ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.89:004.4 **DOI: 10.20998/2411-0558.2019.13.16**

Г. А. САМИГУЛИНА, д-р техн. наук, зав. лаб. "Интеллектуальные системы управления и прогнозирования", Институт информационных и вычислительных технологий, Алматы, Казахстан,
З. И. САМИГУЛИНА, ассоц. проф., Ph.D, Казахстанско-Британский Технический Университет, Алматы, Казахстан

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ МОЛЕКУЛЯРНОГО ДИЗАЙНА ЛЕКАРСТВЕННЫХ СОЕДИНЕНИЙ НА ОСНОВЕ МОДЕЛЬНО-ОРИЕНТИРОВАННОГО ПОДХОДА

Статья посвящена разработке информационной системы ведения научных исследований на базе модельно-ориентированного подхода MDA (Model Driven Architecture) и модифицированного алгоритма искусственных иммунных систем (AIS). Предложен модифицированный алгоритм AIS на базе метода оптимизации серых волков (Gray Wolf Optimization, GWO) для решения задачи прогнозирования зависимости "структура-свойство/активность" лекарственных соединений. Осуществлён сравнительный анализ результатов моделирования с использованием пакета прикладных программ Rapid Miner. Ил.: 5. Табл.: 3. Библиогр.: 10 назв.

Ключевые слова: молекулярный дизайн, модельно-ориентированный подход (MDA), модифицированный алгоритм искусственных иммунных систем, метод оптимизации серых волков.

Постановка проблемы. В последнее время большинство ведущих мировых фармакологических компаний сотрудничают с исследовательскими организациями, занимающимися разработками в области искусственного интеллекта (ИИ). Последние достижения ИИ и современные информационные технологии способствуют значительному прогрессу в области компьютерного молекулярного дизайна лекарственных препаратов. Внедрение инновационных модифицированных алгоритмов с использованием модельно-ориентированного подхода позволяет автоматизировать процесс обработки многомерной химической информации и значительно снизить временные и финансовые затраты при прогнозировании зависимости "структура-свойство/активность" (Quantitative Structure-Activity Relationship, QSAR) лекарственных соединений.

Синтез новых лекарственных препаратов состоит из ряда этапов, в которые вовлечены специалисты из различных областей науки, таких как: биоинформатика, фармакология, химия, хемоинформатика и т.д. Очень часто у исследователей в данной области возникают трудности при работе с современными программными продуктами и IT-

технологиями. Информационная система ведения научных исследований на основе подхода MDA нацелена на специалистов, не владеющих навыками программирования, но позволяющая использовать последние достижения в области искусственного интеллекта для прогнозирования зависимости "структура-свойство/активность" лекарственных соединений.

Методология MDA успешно применяется при разработке лекарств. В работе [1] представлено использование подхода MDA для поиска нового применения существующих лекарственных препаратов и прогнозирования их побочного эффекта. Исследования [2] посвящены объединению методов анализа больших данных "Big Data Analytics" и моделирования больших данных "Implementation of data modeling" под управлением архитектуры MDA для разработки новых лекарственных соединений. В работе [3] рассматривается применение модельно-ориентированной инженерии (Model-Driven Engineering) к сервис-ориентированным разработкам на основе Grid – архитектуры для исследований в области биомедицины (прогнозирование рака молочной железы, исследование заболеваний сердца, разработка систем персонализированной медицины в области педиатрии, создание нейровизуальных биомаркеров и т.д.).

Алгоритмы ИИ являются перспективными, но не универсальными. Актуальны исследования по улучшению прогнозирующих моделей за счёт удаления избыточных данных из исходного набора признаков.

