УДК 004.82:16; 004.8.032.26; 547.9; 577.1

DE NOVO ГЕНЕРАЦИЯ ПОТЕНЦИАЛЬНЫХ ИНГИБИТОРОВ БЕЛКА KASA MYCOBACTERIUM TUBERCULOSIS

А. В. Гончар[1)], К. В. Фурс[1)], А. В. Тузиков[1)], А. М. Андрианов[2)]

[1)] Объединенный институт проблем информатики,
Национальная академия наук Беларуси,
Минск, Республика Беларусь, tuzikov@newman.bas-net.by
[2)] Институт биоорганической химии, Национальная академия наук Беларуси,
Минск, Республика Беларусь, alexande.andriano@yandex.ru

Разработана генеративная состязательная нейронная сеть с частичным привлечением учителя, обученная на графовых эмбеддингах и использованная для генерации потенциальных ингибиторов фермента KasA микобактерии туберкулеза. Проведена оценка потенциала ингибиторной активности новых сгенерированных нейронной сетью соединений против белка KasA методами молекулярного моделирования. На основе анализа полученных данных были идентифицированы шесть соединений, перспективных для создания новых эффективных препаратов для терапии лекарственно-устойчивых форм туберкулеза.

Ключевые слова: Микобактерия туберкулеза; белок KasA; генеративная состязательная нейронная сеть; графовые эмбеддинги; виртуальный скрининг; молекулярный докинг; молекулярная динамика.

DE NOVO GENERATION OF POTENTIAL INHIBITORS OF THE KasA PROTEIN OF MYCOBACTERIUM TUBERCULOSIS

H. V. Hanchar[a)], K. V. Furs[a)], A. V. Tuzikov[a)], A. M. Andrianov[b)]

[a)] United Institute of Informatics Problems, National Academy of Sciences of Belarus,
Minsk, Republic of Belarus, tuzikov@newman.bas-net.by
[b)] Institute of Bioorganic Chemistry, National Academy of Sciences of Belarus,
Minsk, Republic of Belarus, alexande.andriano@yandex.ru

A semi-supervised generative adversarial network was developed, trained on graph embeddings, and applied to the generation of potential inhibitors of the KasA enzyme of Mycobacterium tuberculosis. The inhibitory potential of the newly generated compounds against the KasA protein was evaluated using molecular modeling methods. Based on the analysis of the obtained data, six compounds were identified as promising candidates for the development of new effective drugs for the treatment of drug-resistant forms of tuberculosis.

Keywords: Mycobacterium tuberculosis; KasA protein; generative adversarial network; graph embeddings; virtual screening; molecular docking; molecular dynamics.

1. Введение

Туберкулез (ТБ) по-прежнему входит в число ведущих причин смертности во всем мире и представляет особую угрозу для пациентов с ВИЧ и сахарным диабетом [1]. Рост лекарственной устойчивости существенно осложняет терапию и делает актуальными исследования по созданию новых противотуберкулезных препаратов. На ранних стадиях разработки лекарств все более востребованы технологии виртуального скрининга и методы машинного обучения, позволяющие ускорить поиск активных соединений, сократить время и затраты, необходимые для их создания.

Среди потенциальных мишеней особое значение имеет фермент KasA, играющий ключевую роль в синтезе жирных кислот клеточной стенки микобактерии туберкулеза (МБТ). Известно, что потеря активности белка KasA приводит к лизису клеток бактерии, свидетельствуя о том, что этот фермент имеет ключевое значение для жизненного цикла МБТ и, следовательно, является важной терапевтической мишенью для разработки новых эффективных ингибиторов лекарственно-устойчивого ТБ [2].

Цель настоящего исследования заключалась в создании генеративной модели нейронной сети для de novo дизайна малых молекул, потенциально активных против фермента KasA МБТ. Для достижения этой цели были проведены исследования, которые включали:

- разработку архитектуры нейронной сети для генерации низкомолекулярных химических соединений, обладающих высоким сродством к белку KasA;
- формирование обучающей библиотеки малых молекул, содержащих элементы структуры, способные к селективным взаимодействиям с активным центром фермента;
- обучение модели и ее тестирование на наборе соединений из созданной молекулярной библиотеки;
- генерацию новых молекул, потенциально активных против белка KasA;
- молекулярный докинг сгенерированных молекул с целевым белком;
- молекулярную динамику перспективных соединений;
- оценку физико-химических свойств лучших соединений и отбор потенциальных ингибиторов МБТ.

