УДК: 004.891.2:
615.015.11: 544.165:
544.182.32: 544.187.2
DOI: 10.24412/2075-4094-2023-2-3-7 EDN JRAZFA **

СПОСОБ СВЕРТКИ ПРОСТРАНСТВА ПАРАМЕТРОВ ХИМИЧЕСКОЙ СТРУКТУРЫ ФАРМАКОЛОГИЧЕСКИ АКТИВНЫХ СОЕДИНЕНИЙ

М.А. ПЕРФИЛЬЕВ, П.М. ВАСИЛЬЕВ, А.Н. КОЧЕТКОВ, Д.А. БАБКОВ

ФГБОУ ВО «Волгоградский государственный медицинский университет» Министерства здравоохранения Российской Федерации (ФГБОУ ВО ВолгГМУ Минздрава России), площадь Павших борцов, д. 1, г. Волгоград, 400131, Россия, e-mail: maxim.firu@yandex.com

Аннотация. ***Цель исследования*** – разработать методологию свертки пространства параметров представления химической структуры фармакологически активных соединений, с учетом химико-биологических особенностей этих параметров, и ее апробация в нейросетевом моделировании. ***Материалы и методы исследования***. Обучающая выборка образована из 1283 химических соединений с известным уровнем ингибирования липополисахаридной интоксикации. Каждая структура была описана 119 характеристиками, обеспечивающими достаточно полное представление химико-биологических свойств изучаемых лигандов. Ядром свертки группы параметров одного химического соединения стала одномерная матрица переменного размера, строящаяся по числу сворачиваемых дескрипторов в рядах релевантных химико-биологическому смыслу. Построение ядер свертки проводилось в программе *Microsoft Excel 2010.* Нейросетевые модели были обучены в программном обеспечении *Statistica 8.* Статистический анализ выполнялся в программе *Microsoft Excel 2010 и Statistica 8.* ***Результаты и их обсуждение***. Произведен расчет ряда параметров известных соединений, снижающих липополисахаридную интоксикацию. Разработана методология свертки указанных параметров для получения интегрированных переменных представления химической структуры фармакологически активных соединений. Выполнено итеративное обучение нейросетевых моделей на исходном и свернутом пространстве описания, оценены распознающая и прогностическая точности полученных моделей. Показана целесообразность применения свертки для экономии времени и вычислительных ресурсов при сохранении точности прогноза.

Ключевые слова: свертка, искусственные нейронные сети, *QSAR*, докинг, липополисахаридная интоксикация, одномерная матрица.

METHOD FOR SPACE CONVOLUTION OF THE CHEMICAL STRUCTURE PARAMETERS OF PHARMACOLOGICALLY ACTIVE COMPOUNDS

M.A. PERFILEV, P.M. VASSILIEV, A.N. KOCHETKOV, D.A. BABKOV

Volgograd State Medical University of the Ministry of Health of Russian Federation, Pavshikh Bortsov Square, 1, Volgograd, 400131, Russia, 400131, Russia

Abstract. ***The research aim*** is to develop a methodology of the parameter space convolution for representing the chemical structure of pharmacologically active compounds, taking into account the chemical and biological characteristics of these parameters, and its approbation in neural network modeling. ***Materials and research methods***. The training set was formed from 1283 chemical compounds with a known level of inhibition of lipopolysaccharide intoxication. Each structure was represented by 119 characteristics, providing a sufficiently complete description of the chemical and biological properties of the studied ligands. The convolution kernel of parameter groups of one chemical compound has become a resizable one-dimensional matrix, which is built according to the number of collapsible descriptors in a group relevant to the chemical and biological meaning. Convolution kernels were built in *Microsoft Excel 2010*. Neural network models were trained in *Statistica 8 software*. Statistical analysis was performed using *Microsoft Excel 2010 and Statistica 8*. ***Results and discussion***. A parameters set of known compounds that reduce lipopolysaccharide intoxication were calculated. A methodology has been developed for convolving these parameters to obtain integrated variables representing the chemical structure of pharmacologically active compounds. An iterative method for training neural network models on the initial and convoluted description space showed high recognition and predictive accuracy of the resulting model. Prediction using convolution method has shown to be useful for saving time and computational resources while maintaining accuracy of prediction.

Keywords: convolution, artificial neural networks, QSAR, docking, lipopolysaccharide intoxication, one-dimensional matrix.

