JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Models for Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by Automated Algorithms

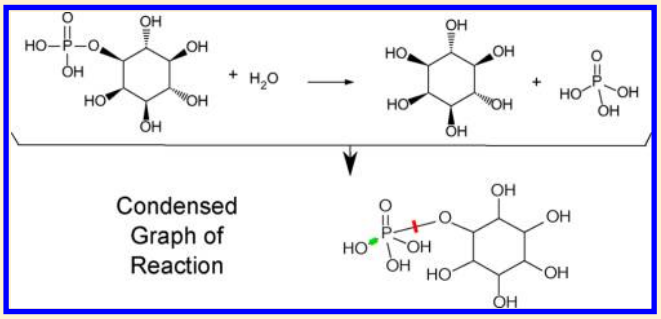
Christophe Muller,[†] Gilles Marcou,[†] Dragos Horvath,[†] João Aires-de-Sousa,[‡] and Alexandre Varnek*[,†]

[†]Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg, 4, rue B. Pascal, Strasbourg 67000, France

[‡]CQFB, REQUIMTE, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Supporting Information

ABSTRACT: Machine learning (SVM and JRip rule learner) methods have been used in conjunction with the Condensed Graph of Reaction (CGR) approach to identify errors in the atom-to-atom mapping of chemical reactions produced by an automated mapping tool by ChemAxon. The modeling has been performed on the three first enzymatic classes of metabolic reactions from the KEGG database. Each reaction has been converted into a CGR representing a pseudomolecule with conventional (single, double, aromatic, etc.) bonds and dynamic bonds characterizing chemical transformations. The ChemAxon tool was used to automatically detect the matching atom pairs in reagents and products. These automated mappings were analyzed by the human expert and classified as "correct" or "wrong". ISIDA fragment descriptors generated for CGRs for both correct and wrong mappings were used as attributes in machine learning. The learned models have been validated in *n*-fold cross-validation on the training set followed by a challenge to detect correct and wrong mappings within an external test set of reactions, never used for learning. Results show that both SVM and JRip models detect most of the wrongly mapped reactions. We believe that this approach could be used to identify erroneous atom-to-atom mapping performed by any automated algorithm.



INTRODUCTION

A chemical reaction is the transformation of particular bonds of reactants resulting in formation of products. Identification of such bonds is possible if a correspondence of atoms in reactants and products (*atom-to-atom mapping*, AAM) is known. The automatized AAM procedures are implemented in most of the chemical database management systems like SYMYX, ChemAxon, and ACD/Labs. However, they do not always correctly identify related atoms in reactants and products. The point is that the mapping should account for the reaction mechanism, which is not easy to implement into the algorithm. Esters hydrolysis is a typical example: in order to perform AAM one should know to which product − acid or alcohol − belongs to the bridging oxygen atom. Since the 1980s, various methods of automated AAM were reported in the literature. Thus, Jochum et al.[1] suggested the AAM algorithm based on the empirical principle that "most chemical reactions proceed along a pathway of minimum chemical distance, i.e. with a redistribution of a minimum number of valence electrons". Unfortunately this simple strategy is not always adequate, because not all chemical reactions follow the minimal chemical distance principle or, in some cases, a combinatorial explosion of possible solutions may occur. Körner and Apostolakis' algorithm to map metabolic reactions[2] assumes that a correct mapping follows the lowest imaginary transition state energy (ITSE), which is computed as the sum of the weights of reacting bonds. The inconvenience of this method is that for some reactions it finds several alternative mappings or no correct solution. ChemAxon, a popular chemoinformatics software provider, uses an algorithm based on the search of isomorph subgraphs in reactant/product pairs.[3] It is not clear whether complementary information about reaction mechanisms is taken into account by it.

To sum up, nowadays there is no unique AAM algorithm which correctly maps all possible chemical reactions. Developing a new algorithm incorporating the knowledge on reaction mechanisms is an extremely challenging task. More realistically and readily feasible, as will be shown in this article, a first step toward more reliable automated AAM could be the learning of models able to distinguish correctly mapped from erroneously mapped reactions. Agreed, these may not tell what the correct mapping should have been, but they may significantly reduce the time invested by a human expert in charge of AAM, by focusing his or her attention on the relevant, problematic cases only.

