

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА
ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ

На правах рукописи

Залевский Артур Олегович

**Атомистический механизм катион-зависимой
активации тромбина**

03.01.09 – Математическая биология, биоинформатика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата биологических наук

Научный руководитель:

д.х.н., профессор Головин Андрей Викторович

Москва – 2019

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ	5
ВВЕДЕНИЕ.	6
1 Актуальность	6
2 Цели и задачи	7
3 Научная новизна	7
4 Практическая значимость	9
5 Положения, выносимые на защиту	9
6 Личный вклад автора	10
7 Степень разработанности темы	10
8 Степень достоверности данных	11
9 Публикации по теме диссертации	11
10 Апробация результатов	11
11 Структура диссертации	12
ОБЗОР ЛИТЕРАТУРЫ	13
1 Тромбин	13
1.1 Механизм катализа тромбина	13
1.2 Катион-зависимая активация	17
2 Кластеризация	21
2.1 Метрика сходства	22
2.2 Алгоритмы кластеризации	23
3 Гибридные КМ/ММ вычисления	24

4	Субстратная специфичность.	29
4.1	Предсказание субстратной специфичности	29
4.2	Докинг пептидов	30
	МАТЕРИАЛЫ И МЕТОДЫ	33
1	Кластеризация структур тромбина	33
1.1	Подготовка структур.	33
2	Расчет RMSD	34
2.1	Кластеризация	34
3	Гибридное КМ/ММ моделирование	36
3.1	Подготовка тромбина для КМ/ММ моделирования.	36
3.2	КМ/ММ моделирование тромбина в GROMACS/DFTB3	37
3.3	КМ/ММ моделирование в GROMACS/MOPAC2012	38
3.4	КМ/ММ метадинамика в GROMACS/GFN2-хТВ	42
4	Моделирование фермент-субстратного комплекса тромбина	42
5	Реконструкция профиля субстратной специфичности	43
6	Визуализация	45
	РЕЗУЛЬТАТЫ	46
1	Кластерный анализ структур тромбина	46
1.1	Кластеризация структур.	46
1.2	Анализ и аннотация кластеров	51
2	КМ/ММ моделирование GROMACS/DFTB3.	55
3	КМ/ММ моделирование GROMACS/MOPAC2012	57
3.1	Тестовые системы	60

3.2	Вычислительное время	63
3.3	Воспроизводимость	65
3.4	Сольватация ионов в GROMACS/MOPAC2012	66
4	QM/MM моделирование GROMACS/GFN2-xTB	67
5	Моделирование взаимодействия с субстратом.	73
6	Предсказание субстратной специфичности	75
6.1	Разработка и оптимизация алгоритма PeptoGrid.	75
6.2	Докинг пептидных библиотек	82
ОБСУЖДЕНИЕ.		85
1	Консистентность базы данных	85
2	Кластеризация структур тромбина	85
3	Ингибирующая водородная связь	90
4	КМ/MM моделирование	90
5	Моделирование комплекса с субстратом.	94
6	Предсказание субстратной специфичности	95
7	Уникальность тромбина.	96
ВЫВОДЫ		98
ЗАКЛЮЧЕНИЕ.		99
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ.		100
ПРИЛОЖЕНИЕ		124

СПИСОК СОКРАЩЕНИЙ

- АПФ** — ангиотензинпревращающий фермент
- ГАМКБ** — рецептор γ -аминомасляной кислоты класса Б
- КМ** — квантово-механические расчеты
- КМ/ММ** — гибридное квантово-механическое/молекулярно-механическое моделирование
- МД** — молекулярная динамика
- ПО** — программное обеспечение
- РСА** — рентгено-структурный анализ
- ЯМР** — ядерно-магнитный резонанс
- API** — программный интерфейс приложения (англ. application programming interface)
- DFT** — теория функционала плотности (англ. density functional theory)
- MHP** — молекулярный гидрофобный потенциал (англ. molecular hydrophobicity potential)
- PDB** — Protein Data Bank, банк данных пространственных структур биомолекул
- RMSD** — среднеквадратичное отклонение позиций атомов (англ. root-mean-square deviation of atomic positions)
- MPI** — интерфейс передачи сообщений между программами (англ. message passing interface)

ВВЕДЕНИЕ

1. Актуальность

Тромбин — ключевой фермент каскада свертывания крови, что делает его привлекательной фармакологической мишенью как для академических исследователей, так и для представителей фармацевтической индустрии. На сегодняшний день накоплен поражающий воображение объем биохимических, генетических и структурных данных, демонстрирующий связь тромбина не только с гемостазом, но и с регуляцией воспаления и регенерации. Все это делает тромбин, упомянутый более чем в 50 000 статей в базе данных PubMed, одним из самых исследованных ферментов. Что, тем не менее, не означает, что мы во всех деталях понимаем все аспекты его работы. Одним из таких белых пятен является механизм катион-зависимой активации тромбина. В отличие от большинства представителей семейства химотрипсиновых протеаз, активность тромбина зависит от присутствия ионов натрия — Na^+ . Причем появление катиона в натрий-связывающем сайте затрагивает как кинетику фермента, увеличивая скорость реакции расщепления белковых и пептидных субстратов почти на порядок, так и субстратную специфичность. Более того, из-за весьма умеренной степени сродства белка к натрию, при физиологических условиях только 60% белка находится в активной “быстрой” форме. Остальные же

40%, как считается, находятся в “медленной” форме, проявляющей антикоагулянтную активность. Таким образом, детальное понимание особенностей обеих форм и механизмов переключения между ними может быть использовано при создании антикоагулянтных препаратов нового поколения.

2. Цели и задачи

Целью данной работы является выявление атомистических деталей механизма влияния катиона натрия на ферментативную активность тромбина.

Для достижения обозначенной цели необходимо решить следующие задачи:

- Проанализировать структуры человеческого тромбина, доступные в банке структур биомолекул PDB, и выделить структурный признак, характерный для быстрой и медленных форм фермента;
- С помощью методов молекулярного моделирования продемонстрировать влияние катиона натрия на поведение обнаруженного признака;
- Продемонстрировать, что переход между формами также вызывает изменение субстратной специфичности фермента.

3. Научная новизна

В данной работе впервые произведен автоматический кластерный анализ всех структур человеческого тромбина, до-

ступных в банке данных PDB. Полученный результат позволил сформулировать гипотезу о наличии ранее не наблюдавшейся водородной связи в активном центре фермента. Для проверки данной гипотезы были специально разработаны инструменты, использующие различные методы молекулярного моделирования: от мультиуровневого гибридного квантово-механического/молекулярно-механического, до пептидного докинга. Все инструменты исходно разрабатывались из соображений оптимизации для работы в массивно-параллельном суперкомпьютерном окружении.

С помощью разработанных инструментов удалось продемонстрировать наличие ингибирующей водородной связи Ser195 OG — Ser214 O в активном центре фермента, отсутствующей в системе с катионом натрия. Продемонстрирована роль сети молекул воды, соединенных водородными связями, в передаче структурирующего сигнала от натрий-связывающего кармана к активному центру фермента.

Наконец, с помощью моделирования молекулярного докинга комбинаторных пептидных библиотек, продемонстрировано, что появление ингибирующей водородной связи также кардинальным образом изменяет профиль специфичности фермента.

Таким образом, совокупность наблюдений, сделанных в ходе вычислительных экспериментов, позволяет расширить представление о медленной форме тромбина, добавив к нему

новое состояние, обладающее характерной ингибирующей водородной связью.

4. Практическая значимость

Данные о новой медленной форме тромбина могут быть использованы при разработке про- или антикоагулянтных препаратов, действие которых будет основано на целенаправленном воздействии на ингибирующую водородную связь. Такие молекулы должны будут избирательно связываться с медленной формой тромбина и фиксировать в ней белок, либо через аллостерические взаимодействия нарушать сеть водородных связей между натрий-связывающим сайтом и активным центром.

5. Положения, выносимые на защиту

1. Новая реализация непараметрического алгоритма кластеризации Affinity Propagation, оптимизированная для структур биомолекул, позволяет выявлять биологически релевантные кластеры.

2. Сравнение кластеров, полученных из 400 записей тромбина в банке структур PDB, выявило возможность существования ранее не описанной ингибирующей водородной связи между O γ каталитического Ser195 и остоным кислородом Ser214.

3. С помощью нового пакета для гибридного КМ/ММ моделирования продемонстрировано влияние наличия иона Na⁺ на образование ингибирующей водородной связи. Дополнительный энергетический барьер для переключения ингибирующей

водородной связи в присутствии пептидного субстрата составляет ~ 0.8 ккал/моль, что соответствует экспериментальным данным о замедлении реакции в 3-10 раз.

4. Предложенный метод ранжирования аффинности библиотек пептидов к белковым мишеням позволяет идентифицировать новые биоактивные пептиды.

5. Образование ингибирующей водородной связи изменяет субстратную специфичность тромбина.

6. Личный вклад автора

Личный вклад автора заключается в разработке инструментов для биоинформатического анализа пространственных структур и гибридного квантово-механического/молекулярно-механического моделирования тромбина, биологической интерпретации полученных результатов, представлении результатов на научных конференциях, публикации в рецензируемых научных журналах.

7. Степень разработанности темы

С момента открытия феномена зависимости активности тромбина от концентрации катионов натрия, в этой области было напечатано большое количество экспериментальных и теоретических работ, получено много качественных структурных данных. Наибольший вклад в описание особенностей отличия быстрой и медленной форм тромбина внесли профессоры E. Di Cera и J. Huntington. Большое количество кристаллографических

данных получено группами V. De Filippis, E. Di Cera, J. Huntington, G. Klebe.

В область разработки быстрых квантово-химических методов большой вклад внесли профессора J. Stewart и S. Grimme.

8. Степень достоверности данных

Данные, представленные в работе, получены с использованием современных вычислительных методик. Результаты воспроизводимы, исходные коды разработанных инструментов опубликованы. Обзор литературы и обсуждение подготовлены с использованием актуальной литературы.

9. Публикации по теме диссертации

По материалам диссертации опубликовано 5 статей в рецензируемых журналах, индексируемых в наукометрических базах данных Web of Science или Scopus.

10. Апробация результатов

Результаты диссертационной работы были представлены на международных конференциях: 12th INTERNATIONAL CONFERENCE “BIOCATALYSIS. FUNDAMENTALS & APPLICATIONS” (Россия, теплоход Леонид Соболев, 2019), Russian Supercomputing Days (Россия, Москва, 2018), Информационные технологии и системы (ИТиС-18) (Россия, Казань, 2018), Bioinformatics of Genome Regulation and Structure/BGRS-2016 (Россия, Новосибирск, 2016).

11. Структура диссертации

Диссертационная работа состоит из введения, обзора литературы, методов и материалов исследования, полученных результатов и их обсуждения, выводов и списка цитируемой литературы. Работа изложена на 125 страницах, иллюстрирована 32 рисунками и 5 таблицами. Список цитируемой литературы включает 145 наименований.

ОБЗОР ЛИТЕРАТУРЫ

1. Тромбин

Система свертывания крови является одним из результатов эволюции системы комплемента и иммунного ответа и важным элементом защитного комплекса организма [1]. Ключевым ферментом каскада свертывания крови является сериновая протеаза — тромбин [2] (Рис. 1). Тромбин образуется из неактивного предшественника — протромбина в результате протеолитического расщепления [3], при этом в качестве интермедиатов могут образовываться мейзотромбин [4], претромбин-1 [5] или претромбин-2 [6].

Тромбин — глобулярный белок, состоящий из двух цепей — “легкой” и “тяжелой” (36 и 259 аминокислотных остатков, соответственно), соединенных дисульфидным мостиком между аминокислотными остатками Cys1 “легкой” и Cys122 “тяжелой”.

1.1. Механизм катализа тромбина

Активный центр фермента представлен канонической каталитической триадой — His57, Asp102, Ser195 [3]. Весь каталитический акт гидролиза пептидной делится на две стадии: ацилирование и деацилирование (Рис. 2). Во время первой стадии происходит перенос протона с каталитического серина на гистидин с образованием нуклеофила, который затем атакует карбонильный углерод разрываемой пептидной связи. В результа-

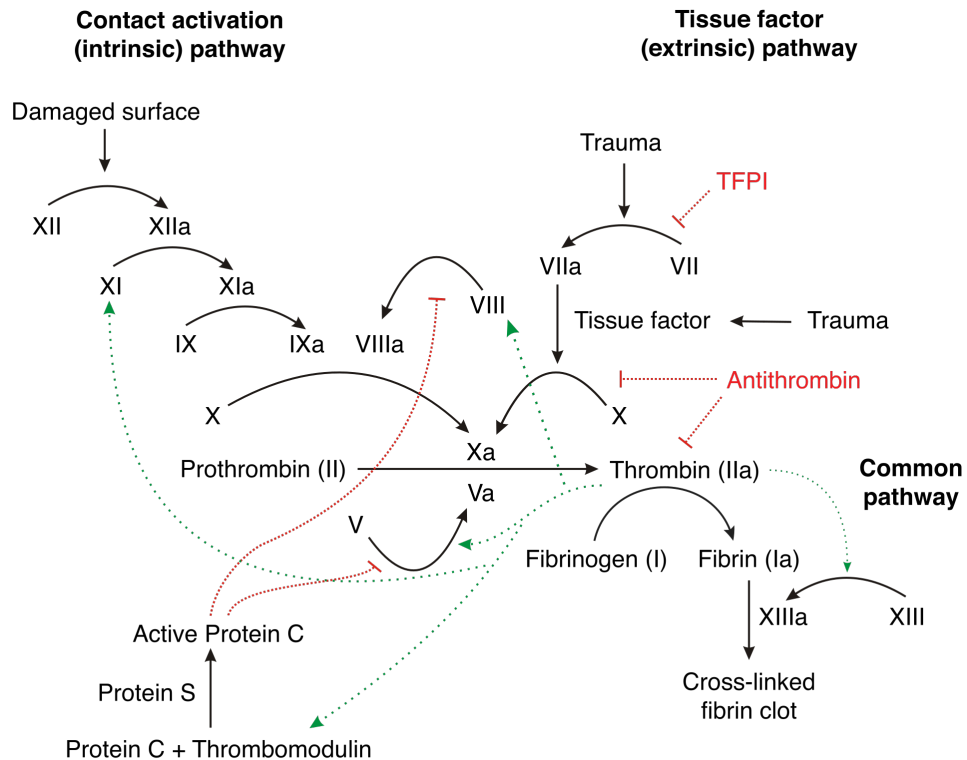


Рисунок 1. Схема каскада свертывания крови человека [7].

те образуется ковалентный конъюгат, а уходящая группа захватывает протон с гистидина. Во время второй стадии нуклеофилом выступает молекула воды, а на последнем шаге каталитический серин возвращает себе протон с гистидина [8]. Несмотря на то, что основные события происходят между серином и гистидином, аспартат также является участником триады, он образует низкобарьерную водородную связь с гистидином, что приводит к сдвигу pK_a гистидина, позволяя ему принимать дополнительный протон с серина [9, 10].

В непосредственной близости от активного центра также расположены области, участвующие в связывании субстрата и его расщеплении (Рис. 4). Первая из них — оксианионный центр, сформированный основными атомами азота остат-

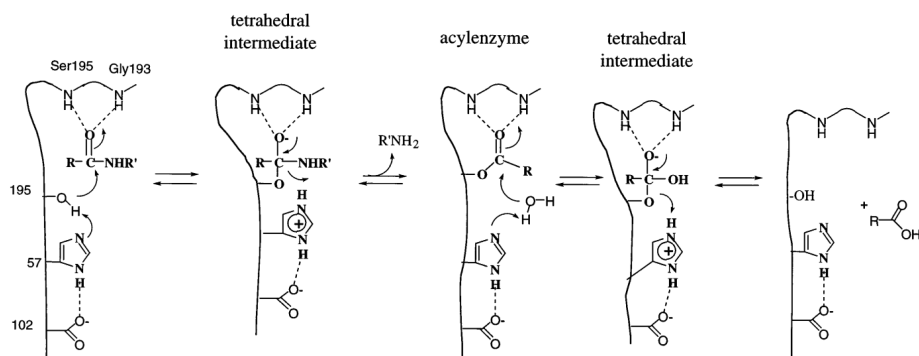


Рисунок 2. Схема механизма катализа сериновых протеаз.
 Воспроизведено из Chem. Rev.2002, 102, 12, 4501-4524 [8].
 ©American Chemical Society (2019).

ков Ser195 и Gly193 и стабилизирующий переходное состояние в ходе протеолитического расщепления субстрата. Вторая — S1 карман специфичности, расположенный между 180-й и 220-й петлями. В глубине этого кармана расположен Asp189, образующий солевой мостик с P1 остатком аргинина субстрата. Гидрофобный карман S2, сформированный остатками 60й петли: Tyr60A, Pro60B, Pro60C, Trp60D, имеет сродство к небольшой гидрофобной боковой цепи остатка в позиции P2 (Рис. 3). Арил-связывающий сайт, образованный остатками Leu99, Ile174 и Trp215, взаимодействуют с ароматическими и гидрофобными остатками в позиции P4, в то время как остаток субстрата в позиции P3 обращен кнаружи от поверхности белка.

На поверхности глобулы тромбина также находятся две положительно заряженные области связывания с лигандами: экзосайты I и II. Экзосайт I расположен в области 70-й петли и, в дополнение к заряду, имеет гидрофобный характер. Его назначение — связывание и ориентация фибриногена, основного суб-

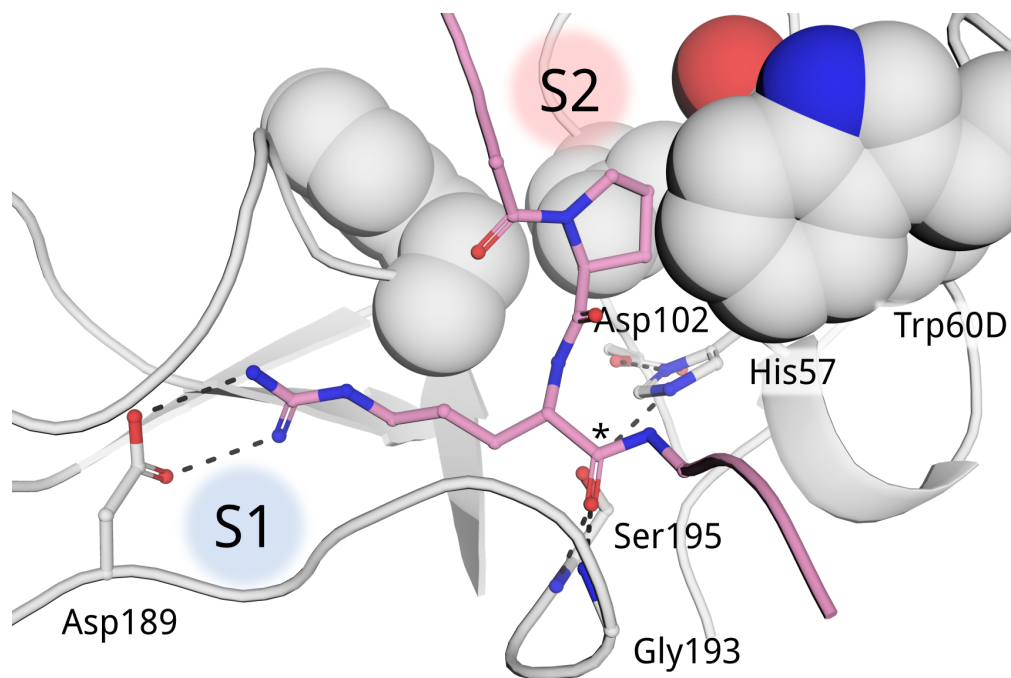


Рисунок 3. Расположение субстрат связывающих карманов S1 и S2. Сферами отмечены атомы радикалов аминокислот, формирующих карман S2. Пунктирными линиями отмечены водородные связи. Символом * отмечен атакуемый атом субстрата.

