БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Hahn C., Bachmann L., Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads – a baiting and iterative mapping approach // Nucleic Acids Res. 2013. Vol. 41. № 13. P. 1–9.
2. Seemann T. Prokka: rapid procaryotic genome annotation // Bioinformatics. 2014. Vol. 30. № 14. P. 2068–2069.
3. Воронова Н. В., Бондаренко Ю. В., Левыкина С. С., Шулинский Р. С. Митохондриальный геном Aphis Fabae Mordvilkoi Borner & Janisch, 1992 // Молекулярная и прикладная генетика: сб. науч. трудов. 2018. Т. 25. С. 73–83.
4. Lee J., Park J., Lee H. et al. The complete mitochondrial genome of *Paracolopha morrisoni* (Baker, 1919) (Hemiptera: Aphididae) // Mitochondrial DNA Part B. 2019. Vol. 4. № 2. P. 3037–3039.
5. Voronova N. V., Warner D., Shulinski R. S. et al. The largest aphid mitochondrial genome found in invasive species *Therioaphis tenera* (Aizerberg, 1956) // Mitochondrial DNA Part B. 2019. Vol. 4. № 1. P. 730–731.
6. Cameron S. L. Insect mitochondrial genomics: implications for evolution and phylogeny // Annu. Rev. Entomol. 2014. Vol. 59. P. 95–117.
7. Voronova N. V., Levykina S. S., Warner D. et al. Characteristic and variability of five complete aphid mitochondrial genomes: Aphis fabae mordvilkoi, Aphis craccivora, Myzus persicae, Therioaphis tenera and Appendiseta robiniae (Hemiptera; Sternorrhyncha; Aphididae) // International Journal of Biological Macromolecules. 2020. V. 149. P. 187–206.
8. Sun W., Huynh B. L., Ojo J. A. et al. Comparison of complete mitochondrial DNA sequences between old and new world strains of the cowpea aphid, *Aphis craccivora* (Hemiptera: Aphididae) // Agri Gene. 2017. Vol. 4. P. 23–29.
9. Chen L., Chen P.-Y., Xue X.-F. et al. Extensive gene rearrangements in the mitochondrial genomes of two egg parasitoids, *Trichogramma japonicum* and *Trichogramma ostriniae* (Hymenoptera: Chalcidoedea: Trichogrammatidae) // Scientific Reports. 2018. Vol. 8. № 1. P. 1–11.

АВТОЭНКОДЕРНАЯ НЕЙРОННАЯ СЕТЬ ДЛЯ ГЕНЕРАЦИИ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ ВИЧ-1 МЕТОДАМИ ГЛУБОКОГО ОБУЧЕНИЯ

Г. И. Николаев[1], Н. А. Шульдов[2], А. И. Анищенко[2], А. В. Тузиков[1], А. М. Андрианов[3]

[1]Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь
[2]Белорусский государственный университет, Минск, Беларусь
[3]Институт биоорганической химии Национальной академии наук Беларуси, Минск, Беларусь
E-mail: reshaemvsem@gmail.com

Методами глубокого обучения разработан генеративный состязательный автоэнкодер для рационального дизайна потенциальных ингибиторов проникновения ВИЧ-1, способных блокировать участок белка gp120 оболочки вируса, критический для

его связывания с клеточным рецептором CD4. Были выполнены исследования, включающие создание архитектуры автоэнкодера, формирование молекулярной библиотеки потенциальных лигандов белка gp120 ВИЧ-1 для обучения нейронной сети, молекулярный докинг лигандов с белком gp120 и расчет свободной энергии связывания, генерацию молекулярных дескрипторов химических соединений обучающего набора данных, обучение нейронной сети, оценку результатов обучения и работы автоэнкодера. Рассмотрены результаты тестирования автоэнкодера на широком наборе соединений из молекулярной библиотеки ZINC. Показано, что совместное использование нейронной сети с виртуальным скринингом баз данных химических соединений формирует продуктивную платформу для идентификации базовых структур, перспективных для создания новых противовирусных препаратов, ингибирующих ранние стадии развития ВИЧ-инфекции.

Ключевые слова: методы глубокого обучения; генеративно-состязательный автоэнкодер; белок gp120; ингибиторы проникновения ВИЧ-1; методы молекулярного моделирования.

Современные методы компьютерного конструирования потенциальных лекарств значительно расширяют возможности фармацевтической индустрии, позволяя существенно сократить время и затраты, необходимые для создания новых терапевтических средств. Несмотря на то что эффективность компьютерных методов в создании лекарственных препаратов в настоящее время является общепризнанной, разработка новых математических подходов в сочетании с доступностью мощных и дешевых вычислительных ресурсов способствует их постоянному совершенствованию. За последние несколько лет идея использования технологий искусственного интеллекта для ускорения процесса создания новых лекарственных препаратов и повышения эффективности программ фармацевтических исследований стала особенно востребованной в области хемоинформатики.