The latter scenario is considered in this paper using as example a set of metabolic reactions from the KEGG database.[4] These data have been selected because any metabolic reaction can be unmistakably mapped thanks to the reactant pairs approach.[5] Here, for a selected set of metabolic reactions we compare a manual mapping with automated AAM performed with the

Received: September 1, 2012
Published: November 20, 2012

dx.doi.org/10.1021/ci300418q | *J. Chem. Inf. Model.* 2012, 52, 3116−3122

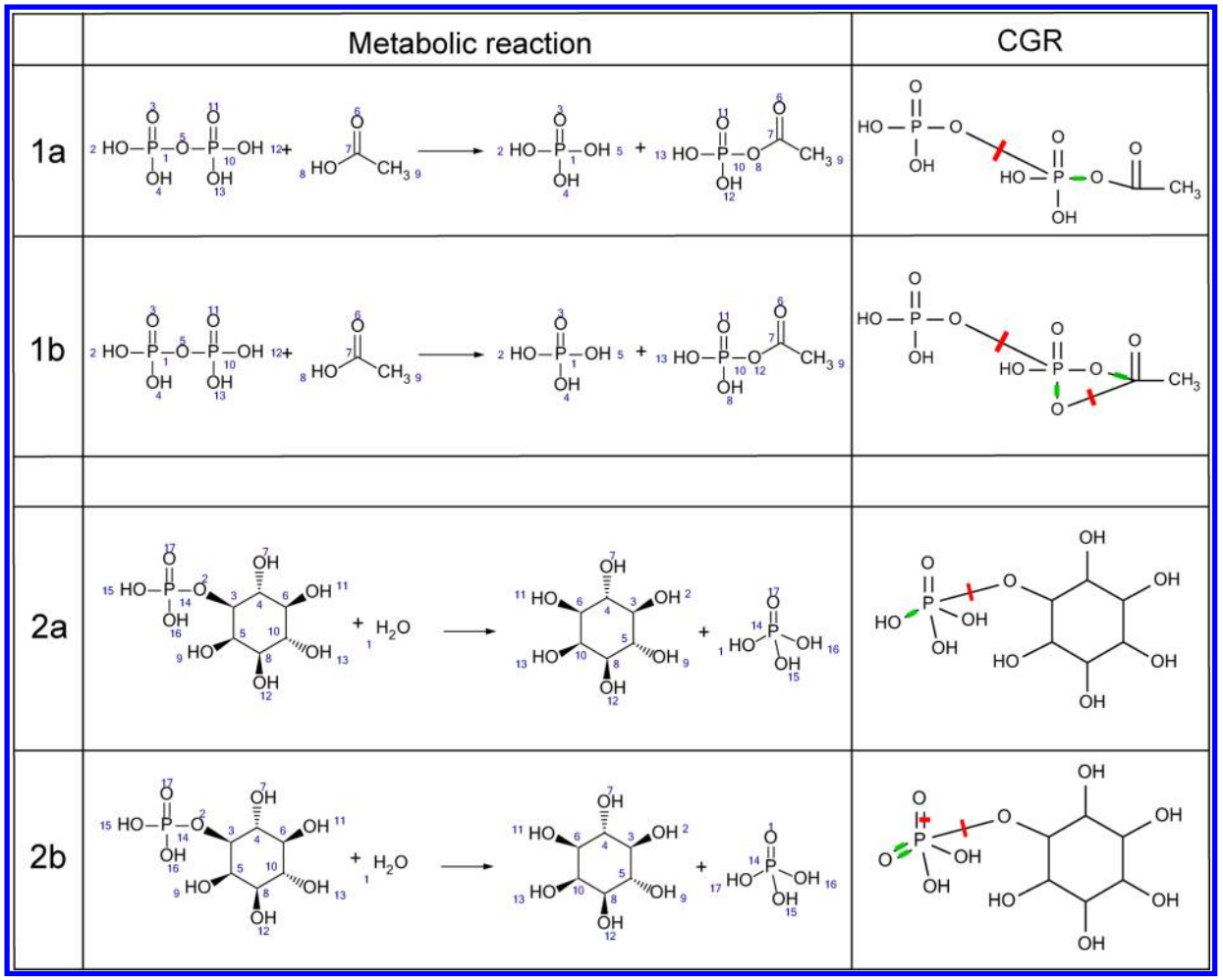


Figure 1. Examples of correct (a) and erroneous (b) atom-to-atom mapping leading to different Condensed Graphs of Reaction. CGR involves both conventional (single, double, ..., etc.) and dynamical bonds describing chemical transformations. Here, and correspond to broken and created single bonds, respectively. One can see that CGRs corresponding to correct AAM contain less dynamical bonds than CGRs corresponding to wrong AAM.

