Atom Identifiers Generated by a Graph Coloring Method Enable Compound Harmonization Across Metabolic Databases

Huan Jin 1, Joshua M. Mitchell 2,3,4 and Hunter N.B. Moseley 2,3,4,5, *

1 Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, KY 40536, USA; huan.jin@uky.edu
2 Department of Molecular & Cellular Biochemistry, University of Kentucky, Lexington, KY 40536, USA
3 Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA
4 Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky, Lexington, KY 40536, USA
5 Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA

* Correspondence: hunter.moseley@uky.edu; Tel.: 859-218-2964

Received: date; Accepted: date; Published: date

Abstract: Metabolic flux analysis requires both a reliable metabolic model and metabolic profiles in characterizing metabolic reprogramming. Advances in analytical methodologies enable production of high-quality metabolomics datasets capturing isotopic flux. However, useful metabolic models can be difficult to derive due to the lack of relatively complete atom-resolved metabolic networks for a variety of organisms, including human. Here, we developed a graph coloring method that creates unique identifiers for each atom in a compound facilitating construction of an atom-resolved metabolic network. What is more, this method is guaranteed to generate the same identifier for symmetric atoms, enabling automatic identification of possible additional mappings caused by molecular symmetry. Furthermore, a compound coloring identifier derived from the corresponding atom coloring identifiers can be used for compound harmonization across various metabolic network databases, which is an essential first step in network integration. With the compound coloring identifiers, 8865 correspondences between KEGG and MetaCyc compounds are detected, with 5451 of them confirmed by other identifiers provided by the two databases. In addition, we found that the Enzyme Commission numbers (EC) of reactions can be used to validate possible correspondence pairs, with 1848 unconfirmed pairs validated by commonality in reaction ECs. Moreover, we were able to detect various issues and errors with compound representation in KEGG and MetaCyc databases by compound coloring identifiers, demonstrating the usefulness of this methodology for database curation.

Keywords: Metabolomics; atom-resolved metabolic network; atom identifier; compound identifier; database harmonization; graph theory; common subgraph isomorphism

1. Introduction

Metabolic flux analysis is an essential approach to access metabolic phenotypes[1-2] that requires both reliable metabolic profiles as well as metabolic models[3-5]. Advances in analytical technologies like mass spectrometry (MS) and nuclear magnetic resonance (NMR) greatly contribute to the detection of thousands of metabolites from biofluids, cells and tissues[6]. Application of those analytical techniques to stable isotope resolved metabolomics (SIRM) experiments facilitates production of high-quality metabolomics datasets capturing isotopic flux through cellular and systemic metabolism[7-8]. Now, the challenge is to construct meaningful metabolic models from the corresponding metabolic profiles for downstream metabolic flux
analysis. A metabolic network is usually represented by compounds connected via biotransformation routes[9]. Obviously, information at the atom level is not represented in such metabolic networks, making it impractical to derive appropriate metabolic models for SIRM datasets. However, currently there is no relatively complete atom-resolved databases of metabolic networks available for human metabolism that can be used to trace individual atom[10].

To construct an atom-resolved metabolic network, compounds and metabolic reactions with detailed documentation at the atom level are required. The molfile description of a compound contains information about atoms, bonds, connectivity, and coordinates[11], which is available in most databases. For atom-resolved metabolic reactions, the KEGG RPAIR database stores patterns of transformations occurring between two reactants in a single reaction[12]. In addition, MetaCyc contains direct atom mappings for every metabolic reaction[13]. Previous work only made use of atom mappings in either the KEGG RPAIR database[14-15] or MetaCyc[16] for atom tracing. The numbers of compounds and atom-resolved reactions in KEGG and MetaCyc databases are summarized in Table 1. MetaCyc has 8.7% more compound entries and 72% more atom-resolved reaction entries than KEGG. Both databases cover metabolism for many common organisms, clearly indicating that these two databases are not independent of each other. Therefore, the necessary first step for constructing a more comprehensive network is to integrate compounds from different databases without redundancy[17].

Table 1. KEGG and MetaCyc databases summary.

| Data types               | KEGG    | MetaCyc |
|--------------------------|---------|---------|
| Compounds                | 18636   | 20264   |
| Reactions                | 11427   | 17203   |
| Atom-resolved reactions  | 10008   | 17203   |

To initially evaluate the level of overlap between KEGG and MetaCyc databases, we used existing identifiers in each database to find the correspondences between KEGG and MetaCyc compounds. Some compounds in MetaCyc have a direct identifier to the corresponding KEGG compound[17]. The results are shown in Table 2. We can see that the number of matched compounds detected by different identifiers are not consistent, with the total less than 5700. Two issues may appear when applying these identifiers to compound integration across various databases. On the one hand, there is no easy way to check if some correspondences are missing. Besides, it is difficult to tell if the results generated by those identifiers are correct, since errors can exist in every database[18-19].