Цель настоящей работы – методами глубокого обучения разработать генеративно-состязательную автоэнкодерную нейронную сеть для дизайна потенциальных ингибиторов ВИЧ-1, способных блокировать участок оболочки вируса, критический для его связывания с клеточным рецептором CD4.

Архитектура разработанного состязательного автоэнкодера состоит из двух нейросетей – автоэнкодера и дискриминатора, работающих во время обучения в соревновательном режиме. Автоэнкодер представляет собой семислойную нейронную сеть, имеющую входной и выходной слои, латентный слой, а также четыре полносвязных слоя (рис. 1).

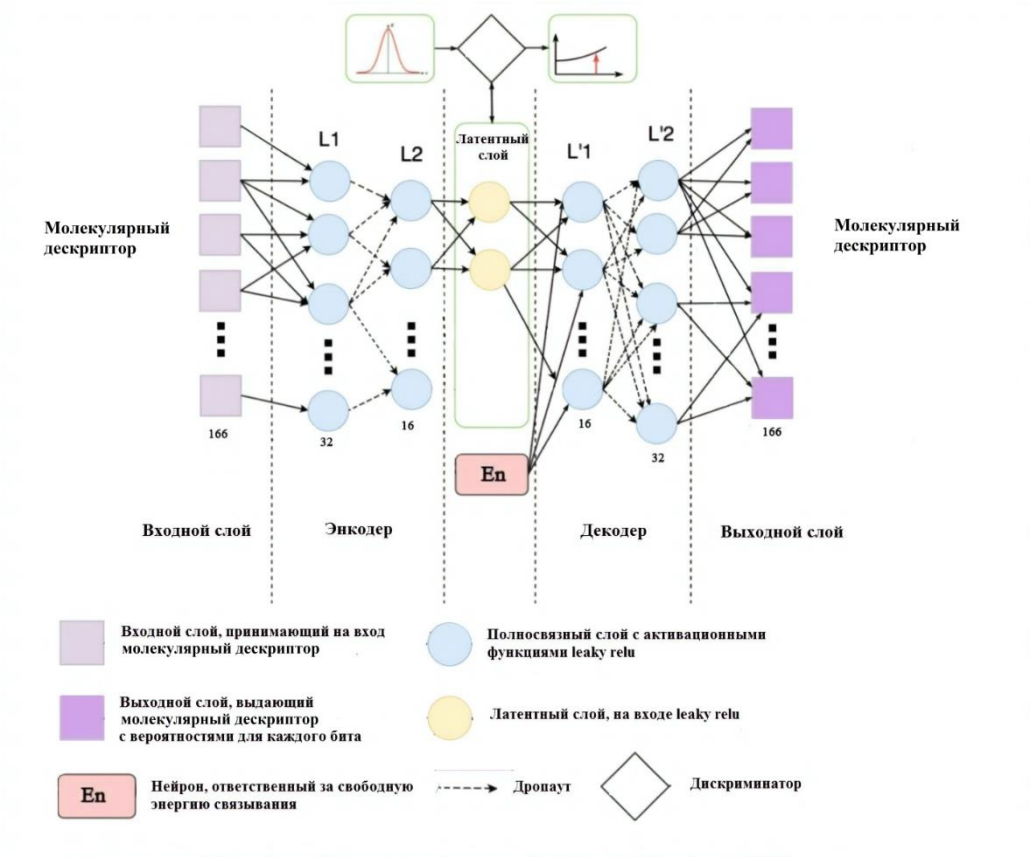

Рис. 1. Архитектура нейронной сети для генерации потенциальных ингибиторов ВИЧ-1, блокирующих CD4-связывающий сайт белка gp120 оболочки вируса

Дискриминатор и автоэнкодер обучались совместно в два этапа – реконструкции и регуляризации, выполняемые в каждом подмножестве из оригинальных данных. На этапе реконструкции автоэнкодер обновлял энкодер и декодер, чтобы минимизировать ошибку восстановления входных и выходных данных. На этапе регуляризации сначала обновлялась сеть дискриминатора, чтобы отличить истинные выборки (полученные с помощью генератора нормального распределения) от сжатых входных данных (данных на латентном слое, вычисленных автоэнкодером), а затем автоэнкодер обновлял свой энкодер, чтобы запутать сеть дискриминатора.

Для обучения автоэнкодера использовались следующие параметры: количество эпох для главной версии модели, используемой для генерации – 400; скорость обучения всего автоэнкодера на первой ступени итерации – 0,005; скорость обучения дискриминатора на второй ступени итерации – 0,001; скорость обучения энкодера на третьей ступени итерации – 0,005; параметр Batch size – 128; оптимизатор – метод Adam [1].