страта тромбина, участвующего в формировании тромба [11]. Помимо фибриногена, по этому же экзосайту связывается тромбомодулин — мембранный белок, рецептор тромбина, который в составе комплекса с тромбином также участвует в каскаде свертывания крови, активируя протеин С [12]. Кроме него, по экзосайту I связывается природный пептидный ингибитор — гирудин [13], выделенный из секрета медицинской пиявки *Hirudo medicinalis*, и протеазо-активируемый рецептор 1 (PAR 1) [14]. Экзосайт II, расположенный на противоположной стороне белка, включает в себя С-терминальную альфа-спираль и окружающие ее остатки 99-й и 170-й петель. Он имеет более основную природу и взаимодействует с полианионными лигандами, такими

как гепарин — серосодержащий гликозаминогликан, используемый как прямой антикоагулянт, и другие гликозаминогликаны [15].

В структуре белка выделяется несколько петель, расположенных в непосредственной близости от активного центра фермента: 37-я, 60-я, 99-я и 148-я — считается, что они участвуют во взаимодействии с субстратами, а последняя также выполняет роль ауторегуляторного сайта [16].

1.2. Катион-зависимая активация

Помимо упомянутых сайтов, большое значение имеет сайт связывания Na^+ , образованный 180-й и 220-й петлями и участвующий в регуляции активности фермента. Катион Na^+ в нем координируется октаэдрически: карбонильными атомами кислорода остатков Arg221A и Lys224, четырьмя молекулами воды, ориентированными боковыми цепями Asp189 и Asp221, и основными атомами Gly223 и Tyr184A. Na^+ -свободный тромбин имеет низкую активность по отношению к некоторым хромогенным субстратам *in vitro*. Связывание натрия повышает протеолитическую активность тромбина приблизительно в 10 раз [17].

Натрий-свободный тромбин существует в равновесии между двумя формами, одна из которых не способна к дальнейшему связыванию катиона, а также некоторых ингибиторов по активному сайту и пептидов по экзосайту I. Вторая форма способна к связыванию катиона и упомянутых лигандов. Ниже приведена

схема, отражающая равновесие между формами *in vivo*: [18]



где E^* — форма, не способная к связыванию натрия, E — медленная форма, способная к связыванию лигандов, $E : Na^+$ — быстрая форма, стабилизированная катионом

При этом “ E^* ” форма представляет собой ансамбль различных конформаций, в отличие от двух других, имеющих “жесткую” структуру и практически не отличающихся между собой [19].

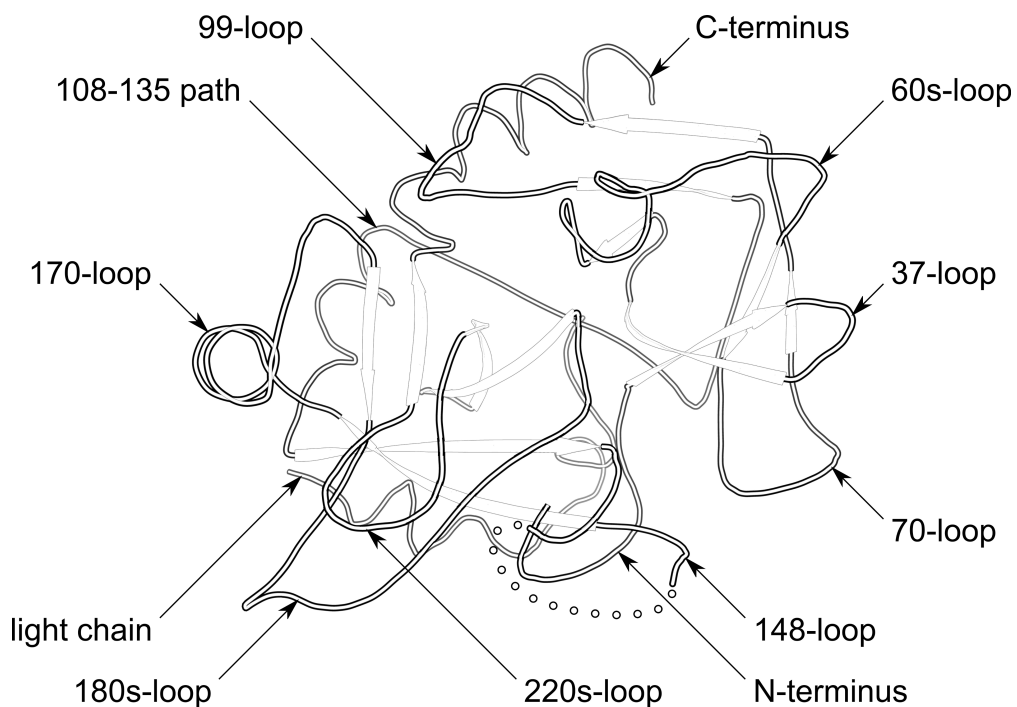


Рисунок 4. Топология тромбина. Молекула тромбина изображена в стандартной ориентации Бодэ (Bode) [20]. Петли изображены в упрощенном виде и выделены темным цветом. Отсутствующие остатки 148-й петли изображены окружностями в нижней части рисунка.

Считается [19], что представителями “*E*” формы Натриесвязывающих структур являются записи PDB 1SGI [21] и 1HXF [22], а “*E**” форма представлена записями 1RD3 [23], 2AFQ [24] и 2GP9 [25]. При этом к ключевым событиям, происходящим при связывании иона натрия, относят непосредственно изменение положения остатков в натрий-связывающем кармане [3], а также переориентацию ароматических остатков, в частности Trp215 в карманах S2 и S3, что существенно уменьшает размер полости активного центра [26]. Выделяют также и менее амплитудные изменения, затрагивающие остатки Asp189 и Glu192, отвечающие за специфичность в S1 и S1/S2 положениях соответственно [3]. Другие структуры, относящиеся к *E* и *E** формам, преимущественно представлены мутантами, например D102N или D189A, и демонстрируют более существенные изменения структуры. Например, в случае D189A, происходит коллапс S1 субстрат-связывающего кармана, что приводит к невозможности продуктивного связывания субстрата [27]. В случае мутанта D102N по каталитическому аспартату, Trp215 занимает полость активного центра, Arg221a смещается более чем на 10 Å, а Arg187 проникает внутрь натрий-связывающего сайта и занимает его место [25]. В целом же, изменения, сопряженные со связыванием натрия, обычно характеризуют как “затвердевание” структуры. Соответственно, все структуры быстрой формы должны быть достаточно однообразны. Медленная форма, по всей видимости, представлена ансамблем

конфигураций, которые составляют сложный конформационный ландшафт и находятся в динамическом равновесии. При этом они могут сочетать уже описанные особенности, но, возможно, иметь и уникальные, еще не описанные [28].

В Protein Data Bank [29] находится более 400 записей, содержащих трехмерные структуры человеческих тромбина и протромбина. Такому пристальному вниманию со стороны исследователей тромбин обязан своей клинической значимостью: тромбин не только является ключевой точкой каскада свертывания крови [3], но и важным элементом регуляции процессов воспаления [30] и регенерации [31], в том числе при травмах головного мозга [32]. Такое огромное функциональное разнообразие делает его популярной фармакологической мишенью для разработки как ингибиторов — антикоагулянтов, так и активаторов — прокоагулянтов, а также тестовых систем активности тромбина (например, для детекции внутренних кровотечений) самых различных конструкций. На сегодняшний день многообразие антикоагулянтов покрывает практически все классы веществ: от малых молекул до пептидов и даже нуклеиновых кислот [33, 34]. В свою очередь, сенсоры и системы *in vivo* визуализации тромбина используют весь доступный арсенал физико-химических методов: от ультрачувствительных флуоресцентных методов [35, 36], до электро-химических сенсоров на основе углеродных нанотрубок или графена [36].

Тем не менее, такое обилие структурной информации, на-

копленной за несколько десятилетий, требует классификации и систематизации. Последний максимально полный разбор этого многообразия был предпринят в 2008 году профессором Ди Сера (Di Cera), соавтором более пятидесяти из обсуждаемых структур [3]. Однако информация продолжает накапливаться и новые попытки систематизации не могут опираться только на экспертное знание. А потому они требуют применения новых вычислительных подходов, которые, возможно, помогут обнаружить особенности структур, ранее не замеченные исследователями.

2. Кластеризация

Одним из таких подходов является кластерный анализ. В его задачи входят [37]:

- группировка объектов на основе некоторой выбранной меры сродства (метрики);
- поиск новизны, то есть объектов, которые нельзя присоединить ни к одной из существующих групп.

Применение кластерного анализа к существующим структурным данным тромбина, в перспективе, может позволить выбрать из большого количества трехмерных структур небольшое подмножество, которое будет описывать основные конформации белка. Подобный набор структур, в свою очередь, может быть использован для оптимизации процесса разработки лекарственных препаратов при помощи интенсивного использования вычислительных методов, таких как молекулярный до-

кинг и моделирование молекулярной динамики.

2.1. Метрика сходства

Кластерный анализ строится на предположении, что, имея некоторую метрику, описывающую меру сродства двух структур, и применяя специальные алгоритмы, можно разделить множество объектов на набор подмножеств [38] объектов, схожих с друг с другом в терминах выбранной метрики. Поэтому выбор метрики, равно как и алгоритма, имеет решающее значение для получения корректных результатов.

Одной из наиболее широко используемых метрик в структурной биоинформатике является RMSD — среднеквадратичное отклонение [39], рассчитываемое согласно выражению (Выражение 2). В случае белков, чаще всего используют значение RMSD, вычисленное по C_{α} атомам, значительно реже — по атомам остова (N, C, C_{α} , и иногда — O) [40] :

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta^2} \quad (2)$$

где δ — расстояние между парой эквивалентных атомов, N — количество атомов.

Однако повсеместное использование этой метрики сопряжено с рядом нюансов: зачастую авторы, подразумевая C_{α} , не указывают, по какому набору атомов или при помощи какого инструмента вычислялось значение. Оба этих нюанса имеют определяющее значение: у структур с одинаковым ходом осто-

ва могут сильно отличаться конформации боковых цепей, а для выравнивая трехмерных структур существует целый ряд алгоритмов [41, 42], результаты работы которых могут отличаться.

К недостаткам RMSD часто относят то, что оно чувствительно к незначительным отклонениям позиций атомов, например, наличие подвижных петель нередко приводит к большим значениям. Это является существенным недостатком при решении стандартных задач поиска близких структур в массиве разнородных объектов: при анализе фолдинга белков и нуклеиновых кислот, при поиске белков с одинаковыми структурными мотивами и т.д. [43]. Для преодоления этой особенности было разработано большое количество различных метрик: GDT [44], TM-Score [45], MaxSub [46] и другие. Как правило, для устранения влияния небольших отклонений, они используют сравнение только по опорным атомам, например C_{α} , а также применяют большое количество нормировочных параметров, чтобы нивелировать эффекты различной длины последовательностей, атомной массы и общего размера, определяемого гидродинамическим радиусом (или радиусом гирации).

2.2. Алгоритмы кластеризации

Вторым важным шагом при кластеризации является выбор подходящего алгоритма. Ранее, для решения подобного рода задач, возникающих, например, в ходе фолдинга белков, применялись разнообразные алгоритмы, от К-средних [47], до различ-

ных вариантов иерархических алгоритмов [48].

Однако эти подходы подразумевают, что исходно имеется представление о структуре данных: количестве групп, их размерах, гипотетических центрах кластеров и т.д., либо же они требуют подбора параметров, которые порой связаны с данными весьма косвенным образом или же просто подбираются эмпирически. К тому же некоторые алгоритмы, например алгоритм К-средних, зависят от выбора стартовых параметров, поэтому для достижения качественного результата необходимо производить множество запусков со случайной инициализацией и последующим отбором консенсусных результатов. При использовании иерархических методов кластеризации неизбежно возникает необходимость разбиения дерева на отдельные кластеры. Разбиения на основе предварительно заданных абсолютных длин ветвей или количества объектов в кластере часто не учитывают внутреннюю структуру кластеров, которые могут сильно различаться в случае работы с биологическими объектами [49].

Альтернативой могут служить непараметрические алгоритмы [50, 51]. В этом случае, полученный результат зависит только от самих данных.

3. Гибридные КМ/ММ вычисления

В данном разделе используются материалы из [52].

Кластерный анализ сам по себе вряд ли сможет перевер-

нуть или хотя бы существенно изменить представления о структурных особенностях тромбина. Тем не менее, он способен предоставить максимально релевантные стартовые точки для методов молекулярного моделирования. Например, гибридного квантово-механического/молекулярно-механического (КМ/ММ) моделирования.

Гибридное КМ/ММ моделирование было предложено в 1976 году [53], но лишь недавно стало активно использоваться для моделирования молекулярной динамики (МД), став популярным инструментом для изучения биомолекулярных систем. Сочетание принципиально разных подходов к моделированию биомолекул, основанных на разных физических моделях, может происходить разными способами, но одной из самых популярных и технически простых схем является ONIOM.

В рамках подхода ONIOM [54] можно разделить систему на несколько слоев с независимым описанием взаимодействий внутри индивидуального слоя. В случае КМ/ММ градиенты силы сначала оцениваются для изолированной квантово-механической (КМ) подсистемы с использованием выбранной полуэмпирической или *ab initio* модели. Затем градиенты и полная потенциальная энергия системы рассчитываются с использованием соответствующего силового поля ММ и добавляются к полученным для изолированной подсистемы КМ. Наконец, чтобы избежать дублирования вклада от подсистемы КМ, выполняется расчет молекулярной механики для изолированной об-

ласти КМ, и результат вычитается из общей суммы (Выражение 3):

$$E_{tot} = E_I^{QM} + E_{I+II}^{MM} - E_I^{MM} \quad (3)$$

где индексы *I* и *II* относятся к подсистемам КМ и ММ соответственно. Верхние индексы указывают на каком уровне теории рассчитываются энергии.

Физическое разделение между системами достигается введением связующих атомов (LA), которые представлены в виде водородов в квантовой части и не вносят дополнительных взаимодействий в механическую часть [55].

Сегодня с помощью методов мультимасштабного моделирования можно изучать поведение биомолекул на временных масштабах нано- и даже миллисекунд, где тепловое движение может внести существенный вклад в химическую реактивность и конформационное пространство системы [56]. Дополнительный импульс развитию гибридных приложений КМ/ММ в исследованиях биосистем дал Паринелло (Parinello) и его коллеги, которые разработали метод под названием “метадинамика” [57–59]. Этот метод позволяет сканировать конформационное пространство биомолекулярных систем в поисках редких событий, таких как реакции, катализируемые ферментами [60].

Как правило, при моделировании ферментативных реакций с использованием метадинамики, квантовая подсистема

включает десятки (реже сотни) атомов, и для моделирования теплового движения системы могут быть сделаны сотни тысяч шагов расчета энергетических градиентов и оптимизации геометрии. Это требует значительных вычислительных ресурсов даже при использовании относительно дешевых методов квантовой механики (КМ), что накладывает ограничение на применимость различных инструментов.

Поэтому такие методы, как полуэмпирические квантово-механические методы семейства PM (PM3, PM6, PM6-D3H4 с поправками для водородной связи и др.), реализованные в пакете MORAS2012 и позволяющие с приемлемой точностью моделировать большой набор элементов и соединений разной химической природы [61], выглядят чрезвычайно привлекательно для данной задачи. В то же время существуют и более сложные методы, основанные на упрощенных модификациях теории функционала плотности (DFT — density functional theory). Например, метод DFTB3, реализованный для наиболее распространенных в биомолекулах элементов C, H, N, O и P, превосходит методы PM6 по точности оценки энергий водородных связей [62] и взаимодействий с участием протонов. Другим представителем этого класса является новое семейство методов общего назначения GFN-xTB, оптимизированных для корректного воспроизведения геометрий, колебательных свойств, а также энергий нековалентных взаимодействий [63, 64]. Благодаря набору оптимизаций, профессору Гримме (Grimme) и коллегам удалось создать

быстрый квантово-химический метод, покрывающий большую часть периодической таблицы и лишенный необходимости параметризации попарных взаимодействий элементов, как это реализовано в DFTB3. Отличительной особенностью GFN2-хТВ в семействе методов GFN-хТВ является использование анизотропической модели электростатических взаимодействий, что должно существенно улучшить энергии нековалентных взаимодействий, в частности, водородных связей и сольватации ионов [63, 64].

Оптимизация вычислительных инструментов для КМ/ММ моделирования является ключевым фактором, влияющим на прогресс в разработке новых ферментов и других биополимеров с желаемыми свойствами.

Сегодня существует множество программных пакетов, применяемых для такого гибридного мультиуровневого моделирования. Популярный программный пакет для моделирования биомолекул GROMACS [65] обладает интерфейсом КМ/ММ для нескольких программных инструментов квантовой химии. Существенным недостатком исходной реализации интерфейса является то, что обмен данными между программами происходит через файловую систему. Это накладывает существенное ограничение на производительность при использовании в кластерных вычислительных системах с распределенной средой хранения [66].

4. Субстратная специфичность

Будучи чрезвычайно мощным инструментом, КМ/ММ расчеты способны предоставить недоступную ранее детализацию процессов, происходящих в активном центре фермента. Однако они, как правило, не могут ответить на вопросы о том, как именно формируется фермент-субстратный комплекс. А этот процесс также может таить в себе много загадок. В частности, некоторые субстраты имеют несколько вариантов связывания в активном центре, и их учет может чрезвычайно сильно усложнить общепринятые кинетические схемы, описывающие поведение ферментов [60, 67].