Table 2. Correspondences between KEGG and Metacyc compounds.

| Identifiers | KEGG    | MetaCyc | Correspondences |
|------------|---------|---------|----------------|
| InChI      | 13216   | 15076   | 2336           |
| ChEBI      | 15353   | 8404    | 3106           |
| KEGG       | 18636   | 5402    | 5402           |
| Either-ID  | 18636   | 15216   | 5681           |

In an atom-resolved metabolic network, each node should include information at both molecule-specific and atom-specific levels. To name each atom in a compound, two rules need to be obeyed: 1) different atoms must have different identifiers; 2) and symmetric atoms must share the same identifier. Previous work used atom index in the molfile associated with a compound in finding atom-specific metabolic pathways without considering symmetry[15-16]. One group tried to assign unique name for every atom in the compound based on the compound’s InChI representation[19], which does not apply to this scenario since symmetric atoms can share the same routes in the metabolic network. No appropriate method has been previously published that provides each atom in a compound a useful identifier for constructing an atom-resolved metabolic network.
In this paper, we developed a novel graph coloring method that creates a unique identifier for each atom in a compound. Moreover, the method is guaranteed to generate the same coloring identifier for symmetric atoms. Furthermore, compound coloring identifiers derived from the corresponding atom coloring identifiers can be used for compound harmonization across metabolic databases. To our knowledge, this is the first attempt to create atom identifiers with the consideration of symmetry and harmonize compounds from KEGG and MetaCyc databases for the construction of a relatively complete atom-resolved metabolic network.

2. Results

2.1. Aromatic substructure detection.

Our graph coloring method described in Section 4.6 is very sensitive to the specific structural representation. Aromatic substructures are not consistently represented in both databases. Instead of being directly labeled as an aromatic bond type, single and double bonds are used alternatively to depict the aromatic substructure. CPD-6962 in MetaCyc has a direct KEGG referencing compound C15523 (Figure 1). We can see that the positions of double bonds and single bonds within the benzene ring vary between these two representations, which can lead to two different sets of atom identifiers. Therefore, we needed to ensure that compound representation is consistent across databases so that each compound will have a single set of atom identifiers. In this case, we first detect the aromatic substructures in all compounds from both databases, and change the single and double bonds within the aromatic substructure to aromatic bond.

![Figure 1. Correspondence between KEGG and MetaCyc compound with different representations.](image)

Two independent aromatic detection methods were used in aromatic substructure detection: our BASS [20] substructure detection method and the aromatic detection facilities in the Indigo package [21]. First, we compared the aromatic substructures derived by these two methods (Table 3). Overall, Indigo appears more conservative than BASS in detecting aromatic substructures, detecting roughly 85% of what the BASS method does. Figure 2 shows an example aromatic substructure that can be missed by Indigo. We assume that Indigo cannot detect aromatic substructures with a double bond connected to atoms outside of the ring. This is not surprising, since BASS leverages the curated set of aromatic substructures in KEGG and has very high precision (99.9%) in the detection of aromatic substructures in KEGG compounds, while Indigo uses a set of simplified aromatic detection heuristics along with hard-coded algorithmic limitations of ring sizes being searched. However, we had concerns that some valid aromatic substructure representations in MetaCyc compounds may not exist in the referencing aromatic substructure set derived from the KEGG database, which would be missed by the BASS method. This was confirmed by Indigo detecting 30 additional MetaCyc compounds with aromatic substructures not detected by the BASS method. Therefore, we supplemented the KEGG aromatic substructures with additional Indigo-detected substructures in MetaCyc. Using both methods, about half of the compounds in either database contain detected aromatic substructures (Table 4).

| Databases | BASS | Indigo |
|-----------|------|--------|
| KEGG      | 0    | ~1500  |

Table 3. Incomplete detection of aromatic substructures by BASS and Indigo.
2.2. Generating identifiers for atoms using a graph coloring method.

Since symmetric atoms share the same neighbors, the graph coloring method is guaranteed to create the same identifier for them. Our concern is whether atoms with the same identifier are actually symmetric. In our graph coloring method, we only include 0-layer identifiers (see Section 4.6 Graph coloring method) in atom coloring to avoid long name strings. In some extreme cases, this shortcut can assign the same identifier to atoms that are asymmetric. An example is shown in Figure 3. We can see that this compound does not contain any symmetric atoms. Without considering the upper right ring, the bottom two rings are symmetric. Therefore, atom 1 and 2 have the same 0-layer identifier, which is the same for atom pairs 4 & 5 and 6 & 7. In addition, once atom 1 and 2 reach atom 3, they will share the same route to the upper right substructure. Finally, atom 1 and 2 will share the same coloring identifier even though they are not symmetric. To deal with this problem, atom coloring validation and recoloring is performed.

