УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ЕСТЕСТВЕННЫЕ НАУКИ

2018, Т. 160, кн. 1
С. 32–39

ISSN 2542-064X (Print)
ISSN 2500-218X (Online)

УДК 547.022+544.412.3

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ ПРОПУЩЕННЫХ РЕАГЕНТОВ И ПРОДУКТОВ В УРАВНЕНИЯХ ХИМИЧЕСКИХ РЕАКЦИЙ

Р.И. Нугманов[1], *Т.И. Маджидов*[1], *И.С. Антипин*[1], *А.А. Варнек*[1,2]

[1]*Казанский* (*Приволжский*) *федеральный университет*, *г. Казань*, *420008*, *Россия*
[2]*Университет Страсбурга*, *г. Страсбург*, *67084*, *Франция*

Аннотация

Одной из проблем при работе с базами данных по химическим реакциям является то, что в них большинство реакций является стехиометрически неуравновешенным, то есть в них отсутствует информация об одном или нескольких реагентах либо продуктах. Неуравновешенность приводит к тому, что некоторые реакции не могут быть найдены в ходе структурных поисков. Было предложено решение этой проблемы, основанное на использовании подхода конденсированного графа реакции. Конденсированный граф реакции является сжатым представлением реакционного превращения, в котором переходящие друг в друга атомы реагента и продукта совмещены. В результате в полученном псевдомолекулярном графе идентифицируются обычные химические связи, а также так называемые динамические связи, соответствующие изменению порядка связи. Ключевое наблюдение, использованное в этом сообщении, заключается в том, что хотя конденсированный граф содержит всю известную структурную информацию о химической реакции, он может быть построен и для незаполненной химической реакции. В ходе обратного преобразования конденсированного графа в химическую реакцию восстанавливается информация о пропущенных реагентах и продуктах. Имеются некоторые ограничения этого подхода, связанные с тем, что предложенным способом не может быть восстановлена информация о пропущенных коэффициентах в реакции. Для решения такой проблемы предложен алгоритм, основанный на использовании изоморфного вложения. Такой алгоритм обнаруживает схожие мотивы в конденсированном графе и при необходимости добавляет коэффициент перед реагентом и/или продуктом либо копирует их в уравнении реакции необходимое количество раз.

Ключевые слова: хемоинформатика, химические реакции, стехиометрическая балансировка реакций, конденсированный граф реакции

Введение

В современном мире поиск информации является неотъемлемой частью эффективной работы во многих сферах деятельности и, в особенности, в науке и наукоемком производстве (химической, фармацевтической промышленности). Поиск информации в базах данных соединений и реакций стал для химиков частью рутинной работы.

Качество введенной в базу данных информации и ее полнота являются ключевыми факторами, определяющими, насколько эффективно проводится поиск,

насколько часто в ходе поиска появляются ошибочные результаты и насколько результаты поиска будут полезными. Если проблемы работы с базами данных химических структур, как правило, связаны с особенностями химических объектов: таутомеризацией [1], проблемами стандартизации представления структур молекул или просто ошибками ввода [2], то для баз данных химических реакций добавляются проблемы неполноты. В настоящее время в доступных коммерческих базах данных химических реакций (таких, как Reaxys [3], поисковая система SciFinder [4]) подавляющее большинство реакций (90–95%) хранится в стехиометрически неуравновешенном виде, то есть некоторые реагенты или продукты в них упущены и приведены только ключевые компоненты [5]. Информация о растворителях и реагентах зачастую бывает приведена либо в специальных текстовых полях, либо в общем описании синтеза. Если какая-то структура оказывается пропущенной в уравнении реакции, то она не может быть найдена в ходе структурного поиска (рис. 1), что приводит к тому, что информация о реакции оказывается недоступной. Идентификация пропущенных в реакционном уравнении структур является достаточно сложной, а иногда и невозможной. До настоящего времени была опубликована только одна работа [6], в которой был описан весьма сложный алгоритм поиска пропущенных реагентов.

