Structure–Reactivity Relationships in Terms of the Condensed Graphs of Reactions

T. I. Madzhidov^a, P. G. Polishchuk^b, R. I. Nugmanov^a, A. V. Bodrov^a, A. I. Lin^a, I. I. Baskin,^{c,d}
A. A. Varnek^d, and I. S. Antipin^a

^a Kazan (Volga Region) Federal University, ul. Kremlevskaya 18, Kazan, 420008 Tatarstan, Russia
e-mail: Timur.Madzhidov@kpfu.ru

^b Bogatskii Physicochemical Institute, National Academy of Sciences of Ukraine,
Lyustdorfskaya doroga 86, Odessa, 65080 Ukraine

^c Faculty of Chemistry, Moscow State University, Vorob'evy gory 1, Moscow, 119991 Russia

^d Chemoinformatics Laboratory, University of Strasbourg,
rue B. Pascal 4, CS 90032 Strasbourg, 67000 France

Received February 3, 2014

Abstract—An approach for the prediction of rate constants of chemical reactions, based on the representation of a chemical reaction as a condensed graph, has been tested on more than 1000 bimolecular nucleophilic substitution reactions with neutral nucleophiles in 38 solvents. Molecular fragment descriptors, temperature, and solvent parameters characterizing solvation power have been used in the reaction modeling. The obtained models ensure a good correlation between the predicted and experimental values; the corresponding deviations are comparable with interlaboratory measurement errors.

DOI: 10.1134/S1070428014040010

Quantitative structure–property (activity) relationship (QSPR, QSAR) models establish a mathematical relation between a physicochemical or biological characteristic and molecular structure represented as a set of parameters called *molecular descriptors* [1, 2]. This approach is widely used in the modeling and prediction of various properties of pure compounds [3]. Its application to chemical reactions is complicated by the fact that each reaction involves several individual molecules of at least two types, reagents and products. Therefore, even for a relatively simple reaction like $A + B \rightarrow C + D$, selection of one of the four molecules, for which descriptors are to be calculated, becomes artificial and is thus seriously restricted. For that reason, QSPR studies were performed either for a reaction series where one reagent was the same (if reagent **B** is constant, descriptors may be calculated only for **A** [4, 5]) or descriptors were calculated for all molecules involved in a reaction with their necessary assignment to a definite type (**A**, **B**, **C**, or **D**), e.g., for nucleophilic reactions, nucleophile, nucleofuge, target product, and by-product [6], which makes the modeling procedure essentially sophisticated. We now pro-

pose a new approach to chemical reaction modeling, which is based on the representation of a chemical reaction as a condensed graph (CGR, Condensed Graph of Reaction).

This approach originates from Vlédutz's paper [7] where it was proposed to represent a chemical reaction via a single graph whose edges denote formation, rupture, and change of bond order instead of an ensemble of molecular graphs mathematically coding structure of molecules (graph vertices correspond to atoms, and edges, to chemical bonds). Fujita [8] extended this approach so as to represent a chemical transformation as a whole (rather than a single reaction center) by a single graph originally called *imaginary transition state*, and later, *condensed graph of reaction* [9]. It was proposed [10, 11] to consider condensed graph of reactions to be a pseudomolecule for which molecular descriptors could be generated, so that standard chemoinformatics methods widely used in the analysis of data for individual molecules could be applied to chemical reactions. As a result, it became possible to classify chemical reactions [10], analyze algorithms of atom–atom mapping of reagents and products [12], and

Fig. 1. Example of condensed graph (bottom) corresponding to nucleophilic substitution reaction with the indicated atom–atom mapping (top). Light circle denotes the newly formed single C–N bond, and the single C–Br bond being broken is struck through.

model structure–reactivity relationships [13]. Condensed graph of reaction includes dynamic bonds that are formed or broken during a reaction, as well as those bonds which remain unchanged (Fig. 1).

Condensed graph of a reaction is generated in two steps. The first step implies atom–atom mapping, i.e., a correspondence between atoms in the reagent and product molecules is established (Fig. 1), and the second step is superposition of atoms with the same number in the reagents and products, which gives a condensed graph. Dynamic bonds can thus be readily identified.

The goal of the present work was to verify the applicability of the CGR approach as a way of structural representation of reactions for modeling rate constants using bimolecular nucleophilic reactions as an example. The main problem was to build up QSAR models for the prediction of rate constants of S_N2 reaction involving a wide variety of nucleophiles, nucleofuges, and products.

