УДК 004.89:004.4 **DOI: 10.20998/2411-0558.2016.21.11**

Г.А. САМИГУЛИНА, д-р техн. наук, зав. лаб. "Интеллектуальные системы управления и прогнозирования", Институт информационных и вычислительных технологий, Алматы, Казахстан,
З.И. САМИГУЛИНА, вед. научн. сотр., Ph.D, лаб. "Интеллектуальные системы управления и прогнозирования", Институт информационных и вычислительных технологий, Алматы, Казахстан

ПОСТРОЕНИЕ ОПТИМАЛЬНОЙ ИММУННОСЕТЕВОЙ МОДЕЛИ ДЛЯ КОМПЬЮТЕРНОГО МОЛЕКУЛЯРНОГО ДИЗАЙНА СУЛЬФАНИЛАМИДОВ НА ОСНОВЕ АЛГОРИТМА RANDOM FOREST

Статья посвящена исследованиям в области компьютерного молекулярного дизайна новых лекарственных препаратов сульфаниламидов на основе технологии иммунносетевого моделирования. В качестве метода предварительной обработки данных для выделения информативных дескрипторов и построения оптимальной иммунносетевой модели представлен алгоритм Random Forest, который позволяет ранжировать переменные по степени значимости. Библиогр.: 19 назв.

Ключевые слова: компьютерный молекулярный дизайн, технология иммуносетевого моделирования, сульфаниламиды, алгоритм Random Forest.

Постановка проблемы и анализ литературы. В настоящее время актуальна разработка и применение новых интеллектуальных технологий для QSAR (прогнозирование зависимости "структура – активность") при целенаправленном синтезе лекарственных соединений, обладающих необходимым комплексом фармакологических свойств. Применение компьютерного молекулярного дизайна на основе интеллектуальных подходов значительно сокращает временные, вычислительные и финансовые ресурсы при создании новых лекарственных соединений.

Современные методы QSAR широко применяют интеллектуальные алгоритмы для синтеза новых лекарственных препаратов. К методам искусственного интеллекта, используемым для прогнозирования зависимости "структура – активность", относятся: искусственные нейронные сети [1], генетические алгоритмы [2], искусственные иммунные системы [3], роевой интеллект [4 – 5] (алгоритм муравьиных колоний, пчелиный алгоритм) и т.д.

В связи с экспоненциальным ростом потребности в новых соединениях в фармацевтической промышленности требуется разработка нетрадиционных методов для обработки многомерной структурной информации. По сведениям одного из главных поставщиков химической

информации CAS число структур химических соединений составляет около 60 млн. [6]. Самый крупный поставщик образцов химических соединений ChemNavigator предлагает около 59,9 млн. уникальных веществ [7]. Число виртуальных молекул (содержащих до 13 атомов) сгенерированных компьютером за последнее время составляет почти миллиард. В связи с этим при компьютерном молекулярном дизайне актуально применение различных интеллектуальных и статистических подходов для решения проблемы выделения информативных дескрипторов, характеризующих химическое вещество. Широкое применение нашли такие статистические подходы как: факторный анализ (Factor Analysis, FA) [8], метод-опорных векторов (Support Vector Machine, SVM) [9], случайный лес (Random Forest, RF) [10] и др.

Алгоритм Random Forest представляет особый интерес в качестве метода обработки химической структурной информации. Например, в работе [11] авторами представлен подход для прогнозирования активности 1-, 3-, 5-триазинов, как каннабиноидных рецепторов (CB2) с использованием RF метода. Вычисляется двадцать молекулярных дескрипторов для набора данных из 58 аналогов. В зависимости от значений дескрипторов осуществляется обучение случайного леса для поиска связи между биологической активностью и молекулярной структурой аналогов. Результаты, полученные с помощью случайного леса сравнивались с методом опорных векторов и деревьями решений (Decision Trees, DT). Наилучший результат (100%) показал метод RF относительно метода SVM (93%) и DT (67%) соотвественно. Работа [12] посвящена исследованиям рецептора эпидермального фактора роста (EGFR) при лечении раковых болезней. Ранее был разработан ряд моделей QSAR, посвященный прогнозированию ингибирующей активности молекул против EGFR. Авторами [12] сделана попытка разработки моделей прогнозирования на большом множестве молекул (~ 3500 молекул), которые включают в себя различные каркасы, такие как хиназолин, пиримидин, хинолин и индол. В работе использовалась база данных PubChem, при этом применение алгоритма Random Forest дало максимальный результат с точностью 83,7%. В исследовании [13] представлено применение алгоритма Random Forest при разработке моделей QSAR предсказания токсичности водной среды химических соединений. Работа [14] посвящена разработке нового алгоритма интерпритации моделей случайного леса. Алгоритм позволяет вычислить вклад каждого дескриптора. Предлагаемая мера информативности дескрипторов не является альтернативой важности переменной Бреймана, но она характеризует вклад конкретной переменной к расчетному значению отклика при исследовании зависимости "структура-активность" химических соединений. Таким образом,