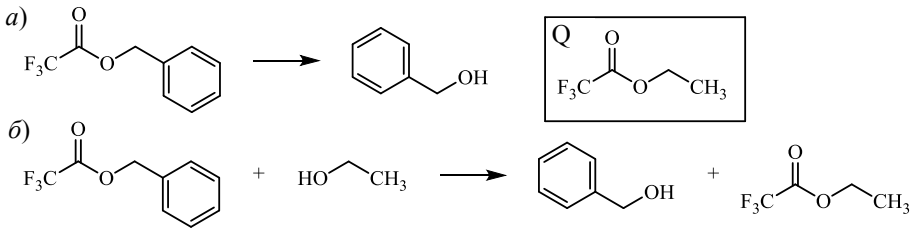

Рис. 1. Пример незаполненной (*а*) и заполненной (*б*) реакции. Если пользователя интересуют реакции, в которые вовлечены молекулы Q, то незаполненная реакция (*а*) не будет приведена в результатах, тогда как заполненная (*б*) будет

Мы предлагаем простой и эффективный способ для поиска структур пропущенных реагентов в химических реакциях. Предлагаемая технология основана на подходе конденсированного графа реакции (КГР) [7, 8]. Получаемые в результате такой обработки стехиометрически уравновешенные реакции, помимо задач хранения и поиска, можно использовать при классификации реакций по типам [9] и установлении количественных соотношений между характеристиками реакций и структурами соединений [10–12]. При этом обращается внимание на то, что в некоторых особо сложных случаях структура реагента не может быть установлена в принципе, поскольку отсутствует необходимая для этого информация. Примером может быть реакция, приведенная на рис. 1, *а*. Поскольку в ней пропущена информация о продукте и реагенте одновременно, на основании приведенного реакционного уравнения невозможно установить, является ли эта реакция гидролизом, сольволизом или переэтерификацией, и, соответственно, невозможно заполнить информацию о пропущенных реагентах и продуктах без дополнительных сведений.

Экспериментальные результаты и их обсуждение

КГР представляет собой визуализацию структурной информации, закодированной в реакционном уравнении, в виде одного псевдомолекулярного графа, в котором вершинами являются атомы, а ребрами, соединяющими атомы, – либо обычные химические связи (одинарные, двойные и т. п.), либо так называемые динамические связи, обозначающие разрыв, образование или изменение порядка связи (рис. 2). КГР (рис. 2, *б*) может быть получен из реакции (рис. 2, *а*) наложением соответствующих друг другу атомов и идентификацией изменившихся в реакции связей после установления соответствия между атомами реагентов и продуктов (атом-атомного отображения). Существует взаимная однозначность между структурным представлением реакции и ее КГР, то есть можно осуществлять конвертацию одного превращения в другое без потери информации.

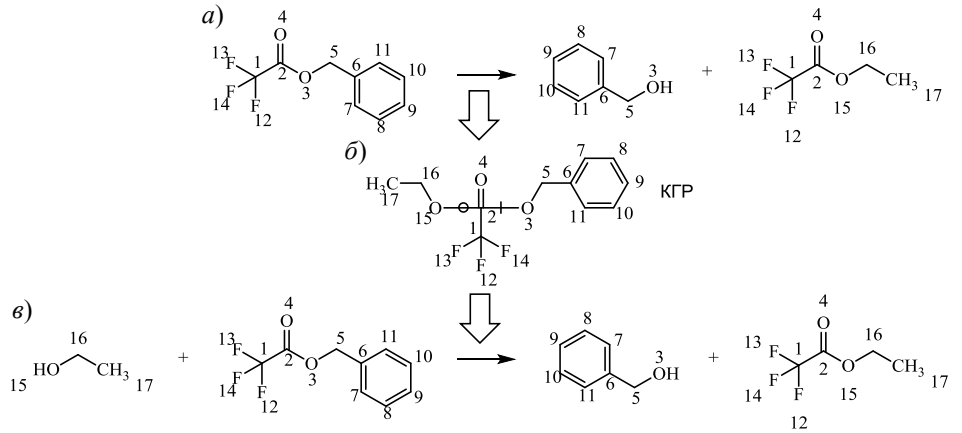

Рис. 2. Химическая реакция (*а*) и ее конденсированный граф (*б*). Номера рядом с атомами соответствуют атом-атомному отображению. В КГР зачеркнутая связь означает разорванную одинарную, кружком помечена образованная одинарная С–О-связь. Превращение КГР в реакцию (*в*) позволяет идентифицировать пропущенные реагенты и продукты

Ключевым наблюдением, полезным для поиска пропущенных реагентов, является то, что КГР может быть построен даже для незаполненной химической реакции. Обратное преобразование графа в реакцию позволяет идентифицировать пропущенные реагенты и продукты (см. рис. 2, *в*). Полученная реакция будет сбалансированной по атомам. В этом и заключается суть предлагаемого нами подхода восстановления информации о пропущенном реагенте или продукте.