Например, одной из особенностей быстрой и медленной форм тромбина считается изменение субстратной специфичности фермента — медленная имеет большее сродство к белку С, демонстрируя прокоагулянтное поведение, в то время как быстрая — к фибриногену, выполняя антикоагулянтную функцию [3]. При этом сами сайты расщепления чрезвычайно похожи.

4.1. Предсказание субстратной специфичности

На сегодняшний день наиболее эффективными и продуктивными способами изучения субстратной специфичности протеаз остаются комбинаторные библиотеки пептидов, получаемых синтетически [68], генерируемых при помощи технологии фагового дисплея [69], встроенных в конструкции с сигнальными пептидами [70] или генетическими конструкциями, акти-

вируемыми после актов протеолиза [71]. Однако у экспериментальных методов есть целый ряд недостатков: сложно обеспечить равномерную представленность всех объектов из библиотеки; поведение пептидов в составе белка и в изолированном виде может существенно различаться; необходимо получения эффективной системы синтеза активной формы протеазы; ограниченность воспроизводимого диапазона условий (например диапазон pH). Что в целом осложняет их массовое применение.

Вычислительные подходы, как правило, опираются на предварительные экспериментальные данные [72] или экспертное мнение [73].

Подходы без использования априорного знания о свойствах мишени (или *ab initio*), основанных исключительно на моделировании молекулярного докинга, встречается чрезвычайно редко [74].

4.2. Докинг пептидов

В данном разделе используются материалы из [75].

Вычислительные методы представляют собой экономически эффективный инструмент для высокопроизводительного скрининга пептидных библиотек [76], среди которых подход виртуального скрининга на основе докинга является наиболее широко используемым [77]. Однако у него есть ряд ограничений. Из-за большого количества торсионных углов, перебор всего пространства конформаций пептидов невозможен. Поэтому

на практике, особенно в контексте комбинаторных библиотек, пределом для современных инструментов являются тетрапептиды [78]. Аккуратный докинг более длинных пептидов является существенно более вычислительно затратной операцией, нередко требующей применения методов моделирования молекулярной динамики [79].

Другой значимой проблемой являются недостатки современных методов оценки качества полученных комплексов [80]. Для простоты вычисления, функции скоринга аффинности обычно являются оценочными, а не производят точного вычисления энергии связывания, что приводит к неточностям в ранжировании результатов [81]. Причина снова кроется в большой вычислительной стоимости данных методов, например, метадинамики с использованием вороночного потенциала [82]. Для облегчения скрининга больших молекулярных библиотек часто используется многоуровневая схема оценки, в которой в качестве быстрого фильтра библиотеки используется простая функция оценки, а более точная и вычислительно дорогая функция оценки применяется для пересчета верхних попаданий для получения окончательного результата. Соответственно, на разных уровнях могут использоваться самые разные методы. Например, методы MM/PBSA и MM/GBSA, позволяющие на основе данных коротких траекторий, полученных методами молекулярной механики, оценить вклад энергии сольватации в общую энергию связывания лиганда [83, 84]. Также применяются мето-

ды, основанные на машинном обучении [85, 86] и метрики, вычисляющие консенсное решение из набора отдельных предсказаний [87].

Надо отметить, что в зависимости от природы исходной функции, некоторые из них могут улучшены. Одна из возможных стратегий — более аккуратный учет гидрофобных взаимодействий, например, с помощью техники молекулярного гидрофобного потенциала (МНП — molecular hydrophobicity potential) [88]. Другая интересная стратегия — учет информации всего ансамбля поз конкретного лиганда или даже всей библиотеки как в случае протокола RosettaLigandEnsemble, который основывается на предположении, что в серии малых молекул с общим химическим скелетом он должен быть расположен в сайте связывания идентичным образом [89].

Таким образом, методы молекулярного докинга могут быть ценным источником информации о субстратной специфичности.

МАТЕРИАЛЫ И МЕТОДЫ

1. Кластеризация структур тромбина

1.1. Подготовка структур

Из банка данных структур биомолекул PDB (по состоянию на 01.05.2019 г.) было получено 400 записей, содержащих человеческие тромбин и про- и претромбин. Одна из них, ЗНТС [90], была исключена из рассмотрения, так как содержала только $C\alpha$ атомы.

Все атомы, не относящиеся непосредственно к тромбину были удалены, а записи, содержащие более одной молекулы белка, были разделены на части так, чтобы каждая содержала ровно одну молекулу. При этом имена новых файлов соответствуют формату XXXX_CC, где XXXX — идентификатор в базе данных PDB (PDB ID) исходной записи, а CC — идентификаторы цепей, попавших в этот файл. В результате было получено 462 индивидуальных структуры.

В случае наличия в PDB альтернативных позиций атомов (alternative locations), использовались позиции с индексом “A”. Все операции по подготовке структур были выполнены при помощи скриптов на языке программирования Python с использованием библиотеки ProDy [91]

2. Расчет RMSD

Попарные полноатомные значения RMSD по тяжелым атомам (N, C, O, S) были посчитаны при помощи утилиты *Mimic* из пакета ProFASI [92]. *Mimic* сопоставляет последовательности и структуры белков, и на выходе предоставляет значение RMSD и список соответствия атомов двух поданных на вход структур.

Так как каждое вычисление независимо от остальных, расчет был оптимизирован за счет применения параллельных вычислений при помощи утилиты GNU Parallel [93]. Это позволило существенно уменьшить время вычислений.

2.1. Кластеризация

Матрица сходства S была рассчитана согласно (Выражение 4):

$$S_{i,j} = -1 * RMSD_{i,j} \quad (4)$$

где $RMSD(i, j)$ — ранее рассчитанное значение RMSD для пары структур i, j

Последняя часть выражения (4) всегда ≥ 1 , таким образом накладывается штраф на плохо выровненные структуры.

Для кластеризации использовалась реализация алгоритма Affinity Propagation [94] из пакета Scikit [95]. При этом значения предпочтений (preference), влияющие на возможность точки стать центром кластера, были поставлены в зависимость от разрешения структур согласно (Выражение 7):

$$P_0 = \text{median}(S) \quad (5)$$

$$P_i = P_0 * \left\langle \frac{\min(\text{len}(i), \text{len}(j))}{\text{len}(\text{match}(i, j))} \right\rangle \quad (6)$$

где P_0 — базовое значение, рассчитанное из (Выражение 5), $\text{len}(i), \text{len}(j)$ — длины соответствующих структур, $\text{len}(\text{match}(i, j))$ — количество совпавших атомов согласно выдаче Mimicq

$$P_i = P_i * \frac{\text{resolution}(i)}{\min(\text{resolution}(\text{all}))} \quad (7)$$

где $\text{resolution}(i)$ — разрешение структуры i , $\min(\text{resolution}(\text{all}))$ — минимальное разрешение среди всех структур в выборке.

таким образом, структуры с лучшим (меньшим) разрешением более вероятно станут центрами кластеров.

Силуэты кластеров рассчитывались при помощи инструментов `silhouette_score` и `silhouette_samples`, а также из пакета Scikit [95]. Иерархическая кластеризация центров кластеров производилась с помощью инструмента `AgglomerativeClustering` методом полной связности на основе матрицы расстояний, сконструированной на базе соответствующих значений RMSD.

3. Гибридное КМ/ММ моделирование

3.1. Подготовка тромбина для КМ/ММ моделирования

Структуры для моделирования были подготовлены на основе модели 1SG8_AB. Был сохранен ион натрия в натрий-связывающем центре, 22 молекулы воды, присутствующие в кристалле и располагающиеся между катионом натрия и активным центром. Также были сохранены все молекулы воды на расстоянии не более 3.5 Å от белка. Недостающие фрагменты (фрагменты легкой и тяжелой цепей: N и C концы, 148-я петля, недостающие радикалы) были восстановлены на основе структуры 1DX5. Структуры располагались в ячейке, соответствующего размера так, чтобы расстояние до границы ячейки составляло не менее 15 Å. Ячейка заполнялась явно заданным растворителем TIP3P (для упрощения последующего выделения квантово-механической системы). Катионы Na^+ помещались в ячейку случайным образом в концентрации 150 мМ. Для нейтрализации заряда системы, аналогичным образом добавляли необходимое количество анионов. Минимизация энергии производилась с использованием алгоритма наискорейшего спуска до значения силы не более 1000 кДж/моль/нм. В завершение подготовки структур производилась оптимизация контактов с растворителем в течение 100 пс молекулярной динамики 300 К с ограничением подвижности атомов белка и кристаллической воды. Моделирование молекулярной динамики проводили в силовом поле amber99sb-star-ildnp с поправками для

торсионных углов радикалов аминокислот [96], остатков пролина [97] и альфа-спиралей [98]. Шаг интегрирования составлял 2 фс. При расчете электростатических взаимодействий использовался метод суммирования по Эвальду (Particle Mesh Ewald, PME), утряска воды — по алгоритму LINCS. Все расчеты выполнялись с помощью пакета GROMACS [99]. Вычисления траекторий молекулярной динамики осуществлялись на суперкомпьютере “Ломоносов-2”.

3.2. КМ/ММ моделирование тромбина в GROMACS/DFTB3

В квантовую систему выделялись остатки активного центра, фрагменты аминокислот, участвующих в координации натрия, и молекул воды. Всего в квантовую систему было включено 466 атомов. КМ/ММ динамика рассчитывалась с помощью пакета GROMACS/DFTB3. Шаг интегрирования при моделировании реакций составлял 0.2 фс. Предварительно перед любым КМ/ММ моделированием проводилась минимизация энергии методом сопряженных градиентов до достижения значения максимальной силы не более 1000 кДж/моль/нм.

Длина траектории составляла 2.54 пс. Суммарно было проведено 28 независимых запусков для систем с ионом и без. Реконструкция профиля свободной энергии осуществлялась с использованием смешанных гауссовских моделей в пакете Scikit [100]. Количество ядер выбиралось по минимуму значения информационного критерия Акаике.

3.3. КМ/ММ моделирование в GROMACS/МОРАС2012

В данном разделе использованы материалы из [52].

Среда сборки

Компиляция МОРАС2012 и GROMACS 5.0.7 в статически связанные двоичные файлы была выполнена на локальной рабочей станции (Intel®Core TM i7-6700K CPU @ 4,00 ГГц, NVIDIA GTX1070, 32 ГБ ОЗУ, A-Data XPG SX8000NP 128 ГБ SSD, Ubuntu 16.04.4 amd64). Из-за зависимостей МОРАС от библиотек Intel®MKL использовалось обновление 1 для Intel®Composer 2015, упакованное с версией 11.2 MKL. Мы также скомпилировали версии GROMACS с поддержкой графических процессоров (NVIDIA ®CUDA toolkit версия 7.5 для локальной системы и версия 8.0 для суперкомпьютера “Ломоносов-2”). Было подготовлено несколько версий GROMACS/ORCA и GROMACS/МОРАС: одинарная точность, одинарная точность с поддержкой графического процессора и двойная точность (двойная точность с графическим процессором не поддерживается). Поддержка MPI была отключена, а поддержка OpenMP была включена во всех версиях. Мы использовали ORCA версии 4.2.1, распространяемую в виде скомпилированных двоичных файлов.

Версии GROMACS/DFTB и GROMACS/GFN2-хТВ собиралась на базе GROMACS 5.0.7 с двойной точностью, поддержкой MPI и PLUMED 2.5.1 — набора инструментов для проведения моделирования молекулярной динамики с повышенной эффективностью сканирования.

“Малая” система

Исходная конформация ферментативного комплекса белок-лиганд была взята из исследований, описанных ранее [101]. Система моделирования была заполнена молекулами воды TIP3P [102], общий заряд был нейтрализован ионами Na^+ или Cl^- . Вода и ионы были уравновешены вокруг белково-параоксонного комплекса с помощью моделирования в 100 пс с ограниченным положением атомов белка и параоксона. Для подсистемы ММ мы использовали параметры из силового поля parm99 с поправками [103]. Подсистема КМ была описана полуэмпирическим гамильтоновым PM3 [104] и состояла из атомов параоксона, боковых цепей Arg35 и Tyr37 и двух ближайших молекул воды.

“Большая” система

Информация о координатах атомов эстеразы были взяты из PDB ID 1LXW [105]. Недостающие остатки V377, D378, D379, Q380 и C66 были добавлены с использованием записи PDB 2XMD [106] в качестве шаблона. Состояние протонирования восстанавливалось по значениям таблицы с помощью инструмента `texttt pdb2gmx` из пакета GROMACS.

Моделирование молекулярного докинга выполняли с использованием AutoDock Vina [107] для помещения лиганда в активный сайт фермента. В качестве лиганда был выбран экотиофат — фосфорорганический пестицид, являющийся модельным соединением для изучения бутирилхолинэстеразы [108]. Подго-

товка входных файлов PDBQT и обработка вывода были выполнены с помощью AutoDock Tools [109]. Первоначальная структура эхотиофата с частичными зарядами Гастейгера была создана с использованием программного обеспечения Avogadro [110]. Ячейка для докинга содержала цельный белок с отступом 5 Å от краевых атомов. Параметр “exhaustiveness”, который влияет на выборку, был установлен равным 64 из-за небольшого числа торсионных углов. Мы провели 20 независимых запусков, в которых белок оставался неподвижным, в то время как лиганд имел все степени свободы.

Для дальнейших расчетов была выбрана одна случайная конфигурация с расстоянием между атомом фосфора эхотиофата и каталитическим ядром менее 3,5 Å.

Дальнейшее уравнивание было выполнено как для предыдущей системы. Конечная система КМ содержала остовные части Tyr111-Phe115, полные остатки (с NH или CO-частями для концевых аминокислот) из петель Gly193-Gly197 и Met434-Phe439, боковые цепи Gln220, Ser221, Asn319, Glu322, Tyr416, лиганда Ech527 и 3 молекулы воды в области активного центра.

Основной запуск

Подготовленные системы подвергались моделированию КМ/ММ с использованием модифицированного пакета GROMACS/ORCA или GROMACS/MOPAC2012 [65, 101]. Используемый временной шаг составлял 0,2 фс. Температура системы поддерживалась на уровне 310K при помощи термостата Nose-

Hoover. Общая длина моделирования была установлена на 100 шагов, и для каждой системы было рассчитано 10 независимых реплик.

КМ параметры

Для КМ вычислений были использованы следующие параметры (Таблица 1).

Таблица 1 . Параметры квантово-механических вычислений

Для GROMACS		
Параметр	Значение	Комментарий
QMMM	yes	Необходим, но значение игнорируется
QMMM-grps	QM	
QMMMscheme	ONIOM	
QMmethod	RHF	
QMbasis	STO-3G	Необходим, но значение игнорируется
QMcharge	XX	Зависит от системы
QMmult	1	
Для MOPAC2012		
PM3 1SCF GRADIENTS CHARGE=XX singlet THREADS=YY		
Для ORCA		
! RHF PM3 NOFROZENCORE CONV HUECKEL		
%rel SOType 1 end		
%elprop Dipole false end		
%scf MaxIter 5000 end		
%output PrintLevel Nothing end		

Заряд и количество потоков MOPAC2012 были изменены в зависимости от тестовой системы.

3.4. КМ/ММ метадинамика в GROMACS/GFN2-хТВ

КМ/ММ динамика рассчитывалась с помощью пакета GROMACS/GFN2-хТВ. Шаг интегрирования при моделировании реакций составлял 0.5 фс. Параметр ассигасу, влияющий на величины барьеров, определяющих сходимость SCC, устанавливали равным 10.

Для преодоления барьера активации использовалась Well-tempered метадинамика с несколькими ходами, прикладываемая к двум коллективным переменным: расстояниям Ser195 OG — His57 NE2 и Ser195 OG — Ser214 O. Потенциал в 0.25 кДж/моль и шириной в 0.015 нм накладывался каждые 50 шагов, biasfactor 40. Дополнительные ограничения накладывались на значения расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 O, а также максимальное из двух: 0.45, 0.55 и 0.35 нм соответственно. Дополнительно было ограничено максимальное расстояние от обоих акцепторов до протона серина — 0.30 нм. Величина всех ограничивающих потенциалов установлена в 100 000 кДж/моль/нм. Профиль реконструировался параллельно 8 ходами для каждой системы, со временем 10 пс на каждую реплику.

4. Моделирование фермент-субстратного комплекса тромбина

Стартовая структура была подготовлена на основе структуры 3LU9_AB. Субстрат был укорочен до фрагмента Ace-PRSFL-

Nme. Остальная процедура аналогична таковой при подготовке для КМ/ММ моделирования. Для создания ингибирующей водородной связи рассчитывали траекторию длиной 1 нс с движущимся потенциалом величиной 100 000 кДж/моль/нм на расстоянии между протоном серина и остовным кислородом Ser214 от 0.35 до 0.2 нм. Реконструкция профиля свободной энергии производилась с помощью 8 независимых реплик (4 из конфигурации быстрой формы и 4 с ингибирующей водородной связью), с длиной траектории 100 нс на реплику. Двумерный профиль был восстановлен с помощью смешанных гауссовых моделей. Количество ядер выбиралось по минимуму значения информационного критерия Акаике.

5. Реконструкция профиля субстратной специфичности

Тестирование эффективности предсказания с использованием QuickVina 2 производилось абсолютно идентично условиям, описанным ранее [75]. Параметр exhaustiveness явно не задавался.

Для докинга пептидов в тромбин использовались структуры, подготовленные на предыдущем шаге, очищенные от лигандов всех типов. Входные файлы были сгенерированы с помощью утилит AutoDock Tools [109]. Для докинга использовали ячейку, покрывающую пептид Ace-PRSFL-Nme в исходной структуре. Докинг и рескоринг производились с помощью пакета инструментов PeptoGrid [75] и QuickVina 2 [111] в качестве програм-

мы докинга. В качестве лигандов использовалась библиотека пептидов с Ace и Nme модификациями на N- и C- концах. Параметр *exhaustiveness* не указывался, максимальное количество поз было установлено в 20.

Для отбора субстратов, имеющих конфигурацию пригодную к катализу, проводилась следующая процедура. Для каждого карбонильного атома углерода в каждом остатке каждой конфигурации субстрата оценивалось расстояние до атакующего кислорода серина. Это расстояние должно было быть не более 3.5 Å. Данное значение выбрано эмпирически на основе нашего опыта моделирования реакций с аналогичным механизмом [60,108]. Для карбонильного кислорода, соединенного с атакуемым углеродом оценивалось расстояние до оксианионного центра, а от предшествующего остоного азота — до остоного кислорода Ser214. Данная пара расстояний должна удовлетворять общепринятому эмпирическому критерию водородной связи — не более 3.5 Å между донором и акцептором. Дополнительные ограничения включали положение атакуемого карбонильного углерода – он должен принадлежать первому или второму остатку трипептида, чтобы сайт расщепления находился между аминокислотами, а не терминальными заглушками и аминокислотой. Наконец, последним условием было значение оценочной функции *PeptoGrid* не менее 0.5. Данный порог был выбран таким, потому как распределение всех значений имеет S образную форму с серединой около 0.5.

