Condensed Graph of Reaction: considering a chemical reaction as one single pseudo molecule

Frank Hoonakker[1,3], Nicolas Lachiche[2], Alexandre Varnek[3], and Alain Wagner[3,4]

[1] Chemoinformatics laboratory, University of Strasbourg, France
[2] LSIIT, University of Strasbourg, France
[3] eNovalys, Illkirch, France
[4] Functional ChemoSystems, University of Strasbourg, France

Abstract. Chemical reactions always involve several molecules of two types reactants and products. On the other hand, Quantitative Structure Activity Relationship (QSAR) methods are used to predict physic-chemical or biological properties of individual molecules. In this article, we propose to use Condensed Graph of Reaction (CGR) approach merging all molecules involved in a reaction into one molecular graph. This allows one to consider reactions as pseudo-molecules and to develop QSAR models based on fragment descriptors. Here Substructure Molecular Fragment descriptors calculated from CGRs have been used to build quantitative models for the rate constant of S_N2 reactions in water. Three common attribute-value regression algorithms (linear regression, support vector machine, and regression trees) have been evaluated.

1 Introduction

Quantitative Structure Activity Relationship (QSAR) consists in predicting some chemical property given the structure of the molecule. It is an important research area in chemistry, and a very challenging application domain for data mining. QSAR typically deals with a single molecule. Chemical reactions usually involve several molecules. As it is possible to predict properties of molecules, the same should be possible with reactions. The problem is to plug several molecules, reactants and products, in a data mining algorithm.

This article points out the use of a Condensed Graph of Reaction (CGR) to represent a reaction involving several molecules as if it was a single molecule, therefore allowing the use of existing techniques dealing with a single molecule. This is illustrated on a real chemical problem.

Chemistry, in particular QSAR, is a main application domain of machine learning and data mining. Inductive Logic Programming and Relational Data Mining can represent and learn from complex structures such as molecules. Moreover they can use background knowledge such as rings, generic atoms[1–4]. However to the best of our knowledge they have not been applied to chemical reactions.

Presented at the 19th International Conference on Inductive Logic Programming (ILP'09).

Some papers related to data mining methods predicting properties of reactions have been published, but they do not really model the reaction. For instance Brauer [5] and Katriski [6] have published papers dealing with Quantitative Structure Reactivity Relationship concerning only one reaction making some parameter (such as solvent) vary. Another attempt has been proposed by Halberstam [7] to model the rate constant of reaction involving two reactants and one product (A +B → C) where the second reactant (B) is always the same. The study was then reduced to a classical QSAR on one compound.

This paper is organised as follows. The condensed graphs of reactions are defined in section 2. Substructure Molecular Fragments are presented in section 3. The prediction of the rate constant of reaction is described in section 4. Section 5 concludes.

2 Condensed Graphs of Reactions

A Condensed Graph of Reaction [8] represents a superposition of reactants and products graphs. A CGR is a complete connected and non oriented graph in which each node represents an atom and each edge a bond. CGR uses both conventional bonds (single, double, aromatic, etc.) which are not transformed in the course of reaction, and dynamical bonds corresponding to those created, broken or modified during the reaction (cf. figure 1).

Actually a CGR is a pseudo molecule in which some new bond types have been added. An editor of CGR has been added to our software environment specialised in chemical data mining: ISIDA (In SIlico Design and Analysis) [9]. Any new type of dynamical bond could be easily added in the list of bond types.

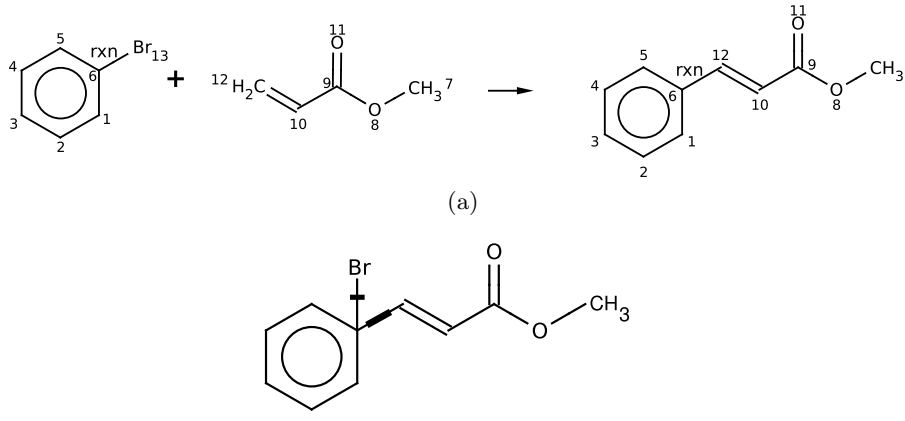

Fig. 1. A reaction (a) and the corresponding Condensed Graph of Reaction (b)

