ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ГЕНЕРАЦИИ НОВЫХ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ ПРОНИКНОВЕНИЯ ВИЧ-1

Е. И. Мордань
Белорусский государственный университет, г. Минск;
yauhenmardan@gmail.com
науч. рук. – А.М. Андрианов, д-р. хим. наук

Работа посвящена разработке алгоритмов генерации новых потенциальных ингибиторов белка gp120 ВИЧ-1. Описывается процесс их создания, вариации и использования. В качестве методов исследования выбраны методы машинного обучения и молекулярного моделирования.

Ключевые слова: машинное обучение, нейронные сети, автоэнкодеры, LSTM сети, ингибиторы проникновения ВИЧ-1, молекулярное моделирование

Вирус иммунодефицита человека типа 1 (ВИЧ-1) поражает иммунные клетки человека, ослабляет иммунную систему и подвергает риску заражения другими инфекциями. Первые пациенты, вызывавшие подозрения на заражение ВИЧ-1, были зарегистрированы в Нью-Йорке и Лос-Анджелесе в 1980х годах прошлого столетия. С этого момента более 70 миллионов человек заразились ВИЧ инфекцией. Заражение ВИЧ происходит при попадании биологической жидкости зараженного человека в кровь здорового. В процессе заражения за механизм присоединения ВИЧ к клетке отвечает белок gp120 оболочки вируса. Этот белок связывается с рецептором CD-4 здоровой клетки. В настоящее время отсутствуют лицензированные лекарственные препараты, терапевтическое действие которых основано на блокаде участка белка gp120, ответственного за связывание вируса с клеточным рецептором CD4. В связи с этим возникает задача поиска ингибиторов проникновения – веществ, задерживающих или подавляющих течение физико-химических или физиологических процессов при проникновении ВИЧ-1 в клетку-мишень.

Предварительным или иногда альтернативным экспериментальному подходу к разработке новых соединений является метод компьютерной генерации. Он позволяет преодолеть физические и ресурсные ограничения высокозатратного экспериментального подхода. Некоторые компьютерные методы основаны на использовании нейронных сетей [1][2], в том числе алгоритмы, разработанные в данной работе.

В работе разработаны два алгоритма генерации новых потенциальных ингибиторов ВИЧ-1: алгоритмы без и с заданием порогового значения энергии связывания белка gp120 с потенциальным ингибитором ВИЧ-1. Идея алгоритма без порогового значения энергии связывания (алгоритм 1) следующая:

1. Исходное соединение – потенциальный ингибитор белка gp120, представленный в некотором формате, кодируется в латентный вектор;

2. В латентный вектор вносится некоторый шум;

3. Далее латентный вектор с шумом декодируется и получается новое соединение, отличное от исходной молекулы.

Основное отличие алгоритма с энергией связывания (алгоритм 2) заключается в том, что на шаге 2 задается некоторая пороговая энергия, вместе с которой латентный вектор декодируется в новое соединение. В качестве формата представления соединений в силу своей простоты был использован формат SMILES [3]. За кодирование-декодирование соединений в формате SMILES отвечает специальная модель нейронной сети – автоэнкодер [4]. Модель для алгоритма 2 представлена на рисунке 1. Мо-

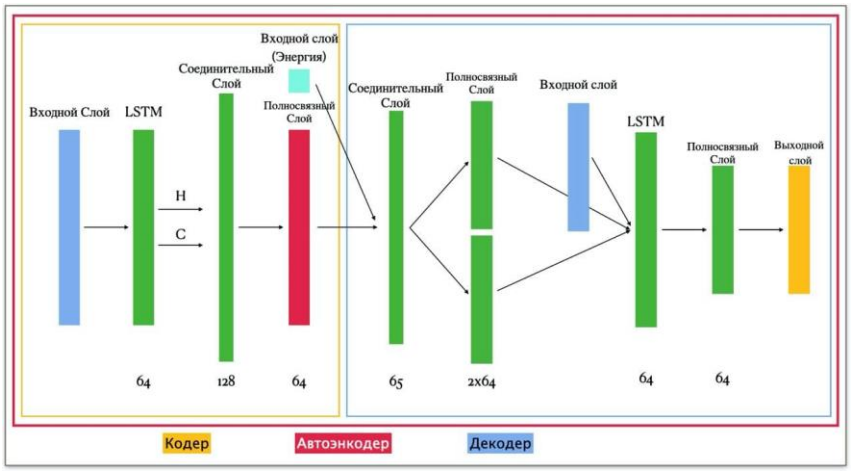

Рис. 1. Модель автоэнкодера с пороговой величиной энергии

дель же для алгоритма 1 отличается лишь отсутствием входного слоя для энергии (голубой цвет) и соединительного слоя (зеленый цвет) после полносвязного латентного слоя (красный цвет). Для полноценной работы двух алгоритмов после обучения соответствующих им моделей требуется их разбиение на кодер (желтая рамка) и декодер (синяя рамка). Кодер и декодер используются для кодирования исходного соединения и декодирования в новое соединение на шагах 1 и 3 соответственно. Данные модели автоэнкодоров используют в своей работе LSTM слои [5]. Такие слои помогают запомнить долгосрочный контекст SMILES-формул соединений и имеют преимущество в запоминании по сравнению с рекуррентными сетями и с сетями прямого распространения.