6. Визуализация

Для визуализации структур биомолекул во всех случаях применялся пакет PyMol [112]. Графики построены с помощью библиотеки Matplotlib [113] и сценариев на языке программирования Python. Визуализация данных иерархической кластеризации выполнялась с помощью ПО FigTree.

РЕЗУЛЬТАТЫ

1. Кластерный анализ структур тромбина

1.1. Кластеризация структур

В результате обработки массива из 400 записей человеческого тромбина, доступных в банке PDB (на 01.05.2019 г.), было получено 462 уникальных структур. Применение процедуры непараметрической кластеризации Affinity Propagation к этому набору структур приводит к выделению 20 кластеров (Таблица 2). При множестве преимуществ, которыми обладает алгоритм Affinity Propagation, нельзя не отметить, что он обладает и одним недостатком — из результатов кластеризации сложно оценить взаимоотношения полученных кластеров.

В частности, сложно оценить расположение центров друг относительно друга. Для визуализации данной картины мы дополнительно применили метод полной связи (дальнего соседа) иерархической кластеризации. На полученной дендрограмме (Рис. 5) можно увидеть как отчетливо выделяются группы структур протромбина, а также медленных E и E* форм. При этом большая группа структур, предположительно в E:Na⁺ форме, соединена относительно короткими ребрами. Мы рассмотрели возможность объединения данных структур в один кластер. Объединение структур вплоть до ветки 1BCU_NL существенно улучшало распределение силуэтов — метрики класте-

ризации, отражающей качество соответствия элемента кластеру (чем ближе к 1, тем лучше). Итоговое распределение силуэтов приведено на (Рис. 6).

Таблица 2 . Описание кластеров структур тромбина

PDB ID центра кластера	Разрешение (Å)	Размер кластера (шт)	Предполагаемая форма
5JDU_CD	1.7	1	Е
1NU9_A	2.2	2	Претромбин-2
5EDM_A	2.2	1	Протромбин
3NXP_A	2.2	4	Претромбин-1
4NZQ_A	2.81	2	Протромбин
6C2W_A	4.12	3	Протромбин
2AFQ_CD	1.93	4	Е
3BEI_AB	1.55	5	ЕЕ*
3K65_AB	1.85	7	Протромбин-2
4CH8_AB	1.75	46	?
3VXE_HL	1.25	6	?
1NT1_A	2.0	12	Е:Na ⁺
1A4W_HL	1.8	49	Е:Na ⁺
1BB0_AB	2.1	70	Е:Na ⁺
3EQ0_HL	1.53	55	Е:Na ⁺
5AFZ_HL	1.53	78	Е:Na ⁺
4YES_AB	1.5	39	Е:Na ⁺
5AFY_HL	1.12	45	Е:Na ⁺
2CN0_HL	1.3	20	Е:Na ⁺

Продолжение на следующей странице

Таблица 2 – Продолжение

PDB ID	Разрешение	Размер	Предполагаемая
центра	(Å)	кластера	форма
кластера		(шт)	
1BCU_HL	2.0	7	E:Na ⁺

Знаком ? отмечены структуры, которые, предположительно, могут также принадлежать к медленной E форме.

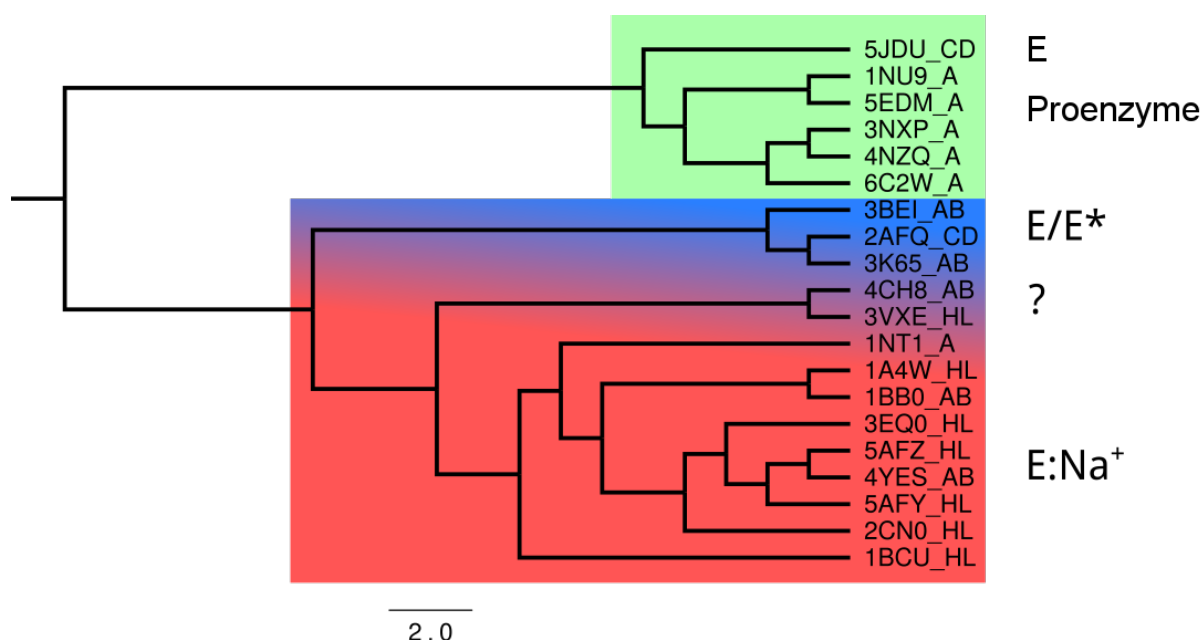


Рисунок 5. Иерархическая кластеризация центров кластеров, полученных методом непараметрической кластеризации Affinity Propagation. Структуры отмеченные зеленым являются протромбином. Отмеченные красным аннотированы как быстрая E:Na⁺ форма, синим — медленные E и E* формы. Знаком ? отмечены структуры, которые, предположительно, могут также принадлежать к медленной E форме.

Для оценки устойчивости результатов к небольшим изменениям в матрице расстояний, аналогичная процедура бы-

ла повторена для структур, препроцессированных при помощи программы Reduce из пакета MolProbity [114]. Данная утилита использует набор геометрических критериев для оптимизации положения радикалов аминокислот, содержащих амидную связь: глутамин и аспарагин, а также гистидина. В силу особенностей метода PCA, нередко невозможно достоверно определить их ориентацию, однако по косвенным данным, таким как наличие донор/акцепторов, заряженных групп и т.д. в окружении, можно сделать разумную догадку о корректной ориентации радикала. Таким образом, не внося существенных изменений в общую структуру, обработка при помощи программы Reduce внесет небольшой числовой шум в матрицу расстояний (медиана изменилась с -1.196 до -1.189). При этом результаты кластеризации были абсолютно идентичны полученным ранее.

Также мы использовали еще один вариант предобработки структур — проект PDB Redo использует стандартизированный протокол создания пространственных структур на основе данных электронной плотности, размещенных в исходном банке PDB [115]. К сожалению, далеко не все исследователи размещают в PDB исходные данные, поэтому в данном случае в выборке оказались только 301 уникальная структура (почти на треть меньше, чем в исходной). При этом медиана матрицы расстояний практически не изменилась (-1.191 против -1.196). Сохранились и ключевые особенности кластеризации: состав протромбиновых кластеров, состав кластеров E/E*, наличие погранич-

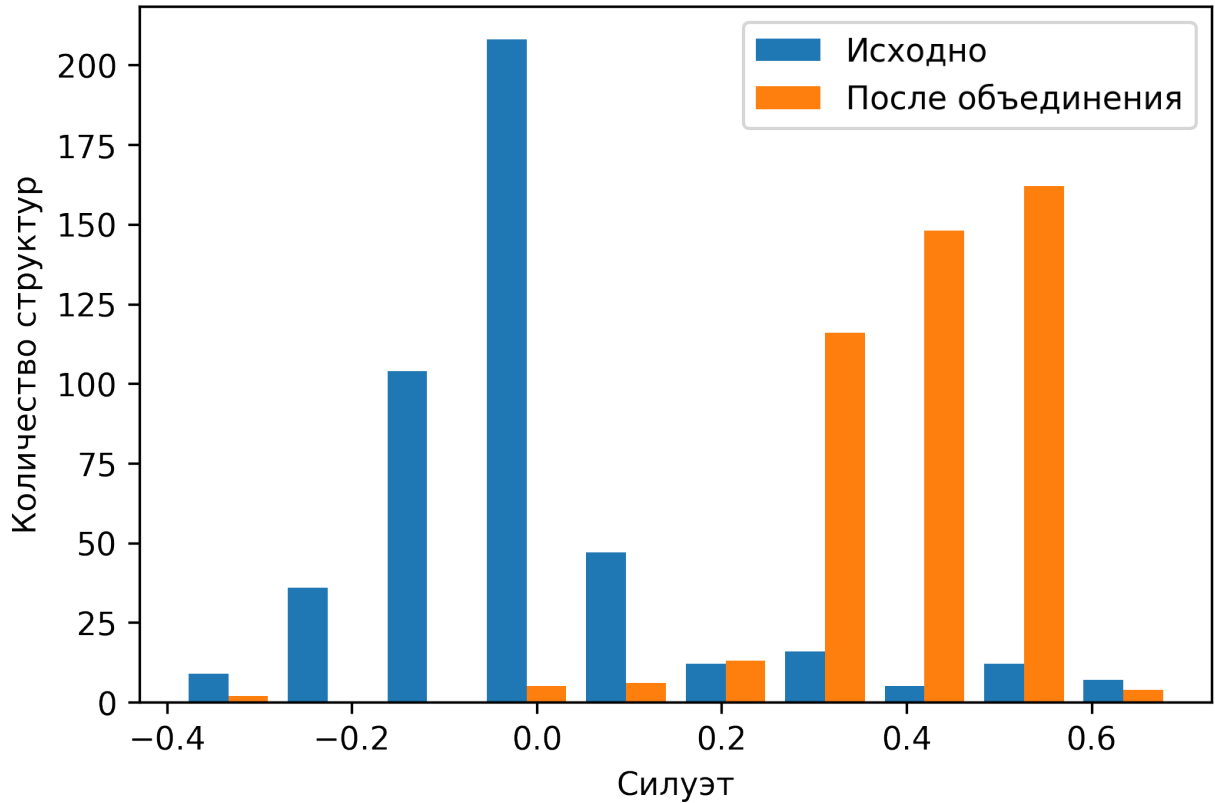


Рисунок 6. Объединение E:Na⁺ структур в один кластер улучшает метрику кластеризации.

ных кластеров 4CH8 и 3EQ0.

Таким образом, можно заключить, что результаты кластеризации устойчивы к небольшим изменениям в структурах, индуцированных вращением радикалов или же полной реоптимизацией структуры. При этом, полученным кластерам можно сопоставить биологически релевантные метки: тип фермента (процессированный или профермент) или же форму. Последнее обстоятельство позволяет, во-первых, заключить, что вся процедура имеет биологический смысл, а во-вторых, позволяет предположить, что анализ промежуточных кластеров может привести к получению новой информации о структуре медленной E

формы.

1.2. Анализ и аннотация кластеров

В рамках темы данной работы наибольший интерес представляют промежуточные кластеры 3VXE_HL и 4CH8_AB, расположенные между объединенным кластером структуры в активной форме E:Na⁺ и кластерами в E/E* форме. Для детального сравнения структур мы также выбрали референсную модель — это такая структура из всего набора, что сумма расстояний от нее до остальных является минимальной. Референсной моделью оказалась 5AFZ_HL, обладающая разрешением 1.53 Å, что является пусть и не лучшим показателем в выборке (Рис. 7), однако все равно достаточно высоко и входит в топ-50 структур с лучшим разрешением. Также данная структура представляет собой один из 20 исходных центров кластеров и принадлежит подмножеству активной E:Na⁺ формы (Рис. 5).

Можно предположить, что наиболее вероятным местом наличия небольших, но критичных для работы фермента, изменений будет окружение активного центра. Тем более, что мажорные изменения, характеризующиеся, например, разрушением геометрии Na⁺-связывающего центра, как следует из результатов кластеризации, выделяются в отдельные хорошо различимые кластеры на границе со структурами профермента.

На (Рис. 8) приведена геометрия активного центра быстрой E:Na⁺ формы из структуры 5AFZ_HL. Ключевым элементом

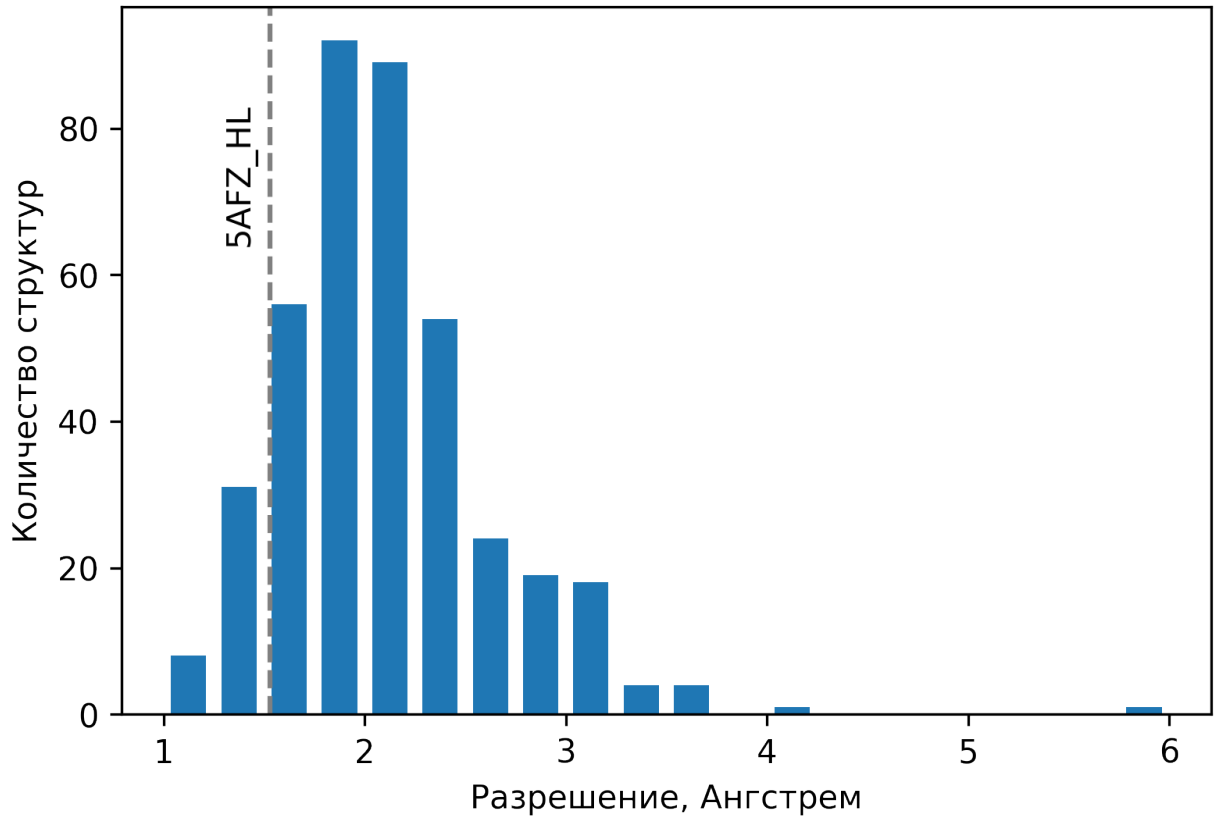


Рисунок 7. Гистограмма значений разрешения структур тромбина. Прерывистая серая линия на значении 1.53 демонстрирует положение выбранной референсной структуры 5AFZ_HL.

является каталитическая триада: Ser195, His57, Asp102 — остатки которой связаны цепочкой водородных связей.

При сравнении референсной структуры 5AFZ_HL, структуры в медленной E форме 3BEI_AB и структуры из промежуточного кластера 3VXE_HL (Рис. 9), в глаза бросается разница в положении основного кислорода остатка Ser214. Причем наиболее выраженное смещение в наблюдается структуре 3BEI_AB. При этом, в результате этого смещения расстояние между кислородом OG каталитического Ser195 до основного кислорода Ser214 уменьшается с 4.2 Å до 3.3 Å, что попадает в эмпирические гра-

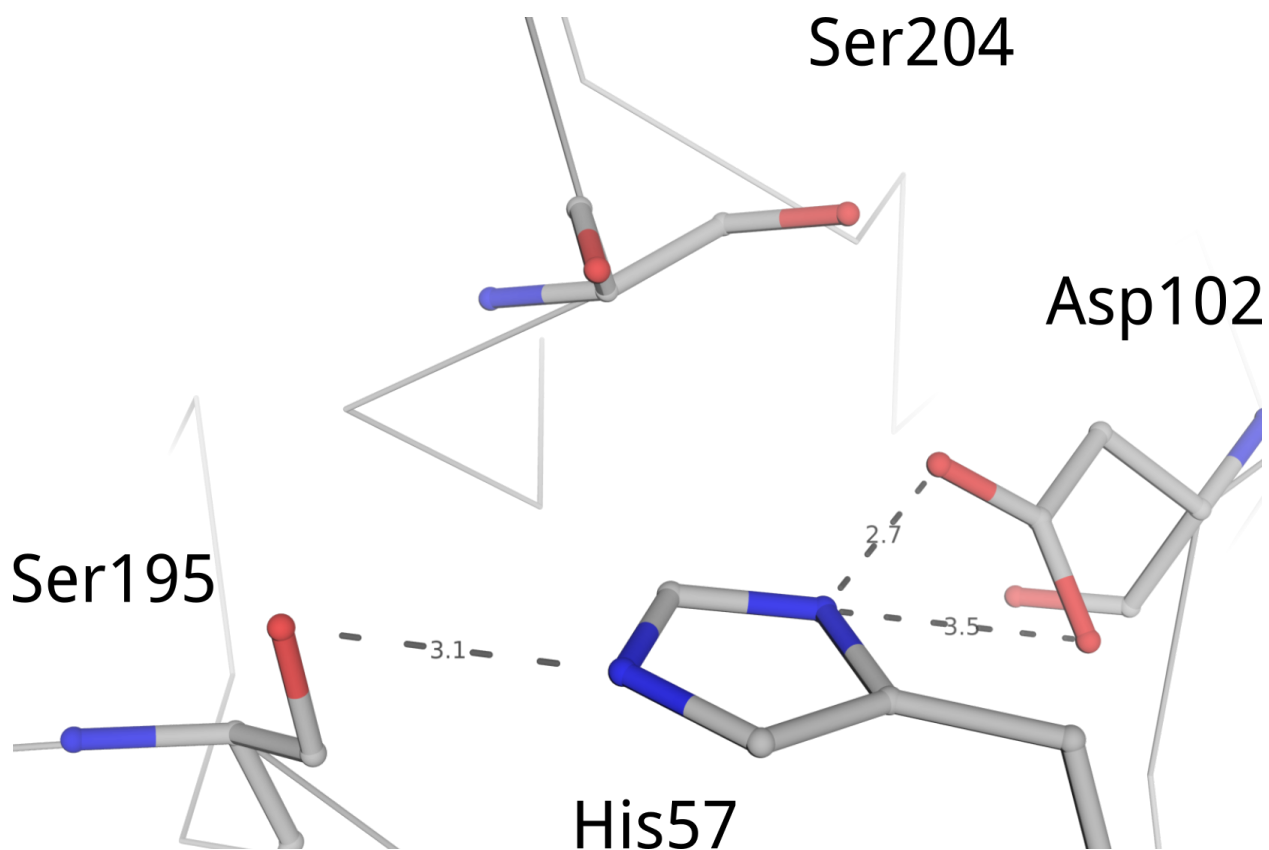


Рисунок 8. Каталитическая триада (Ser195, His57, Asp102) и соединяющие водородные связи в активном центре референсной структуры 5AFZ_HL.