Однако у предложенного подхода есть некоторые ограничения, приводящие в некоторых случаях к ошибкам в структурах реагентов и продуктов: зачастую в реакции участвует несколько реагентов одного типа, но коэффициенты в реакции не указаны (рис. 3, *а*). В этом случае предложенный алгоритм не способен полностью уравновесить реакцию. Для балансировки коэффициентов реакции был реализован алгоритм, который с помощью подструктурного поиска идентифицирует, сколько раз в КГР повторяется один и тот же фрагмент (выделен пунктирной линией на рис. 3, *а*), и дублирует соответствующие реагенты и продукты требуемое число раз. Вторую проблему для предложенного подхода представляют изменения в формальных зарядах и степенях окисления атомов в реагентах

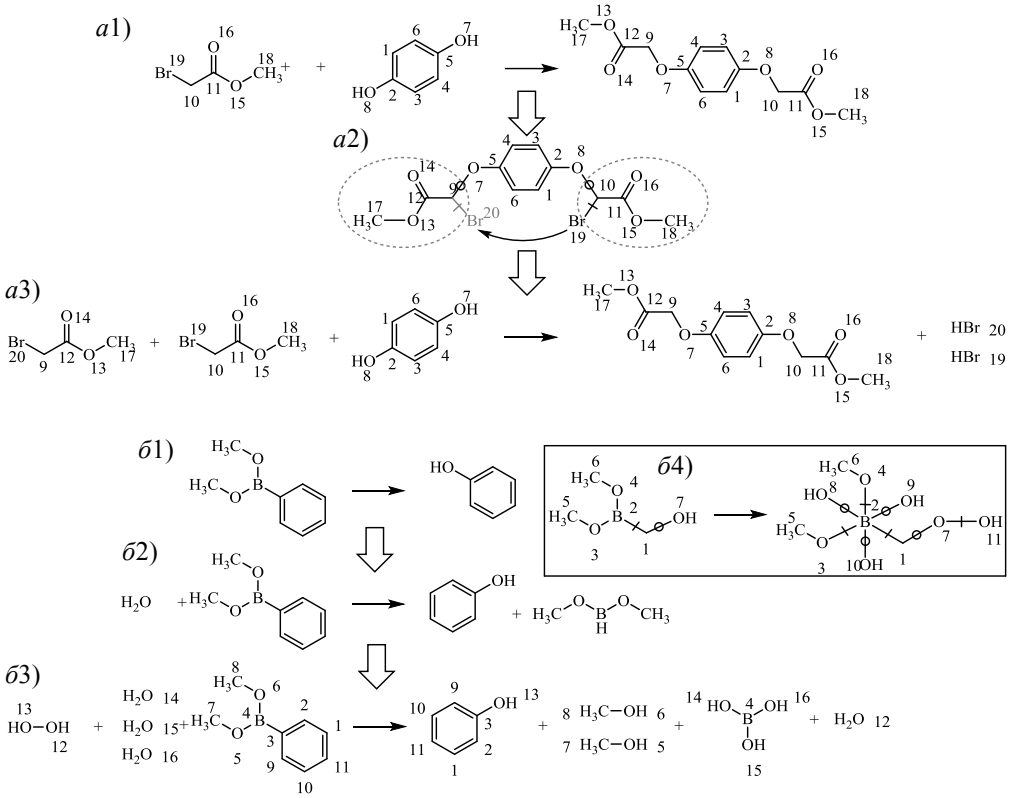

Рис. 3. *а*) Несбалансированная по коэффициентам и продуктам химическая реакция (*а*1), ее конденсированный граф (*а*2) и конечная сбалансированная реакция (*а*3). В КГР выделены общие фрагменты и показан процесс копирования утерянной информации об уходящей группе. *б*) Несбалансированная реакция (*б*1), некорректно сбалансированная (*б*2) и восстановленная (*б*3) на основе эмпирического правила (*б*4), вносящего информацию о пропущенных реагентах и ряде продуктов