Moreover we developed an algorithm that generates a CGR from the file formats (RXN and RD) usual in chemoinformatics to model reactions. This programs requires the information about the atom mapping in reactants and products. This information is often available from existing software to edit and manage chemical data, otherwise it can be added thanks to our editor. Indeed the key point to produce a CGR is to map the reaction, that means that each atom on the left of the arrow corresponds to an atom on the right of the arrow. On figure 1.(a) each atom is uniquely numbered in order to assign the same number to the same atom on both sides of the arrow, for instance the atom numbered 12 on figure 1.(a). Moreover, some flags are added to describe the bonds that change, as described in the "CTFile format" document [10] from Elsevier MDL©. In the case of our example reaction a "*rxn*" flag is drawn beside the created bond between the atoms mapped 6 and 12 and for the broken bond between atoms 6 and 13.

In most of the database, the mapping is automatically done, with some errors due to mismatching of the atoms on each side of the arrow. For our dataset, the mapping was manually done and verified by a chemist to guarantee avoiding mismatch. Once the reactions are correctly mapped, the CGR are created. The algorithm consists in gathering the atoms of all the compounds of the reaction without duplication of the mapped atom. Then the connection table of reactants and products are examined to find the reactivity flag and write the dynamical bond in the CGR. Figure 1.(b) shows the CGR corresponding to the reaction above. Let us emphasize that the bond types assigned between the carbon 6 and the brome 13 denotes a broken single bond and the bond type between carbons 6 and 12 denotes the creation of a single bond.

This change of representation allows one to store the reaction database in the format (SD) usual in chemoinformatics to represent individual molecules.

3 Substructure Molecular Fragments

For each compound, the substructural molecular fragments (SMF) [11, 12] produce a vector of integers counting the occurrences of molecular fragments. The nature of each descriptor is a molecular fragment, as detailed below, and its value is the count of this fragment in a molecule.

In this paper, fragments were constructed by computing the shortest paths in the molecular graph between two atoms -in terms of the number of nodes passed through. The fragment is a representation of the Atoms and Bonds (AB) traversed by this path. The SMF descriptors apply to the CGR. A fragment is the shortest path between two atoms. For the reactions only fragments containing at least one dynamical bond are selected. Some example of fragments in their linear notation are shown in figure 2. The first example (Cl⁄–S-C*C*C*C) represent the shortest path between the two marked atoms (length = 6). If several shortest paths could be found, all of them are take into account. Symmetric fragments, for example the C-C-N and N-C-C, are considered as a single descriptor.

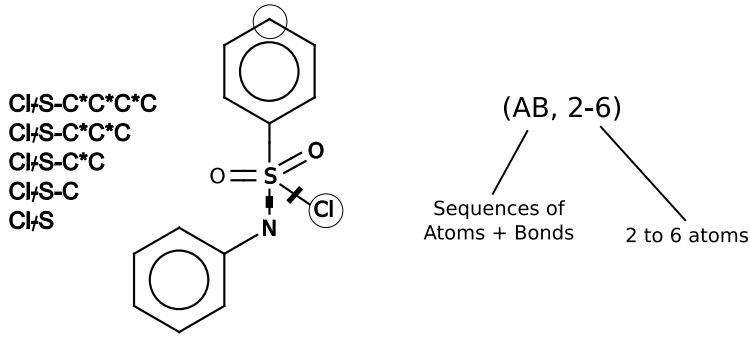

Fig. 2. Example of Substructure Molecular Fragments applied to a Condensed Reaction Graph

The fragmentation takes a minimum and a maximum length as parameters, for instance (AB,2-6) means all fragments having from 2 to 6 atoms.

This kind of fragmentation produces a large number of descriptors some of which are dependent, eg. Cl/−S and Cl/−S-C on figure 2. Consequently the correlated descriptors are eliminated, using the Pearson's correlation coefficient (R) of each pair of descriptor and keeping only one of them (the first one). The threshold used to determine if two descriptors are correlated is $R > 0.99$. Table 1 shows the number of fragments before and after attribute selection.

