На правах рукописи

Ромеро Рейес Илякай Владиславовна

ОЦЕНКА АФФИННОСТИ КОМПЛЕКСОВ БЕЛОК–ЛИГАНД С ПРИМЕНЕНИЕМ НЕЙРОННЫХ СЕТЕЙ

Специальность: 05.13.18 – математическое моделирование, численные методы и комплексы программ

Автореферат диссертации на соискание ученой степени кандидата физико-математических наук

Москва – 2014

Работа выполнена в Федеральном государственном бюджетном учреждении «Научно-исследовательский институт биомедицинской химии имени В. Н. Ореховича» Российской академии медицинских наук.

Научный руководитель: кандидат физико-математических наук Филимонов Дмитрий Алексеевич

Официальные оппоненты: Баскин Игорь Иосифович, доктор физико-математических наук, МГУ имени М.В. Ломоносова, Физический факультет, ведущий научный сотрудник

Макеев Всеволод Юрьевич, доктор физико-математических наук, Институт общей генетики им. Н.И. Вавилова РАН, зав. отд. вычислительной системной биологии

Ведущая организация: Национальный исследовательский ядерный университет «МИФИ»

Защита состоится «____»_____2014 г. в_____часов на заседании Диссертационного совета Д 720.001.04 в Лаборатории информационных технологий Объединенного института ядерных исследований, г. Дубна Московской области.

С диссертацией можно ознакомиться в библиотеке ОИЯИ.

Автореферат разослан «__»_____2014 г.

Ученый секретарь Диссертационного совета,
доктор физико-математических наук,
профессор Иванченко И. М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. В настоящее время при создании и поиске новых лекарственных соединений активно применяют различные компьютерные методы поиска, молекулярного моделирования и конструирования фармакологически перспективных соединений *de novo*. Применение данных методов позволяет существенно ускорить процессы разработки и внедрения и снизить их стоимость.

Одной из важных задач является компьютерный расчет аффинности предполагаемых лигандов к макромолекуле-мишени. Наиболее распространены два подхода: расчет изменения свободной энергии при связывании лиганда с макромолекулой-мишенью методами молекулярного моделирования и подбор для каждой мишени оценочной функции. Однако компьютерные оценки изменения свободной энергии сопряжены с фундаментальными ограничениями и имеют низкую точность, т.к. при данном подходе часто сложно учесть энтропийную составляющую энергии взаимодействия. Второй подход основан на использовании выборок соединений с соответствующими экспериментально измеренными величинами для подбора параметров оценочных функций. Для этих методов основное ограничение заключается в том, что такие модели дают хорошие предсказания только для лигандов того же химического класса, что и соединения из обучающей выборки. При этом дескрипторы для лигандов каждого класса необходимо подбирать индивидуально.

В данной работе предложен комбинированный метод на основе объединения результатов молекулярного моделирования и подхода на основе лигандов с известными свойствами. Для решения задачи построения оценочной функции в данной работе применяются искусственные нейронные сети (ИНС).

Цель работы: разработка численных методов для оценки аффинности комплексов белок–лиганд с применением нейронных сетей и их реализация в виде комплекса программ на графических процессорах.

Задачи исследования:

1. Подготовить выборки данных по белкам, лигандам и комплексам белок–лиганд и выполнить молекулярное моделирование отобранных комплексов и расчет энергетических параметров их взаимодействия по результатам молекулярной динамики.
2. Разработать численный метод для оценки аффинности лигандов к ядерным рецепторам стероидных гормонов на основе комплексного подхода, совмещающего методы молекулярного моделирования и искусственных нейронных сетей и провести тестирование на независимой выборке.
3. Разработать высокопроизводительную программную реализацию метода с применением параллельных вычислений с использованием графических процессоров.

Научная новизна. Впервые предложен метод оценки аффинности нестероидных лигандов к рецепторам глюкокортикоидов и прогестерона с использованием методов нелинейного снижения размерности и ИНС, и показана возможность применения метода расширения вложения для новых точек[1] в задаче оценки параметров взаимодействия комплексов белок–лиганд. Разработанные методы реализованы в виде программ, поддерживающих параллельные вычисления на основе графических процессоров.

Практическая значимость работы. Результаты диссертации могут быть использованы для оценки аффинности лигандов к ядерным рецепторам стероидных гормонов на основе физико-химических дескрипторов лигандов и составляющих изменения энергии взаимодействия комплексов белок–лиганд. Предлагаемый метод позволяет получить статистически значимые модели оценки аффинности для рассматриваемых рецепторов (коэффициент детерминации $\overline{R^2} = 0{,}94$) в отличие от моделей по оценочным функциям

[1] Bengio Y., Paiement J.-F., Vincent P., Delalleau O., Le Roux N., Ouimet M. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. // Advances in Neural Information Processing Systems. 2004. V. 16. P. 177-184.

молекулярного моделирования (коэффициент детерминации $\overline{R^2} < 0{,}1$). Разработанный автором метод был применен в лаборатории структурной биоинформатики ФГБУ «ИБМХ» РАМН для оценки аффинности стероидных лигандов ко всем ядерным рецепторам стероидных гормонов и показал хорошую предсказательную способность[2]. Также совместно с вычислительным экспериментом для ряда лигандов-пентаранов рецептора прогестерона был проведен синтез и тестирование *in vitro* в Институте органической химии имени Н. Д. Зелинского РАН и Московском государственном университете имени М. В. Ломоносова. Предсказанные значения аффинности комплексов рецептор–лиганд хорошо согласуются с результатами экспериментальной проверки.

Положения, выносимые на защиту.

- Новый метод оценки аффинности лигандов к внутриклеточным рецепторам глюкокортикоидов и прогестерона с использованием нелинейного снижения размерности и искусственных нейронных сетей на основе физико-химических параметров лиганда и составляющих изменения энергии взаимодействия комплексов белок–лиганд.
- Применение метода расширения вложения для новых точек в задаче оценки аффинности комплексов белок–лиганд.
- Параллельная программная реализация разработанных методов с использованием графических процессоров.