Анализ литературы. Хорошо зарекомендовали себя генетические алгоритмы, а также модификации на их основе для решения подобного рода задач. В работе [4] представлен гибридный генетический алгоритм, в котором используется алгоритм локального поиска всемирного потопа (Great deluge algorithm, GDA) вместо операции мутации. Произведён сравнительный анализ полученных данных с использованием различных классификаторов (*k*-ближайших соседей, многослойного персептрона, метода опорных векторов и т.д.), который показал эффективность предложенного алгоритма. В статье [5] представлен гибридный бинарный алгоритм чёрной дыры (Black hole algorithm, BHA) и модифицированный бинарный алгоритм оптимизации роя частиц (Particle swarm optimization, PSO) для решения задачи прогнозирования возникновения рака. Исследования [6] посвящены разработке модифицированного алгоритма искусственной иммунной сети для выделения информативных признаков с целью применения в качестве метода предварительной обработки данных с функциями сжатия и очистки данных. В работе [7] рассмотрены существующие алгоритмы AIRS и предложена модификация, которая учитывает параметр

numRepAg (количество обучающих агентов), не используемый в предыдущих версиях. Моделирование проводилось на основе данных репозитория "UCI machine learning" и показало хороший результат.

Таким образом, анализ литературы подтверждает актуальность разработок в области искусственного интеллекта для компьютерного молекулярного дизайна новых лекарственных препаратов.

Цель статьи – разработка информационной системы ведения научных исследований на основе подхода MDA и модифицированного алгоритма искусственных иммунных систем для прогнозирования зависимости "структура-свойство/активность" лекарственных соединений. Предложенный модифицированный алгоритм состоит из предварительной обработки химической информации на основе алгоритма оптимизации серых волков (GWO) и решения задачи прогнозирования с помощью AIS.

Информационная система ведения научных исследований на основе подхода MDA и модифицированного алгоритма AIS.

Подход MDA впервые был предложен консорциумом Object Management Group (OMG) и является удобным инструментом для реализации информационных систем на любых платформах. Построение архитектуры MDA основывается на разработке моделей, описываемых с помощью UML (Unified Modeling Language) языка. Разработка MDA архитектуры делится на два этапа: создание платформо-независимой модели (Platform Specifical Model, PSM) и платформо-зависимой модели (Platform Specifical Model, PSM). Данная стратегия позволяет специалистам в определённой предметной области, не владеющим навыками программирования разрабатывать информационные системы для различных приложений в фармакологии, биоинформатике и биомедицине.

Модифицированные алгоритмы искусственного интеллекта описываются в виде моделей предметной области, которые составляют основу MDA архитектуры. В качестве научно-исследовательской платформы для моделирования разработанных алгоритмов удобно использовать программный продукт Rapid Miner. Данная программа в настоящий момент занимает лидирующие позиции на мировом рынке для работы с большими данными, содержит в себе статистические методы анализа данных, оптимизационные модели, современные модели прогнозирования на основе алгоритмов искусственного интеллекта и т.д. Программное обеспечение Rapid Miner позволяет разрабатывать графические модели, имеет встроенные функции предварительной

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

обработки данных (Turbo Preparation), а также мощный инструмент визуализации.

Рассмотрим разработанную MDA архитектуру для компьютерного молекулярного дизайна лекарственных препаратов на базе Rapid Miner (рис. 1).