2. Материалы и методы

Для решения поставленных задач нами была разработана модель генеративной состязательной нейронной сети с частичным привлечением учителя (SGAN, Semi-Supervised Generative Adversarial Network) (рис. 1), которая использует графовые эмбеддинги, полученные из латентного

пространства вариационного автоэнкодера JTVAE (Junction Tree Variational Autoencoder) [3]. Обучение SGAN проводили на наборе молекул из обучающей выборки с использованием в неявном виде значений энергии связывания с белком KasA. Молекулы в обучающем наборе данных разделяли на две группы, включавшие соответственно соединения с низкими (ниже –8,2 ккал/моль) и высокими значениями энергии связывания с целевым белком, рассчитанными с помощью методов молекулярного докинга. Это позволило SGAN генерировать новые молекулы, похожие на соединения с высоким сродством к мишени.

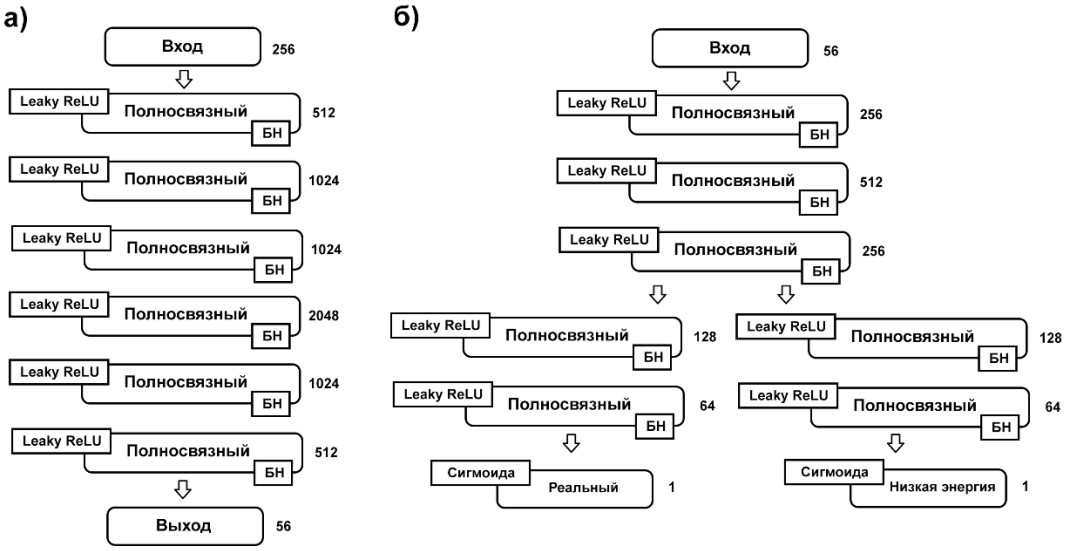

Рис. 1. Архитектура SGAN:
а – генератор; *б* – дискриминатор

Формирование обучающего набора данных

Для формирования обучающего набора данных был проведен виртуальный скрининг библиотек Zinc15, ChemSpace и ChemDiv на веб-сервере Pharmit с использованием трех фармакофорных моделей, построенных на основе комплексов белка C171Q KasA с ингибиторами TLM и TLM5 (PDB ID: 4C6X и 4C72 соответственно). С использованием Python 3 и программного пакета RDKit (https://www.rdkit.org) из отобранных молекул были удалены дубликаты, и для каждого соединения получены канонические представления SMILES. В результате размер обучающей молекулярной библиотеки, сформированной с помощью фармакофорного анализа баз данных веб-сервера Pharmit, составил 58 815.

Для генерации графовых эмбеддингов, необходимых для обучения SGAN, из формата SMILES, полученного на этапе фармакофорного поиска, использовали предобученную модель нейронной сети JTVAE,

которую разработчики обучали на выборке из 250 000 соединений, отобранных случайным образом в библиотеке Zinc15.

Для расчета значений свободной энергии связывания молекул из обучающего набора данных с целевым белком с помощью программы QuickVina2 был проведен молекулярный докинг этих лигандов с ферментом C171Q KasA в свободном состоянии. Структура белка извлекалась из комплекса C171Q KasA в кристалле с ингибитором TLM5 (ID PDB: 4C72). Структуры белка и лигандов были подготовлены к расчетам с помощью программного пакета MGLTools (https://ccsb.scripps.edu/mgltools/). Ячейка для докинга включала малонил-связывающий сайт KasA и имела следующие размеры: ΔX = 20,67 Å, ΔY = 24,8 Å, ΔZ = 16,46 Å с центром X = = – 7,24 Å, Y = –19,9 Å, Z = 6,75 Å. Значение параметра охвата конформационного пространства, определяющего широту поиска, было установлено равным 100.