We validate symmetry after a first round of coloring, recolor the compound if asymmetric atoms have the same identifier, and verify symmetry again. After this coloring-validation-recoloring-validation process, our results indicated that the graph coloring method is able to generate the same identifier for symmetric atoms and asymmetric atoms have unique identifiers for all compounds in both KEGG and MetaCyc databases.

2.3. Detection of correspondences between KEGG and MetaCyc compounds via coloring identifiers.
After creating a single set of atom identifiers for each compound, we were able to derive ordered compound coloring identifiers at different levels of chemical specificity, which can be used to harmonize compounds across databases. Since KEGG and MetaCyc can include different numbers of H (hydrogen atoms) in the molfile, we exclude H in coloring at this point. We first tried to include information of bond stereochemistry, atom charge, atom stereochemistry, and isotope stereochemistry in coloring to ensure each compound has a unique name. With the relatively specific coloring identifiers, 1762 correspondences between KEGG and MetaCyc compounds can be detected (see Table 5), which is not satisfactory compared to 5681 pairs discovered by other identifiers. This lack of correspondence is due to the inconsistencies in bond stereochemistry, atom charge, atom stereochemistry, and isotope stereochemistry information between these two databases. An example shown in Figure 4, where compound CPD-20570 in MetaCyc has a direct KEGG referencing compound C13014.

![Figure 4. Example of charge inconsistency of compound representations between databases.](image)

For the following analysis, we only included information of atom type and bond type to keep the backbone of a compound in atom naming. About 8865 correspondences between KEGG and MetaCyc are discovered (see Table 5), and 5451 of them can be confirmed by other identifiers. The question then becomes how to validate the remaining 3414 possible pairs. Matched compounds are supposed to take part in the same metabolic reactions. The Enzyme Commission (EC) number is a numerical classification scheme for enzymes, playing a key role in classifying enzymatic reactions[22]. We expected matched compounds to take part in metabolic reactions with similar EC numbers.

| Identifiers                  | Color matched pairs | ID verified pairs |
|------------------------------|---------------------|------------------|
| Tight coloring identifier    | 1763                | 1448             |
| Loose coloring identifier    | 8865                | 5451             |

Table 5. Matched compounds detected by the compound coloring identifiers

Then, we analyzed the metabolic reactions in KEGG and MetaCyc databases (see Table 6). We can see that the documentation of EC number in KEGG is more complete compared to MetaCyc, but the number of metabolic reactions in MetaCyc is 50% larger than in KEGG. Around 80% of reactions in both databases can be related to at least a 3-leveled EC number.

| EC types         | KEGG (count / percentage) | MetaCyc (count / percentage) |
|------------------|---------------------------|-------------------------------|
| No EC            | 1263 / 11.05%             | 3427 / 19.92%                 |
| 4-leveled EC     | 8933 / 78.17%             | 10740 / 62.43%               |
| 3-leveled EC     | 1081 / 9.46%              | 2958 / 17.19%                |

Table 6. Analysis of EC involved in reactions in KEGG and MetaCyc

Next, we tested how well EC numbers work in the validation of correspondences between KEGG and MetaCyc compounds (See Table 7). We first sifted out color-harmonized pairs that both take part in some reactions in their respective database. There are 4227 ID confirmed pairs and 2292 possible pairs involved in the metabolic reactions. We further investigated if those pairs participate into the same type of reaction indicated by EC number. If we used the first 3 levels of the sectioned EC number
as the standard, 3810 (90.13%) ID-confirmed pairs are verified by 3-leveled EC numbers. In addition, 3850 of them can be further confirmed by 4-leveled EC numbers. Furthermore, 1848 and 1540 possible pairs are confirmed by 3-leveled and 4-leveled EC numbers, respectively. These results suggest that EC numbers may be useful in validating possible pairs that have slight coloring deviations.

Table 7. Correspondences between KEGG and MetaCyc compounds verified by reactions.

| Conditions                  | ID-confirmed pairs | Possible pairs |
|-----------------------------|--------------------|---------------|
| Pairs not in reaction       | 1224               | 1122          |
| Pairs in reactions          | 4227               | 2292          |
| Verified by 3-leveled EC    | 3810               | 1848          |
| Verified by 4-leveled EC    | 3580               | 1540          |

2.4. Compound representation errors and issues detected in the KEGG and MetaCyc databases.

When harmonizing compounds between KEGG and MetaCyc databases, we found that there are various compound representation issues and errors existing in both databases, which can be grouped into several categories like mismatch between compound image and molfile, incorrect cross-referencing, and different bonds attached to metal ions. Here, we give a brief description with some examples, and all the detected inconsistency is documented in Supplementary Spreadsheet S1.