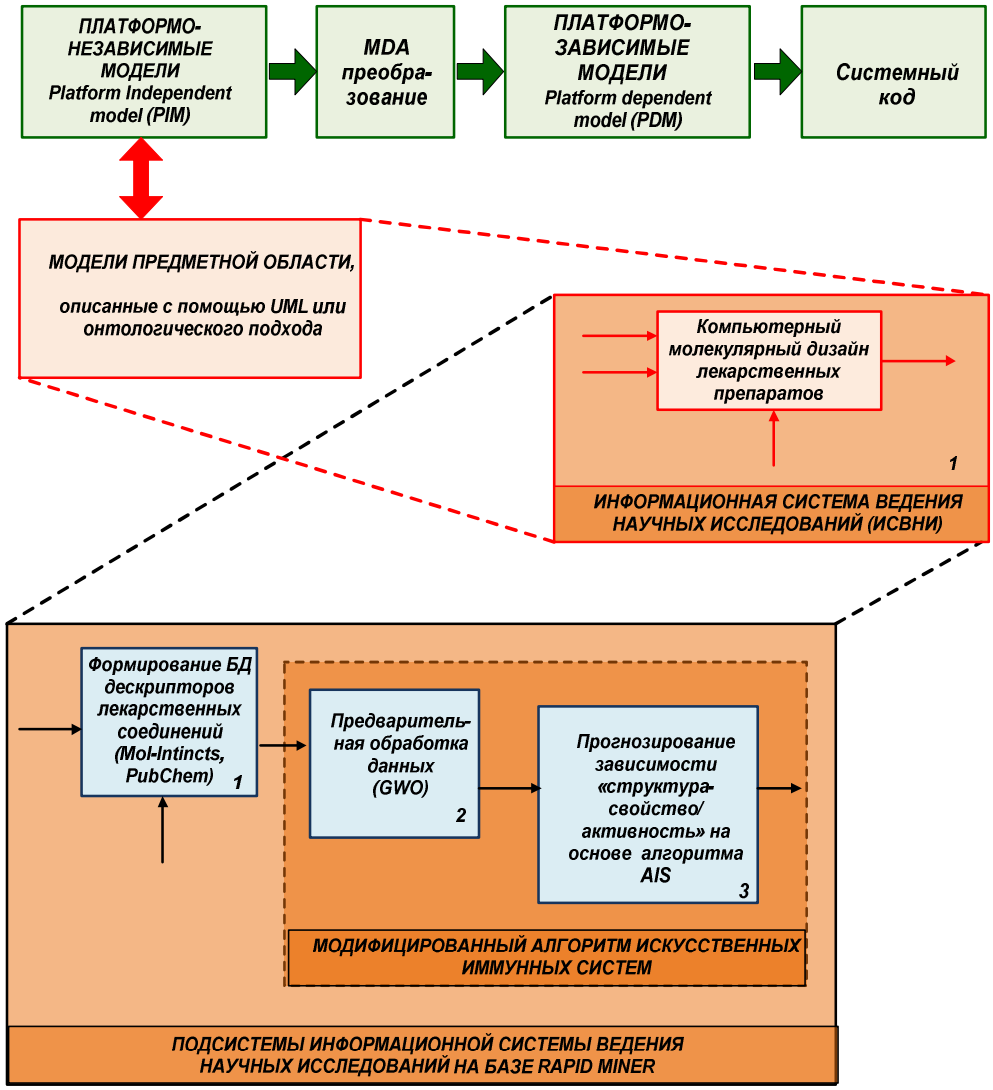

Рис.1. Архитектура информационной системы ведения научных исследований на основе MDA технологии для компьютерного молекулярного дизайна лекарственных препаратов

Модель предметной области для информационной системы ведения научных исследований состоит из следующих этапов: 1 этап – формирование базы данных дескрипторов, описывающих химическое соединение (подключение к мировым банкам данных химической информации); 2 этап – предварительная обработка данных с помощью алгоритма оптимизации серых волков, 3 этап – прогнозирование зависимости "структура-свойство/активность" на основе алгоритма AIS. Этапы 2 и 3 формируют модифицированный алгоритм GWO-AIS для прогнозирования зависимости "структура-свойство/активность" лекарственных соединений.

Рассмотрим алгоритм GWO для предварительной обработки данных. Метод оптимизации серых волков впервые был предложен в 2014 году Mirjalili et al [8] и описывает поведение серых волков в природе. Иерархия волков представлена в виде следующей структуры: α – волки лидеры; β – волки советчики, помогающие α в принятии решений; δ – подчиняются группе α и β, но доминируют над последним уровнем; ω – последний уровень иерархии, подчиняется всем вышестоящим уровням.

Математическая модель поведения волков во время охоты имеет следующий вид:

$$\begin{aligned} \vec{D} &= \left|\vec{C}\cdot\vec{X}_p(t)-\vec{X}(t)\right|, \\ \vec{X}(t+1) &= \vec{X}_p(t)-\vec{A}\cdot\vec{D}, \end{aligned} \qquad (1)$$

где t – текущая итерация; $\vec{A},\vec{C}$ – векторы коэффициенты, рассчитываемые по формулам: $\vec{A}=2\vec{a}\cdot\vec{r}_1-\vec{a}$; $\vec{C}=2\cdot\vec{r}_2$, значение $\vec{a}$ линейно уменьшается от 2 до 0, согласно порядку итерации и $\vec{r}_1$, $\vec{r}_2$ – рандомные векторы $[0,\ 1]$; $\vec{X}_p$ – позиция вектора добычи, $\vec{X}$ – позиция вектора волков [9].