Апробация результатов работы. Основные положения и результаты диссертационной работы докладывались на XVI Российском национальном конгрессе «Человек и лекарство», Москва, Россия, 2010; IV сессии научной школы-практикума молодых ученых и специалистов в рамках VIII Всероссийской межвузовской конференции молодых ученых, Санкт-Петербург, Россия, 2011; XVI Международной конференции по нейрокибернетике, Ростов-на-Дону, 2012; Международной суперкомпьютерной конференции «Научный сервис в сети

[2] Федюшкина И.В. Предсказание аффинности и спектра действия лигандов ядерных рецепторов стероидных гормонов методами компьютерного моделирования: дис. на соискание ученой степени канд. биол. наук: 03.01.09 / Федюшкина Ирина Викторовна. – М., 2013 – 115 с.

Интернет: поиск новых решений», Новороссийск, Россия, 2012; XVII научной конференции молодых ученых и специалистов (ОМУС-2013) к 100-летию В. П. Джелепова, Дубна, Россия, 2013.

Публикации. Основные положения диссертационного исследования опубликованы в 9 научных трудах, в том числе 3 статьи в рецензируемых изданиях.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка сокращений и списка литературы, включающего 159 источников. Общий объем работы составляет 106 страниц, в том числе 26 рисунков и 12 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении по основным характеристикам существующих вычислительных методов оценки изменения свободной энергии комплексов белок–лиганд обоснована необходимость создания новых методов на основе объединения результатов молекулярного моделирования и подхода на основе лигандов с известными свойствами.

Первая глава посвящена обзору существующих вычислительных методов оценки аффинности лиганда к рецептору по изменению свободной энергии их взаимодействия; искусственных нейронных сетей; методов снижения размерности; основных принципов параллельных вычислений и технологии параллельных вычислений CUDA при использовании графических процессоров NVIDIA.

Задача оценки аффинности комплексов белок–лиганд представляет собой частный случай общей задачи поиска зависимостей структура–свойство[3]. В этой главе представлено описание двух основных групп методов для таких задач: на основе структуры лигандов (LBDD) и на основе структуры белка-мишени (SBDD)[4]. Первая группа методов актуальна в случае неизвестной

[3] Veselovsky A.V., Ivanov Y.D., Ivanov A.S., Archakov A.I., Lewi P., Janssen P. Protein–protein interactions: mechanisms and modification by drugs. // Journal of Molecular Recognition. 2002. V. 15. P. 405-422.
[4] Ivanov A.S., Veselovsky A.V., Dubanov A.V., Skvortsov V.S. Bioinformatics Platform Development. Springer, 2006. P. 389-431.

пространственной структуры мишени. К ним относят фармакофорные модели и модели «псевдорецептора», а также методы на основе регрессионного анализа взаимосвязи биологической активности лигандов и их молекулярных дескрипторов. Вторая группа методов требует анализа структуры белка-мишени с определением потенциального места связывания рассматриваемых лигандов. К ним относят методы молекулярного моделирования – докинг с последующим моделированием молекулярной динамики и расчётом составляющих изменения энергии взаимодействия комплексов, но оценки энергии Гиббса таким способом далеки от реальных величин[5].

Одно из наиболее важных направлений вычислительных методов связано с применением искусственных нейронных сетей[6]. В теореме Колмогорова[7] и многочисленных ее развитиях[8] представлено утверждение об универсальных аппроксимационных возможностях произвольной нелинейности: с помощью линейных операций и единственного нелинейного элемента φ можно спроектировать систему, вычисляющую любую непрерывную функцию с любой желаемой точностью. При этом указанная система имеет структуру нейронной сети с одним скрытым слоем, поэтому ИНС являются универсальными аппроксиматорами непрерывных функций.

ИНС успешно применяют в задачах прогнозирования физико-химических свойств органических соединений и их биологической активности. Нейронные сети с обратным распространением ошибки (Back propagation neural network, BPNN) являются очень мощным инструментом в исследованиях QSAR[9]. Так как настройка BPNN осуществляется процедурой «обучения с учителем»[10], то ее можно обучить по известным примерам, а потом использовать ее в качестве

---

[5] Hubbard R.E. Can drugs be designed? // Current Opinion in Biotechnology. 1997. V. 8. P. 696-700.

[6] Баскин И.И., Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биохимических исследованиях. // Вестник Московского университета. Серия 2. Химия. 1999. № 40. С. 323-326.

[7] Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного. // Докл. АН СССР. 1957. № 114. С. 953-956.

[8] Горбань А.Н., Дунин-Барковский В.Л., Кирдин А.Н., Миркес Е.М., Новоходько А.Ю., Россиев Д.А., Терехов С.А., Сенашова М.Ю., Царегородцев В.Г. Нейроинформатика. Новосибирск: Наука. Сибирское предприятие РАН, 1998. 296 с.

[9] Devillers J. Neural Networks in QSAR and Drug Design. London: Academic Press, Inc., 1996. 304 p.

[10] Haykin S. Neural Networks: A Comprehensive Foundation. New Jersey: Prentice Hall International, 1999. 842 p.

регрессионной модели. Но, у нее есть ряд недостатков: длительная по времени процедура обучения, скрытый характер функционирования, проблема переобучения. В главе представлено описание методов преодоления указанных недостатков, а также рассмотрены различные алгоритмы обучения с обратным распространением ошибки и методы перекрестного контроля.

Обязательным этапом предварительной обработки набора данных в данном исследовании было выявление наличия зависимостей входных параметров между собой для дальнейшего снижения размерности входного множества для нейронной сети.