Кроме двух изложенных выше алгоритмов, для каждого из них был разработан особый режим генерации новых потенциальных ингибиторов. Такой режим не требует исходного соединения, задаваемого на шаге 1. Вместо этого латентный вектор сразу берется из некоторого распределения X и, если речь идет об алгоритме с энергией связывания, вместе с

пороговой энергией подается на декодер, в результате имея новое соединение. Для получения представления о распределении латентных векторов X была проанализирована их выборка, в результате чего было замечено, что некоторые компоненты латентных векторов либо всегда имеют нулевое значение, либо имеют распределение, напоминающее гамма-распределение с некоторыми параметрами. Таким образом, для получения нового латентного вектора из неизвестного распределения X, для всех ненулевых компонент берется значение из соответствующего им гамма-распределения, а для нулевых компонент – значение 0. Минусами такого режима работы является по-прежнему неизвестное, лишь приближенное распределение X. Для контроля данного распределения может быть использован дискриминатор. Однако в качестве плюсов стоит отметить почти бесконечное потенциальное количество новых латентных векторов, а, следовательно, и новых соединений, а также повышенную скорость работы ввиду отсутствия кодировщика.

В результате работы алгоритмов 1 и 2 были получены соответственно 1560 и 1662 новых уникальных соединения – потенциальных ингибиторов белка gp120 ВИЧ-1. Для каждого исходного соединения были идентифицированы три новых соединения. Далее для новых соединений был проведен молекулярный докинг и оценена энергия их связывания с белком gp120. Следует отметить, что в среднем значения энергии связывания для новых соединений лучше, чем для исходных молекул. Это же можно сказать и о значениях медиан. Всего 1090 из 1560 соединений алгоритма 1 и 1211 из 1692 соединений алгоритма проявляют энергию связывания с белком gp120 ниже, чем у исходных соединений.

Таким образом, в данной работе нами была поставлена задача по разработке алгоритма генерации новых потенциальных ингибиторов белка gp120 ВИЧ-1, блокирующих процесс прикрепления вируса к клетке хозяина. Разработка алгоритма идентификации проводилась с помощью методов машинного обучения и молекулярного моделирования. В рамках работы были разработаны два алгоритма генерации: не использующий и использующий данные о пороговой величине энергии связывания. Также был предложен вариант работы этих алгоритмов с генерацией ингибиторов из шума. В результате применения алгоритмов для генерации новых потенциальных ингибиторов проникновения ВИЧ-1 было идентифицировано более 3000 новых соединений. Для их первичного анализа был проведен молекулярный докинг и оценена энергия связывания сгенерированных соединений с белком gp120. В дальнейшем некоторые из полученных соединений могут быть более детально проанализированы и, в конечном итоге,

синтезированы и протестированы на противовирусную активность, разработанные алгоритмы могут быть использованы для поиска новых потенциальных ингибиторов проникновения ВИЧ-1.

Библиографические ссылки

1. Esben Jannik Bjerrum, Boris Sattarov. Improving Chemical Autoencoder Latent Space and Molecular De-novo Generation Diversity with Heteroencoders // Journal of Biomolecules 2018, 8(4), 131. arXiv: 1806.09300.
2. www.cheminformania.com - Chemical blog [Electronic resource] // - Mode of access: https://www.cheminformania.com/master-your-molecule-generator-seq2seq-rnnmodels-with-smiles-in-keras/. - Date of access: 5.04.2021.
3. Чумаков А.А., Слижов Ю.Г. Система SMILES-кодирования молекулярных структур и её применение для решения научно-исследовательских задач // Национальный исследовательский Томский государственный университет, 2017.
4. www.paperspace.com - Blog [Electronic resource] // - Mode of access: https:// blog.paperspace.com/autoencoder-image-compression-keras/. - Date of access: 03.05.2021
5. neurohive.io - Neural networks website [Electronic resource] // - Mode of access: https://neurohive.io/ru/osnovy-data-science/lstm-nejronnaja-set/. - Date of access: 03.12.2020.