ChemAxon Standardizer software (release 2009).[6] Reactions for which automated and manual mapping differ represent the "wrong" AAM class of a modeling data set, serving to learn classification models.

Typically in structure−property modeling, machine-learning methods are applied to individual molecules represented by descriptor vectors. Since any chemical reaction involves several molecular species (at least, two), at the first step they were transformed into Condensed Graphs of Reaction (CGR) − some sort of pseudomolecules condensing information about all reactants and products[7] (Figure 1). Atoms' connectivity in CGR depends on atom−atom mapping, thus different mappings produce different graphs. At the second step, ISIDA fragment descriptors[8,9] have been generated for CGRs, followed by their use in model building (Figure 2). Finally, SVM[10] and JRip[11] models distinguishing correct and wrong AAM were obtained and validated.

METHOD

Condensed Graphs of Reaction. The "Condensed Graph of Reaction" (CGR) is a pseudomolecule which results from the superposition of reactant(s) and product(s) molecular graphs into one single graph.[12] The chemical transformations are explicitly taken into account through dynamic bonds. The latter represent covalent bonds being broken, formed, or transformed during the reaction (Figure 1). Atom-to-atom mapping must be performed in order to identify superposed atoms in reactants and products. AAM is a crucial stage of CGR construction: different mappings lead to different graphs (Figure 1).

Data Processing. 850 metabolic reactions from the three first enzymatic classes - EC 1 (oxidoreductases), EC 2 (transferases), and EC 3 (hydrolases) have been retrieved from the KEGG LIGAND Database.[4] All reactions were standardized with ChemAxon Standardizer (release 2009) tool to remove explicit hydrogens and stereochemistry. Stoichiometry of all reactions has been respected. Moreover, a reactant needs to be duplicated in the reaction if it enters into the product twice. Certain reactions have been discarded either by ChemAxon or manually. This concerns the following:

1. Too complex reactions for which ChemAxon Standardizer gives a warning "Reaction is too complex to be automapped". In case of metabolic reactions, it may happen if a sum of atoms in the reagents and products is larger than 130.
2. Reactions having an unbalanced number of atoms or an unbalanced total charge.
3. Reactions involving metals.
4. The duplicates have been discarded.

dx.doi.org/10.1021/ci300418q | J. Chem. Inf. Model. 2012, 52, 3116−3122

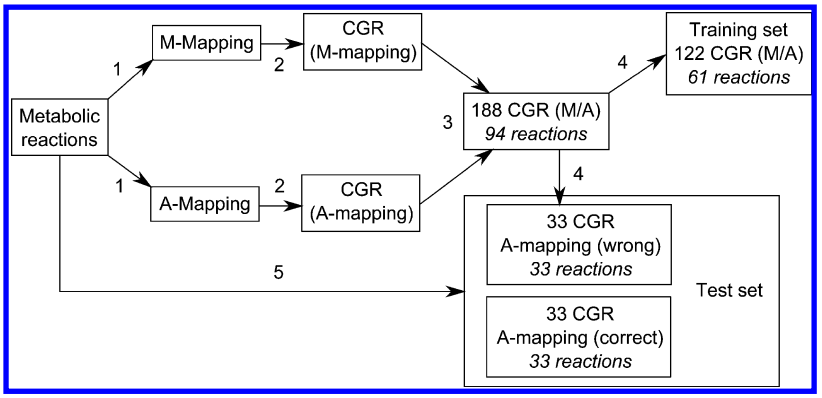


Figure 2. Dataflow. 1. Selected set of metabolic reactions is mapped both manually (M-mapping) or automatically (A-mapping). 2. Two sets of Condensed Graphs of Reactions corresponding to different mapping procedures are generated. 3. Their intersection resulted in a limited number of reactions for which A- and M-mapping differ. The resulting set of CGR (M/A) forms a modeling set which is then split into training and external test sets. 4. 122 CGRs (M/A) are moved to the training set and remaining CGRs of automatic erroneous mapping are added to test set. 5. CGRs of automatic correctly mapped reactions from initial set are generated and added to test set.

