all assigned the same partial EC number, are assigned to multiple distinct reactions across several KEGG metabolic maps. We present one example for *E.coli*. Please refer to the Supplementary Materials for additional examples from *E.coli*, *H.sapiens* and *C.crescentus*.

E.coli EC 4.2.1.-

EC 4.2.1.- appears in 10 KEGG pathway maps. In eight of these maps, two genes b0036 [Swiss-Prot entry P31551, Carnitinyl-CoA dehydratase (EC 4.2.1.-)] (example URL: http://www.genome.jp/dbget-bin/www_bget?eco:b0036) and b1517 [Swiss-Prot entry P76143, Putative aldolase yneB (EC 4.2.1.-); hypothetical 31.9 kDa protein in hipB–uxaB intergenic region] are both assigned to catalyze a set of 16 distinct reactions as shown in the Supplementary Materials. According to EcoCyc, b0036 and b1517 encode a carnitine racemase (EC 5.-.-.-) and a putative aldolase, respectively. The function of b0036 is supported by experimental evidence (11). VIMSS also assigns the same two genes to multiple different reactions in nine KEGG maps (example URL: http://www.microbesonline.org/cgi-bin/fetchEC2.cgi?ec=4.2.1.-&taxId=83333). IMG assigns the same two genes to multiple different reactions in nine KEGG maps (example URL: http://img.jgi.doe.gov/v1.0/main.cgi?page=keggMapTaxonGenes&taxon_oid=65&map_id=map00120&ec_number=EC:4.2.1.-&gene_oid=6000860).

Systematic analysis of the *E.coli* subset of KEGG

The above example represents one instance where the assignment of a set of genes to the same partial EC number results in erroneous conclusions about the functions of those genes. This is one instance of a systematic problem. The KEGG *E.coli* dataset assigns 869 genes to reactions with full or partial EC numbers. Table 1 describes the number of genes with partial EC numbers, and the number of genes determined to be correctly or incorrectly assigned to a set of reactions in KEGG based on our criteria as outlined in Methods. Out of the 869 genes, 135 genes (16%) are assigned to reactions with partial EC numbers. Of the 135 genes with partial EC number reactions, 58 genes are incorrectly assigned to a set of reactions that is either inconsistent with their functional annotation or, in spite of the gene's non-specific annotation (e.g. putative acetyltransferase), it is unlikely that the gene catalyzes each of the disparate reactions to which it has been assigned. Thirty-nine genes with partial EC numbers appear to be correctly assigned to a set of reactions that are consistent with their functional annotation. We found no significant difference in the average number of 4-digit EC numbers between the correct and the incorrect classes (i.e. the incorrect genes were not in 'more difficult' EC sub-subclasses than the correct genes). Fourteen genes with partial EC numbers seem to be unfinished in some way. For these 14 genes, each gene, G, appears in the list of genes for a specific pathway, say, pathway P. The GENES page for G is lacking an entry in its pathway field. Further, the gene is also absent from the KGML entry for pathway P. We were unable to discern the cause of these missing genes.

Treatment of partial EC numbers in the three DBs often leads to inconsistencies. Consider *E.coli* gene b3787 [Swiss-Prot entry P27829, UDP-*N*-acetyl-D-mannosamine

Table 1. Summary of systematic analysis of the KEGG *E.coli* dataset

Group of genes	Number of genes	Percentage of partial ECs
All *E.coli* genes in KEGG	4411	
With EC numbers (full or partial)	869	
With partial EC numbers	135	
Correct (consistent with functional annotation)	38	28.1
Incorrect	59	43.7
Unfinished/missing	14	10.4
No associated reaction	21	15.6
Unable to determine correctness	3	2.2

dehydrogenase (EC 1.1.1.-)]. The function listed for the product of this gene on the KEGG gene page at URL http://www.genome.jp/dbget-bin/www_bget?eco:b3787 is UDP-*N*-acetyl-D-mannosaminuronic acid dehydrogenase. However, KEGG assigns this gene to 15 different reactions, none of which match this enzymatic activity, meaning that different parts of the KEGG DB make incongruent assertions about the function of this gene. These same inconsistencies are present in VIMSS and IMG.

DISCUSSION

As noted above, not all genes with a partial EC number have been incorrectly assigned. In fact, there appear to be two different 'classes' of KO entries associated with partial EC reactions in KEGG. Reactions with the same partial EC number may be part of different KO entries. In the cases we examined, the first class of KO groups appears to describe a specific function and is associated with only one or a few specific reactions. For these class 1 KO entries, the descriptions of the associated genes typically agree with the activity of the reactions. For class 2, each KO entry includes multiple disparate reactions, the descriptions of the genes in a KO grouping vary, and may be inconsistent with the functions implied by the reactions in the entry. As an example of class 1, the definition of K03473 is 'erythronate-4-phosphate dehydrogenase'. The entry includes one reaction in the 'Vitamin B6 metabolism' pathway, and one *E.coli* gene is associated with this reaction/KO, b2320 [Swiss-Prot entry P05459, erythronate-4-phosphate dehydrogenase (EC 1.1.1.-)]. Entry K00100 is an example of class 2. It has no definition field, includes 17 reactions and is associated with three genes in *E.coli* as discussed above. We were unable to locate any explanation for the differences in these KO entries.

The exact processing strategy used by KEGG to assign genes to metabolic reactions is not clearly defined in any publication we have examined, so it is difficult to be certain of the source of the errors we have presented. Our current hypothesis is that these errors are caused by a partially computational reannotation pipeline in which genes are mapped onto orthology groups, which in turn are associated with EC numbers within KEGG. We hypothesize that the KEGG software computes that when two genes belong to the same orthology group, the genes catalyze all reactions associated with that orthology group. That processing is probably correct for full EC numbers, but it is erroneous for partial EC numbers.

