

Christophe MULLER, Gilles MARCOU and Alexandre VARNEK*

Laboratoire d'Infochimie, Université de Strasbourg, 4 rue B. Pascal, Strasbourg, 67000, France

Summary: Atom-Atom Mapping of chemical reactions represents a difficult task. In fact, there exists no unique algorithm providing with a definite solution of this problem; one can speak only about more or less successful techniques. Here, incorrect mapping has been identified using SVM and a JRip models involving fragment descriptors generated from the Condensed Graphs of Reaction. As an example, we used mapping of metabolic reactions from the KEGG database performed with the ChemAxon tools. Developed classification models retrieve incorrectly mapped reactions with very high rate (up to 100%).

Data Preparation

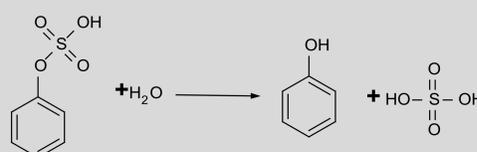
ChemAxon Standardizer has been used for atom-atom mapping of the dataset of ≈ 850 reactions for the 3 first enzymatic classes (*i.e.*, oxidoreductase, transferase and hydrolase) extracted from KEGG. All these reactions were also manually mapped taking into account the published mechanisms. In such a way, 95 incorrectly mapped reactions were detected.

1) Training set: 62 incorrectly mapped reactions (IMR) + 62 correctly mapped reactions (CMR).

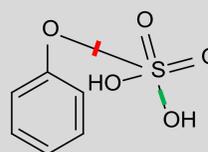
2) Test set: 33 IMR + 33 CMR

-All reactions were transformed into Condensed Graphs of Reaction (CGR), from which ISIDA fragment descriptors were generated

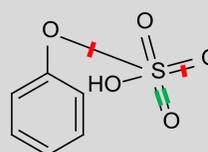
CGR



CGR of correct mapping:



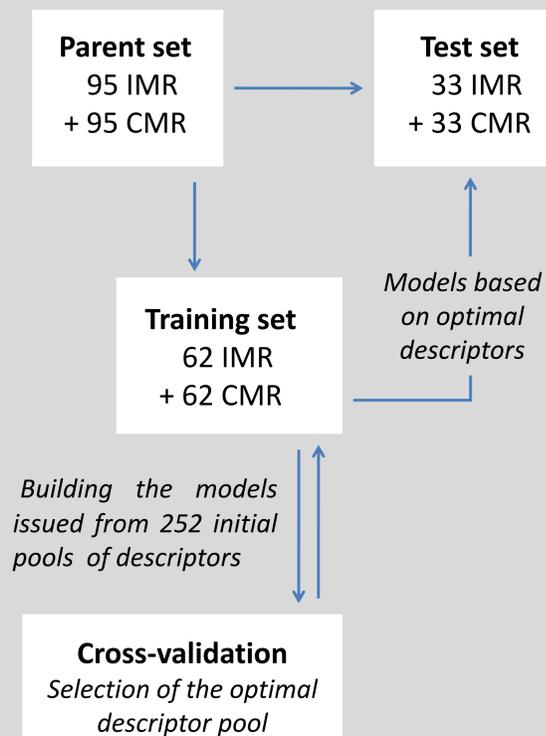
CGR of incorrect mapping:



Dynamic bonds:

- formed bonds: green lines
- broken bonds: red lines

Modeling Workflow



Obtaining and validation of models

Descriptors: ISIDA fragment descriptors: « sequences » of atoms and bonds, and « augmented atoms » containing up to 10 atoms and, at least, one dynamical bond.

Machine learning methods: SVM and JRip (rule learner).

Validation and selection of models:

- 20x5 fold external CV
- Scrambling

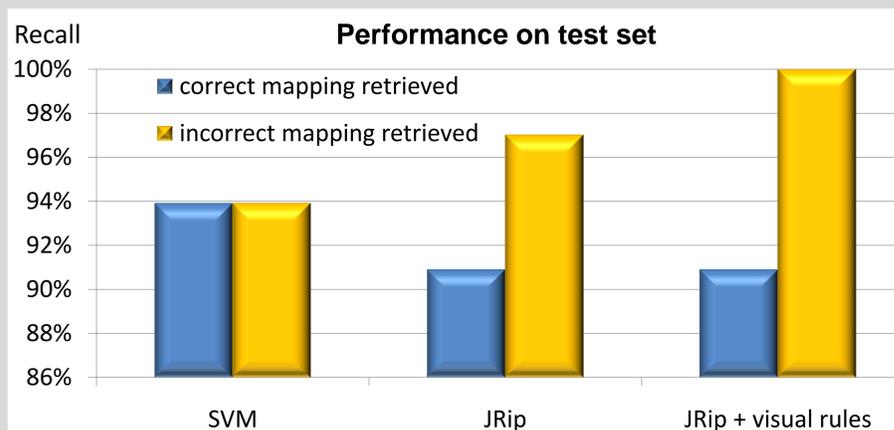
Rules based model for incorrect mapping:

- $\text{---} > 0$ and $\text{---} \neq 1$ and $\text{---} > 0$
- $\text{---} \text{---} \text{---} \text{---} \text{---} \text{---} < 1$
- $\text{---} \text{---} \text{---} < 1$
- $\text{---} \text{---} \text{---} \text{---} = 0$
- $\text{C} \text{---} \text{O} \text{---} \text{C} > 0$
- $\text{C} \text{---} \text{CH}_2 \text{---} \text{C} > 0$

JRip (1-4): Fragments involving only bonds

Visual observations (5-6): Fragments involving both atoms and bonds

Predictive performance of the models



References:

¹ "Chemoinformatics Approaches to Virtual Screening", A. Varnek and A. Tropsha, Eds., RSC Publishing, 2008

Conclusion:

- Developed classification models retrieve incorrect mapping with a very high rate (*Recall* = 94 – 100%).
- The "JRip + visual rules" model has been found more predictive than SVM and JRip models.