2.4.1 Incomplete KEGG aromatic atom types.

KEGG atom types annotate every atom in every compound of the KEGG Compound database. The KEGG atom type of an atom maps that atom to a unique chemical substructure and these substructures often map to functional groups (e.g. the atom type “O1a“ represents an oxygen of a hydroxyl group). However, the set of KEGG atom types are not complete, especially with regard to aromatic heterocycle atoms. In particular, there are no oxygen and sulfur aromatic KEGG atom types defined, which prevents full automation of aromatic substructure determination based on KEGG atom type alone. We used a simple heuristic to consider oxygen and sulfur atoms as aromatic when they are part of a ring where all other carbon and nitrogen atoms are labeled as aromatic, based on KEGG atom types. But this aromatic substructure detection approach has limitations that requires some manual inspection, as highlighted in Figure 5. KEGG Compound entry C03861 contains a 1,4-dioxin flanked by aromatic rings. The 1,4-dioxin is not aromatic. In a counter-example KEGG Compound entry C07729 contains an aromatic pyridine substructure flanked by benzyl rings. The presence of both examples illustrates why aromatic substructure detection cannot be fully automated based on the current set of KEGG aromatic atom types.

As an aside, the quinoid fragment in KEGG Compound entry C03861 is likely mislabeled as aromatic, since quinoid fragments are standardly antiaromatic[23]. This quinoid fragment was likely mislabeled as aromatic due to the whole three-ring structure obeying Huckel’s rule. While we treated KEGG-identified aromatic substructures as completely correct, this example does indicate the presence of some error in KEGG’s aromatic substructure detection methods. Comparison of Indigo to KEGG may provide means for detecting suspect KEGG aromatic substructures, but a manual inspection of all suspect substructures is not practical, especially from an automated analysis perspective. Moreover, aromatic mislabeling should not impact compound harmonization if applied consistently across databases.

![Figure 5. Compound with incomplete KEGG aromatic atom types.](image)
2.4.2 Inconsistent compound representations.

Using ID-based compound harmization, we found that there are about 10 MetaCyc compounds that contain valid aromatic substructures not detected by either the BASS or Indigo methods (Figure 6). To deal with this problem, we incorporated those valid aromatic substructures into the referencing aromatic substructure set.

![Figure 6. Compound with different aromatic representations.](image)

2.4.3. Incorrect cross-referencing.

There are some matched compounds detected by other identifiers that don’t have the same coloring identifier. Compound CPD-19437 in MetaCyc has a direct KEGG referencing compound C12187, but their coloring identifiers are different (see Figure 7). We can see that the compound representation in MetaCyc is not consistent with its counterpart in KEGG. In addition, CPD-19437 and C12187 have the same ChEBI reference compound 32074, and the representation in ChEBI is the same with that of KEGG, suggesting the representation in MetaCyc may be incorrect.

![Figure 7. Inconsistent compound representations between KEGG and MetaCyc.](image)

2.5. Estimating the error rate of the graph coloring method.

2.5.1 Ambiguous coloring identifiers.

During the compound harmonization process, tight atom and compound coloring was loosened, keeping only atom type and bond type in the atom coloring for the final steps in compound harmonization. With the loose coloring, multiple compounds in one database can have the same coloring identifier. We first tested if a compound can have a unique coloring identifier when all information is included in the atom coloring with hydrogen (H) atoms excluded. Here, we don’t count compounds with R generic group (Table 8), and results that include all compounds are described in Supplementary Table 1. Several types of compounds cannot be distinguished by the tight coloring identifier except for those duplicates (see Supplementary Figure 1). When we only include atom type and bond type in the atom coloring, many more compounds share the same coloring identifier. We are able to tell compounds with the same coloring identifier via the compound identifier from the source database after compound harmonization.

![Table 8. Compounds with the same coloring identifiers, excluding R groups.](image)
When the compound identifier is ambiguous, a compound in one database can be mapped to several different compounds in the other database during compound harmonization. For ID confirmed pairs, 28 MetaCyc compounds can be linked to more than one KEGG compounds, which is caused by inconsistency of different ID references. Also, about 478 MetaCyc compounds have several KEGG correspondences among the 1848 pairs verified by 3-leveled EC. This highlights the value in leveraging metabolic reactions and the corresponding atom mappings to disambiguate multiple possible mappings while constructing an integrated metabolic network.