и продуктах (см. рис. 3, *б*). В результате, в некоторых случаях (реакциях окисления-восстановления) структура реагента или продукта устанавливается некорректно. Для этого нами был создан основанный на эмпирических правилах алгоритм, который применяет предложенные правила для трансформации КГР в полностью сбалансированную реакцию с корректными структурами реагентов и продуктов. Принцип работы алгоритма заключается в изменении частей КГР, удовлетворяющих определенному правилу, в соответствии с заранее определенным шаблоном (рис. 3, *б*4). С помощью изоморфного вложения левой части шаблона идентифицируется часть КГР, требующая модификации, после применения правила создается скорректированный КГР, конвертация которого обратно в реакцию приводит к корректно стехиометрически уравновешенной реакции (рис. 3, *б*3). Набор правил замены создавался вручную и включал в себя пять самых распространенных типов реакций: гидролиза некоторых групп и этерификации. Для расширения применимости подхода возможно добавление новых правил. Эти технологии легли в основу разработанного нами алгоритма для стехиометрического уравновешивания реакций CGRBalancer.

Заключение

С использованием технологии конденсированного графа реакции был реализован достаточно простой и эффективный алгоритм, способный идентифицировать структуры пропущенных реагентов и продуктов в химических реакциях. Процесс поиска занимает доли секунд на одну реакцию, не требует анализа механизма реакции и возвращает реалистичные стехиометрически уравновешенные реакции, если в исходной реакции не было упущено слишком много информации. Данный подход, лежащий в основе алгоритма CGRBalancer, может быть использован для заполнения пропущенной информации в больших базах данных химических реакций.

Благодарности. Работа выполнена при финансовой поддержке Российского научного фонда (проект № 14-43-00024).

Литература

1. *Sitzmann M., Ihlenfeldt W.-D., Nicklaus M.C.* Tautomerism in large databases // J. Comput.-Aided Mol. Des. – 2010. – V. 24, No 6–7. – P. 521–551. – doi: 10.1007/s10822-010-9346-4.
2. *Williams A.J., Ekins S., Tkachenko V.* Towards a gold standard: Regarding quality in public domain chemistry databases and approaches to improving the situation // Drug Discovery Today. – 2012. – V. 17, No 13. – P. 685–701. – doi: 10.1016/j.drudis.2012.02.013.
3. Reaxys, version 1.7.8. – Elsevier. – URL: www.reaxys.com.
4. *Ridley D.* Information Retrieval: SciFinder. – John Wiley & Sons, 2009. – 226 p. – doi: 10.1002/9780470749418.
5. *Kraut H., Eiblmaier J., Grethe G., Löw P., Matuszczyk H., Saller H.* Algorithm for reaction classification // J. Chem. Inf. Model. – 2013. – V. 53, No 11. – P. 2884–2895. – doi: 10.1021/ci400442f.
6. *Patel H., Bodkin M.J., Chen B., Gillet V.J.* Knowledge-based approach to de Novo design using reaction vectors // J. Chem. Inf. Model. – 2009. – V. 49, No 5. – P. 1163–1184. – doi: 10.1021/ci800413m.
7. *Hoonakker F., Lachiche N., Varnek A.* Condensed Graph of Reaction: Considering a chemical reaction as one single pseudo molecule // Int. J. Artif. Intell. Tools. – 2011. – V. 20, No 2. – P. 253–270.
8. *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. – V. 19, No 9–10. – P. 693–703. – doi: 10.1007/s10822-005-9008-0.
9. *Schneider N., Lowe D.M., Sayle R.A., Landrum G.A.* Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity // J. Chem. Inf. Model. – 2015. – V. 55, No 1. – P. 39–53. – doi: 10.1021/ci5006614.
10. *Madzhidov T.I., Bodrov A. V., Gimadiev T.R., Nugmanov R.I., Antipin I.S., Varnek A.* Structure–reactivity relationship in bimolecular elimination reactions based on the condensed graph of a reaction // J. Struct. Chem. – 2015. – V. 56, No 7. – P. 1227–1234. – doi: 10.1134/S002247661507001X.
11. *Madzhidov T.I., Polishchuk P.G., Nugmanov R.I., Bodrov A. V., Lin A.I., Baskin I.I., Antipin I.S., Varnek A.* Structure-reactivity relationships in terms of the condensed graphs of reactions // Russ. J. Org. Chem. – 2014. – V. 50, No 4. – P. 459–463. – doi: 10.1134/S1070428014040010.