The reaction types, conditions, and $\log k$ values (k is the rate constant) were taken from [14]. The selected reactions may be divided into two main groups where the nucleophile is either neutral or negatively charged. In the latter case, the rate constants may depend on the nucleophile concentration due to complexity of processes occurring in solution (dissociation, ion association, aggregation, etc.); therefore, for the present study we selected only reactions involving attack by neutral nucleophile as shown in the following general scheme:

Here, X is a leaving group (most frequently, a halogen atom), U is the central carbon atom in the nucleophile (N, P), and R^1 – R^6 are substituents. Total of 1041 reaction rate constants in 38 solvents were collected.

As descriptors we used ISIDA fragment descriptors [10] and SiRMS simplex descriptors [15]. The values of both these are equal to the occurrence of a given fragment in a molecule. ISIDA descriptors count fragments that are chains of a given length with indication of the atom and bond types. Only most important fragments, namely chains containing at least one dynamic bond, were used to simulate chemical reactions encoded as condensed graphs (Fig. 2). SiRMS descriptors (Fig. 2) count all possible four-atom units, even those in which atoms are not linked directly to each other by covalent bonds, with account taken of the atom and bond types. Insofar as $\log k$ also depends on the temperature and solvent nature, the temperature and 15 solvent characteristics were also included into the set of descriptors. In particular, polarity, polarizability, and proton-acceptor and proton-donor abilities of solvent were considered (see Experimental).

Fig. 2. Examples of ISIDA and SiRMS descriptors; some fragments used for the generation of descriptors are shown.

Relationships between the reaction parameters and rate constants were found using the Random Forest algorithm [16] implying a set of so-called decision trees. A random subset of initial objects, generated by the bootstrap technique, was used as training sample for each tree. Objects not included into the training sample, were placed into so-called *out-of-bag* sample for that tree, which can be used to estimate the stability and predicting ability of the Random Forest model.

The model performance was assessed by the determination coefficient Q^2 (squared correlation coefficient) between the predicted and experimental values, as well as by the root mean square errors (RMSE). It should be emphasized that the predicted rate constants were obtained on the out-of-bag sample, i.e., with the use of only those decision trees which were built up without considering a given reaction. Therefore, Q^2 and RMSE values characterize the quality of predictions made by the given model rather than its descriptive ability (goodness of fit).

The best models are obtained by simulation of rate constants of chemical reactions with joint consideration of descriptors characterizing condensed graph of reaction, temperature, and solvent. In this case, fairly high correlation between the predicted and experimental rate constants is observed ($Q^2 = 0.67$, RMSE = 0.51 for ISIDA descriptors and $Q^2 = 0.70$, RMSE = 0.50 for SiRMS descriptors). Despite different principles of selection of molecular fragments, the use of the ISIDA and SiRMS descriptors gives very similar results. As follows from Fig. 3, the calculated $\log k$ values ob-

tained using the ISIDA and SiRMS descriptors on the basis of the out-of-bag sample show a fairly good agreement with the experimental values.

The RMSE value (~ 0.5 in both models) is quite close to the experimental error in the determination of rate constants. Although the measurement errors reported in the literature usually do not exceed 5% (which corresponds to no more than 0.35 log unit), the reproducibility of data under different conditions is somewhat lower, and it reaches 0.4–0.5 or (in some rather rare cases) may exceed this level. Although it is fairly difficult to compare different experimental methods used for the determination of rate constants of the reactions under study, we have revealed some strong deviations which are illustrated by the scheme shown below. Here, *a* is the reaction for which discrepancies in the experimental rate constants are related to different methods of their determination; *b* and *c* are the reactions for which rate constants were measured by the same method but by different authors; and *d* is a rare case when the experimental $\log k$ value depended on the initial reactant concentration (as a rule, no such differences are observed for reactions with neutral nucleophiles). Introduction of the reagent concentration as an additional descriptor leads to some reduction of Q^2 and increase of RMSE, which is likely to indicate insignificant effect of that parameter on the reaction rate.

Thus, following the CGR approach and using fragment descriptors based thereon, we have constructed models for prediction of the reaction rate constants

Fig. 3. Predicted versus observed plots for the rate constants ($\log k$) predicted on the basis of out-of-bag samples using (a) ISIDA and (b) SiRMS descriptors.