5. Reactions involving only displacement of hydrogen atoms. These transformations are not visible on hydrogen suppressed graphs.

6. Reactions containing highly symmetrical reactants or products, because the exact positioning of reaction center could not be deduced by the reaction pairs approach (see below).

7. Reactions in which AAM algorithm assigns to products' atoms numbers which were not assigned to reactants' atoms.

The remaining 630 reactions were both automatically mapped with ChemAxon Standardizer and manually mapped using reactant pairs approach (Figure 2). Manual mapping of metabolic reactions has been performed using the KEGG RPAIR database, which allows users to perform an alignment of reactant pairs. The latter is defined by Kotera et al.[13] " as pairs of compounds that have atoms or atom groups in common on two sides on a reaction". In most cases, the information about reactant pairs is sufficient to perform an unambiguous manual AAM. All manually mapped reactions were considered as correct AAM. If ChemAxon's mapping differed from the manual one, it was considered as erroneous.

All reactions (RD file) have been transformed into CGR using the ISIDA-CGR Designer program and saved as an SD file. Thus, 2 subsets of CGRs corresponding to ChemAxon and manual AAM, respectively, have been generated (Figure 2). These subsets were merged into one SD file in which the duplicates, coming from correct ChemAxon mapping perfectly matching manual AAM, were identified with the EdiSDF program[14] freely available at http://infochim.u-strasbg.fr/.

Modeling Data Sets. Comparison of manual and automated AAM has shown that 94 out of 630 reactions were wrongly mapped by ChemAxon Standardizer, which represents 15% of the initial data set. These reactions and their correctly mapped counterparts have been transformed in 188 Condensed Graphs of Reaction which formed a modeling set. Thus, for each metabolic reaction, 2 CGRs corresponding to correct and erroneous AAM were generated. All reactions forming the modeling set have been split into training and test sets in proportion 2:1. The training set contained 122 CGRs issued from 61 reactions, including 4 reactions of the class EC 1, 29 of the class EC 2, and 28 of the class EC 3. An external test set was composed of 66 CGRs issued from 66 reactions automatically mapped by ChemAxon, namely 33 out of 94 wrongly mapped reactions, and 33 reactions initially correctly mapped. This included 3 reactions of the class EC 1, 32 of the class EC 2, and 31 of the class EC 3. Thus, there is no particular reaction type which represents a problem for ChemAxon AAM: wrongly mapped reactions were detected for all studied classes.

Descriptors. ISIDA fragment descriptors[8,9] were used for the modeling. Each fragment represents a subgraph of a CGR, whereas its occurrence is a descriptor value. Molecules were represented with implicit hydrogen atoms. For each class, four subtypes are defined AB, A, B, and AP. Sequences represent the shortest paths between two selected atoms and may include both atoms and bonds (AB), atoms only (A), and bonds only (B). For the Atoms Pairs (AP) subtype, only terminal atoms and the topological distance between them are represented explicitly. An "extended augmented atom" represents a selected atom with its environment including sequences of AP, AB, A, and B types issued from this atom. Only fragments having at least one dynamical bound were considered. In extended augmented atoms, the branches not containing dynamic bonds were omitted.

For every subclass, the minimal (n_{min}) and maximal (n_{max}) numbers of atoms varied from 2 to 10 for the sequences and from 2 to 6 atoms for the augmented atoms. For any combination of n_{min} and n_{max} all intermediate shortest paths with n atoms ($n_{min} \leq n \leq n_{max}$) are also generated (Figure 3). Varying n_{min} and n_{max}, as well as different subclasses, 240 descriptors' pools were generated with the ISIDA Fragmentor program.[15] Fragmentation types leading to the most performing models have been identified in a cross-validation procedure.

Obtaining and Validation of Models. Two machine learning methods - Support Vector Machine[10] and a propositional rule learner JRip - have been used to build and validate the models distinguishing correct and incorrect AAM considered as two different classes. SVM modeling was performed with the LibSVM[16] software. Tanimoto similarity coefficient was used as kernel. The cost parameter was optimized to achieve the best class separation.