ницы расстояния между донором и акцептором водородной связи, верхняя граница для которых общепринято считается на расстоянии до 3.5 Å [116]. При этом промежуточный кластер 3VXE_HL также занимает некоторое промежуточное состояние с точки зрения положения треугольника Ser195 OG, His57 NE2, Ser214 O.

Чтобы проверить, насколько это смещение — воспроизводимое явление, мы провели анализ расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 O во всех структурах тромбина, использованных для кластеризации. Как видно, существует значимое количество структур (29), в которых расстояние Ser195

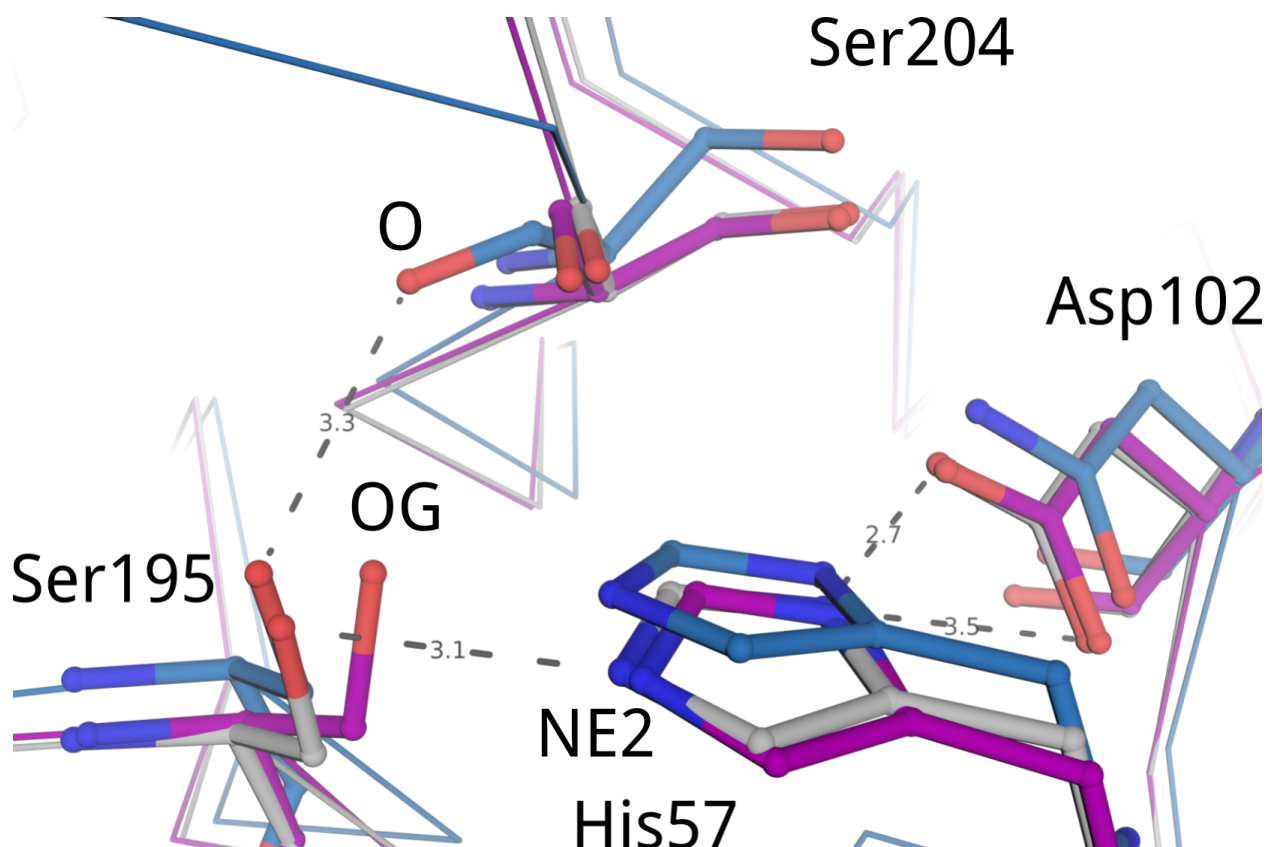


Рисунок 9. Наложение референсной структуры 5AFZ_HL (серый), структуры в медленной E форме 3BEI_AB (голубой) и структуры из промежуточного кластера 3VXE_HL (фиолетовый). Атомы азота, кислорода и серы окрашены в синий, красный и желтый соответственно.

OG — Ser214 O меньше эмпирической границы 3.5 Å (Рис. 10) . Более того, выделяется группа из трех структур, у которых при этом расстояние Ser195 OG — His57 NE2 больше данной границы, что может свидетельствовать об формировании новой водородной связи. Отметим, что данные структуры относятся к группе медленных. Таким образом, возможно, смещение Ser195 и Ser214 может быть следствием формирования новой водородной связи. При этом может разрушаться ключевая каталитическая водородная связь Ser195 OG — His57 NE2. Возможность ослабления или полного разрушения каталитической водород-

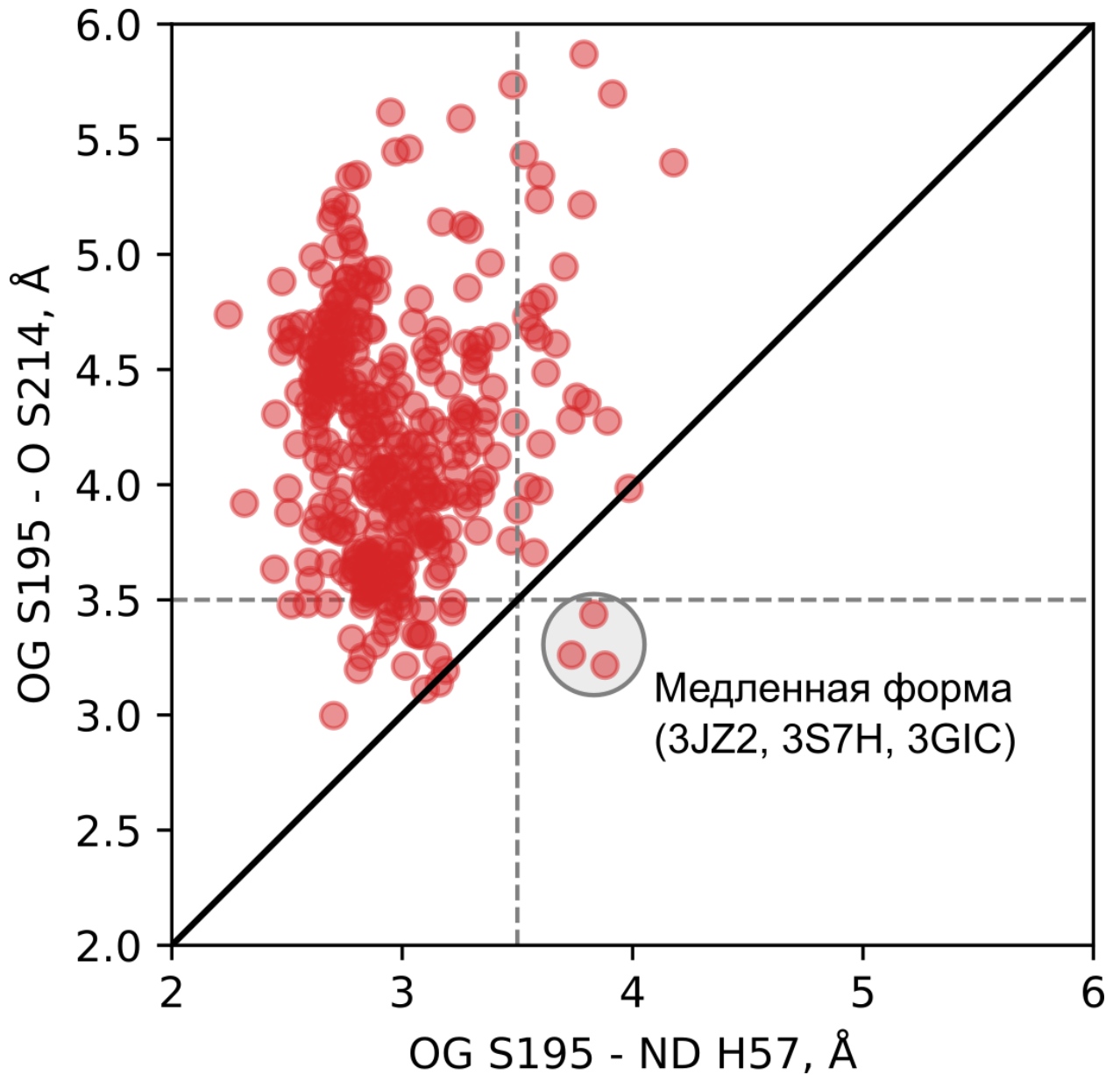


Рисунок 10. Распределение расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 в структурах тромбина.

ной связи ранее допускалась исследователями [18].

2. КМ/ММ моделирование GROMACS/DFTB3

Для проверки гипотезы о возможности формирования ранее неизвестной водородной связи в активном центре тромбина, мы решили прибегнуть к методам молекулярного моделирования. С помощью гибридного КМ/ММ моделирования

мы проанализировали поведение расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 в системах с ионом Na^+ в натрий-связывающем сайте и без него. В квантовую систему — часть системы, которая описывается с помощью квантовомеханических подходов, в противовес остальной, описываемой с помощью уравнений классической механики, были включены: активный центр, катион-связывающий карман, соединяющая их сеть молекул воды, а также фрагменты аминокислот, участвующих в их координации. Всего в квантовую систему вошло более 460 атомов. На (Рис. 11) приведены квантовая система, квантовая система в составе белка и полная система, использованные в моделировании.

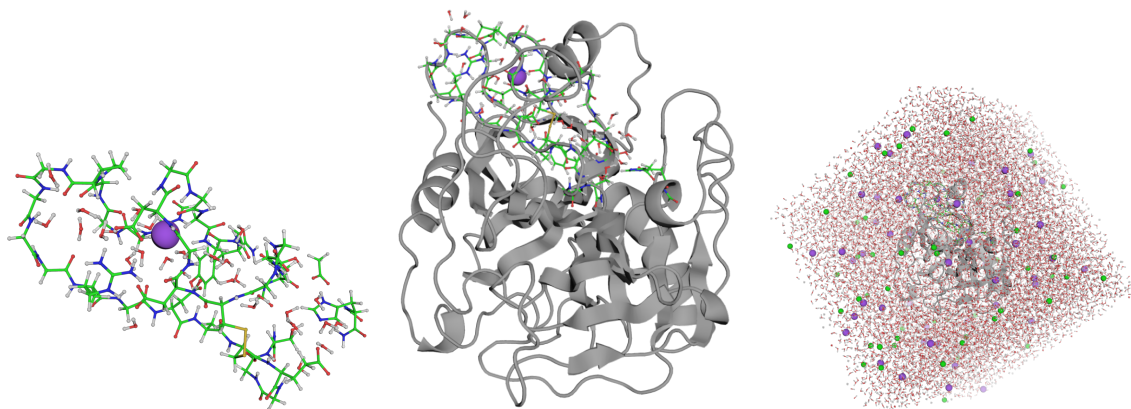


Рисунок 11. Квантовая система, квантовая система в составе белка и полная система, использованные в моделировании.

На (Рис. 12) можно видеть заметную разницу в поведе-

нии систем с ионом и без. Наличие иона натрия существенно ограничивает доступную область. При этом, надо отметить, что данная картина наблюдалась уже в ходе достаточно короткого моделирования (28 независимых реплик, 2.54 пс каждая) без использования методов повышения эффективности сканирования фазового пространства. При детальном анализе событий, произошедших в данной системе, мы обнаружили, что столь быстрые изменения в системе без иона могут быть следствием некорректно заданного протонирования остатков, окружающих квантовую систему. Причиной некорректного протонирования Asp240 (избыточный протон) стала некорректная интерпретация нами наблюдений, описанных в работе, авторы которой пытались оценить состояния протонирования на основе изменений в длинах связей при разных вариантах оптимизации структуры в электронной плотности [117]. При этом, система с ионом все равно сохраняла геометрию, характерную для кристаллических структур, а в системах без иона происходил поворот вокруг остова Ala190, который в результате принимал конфигурацию, характерную для протромбиновых структур.

3. КМ/ММ моделирование GROMACS/MOPAC2012

Потому как, возможно, такое сильное и, самое главное, быстрое накопление отличий в поведении могло быть следствием некорректно заданного протонирования окружающих остатков, мы решили валидировать данный результат с помощью по-

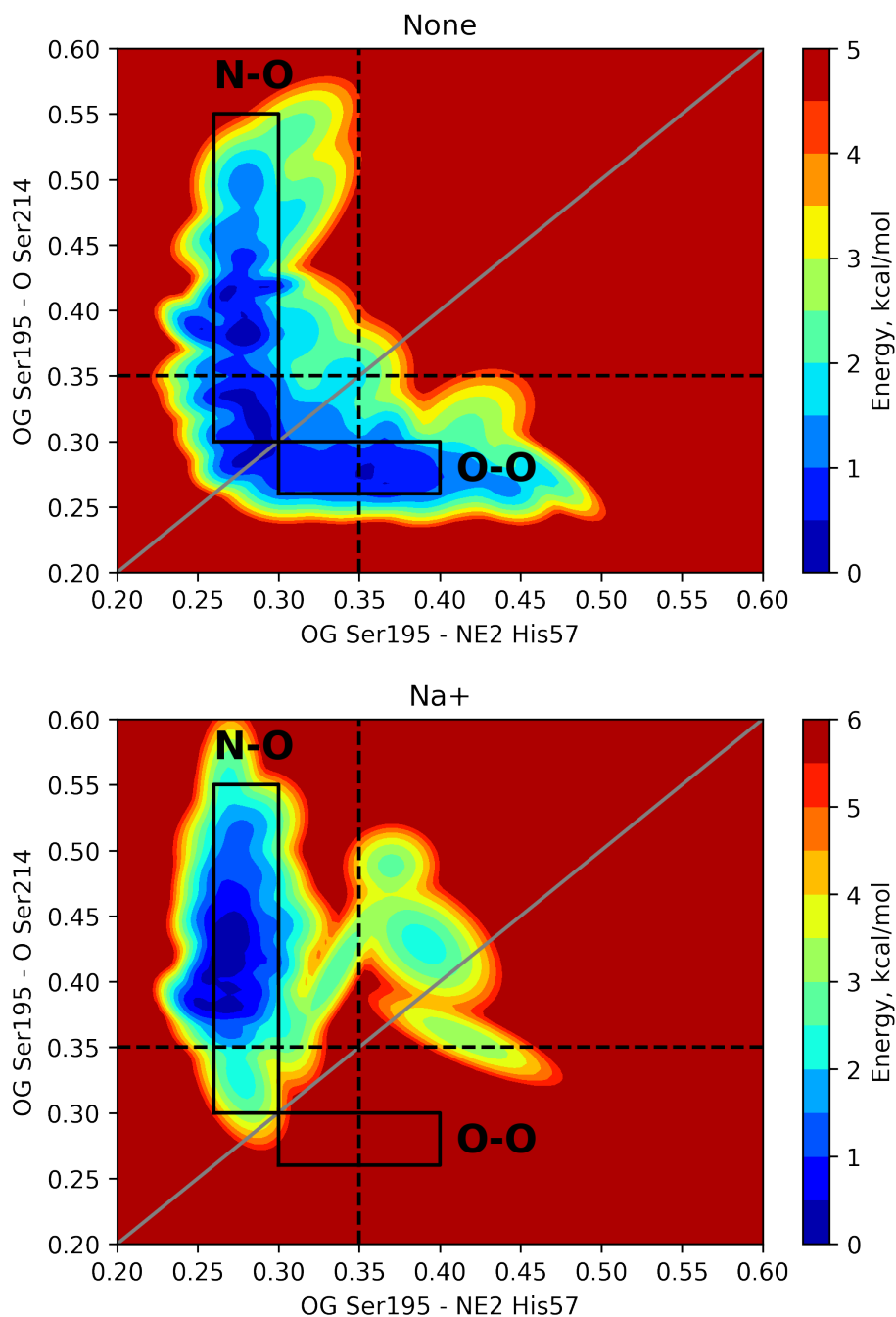


Рисунок 12. Ландшафт свободной энергии для расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 в зависимости от отсутствия (сверху) или наличия (снизу) иона Na^+ в натрий-связывающем сайте. Ион существенно ограничивает доступную область

вторного моделирования с другим паттерном протонирования, основанном на табличных значениях pK_a аминокислот и учете окружения, способного повлиять на формирование водородных

связей. Однако GROMACS/DFTB3, несмотря его неоспоримые достоинства, является чрезвычайно низкопроизводительным решением. Расчет 2.54 пс на 1 ядре с шагом 0.2 фс занял порядка 14 дней. При этом бутылочным горлышком, лимитирующим производительность, является квантово-механическая часть, которая не поддерживает параллелизацию.