исследования в области применения современных методов искусственного интеллекта и статистических подходов при построении QSAR моделей являются актуальной задачей.

Цель статьи – построение оптимальной иммунносетевой модели при прогнозировании зависимости "структура – активность" лекарственных соединений сульфаниламидов на основе выделения информативных дескрипторов с использованием алгоритма Random Forest. Критерием оптимальности является максимальное сохранение информации о химическом соединении при минимальном количестве дескрипторов.

Реализация алгоритма Random Forest для построения оптимальной иммунносетевой модели. Рассмотрим иммунносетевую технологию прогнозирования зависимости "структура – активность" лекарственных соединений (сульфаниламидов) на основе искусственных иммунных систем [15]. На первом шаге осуществляется выбор химических соединений для исследования, затем описание структуры соединений на основе дескрипторов и их классификация по прогностическим группам. Далее осуществляется построение оптимальной математической модели (временных рядов – эталонов, состоящих из дескрипторов, описывающих структуру выбранных химических веществ с известными свойствами). На следующем шаге осуществляется обучение искусственной иммунной системы по эталонам. Затем реализуется алгоритм распознавания образов на основе сингулярного разложения матриц для определения класса, к которому принадлежат рассматриваемые образы. Оценка энергетических погрешностей по гомологам позволяет определить эффективность распознавания образов. После этого производится отбор кандидатов лекарственных соединений с заданными свойствами для дальнейших исследований.

При построении оптимальной иммунносетевой модели применяется мультиалгоритмический подход, который заключается в следующем [16]. На этапе предварительной обработки исходного набора дескрипторов сульфаниламидов выбираются несколько интеллектуальных или статистических алгоритмов [17], которые обрабатывают исходный набор дескрипторов. Далее осуществляется выбор наилучшего из них, то есть выбирается алгоритм с наименьшей ошибкой обобщения.

В качестве одного из алгоритмов, используемых в мультиалгоритмическом подходе, выбран Random Forest. Метод основан на построении ансамбля деревьев решений.

Рассмотрим следующие понятия [18 – 19]: $T_i(x),\ i=\overline{1,\ B}$ – это ансамбль деревьев решений, где x – вектор дескрипторов размерности N; B – количество деревьев в ансамбле; $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ – обучающая выборка данных, где y_j, $j=\overline{1,\ n}$ – класс.

Далее из исходной выборки данных выбирается n случайных объектов с повторениями (bootstrap sample), принимающих значение D_i. Затем строится дерево решений для D_i (полное построение без отсечения ветвей) с повтором алгоритма B раз. Мера информативности дескрипторов, основанная на алгоритме Random Forest находится по следующему алгоритму.

Алгоритм.

1. Пусть x_i – некоторый дескриптор. Необходимо построить случайный лес и получить оценку вероятности ошибочной классификации E_i методом Out-Of-Bag, предложенным Брейманом [19] на основе наблюдений, не входящих в bootstrap выборки. Данные выборки называются выборками Out-Of-Bag.

2. В выборках Out-Of-Bag осуществить случайную перестановку значений дескриптора x_i для каждого дерева из построенного случайного леса.

3. По модифицированным Out-Of-Bag определяется оценка вероятности ошибочной классификации $\widehat{E}_i$.

4. Определение информативности дескриптора x_i реализуется как значение разницы: $I(x_i)=\max(0,\widehat{E}_i-E_i)$.

В качестве достоинств данного метода можно представить: высокое качество получаемых моделей; эффективная обработка данных с большим числом дескрипторов и классов; возможность распараллеливания алгоритмов для работы с большими данными; деревья решений являются удобной моделью для представления знаний в экспертных системах; нечувствительны к масштабированию; обучающая выборка может содержать признаки измеренные как в числовой, так и в номинальной шкале; наличие методов оценивания информативности отдельных дескрипторов за счет внутренней оценки Out-Of-Bag и т.д.