2.5.2 Pseudosymmetric atoms.

Omitting information in the atom coloring can also lead to pseudosymmetric atoms. We tested if incorporation of atom charge, atom stereochemistry, or bond stereochemistry in the atom coloring will erase some symmetric atoms (Table 9). After addition of atom charge, 148 MetaCyc and 38 KEGG compounds lose symmetry. Most of them are caused by terminal atoms, like CPD-321 (Figure 8). Since either symmetric atom can be labeled with charge, asymmetry caused by atom charge can be ignored in constructing metabolic network. In addition, both databases contain compounds affected by bond and atom stereochemistry. We need to take bond and atom stereochemistry into consideration, since some enzymes are stereochemically specific. A heuristic method can be used to test if symmetric atoms are affected by bond and atom stereochemistry, and then atom coloring identifiers incorporated with bond and atom stereochemistry will be generated to overcome this issue.

Table 9. Compounds gaining asymmetry after addition of extra information in the atom naming.

| Databases | Atom stereochemistry | Atom charge | Bond stereochemistry |
|-----------|----------------------|-------------|---------------------|
| KEGG      | 232                  | 38          | 169                 |
| MetaCyc   | 219                  | 148         | 227                 |

Figure 8. Example compounds gaining asymmetry after addition of extra information.

2.5.3 Changeable graph representation.

There are two types of matched compounds that cannot be detected by coloring identifiers. One group of compounds can have either linear or circular representations (see Figure 9), and there are about 26 examples in this category. The other group is caused by resonance structures (see Figure 10), and we discovered about 46 similar cases. Artificial sets of atom mappings can be created to represent chemical transformations that are spontaneous.
3. Discussion

Here, we have developed a graph coloring method that creates unique identifiers for each atom in a compound with consideration for molecular symmetry. The atom-specific identifiers can capture additional cross-reaction atom mappings caused by symmetric atoms, which will contribute to the construction of a more complete atom-resolved metabolic network requiring information at both the compound and atom levels. Towards this overall goal, the ordered compound coloring identifiers derived from the corresponding atom coloring identifiers facilitate compound harmonization across metabolic databases, which is an essential first step in cross-database network integration. Different databases can have distinct preference in compound representations, especially for aromatic substructures. To overcome inconsistent aromatic representations between databases, we devised a pragmatic BASS method for aromatic substructure detection that leverages the labeled aromatic substructures in KEGG. Application of BASS to KEGG validated the method, providing confidence in its application to the MetaCyc database. The automatic aromatic atom detection method in Indigo further validated the comprehensiveness of our BASS aromatic substructure detection method that leverages KEGG's curated aromatic substructures, and the combination of BASS and Indigo can achieve good performance in aromatic substructure detection. This was further augmented by detecting additional aromatic substructure representations in MetaCyc through ID-based compound harmonization. In addition, compound states such as atom charge are not always the same between KEGG and MetaCyc. Therefore, identifiers like InChI that include these details to achieve an unambiguous label are not a good choice for maximizing cross-database compound harmonization in this situation. However, InChI is very useful for validation of the presented methods development.

The graph coloring method is flexible in adjusting information used in atom coloring, which can help detect more possible matched compounds with a higher false positive rate. With the coloring identifiers, we were able to detect 8865 correspondences between KEGG and MetaCyc compounds, and 5451 of them can be confirmed by other identifiers. In addition, commonality in EC numbers associated with reactions and compounds provided another avenue for both validating and predicting possible correspondence pairs. This method validated 1848 pairs unconfirmed by other identifies. While harmonizing compounds between KEGG and MetaCyc, we detected various errors in the databases by coloring identifiers, suggesting that this method can also be used for curation of current metabolic databases. Furthermore, the graph coloring method and compound harmonization approach can be used to integrate any metabolic database that provides a molfile representation of compounds, greatly facilitating future construction of more complete integrated metabolic networks.

4. Materials and Methods
4.1 Compound and metabolic reaction data

All data were downloaded directly from the corresponding databases. The KEGG COMPOUND and KEGG REACTION data is in version May 2019. MetaCyc compound and reaction data is in version 23.0, downloaded from BioCyc.

4.2 Overview of major analysis steps.

A compound can be represented as a graph where each node is an atom in the compound and each edge between atoms is a chemical bond. Based on the molfile, we are able to create a graph representation for the corresponding compound. After we detect the aromatic substructures for a compound, we can change the bonds within the aromatic substructures to aromatic type (molfile bond designation 4). After curation of aromatic substructures and double bond stereochemistry, we performed atom coloring and validation to guarantee that symmetric atoms share the same identifier and different atoms have different identifiers. Each set of atom identifiers for a compound is used to derive the corresponding compound coloring identifier. Finally, we detect corresponded pairs of compounds across two databases using ordered compound identifiers for each compound in each database. The flowchart of the overall compound harmonization procedure is shown in Figure 11.

