Интеллектуальная среда прогнозирования свойств антибиотиков по их химической структуре

Т. Б. Чистякова¹, Р. В. Макарук², Э. Э. Мусаев³

Санкт-Петербургский государственный технологический институт (технический университет) 1 chistb@mail.ru, 2 makaruk_rv@mail.ru, 3 eldarm@uw.edu

Аннотация. В статье предложена функциональная структура интеллектуальной среды прогнозирования свойств антибиотиков по их химической структуре, позволяющая прогнозировать их основные полезные свойства и токсичность. В состав системы входят две нейронные сети: для прогнозирования токсичности применяется сверточная сеть с прямой связью, где молекула антибиотика описывается с помощью молекулярных дескрипторов; для прогнозирования полезных свойств антибиотика используется рекурентная нейронная сеть с нейронами GRU, на вход которой подаётся структура молекулы в нотации SMILES учитывающая расположение элементов молекулы и связи между ними в векторном виде. В состав среды также входит база знаний, позволяющая исследователю осуществлять выбор антибиотика направление его синтеза в зависимости от структуры, что в свою очередь позволяет значительно снизить стоимость новых антибиотиков. разработанной среды протестирована на примере противогрибковых полиеновых макролидных антибиотиков.

Ключевые слова: интеллектуальная среда; нейронные сети; база знаний; прогнозирование; антибиотики; words2vec; QSAR

І. Введение

Синтез новых антибиотиков является одним из основных направлений работы современной медицины. Это связано с тем, что патогены становятся более устойчивыми к использующимся против них препаратам. Созданию новых антибиотиков препятствуют высокая стоимость создания новых лекарственных препаратов, а также вероятность их получения с высокой токсичностью и/или плохой связываемостью с целевым белком, влияние на который и оказывает антибиотическое действие лекарства.

В настоящее время борьба с грибковыми инфекциями является одной из важнейших проблем здравоохранения. Среди грибковых заболеваний инвазивные микозы (заболевания, вызванные грибами) становятся все более актуальной проблемой практической медицины, что связано с увеличением численности иммуноскомпроментированных популяций пациентов. Ввиду этого эффективное применение антифунгальных

В. В. Белахов

Технион – Израильский институт технологии, Хайфа, Израиль chvalery@technion.ac.il

(противогрибковых) препаратов, и прежде всего полиеновых макролидных антибиотиков (ПМА), в терапии различных форм микозов является важным фактором борьбы с грибковыми инфекциями [1].

ПМА, применяющиеся в медицинской практике, не в полной мере удовлетворяют потребности врачей-клиницистов из-за их ограниченной эффективности (сильная токсичность, невысокая антифунгальная активность и т.д.). Ввиду этого поиск новых производных ПМА (ППМА) с улучшенными медико-биологическими свойствами является актуальной задачей для химиковбиологов, однако это долгий и дорогостоящий процесс, связанный с различными рисками [1].

Развитие медицины И применение информационных технологий привело к появлению и стремительному совершенствованию различных моделей прогнозирования химических и фармацевтических свойств синтезируемых препаратов, накоплению медицинских баз данных и повышению их доступности в открытых государственных и коммерческих источниках. Используя эту информацию, а также опыт разработчиков системы в синтезе и модификации ППМА [1-4], нами разработана интеллектуальная среда прогнозирования антибиотиков по их химической структуре, включающая такие модели представления знаний как продукционнофреймовая и нейронные сети, позволяющие снизить время и стоимость синтеза ППМА путем прогнозирования таких параметров как токсичность и полезные свойства по заданной структуре молекулы.

II. РЕШЕНИЕ ПРОБЛЕМЫ

Антибиотик, с точки зрения его использования, можно описать с помощью следующего вектора: $A = \{LD_{50}, Ab_{into}B, Se, R\}$, где LD_{50} – уровень острой токсичности LD_{50} (мг/кг, средняя доза вещества, вызывающая 50%-ю гибель лабораторных животных при изучении токсичности соединений), Ab_{int} — всасываемость из желудочнокишечного тракта, B — белки, против которых направлено действие антибиотика, Se — побочные действия, R — виды микроорганизмов с известной резистентностью к данному виду антибиотика. Одними из важнейших характеристик антибиотика являются LD_{50} и B.