Постановка задачи снижения размерности (задача вложения)[11]

Пусть $X_m = \{x_1, x_2, \dots, x_m\}$ – выборка из множества данных $X \subset \mathbb{R}^D$. По точкам множества X_m построить отображение

$$h: X_m \to Y_m = h(X_m) = \{y_1, y_2, \dots, y_m\} \subset \mathbb{R}^d$$

точек множества X_m во множество точек Y_m, лежащее в пространстве меньшей размерности $d < D$ и сохраняющее заданные соотношения между точками множеств X_m и Y_m. Полученные точки y_i, рассматриваемые как низкоразмерные представления многомерных векторов $x_i \in X_m$, должны «достоверно представлять» выборку X_m.

Обычно от отображения h требуют, чтобы точки множества Y_m сохраняли геометрическую структуру точек множества X_m (локальную геометрию, отношения близости, геодезические расстояния и др.).

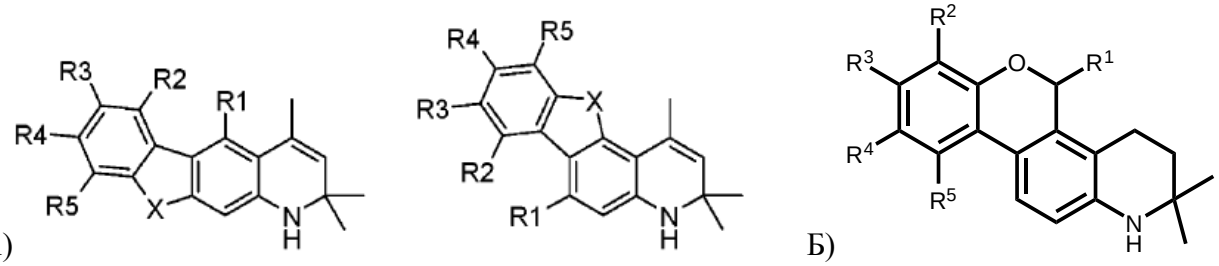
В этом разделе дается описание линейного метода снижения размерности – метода главных компонент, и четырех нелинейных методов – многомерного шкалирования, изометрического отображения, локально-линейного вложения и карт собственных значений лапласиана. Представлено описание **метода расширения вложения для новых точек** – задачи вложения для произвольной точки $x_{new} \in X/X_m$ и построение вложения h для точек множества $\{X_m \cup \{x_{new}\}\}$, сохраняющего множество $Y_m = h(X_m)$, – и его частные случаи

[11] Бернштейн А.В., Кулешов А.П. Снижение размерности при наличии предикатов. // Информационные процессы. 2008. № 8. С. 47-57.

для рассматриваемых методов снижения размерности.

Также в этой главе представлен обзор основных принципов параллельных вычислений и технологии параллельных вычислений CUDA[12] при использовании графических процессоров NVIDIA и подробное описание архитектуры графических процессоров NVIDIA Fermi GPU.

Вторая глава посвящена разработке вычислительных методов оценки аффинности комплексов белок–лиганд. Для рецептора прогестерона был взят набор из 63 нестероидных лигандов[13] с известными значениями $pK_i = -\lg(K_i)$, где K_i – константа ингибирования, а для рецептора глюкокортикоидов – набор из 69 нестероидных лигандов[14] с известными величинами pK_i. Общая структура для этих наборов представлена на рисунке 1.



А) Б)

Рисунок 1. Общая базовая структура рассматриваемых лигандов рецепторов прогестерона (А) и глюкокортикоидов (Б).

Входными параметрами для моделей оценки аффинности послужили 11 молекулярных дескрипторов – физико-химические параметры лигандов: молекулярный вес, [г/моль]; площадь поверхности, [Å^2]; площадь полярной поверхности, [Å^2]; полярный объём, [Å^3]; общий объём, [Å^3]; и составляющие энергии Гиббса [ккал/моль] комплексов белок–лиганд: изменение величины электростатического взаимодействия; изменение величины ван-дер-ваальсовых взаимодействий; вклады гидрофобных взаимодействий и сольватации,

[12] Боресков А.В., Харламов А.А. Основы работы с технологией CUDA. М.: ДМК Пресс, 2010. 232 с.
[13] Söderholm A.A., Lehtovuori P.T., Nyrönen T.H. Docking and three-dimensional quantitative structure-activity relationship (3D QSAR) analyses of nonsteroidal progesterone receptor ligands. // Journal of Medicinal Chemistry. 2006. V. 49. P. 4261-4268.
[14] Xu Y., Zhang T., Chen M. Combining 3D-QSAR, docking, molecular dynamics and MM/PBSA methods to predict binding modes for nonsteroidal selective modulator to glucocorticoid receptor. // Bioorganic & Medicinal Chemistry Letters. 2009. V. 19. P. 393-396.

рассчитанные по уравнению Пуассона-Больцмана и по обобщенной модели Борна.

Расчет составляющих энергии взаимодействия осуществлялся посредством продуктивной молекулярной динамики комплексов методами MM-PBSA/MM-GBSA[15] (программный пакет AMBER 9.0) после успешного докинга (программный пакет DOCK 6.5) молекул лигандов к участку связывания природного лиганда лиганд-связывающего домена рассматриваемого рецептора. На этапе построения модели значения величин аффинности рассматриваемых комплексов из литературных источников, были использованы как целевые, а для решения задачи нелинейной регрессии были использованы искусственные нейронные сети.

На этапе предварительной обработки данных для получения линейно-независимых входных параметров и снижения размерности входного множества был использован метод главных компонент с предварительной стандартизацией данных. По результатам этого метода размерность входного множества удалось снизить на две единицы, что объясняется наличием корреляций между вкладами гидрофобных взаимодействий и сольватации, рассчитанными по уравнению Пуассона-Больцмана (MM-PBSA) и по обобщенной модели Борна (MM-GBSA).