Введение. Базы данных известных фармакологически активных химических соединений могут содержать тысячи записей. При этом возможно разнообразное представление структуры этих лигандов, обусловленное их биологической ролью. Выполнение на выборках значительного объема нейросетевого моделирования занимает очень длительное время. Исходя из этого, возникла необходимость применения операций свертки пространства. Традиционная область, где наиболее часто применяются *сверточные нейронные сети, [convolution neural networks]* (*CNN*) – это задачи классификации изображений. Впервые идея свертки была применена в нейрокогнитроне, разработанным японским ученым Кунихикой Фукусимой в 1980 году на основе нейрофизиологических исследований зрительной коры головного мозга [3]. Свое развитие и современное название эта архитектура получила в 1989 году благодаря Яну Лекуну [5]. В задачах классификации изображений применяются двумерные рамки свертки, называемые ядром или маской. Свертка – это один из наиболее универсальных и мощных инструментов обработки данных, главная задача которой заключается в распознавании любых скрытых образов, в том числе описываемых совокупностью химико-биологических параметров. Отсюда возникла задача адаптировать способы свертки, преимущественно применяемые при распознавании скрытых образов в изображениях, для создания методологии нейросетевого моделирования на интегрированных переменных с учетом химико-биологического наполнения этих параметров.

Цель исследования – разработка методологии свертки пространства параметров представления химической структуры фармакологически активных соединений, с учетом химико-биологических особенностей этих параметров, и ее апробация в нейросетевом моделировании.

Материалы и методы исследования. В *CNN* применяются несколько общепринятых вариантов компоновки ядра свертки, различающихся по значениям весов ячеек. В настоящем исследовании ядром свертки группы параметров одного химического соединения была выбрана матрица размером 1 *x m* (вектор), где *m* – число сворачиваемых переменных в группе. Таким образом, при свертке пространства переменных каждой группы параметров применялись ядра свертки различного размера, в соответствии с химико-биологическим смыслом сворачиваемых параметров химической структуры. На первом шаге свертки центром маски служила первая ячейка сверточного вектора. В этом случае вес ядра свертки в центре равен 5, а коэффициенты свертки для остальных ячеек чередуясь принимают значения -1 и 0.1. При движении маски центр свертки переходит на следующее значение в векторе. После наложения и прохождения ядра по всем переменным в сворачиваемом ряду, полученные значения суммируются в одну интегрированную переменную.

Обучающая выборка была образована из 1283 известных химических соединений ингибирующих *липополисахаридную, [lipopolysaccharide]* (*LPS*) интоксикацию, кластеризованных по уровню активности на классы высокой и иной активности (*high*, *not_high*). Ансамблевый докинг в релевантные биомишени аденозинового рецептора A_{2A} – *ADORA2A*, тирозинкиназы Брутона – *BTK*, гликогенсинтазы киназы 3 β – *GSK3B*, киназы ассоциированной с рецептором ИЛ-1 – *IRAK1*, Янус-киназы 1 – *JAK1*, Янус-киназы 3 – *JAK3* и тирозинкиназы 2 – *TYK2* был выполнен с помощью программ *LigPlot+ 1.4.5* [6], *MarvinSketch 17.1.23* [2], *MOPAC2016* [7] и *AutoDock Vina 1.1.2* [9] по методике, описанной в работе [1]. Перечисленные белки играют важные роли в регуляции иммунного ответа, в том числе в сигнальных путях, приводящих к возникновению цитокинового шторма. В качестве параметров аффинности каждого химического соединения были взяты 7 минимальных энергий докинга ΔE в специфические сайты связывания указанных релевантных биомишеней. В качестве фрагментных параметров описания каждой химической структуры с помощью системы *IT Microcosm 7.3* были рассчитаны значения встречаемости 92 видов *(язык описания количественных соотношений структура-свойство, [Quantitative Structure-Activity Relationship Language]*) *QL*-дескрипторов первого ранга. Квантово-химические параметры были рассчитаны с помощью полуэмпирического метода *PM3* с использованием программы *HyperChem 8.0* [4] и специально написанного скрипта на *Visual Basic* в виде 20 значений энергий граничных *молекулярных орбиталей для 10 верхних заполненных* (МО E_{HOMO}) и для *10 нижних пустых* (МО E_{LUMO}).

В итоге каждая химическая структура была представлена 119 характеристиками, обеспечивающими достаточно полное представление химико-биологических свойств изучаемых лигандов. После удаления плохо определенных переменных, в итоговую обучающую выборку для классической *многослойной перцептронной нейросети,[multilayer perceptron]* (*MLP*) с архитектурой узкого горла вошли 108 переменных.