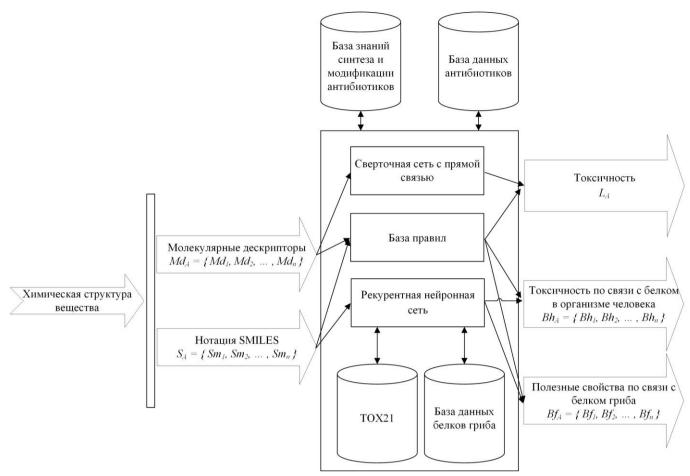


Рис. 1. Структура интеллектуальной среды

Токсичность (LD_{50}) можно описать как связь молекулы (в нашем случае, противогрибкового антибиотика) с белком, или иной молекулой в человеческом теле. При связи с белком, антибиотик нарушает его поведение. Для макролидных противогрибковых антибиотиков, нарушение поведения, главным образом, происходит в результате того, что макролидные противогрибковые антибиотики могут занять место фосфолипида в мембране человеческой клетки, нарушить плотность мембраны, и таким образом привести к потере ионов вне клетки. Противогрибковое свойство опирается на тот же самый механизм, но вместо фосфолипидов человеческой клетки, антимикотик связывается с мембраной клетки грибка, которая состоит из эргостерола. Присутствие большого липофильного участка во всех полиеновых макролидных антибиотиках позволяет им тесно связаться эргостеролом, и нарушать структуру клеточной мембраны грибка.

Так как эргостерол выполняет роль, которую в человеческом теле играют фосфолипиды, задача определения токсичности для человека и противогрибковой активности тесно связана. Проблема синтеза макролидных противогрибковых лекарств состоит в том, чтобы найти такую структуру $\Pi\Pi MA$, которая связывалась бы с эргостеролом (элемент B в векторе A), но не с фосфолипидами мембраны клеток человека. Другие

противогрибковые антибиотики связываются с белками грибковой клетки, и таким образом останавливают синтез элементов стенки мембраны (глюканов составляющих стенку клетки для эхинокандинов, эргостерола составляющего мембрану клетки для азолов) [3, 5].

Целью разработанной среды является прогноз свойств антибиотика А. Используя молекулярную структуру разработанная противогрибкового препарата, интеллектуальная среда прогнозирует токсичность (LD_{50}) и полезные свойства (*B*) синтезируемого модифицируемого антимикотика. Среда включает в себя следующие модули прогнозирования: токсичности по молекулярным дескрипторам; связи молекул с целевыми белками, для прогнозирования полезных свойств (связь с целевым белком гриба) и дополнительно токсичности (связь с белками организма человека). Кроме того, система себя базу знаний, позволяющую включает В исследователю осуществлять выбор антибиотика и направление его синтеза в зависимости от химической структуры вещества и его прогнозируемых свойств. Модуль базы знаний включает в себя накопленные мировой опыт и экспертные знания в области синтеза и модификации антибиотиков, что позволяет уточнить прогноз свойств получаемого препарата. Общая схема разрабатываемой системы описана в [2], а описываемая подсистема предоставлена на рисунке.

Разработанная интеллектуальная среда использует открытые государственные источники химикофармацевтических данных о лекарственных препаратах. Полученная из этих баз данных информация приводится к двум используемым в среде форматам, которые подаются на вход нейронным сетям и базе правил. Первый формат, это молекулярные дескрипторы, которые в векторном виде имеют следующий вид: $Md_A = \{ Md_1, Md_2, ..., Md_n \}$, где n= 99, по количеству дескрипторов, Md_n – молекулярный дескриптор [2]. Md_A формируется с использованием Python-библиотеки для прикладной информатики в химии RDKit [2]. Второй формат представляет химическую структуру антибиотика в нотации SMILES (Simplified Molecular Input Line Entry System, система упрощённого представления молекул в строке ввода), которая предварительно обрабатывается алгоритмом Word2Vec (программный инструмент анализа семантики естественных языков) [6], но для элементов, составляющих запись SMILES, вместо слов естественного языка. Таким образом второй используемый формат в векторной форме имеет следующий вид: $S_A = \{ Sm_1, Sm_2, ..., Sm_n \}$, где n = 15(количество свойств в векторном формате Word2Vec), $Sm_n = \{ Sv_1, Sv_2, ..., Sv_j \}$ — записи в векторном формате нотации SMILES, $z \partial e j = 250$ (максимальная длина записи), Sv_i – координата элемента.