Также для всех молекул-лигандов в исходном наборе был вычислен коэффициент T_c молекулярного подобия Танимото[16] для всех пар соединений. По сумме $\sum_j T_c^j$ было выбрано соединение с наибольшим значением суммы, а значения $\tilde{T}_c^j$ из соответствующей ему строки, в дальнейшем, были использованы при разбиении исходного набора на три выборки: обучающую (70%), контрольную (15%) и тестовую (15%). Разбиение на эти выборки происходило случайным образом, но так, чтобы точки покрывали весь диапазон изменения коэффициента Танимото.

[15] Kollman P.A., Massova I., Reyes C., Kuhn B., Huo S., Chong L., Lee M., Lee T., Duan Y., Wang W. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. // Accounts of Chemical Research. 2000. V. 33. P. 889-897.

[16] Wold S. Pattern recognition by means of disjoint principal components models. // Pattern recognition. 1976. V. 8. P. 127-139.

На этапе построения модели полученные выборки сжатых данных с линейно независимым переменными были использованы для настройки ИНС. В основу структуры нейронной сети легла однонаправленная нейронная сеть с сигмоидальной функцией активации в одном скрытом слое и линейной функцией передачи в выходном слое. Структура сети представлена на рисунке 2.

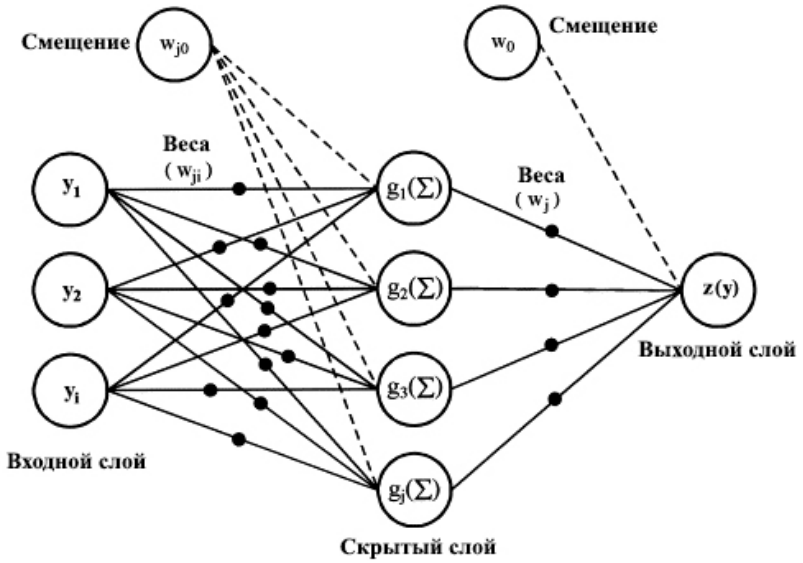


Рисунок 2. Архитектура однонаправленной нейронной сети с сигмоидальной функцией активации $g_j(v) = \tanh(v)$ в скрытом слое и линейной функцией $z(v) = av + b$ передачи в выходном слое.

Для настройки весов сети был использован пакетный алгоритм обучения с обратным распространением ошибки и минимизацией невязки методом Левенберга-Марквардта[17]. Входной слой содержал 9 нейронов, соответствующих линейно независимым дескрипторам, полученным на этапе предварительной обработки данных, и вектор смещений. Число нейронов в скрытом слое варьировалось, наилучший вариант для обоих рецепторов был получен при числе нейронов $N_{hidden} = 8$. Скрытый слой также содержал вектор смещений. В ходе настройки сети тестовая выборка также была использована для предотвращения переучивания сети. Для этого отслеживалось изменение среднеквадратичной ошибки на тестовой выборке. Если ошибка в течение некоторого числа эпох переставала уменьшаться, то обучение прерывалось, и

[17] Hagan M.T., Menhaj M.B. Training feedforward networks with the Marquardt algorithm. // Neural Networks, IEEE Transactions on. 1994. V. 5. P. 989-993.

использовалась та сеть, которая показывала минимум ошибки на тестовой выборке. Также для дополнительной оценки модели был проведён перекрестный контроль с исключением по одному (LOO). Результаты настройки и тестирования сети представлены в таблице 1 и на рисунке 3 (для рецептора прогестерона).

Для оценки модели были выбраны следующие статистические параметры: R^2 – коэффициент детерминации; Q^2 – коэффициент детерминации предсказания; *RMSE* – среднеквадратичная ошибка.

Таблица 1. Статистические параметры моделей «молекулярный докинг + молекулярная динамика + ИНС» при использовании метода главных компонент для исходных данных.

Статистические параметры	Рецептор прогестерона	Рецептор глюкокортикоидов
Количество входных параметров	9	9
R^2 для обучающей выборки	0,95	0,96
RMSE для контрольной выборки	0,14	0,09
Q^2 при контроле LOO	0,95	0,93
RMSE при контроле LOO	0,17	0,21

Дополнительно для обоих рецепторов было выполнено сравнение оценки аффинности по предложенному методу и по оценочным функциям изменения энергии взаимодействия комплекса белок-лиганд в результате молекулярного моделирования. Предлагаемый метод позволяет получить статистически значимые модели для рассматриваемых рецепторов (среднее значение $\overline{R^2} = 0{,}94$) в отличие от моделей по оценочным функциям молекулярного моделирования (среднее значение $\overline{R^2} < 0{,}1$).

Следующий этап исследования был посвящен применению нелинейных методов снижения размерности для входного множества. Была сформулирована расширенная задача вложения (см. выше), состоящая из двух пунктов:

- Задача вложения для входного множества $X_m \subset \mathbb{R}^D$.
- Задача вложения для произвольной точки $x_{new} \in X/X_m$.

