# Efficient Validation of Metabolic Pathway Databases

**Liliana Félix** **Gabriel Valiente**

Department of Software
Technical University of Catalonia
E-08034 Barcelona, Spain

## Abstract

Metabolic pathway databases such as KEGG contain information on thousands of enzymatic reactions drawn from the biomedical literature. Ensuring consistency of such large metabolic pathways is essential to their proper use. In this paper, we develop an efficient method to determine consistency of an important class of enzymatic reactions, and test the method on the latest release of the KEGG LIGAND database.

**Keywords**: metabolic pathway, enzymatic reaction, classification, consistency, validation, KEGG.

## 1. Introduction

Biochemical pathways, such as metabolic, regulatory, and signal transduction pathways, constitute complex networks of functional and physical interactions between molecular species in the cell [3]. Current knowledge on chemical compounds, biochemical reactions, and biochemical pathways in cellular processes, is accumulated in several biological databases. In particular, metabolic pathway databases such as KEGG [7] contain information on thousands of enzymatic reactions drawn from the biomedical literature. However, no thorough validation of the information contained in these biological databases has been performed yet.

Validation of a metabolic pathway database can be made by metabolic reconstruction [2] or by comparison against the artificial chemistry defined by the chemical compounds and enzymatic reactions that are stored in the database [8, 9]. In this paper, we focus on the problem of ensuring consistency of the enzymatic reactions that are stored in a metabolic pathway database. This is known as the automatic mapping problem: to determine the optimal atom-atom mapping between substrate and product of an enzymatic reaction.

Unfortunately, the automatic mapping problem is NP-hard, even in a constrained form [1]. Therefore, previous work on the automatic mapping problem is either centered on heuristic maximum common subgraph algorithms [2, 6] or focused on particular cases [1].

In this paper, we exploit the knowledge of the atomic rearrangement pattern in the enzymatic reactions stored in a metabolic pathway database, to reduce the automatic mapping problem to a series of chemical substructure searches between the substrate and the product chemical graph [8, 9]. The advantage of our approach is the availability of fast subgraph isomorphism and polynomial-time isomorphism algorithms for chemical graphs [13].

We have implemented tool support for our method, and performed an exhaustive validation of a substantial portion of the KEGG LIGAND [5] database. As a result, we have ensured consistency of a large number of enzymatic reactions, and have identified several inconsistencies in the information about chemical compounds and enzymatic reactions stored in KEGG.

## 2. Materials and Methods

Most enzymatic reactions can be classified according to the pattern of atomic rearrangement, into the following four main classes [9]: combination, decomposition, displacement, and exchange reactions. In c*ombination* reactions, two or more substrates combine to form a single product, according to the pattern: A + B <=> AB. In d*ecomposition* reactions, a single substrate is decomposed or broken down into two or more products, according to the pattern: AB <=> A + B. In d*isplacement* reactions, also called single replacement reactions, one of the substrates is displaced into another one, according to the pattern: A + B−C <=> A−C + B. Finally, in e*xchange* reactions, also called double replacement reactions, one of the substrates is exchanged by another one, according to the pattern: A−B + C−D <=> A−D + C−B.

We have observed several classes of decomposition reactions. For instance, those reactions in which a single bond in the substrate is broken and another single bond is turned into a double bond, according to the pattern: A−B−C <=> A + B=C. Besides, there are also pseudo-exchange reactions, in which some single and double bonds in the substrate

are exchanged, according to the pattern: A−B + C=D <=> A=B + C−D.

The automatic mapping problem can be solved for decomposition reactions as follows. Given three chemical graphs X, Y, Z, where X is the substrate and Y, Z are the products, there is an enzymatic reaction decomposing X into Y and Z according to the pattern: A−B−C <=> A + B=C if and only if there is a subgraph X' of X isomorphic to Y such that X minus X' is isomorphic to a subgraph of Z.

Combination and decomposition reactions are symmetrical to each other, and the automatic mapping problem can be solved for combination reactions by a similar procedure.