![Flowchart of compound harmonization procedure](image)

**Figure 11.** Overview of major compound harmonization steps.

4.3 Molfile parser.

We used a modified ctfile Python 3 package[24] to parse a molfile into atom and bond blocks, and save them into the JavaScript Object Notation (JSON) format[25], facilitating access and modification.

4.4 Aromatic substructure detection.

We used two methods in aromatic substructure detection. One is based on the maximum common subgraph isomorphism, and the other is an automatic aromatic atom detection method in Indigo packages[21]. In the KEGG database, aromatic atoms in a compound are specified in its KCF file[26]. Based on the aromatic atoms, we were able to extract the aromatic substructures present within a compound, and then saved every substructure into a separate molfile. If several aromatic rings are connected, we would fuse them together as one substructure. Then, we built a set of all aromatic
substrutures detected from the KEGG compounds without duplication. Furthermore, we manually inspected the set of aromatic substructures to ensure data quality. With this curated set of referencing aromatic substructures, we tested each compound in a database for the presence of any of these aromatic substructures using the Biochemically Aware Substructure Search (BASS) method[20]. We analyzed KEGG to validate the aromatic substructure detection method itself. Then, we analyzed MetaCyc and labeled the bonds of detected aromatic substructures as aromatic. Furthermore, valid aromatic substructures in MetaCyc compounds can be detected by Indigo and other IDs. Finally, we created 366 KEGG-derived and 21 MetaCyc-derived aromatic substructures in the referencing aromatic substructure set.

4.5. Identification of double bond stereochemistry

The C=C double bond stereochemistry is not clearly specified in the molfile in both databases. To distinguish cis/trans stereoisomers, we adopted a method for automated identification of double bond stereochemistry[27]. This method requires fully hydrogenated compounds. Therefore, we first used Open Babel[28] to add hydrogen atoms for every compound, and then performed the calculation.

4.6. Graph coloring method

The graph coloring method is based on a breadth first search algorithm[29]. This method names each atom based on its own and neighbors’ chemical information, which can include atom type, atom charge, atom stereochemistry, isotope, bond type, and bond stereochemistry. The method is flexible in adjusting the chemical information included in the atom coloring. A flowchart of the graph coloring method is shown in Figure 12. First, the method named each atom with its own chemical information, which will be saved as the 0_layer identifier and the start of the current atom identifier. Then, the method builds a dictionary that relates each atom with its 0_layer identifier and directly linked atoms. Directly bonded atoms of each atom are initialized as its neighbors. The method continues to extend the name of each atom, adding information about its neighbors into the 0_layer dictionary to its current identifier, and updating neighbors with neighbors’ neighbors that have not been used in extending the name of that atom. The method first repeats this process 3 times for all the atoms to avoid early stopping that can lead to non-unique compound coloring identifiers. Then, the method checks if an atom has a unique identifier. Atom naming will continue for those atoms that still share the same identifiers with other atoms until all the atoms in the compound have been used in name extension. Finally, the current name for each atom will be its coloring identifier. Compound C00047 in KEGG database (S Figure 2) is used as an example to illustrate how it works (S Table 2).
4.7. Atom coloring validation and recolor.

The atom coloring validation and recoloring are also based on a breadth first search algorithm. The atom coloring validation flowchart is shown in Figure 13. For atoms with the same coloring identifier, we checked if neighbors of these atoms are also the same, layer by layer, until all the atoms in the compound have been tested. Then, the recoloring method will correct atoms with the same identifier that don’t have the same neighbors. The recoloring process is similar to the graph coloring method. Instead of creating a 0_layer identifier dictionary, we will use a full identifier dictionary. In addition, we only color atoms to where they have different neighbors to distinguish between them.
4.8. Creation of compound identifier bases on atom identifiers.

Once we create the identifiers for all the atoms in a compound, we can combine the number of atoms with the same identifier along with the atom coloring identifier. We sorted all the substrings, and then concatenate them together to form an ordered coloring identifier for the compound. The formulation is shown in Equation 1.

\[
\text{Compound color identifier} = (n_1)[a_1](n_2)[a_2](n_3)[a_3] \ldots (n_k)[a_k]
\] (1)

4.9. Prediction of possible compound correspondence via metabolic reactions.

We connected each compound with the metabolic reactions it is a part of. For matched compounds between KEGG and MetaCyc, we tested if the compound shares at least one metabolic reaction indicated by the EC number in both databases.

Data Availability: All data used and the results generated in this manuscript are available on: https://figshare.com/s/fab8568172b55d89e97.

Supplementary Materials: Table S1: Compounds with the same coloring identifiers, which includes R groups, Figure S1: Representative compounds that cannot be distinguished by coloring identifier, Figure S2: KEGG Compound C00047, Table S2: Generation of atom identifiers for compound C00047 via graph coloring method, Spreadsheet S1: Inconsistency between KEGG and MetaCyc.