В качестве недостатков можно выделить: для некоторого типа задач (с наличием сильной зашумленности) возможно переобучение алгоритма; модели занимают много памяти для хранения данных.

Выводы. В настоящее время не существует универсальных методов для оценки меры информативности дескрипторов в исходной выборке данных, а поскольку при решении задач синтеза новых лекарственных

препаратов используется огромное количество структур химической информации, применение мультиалгоритмического подхода является удобным с точки зрения подбора наилучшего алгоритма для построения оптимальной иммунносетевой модели.

Алгоритм Random Forest с его достоинствами и малым количеством недостатков успешно применяется в качестве инструмента для оценки степени важности дескрипторов сульфаниламидов при построении оптимальной иммунносетевой модели для компьютерного молекулярного дизайна лекарственных препаратов на основе искусственных иммунных систем.

Исследования проводятся по гранту №ГР 0115РК00549 МОН РК по теме: Компьютерный молекулярный дизайн лекарственных препаратов на основе иммунносетевого моделирования (2015-2017 гг.).

References:

1. Montañez-Godínez, N., Martínez-Olguín, A.C., Deeb, O., Garduño-Juárez, R. and Ramírez-Galicia, G. (2015), "QSAR/QSPR as an Application of Artificial Neural Networks", *Journal Artificial Neural Networks,* Vol. 1260, pp. 319-333.

2. Nekoei, M., Mohammadhosseini, M. and Pourbasheer, E. (2015) "QSAR study of VEGFR-2 inhibitors by using genetic algorithm-multiple linear regressions (GA-MLR) and genetic algorithm-support vector machine (GA-SVM): a comparative approach", *Journal of Medicinal Chemistry Research,* Vol. 24, No. 7, pp. 3037-3046.

3. Ivanciuc, O. (2009), "Drug Design with Artificial Neural Networks", *Encyclopedia of Complexity and Systems Science,* pp. 2139-2159.

4. Khajeh, A. Modarress, H. and Zeinoddini-Meymand, H. (2013), *"*Modified particle swarm optimization method for variable selection in QSAR/QSPR studies*", Journal of Structural Chemistry*, Vol. 24, No. 5, pp. 1401-1409.

5. Goodarzi, M., Freitas, M.P. and Jensen, R. (2009), "Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3,5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions", *Journal of Chemometrics and Intelligent Laboratory Systems*, Vol. 98, No. 2., pp. 123-129.

6. Blum, L. and Reymond, J. (2009), "970 million drug like small molecules for virtual screening in the chemical universe database GDB -13", *Journal of American Chemical* Society, Vol. 131 (25), pp. 8732-8733**.**

7. Nonell-Canals, A. and Mestres, J. (2011), "In silico target profiling of one billion molecules", *Molecular Informatics*, Vol. 30 (5), pp. 405-409.

8. Shahlaie, M., Fassihi, A., Pourhossein, A., Arkan, E. (2013), "Statistically validated QSAR study of some antagonists of the human CCR5 receptor using least square support vector machine based on the genetic algorithm and factor analysis", *Journal of Medicinal Chemistry Research*, Vol. 22, No. 3, pp. 1399-1414.

9. Zhou, X.B., Han, W.J., Chen, J. and Lu, X.Q. (2011), "QSAR study on the interactions between antibiotic compounds and DNA by a hybrid genetic-based support vector machine", *Monatshefte für Chemie - Chemical Monthly*, Vol. 142, No. 9, pp. 949-959.

10. Sprague, B., Shi, Q., Kim, M.T., Zhang, L., Sedykh, S., Ichiishi, E., Tokuda, H., Lee, K. and Zhu, H. (2014), "Design, synthesis and experimental validation of novel potential chemopreventive agents using random forest and support vector machine binary classifiers", *Journal of Computer-Aided Molecular Design*, Vol. 28, No. 6, pp 631-646.