12. *Nugmanov R.I., Madzhidov T.I., Khaliullina G.R., Baskin I.I., Antipin I.S., Varnek A*. Development of "structure-property" models in nucleophilic substitution reactions involving azides // J. Struct. Chem. – 2014. – V. 55, No 6. – P. 1026–1032. – doi: 10.1134/S0022476614060043.

Поступила в редакцию
16.11.17

Нугманов Рамиль Ирекович, кандидат химических наук, старший научный сотрудник отдела органической химии

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *stsouko@live.ru*

Маджидов Тимур Исмаилович, кандидат химических наук, старший научный сотрудник отдела органической химии

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *tmadzhid@gmail.com*

Антипин Игорь Сергеевич, доктор химических наук, заведующий кафедрой органической химии

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *iantipin54@ya.ru*

Варнек Александр Алексеевич, доктор химических наук, заведующий лабораторией хемоинформатики

Университет Страсбурга
ул. Рене Декарта, д. 5, г. Страсбург, 67084, Франция
E-mail: *varnek@unistra.fr*

ISSN 2542-064X (Print)
ISSN 2500-218X (Online)

UCHENYE ZAPISKI KAZANSKOGO UNIVERSITETA. SERIYA ESTESTVENNYE NAUKI
(Proceedings of Kazan University. Natural Sciences Series)

2018, vol. 160, no. 1, pp. 32–39

An Approach for Automated Detection of Missing Reagents and Products in Chemical Reaction Equations

R.I. Nugmanov[a*], *T.I. Madzhidov*[a**], *I.S. Antipin*[a***], *A.A. Varnek*[a,b****]

[a]*Kazan Federal University*, *Kazan*, *420008 Russia*
[b]*University of Strasbourg*, *Strasbourg*, *67084 France*
E-mail: [*]*stsouko@live.ru*, [**]*tmadzhid@gmail.com*, [***]*iantipin54@ya.ru*, [****]*varnek@unistra.fr*

Received November 16, 2017

Abstract

One of the problems of chemical reactions databases management is the fact that most reactions in them are stoichiometrically unbalanced, i.e., they lack information about one or more reagents or products. It leads to the problem that some reactions cannot be found in structural searches. A solution of the problem based on the use of the Condensed Graph of Reaction (CGR) approach has been suggested. The CGR is a compressed representation of the reaction transformation in which the atoms of the reagent and the product are aligned. As a result, so-called dynamic bonds corresponding to changes in the bond

order during the chemical transformation are identified in the pseudomolecular graph along with the usual chemical bonds. The key observation is that the Condensed Graph, on the one hand, contains all required structural information on chemical reaction and, on the other hand, can be constructed for an unbalanced chemical reaction with missing reagents or products. During the inverse transformation of the Condensed Graph into a chemical reaction, the information on the missing reagents and products can be restored. There are some limitations to this approach, because information about coefficients missing in the reaction cannot be retrieved by the proposed method. To solve such problems, an algorithm based on the use of isomorphic embedding has been developed. This algorithm reveals similar motifs in the Condensed Graph and, if necessary, adds a coefficient before the reagent and/or product, or copies them the required number of times in the reaction equation.

Keywords: chemoinformatics, chemical reactions, stoichiometric reaction balancing, Condensed Graph of Reaction

Acknowledgments. The study was supported by the Russian Science Foundation (project no. 14-43-00024).

Figure Captions

Fig. 1. The example of unbalanced (*a*) and balanced (*b*) reactions. If the user is interested in the reactions involving molecule Q, the unfilled reaction (*a*) will not be retrieved as the results of the search, while the filled reaction (b) will be.