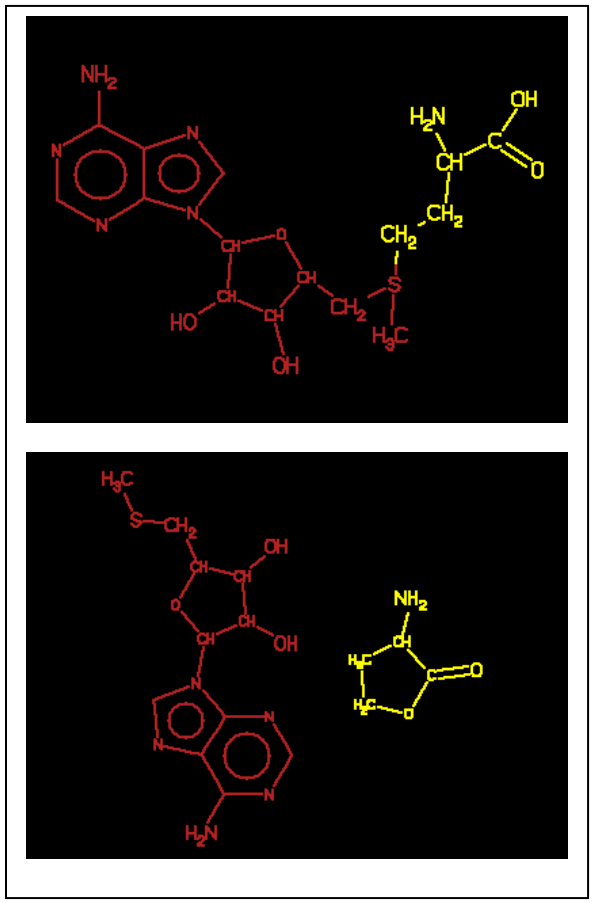

Fig. 1: Visualization of a combination reaction, KEGG LIGAND R00180. Color is used to show the correspondence between substrate (top) and product (bottom), and atom numbering is omitted for readability purposes.

Finally, the automatic mapping problem can be solved for displacement reactions as follows. Given four chemical graphs X, Y, Z, W, where X, Y are the substrates and Z, W are the products, there is an enzymatic reaction displacing X and Y into Z and W if and only if there is a subgraph Y' of Y isomorphic to W, there is a subgraph Z' of Z isomorphic to X, and Y minus Y' and Z minus Z' are nonempty and isomorphic. Otherwise, if Y minus Y' and Z minus Z' are empty, there is a pseudo-exchange reaction between X, Y and Z, W.

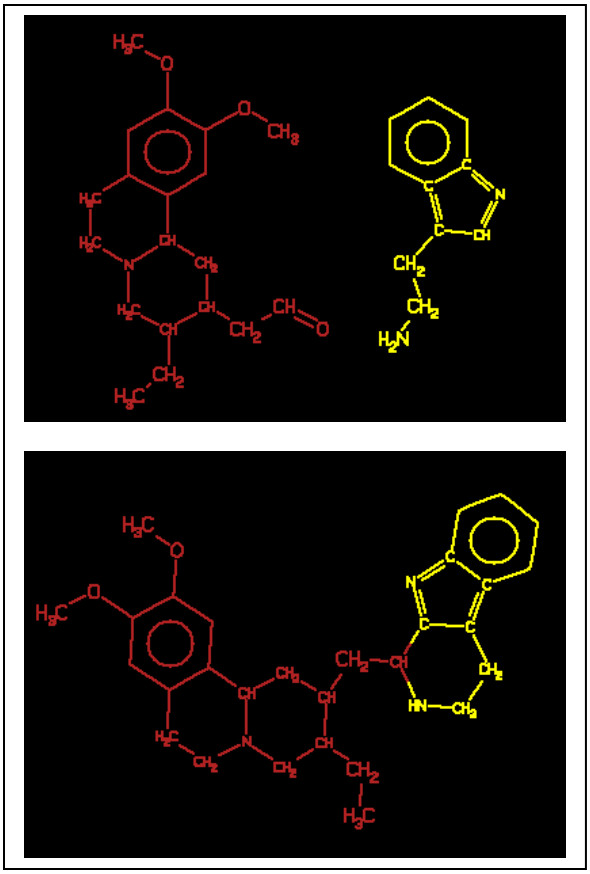

Fig. 2: Visualization of a decomposition reaction, KEGG LIGAND R05895.

The previous descriptions translate into detailed algorithms in a straightforward way. For instance, the detailed description in pseudo-code of the automatic mapping algorithm for decomposition reactions is the following. Given three chemical graphs X, Y, Z, where X is the substrate and Y, Z are the products, automap (X, Y, Z) returns an optimal mapping M of substrate to product atoms, or an empty atom mapping if there is no subgraph X' of X isomorphic to Y, or X minus X' and Z are not isomorphic.

```
automap (X, Y, Z)
  for each occurrence X’ of Y in X
    let M be the atom mapping of Y to X
    let X’’ := X minus X’
    if X’’ and Z are isomorphic then
```

```
            return M
        end if
    end for
    return an empty atom mapping
end
```