Для обхода данного ограничения мы реализовали альтернативную версию КМ/ММ пакета — GROMACS/MOPAC2012. Используя в качестве эталона реализацию интерфейса GROMACS/ORCA, мы модифицировали соответствующие части GROMACS (модуль MD, модуль ввода-вывода и файлы конфигурации CMake) и MOPAC2012 (модули ввода-вывода и модули проверки параметров). В эталонной реализации обмен между пакетами GROMACS и ORCA выполняется через текстовые файлы, что ограничивает точность: данные печатаются и считываются с шаблоном “ % 10.7f ”, который предоставляет только 7 десятичных цифр, соответствующих одинарной точности, в то время как внутри себя ORCA использует двойную точность. В нашей реализации используется MOPAC2012, который был скомпилирован и собран в статическую библиотеку и напрямую связан со статическим двоичным файлом GROMACS во время компиляции, что позволяет осуществлять прямую передачу данных между MOPAC2012 и GROMACS с двойной точностью. Для оптимизации управления вычислениями QM была добавлена переменная среды `GMX_QM_MOPAC2012_KEYWORDS`,

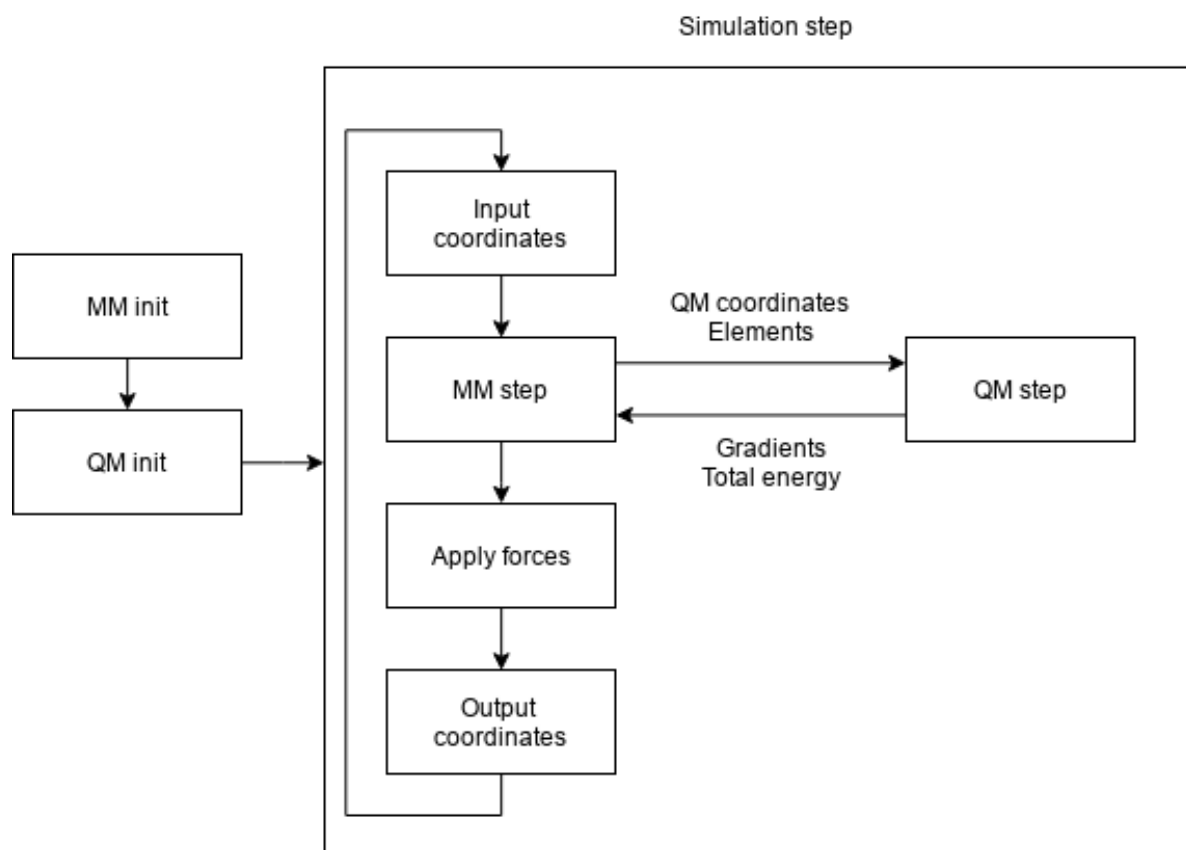


Рисунок 13. Схема универсального интерфейса КМ/ММ

которая содержит ключевые слова MORAC2012 и считывается на начальном этапе моделирования (Рис. 13).

3.1. Тестовые системы

Тест производительности инструментов КМ/ММ проводился на двух модельных системах: “малая” — искусственный фермент с параоксоновым субстратом [118] и “большая” — комплекс бутирилхолинэстеразы с экотиофатом. Размер подсистемы КМ составлял 64 и 408 атомов для систем “малая” и “большая”, соответственно.

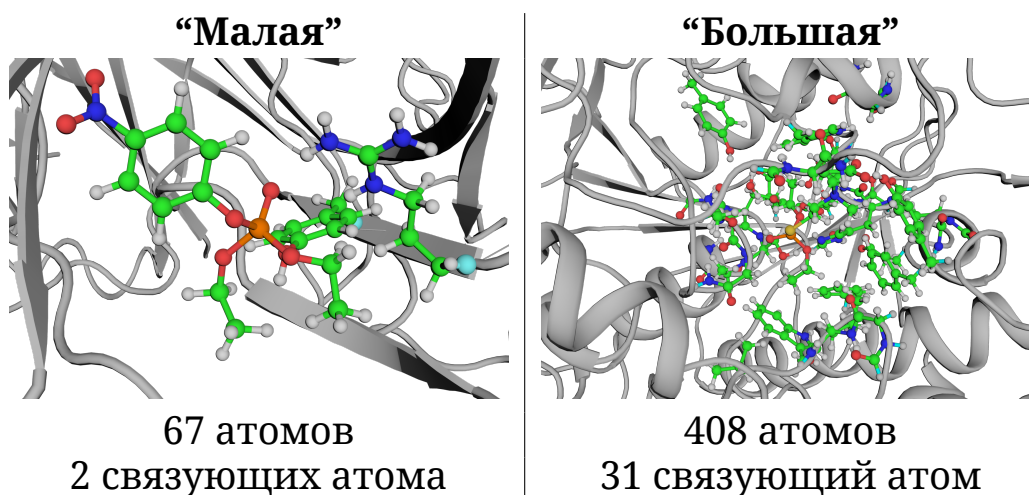


Рисунок 14. Состав тестовых систем.
Молекулярно-механическая часть представлена серым, квантово-механическая выделена цветом. Связывающие атомы окрашены в голубой цвет.

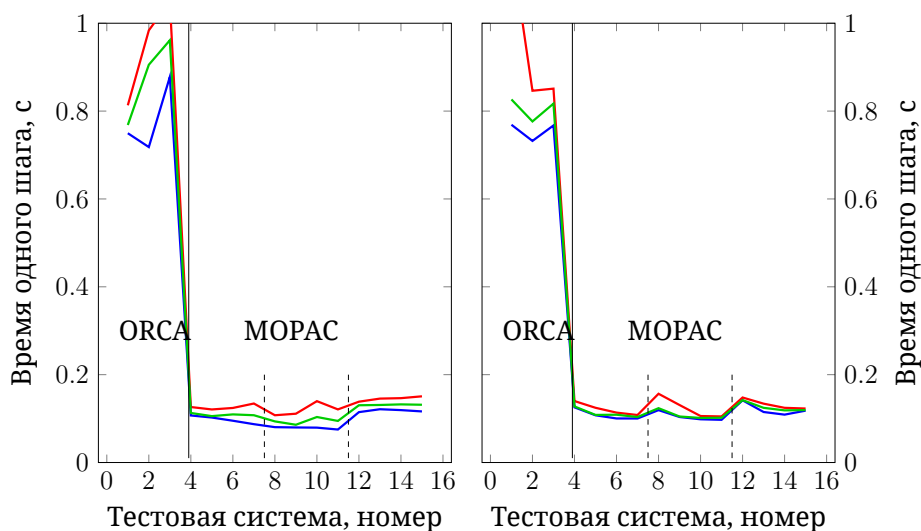


Рисунок 15. Производительность тестовых систем для “малой” системы на локальной рабочей станции (слева) и суперкомпьютере “Ломоносов-2” с файловой системой Lustre®. Максимальное значение на десяти репликах обозначено красной линией. Зеленый — средние значения времени для десяти реплик. Синий — минимальные значения времени для десяти реплик. Описание систем приведено в (Таблица 3).

Таблица 3 . Описание тестовых систем. Подсчет потоков относится как к подсистемам ММ, так и к QM. Точность описания описывает точность Gromacs. MM GPU относится к ускорению механической подсистемы в GROMACS с помощью GPU.

Номер системы	Описание
1	Orca single precision
2	Orca single precision and MM GPU
3	Orca double precision
4	Mopac single precision; 1 thread
5	Mopac single precision; 2 threads
6	Mopac single precision; 4 threads
7	Mopac single precision; 8 threads
8	Mopac single precision; 1 thread and MM GPU
9	Mopac single precision; 2 threads and MM GPU
10	Mopac single precision; 4 threads and MM GPU
11	Mopac single precision; 8 threads and MM GPU
12	Mopac double precision; 2 thread
13	Mopac double precision; 2 threads
14	Mopac double precision; 4 threads
15	Mopac double precision; 8 threads

3.2. Вычислительное время

Гибридный расчет КМ/ММ “малой” системы с библиотекой МОРАС показал восьмикратное ускорение по сравнению с вызовом исполняемого файла ORCA (Рис. 16), расшифровка тестовых систем приведена в таблице (Таблица 3). Разница в производительности была связана главным образом с необходимостью обмена данными между GROMACS и ORCA через файловую систему; энергетические градиенты квантовой подсистемы рассчитывались с одинаковой скоростью с помощью инструментов МОРАС и ORCA. Использование версии GROMACS с двойной точностью замедлило скорость вычислений на локальной рабочей станции примерно в 1,5 раза, но разница была полностью нивелирована, когда вычисления выполнялись на суперкомпьютере. Удивительно, но максимальная производительность была достигнута на локальном компьютере с быстрой файловой системой (NVMe 1.2). Использование потоков для подсистем QM и MM дало прирост производительности на 10%, наиболее существенный эффект наблюдался для медленных шагов SCF (Self-Consistent-Field) на суперкомпьютере “Ломоносов-2”.

По сравнению с “малой” системой, мы наблюдали нелинейное увеличение вычислительного времени. В то время как размер системы увеличился в 6,5 раз, время вычислений увеличилось в 30 раз для версий ORCA и МОРАС2012. Тем не менее, соотношение внутренних показателей между ORCA и МОРАС2012 осталось прежним. С полученной производитель-

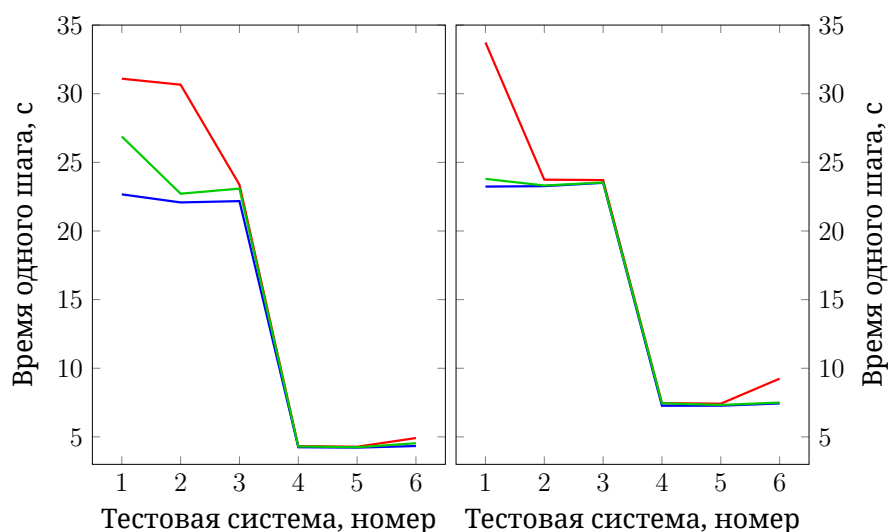


Рисунок 16. Производительность тестовых систем для “большой” системы на локальной рабочей станции (слева) и суперкомпьютере “Ломоносов-2” с файловой системой Lustre®. Максимальное значение на десяти репликах обозначено красной линией. Зеленый — средние значения времени для десяти реплик. Синий — минимальные значения времени для десяти реплик. Описание систем приведено в (Таблица 4).

Таблица 4 . Описание тестовых систем. Подсчет потоков относится как к подсистемам ММ, так и к QМ. Точность описывает точность GROMACS. ММ GPU относится к ускорению механической подсистемы в GROMACS с помощью GPU

Номер системы	Описание
1	Orca single precision
2	Orca single precision and MM GPU
3	Orca double precision
4	Mopac single precision; 1 thread
5	Mopac single precision; 1 thread and MM GPU
6	Mopac double precision; 1 thread

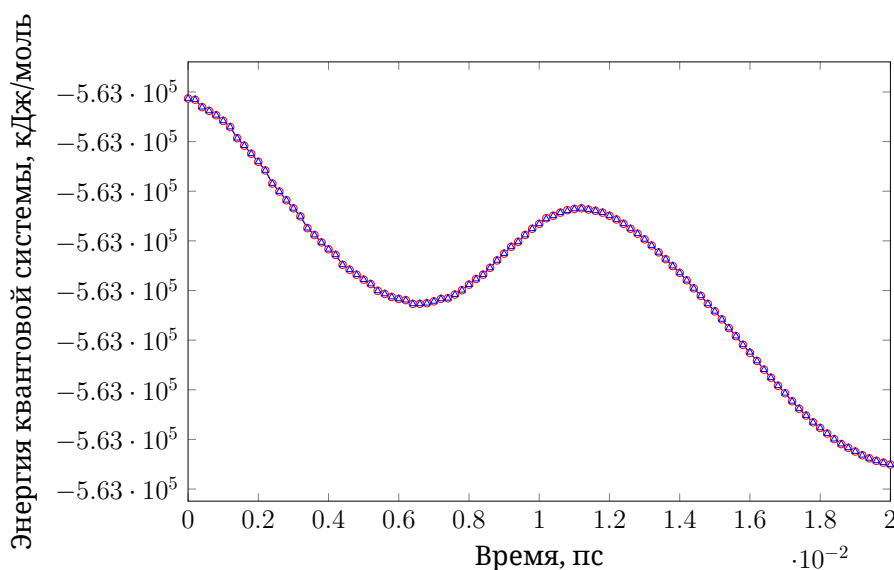


Рисунок 17. Сравнение значений энергии для квантовой подсистемы GROMACS/MOPAC2012: красные круги и GROMACS/ORCA: синие треугольники.

ностью, GROMACS/MOPAC2012 приблизительно на 1-1.5 порядка быстрее, чем GROMACS/DFTB при идентичном наборе ресурсов.

3.3. Воспроизводимость

Поскольку реализации вычислительных протоколов в разных пакетах могут различаться, очень важно проверить воспроизводимость результатов в разных программах. Чтобы убедиться, что наш интерфейс дает правильные результаты, мы сравнили значения энергии между соответствующими запусками с MOPAC2012 и ORCA. Типичная ошибка составляет менее 0,1% в масштабе 100 шагов (Рис. 17) и может быть объяснена разницей в точности передачи данных (одинарная в ORCA против двойной для MOPAC2012).

3.4. Сольватация ионов в GROMACS/МОРАС2012

При подготовке систем с тромбином для моделирования с помощью новой реализации GROMACS/МОРАС2012 мы обратили внимание, что ион металла в натрий-связывающем кармане практически с первых шагов теряет свою координационную сферу. Тестирование на модельной системе — катион натрия в кубике воды — продемонстрировало удивительную проблему: все семейство полуэмпирических методов PM3, PM6-X (PM6, PM6-D3, PM6-D3H4), PM7 не способно воспроизвести координацию молекул воды ионом металла. Типичный результат оптимизации геометрии такой системы приведен на (Рис. 18). При этом, даже без использования поправок DX для водородной связи, выстраивается хорошо заметная сеть водородных связей между молекулами воды. По всей видимости, существенная ошибка в энергии сольватации ионов металлов является характерной чертой семейства методов с пренебрежением двухатомным дифференциальным перекрытием (NDDO), к которым принадлежат все методы семейства PM* [119].

Таким образом, несмотря на высокую производительность GROMACS/МОРАС2012 (приблизительно 1-1.5 порядка на системе 460 атомов, по сравнению с GROMACS/DFTB), для задачи, связанной с учетом сольватации иона металла в составе белка, данный инструмент непригоден.

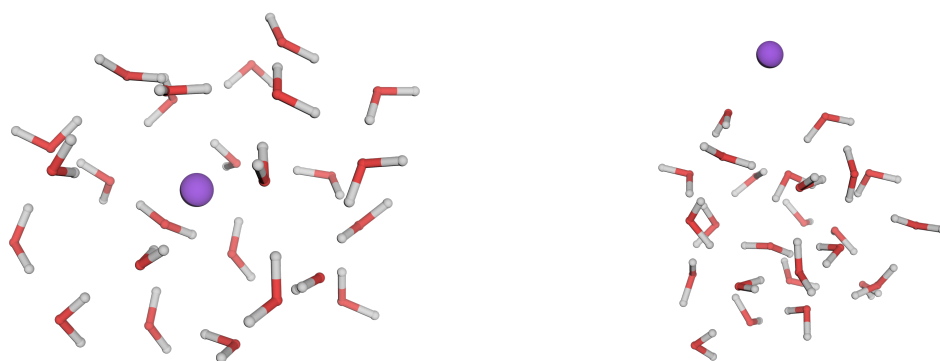


Рисунок 18. Стартовая и оптимизированная геометрии системы вода-катион. Использован полуэмпирический метод PM6-D3.

4. QM/MM моделирование GROMACS/GFN2-xTB

Так как GROMACS/MOPAC2012 оказался непригоден для задачи описания иона в составе натрий-связывающего сайта в тромбине, нам пришлось разработать еще один инструмент для гибридного КМ/ММ моделирования. На этот раз мы использовали метод GFN-xTB2, входящий в новое семейство полуэмпирических DFTB методов общего назначения, оптимизированных для корректного воспроизведения геометрий, колебательных свойств, а также энергий нековалентных взаимодействий [63, 64].

При оптимизации геометрии системы, состоящей из кубика воды и катиона натрия, методом GFN-xTB2, наблюдалось координационное число 6.0 (Рис. 19), характерное для этого иона [120]. Однако расстояния в статичной системе составляют 2.2-2.4Å, при характерном диапазоне 2-3Å с пиком в области 2.45-2.55 [120].