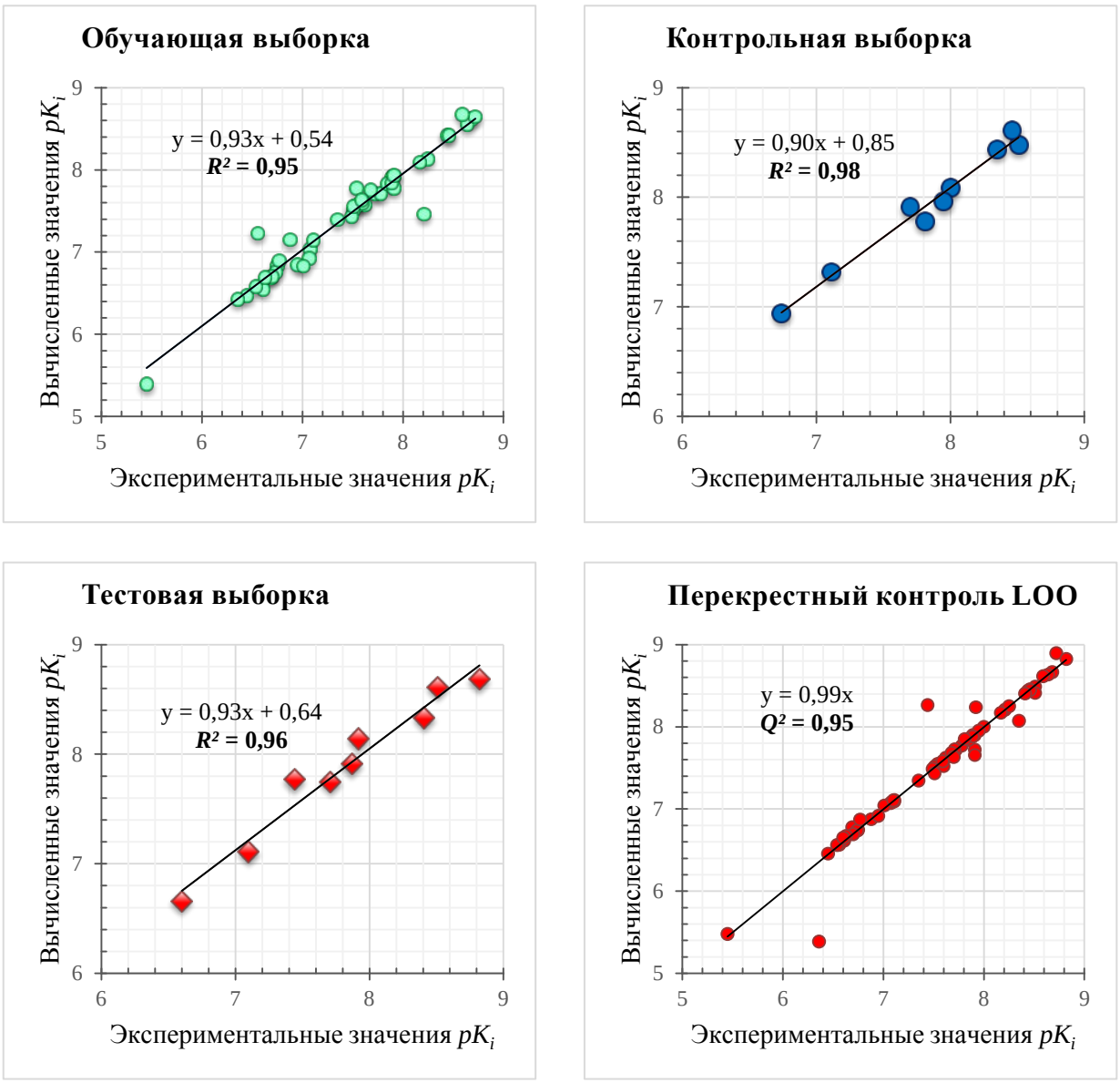


Рисунок 3. Сравнение вычисленных значений аффинности по модели со снижением размерности методом главных компонент и экспериментальных данных величины аффинности нестероидных лигандов к рецептору прогестерона для обучающей, контрольной и тестовой выборок, а также при кросс-валидации с исключением по одному.

Для решения задачи вложения для входного множества было произведено сравнение нелинейных методов снижения размерности на основе подходов глобальной (неметрическое многомерное шкалирование – MDS, и изометрическое отображение – Isomap) и локальной (локально-линейное вложение – LLE, и карты собственных значений лапласиана – LE) нелинейности.

Для этого на стандартизованном наборе исходных данных поочередно были применены исходные методы с указанием числа нелинейных компонент $N_{nonlinear}$, и для каждого варианта была проведена процедура построения сети, изложенная выше. Результаты представлены в таблицах 2 и 3.

Таблица 2. Коэффициент детерминации R^2 обучения ИНС при использовании нелинейных методов снижения размерности исходных данных для рецептора прогестерона.

$N_{nonlinear}$	5	6	7	8
MDS	0,73	0,91	**0,92**	0,92
Isomap	0,82	0,82	0,82	0,82
LLE	0,82	0,82	0,82	0,82
LE	0,70	0,72	0,72	0,74

Таблица 3. Коэффициент детерминации $\boldsymbol{R^2}$ обучения ИНС при использовании нелинейных методов снижения размерности исходных данных для рецептора глюкокортикоидов.

$N_{nonlinear}$	5	6	7	8
MDS	0,68	**0,91**	0,79	0,91
Isomap	0,56	0,85	0,67	0,81
LLE	0,67	0,71	0,77	0,82
LE	0,46	0,56	0,43	0,71

На основе этого анализа было выявлено, что наилучший результат достигается при использовании метода неметрического многомерного шкалирования:

- для рецептора прогестерона $N_{nonlinear} = 7, N_{hidden} = 6$
- для рецептора глюкокортикоидов $N_{nonlinear} = 6, N_{hidden} = 5$

Таким образом, размерность входного множества была снижена до 7 и 6 для рецептора прогестерона и рецептора глюкокортикоидов, соответственно.