In a similar vein, the detailed description in pseudo-code of the automatic mapping algorithm for displacement reactions is the following. Given four chemical graphs X, Y, Z, W, where X, Y are the substrates and Z, W are the products, automap (X, Y, Z, W) returns an optimal mapping M of substrate to product atoms, or an empty atom mapping if there is no subgraph Y' of Y isomorphic to W or no subgraph Z' of Z isomorphic to X, or Y minus Y' and Z minus Z' are not isomorphic.

```
automap (X, Y, Z, W)
    for each occurrence Y' of W in Y
        let M be the atom mapping of W to Y
        let Y'' := Y minus Y'
        for each occurrence Z' of X in Z
            let M' := M
            extend M' with the mapping of X to Z
            let Z'' := Z minus Z'
            if Y'' and Z'' are isomorphic then
                return M'
            end if
        end for
    end for
    return an empty atom mapping
end
```

## 3. Results

We have implemented some tool support for our method, using the PerlMol collection of Perl modules for computational chemistry [11].

On the one hand, given a candidate enzymatic reaction, a first tool solves the automatic mapping problem of the substrate to the product, if their chemical formulas are compatible. In such a case, the output takes the form of a mapping of substrate atoms to product atoms with the least possible number of broken and created bonds.

On the other hand, given an enzymatic reaction and a mapping of substrate atoms to product atoms, a second tool produces a pair of GIF files containing a diagram of the reaction, with appropriate atom numbering to illustrate the correspondence among substrate and product atoms. Samples of the GIF files obtained with this tool are shown in Fig. 1 to Fig. 4. This tool relies on the DEPICT algorithm and tool support [12].

Using these tools, we performed an exhaustive validation of a substantial portion of the KEGG LIGAND [5] database, release 29.0, which comprises a total of 25,930 occurrences of 5,302 compounds in 6,304 enzymatic reactions. We have *cured* the database by discarding the 344 reactions that involve carbohydrate structures from the KEGG GLYCAN database, discarding the 589 reactions that involve compounds of unknown structure (for which no MOL file is available in KEGG LIGAND), and instantiating all wildcards in the compounds (substituting carbon dihydride for an * group or an X group; cycloheptane for an R group; and 1 for *n*, the degree of polymerization). The resulting database comprises a total of 21,961 occurrences of 4,493 compounds in 5,371 enzymatic reactions.

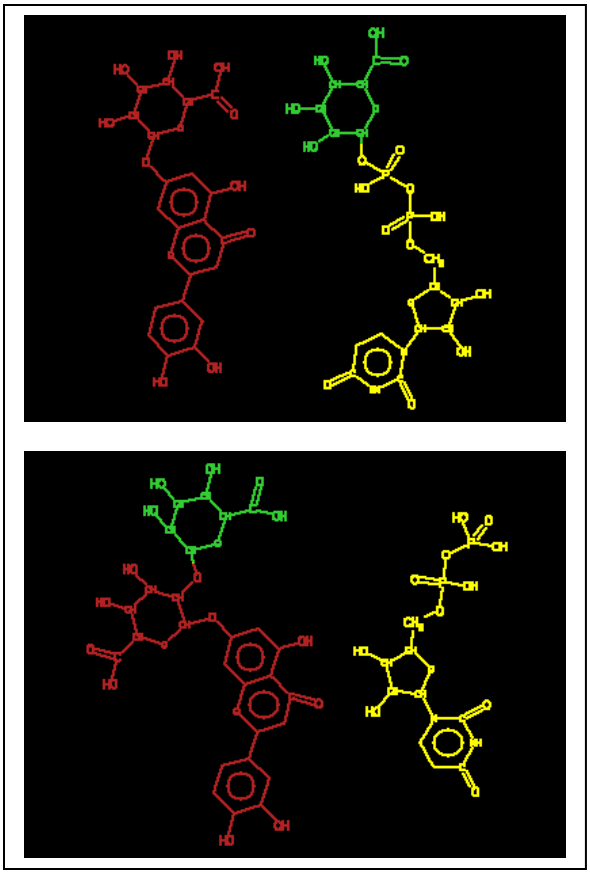

Fig. 3: Visualization of a displacement reaction, KEGG LIGAND R06827.

We have focused on the two groups of most frequent enzymatic reactions: 2,598 reactions with four compounds and 651 reactions with three compounds among their substrate and product. Among the former, 212 reactions are incompatible, and 1,661 reactions were found to be consistent (6 of them are

decomposition reactions, of the form A−B−C−D <=> A + B + C−D, 60 are pseudo-exchange reactions of the form A−B + C=D <=> A=B + C−D, and 1,595 are displacement reactions, of the form A + B−C <=> A−C + B.

Among the latter, 54 are incompatible, and 426 reactions were found to be consistent (they are all decomposition reactions of the form A−B−C <=> A + B=C). The remaining 725 reactions with four compounds and 171 reactions with three compounds require further analysis.