Обучение SGAN

Для 58815 малых молекул из сформированного обучающего набора данных с помощью сети JTVAE были получены графовые эмбеддинги, а затем эта виртуальная библиотека была разделена на две выборки – тренировочную и тестовую, состоящие из 47052 и 11763 векторов соответственно. Путем многочисленных экспериментов параметр, задающий количество эпох обучения, был выбран равным 150. Отношение частоты обучения генератора к дискриминатору было выбрано равными 0,3 к 0,7.

Функции потерь дискриминатора и генератора представлены в формулах (1) и (2):

$$D_{loss} = BCE\left(D_{out1}\left(G(noise)\right),0\right) + BCE\left(D_{out2}(real\_data), energy\_class\right), \quad (1)$$

$$G_{loss} = BCE\left(D_{out1}\left(G(noise)\right),1\right), \quad (2)$$

где BCE – функция потерь бинарная кросс-энтропия; D – дискриминатор; D_{out1} – выход дискриминатора, отвечающий за предсказание реальности молекулы; D_{out2} – выход дискриминатора, отвечающий за предсказание класса энергии молекулы; G – генератор; $noise$ – вектор гауссовского шума размерности 256; $real\_data$ – вектор, соответствующий молекуле из тренировочной выборки; $energy\_class$ – 1 или 0 в зависимости от того, принадлежит ли молекула классу с низкой энергией связывания или нет.

Графики функций потерь генератора и дискриминатора SGAN представлены на рис. 2. Сплошная синяя линия соответствует потерям генератора, а красная пунктирная линия – потерям дискриминатора. Анализ

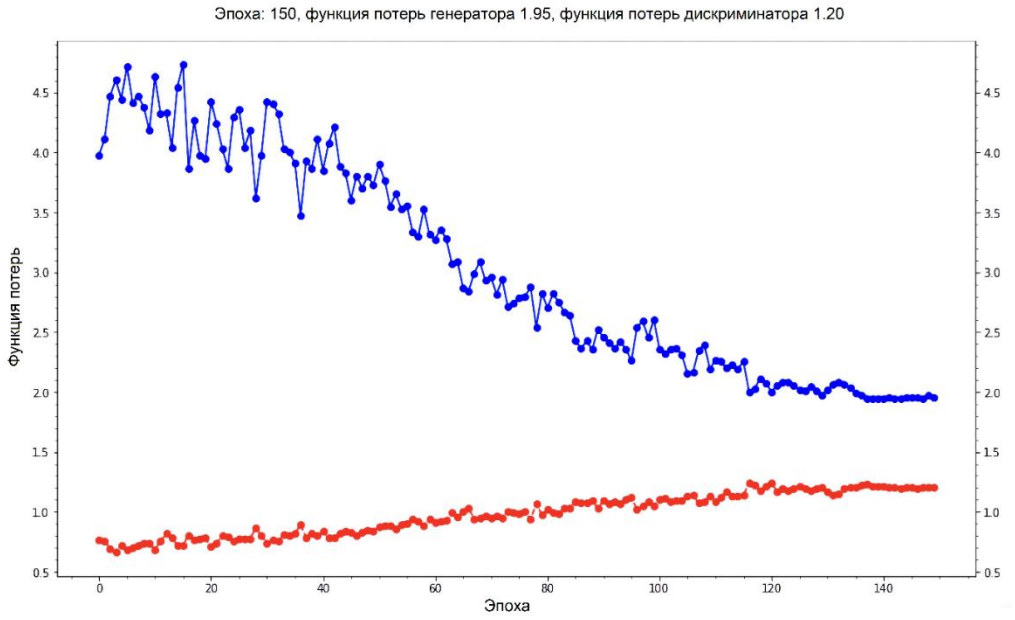

Рис. 2. Функции потерь генератора (синяя линия)
и дискриминатора (красная линия) SGAN

графиков функций потерь дает основания полагать, что модель была обучена корректно, так как функции потерь сходятся, но не пересекаются.