Использование формы нотации SMILES [7] вызвано сложностью ввода химических структур молекул в нейронную сеть, а также в связи с тем, что химическая простой форме формула В (например, противогрибкового антибиотика нистатина, C47H75NO17) не предоставляет достаточно информации о связи между атомами, и не является уникальной (множество молекул могут иметь одну и ту же химическую формулу). Ценность канонической нотации SMILES в том, что предоставляет уникальную, однозначную форму записи (одна форма для каждой молекулы, например для нистатина,

C[C@@H]IOC(=O)C[C@H](O)C[C@H](O)C[C@H](O)CC[C@@H](O)[C@H](O)C[C@H](O)[C@H](O)[C@H](C]C@H](O)[C@H](C]C@H](O)[C@H](O)[C@H](N)[C@H]3O](C=C)(C=C)(C=C)(C=C)[C@H](C)[C@H](C)[C@H](O)[C@H](C)[C@H](O)[C@H]1C)O2)C(=O)O). Недостатком SMILES является разделение атомов углевода и азота в ароматическом кольце и вне ароматического кольца [8].

Кроме стандартных молекулярных дескрипторов, доступны два дополнительных вида представления молекулы: химическая формула и рисунок. Хотя рисунок является стандартизированным, различия в размере рисунка ограничивают возможность применения сверхточных нейронных сетей, так как те обычно используют рисунки одинакового размера.

Химические формулы в форме нотации SMILES преобразуются в вектора (атомов в химической формуле) в форме, образованной статистически по элементам окружающим его. В обработке естественного языка, этот подход известен как Words2Vec [6]. Статистическое определение местоположения каждого элемента в векторном пространстве возможно определять двумя способами: «Continuous-bag-of-words» (CBOW) и «Skip-

Grams». «Skip-grams» тренирует небольшую нейронную сеть предсказывать элементы, окружающие выбранный, а СВОW тренирует предсказывать текущий используя окружающие его элементы. применяются оба алгоритма для решения проблемы применения и нормализации функции активации Softmax для большого количества выходных параметров. Такой подход уместен, когда количество выходных параметров является количеством слов в естественном языке, но не является проблемой для небольшого количества возможных составляющих химической формы записи. Поэтому в интеллектуальной среде используется только метод «Skip-Gram».

Модель Words2Vec для SMILES состоит из трех слоев, слоя внедрения размером 1x78 нейронов (количество уникальных свойств в форме записи SMILES), скрытого слоя размером 15x78 нейрона, и слоя вывода размером 1x78 нейронов. Размер движущегося окна — 11 элементов (5 с каждой стороны текущего). Скрытый слой предоставляет векторное представление каждого элемента формы записи SMILES.

В среде используются две нейронные сети. Первая прогнозирует токсичность L_A (LD_{50}) по молекулярным дескрипторам Md_A [2]. Вторая прогнозирует связь антибиотика по химической структуре в форме нотации SMILES с белком в: организме человека $Bh_A = \{Bh_I, Bh_2, \dots, Bh_n\}$, где n=2I, по количеству белков в базе TOX21 (https://ntp.niehs.nih.gov/results/tox21/index.html), Bh_n — наличие или отсутствие связи с n-ым белком; клетке грибка $Bf_A = \{Bf_I, Bf_2, \dots, Bf_n\}$, где n — количество белков, повреждение которых убивает грибок, Bf_n — наличие или отсутствие связи с n-ым белком. Вывод обеих нейронных сетей может быть уточнён с помощью базы знаний в области синтеза и модификации антибиотиков.

Для прогнозирования полезных свойств антибиотика (В) используется рекурентная нейронная сеть с нейронами GRU (Gated Recurrent Units, управляемые рекуррентные блоки), с существующей архитектурой, которая состоит из одного вводного слоя размером на 50 нейронов, двух рекуррентных слоев размером на 224 и 384 GRU нейрона, и вывод, который для TOX21 состоял из 21 нейрона [9]. Такой подход обоснован мировой практикой применения нейронных сетей для прогнозирования химических свойств [10], а также развитием подходов для обработки данных сфере последовательных В обработки естественного языка, которые применимы для обработки химической формулы в форме нотации SMILES.