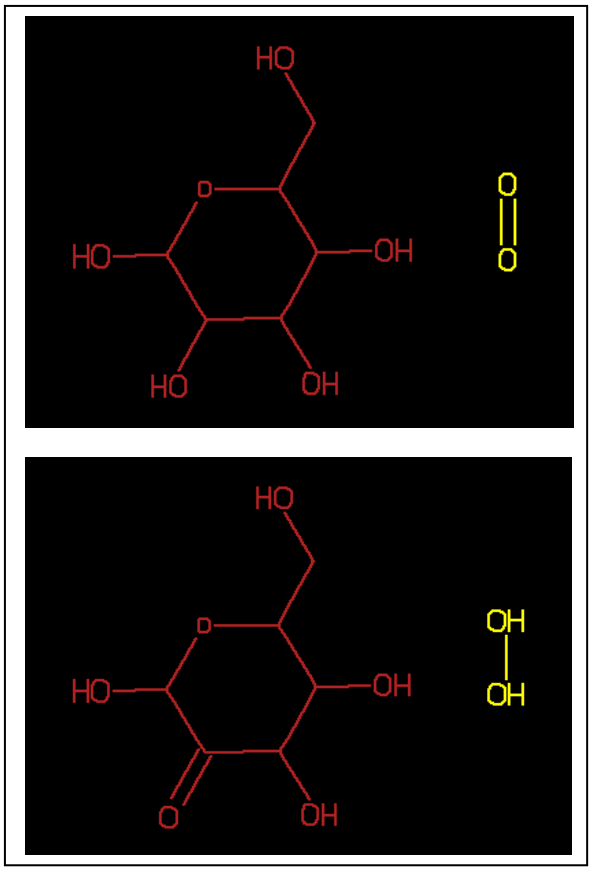


Fig. 4: Visualization of a pseudo-exchange reaction, KEGG LIGAND R00302.

## 4. Acknowledgement

The research described in this paper has been partially supported by the Spanish CICYT through project GRAMMARS (TIN2004-07925-C03-01), and by the Japan Society for the Promotion of Science through Long-term Invitation Fellowship L05511 for visiting JAIST (Japan Advanced Institute of Science and Technology).

## 5. References

[1] T. Akutsu, "Efficient Extraction of Mapping Rules from Enzymatic Reaction Data," J. Comput. Biol., vol. 11, no. 2-3, pp. 449-462, 2004.

[2] M. Arita, "Metabolic Reconstruction using Shortest Paths," Simulat. Pract. Theory, vol. 8, no, 2, pp. 109-125, 2000.

[3] Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak, "An Overview of Data Models for the Analysis of Biochemical Pathways," Briefings in Bioinformatics, vol. 4, no. 3, pp. 246-259, 2003.

[4] L. Félix, F. Rosselló, and G. Valiente, "Artificial Chemistries and Metabolic Pathways," *Proc. 5th Annual Spanish Bioinformatics Conf.*, pp. 56-59, 2004.

[5] S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa, "LIGAND: Database of Chemical Compounds and Reactions in Biological Pathways," Nucleic Acids Res., vol. 30, no. 1, pp. 402-404, 2002.

[6] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a Chemical Structure Comparison Method for Integrated Analysis of Chemical and Genomic Information in the Metabolic Pathways," J. Am. Chem. Soc., vol. 125, no. 1, pp. 11853-11865, 2003.

[7] M. Kanehisa and S. Goto, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Res., vol. 28, no. 1, pp. 27-30, 2000.

[8] F. Rosselló and G. Valiente, "Analysis of Metabolic Pathways by Graph Transformation," *Proc. 2nd Int. Conf. Graph Transformation*, Lect. Notes Comput. Sci., vol. 3256, pp. 70-82, 2004.

[9] F. Rosselló and G. Valiente, "Chemical Graphs, Chemical Reaction Graphs, and Chemical Graph Transformation," *Proc. 2nd Int. Workshop Graph-Based Tools*, Electr. Notes Theor. Comput. Sci., vol. 127, no. 1, pp. 157-166, 2005.

[10] F. Rosselló and G. Valiente, "Graph Transformation in Molecular Biology," *Formal Methods in Software and System Modeling*, Lect. Notes Comput. Sci., vol. 3393, pp. 116-133, 2005.

[11] I. Tubert-Brohman, "Perl and Chemistry," The Perl Journal, vol. 8, no. 6, pp. 3-5, 2004.

[12] D. Weininger, "SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures," J. Chem. Inf. Comp. Sci., vol. 30, no. 3, pp. 237-243, 1990.

[13] D. Weininger, "SMILES. 2. Algorithm for Generation of Unique SMILES Notation," J. Chem. Inf. Comp. Sci., vol. 29, no. 1, pp. 97-101, 1989.