Формирование обучающего набора данных выполнено в рамках подхода, использующего методологию клик-химии [2] для генерации наиболее вероятных структур-кандидатов биологически активных со-

единений. В результате имитации реакции азид-алкинового циклоприсоединения для молекул из двух библиотек, сформированных из низкомолекулярных химических соединений (с молекулярной массой менее 250 Да) из базы данных ZINC [3], был получен набор из 120 000 соединений, удовлетворяющие «правилу пяти» Липинского. После оценки методом молекулярного докинга энергии связывания отобранных молекул с белком gp120 и генерации для них молекулярных дескрипторов MACCS, эти соединения были включены в обучающий набор данных.

Для тестирования работы автоэнкодера была создана библиотека молекулярных дескрипторов MACCS для 21 325 567 соединений из библиотеки Drug-Like базы данных ZINC [3] и рассчитаны пять молекулярных дескрипторов для сгенерированных автоэнкодером молекул при пороговом значении энергии связывания с белком gp120, равном –5 ккал/моль. В результате виртуального скрининга созданной библиотеки для каждой из этих молекул с подобными молекулярными дескрипторами были найдены лиганды, которые показаны в таблице.

Таблица

Результаты тестирования нейронной сети

Коды соединений в базе данных ZINC	Расстояние Хэмминга *R*	Энергия связывания, ккал/моль
ZINC000026430653	*R* = 3	–8,8
ZINC000037104033	*R* = 3	–7,1
ZINC000002786698	*R* = 4	–7,0
ZINC000055836809	*R* = 3	–6,9
ZINC000055843838	*R* = 4	–6,5
ZINC000026430653	*R* = 5	–8,8
ZINC000037104033	*R* = 6	–7,1
ZINC000002786698	*R* = 6	–7,0
ZINC000163393594	*R* = 4	–6,0
ZINC000128895014	*R* = 4	–5,9
ZINC000035245594	*R* = 6	–6,6
ZINC000163489237	*R* = 7	–6,6
ZINC000052221501	*R* = 7	–6,6
ZINC000600676089	*R* = 6	–6,4
ZINC000004006242	*R* = 7	–5,9
ZINC000026430653	*R* = 4	–8,8
ZINC000002786698	*R* = 5	–7,0
ZINC000055836809	*R* = 4	–6,8
ZINC000037104033	*R* = 4	–6,7
ZINC000685198234	*R* = 4	–6,0

ZINC000182934853	$R = 7$	–7,4
ZINC000771860139	$R = 5$	–6,6
ZINC000012991344	$R = 7$	–6,0
ZINC000128895014	$R = 7$	–6,0
ZINC000163393499	$R = 7$	–6,0

При этом в качестве меры подобия молекулярных дескрипторов использовалось расстояние Хэмминга.

Анализ результатов молекулярного докинга найденных соединений с белком gp120 показал (таблица), что совместное использование нейронной сети с виртуальным скринингом библиотеки молекулярных дескрипторов позволяет идентифицировать лиганды с более низкой по сравнению с заданным пороговым значением энергией связывания. Результаты исследования свидетельствуют о том, что разработанная нейронная сеть представляет собой эффективную математическую модель для виртуального скрининга баз данных химических соединений, направленного на поиск малых молекул с высоким сродством к белку gp120 и разработку на их основе новых анти-ВИЧ препаратов широкого спектра действия.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Kingma D. P., Ba J. L. Adam: A Method for Stochastic Optimization [Электронный ресурс] – Режим доступа: https://arxiv.org/pdf/1412.6980.pdf. Дата обращения: 13.02.2020.
2. Kolb H. C., Finn M. G., Sharpless K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions // Angew. Chem. Int. Ed. Engl. 2001. Vol. 40. № 11. P. 2004–2021.
3. Irwin J. J., Shoichet B. K. ZINC – a Free Database of Commercially Available Compounds for Virtual Screening // J. Chem. Inf. Model. 2005. Vol. 45. №1. P. 177–182.

МЕТОД ОЦЕНКИ ПОЛНОТЫ НУКЛЕОТИДНЫХ ДАННЫХ ДЛЯ СБОРКИ ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ НА ОСНОВЕ РАСЧЁТА ДОЛИ СМЕЖНЫХ КОНТИГОВ

М. А. Сиколенко[1], Р. С. Сергеев[2], Л. Н. Валентович[1, 3]

1 Белорусский государственный университет, Минск, Беларусь;
[2] Объединённый институт проблем информатики Национальной академии наук Беларуси, Минск, Беларусь;
[3] Институт микробиологии Национальной академии наук Беларуси, Минск, Беларусь;
E-mail: maximdeynonih@gmail.com