В сфере медицинских информационных систем разработано большое количество разнообразных баз данных, которые позволяют получить информацию о медицинских свойствах различных препаратов, в том числе и противогрибковых антибиотиках. Химические формулы в форме нотации SMILES были получены из базы данных Европейской Лаборатории Молекулярной Биологии (ChEMBL) [11].

В качестве исходных данных для обучения и тестирования нейронной сети для прогнозирования полезных свойств антибиотика использовались базы

данных ChEMBL (1727081 молекул с известной структурой и свойствами), набор данных TOX21 (10,000 молекул), а также полученные авторами ППМА [1, 3, 7, 12].

III. ЗАКЛЮЧЕНИЕ

Тестирование разработанной интеллектуальной среды прогнозирования свойств антибиотиков по их химической структуре на примере ПМА нистатина (nystatin, https://www.drugbank.ca/drugs/DB00646) показало состоятельность такого подхода и возможность его использования для прогнозирования антифунгальных свойств и токсичности препарата в результате связи препаратов с белками в клетках гриба и организме человека.

Разработанная интеллектуальная среда прогнозирования свойств антибиотиков по их химической структуре направлена на интеграцию в единый программный комплекс подсистемы интеллектуального анализа и оценки характеристик антимикотика применяя такие модели представления знаний как продукционнофреймовая и нейронные сети. Среда обладает гибкостью и дополняемостью, позволяет обучать разработанные нейронные сети с помощью синтезированных ППМА, добавлять новые антибиотики, что в результате позволяет снизить стоимость и время синтеза новых ППМА.

Список литературы

[1] Белахов В.В., Колодязная В.А., Гарабаджиу А.В., Чистякова Т.Б., Смирнов И.А. Диалкил(диарил)амидофосфатные производные тетраенового макролидного антибиотика люцензомицина: получение, противогрибковая активность, разработка интеллектуальной информационной системы // Успехи медицинской микологии. 2016. Т. XVI: Материалы мемориальной конференции по медицинской микологии. Москва: издательство Национальной Академии Микологии. С. 114–118. URL: http://elibrary.ru/download/34337136.pdf.

- [2] Chistyakova T.B., Belakhov V.V., Makaruk R.V., Musayev, E.E. Computer-Aided solution for intellectual analysis and judicious selection of medically advanced antifungals synthesis conditions // Proceedings of the XIX IEEE International Conference on Soft Computing and Measurements (SCM'2017). SPb.: SPb Electrotechn. Univ., 2017. P. 516-518
- [3] V. Fanos & L. Cataldi (2000). Amphotericin B-Induced Nephrotoxicity: A Review. Journal of Chemotherapy, 12, 463-470. (doi: 10.1179/joc.2000.12.6.463)
- [4] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, Wang J, Yu B, Zhang J, Bryant SH. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4; 44(D1):D1202-13. Epub 2015 Sep 22 [PubMed PMID: 26400175] doi: 10.1093/nar/gkv951.
- [5] "Frank C. Odds, Alistair J.P. Brown & Neil A.R. Gow" ("2003"). "Antifungal agents: mechanisms of action". "Trends in Microbiology", "11", "272-279". (doi: 10.1016/S0966-842X(03)00117-3)
- [6] Mikolov, Tomas; et al. "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781
- [7] Weininger, David (1988). SMILES, a chemical language and information system. 1. Introduction to methodolo-gy and encoding rules. Journal of Chemical Information and Computer Sciences, 28, 31-36.
- [8] O'Boyle, N. M. (2012). Towards a Universal SMILES representation -A standard method to generate canonical SMILES based on the InChI. Journal of Cheminformatics, 4, 22. http://doi.org/10.1186/1758-2946-4-22
- [9] B. Goh, Garrett & O. Hodas, Nathan & Siegel, Charles & Vishnu, Abhinav. (2017). SMILES2Vec: An Interpreta-ble General-Purpose Deep Neural Network for Predicting Chemical Properties.
- [10] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules. arXiv preprint arXiv:1703.07076.
- [11] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos and J.P. Overington (2014) 'The ChEMBL bioactivity database: an update.' Nucleic Acids Res., 42 1083-1090. (doi: 10.1093/nar/gkt1031)
- [12] Belakhov V.V., Kolodyaznaya V.A., Garabadzhiu A.V., Chistyakova T.B., Smirnov I.A. Application of the Todd-Atherton Synthetic Approach for Chemical Modification of Tetraene Macrolide Antibiotic Lucensomycin // Russian Journal of General Chemistry. 2016. Vol. 86, № 3. P. 570-578. (doi: 10.1134/S1070363216030129).