11. Abu El-Atta, A. and Hassanien, A.E. (2014), "Predicting Biological Activity of 2,4,6-trisubstituted 1,3,5-triazines Using Random Forest", Advances in Intelligent Systems and Computing, *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014,* Vol. 303, pp. 101-110.
12. Singh, H., Singh S., Singla, D., Agarwal, S.M. and Raghava, G.P. (2015), "QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest", *Biology Direct*, Vol. 10, No. 10, pp. 1-10.
13. Polishchuk, P.G., Muratov, E.N., Artemenko, A.G., Kolumbin, O.G., Muratov, N.N. and Kuz'min, V.E. (2009), "Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity", *American Chemical Society*, Vol. 49, No. 11, pp. 2481-2488.
14. Kuz'min, V.E., Polishchuk, P.G., Artemenko, A.G. and Andronati, S.A. "Interpritation of QSAR Models Based on Random Forest Methods", *Molecular Informatics*, Vol. 30, No. 6-7, pp. 593-603.
15. Samigulina, G.A. and Samigulina, Z.I. (2015), "Computational molecular design of sulfanomides based on immune network modeling", *Twelve International Conference on Electronics Computer and Computation (ICECCO)*, pp. 1-5.
16. Samigulina, G.A., Samigulina, Z.I., Wuizik, W. and Krak, Yu. (2015), "Prediction of "structure – property" Dependence of New Organic Compounds on the basis of Artificial Immune Systems", *Journal of Automation and Information Sciences*, Vol. 47, No. 4, pp. 28-35.
17. Samigulina, G.A. Building immunnosetevoy optimal model to predict the unknown properties of medicinal compounds based approach multialgoritmicheskogo (2013), *Informatics Problems*, Novosibirsk, No. 2, pp. 21-29.
18. Breiman, L. (2001), "Random forests", *Machine Learning*, Vol 45, No. 1, pp. 5-32.
19. Breiman, L. Cutler, A. (2005), *Random forests,* Berkley, 56 p.

Поступила (received) 6.05.2016

Статью представил д-р техн. наук, проф. НТУ "ХПИ" Леонов С.Ю.

Samigulina Galina, Dr. Sci. Tech.
Institute of Information and Computing Technologies
Str. Pushkeen, 125, Almaty, Kazakhstan, 050010
Tel:+7(777)244-43-67, e-mail: galinasamigulina@mail.ru
ORCID ID: 0000-0003-1798-9161

Samigulina Zarina, Cand. Sci. Tech.
Institute of Information and Computing Technologies
Str. Pushkeen, 125, Almaty, Kazakhstan, 050010
Tel:+7(702)218-97-73, e-mail: zarinasamigulina@mail.ru
ORCID ID: 0000-0002-5862-6415

УДК 004.89:004.4

Побудова оптимальної іммунносетевої моделі для комп'ютерного молекулярного дизайну сульфаніламідів на основі алгоритму Random Forest / Самігуліна Г.А., Самігуліна З.І. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2016. – № 21 (1193). – С. 102 – 108.

Стаття присвячена дослідженням в області комп'ютерного молекулярного дизайну нових лікарських препаратів сульфаніламідів на основі технології імунносетевого моделювання. Як метод попередньої обробки даних для виділення інформативних дескрипторів і побудови оптимальної іммунносетевой моделі представлений алгоритм Random Forest, який дозволяє ранжувати змінні за ступенем значущості. Бібліогр.: 19 назв.

Ключові слова: комп'ютерний молекулярний дизайн, технологія імунносетевого моделювання, сульфаніламіди, алгоритм Random Forest.

УДК 004.89:004.4

Построение оптимальной иммунносетевой модели для компьютерного молекулярного дизайна сульфаниламидов на основе алгоритма Random Forest / Самигулина Г.А., Самигулина О.И. // Весник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2016. – № 21 (1193). – С. 102 – 108.

Статья посвящена исследованиям в области компьютерного молекулярного дизайна новых лекарственных препаратов сульфаниламидов на основе технологии иммунносетевого моделирования. В качестве метода предварительной обработки данных для выделения информативных дескрипторов и построения оптимальной иммунносетевой модели представлен алгоритм Random Forest, который позволяет ранжировать переменные по степени значимости. Библтогр.: 19 назв.

Ключевые слова: компьютерный молекулярный дизайн, технология иммунносетевого моделирования, сульфаниламиды, алгоритм Random Forest.

UDC 004.89:004.4

Building of the immune network optimal model for computer molecular design sulfonamides based on random forest algorithm /Samigulina G.A., Samigulina Z.I. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2016. – № 21 (1193). – P. 102 – 108.

The article is devoted to research in the field of the computer molecular design of the new drugs sulfonamides on the basis of immune network technology. As a method of the data pre-processing to extraction of the informative descriptors and building of the optimal immune network model presented the Random Forest algorithm, which allows to rank variables in the order of importance. Refs.: 19 titles.

Keywords: computer molecular design, immune network technology, sulfonamides, algorithm Random Forest.