MeOH, 30°C, $\log k = -3.25, -3.60$

MeOH, 40°C, $\log k = -2.94, -3.34$

MeOH, 25°C, $\log k = -3.69, -4.02$

MeOH, 35°C, $\log k = -2.92, -3.46$

PhNO₂, 25°C, $\log k = -4.08, -4.157, -4.16, -4.18$

C₆H₁₂, 25°C, $\log k = -4.16, -4.87, -5.62, -5.95, -6.04, -6.48, -6.88$ (at different reactant concentrations)

from the corresponding condensed graphs and reaction conditions. The prediction error of the proposed models is comparable with the experimental error. The accuracy in the prediction of rate constants of S_N2 reactions in terms of the CGR approach is almost the same as that reported in [17] ($Q^2 = 0.79$, RMSE = 0.58 for test set), but our approach is more advantageous due to simplicity of the calculation of the required descriptors (their calculation is faster by several orders of magnitude than the calculation of quantum chemical descriptors used in [17]) and their mechanistic interpretation. The CGR approach may be useful in the modeling of non-uniform reaction series where the number of reagents or products is not constant.

EXPERIMENTAL

ChemAxon InstantJChem [18] was used as database management system for storage reaction data. The structures of chemical compounds participating in the reactions were standardized (aromatization, standardization of the structure of substituents such as nitro group, nitroso group, etc.) by Standardizer tool of JChem [19]. The same tool was also used for atom-atom mapping. The errors in the atom-atom mapping were revealed using in-house ReactMap program

which determined the number of formed and broken bonds at a given mapping. This number for S_N2 reactions (neglecting hydrogen atoms) should be equal to 2; larger numbers indicate mapping error. The errors were corrected either automatically (on the basis of the principles given in [20]) or manually (if automatic correction was impossible). The atom-atom mapping thus obtained was used to generate condensed graph of reaction with the aid of in-house Condenser program. Condensed graphs of reactions were stored in a modified SDF format described in the Fragmenter manual [21].

ISIDA descriptors for condensed graphs were generated by Fragmenter [10] for chains with a length of 2 to 10 atoms with labeled atom and bond types; all possible chains containing at least one dynamic bond were counted (DoAllWays option). Total of 490 descriptors were generated. SiRMS descriptors (981 in total) were generated using SiRMS program [22]. As solvent descriptors we used Catalan SPP [23], SA [24], and SB constants [25], Kamlet-Taft constants α [26], β [27], and π^* [28], and the following parameters characterizing solvent polarity and polarizability:

(1) functions depending on the dielectric constant ϵ : Born function $f_B = (\epsilon - 1)/\epsilon$, Kirkwood function $f_K = (\epsilon - 1)/(2\epsilon + 1)$, and functions $f_1 = (\epsilon - 1)/(\epsilon + 1)$ and $f_2 = (\epsilon - 1)/(\epsilon + 2)$;

(2) functions depending on the refractive index n_D^{20} (denoted as n for the sake of simplicity): $g_1 = (n^2 - 1)/(n^2 + 1)$, $g_2 = (n^2 - 1)/(2n^2 + 1)$, and $h = (n^2 - 1)(\epsilon - 1)(n^2 + 1)^{-1}(2\epsilon + 1)^{-1}$.

No preliminary selection of descriptors was performed. All in-house program code was written in Python.

Random Forest models were built up with the aid of RandomForest package [29] of R program for statistical calculations [30] using default parameters, and each model included 500 trees. The prediction quality was assessed on the out-of-bag sample. The predicted $\log k$ values were compared with those found experimentally using the following statistical parameters:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (x_i^{\text{pred}} - x_i^{\text{exp}})^2}{N}},$$

$$Q^2 = 1 - \frac{\sum_{i=1}^N (x_i^{\text{pred}} - \bar{x}_i^{\text{exp}})^2}{\sum_{i=1}^N (x_i^{\text{exp}} - \bar{x}_i^{\text{exp}})^2}.$$

where x_i^{pred} and x_i^{exp} are, respectively, the predicted and experimental $\log k$ values for i th reaction, $\bar{x}_i^{\text{exp}}$ is the average experimental rate constant, and N is the number of elements in the sample.

This work was funded by the subsidy of the Russian Government to support the Program of Competitive Growth of Kazan Federal University among World's Leading Academic Centers.