Процедуре свертки подвергались только объединенные в группы конвергентные параметры. К таким относятся *QL*-дескрипторы длины углеродной цепи и *QL*-дескрипторы, отвечающие за наличие алициклических и ароматических циклов. Сворачивались по двум группам переменных отдельно энергии 10 верхних заполненных МО E_{HOMO} и энергии 10 нижних пустых МО E_{LUMO}. Соединения были также описаны энергиями докинга в релевантные биомишени, однако из-за неясных взаимозависимостей между исследуемыми киназами энергии докинга не подвергались свертке. Таким образом, после свертки на вход нейронов в *CNN* подавались значения 38 переменных.

Нейросетевые модели с архитектурой многослойного перцептрона с узким горлом были обучены в

программе *Statistica 8* [8] отдельно на исходной полной обучающей выборке со всеми 108 входными нейронами и на свернутой обучающей выборке с 38 входными нейронами. В ходе итеративного обучения на первой стадии было построено 200 нейронных сетей. В последующих итерациях обучение проводилось циклами с начальными параметрами лучших нейросетей, отобранных на предыдущем этапе. Всего для каждой архитектуры было обучено около 500 нейросетевых моделей. В финале среди сетей последнего цикла обучения отбиралась одна наилучшая по точности нейронная сеть. Описание лучшей нейросетевой модели включает количество нейронов в слоях, параметр алгоритма обучения *Бройдена – Флетчера – Гольдфарба – Шанно* (*BFGS*), тип функции ошибки, типы активационных функций скрытого и выходного нейронов. Обучение выполнялось в режиме параллельных вычислений на двух одинаковых персональных суперкомпьютерах гибридной архитектуры на базе процессора *IntelCore i7-3930K 3.2* ГГц *6 core* с 24 Гб оперативной памяти, вычислительная нагрузка при обучении колебалась от 9% до 11%. Статистический анализ проводился в программах *Microsoft Excel 2010* и *Statistica 8.0* с помощью *F* критерия – точного критерия Фишера. Различия в точности прогнозов полученных искусственных нейросетей считались достоверными при критическом уровне значимости в проверке статистических гипотез $p \leq 0{,}05$.

Результаты и их обсуждение. Найдены лучшие по точности нейросети с архитектурой: *MLP*-108-30-2 (*BFGS61*, *Entropy*, *Tanh*, *Softmax*); *CNN*-38-22-2 (*BFGS57*, *Entropy*, *Logistic*, *Softmax*). Время обучения нейронной сети с архитектурой многослойного перцептрона с узким горлом на полной обучающей выборке для классификации соединений с высоким уровнем ингибирования *LPS*-интоксикации составило более 9 часов. Точность валидации при обучении равна 90%. Точность классификации на объединенной выборке высокоактивных соединений составила 92%, не высокоактивных соединений – 98%. Нейронная сеть с архитектурой многослойного перцептрона с узким горлом на свернутой выборке обучилась за 62 минуты, с точностью валидации при обучении 89%. Точность классификации на объединенной выборке высокоактивных соединений составила 90%, против 97% при классификации не высокоактивных соединений.

Разница в точности прогноза высокоактивных соединений между итоговыми нейросетевыми моделями двух исследуемых архитектур статистически недостоверна по точному критерию Фишера F=0.86, p=0,05, рассчитанному в электронных таблицах *Microsoft Excel 2010* и программе *Statistica 8.0*.

При прогнозе уровня активности 10 структур, не вошедших в обучающую выборку, точность классификации для нейросети на основе полного описания составила 60%, а на основе свернутого описания 50%, что статистически неразличимо по точному критерию Фишера F=1.00, p=0,05.

Выводы. Выполнено нейросетевое моделирование двух зависимостей высокого уровня активности по снижению липополисахаридной интоксикации от структуры химических соединений на основе полной и свернутой входных матриц данных. Распознающая и прогностическая способности двух моделей статистически не различаются. Продемонстрирована значительная экономия времени обучения при сохранении точности прогноза с применением архитектуры сверточной нейронной сети.

Разработана методология свертки пространства представления химической структуры фармакологически активных соединений.