Процесс охоты инициируется α, в то время как β и δ могут помогать. В математической модели (1) показано, что α, β и δ представляют собой наилучшее решение относительно потенциального расположения добычи. Первые три лучших решения сохраняются и другие агенты (agents) обязаны обновлять свои позиции согласно позиции агентов наилучшего поиска (best search agents) на основе уравнений вида:

$$\vec{D}_\alpha=\left|\vec{C}_1\cdot\vec{X}_\alpha-\vec{X}\right|,\ \vec{D}_\beta=\left|\vec{C}_2\cdot\vec{X}_\beta-\vec{X}\right|,\ \vec{D}_\delta=\left|\vec{C}_3\cdot\vec{X}_\delta-\vec{X}\right|, \qquad (2)$$

$$\vec{X}_1=\vec{X}_\alpha-\vec{A}_1\cdot(\vec{D}_\alpha),\ \vec{X}_2=\vec{X}_\beta-\vec{A}_2\cdot(\vec{D}_\beta),\ \vec{X}_3=\vec{X}_\delta-\vec{A}_3\cdot(\vec{D}_\delta), \qquad (3)$$

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3}, \quad (4)$$

где вектор $\vec{A}$ представляет собой случайное значение, лежащее в диапазоне $[-2a, 2a]$, а значение параметра a уменьшается от 2 до 0, согласно текущей итерации.

Далее прогнозирование QSAR осуществляется с помощью алгоритма AIS. Механизм AIS основан на принципах теоретической иммуннологии и описывает реакцию организма на попадание в него различного рода патогенов. В настоящее время наиболее известными подходами AIS являются: клональная селекция (Clonal Selection, CS), отрицательный отбор (Negative selection, NS) и иммунные сети. Широко известны следующие алгоритмы AIS: алгоритм распознавания искусственной иммунной системой (Artificial Immune Recognition Systems, AIRS), предложенный Andrew B. Watkins и Jon Timmis; алгоритм клональной селекции (CLONALG), разработанный Leandro N. de Castro и Fernando J. Von Zuben; алгоритм классификации на основе клональной селекции (Clonal Selection Classification Algorithm, CSCA) созданный Jason Brownlee и т.д.

Перспективным является иммунносетевой подход, в котором используется механизм молекулярного узнавания [10]. В основе алгоритма лежит понятие формального пептида и определение минимума энергии связи между формальными пептидами (антиген-антитело) на основе сингулярного разложения матриц. Данный метод используется для разработки модифицированного алгоритма GWO-AIS при прогнозировании зависимости "структура-свойство/активность" лекарственных соединений.

Моделирование и сравнительный анализ работы модифицированного алгоритма GWO-AIS.

Рассмотрим моделирование модифицированного алгоритма GWO-AIS на примере лекарственных соединений сульфаниламидной группы. Сульфаниламиды представляют собой антибактериальные средства широкого спектра действия. В таблице 1 представлен фрагмент базы данных (БД) сульфаниламидов, составленный на основе крупнейшего мирового репозитория химической информации Mol-Instincts. База данных состоит из дескрипторов различного уровня, описывающих структуру химических соединений, например: number of atoms – число атомов, molecular weight – молекулярный вес, Gravitation index – гравитационный индекс; Cubic root of Gravitation index – корень кубический гравитационного индекса и т.д. Размерность БД составляет R_1=15 × 1500, всего 22 000 атрибутов данных.