До недавнего времени GFN2-xTB распространяется в виде

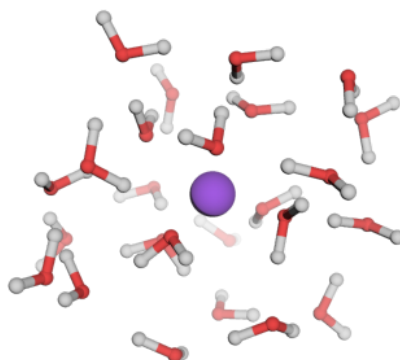


Рисунок 19. Координационная сфера Na^+ в кубе воды, оптимизированная в GFN-xTB2

бинарного исполняемого файла (исходные коды и программный интерфейс API были открыты 02.10.2019 г.), поэтому просто перенести на него использованный ранее с MORAC2012 подход невозможно. Однако, используя ту же идею и архитектуру, мы реализовали интерфейс общего назначения с подключаемыми плагинами для различных квантово-механических пакетов. Данный интерфейс устроен аналогично (Рис. 13). На языке программирования Cython — диалекте языка программирования Python, конвертируемом в код на языке C, пишется сценарий-плагин, отвечающий за а) запись входного файла для квантово-механического пакета; б) вызов исполняемого файла с необходимыми параметрами; в) чтение выходного файла. Для удобства управления и тонкой настройки квантово-механического пакета, можно опционально добавлять переменные окружения, которые можно изменять “на лету”. Так как

плагин компилируется в разделяемую библиотеку, то, благодаря механизмам управления поиска путей к библиотекам (с помощью переменной окружения `LD_LIBRARY_PATH`), можно быстро переключать используемый инструмент без пересборки всего пакета GROMACS. Это чрезвычайно удобная возможность, когда необходимо валидировать полученный результат с помощью квантово-химических методов различного уровня точности. Чтобы минимизировать задержки, связанные с использованием сетевых файлов систем, особенно на кластерах и суперкомпьютерах (как мы продемонстрировали на примере GROMACS/ORCA/MOPAC2012, такие задержки могут замедлять расчет практически на порядок), обмен данными происходит через временную локальную файловую систему, размещенную в оперативной памяти (TMPFS).

Также такая реализация интерфейса не накладывает никаких ограничений на исходную кодовую базу, поэтому можно использовать дополнительные плагины для пакета GROMACS. В частности, набор инструментов PLUMED2 для проведения расчетов с использованием методов повышения эффективности сканирования фазового пространства [121, 122].

В данном случае мы использовали модификацию метода метадинамики (well-tempered metadynamics). На основе статистики распределения расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 в PDB, а также предварительных данных моделирования в GROMACS/DFTB3 (Рис. 12), мы дополнительно ограничи-

ли фазовое пространство при помощи гармонических потенциалов.

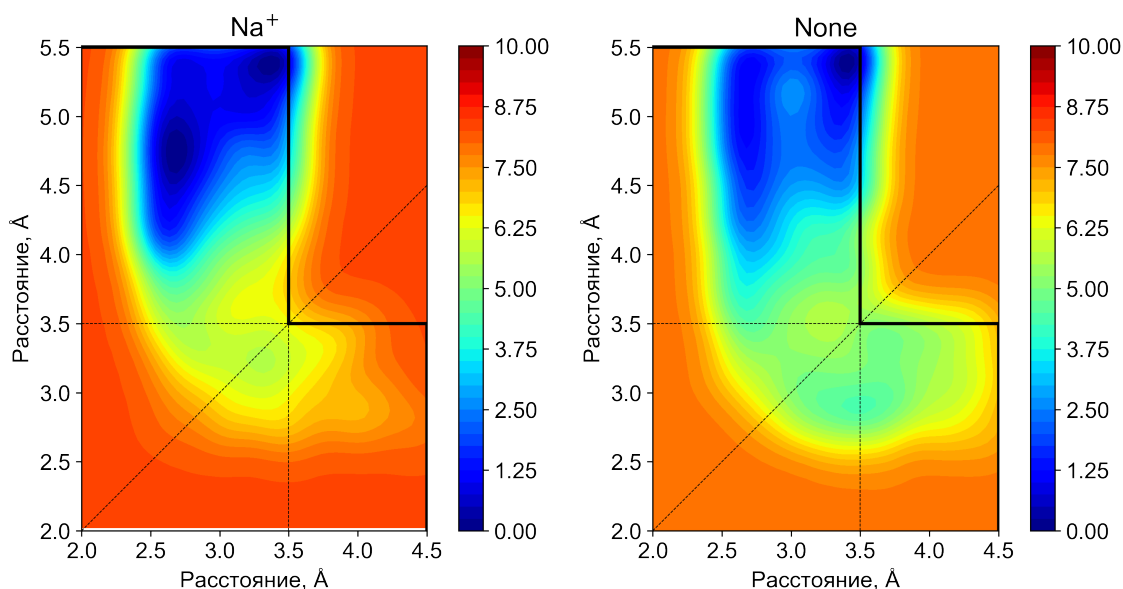


Рисунок 20. Ландшафт свободной энергии для расстояний Ser195 OG — His57 NE2 и Ser195 OG — Ser214 O в зависимости от наличия (слева) или отсутствия (справа) иона Na^+ в натрий-связывающем сайте. Наличие катиона существенно ограничивает доступную область конформационного пространства.

Как видно на (Рис. 20), системы с катионом и без продолжают демонстрировать различное поведение. В случае системы без иона, присутствует несколько локальных минимумов ((3.4, 3.7), (3.5, 2.8)) в области, находящейся за пределами границ водородной связи O-N (Ser195 OG — His57 NE2). Причем глубина последнего на (3.5, 2.8) составляет около ~ 0.8 ккал/моль. Непосредственное наличие в данном минимуме водородной связи с

Ser214 O подтверждается визуальным анализом (Рис. 22).

Важной характеристикой моделирования с использованием метадинамики является сходимость полученных результатов. В данном случае мы отслеживали величину барьера (Рис. 21) на пути между двумя локальными минимумами, обнаруженными в системе без иона: (2.7, 5.3) (характерный для исходной водородной связи) и (3.5, 2.8) (связь с Ser214 O). Барьер вычислялся как разница максимального и минимального значений на данном пути. Помимо процедуры реконструкции поверхности свободной энергии на основе суммирования наложенных в ходе метадинамики потенциалов, мы также применили реконструкцию с использованием взвешенного гистограммного анализа.

Анализ водородных связей внутри квантовой системы между молекулами воды и белковой частью системы также демонстрирует, что общее пространство водородных связей сужено: в системах без катиона, в среднем, 26 водородных связей, против 23 в системе с катионом (Рис. 23). Амплитуды изменений в системе без иона тоже больше. Все вместе это можно интерпретировать следующим образом: в системе с катионом поддерживается стабильная структура водородных связей, которые гораздо реже обмениваются.

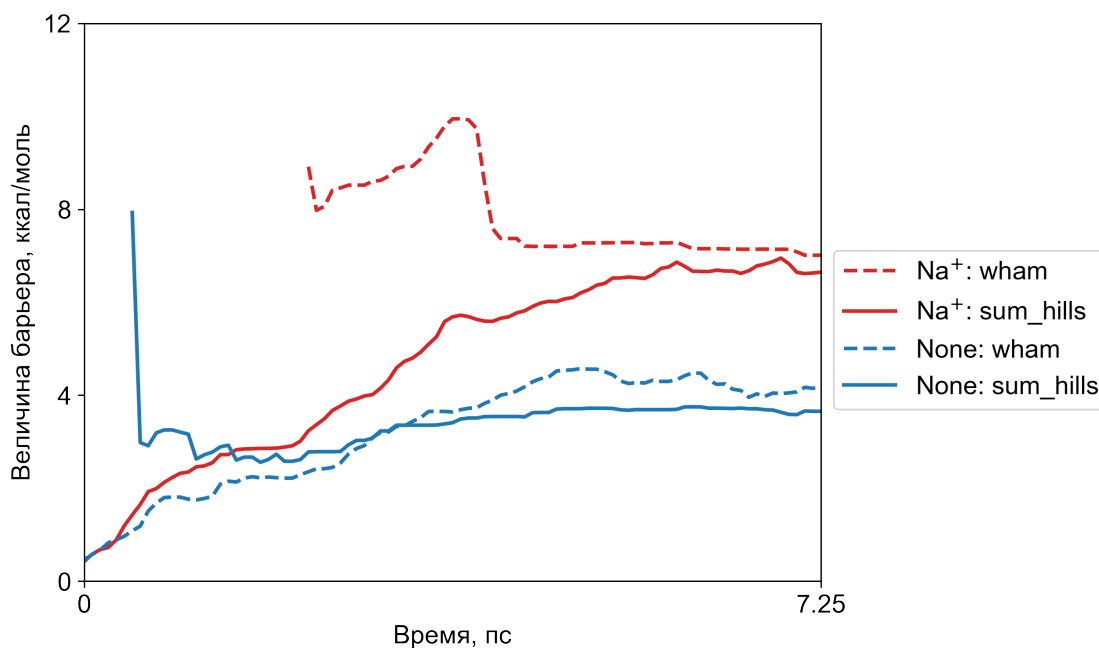


Рисунок 21. Величина энергетического барьера на пути в локальный минимум, характерный для системы без катиона. Система с ионом Na^+ в натрий-связывающем сайте изображена красным, без катиона — синим. Прерывистая линия отображает анализ на основе реконструкции поверхности свободной энергии с использованием взвешенного гистограммного метода.

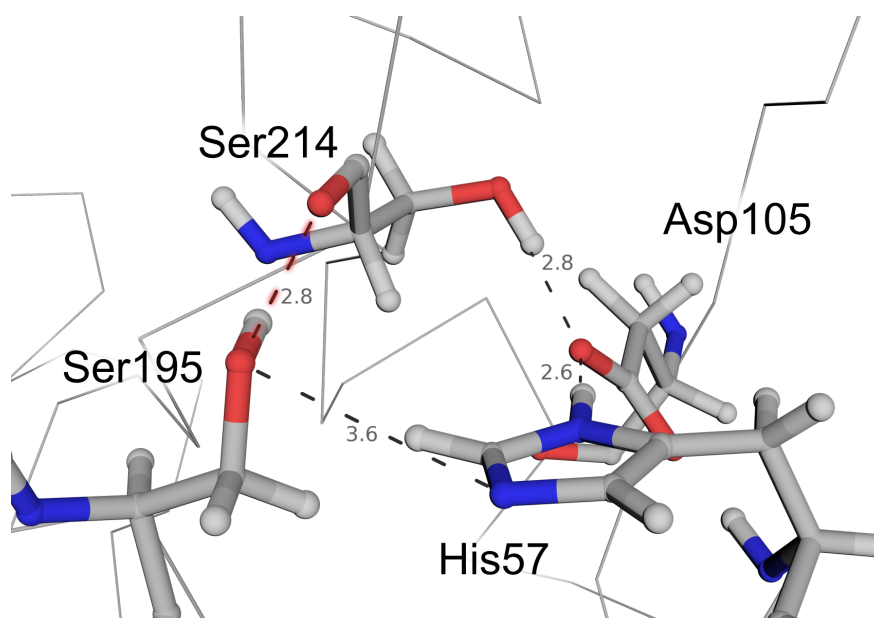


Рисунок 22. Водородная связь Ser195 OG — Ser214 O в системе без катиона Na^+ .

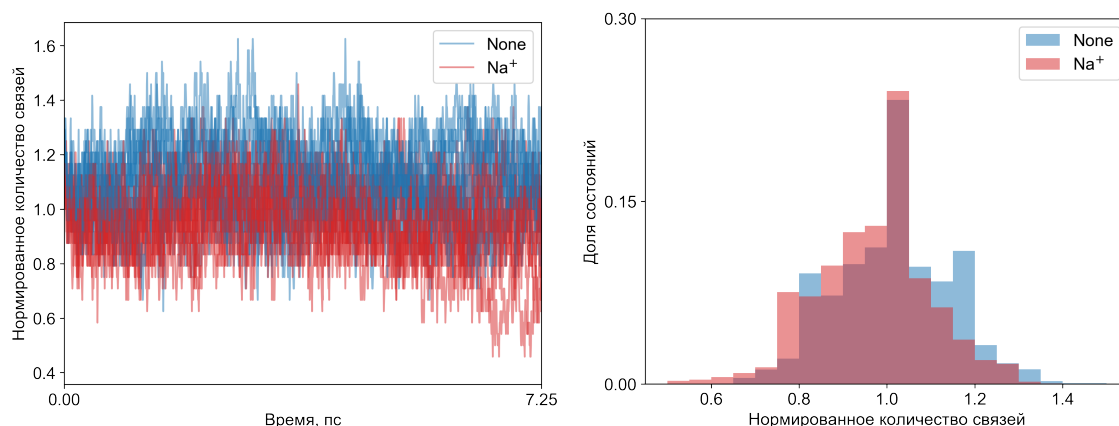


Рисунок 23. Количество водородных связей в система с Na^+ (красный) и без (синий). Слева приведено изменение количества связей во времени, справа в виде гистограммы.

5. Моделирование взаимодействия с субстратом

Учитывая наличие альтернативных конфигураций активного центра, нам было интересно узнать, могут ли они оказывать эффект на взаимодействие с субстратом. Для этой задачи мы провели классическое молекулярно-механическое моделирование комплекса тромбина с фрагментом нативного пептидного субстрата Ace-PRSFL-Nme. Нами были созданы две стартовых конфигурации: исходная, соответствующая $\text{E}:\text{Na}^+$ форме из кристалла, и альтернативная — со связью Ser195 OG – Ser214 O. Последняя была получена путем применения к исходной гармонического потенциала, ограничивающего расстояние между протоном каталитического серина и остовным кислородом Ser214 на расстоянии 1.8 Å.

В результате моделирования классической (без применения методов повышения эффективности сканирования) молекулярной динамики, мы получили двумерный ландшафт сво-

бодной энергии в уже многократно использованных ранее координатах: Ser195 OG – His57 NE2, Ser195 OG – Ser214 O (Рис. 24). Данный ландшафт качественно воспроизводит картину КМ/ММ моделирования: минимум, соответствующий каталитически активной $E:Na^+$ форме существенно глубже, нежели E форме, образованной за счет появления ингибирующей водородной связи с остовом Ser214. Отдельно можно отметить появление еще одного состояния – вытесненного (или Expelled).

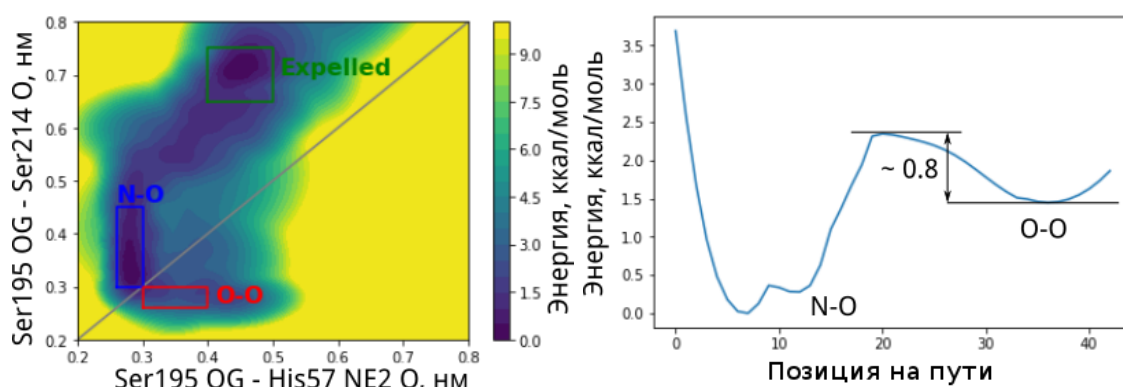


Рисунок 24. Поверхность свободной энергии комплекса тромбина с пептидным субстратом. Выделенные области соответствуют: N-O — каталитическая конфигурация, O-O — ингибирующая водородная связь, Expelled — вытесненная конфигурация (слева). Одномерный профиль энергии вдоль пути между минимумами зон O-N и O-O. (справа)

Более детальный анализ фрагмента ландшафта, соединяющего области N-O и O-O (Рис. 24), демонстрирует, что если фермент-субстратный комплекс образовался с конформацией фермента, имеющей водородную связь с остовом Ser214, то для

выхода из этого состояния необходимо преодолеть барьер ~ 0.8 ккал/моль. А это, очевидным образом, будет вносить вклад в общую кинетику процесса и соответствует уменьшению наблюдаемой скорости приблизительно на порядок.

6. Предсказание субстратной специфичности

6.1. Разработка и оптимизация алгоритма PeptoGrid

Одно из экспериментальных наблюдений, касающихся медленной формы, состоит в том, что она имеет измененную субстратную специфичность. При этом, по всей видимости, происходит не смена профиля к другому набору субстратов, а скорее падение специфичности к фибриногену. В пользу этой гипотезы говорит тот факт, что взаимодействие медленной формы тромбина с протеином С (лиганд, характерный для медленной формы) улучшается в 1 700 раз за счет образования тройного комплекса с тромбомодулином [123].

Взаимодействие белка с пептидными лигандами может быть изучено с помощью моделирования молекулярного докинга. Одной из лучших оценочных функций, специально сконструированных для пептидных библиотек, на сегодняшний день является разработанный нами алгоритм PeptoGrid [75]. Он использует набор результатов докинга, полученных с помощью любого из популярных пакетов: AutoDock Vina [107], PLANTS [124] или LeDock [125]. Процедура основана на анализе частот встречаемости атомов определенного типа в различных обла-

стях пространства и последующем ранжировании структур в соответствии с этими пространственными частотными картами. Структуры, лучше соответствующие такому консенсусному облаку контактов, получают более высокие значения оценочной функции (Рис. 25).

Процедура была валидирована на библиотеке трипептидов, для которых известны экспериментальные константы связывания с ангиотензинпревращающим ферментом (АПФ). А ее эффективность подтверждена результатами виртуального скрининга биогенных пептидов, способных взаимодействовать с внеклеточный доменом ГАМКБ рецептора — фармакологической мишени для лечения зависимостей, депрессии и тревожных расстройств [126]. Отобранные в результате вычислительного предсказания пептиды продемонстрировали выраженное влияние на поведение рыбок *Danio rerio in vivo* [75].