Для отобранных моделей была проведена процедура перекрестного контроля с исключением по одному, аналогичная варианту при использовании метода главных компонент. Результаты настройки и тестирования сети представлены в таблице 4.

Для решения задачи вложения для новых точек был использован метод расширения вложения за пределами выборки для многомерного шкалирования, описанный в первой главе. Снижение размерности входного множества нелинейным методом MDS приводит к большему снижению размерности (с 11 дескрипторов до 7 и 6) и позволяет получить модели с хорошей прогностической способностью ($\overline{Q^2} = 0{,}90$), но с небольшой потерей точности по сравнению с результатами при использовании метода главных компонент ($\overline{Q^2} = 0{,}94$), который позволяет снизить размерность с 11 дескрипторов только до 9.

Таблица 4. Статистические параметры моделей «молекулярный докинг + молекулярная динамика + ИНС» при использовании метода неметрического многомерного шкалирования исходных данных.

Статистические параметры	Рецептор прогестерона	Рецептор глюкокортикоидов
Количество входных параметров	7	6
R^2 для обучающей выборки	0,92	0,91
RMSE для контрольной выборки	0,19	0,21
Q^2 при контроле LOO	0,90	0,90
RMSE при контроле LOO	0,26	0,25

Третья глава посвящена проверке разработанной методики. Дополнительным этапом в данном исследовании послужила экспериментальная проверка и сравнительный анализ оценки аффинности по разработанной модели и основным методам 3D QSAR на основе структур известных лигандов. В качестве объектов были рассмотрены: трехмерные структуры комплексов лиганд-связывающего домена рецептора прогестерона, для которых были проведены процедуры докинга и продуктивной молекулярной динамики; 42 аналога природного лиганда рецептора прогестерона – прегна-D´-пентараны (общая формула представлена на рисунке 4), синтез и биологическое тестирование этих пентаранов было выполнено ранее для указанного рецептора крысы и

кролика[18], этот набор был использован для настройки моделей; 8 дополнительных прегна-D´-пентаранов, для которых не было данных об экспериментальной оценке их аффинности к рецептору прогестерона, для этой тестовой выборки были рассчитаны оценки аффинности по разработанной модели и методам 3D QSAR.

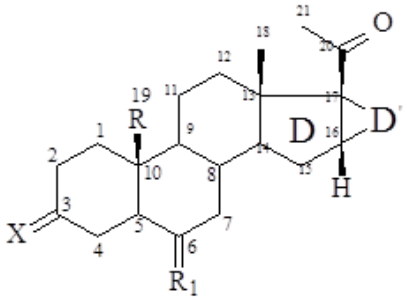


Рисунок 4. Общая формула производных 16α,17α-циклоалканопрогестерона (прегна-D´-пентаранов).

Оценка аффинности рассматриваемых объектов была выполнена с помощью методов CoMFA/CoMSIA, а также по линейно-регрессионной модели и модели, изложенной в данной работе с использованием метода главных компонент, с идентичными входными параметрами (физико-химические параметры лигандов и составляющие энергии Гиббса комплексов рецептор-лиганд). Статистические параметры настройки моделей представлены в таблице 5.

Для методов CoMFA/CoMSIA представлено два набора моделей, рассчитанных для полной выборки и с учётом так называемых «выбросов». Наличие «выбросов» очень характерно для методов CoMFA и CoMSIA, так как оценка качества моделей ведётся по процедуре перекрестного контроля.

[18] Smirnov A.N., Pokrovskaya E.V., Kogteva G.S., Shevchenko V.P., Levina I.S., Kulikova L.E., Kamernitzky A.V. The size and/or configuration of the cycloalkane D′ ring in pentacyclic progesterone derivatives are crucial for their high-affinity binding to a protein in addition to progesterone receptor in rat uterine cytosol☆. // Steroids. 2000. V. 65. P. 163-170.
Levina I., Kulikova L., Kamernitskii A., Shashkov A., Smirnov A., Pokrovskaya E. Synthesis of 6 (E)-and 6 (Z)-(3-ethoxycarbonylpropyl) oximes of 16α, 17α-cyclohexanopregn-4-ene-3, 6, 20-trione and study of their interaction with proteins of the rat uterine cytosol and blood serum. // Russian Chemical Bulletin. 2002. V. 51. P. 703-708.
Левина И.С., Куликова Л.Е., Камерницкий А.В., Покровская Е.В., Смирнов А.Н. Синтез 19-замещенных стероидов ряда 16α,17α-циклогексанопрегнанов и изучение их взаимодействия с белками цитозоля матки и сыворотки крови крысы. // Известия Академии Наук, Серия Химическая. 2005. № 11. С. 2579-2584.
Kamernitskii A.V., Levina I.S. Pregna-D'-pentaranes--progestins and antiprogestins: I. Differentiation of biological functions of steroid hormones. // Bioorganicheskaia Khimiia. 2005. V. 31. P. 115-129.

В этом случае наличие неудачно выравненных молекул или молекул, имеющих уникальные боковые радикалы, может давать существенные флуктуации. Как правило, их отбрасывают. Для моделей на основе методов «Докинг + Молекулярная динамика» одно соединение было отброшено, так как процедура докирования не смогла найти для него решения, поэтому для этой группы методов выборка содержит 41 соединение. Из всех построенных моделей ИНС с предварительным снижением размерности по методу главных компонент даёт самые лучшие результаты ($\boldsymbol{Q^2} = 0{,}91$), а среднеквадратичная ошибка предсказания при перекрестном контроле в 3,7 раза меньше, чем в лучшем варианте моделей 3D QSAR.

Таблица 5. Статистические параметры настройки моделей для прегна-D´-пентаранов.