The JRip method realizes the RIPPER algorithm[11] (Repeated Incremental Pruning to Produce Error Reduction) implemented in the Weka software.[17] In RIPPER, the training set is randomly split into growing and pruning sets. An initial rule set is generated on a growing set followed by its further simplifications by applying pruning operation. The pruning stops when simplification yields to an increase of the error on the pruning set. Default parameters from Weka were kept for building models.

Balanced accuracy (BA) was used to assess the performance of the models. BA is calculated as a function of true positive (TP),

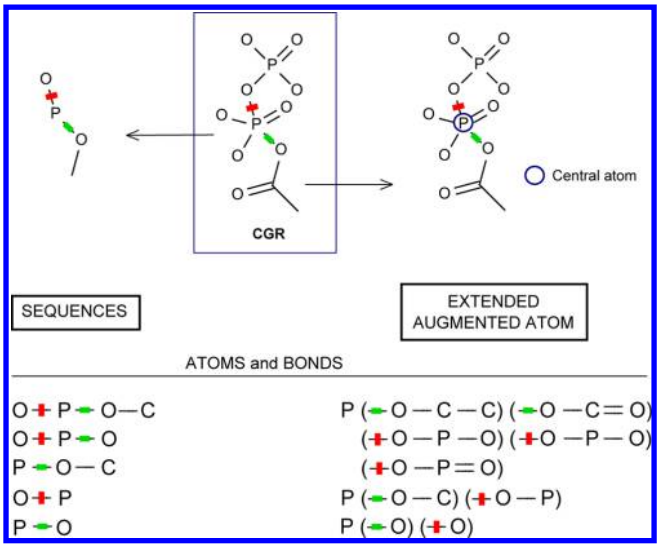


Figure 3. Different types of ISIDA fragment descriptor used in this study: only sequences containing at least one dynamic bond (red if broken, green if created) are considered. This example shows atoms/bonds sequences (subclass AB) of length from 2 to 4 atoms are generated and extended augmented atoms of length from 2 to 3 atoms. Fragments of subclasses A (atoms only) and B (bonds only) could be derived from AB fragments by omitting symbols of atoms or bonds, respectively.[7]

false positive (FP), true negative (TN), and false negative (FN) examples retrieved with the model:

$$BA = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

Selection of Optimal Types of Fragment Descriptors and Model Validation. The goal of this stage was to select among different initial descriptor pools the one providing the highest predictive performance of the models. The workflow is shown in Figure 4. At the first step, few fragmentation types performing the best in 5 × 5-folds cross-validation (5 × 5 CV) have been selected according to BA. They have been tested additionally in 20 × 5 CV. For each calculation, the average value <BA> and the standard deviation Δ(BA) have been assessed. Smaller Δ(BA) corresponds to more stable models. Finally, *Y*-scrambling has been performed in order to check a chance correlation problem. At this step, 5-fold cross-validation has been performed on a data set with the reshuffled labels "correct/erroneous AAM". This procedure has been repeated 20 times. For robust models, BA (scrambling) is expected to be much lower than <BA>.

Selected fragmentations have been used to build the models on the entire training set followed by their validation on the external test set. Applicability domain has not been used because all reactions belong to the three first enzymatic classes.

RESULTS AND DISCUSSION

In SVM calculations, sequences of bonds of length 2 to 3 atoms perform better than other fragments. In 20 × 5 CV, the models built on these descriptors result in <BA> = 88.5% and Δ(BA) = 1.1. In scrambling calculations, BA varied from 38.7 to 58.1% which is significantly smaller compared to <BA>. In JRip modeling, the optimal fragment descriptors were sequences of bonds of length 2 to 8 atoms. In 20 × 5 CV, it performs similarly to SVM: <BA> = 88.7%, Δ(BA) = 3.4, and BA (scrambling) = 40.4 to 56.5%.

Finally, the models for correct and wrong AAM has been obtained on the entire training set using optimal descriptor pools for SVM and JRip. Both machine-learning methods perform well on the external test set achieving BA = 0.94. Notice that SVM retrieves correctly mapped reactions slightly better than JRip, whereas the opposite trend is found for the retrieval of wrongly mapped reactions where JRip model performs slightly better (Table 1).