Таблица 1

Фрагмент базы данных сульфаниламидов

Вещество	Дескрипторы, описывающие структуру химического соединения						
	Number of atoms	Molecular weight	Average molecular weight	Gravitation index	Cubic root of Gravitation index	…	Polarity parameter
Sulfadiazine	27	250.2751	9.2700	1980.800	12.5588	…	0.1576
Sulfadimidine	33	278.3287	8.4400	1916.900	12.4222	…	0.2407
Sulfafurazole	31	267.3018	8.6200	1874.900	12.3308	…	0.2798
Sulfamethizole	27	270.3240	10.0100	1856.900	12.2913	…	0.2712
Sulfamethoxazole	46	311.4416	6.7700	2335.900	13.2684	…	0.1854
...	...	...	...	...	...	…	…
Sulfaperin	30	264.3018	8.8100	1843.000	12.2605	…	0.2439

Для решения задачи прогнозирования сульфаниламиды можно разделить на следующие классы: 1 класс – сульфаниламиды короткого действия (менее 10 ч); 2 класс – сульфаниламиды средней длительности действия (10 – 24 ч); 3 класс – сульфаниламиды длительного действия (24 – 48 ч). Визуализация фрагмента БД сульфаниламидов в 2D форме для соединения sulfadiazine средней продолжительности действия представлена на рис. 2.

Для тестирования эффективности применения алгоритма оптимизации серых волков в качестве метода для редукции малоинформативных дескрипторов сульфаниламидов и формирования оптимальной БД дескрипторов рассмотрим сравнительный анализ с различными алгоритмами машинного обучения. В качестве алгоритмов сравнения выбраны следующие методы: наивный Байесовский алгоритм (Naive Bayes), глубокое обучение (Deep Learning), деревья решений (Decision Tree), случайный лес (Random Forest), метод опорных векторов (Support Vector Machine). Моделирование осуществлялось в среде Rapid Miner.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

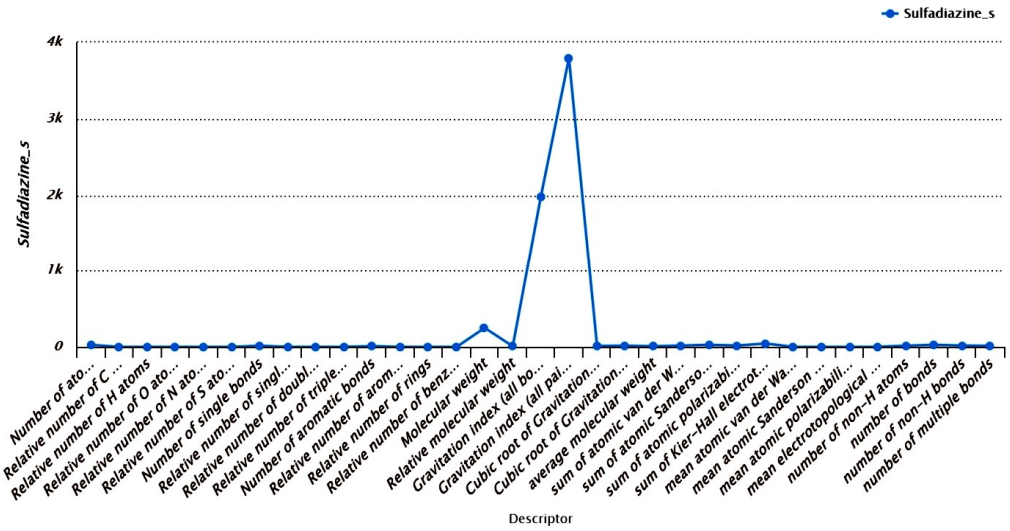

Рис. 2. Фрагмент базы данных дескрипторов sulfadiazine

На рис. 3 представлены графики эффективности применения представленных алгоритмов по критериям: точность (accuracy) и быстродействие (Runtime) для полной БД дескрипторов сульфаниламидов размерности R_1 без предварительной обработки данных на основе GWO.

В табл. 2 представлена детализация результатов моделирования.

Таблица 2

Сравнительный анализ результатов прогнозирования

Модель прогнозирования	Точность распознавания	Ошибка распознавания	Время моделирования
Наивный Баесовский алгоритм	62,5%	37,5%	2 сек.
Глубокое обучение	69,8%	30,2%	34 сек.
Деревья решений	64,2%	35,8%	3 сек.
Случайный лес	86,4%	13,6%	1 мин. 1 сек.
Метод опорных векторов	87,8%	12,2%	15 сек.