Статистические параметры	Полная выборка		-10% выбросов		Докинг+Молек. динамика+	
	CoMSIA	CoMFA	CoMSIA	CoMFA	+лин.регр.	+PCA+ +ИНС
Число соединений	42	42	37	37	41	41
R^2 обучения	0,82	0,92	0,91	0,93	0,60	0,98
Q^2 при контроле LOO	0,37	0,38	0,59	0,57	0,41	0,91
RMSE при контроле LOO	1,28	1,30	1,08	1,12	1,13	0,29

Для оценки предсказательной способности полученных моделей была использована тестовая выборка из 8 описанных выше пентаранов. Одновременно с вычислительным экспериментом данные вещества из тестовой выборки были синтезированы и изучены *in vitro*[19] в Институте органической химии имени Н.Д. Зелинского РАН и Московском государственном

[19] Levina I.S., Pokrovskaya E.V., Kulikova L.E., Kamernitzky A.V., Kachala V.V., Smirnov A.N. 3- and 19-oximes of 16alpha,17alpha-cyclohexanoprogesterone derivatives: synthesis and interactions with progesterone receptor and other proteins. // Steroids. 2008. V. 73. P. 815-827.
4-Гетеро-16α, 17α-Циклогексанопрегнаны: пат. 2426737 Рос. Федерация: МПК C07J 53/00 C07J 73/00 A61P 35/00 / Левина И.С., Куликова Л.Е., Смирнов А.Н., Шимановский Н.Л., Семейкин А.В., Карева Е.Н., Федотчева Т.А., Болотова Е.Н.; заявитель и патентообладатель ИОХ РАН, ГОУ ВПО РГМУ Росздрава. – № 2009146877/04; заявл. 17.12.2009; опубл. 20.08.2011, Бюл. № 23. – 7 с.
Левина И.С., Куликова Л.Е., Шулишов Е.В., Томилов Ю.В., Смирнов А.Н. Синтез, структура и биологические свойства замещенных [16α,17α]-цикло-пропапрегн-4-ен-3,20-дионов. // Известия Академии Наук, Серия Химическая. 2013. № 6. С. 1449-1453.

университете имени М. В. Ломоносова. Все данные об этих соединениях были любезно предоставлены ведущим сотрудником лаборатории химии стероидных соединений ИОХ РАН д.х.н. И. С. Левиной.

При сравнении с экспериментальными данными вычисленные значения десятичного логарифма относительной конкурентной активности в случае ИНС дают $R^2{}_{test}$ тестирования $= 0{,}77$, а для 3D QSAR и линейной модели эта величина составляет всего 0,39 и 0,37. Но при этом две последние модели способны отличить лиганды с высоким сродством от лигандов с низким сродством. Нейронная же сеть даёт очень хороший результат ($R^2{}_{test} = 0{,}77$), что позволяет использовать ее для точного ранжирования лигандов по связыванию в ряду исследуемых соединений.

Таким образом, разработанные модели могут быть использованы для отбора новых кандидатов с заданной аффиностью к рецепторам прогестерона и глюкокортикоидов при направленном конструировании и/или поиске новых лекарственных соединений.

Четвертая глава посвящена разработке программной реализации созданной модели. Первоначально различные модификации модели были реализованы с помощью математического пакета MATLAB 2012b. По итогам исследования была установлена модель с конечной структурно-функциональной схемой, для которой была разработана параллельная реализация с использованием графических процессоров.

Итоговая модель оценки аффинности комплексов белок–лиганд включает в себя следующие компоненты:

1. Стандартизация данных
2. Два метода снижения размерности:
   2.1. Линейный – метод главных компонент;
   2.2. Нелинейный – неметрическое многомерное шкалирование.
3. Однонаправленная нейронная сеть с сигмоидальной функцией активации в одном скрытом слое и линейной функцией передачи в выходном слое:

3.1. Разбиение исходного набора на обучающую (70%), контрольную (15%) и тестовую (15%) выборки с учетом коэффициента молекулярного подобия Танимото.

3.2. Обучение с минимизацией невязки по методу Левенберга-Марквардта;

3.3. Вложение дескрипторного описания нового лиганда в пространство сниженной размерности и оценка его аффинности к рассматриваемому рецептору.

На этапе разработки параллельной реализации модели была создана параллельная форма алгоритма итоговой модели. Для этого в пунктах 2 и 3.2 структурно-функциональной схемы модели были выделены независимые операции, по которым вычисления можно проводить по отдельности – в параллельном режиме. Так как параллельный алгоритм был реализован на технологии NVIDIA CUDA, то на этапе построения алгоритма были учтены архитектурные особенности графических карт – SIMT (single-instruction, multiple-thread), пропускная способность передачи данных между центральным процессором и графической картой, объем и скорость чтения/записи различных типов памяти графической карты. В параллельной программной реализации были использованы следующие библиотеки: cuBLAS (NVIDIA CUDA Basic Linear Algebra Subroutines) для вычисления матричных операций; cuRAND (NVIDIA CUDA Random Number Generation library) для генерации случайных величин при инициализации весов ИНС; CULA dense (библиотека линейной алгебры LAPACK, оптимизированная под CUDA) для распараллеливания метода главных компонент; HiT-MDS (High-Throughput Multidimensional Scaling) в качестве основы для параллельной реализации неметрического многомерного шкалирования; LevMar (Levenberg-Marquardt optimization algorithm library) в качестве основы для параллельной реализации обучения ИНС.

Расчеты проводились на гибридной вычислительной системе на базе серверной платформы HP Proliant G7 (AMD Opteron 6100) и вычислительной

системы Tesla S2050 с использованием технологии NVIDIA CUDA 5.0. Результаты вычислений параллельной версии практически идентичны результатам, полученным в реализации MATLAB.

Характеристики скорости вычислений описанными способами представлены в таблице 6. Для обучения сети замерялось время, которое необходимо для обучения 20 эпох. На основании этого времени было вычислено количество эпох обучения одной сети, приходящееся в среднем на единицу времени.