Although SVM and JRip models have similar predictive performances, the latter approach has some additional benefits.

Table 1. Number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) Examples Retrieved with SVM and JRip Models from the External Test Set[a]

	TP	FP	TN	FN
SVM	31	2	31	2
JRip	30	1	32	3

[a]Here, reactions with correct and erroneous AAM correspond, respectively, to positive and negative examples.

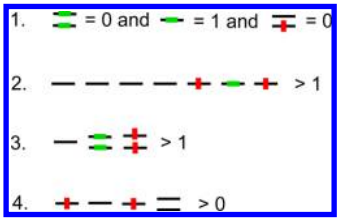


Figure 5. Rules generated by JRip to filter correctly mapped metabolic reactions. In selected structural patterns, only bonds are represented explicitly.

Table 2. True Positives (TP) and False Positives (FP) Retrieved by Particular JRip Rules on the Training and Test Sets

	training set		test set	
rule #	FP	TP	FP	TP
1	1	45	0	28
2	0	5	0	0
3	1	4	1	2
4	0	2	0	0

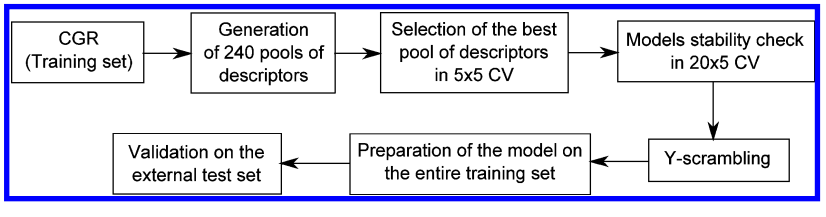


Figure 4. Modeling workflow used in this study.