Author Contributions: HJ and HNBM worked together on the design of the experiments and the analysis of the results. HJ wrote the manuscript and HNBM revised it. JMM contributed to the detection of aromatic substructures. All authors have read and approved the manuscript.

Funding:

The work was supported in part by grant NSF 1419282 (PI Moseley).

Acknowledgments: Thanks for Dr. Robert B. Grossman for examining aromatic substructures.

Conflicts of Interest: The authors declare that they have no competing interests.
References

1. Young JD. INCA: a computational platform for isotopically non-stationary metabolic flux analysis. Bioinformatics. 2014;30(9):1333-1335. doi:10.1093/bioinformatics/btu015

2. Antoniewicz MR. Methods and advances in metabolic flux analysis: a mini-review. J Ind Microbiol Biotechnol. 2015;42(3):317-325. doi:10.1007/s10295-015-1585-x

3. Moseley HN, Lane AN, Belshoff AC, Higashi RM, Fan TW. A novel deconvolution method for modeling UDP-N-acetyl-D-glucosamine biosynthetic pathways based on 13C mass isotopologue profiles under non-steady-state conditions. BMC Biol. 2011;9(1):37. doi:10.1186/1741-7007-9-37

4. Jin H, Moseley HNB. Moiety modeling framework for deriving moiety abundances from mass spectrometry measured isotopologues. BMC Bioinformatics. 2019;20(1):524. doi:10.1186/s12859-019-3096-7

5. Jin H, Moseley HNB. Robust Moiety Model Selection Using Mass Spectrometry Measured Isotopologues. Metabolites. 2020;10(3):118. doi:10.3390/metabo10030118

6. Rathahao-Paris E, Alves S, Junot C, Tabet J-C. High resolution mass spectrometry for structural identification of metabolites in metabolomics. Metabolomics. 2016;12(1):10. doi:10.1007/s11306-015-0882-8

7. Fan TW-M, Lorkiewicz PK, Sellers K, Moseley HNB, Higashi RM, Lane AN. Stable isotope-resolved metabolomics and applications for drug development. Pharmacol Ther. 2012;133(3):366-391. doi:10.1016/j.pharmthera.2011.12.007

8. Creek DJ, Chokkathukalam A, Janekovics A, Burgess KE V., Breitling R, Barrett MP. Stable Isotope-Assisted Metabolomics for Network-Wide Metabolic Pathway Elucidation. Anal Chem. 2012;84(20):8442-8447. doi:10.1021/ac3018795

9. Aittokallio T. Graph-based methods for analysing networks in cell biology. Brief Bioinform. 2006;7(3):243-255. doi:10.1093/bib/bbl022

10. Chokkathukalam A, Kim D, Barrett MP, Breitling R, Creek DJ. Stable isotope-labeling studies in metabolomics: new insights into structure and dynamics of metabolic networks. Bioanalysis. 2014;6(4):511-524. doi:10.4155/bio.13.348

11. Dalby A, Nourse JG, Hounshell WD, et al. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Model. 1992;32(3):244-255. doi:10.1021/ci00007a012

12. Kotera M, Hattori M, Oh M-A, et al. RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. Genome Informatics. 2004;15(1):P062.

13. Latendresse M, Malerich JP, Travers M, Karp PD. Accurate Atom-Mapping Computation for Biochemical Reactions. J Chem Inf Model. 2012;52(11):2970-2982. doi:10.1021/ci3002217

14. Pitkänen E, Jouhten P, Rousu J.Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol. 2009;3(1):103. doi:10.1186/1752-0509-3-103

15. Heath AP, Bennett GN, Kavraki LE. Finding metabolic pathways using atom tracking. Bioinformatics. 2010;26(12):1548-1555. doi:10.1093/bioinformatics/btq223

16. Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. Bioinformatics. 2014;30(14):2043-2050. doi:10.1093/bioinformatics/btu150

17. Altman T, Travers M, Kothari A, Caspi R, Karp PD. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinformatics. 2013;14(1):112. doi:10.1186/1471-2105-14-112

18. Dashti H, Wedell JR, Westler WM, Markley JL, Eghbalnia HR. Automated evaluation of consistency within
the PubChem Compound database. Sci Data. 2019;6(1):190023. doi:10.1038/sdata.2019.23

19. Dashti H, Westler WM, Markley JL, Eghbalnia HR. Unique identifiers for small molecules enable rigorous labeling of their atoms. Sci Data. 2017;4(1):170073. doi:10.1038/sdata.2017.73

20. Mitchell JM, Fan TW-M, Lane AN, Moseley HNB. Development and in silico evaluation of large-scale metabolite identification methods using functional group detection for metabolomics. Front Genet. 2014;5. doi:10.3389/fgene.2014.00237