Fragmentation	Total number	Selected attributes
(AB,2-6)	2066	448
(AB,2-8)	2784	622
(AB,2-10)	4458	698

Table 1. Numbers of attributes before and after removing correlated attributes, for three fragmentation lengths

4 Prediction of the rate constant of reaction

The data used for the computation of the rate constant comes from a compilation [13] of the rate and equilibrium constants of heterolytic organic reactions. The selected reactions concern Nucleophile Substitution 2 (S_N2) water. The database[5] was manually built and contains 1014 instances described by: (i) the reaction, (ii) the temperature represented as 1/T with T in Kelvin, (iii) the $log(k)$ where k is the rate constant.

[5] Available on demand at Laboratoire d'Infochimie, 4 rue Blaise Pascal 67000 Strasbourg France. varnek-at-infochim.u-strasbg.fr

Three methods were used to model our data : (*i*) M5P (model tree), (*ii*) SVMreg (an SVM method for regression problems) using a RBF kernel and (*iii*) linear regression (LR), from *WEKA* [14], all with their default parameters.

The average of the Root Mean Squared Error (RMSE) and of the correlation coefficient, using ten times a ten-fold cross-validation, for the three methods on three fragmentations are reported in table 2.

	(AB, 2-6)		(AB, 2-8)		(AB, 2-10)	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
SVMreg	0.52	1.27	0.53	1.26	0.53	1.26
M5P	0.47	1.33	0.5	1.30	0.47	1.33
LR	0.34	1.49	0.37	1.45	0.32	1.51

Table 2. Values of R^2 and $RMSE$ for the three methods (SVMreg, M5P, LR) associated with three fragmentation lengths (I(AB, 2-6), I(AB, 2-8), I(AB, 2-10)).

Finally CGR and SMF made possible to model the rate constant of reaction. The best statistical results are measured for SVM ($R^2 = 0.53$), but according to the REC curves (not reported in this extended abstract) all the methods are close even if R^2 for linear regression is lower than the other. It is not surprising, because the data are not linear and the other methods (SVM and M5P) are more able to deal with non-linear problems. Tuning the fragment length has little influence on accuracy of the prediction.

5 Conclusion

Condensed Graphs of Reactions enable any existing QSAR technique to be applied to chemical reactions. This approach has been successfully experimented on a real chemical problem, using Substructure Molecular Fragments to generate an attribute-value representation of the chemical reactions, and out-of-the-box regression techniques from Weka.

References

1. Dzeroski, S.: Relational Data Mining Applications: An Overview. In: Relational Data Mining. Springer (2001)
2. Kramer, S., Frank, E., Helma, C.: Fragment generation and support vector machines for inducing sars. SAR and QSAR in Environmental Research **13**(5) (2002) 509–523
3. Helma, C., Cramer, T., Kramer, S., Raedt, L.D.: Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationship of noncongeneric compounds. J. Chem. Inf. Comput. Sci. **44** (2004) 1402–1411

4. Cannon, E.O., Amini, A., Bender, A., Sternberg, M.J.E., Muggleton, S.H., Glen, R.C., Mitchell, J.B.O.: Support vector inductive logic programming outperforms the naive bayes classifier and inductive logic programming for the classification of bioactive compounds. J. Comput. Aided Mol. Des. **21** (2007) 269–280
5. M. Brauer, J. L. Péres-Lustres, J. Weston, E. Anders: Quantitative Reactivity model for the hydratation of carbon dioxide by Biometric Zinc Complexes. Inorg. Chem. **41** (2002) 1454–1463
6. A. R. Katritzky, S. Perumal, R. Petrukhin: A QSRR Treatment of Solvent Effects on the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylates Employing Molecular Descriptors. J. Org. Chem. **66**(11) (2001) 4036 –4040
7. N. M. Halberstam, I. I. Baskin, V. A. Palyulin, N. S. Zefirov: Neural networks as a method for elucidating structure-property relationships for organic compounds. Russ. Chem. Rev. **72**(7) (2003) 629–649
8. S. Fujita: Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. J. Chem. Inf. Comput. Sci. **26**(4) (1986) 205
9. A. Varnek: ISIDA software - http://infochim.u-strasbg.fr/recherche/isida/index.php
10. Elsevier MDL: CTfile Format - http://www.mdli.com/downloads/public/ctfile/ctfile.jsp (2007)
11. A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev: Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. J. Comput. Aided. Mol. Des. **19**(9-10) (2005) 693–703
12. V. P. Solov'ev, A. Varnek, G. Wipff: Modeling of ion complexation and extraction using substructural molecular fragments. J. Chem. Inf. Comput. Sci. **40**(3) (2000) 847–58
13. Laboratory of chemical kinetics and catalysis. Tartu State University: Table of rate and equilibrium constants of heterolytic organic reactions. (1977)
14. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Second edition edn. Morgan Kaufmann (2005)