Fig. 2. Chemical reaction (*a*) and its Condensed Graph (*b*). The numbers next to atoms correspond to atom-to-atom mapping. In CGR, a crossed bond means a cleaved bond, a circle marks formed single C–O bond. Converting CGR into reaction (*c*) allows identification of the missed reagents and products.

Fig. 3. *a*) Unbalanced reaction with missed coefficients and products (*a*1), its Condensed Graph (*a*2) and the final balanced reaction (*a*3). Using CGR, common fragments are identified. The process of restoration of the lost information about the outgoing group is shown. *b*) An unbalanced reaction (*b*1), incorrectly balanced (*b*2), and restored (*b*3) on the basis of the empirical rule (*b*4), which provides information about the missing reagents and a number of products.

References

1. Sitzmann M., Ihlenfeldt W.-D., Nicklaus M.C. Tautomerism in large databases. *J. Comput.-Aided Mol. Des.*, 2010, vol. 24, nos. 6–7, pp. 521–551. doi: 10.1007/s10822-010-9346-4.
2. Williams A.J., Ekins S., Tkachenko V. Towards a gold standard: Regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug Discovery Today*, 2012, vol. 17, no. 13, pp. 685–701. doi: 10.1016/j.drudis.2012.02.013.
3. Reaxys, version 1.7.8. Elsevier. Available at: www.reaxys.com.
4. Ridley D. *Information Retrieval: SciFinder*. John Wiley & Sons, 2009. 226 p. doi: 10.1002/9780470749418.
5. Kraut H., Eiblmaier J., Grethe G., Löw P., Matuszczyk H., Saller H. Algorithm for reaction classification. *J. Chem. Inf. Model.*, 2013, vol. 53, no. 11, pp. 2884–2895. doi: 10.1021/ci400442f.
6. Patel H., Bodkin M.J., Chen B., Gillet V.J. Knowledge-based approach to de Novo design using reaction vectors. *J. Chem. Inf. Model.*, 2009, vol. 49, no. 5, pp. 1163–1184. doi: 10.1021/ci800413m.
7. Hoonakker F., Lachiche N., Varnek A. Condensed Graph of Reaction: Considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools*, 2011, vol. 20, no. 2, pp. 253–270.
8. 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, vol. 19, nos. 9–10, pp. 693–703. doi: 10.1007/s10822-005-9008-0.
9. Schneider N., Lowe D.M., Sayle R.A., Landrum G.A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.*, 2015, vol. 55, no. 1, pp. 39–53. doi: 10.1021/ci5006614.
10. Madzhidov T.I., Bodrov A. V., Gimadiev T.R., Nugmanov R.I., Antipin I.S., Varnek A. Structure–reactivity relationship in bimolecular elimination reactions based on the condensed graph of a reaction. *J. Struct. Chem.*, 2015, vol. 56, no. 7, pp. 1227–1234. doi: 10.1134/S002247661507001X.

11. Madzhidov T.I., Polishchuk P.G., Nugmanov R.I., Bodrov A. V., Lin A.I., Baskin I.I., Antipin I.S., Varnek A. Structure-reactivity relationships in terms of the condensed graphs of reactions. *Russ. J. Org. Chem.*, 2014, vol. 50, no. 4, pp. 459–463. doi: 10.1134/S1070428014040010.

12. Nugmanov R.I., Madzhidov T.I., Khaliullina G.R., Baskin I.I., Antipin I.S., Varnek A. Development of "structure-property" models in nucleophilic substitution reactions involving azides. *J. Struct. Chem.*, 2014, vol. 55, no. 6, pp. 1026–1032. doi: 10.1134/S0022476614060043.

***Для цитирования*:** *Нугманов Р.И., Маджидов Т.И., Антипин И.С., Варнек А.А.* Автоматическое определение пропущенных реагентов и продуктов в уравнении химических реакций // Учен. зап. Казан. ун-та. Сер. Естеств. науки. – 2018. – Т. 160, кн. 1. – С. 32–39.

***For citation*:** Nugmanov R.I., Madzhidov T.I., Antipin I.S., Varnek A.A. An approach for automated detection of missing reagents and products in chemical reaction equations. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Estestvennye Nauki*, 2018, vol. 160, no. 1, pp. 32–39. (In Russian)