Далее рассмотрим статистику после обработки БД дескрипторов сульфаниламидов с помощью алгоритма оптимизации серых волков.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Размерность новой БД дескрипторов сульфаниламидов составляет $R_2 = 15 \times 200$, 3000 аттрибутов. На рис. 4 представлены результаты моделирования с использованием БД размерности R_2.

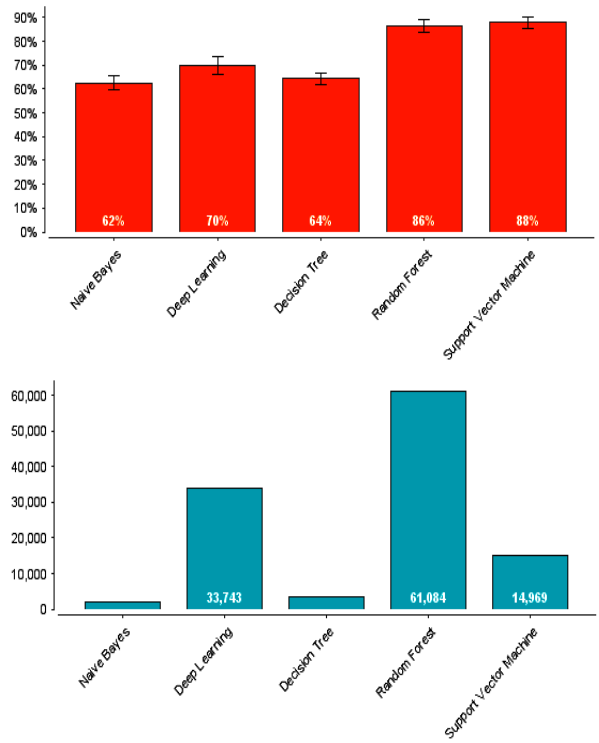

Рис. 3. Решение задачи распознавания образов с использованием БД дескрипторов сульфаниламидов без предварительной обработки данных

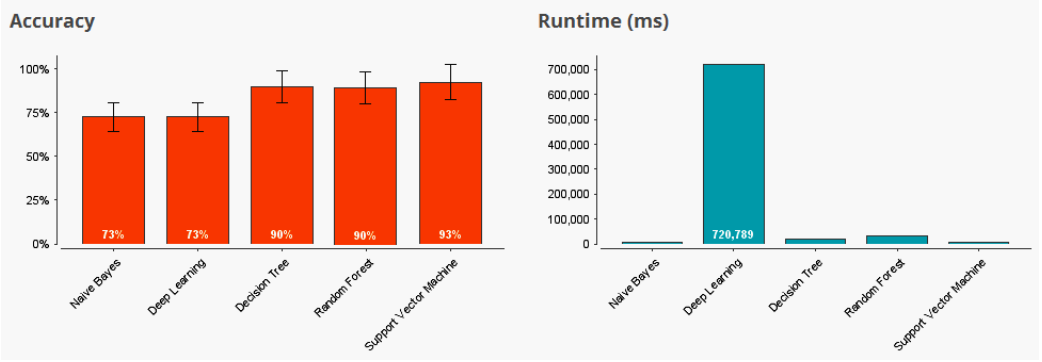

Рис. 4. Решение задачи распознавания образов с использованием БД дескрипторов сульфаниламидов после предварительной обработки данных на основе GWO

Статистика эффективности применения рассмотренных алгоритмов представлена в табл. 3.

Таблица 3

Сравнительный анализ результатов прогнозирования БД сульфаниламидов

Модель прогнозирования	Точность распознавания	Ошибка распознавания	Время моделирования
Наивный Баесовский алгоритм	72,7%	27,3%	7 сек
Глубокое обучение	72,7%	27,3%	12 мин 0 сек
Деревья решений	90%	10%	20 сек
Случайный лес	90%	10%	32 сек
Метод опорных векторов	92,7%	7,3%	7 сек

Таким образом, алгоритм оптимизации серых волков может быть применён при разработке модифицированного алгоритма AIS в рамках архитектуры MDA для молекулярного дизайна лекарств.