На основе случайных векторов длины 256, имеющих гауссовское распределение, с помощью генератора SGAN было сгенерировано 200 000 векторов длиной 56 и получены соответствующие предсказания дискриминатора. Для оценки генеративных возможностей SGAN были выбраны два региона векторов. Первый (перспективный) регион включал вектора с вероятностью реальности выше 50% и вероятностью низкого энергетического класса выше 50%, а второй регион представлял область с вероятностью реальности молекул выше 50% и вероятностью низкого энергетического класса ниже 50%. Всего было получено 2565 и 2745 графовых эмбеддингов для первого и второго регионов соответственно, которые после декодирования в JTVAE, удаления дубликатов и проверки корректности дали 1755 и 1882 уникальные молекулы.

Для выбора наиболее перспективных лигандов среди всех сгенерированных SGAN соединений (из обеих областей), значения свободной энергии связывания были переоценены с использованием оценочных функций RFScore-4 и NNScore 2.0. На основе предсказанных с помощью оценочных функций AutoDock Vina, RFScore-4, NNScore 2.0 значений энергии связывания для каждого соединения был рассчитан экспоненциальный консенсусный ранг (ECR) [4] и отобраны 20 лучших соединений для молекулярно-динамических расчетов.

Молекулярно-динамическое (МД) моделирование комплексов лиганд/белок проводили в воде с использованием Amber18 [5] и силовых

полей Amber ff14SB (белок) и GAFF (лиганды). После минимизации, постепенного нагрева до 300 K, уравновешивания давления и достижения равновесия система моделировалась в ансамбле NPT при 300 K и 1 атм в течение 200 нс. В качестве контрольного соединения использовали ингибитор KasA TLM5 (Kd = 25,6 мкМ).

Свободную энергию связывания рассчитывали методом MM/GBSA для 40 комплексов, извлечённых из последних 80 нс траекторий (шаг 2 нс). Энтропийный вклад определяли с помощью Nmode из AmberTools19, а анализ траекторий проводили модулем CPPTRAJ.

3. Результаты и обсуждение

Результаты молекулярного докинга для новых сгенерированных соединений

Для оценки потенциала разработанной модели нейронной сети были получены прогнозные показатели дискриминатора для 11 763 векторов из тестового набора данных, из которых в перспективную и вторую области вошли соответственно 3438 и 4448 молекулярных эмбеддингов. Для сгенерированных молекул и соединений из тестового набора был проведен молекулярный докинг с белком KasA с использованием того же вычислительного протокола, что и для соединений из обучающего набора данных. Результаты молекулярного докинга представлены в табл. 1.

Таблица 1

Результаты молекулярного докинга для сгенерированных соединений и соединений из тестовой выборки

1 – перспективная область 2 – вторая область		Количество соединений	Процент соединений с низкой энергией связывания (< –8,2 ккал/моль)	Средняя энергия связывания, (ккал/моль)
Сгенерированные соединения	1	1755	67	–8,5
	2	1882	40	–8,0
Соединения из тестовой выборки	1	3438	76	–8,7
	2	4448	21	–7,6

Данные табл. 1 указывают на то, что дискриминатор SGAN может с высокой точностью предсказывать класс энергии на реальных (не сгенерированных) данных, что подтверждает корректность обучения модели. Полученные результаты свидетельствуют о способности сети SGAN генерировать различные химические соединения с высоким сродством к белку KasA, особенно в перспективном регионе. Эти данные также указывают на то, что разделение молекул на классы на основе энергии связывания при обучении модели является эффективной стратегией для создания соединений с высоким сродством к целевому белку.

Анализ полученных траекторий молекулярной динамики позволил идентифицировать 12 соединений, у которых значения свободной энергии связывания (ΔG) оказались ниже, чем у контрольного соединения TLM5, рассчитанного по тому же вычислительному протоколу. Ниже представлены значения свободной энергии связывания и их стандартные отклонения по результатам молекулярной динамики для шести наиболее перспективных соединений (рис. 3), физико-химические параметры которых представлены в табл. 2. Учитывая стандартную ошибку метода MM/GBSA, составляющую около 2,9 ккал/моль, полученные данные указывают на то, что сродство этих соединений к KasA значительно выше по сравнению с ингибитором KasA TLM5. Также из табл. 2 видно, что отобранные соединения полностью соответствуют требованиям, предъявляемым к потенциальному лекарству согласно «правилу пяти» Липинского [6].