21. Indigo Toolkit. Available online: https://lifescience.opensource.epam.com/indigo/index.html (accessed on 30-Apr-2020)

22. Matsuta Y, Ito M, Tohsato Y. ECOH: An Enzyme Commission number predictor using mutual information and a support vector machine. Bioinformatics. 2013;29(3):365-372. doi:10.1093/bioinformatics/bts700

23. Szatylowicz H, Krygowski TM, Solà M, et al. Why 1,2-quinone derivatives are more stable than their 2,3-analogues? Theor Chem Acc. 2015;134(3):35. doi:10.1007/s00214-015-1635-5

24. Andrey S. ctfile. Available online: https://github.com/MoseleyBioinformaticsLab/ctfile (accessed on 26-Nov-2019)

25. Bray T, ed. The JavaScript Object Notation (JSON) Data Interchange Format.; 2014. doi:10.17487/rfc7159

26. Hattori M, Okuno Y, Goto S, Kanehisa M. Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways. J Am Chem Soc. 2003;125(39):11853-11865. doi:10.1021/ja036030u

27. Teixeira AL, Leal JP, Falcao AO. Automated Identification and Classification of Stereochemistry: Chirality and Double Bond Stereoisomerism. March 2013. http://arxiv.org/abs/1303.1724. O’Boyle

28. NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminform. 2011;3(1):33. doi:10.1186/1758-2946-3-33

29. Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms. 2nd ed. MIT Press and McGraw-Hill; 2001.
Supporting Information for

Atom Identifiers Generated by a Graph Coloring Method Enable Compound Harmonization Across Metabolic Databases

Huan Jin 1, Joshua M. Mitchell 23,4 and Hunter N.B. Moseley 2, 3, 4, 5, *

1 Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, KY 40536, USA; huan.jin@uky.edu
2 Department of Molecular & Cellular Biochemistry, University of Kentucky, Lexington, KY 40536, USA
3 Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA
4 Resource Center for Stable Isotope Resolved Metabolomics, University of Kentucky, Lexington, KY 40536, USA
5 Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA

* Correspondence: hunter.moseley@uky.edu; Tel.: 859-218-2964
**S Table 1.** Compounds with the same coloring identifiers, which includes R groups.

| Databases | Tight coloring identifier | Loose coloring identifier |
|-----------|---------------------------|---------------------------|
| KEGG      | 209                       | 1132                      |
| MetaCyc   | 449                       | 1638                      |

**S Figure 1.** Representative compounds that cannot be distinguished by coloring identifier. A) Compound and its radical form; B) Compound containing repeated substructure; C) Compound with R representing a generic group; D) Isomers containing C=N; E) Compounds after curation of aromatic substructures.

**S Figure 2.** KEGG Compound C00047.

**S Table 2.** Generation of atom identifiers for compound C00047 via graph coloring method.

| Round | Atom identifier | Atom index       |
|-------|-----------------|------------------|
| 1     | C               | 1, 2, 3, 5, 8, 9 |
|       | N               | 4, 10            |
| Atom | 2 | 3 |
|------|---|---|
| O    | 6, 7 |  |
| C(C(C,1)(C,1)(N,1)) | 1 |  |
| C(C(C,1)(C,1)) | 2, 5, 8 |  |
| C(C(C,1)(O,1)(O,2)) | 3 |  |
| O(N(C,1)) | 4, 10 |  |
| O(O(C,1)) | 6 |  |
| O(O(C,2)) | 7 |  |
| C(C(C,1)(N,1)) | 9 |  |
| C(C(C,1)(C,1)(N,1))(C(C,1)(C,1)(C,1)(O,1)(O,2)N(C,1)) | 1 |  |
| C(C(C,1)(C,1))(C(C,1)(C,1)(C,1)(C,1)(N,1)) | 2 |  |
| C(C(C,1)(O,1)(O,2))(C(C,1)(C,1)(N,1)O(C,1)O(C,2)) | 3 |  |
| N(N(C,1))(C(C,1)(C,1)(N,1)) | 4 |  |
| C(C(C,1)(C,1))(C(C,1)(C,1)(C,1)(C,1)) | 5 |  |
| O(O(C,1))(C(C,1)(O,1)(O,2)) | 6 |  |
| O(O(C,2))(C(C,1)(O,1)(O,2)) | 7 |  |
| C(C(C,1)(C,1))(C(C,1)(C,1)(C,1)(N,1)) | 8 |  |
| C(C(C,1)(N,1))(C(C,1)(C,1)(N(C,1)) | 9 |  |
| N(N(C,1))(C(C,1)(N,1)) | 10 |  |

Only chemical information of atom type and bond type is included in atom naming. The first three rounds of naming are shown above.