Работа выполнена в рамках государственного задания Министерства здравоохранения Российской Федерации №121060700050-2 «Разработка методологии компьютерного поиска фармакологически активных соединений на основе множественного докинга и технологии искусственных нейронных сетей»

Литература

1. Перфильев М.А., Васильев П.М., Бабков Д.А., Спасов А.А., Кочетков А.Н., Королева А.Р., Голубева А.В. Корреляционная сеть ингибиторов киназ, регулирующих снижение LPS-интоксикации. Сборник научных трудов XXVII симпозиума «Биоинформатика и компьютерное конструирование лекарств». 2021. С. 65.

2. ChemAxon: официальный сайт. Будапешт. URL: https://chemaxon.com/products/marvin (дата обращения: 11.10.2022).

3. Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position // Biological Cybernetics. 1980. Vol.36, №4. P. 193–202. DOI: 10.1007/BF00344251

4. HyperChem: официальный сайт. Гейнсвилл. URL: http://www.hypercubeusa.com (дата обращения: 02.10.2022).

5. LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D. Backpropagation Applied to Handwritten Zip Code Recognition // Neural Computation. 1989. №1. P. 541–551. DOI: 10.1162/neco.1989.1.4.541

6. Laskowski R.A., Swindells M.B. LigPlot+: multiple ligand-protein interaction diagrams for drug dis-

covery // Journal of Chemical Information and Modeling. 2011. № 51. P. 2778–2786. DOI: 10.1021/ci200227u

7. MOPAC: официальный сайт. Глазго. URL: http://openmopac.net (дата обращения: 11.10.2022).

8. Statistica: официальный сайт. Пало-Альто. URL: http://www.statsoft.com (дата обращения: 14.10.2022).

9. Trott O., Olson A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading // Journal of Computational Chemistry. 2010. Vol.31, №2. P. 455–461. DOI: 10.1002/jcc.21334

References

1. Perfilev MA, Vassiliev PM, Babkov DA, Spasov AA, Kochetkov AN, Koroleva AR, Golubeva AV. Korrelyacionnaya set' ingibitorov kinaz, regulirujyschih snijenie LPS-intoksikacii [Correlation network of kinase inhibitors regulating the reduction of LPS intoxication]. Sbornik nauchyh trudov XXVII simpoziuma «Bioinformatika i komp'uternoe konstruirovanie lekarstv». 2021. Russian.

2. ChemAxon: official site. Budapest. URL: https://chemaxon.com/products/marvin (date of the application: 11.10.2022)

3. Fukushima K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics. 1980;36(4):193-202. DOI: 10.1007/BF00344251

4. HyperChem: official site. Gainesville. URL: http://www.hypercubeusa.com (date of the application: 02.10.2022).

5. LeCun Y, Boser B, Denker JS, Henderson D., Howard RE, Hubbard W, Jackel LD. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation. 1989;1:541-51. DOI: 10.1162/neco.1989.1.4.541

6. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. Journal of Chemical Information and Modeling. 2011;51:2778-86. DOI: 10.1021/ci200227u

7. MOPAC: official site. Glasgow. URL: http://openmopac.net (date of the application: 11.10.2022).

8. Statistica: official site. Palo Alto. URL: http://www.statsoft.com (date of the application: 14.10.2022).

9. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2010;31(2):455-61. DOI: 10.1002/jcc.21334

Библиографическая ссылка:
Перфильев М.А., Васильев П.М., Кочетков А.Н., Бабков Д.А. Способ свертки пространства параметров химической структуры фармакологически активных соединений // Вестник новых медицинских технологий. Электронное издание. 2023. №2. Публикация 3-7. URL: http://www.medtsu.tula.ru/VNMT/Bulletin/E2023-2/3-7.pdf (дата обращения: 24.04.2023). DOI: 10.24412/2075-4094-2023-2-3-7. EDN JRAZFA*

Bibliographic reference:
Perfilev MA, Vassiliev PM, Kochetkov AN, Babkov DA. Sposob svertki prostranstva parametrov himicheskoj struktury farmakologicheski aktivnyh soedinenij [Method for space convolution of the chemical structure parameters of pharmacologically active compounds]. Journal of New Medical Technologies, e-edition. 2023 [cited 2023 Apr 24];2 [about 4 p.]. Russian. Available from: http://www.medtsu.tula.ru/VNMT/Bulletin/E2023-2/3-7.pdf. DOI: 10.24412/2075-4094-2023-2-3-7. EDN JRAZFA

* номера страниц смотреть после выхода полной версии журнала: URL: http://medtsu.tula.ru/VNMT/Bulletin/E2023-2/e2023-2.pdf

**идентификатор для научных публикаций EDN (eLIBRARY Document Number) будет активен после выгрузки полной версии журнала в eLIBRARY