dx.doi.org/10.1021/ci300418q | J. Chem. Inf. Model. 2012, 52, 3116−3122

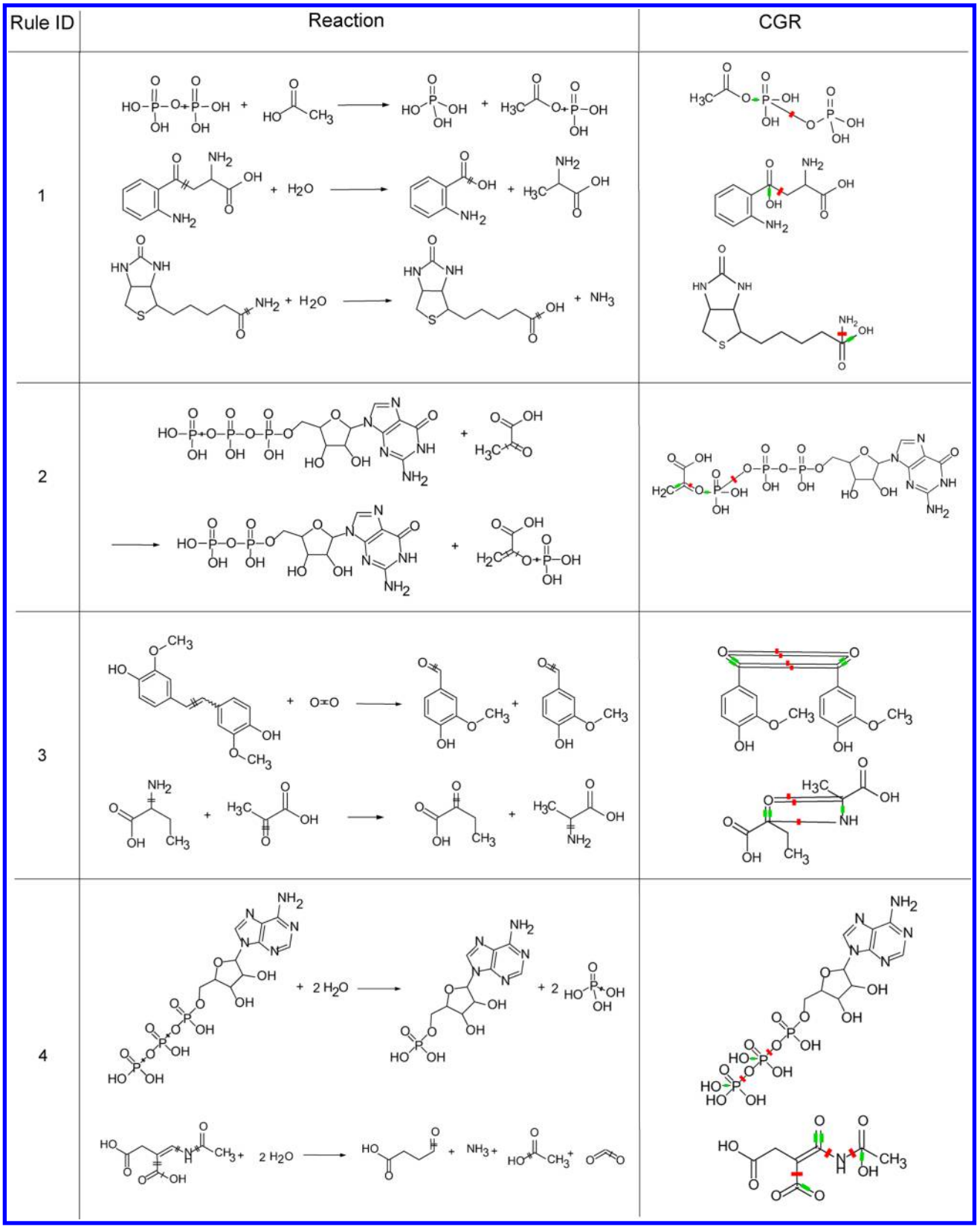


Figure 6. Examples of reactions retrieve by each rule generated by JRip to filter correctly mapped metabolic reactions. Crossed bonds in reactants and products structures represent the ChemAxon annotation of chemical transformations.

Generally, JRip models could be easily interpreted. In this study, it involves four simple rules based on occurrences of particular structural patterns (Figure 5). For example, the first rule states that correctly mapped reaction *must not* involve formation of a double bond AND transformation of double bond to single bond AND *must* involve formation of one single bond. Notice that this rule allows one to retrieve more than 74% and 84% of correctly mapped reactions in the training and test set, respectively (Table 2).

dx.doi.org/10.1021/ci300418q | *J. Chem. Inf. Model.* 2012, 52, 3116−3122

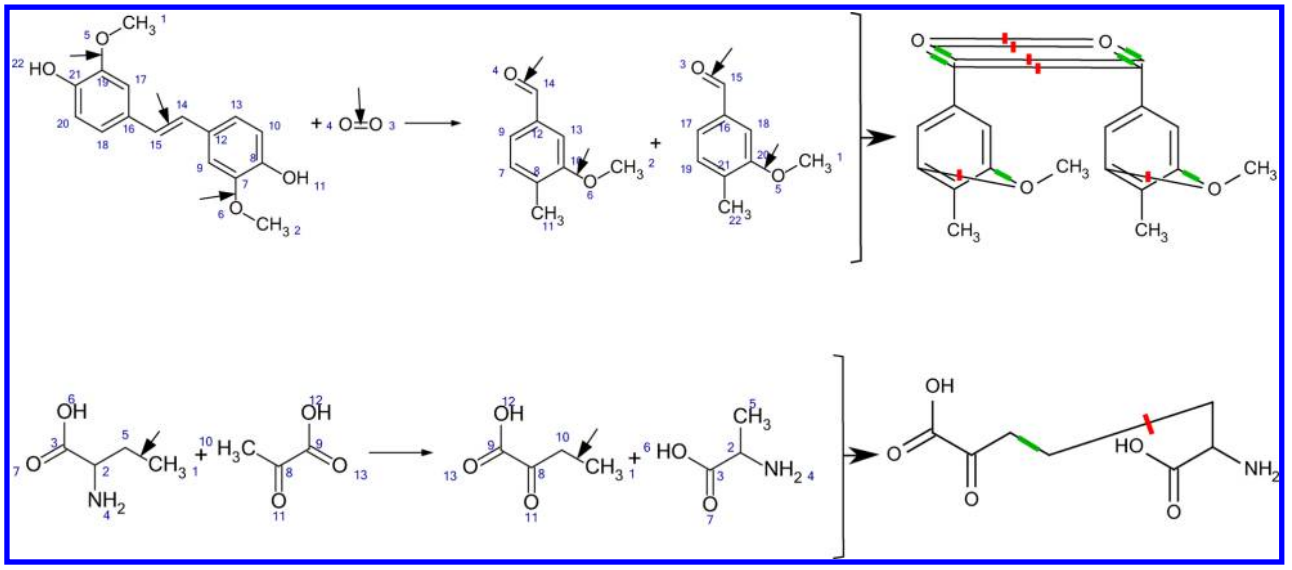


Figure 7. Reactions from training set erroneously classified by JRip model as correctly mapped ones. Black arrows point to the reacting centers.