Таблица 6. Характеристики выполнения последовательных и параллельных расчетов отдельных функционалов модели для рецептора прогестерона.

Реализация	MATLAB	C++ CPU	C++ CPU + CUDA GPU
Метод главных компонент PCA (сек.)	0,030	0,014	0,002
Неметрическое многомерное шкалирование MDS (сек.)	0,154	0,103	0,028
Обучение ИНС при PCA (эпох/ед.врем.)	54/1 сеть	78/1 сеть	5396/256 сетей
Обучение ИНС при MDS (эпох/ед.врем.)	61/1 сеть	82/1 сеть	5524/256 сетей

По указанным характеристикам видно, что использование графических процессоров позволяет ускорить процедуру сжатия данных в 7 раз для метода главных компонент и в 3,7 раз для неметрического многомерного шкалирования по сравнению с C++ реализацией на центральном процессоре CPU. Также видно, что количество эпох обучения для одной сети на CPU+GPU в 3,8 раз ниже, чем на CPU, но из-за распараллеливания на 256 сетей, число эпох обучения в единицу времени всех 256 сетей в ~69 раз превосходит CPU реализацию. Данный результат согласуется с архитектурой GPU: большое количество более медленных по сравнению с CPU вычислительных ядер.

Также для ускорения процедуры обучения в реализацию была включена поддержка расчетов на нескольких графических картах, которая была использована при обучении сети с различными вариантами

первоначальных весов для модели с использованием метода главных компонент. Это позволило получить дополнительное ускорение, соразмерное с количеством используемых графических карт.

Таким образом, был разработан эффективный параллельный алгоритм созданной модели и его программная реализация с применением графических карт NVIDIA и технологии CUDA.

ВЫВОДЫ

1. Предложен численный метод оценки аффинности комплексов лиганд-белок на основе комплексного подхода, совмещающего методы молекулярного моделирования, искусственных нейронных сетей и нелинейного снижения размерности.
2. На основе предложенного метода разработаны модели оценки аффинности лигандов к внутриклеточным рецепторам прогестерона и глюкокортикоидов с высокой предсказательной силой ($\overline{Q^2} = 0{,}94$ при снижении размерности входного множества методом главных компонент, $\overline{Q^2} = 0{,}90$ – методом неметрического многомерного шкалирования).
3. Вычислительный эксперимент по оценке величины связывания прегна-D´-пентаранов с рецептором прогестерона с последующей экспериментальной проверкой показал, что результаты вычислений модели на основе разработанного метода хорошо согласуются с экспериментальными данными и дают существенно лучший результат ($R^2{}_{test} = 0{,}77$) по сравнению с предсказаниями по методам 3D QSAR ($R^2{}_{test} = 0{,}37$).
4. Разработан эффективный алгоритм и программная реализация численного метода с применением параллельной технологии CUDA. Данная реализация позволяет ускорить процедуру сжатия данных в 7 и 3,7 раз для метода главных компонент и неметрического многомерного шкалирования, соответственно; и ускорить процедуру обучения в ~69 раз при использовании GPU ускорителей.

Основные положения диссертации изложены в следующих научных трудах и публикациях:

1. Федюшкина И.В., Скворцов В.С., **Ромеро Рейес И.В.**, Левина И.С. Молекулярный докинг и 3D QSAR производных 16α,17α-циклоалканопрогестерона как лигандов рецептора прогестерона // Биомедицинская химия. 2013. Т. 59. № 6. С. 622-635.
2. Федюшкина И.В., **Ромеро Рейес И.В.**, Лагунин А.А., Скворцов В.С. Предсказание спектра действия лигандов рецепторов стероидных гормонов // Биомедицинская химия. 2013. Т. 59. № 5. С. 591-599.
3. **Romero Reyes I.V.**, Fedyushkina I.V., Skvortsov V.S., Filimonov D.A. Prediction of progesterone receptor inhibition by high-performance neural network algorithm // International Journal of Mathematical Models and Methods in Applied Sciences. 2013. V. 7. P. 304-310.
4. Fedyushkina I.V., **Romero Reyes I.V.** Prediction of Glucocorticoid Receptor Inhibition by High-Performance Neural Network Algorithm // Advances in Mathematical and Computational Methods, 2012. V. 4. P. 203-208.
5. **Ромеро Рейес И.В.** Использование нейронной сети для оценки проницаемости гематоэнцефалического барьера. // Семнадцатая научная конференция молодых ученых и специалистов (ОМУС-2013) к 100-летию В.П. Джелепова. Сборник аннотаций докладов. Дубна. 2013. С. 46.
6. **Ромеро Рейес И.В.** Разработка нейронной сети с использованием графических процессоров для оценки проницаемости гематоэнцефалического барьера. 2012 год // Научный сервис в сети Интернет: поиск новых решений: Труды Международной суперкомпьютерной конференции. Новороссийск. 2012. С. 699.
7. **Ромеро Рейес И.В.**, Скворцов В.С., Филимонов Д.А. Использование нейронной сети для оценки проницаемости гематоэнцефалического барьера. // Материалы XVI Международной конференции по нейрокибернетике. Ростов-на-Дону. 2012. Т. 2. С. 188-190.
8. **Ромеро Рейес И.В.**, Чернобровкин А.Л. Реализация идентификации пептидов по масс-спектрометрическим пикам. // Материалы IV сессии научной школы-практикума молодых ученых и специалистов в рамках VIII Всероссийской межвузовской конференции молодых ученых. Санкт-Петербург. 2011.
9. **Ромеро Рейес И.В.**, Скворцов В.С. Оценка острой токсичности у грызунов с использованием нейросетевых моделей. // Сборник тезисов XVI Российского национального конгресса «Человек и лекарство». Москва. 2010. С. 711.