These errors in KEGG, VIMSS and IMG will clearly impact scientists who use these DBs as encyclopedias to manually look up information about the relationships among genes, reactions and metabolic pathways.

These errors affect another class of users when KEGG data are used as a gold standard for training and validation of computational methods (1–8). For example, when developing genome context methods for predicting functional associations between genes, metabolic pathways are often used as one definition of functional association. Both KEGG and EcoCyc define sets of genes that are functionally related based on their involvement in the same metabolic pathway. The erroneous gene-reaction assignments in KEGG will result in large numbers of false-positive functional associations in KEGG. For example, KEGG shows b0207 [Swiss-Prot entry P30863, 2,5-diketo-D-gluconic acid reductase B (EC 1.1.1.274)], b1580 [Swiss-Prot entry P38105, starvation-sensing protein rspB (EC 1.1.1.-)] and b3787 [Swiss-Prot entry P27829, UDP-*N*-acetyl-D-mannosamine dehydrogenase (EC 1.1.1.-)] as 1.1.1.- in several pathways. The nucleotide sugars metabolism pathway (eco00520) in the *E.coli* dataset has 21 genes assigned to 15 reactions in the pathway, including the three genes for EC 1.1.1.-. The genes in this pathway specify 210 unique pair-wise functional associations (i.e. 210 unique pairs of genes appear in this pathway). If one of the three genes is incorrectly assigned to the pathway, 20 of these pair-wise associations are incorrect. Since the time of our analysis, b0207 has been removed from the K00100 entry. Hence, all the methods trained here include at least these 20 incorrect associations. If a second gene, for example b3787, is incorrectly assigned to 1.1.1.- a total of 39 of the 210 pair-wise associations are wrong.

Recommendations for the proper handling of partial EC numbers in metabolic pathway databases

The BioCyc DB collection consists of 160 organism-specific pathway/genome DBs (PGDBs), most of which include metabolic pathway predictions generated by SRI's Pathway Tools software (10). Because Pathway Tools assigns genes to the reactions that their products catalyze by matching only fully qualified EC numbers provided in the genome annotation, BioCyc PGDBs do not contain the type of error with which this article is concerned. Because Pathway Tools also assigns genes to reactions by matching based on enzyme names, using a comprehensive dictionary of enzyme names within the MetaCyc DB, pathway tools is able to correctly make assignments for reactions that lack fully qualified EC numbers. For example, the *E.coli* gene b0207 was originally annotated as 2,5-diketo-D-gluconate reductase B (EC 1.1.1.-); if we match the name of the enzyme to the reaction in MetaCyc, the gene is associated with only one of the 31 reactions for EC 1.1.1.- in MetaCyc (six EcoCyc reactions have that partial EC).

Recommendations for an explicit specification of partial EC numbers

To mitigate the future effects of the semantic ambiguity in partial EC numbers, we propose a change in the specification of these numbers. As discussed above, a partial EC number may have one of the two meanings, the first being 'I don't know the exact activity of this enzyme and therefore, cannot specify the fourth number', and the second being 'I know the exact activity of this enzyme, but the NC-IUBMB has not yet assigned a sequence number'. In the first case, we propose that these instances should be indicated with a '?' in the fourth position, e.g. EC 2.3.4.?, meaning 'unknown', while instances of the second case should be indicated with an 'n' in the fourth position, e.g. EC 2.3.4.n, meaning 'not available yet'.

Summary

Our analysis has identified a new class of misannotation within genome and pathway DBs that is due to misinterpretation of partial EC numbers. We illustrated these errors through

examples from the KEGG, VIMSS and IMG DBs, which systematically misassign genes to reactions with partial EC numbers. Each gene that is assigned a K number for a KO entry associated with a generalized group of reactions is assigned to each of those reactions (http://www.genome.jp/kegg/ko.html), which is an incorrect inference. Furthermore, different parts of KEGG, VIMSS and IMG contain inconsistent information about the functions of these genes; one part of KEGG contains the gene function from the original genome annotation; another part contains the new erroneous multiple reaction assignments.

We assert that bioinformaticists should exercise more caution in utilizing these DBs as training datasets because these gene-reaction assignments generate many incorrect data points in those training data.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

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 numbers RR07861 and GM70065 from the National Institutes of Health. This financial support does not constitute an endorsement of the views expressed herein. Funding to pay the Open Access publication charges for this article was provided by NIH RR07861 and GM70065.

Conflict of interest statement. None declared.

REFERENCES

1. Tringe,S.G., von Mering,C., Kobayashi,A., Salamov,A.A., Chen,K., Chang,H.W., Podar,M., Short,J.M., Mathur,E.J., Detter,J.C. *et al*. (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
2. von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A. and Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.
3. Wu,J., Kasif,S. and DeLisi,C. (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**, 1524–1530.
4. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*., **18**, 176–179.
5. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
6. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res*., **32**, W336–W339.
7. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*., **31**, 258–261.
8. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
9. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res*., **33**, D334–D337.
10. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18** (Suppl. 1), S225–S232.
11. Eichler,K., Buchet,A., Lemke,R., Kleber,H.P. and Mandrand-Berthelot,M.A. (1996) Identification and characterization of the caiF gene encoding a potential transcriptional activator of carnitine metabolism in *Escherichia coli*. *J. Bacteriol*., **178**, 1248–1257.