Рассмотрим результаты моделирования модифицированного алгоритма GWO-AIS для решения задачи прогнозирования QSAR сульфаниламидов. Для оценки эффективности работы предложенного модифицированного алгоритма GWO-AIS проведен сравнительный анализ (рис.5) результатов моделирования с алгоритмом GWO-AIRS в программной среде WEKA (Waikato Environment for Knowledge Analysis).

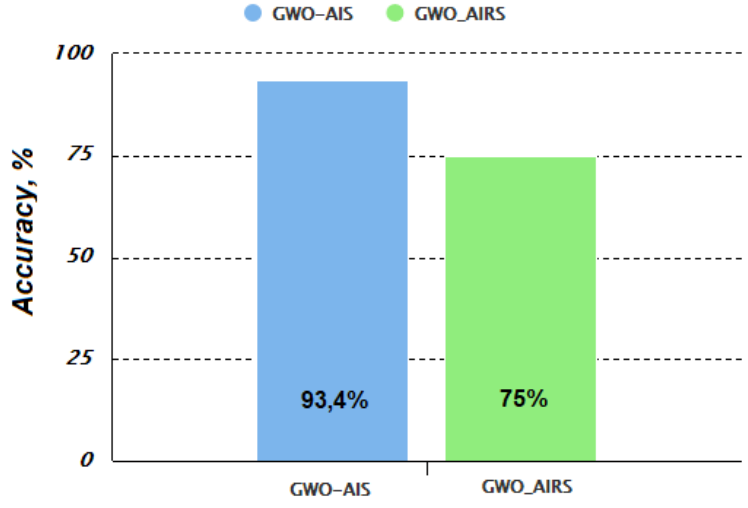

Рис. 5. Прогнозирование сульфаниламидов с помощью модифицированных алгоритмов GWO-AIS и GWO-AIRS

Точность GWO-AIS составляет 93,4% (время моделирования 3 сек.), эффективность алгоритма GWO-AIRS составляет 75% (время моделирования 5 сек).

Полученные результаты показывают перспективность применения предложенного модифицированного алгоритма GWO-AIS для прогнозирования QSAR новых лекарственных соединений сульфаниламидной группы.

Выводы. Разработанная архитектура информационной системы ведения научных исследований на основе подхода MDA и модифицированного алгоритма ИИС является эффективным инструментом для исследователей в области компьютерного молекулярного дизайна новых лекарственных препаратов, не владеющих навыками программирования. Предложенная архитектура, на основе разработки моделей предметной области может дополняться новыми перспективными алгоритмами и успешно подходит для исследований в области фармакологии и биоинформатики.

Работа выполнена по гранту Комитета Науки Министерства Образования и Науки Республики Казахстан AP05130019 по теме: "Разработка и анализ баз данных для информационной системы прогнозирования зависимости "структура-свойство" лекарственных соединений на основе алгоритмов искусственного интеллекта" (2018 – 2020 гг.).

References:

1. Banpatte, S., Shinde, U., Patil, R., Manole, K., Patil, R., and Aldar, K.M. (2017), "Drug Discovery Based On Model Driven Architecture", *International Research Journal of Engineering and Technology,* vol. 4, № 4, pp. 886-893.

2. Etani, N. (2015), "Database application model and its service for drug discovery in Model-driven architecture", *Journal of Big Data,* vol. 2, № 16, pp. 1-17.

3. Manset, F.M. (2012), "A formal architecture-centric and model-driven approach for the engineering of science gateways", *Centre for Complex Cooperative Systems,* pp.1-71.

4. Guha, R., Ghosh, M., Kapri, S., Shaw, S., Mutsuddi, S., Bhateja, V., and Sarkar, R. (2019), "Deluge based Genetic Algorithm for feature selection", *Evolutionary intelligence*, pp. 1-11.

5. Elham, E., and Aydin, P. (2018), "Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization", *Genomics,* pp. 1-18.

6. Ge, H., and Yan, X. (2011), "A Modified Artificial Immune Network for Feature Extracting", Advances in Swarm Intelligence, pp. 408-415.

7. Jenhani, I., and Elouedi, Z. (2014), "Re-visiting the artificial immune recognition system: a survey and an improved version", *Artificial Intelligence Review,* vol. 42, № 4., pp. 821-833.