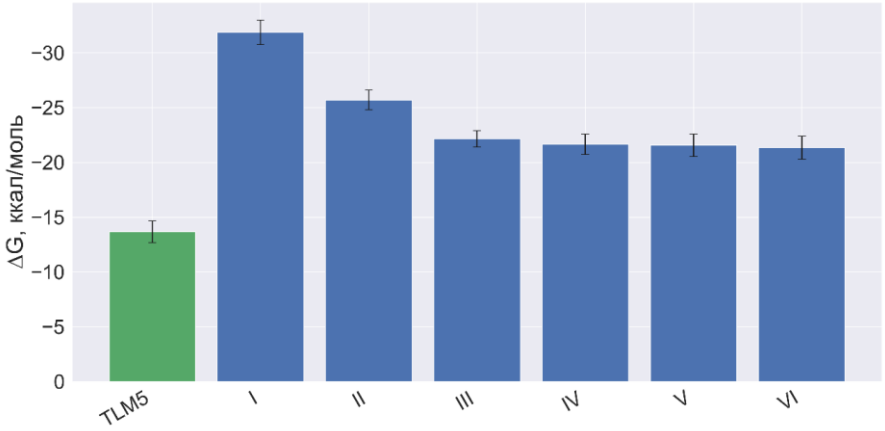

Рис. 3. Средние значения ΔG, рассчитанные для динамических моделей комплексов лиганд/KasA

Таблица 2

Физико-химические параметры идентифицированных соединений, связанные с «правилом пяти» Липинского [2; 4], и их синтетическая доступность

Лиганд	Химическая формула	Молекулярная масса, Да	$LogP_{o/w}$	Синтетическая доступность	Число доноров водородной связи	Число акцепторов водородной связи
I	$C_{23}H_{29}N_3O_4S$	443,56	3,14	5,15	1	5
II	$C_{23}H_{30}N_6O_3$	438,52	2,58	4,22	0	6
IV	$C_{18}H_{22}N_6O_2S$	386,47	1,78	4,06	1	5
V	$C_{19}H_{14}FN_5O_2S$	395,41	2,81	3,47	1	6
III	$C_{23}H_{26}N_3O_3+$	392,47	2,69	3,8	2	5
VI	$C_{23}H_{35}N_5O_2$	413,56	3,22	4,95	1	5

Примечание. Приведенные данные получены с помощью веб-сервера открытого доступа SwissADME (http://www.swissadme.ch); $LogP_{o/w}$ – липофильность соединения.

4. Заключение

В ходе проведенного исследования была разработана генеративная состязательная нейронная сеть (SGAN) с частичным привлечением учителя, адаптированная для поиска потенциальных ингибиторов фермента KasA, являющегося ключевым элементом биосинтеза клеточной стенки микобактерии туберкулеза.

Для реализации проекта была сформирована молекулярная библиотека соединений, содержащая низкомолекулярные аналоги известных ингибиторов KasA. В процессе исследования проведена оценка генеративных возможностей модели, выполнен молекулярный докинг сгенерированных соединений, а также оценка их сродства к мишени с использованием оценочных функций AutoDock Vina, RFScore-4 и NNScore 2.0. Проведенные МД расчеты подтвердили стабильность взаимодействия перспективных соединений с ферментом KasA, что делает их перспективными кандидатами на дальнейшие экспериментальные исследования.

Результаты работы позволили выделить шесть соединений, обладающих наилучшими характеристиками по сродству к KasA и удовлетворяющих требованиям фармакокинетики. Эти соединения могут стать основой для дальнейших экспериментальных исследований, направленных на разработку новых эффективных препаратов для терапии лекарственно-устойчивых форм туберкулеза.

Работа выполнена при поддержке Консорциума и Программы по разработке портала по лекарственно устойчивому туберкулезу (https://tbportals.niaid.nih.gov) и поддержана Белорусским республиканским фондом фундаментальных исследований (проекты Ф23-007, Ф24КИТГ-016).

Библиографические ссылки

1. WHO. Global tuberculosis report 2024. World Health Organization, 2024. URL: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024 (date of access 01.09.2025).

2. Identification of KasA as the cellular target of an anti-tubercular scaffold / K. A. Abrahams [et al.] // Nat. Commun. 2016. Vol. 7. Article no. 12581. DOI: 10.1038/ncomms12581.

3. *Jin W., Barzilay R., Jaakkola T.* Junction tree variational autoencoder for molecular graph generation // International conference on machine learning. 2018. P. 2323–2332.

4. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking / K. Palacio-Rodriguez [et al.] // Sci. Rep. 2019. Vol. 9. Article no. 5142.

5. AMBER 2018 / D. A. Case [et al.]. San Francisco : University of California, 2018.

6. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings / C. A. Lipinski [et al.] // Advanced Drug Delivery Reviews. 2001. Vol. 46, No. 1–3. P. 3–26. DOI: 10.1016/s0169-409x(00)00129-0.