Examples of reactions retrieve by the JRip rules are given in Figure 6. Rule 1 retrieves both transferase and hydrolase reactions. Rule 2 retrieves exclusively transferase reactions of transfer of phosphate groups using alcohol or a carboxyl groups as acceptor. Rule 3 concerns oxidoreductases and transferases reactions. Finally, the last rule retrieves only hydrolases reactions.

Another advantage of JRip models concerns the possibility of completing them by manually designed rules. In Table 2, two False Positives in the training set correspond to wrongly mapped reactions given in Figure 7. There are very few similar reactions in the training set, and therefore the model has not captured their structural patterns. Visual inspection of CGRs corresponding to the above reactions suggests two additional rules C − CH_2 + C > 0 and C − O + C > 0 which improve the model's performance on the external test sets (BA = 0.95).

CONCLUSION

Automated atom-to atom mapping of chemical reactions represents a difficult task because in some cases the computer algorithm based on the graph theory is not sufficient and a reaction mechanism must be taken into account. Nowadays, there exists no automated algorithm perfectly performing AAM. In this paper, we demonstrated that statistical models built on fragment descriptors issued from Condensed Graphs of Reactions represent an efficient way to detect erroneous AAM made by automated algorithms. Although, any machine-learning methods can be used for the modeling, those deriving associative rules (e.g., JRip) have some preference because automatically derived rules can be completed by manually derived custom ones.

ASSOCIATED CONTENT

Supporting Information

630 enzymatic reactions used in the modeling. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: varnek@chimie.u-strasbg.fr.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Jochum, C.; Gasteiger, J.; Ugi, I. The principle of minimal chemical distance. *Angew. Chem., Int. Ed. Engl.* **1980**, *19*, 495−505.

(2) Korner, R.; Apostolakis, J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J. Chem. Inf. Model.* **2008**, *48* (6), 1181−1189.

(3) *AutoMapper*, version 5.1.1; ChemAxon: Budapest, Hungary, 2009.

(4) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40* (D1), D109−D114.

(5) 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.

(6) *Standardizer*, version 5.1.1; ChemAxon: Budapest, Hungary, 2009.

(7) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19* (9−10), 693−703.

(8) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I.; Marcou, G. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008**, *4* (3), 191−198.

(9) Horvath, D.; Bonachera, F.; Solovev, V.; Gaudin, C.; Varnek, A. Stochastic versus stepwise strategies for quantitative structure-activity relationship generation − how much effort may the mining for successful QSAR models take? *J. Chem. Inf. Model.* **2007**, *47* (3), 927−939.

(10) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cristianini, N., Shawe-Taylor, J., Eds.; Cambridge University Press: Cambridge, United Kingdom, 2000.

(11) Cohen, W. W. Fast Effective Rule Induction. In *Machine learning: proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, July 9−12, 1995*; Prieditis, A., Eds.; The Morgan Kaufmann series in machine learning; Morgan Kaufmann Publishers: Burlington, 1995; pp 115−123.

(12) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools* **2011**, *20* (2), 253−270.

(13) Kotera, M.; Okuno, Y.; Hattori, M.; Goto, S.; Kanehisa, M. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.* **2004**, *126* (50), 16487−16498.

(14) de Luca, A.; Horvath, D.; Marcou, G.; Solov'ev, V.; Varnek, A. Mining chemical reactions using neighborhood behavior and condensed

graphs of reactions approaches. *J. Chem. Inf. Model.* **2012**, *52* (9), 2325−2338.
(15) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA property-labelled fragment descriptors. *Mol. Inf.* **2010**, *29* (12), 855−868.
(16) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intelligent Systems Technol.* **2011**, *2* (3), 27:1−27:27.
(17) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: an update. *SIGKDD Explorations* **2009**, *11* (1), 10−18.