8. Mirjalili, S., Mirjalili, and S.M, Lewis, A. (2014), "Gray wolf optimizer", *Advances in Engineering software*, vol. 69, pp. 46-61.

9. Faris, H., Aljarah, I., Al-Betar, M., and Mirjalili, S. (2018), "Gray wolf optimizer: a review of recent variants and applications", *Neural Computing and Applications,* vol. 30, №2., pp. 413-435.

10. Samigulina, G., and Samigulina, Z. (2017), "Immune Network Technology on the basis of random forest algorithm for computer-aided drug design", *Bioinformatics and Biomedical Engineering,* vol. 10208, pp. 50-61.

Статью представил д.т.н., проф. Национального технического университета "Харковский политехнический институт" С.Ю. Леонов

Поступила (received) 27.05.2019

Samigulina Galina, Dr. Sci. Tech,
Institute of Information and Computing Technologies,
Str. Pushkeen, 125, Almaty, Kazakhstan, 050010,
Tel:+7(777)244-43-67, e-mail: galinasamigulina@mail.ru
ORCID ID: 0000-0003-1798-9161

Samigulina Zarina, Ph.D,
Kazakh British Technical University
Str. Tole bi, 59, Almaty, Kazakhstan, 050000,
Tel:+7(702)218-97-73, e-mail: zarinasamigulina@mail.ru
ORCID ID: 0000-0002-5862-6415

УДК 004.89:004.4

Інформаційна система для молекулярного дизайну лікарських сполук на основі модельно-орієнтованого підходу / Самігуліна Г.А., Самігуліна З.І. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2019. – № 1. – С. 140 – 152.

Стаття присвячена розробці інформаційної системи ведення наукових досліджень на базі модельно-орієнтованого підходу MDA (Model Driven Architecture) і модифікованого алгоритму штучних імунних систем (AIS). Запропоновано модифікований алгорітм AIS на базі методу оптимізації сірих вовків (Gray Wolf Optimization, GWO) для вирішення задачі прогнозування залежності "структура-властивість/активність" лікарських сполук. Здійснено порівняльний аналіз результатів моделювання з використанням пакета прикладних програм Rapid Miner. Іл.: 5. Табл.: 3. Бібліогр.: 10 назв.

Ключові слова: молекулярний дизайн; модельно-орієнтований підхід (MDA); модифікований алгоритм штучних імунних систем; метод оптимізації сірих вовків.

УДК 004.89:004.4

Информационная система для молекулярного дизайна лекарственных соединений на основе модельно-ориентированного подхода / Самигулина Г.А., Самигулина О.И. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2019. – № 1. – С. 140 – 152.

Статья посвящена разработке информационной системы ведения научных исследований на базе модельно-ориентированного подхода MDA (Model Driven Architecture) и модифицированного алгоритма искусственных иммунных систем (AIS). Предложен модифицированный алгоритм AIS на базе метода оптимизации серых волков (Gray Wolf Optimization, GWO) для решения задачи прогнозирования зависимости "структура-свойство/активность" лекарственных соединений. Осуществлён сравнительный анализ результатов моделирования с использованием пакета прикладных программ Rapid Miner. Ил.: 5. Табл.: 3. Библиогр.: 10 назв.

Ключевые слова: молекулярный дизайн; модельно-ориентированный подход (MDA); модифицированный алгоритм искусственных иммунных систем; метод оптимизации серых волков.

UDC 004.89:004.4

Information system for molecular design of drug compounds based on a model-based approach / Samigulina G.A., Samigulina Z.I. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2019. – №.1. – C. 140 – 152.

The article is devoted to the development of an information system for conducting scientific research based on the model-based approach MDA (Model Driven Architecture) and the modified algorithm of artificial immune systems (AIS). A modified AIS algorithm based on the Gray Wolf Optimization (GWO) is optimization method for solving the problem of predicting the "structure-property/activity" dependence of medicinal compounds is proposed. A comparative analysis of simulation results using the Rapid Miner application software package was carried out. Figs.: 5. Tabl.: 3. Refs.: 10 titles.

Keywords: molecular design of drug; model-based approach MDA; modified algorithm of artificial immune systems; gray wolf optimization.