REFERENCES

1. *Recent Advances in QSAR Studies: Methods and Applications*, Puzyn, T., Leszczynski, J., and Cronin, M.T.D., Eds., Dordrecht: Springer, 2010.
2. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, Dehmer, M., Varmuza, K., and Bonchev, D., Eds., Weinheim: Wiley–Blackwell, 2012.
3. Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*, Weinheim: VCH, 1993.
4. Halberstam, N.M., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Mendeleev Commun.*, 2002, vol. 12, p. 185.
5. Zhokhova, N.I., Baskin, I.I., Palyulin, V.A., Zefirov, A.N., and Zefirov, N.S., *Dokl. Chem.*, 2007, vol. 417, p. 282.
6. Kravtsov, A.A., Karpov, P.V., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Dokl. Chem.*, 2011, vol. 441, p. 314.
7. Vlédutz, G.É., *Inform. Storage Retr.*, 1963, vol. 1, p. 117.
8. Fujita, S., *J. Chem. Inf. Comput. Sci.*, 1986, vol. 26, p. 205.
9. Jauffret, P., Tonnelier, C., and Kaufmann, G., *New J. Chem.*, 1990, vol. 14, p. 945.
10. Varnek, A., Fourches, D., Hoonakker, F., and Solov'ev, V.P., *J. Comput.-Aided Mol. Des.*, 2005, vol. 19, p. 693.
11. Varnek, A., *Chemoinformatics and Computational Chemical Biology*, Bajorath, J., Ed., New York: Humana, 2010.
12. Muller, C., Marcou, G., Horvath, D., Aires-de-Sousa, J., and Varnek, A., *J. Chem. Inf. Model.*, 2012, vol. 52, p. 3116.
13. Hoonakker, F., Lachiche, N., Varnek, A., and Wagner, A., *Int. J. Artif. Intell. Tools*, 2011, vol. 20, p. 253.
14. *Tablitsy konstant skorosti i ravnovesiya geteroliticheskikh organicheskikh reaktsii* (Tabulated Rate and Equilibrium Constants of Heterolytic Organic Reactions), Pal'm, V.A., Ed., Moscow: VINITI, 1977, vol. 2.
15. Kuz'min, V.E., Artemenko, A.G., Polischuk, P.G., Muratov, E.N., Khromov, A.I., Liahovskiy, A.V., Andronati, S.A., and Makan, S.Y., *J. Mol. Model.*, 2005, vol. 11, p. 457.
16. Breiman, L., *Machine Learning*, 2001, vol. 45, p. 5.
17. Kravtsov, A.A., Karpov, P.V., Baskin, I.I., Palyulin, V.A., and Zefirov, N.S., *Dokl. Chem.*, 2011, vol. 440, p. 299.
18. *InstantJChem 6.0.6*, 2013, ChemAxon. <http://www.chemaxon.com>
19. *JChem 6.0.6*, 2013, ChemAxon. <http://www.chemaxon.com>
20. Fontain, E., *Anal. Chim. Acta*, 1992, vol. 265, p. 227.
21. Marcou, G., Solov'ev, V., Horvath, D., and Varnek, A., *ISIDA Fragmentor2011. User Manual. 2012*. http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/Fragmentor2011_Manual.pdf
22. Polishchuk, P., *SIRMS. 2013*, <http://github.com/DrrDom/sirms>
23. Catalán, J., López, V., Pérez, P., Martín-Villamil, R., and Rodríguez, J.G., *Justus Liebigs Ann. Chem.*, 1995, p. 241.
24. Catalán, J. and Díaz, C., *Justus Liebigs Ann. Chem.*, 1997, p. 1941.
25. Catalán, J., Díaz, C., López, V., Pérez, P., de Paz, J.L.G., and Rodríguez, J.G., *Justus Liebigs Ann. Chem.*, 1996, p. 1785.
26. Taft, R.W. and Kamlet, M.J., *J. Am. Chem. Soc.*, 1976, vol. 98, p. 2886.
27. Kamlet, M.J. and Taft, R.W., *J. Am. Chem. Soc.*, 1976, vol. 98, p. 377.
28. Kamlet, M.J., Abboud, J.L., and Taft, R.W., *J. Am. Chem. Soc.*, 1977, vol. 99, p. 6027.
29. Liaw, A. and Wiener, M., *R News*, 2002, vol. 2, p. 18.
30. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012.