Также аналогичная процедура может использоваться и для олигомеров нуклеиновых кислот. Например, с помощью докинга комбинаторной библиотеки тринуклеотидов в структуру тромбина и используя для построения не индивидуальные типы атомов, а все атомы оснований можно выделить области наиболее предпочтительного расположения азотистых оснований на его поверхности (Рис. 26). Причем оказывается, что в структурах тромбина с ДНК-аптамером 15ТВА и его производными, в первом локальном максимуме (слева) действительно располагается один из тиминов ТТ-петли. А во второй области мо-

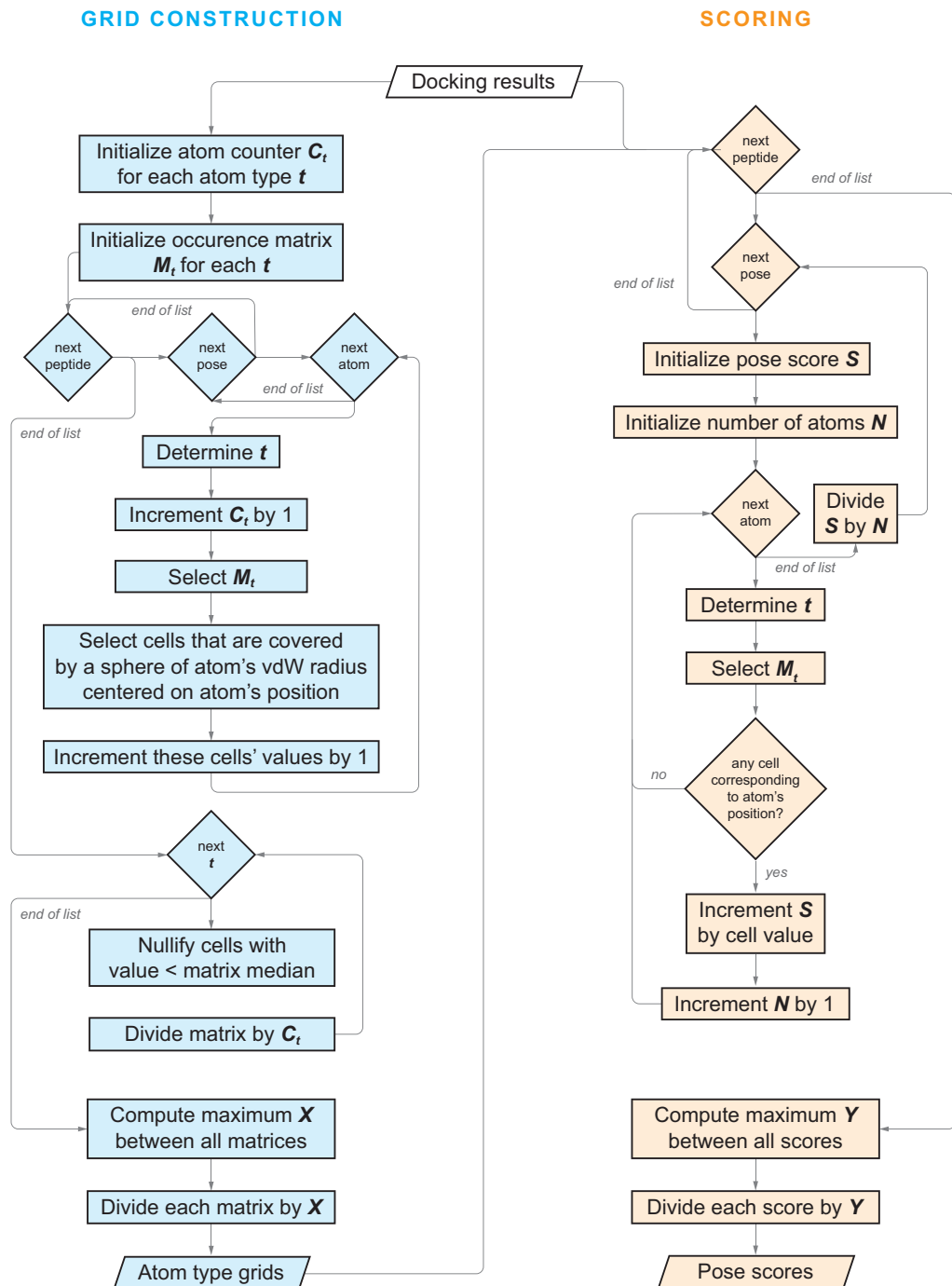


Рисунок 25. Блок-схема алгоритма оценочной функции PeptoGrid

жет располагаться ароматическое кольцо из химической модификации по 12 нуклеотиду — 5-(3-(2-(1H-индол-3-ил)ацетамид-N-ил)-1-пропен-1-ил)-2'-дезоксиуридин. Это означает, что информация, получаемая при помощи подхода ReptoGrid, может использоваться не только для дизайна новых молекул, но также для подбора мест введения химических модификаций.

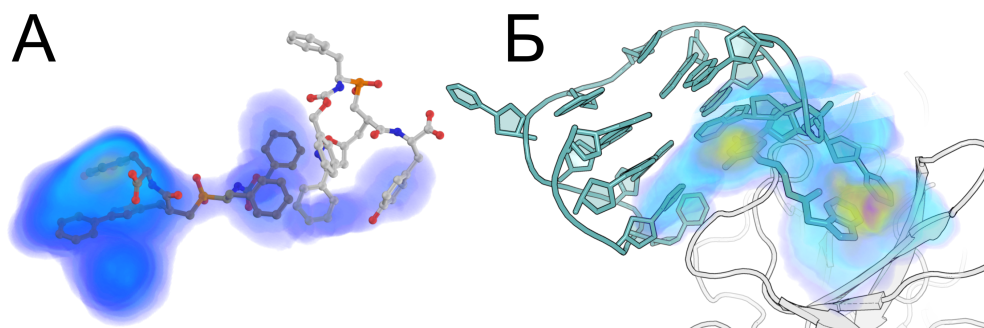


Рисунок 26. Совмещение карты пространственной плотности и структур лигандов. (А) Карта для ароматических углеродов и структура пептидомиметика в активном центре ангиотензинпревращающего фермента (PDB ID 2XY9). (Б) Карта для тимина и структура химически модифицированного ДНК-аптамера к тромбину (PDB ID 6EO6). Увеличение частоты встречаемости атомов происходит от синего, через желтый, к пурпурному.

Таким образом, набор инструментов ReptoGrid может оказаться полезным для изучения взаимодействия тромбина с пептидными субстратами. Однако для предсказания субстратной специфичности необходимо иметь возможность многократно производить массовый параллельный докинг пептидных биб-

лиотек. Для решения данной задачи мы оптимизировали разработанный ранее подход по следующим направлениям:

- Оптимизация структуры хранения данных;
- Оптимизация для высокопараллельных окружений;
- Увеличение скорости моделирования докинга;

Нативные пакеты для моделирования молекулярного докинга, такие как AutoDock Vina, PLANTS и LeDock в процессе работы генерируют большое количество промежуточных и итоговых файлов. В случае AutoDock Vina, для полной комбинаторной библиотеки трипептидов из 8 000, общее количество поз будет более 160 000 (при 20 индивидуальных позах для каждого пептида). При работе в суперкомпьютерных окружениях, использующих распределенные сетевые файловые системы, такое большое количество файлов, расположенных, к тому же, в небольшом количестве директорий способно существенно понизить производительность [66]. Для обхода данной проблемы мы реализовали хранение полученных координат в виде единой матрицы формата HDF5. Помимо уменьшения количества файлов, HDF5 также позволяет работать в нескольких режимах: один писатель — множество читателей и множество писателей — множество читателей. Таким образом, при использовании технологии параллельного программирования MPI мы получаем возможность запускать до нескольких сотен параллельных процессов, независимо осуществляющих молекулярный докинг, и запись результатов в один общий файл, без необ-

ходимости дополнительного постпроцессинга по объединению результатов.

Скорость молекулярного докинга лимитирована выбранным пакетом. При этом лучший результат по качеству предсказаний ранее продемонстрировала AutoDock Vina [75], которая является самой медленной из тройки AutoDock Vina, PLANTS, LeDock.

К счастью, некоторое время назад была разработана пакет Quick Vina 2, имплементирующий логику и скоринг функцию AutoDock Vina, но с несколькими эвристиками, значительно ускоряющими моделирование (до 20.5 раз) [111]. При этом, так как Quick Vina 2 имеет полностью совместимый синтаксис с Autodock Vina, то ее внедрение в PeptoGrid элементарно. Из-за особенностей модификаций Quick Vina 2, зависимость качества результатов от параметра exhaustiveness, влияющего на эффективность сканирования конформационного пространства существенно отличается от исходной. Поэтому мы сравнили эффективность предсказания при автоматически выбираемом значении данного параметра с полученными ранее результатами. Сравнение производилось путем сопоставления коэффициентов корреляции с экспериментальными данными. Каждый раз случайно выбиралось 5 точек из разных диапазонов константы связывания, процедура была повторена 1000 раз. Как видно из (Рис. 27), даже исходные результаты Quick Vina 2 оказываются лучше, чем AutoDock Vina. Применение же рескоринг-

функции PeptoGrid дополнительно улучшает медиану коэффициента корреляции Спирмена на 6.4 процентных пункта. Исходный код обновленного пакета PeptoGrid доступен по ссылке <https://github.com/aozalevsky/peptogrid2/>

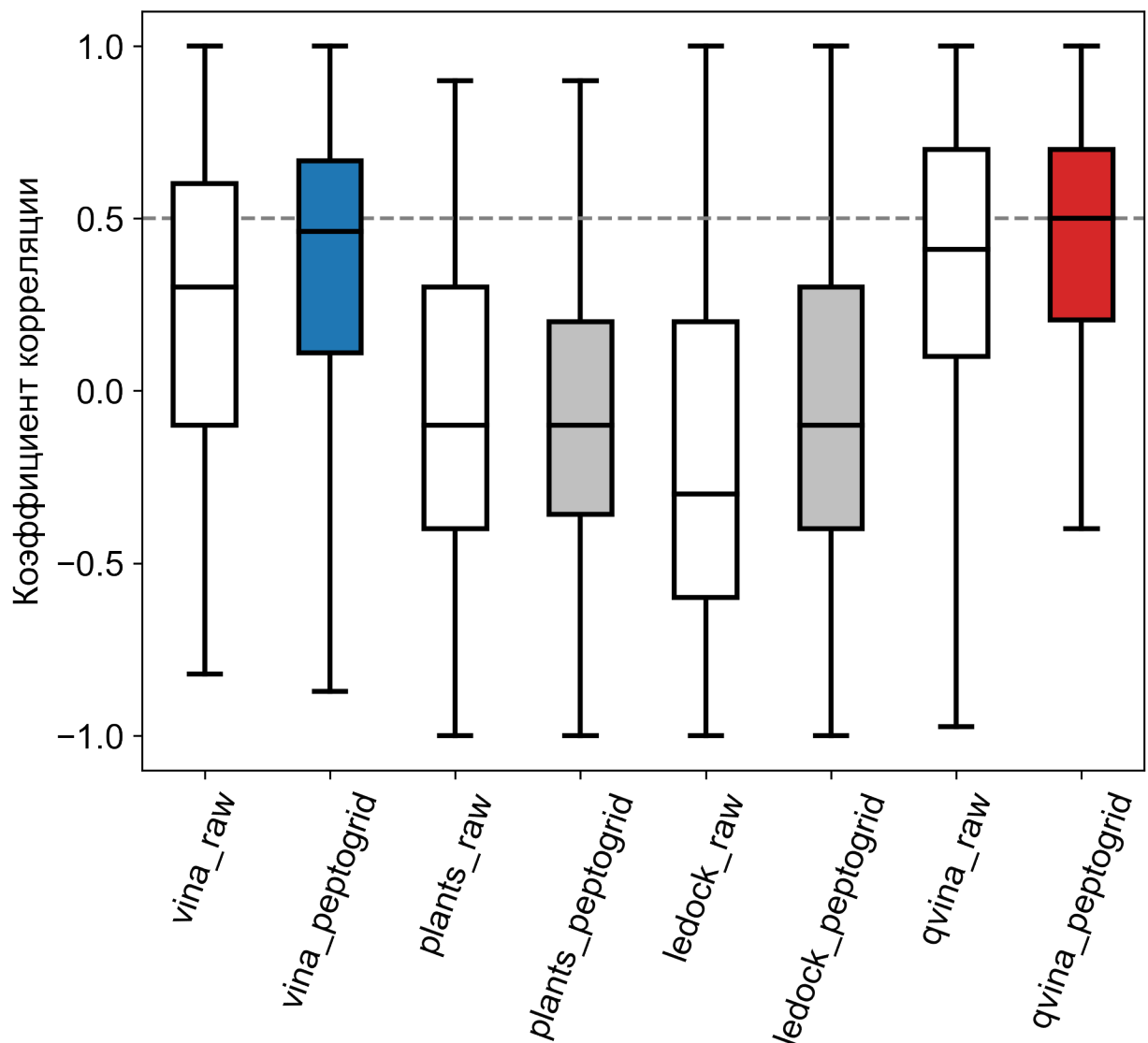


Рисунок 27. Качество предсказаний различных скоринг функций для пептид-белкового докинга. Синим и красным отмечены результаты применения рескоринг функции PeptoGrid к Vina и Qvina 2, соответственно.

С использованием Quick Vina 2 и 15 узлов суперкомпьютера “Ломоносов-2”, сканирование полной комбинаторной библиотеки трипептидов занимает порядка 100 минут.

6.2. Докинг пептидных библиотек

Для анализа субстратной специфичности с помощью разработанного нами инструмента PeptoGrid [75], мы провели массовый параллельный докинг библиотеки трипептидов в структуры тромбина, полученные в предыдущей главе.

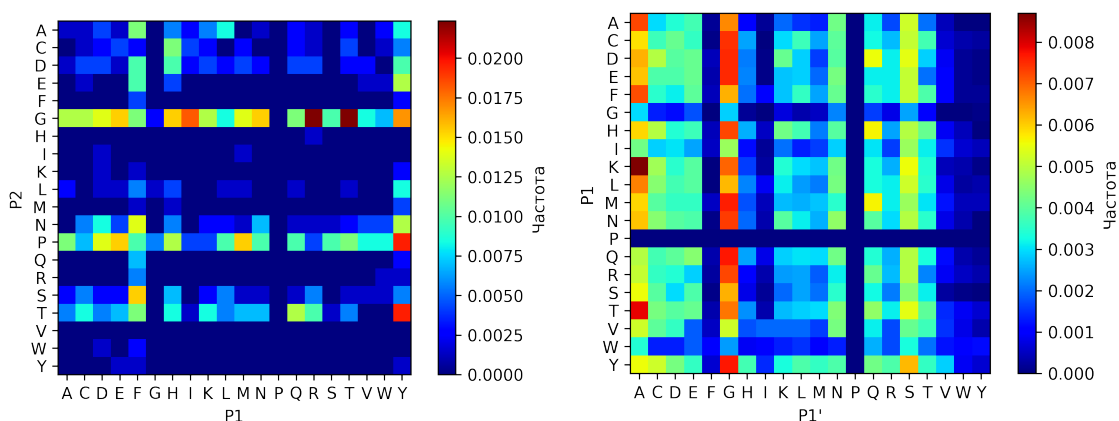


Рисунок 28. Профиль специфичности тромбина в положениях P2, P1, P1' быстрой формы тромбина

Реконструированный профиль специфичности, основанный на пептидах, чьи позы имеют конфигурацию, пригодную для гидролиза (всего 11 153 из 160 000 поз) демонстрирует, что наиболее частое сочетание в положениях P2-P1 — GR. В целом же, в P2 встречаются аминокислоты GPST, в P1 — R, в P1' —

AGS, что хорошо соотносится с экспериментальными данными о субстратной специфичности [69]. Отдельно интересно отметить наличие H, Y и F в P1 положении. Для фактора свертывания VIII мутация A372N является одним из природных вариантов, который вызывает гемофилию из-за того, что скорость активации такого белка тромбина падает в 80 раз [127]. Расположение же тирозина и фенилаланина (и его аналогов) в P1 положении является характерной особенностью целого ряда высокоселективных ингибиторов тромбина [128, 129].

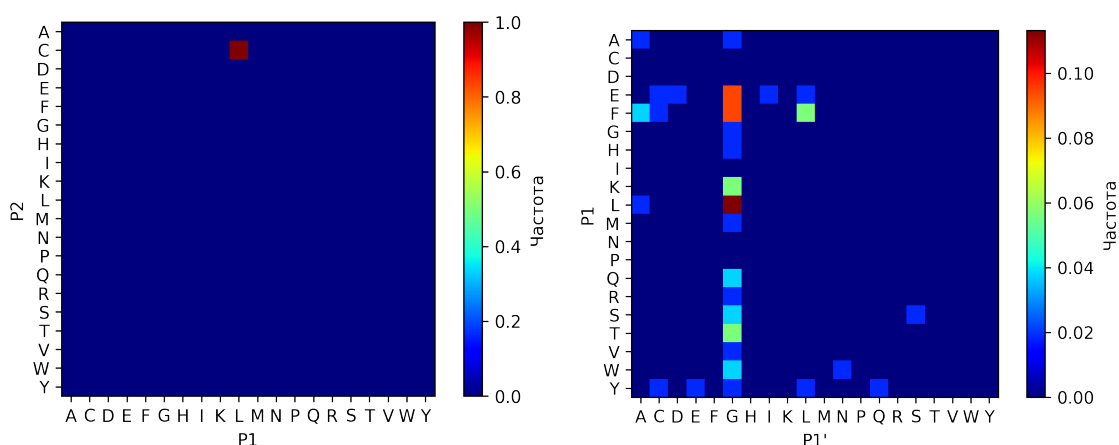


Рисунок 29. Профиль специфичности тромбина в положения P2, P1, P1' тромбина с ингибирующей водородной связью.

Для аналогичного профиля тромбина с ингибирующей водородной связью подходит всего 53 позы из 160 000 поз (и всего 1 поза в P2 положении), что скорее можно трактовать как отсутствие пригодных для катализа поз вообще. Такое изменение

геометрии активного центра вызывает общий сдвиг специфичности (коэффициент корреляции Пирсона между значениями оценочной функции ReptoGrid для двух форм тромбина равен 0.3). После чего данный фермент-субстратный комплекс все равно должен перейти в каталитическую конфигурацию. На кривой накопление продукта в ходе реакции данный переход может быть детектирован в виде дополнительной лаг-фазы при старте из условий, в которых заведомо превалирует медленная форма фермента [130].

ОБСУЖДЕНИЕ

1. Консистентность базы данных

PDB — уникальный источник структурной информации о белках. Однако автоматическая обработка его данных может быть сложной, в силу ряда причин, таких как меняющиеся со временем требования к оформлению записей, записи с несколькими молекулами, разнообразие именования определенных фрагментов молекул, в зависимости от текущих взглядов в данной области и т.д. Например, в случае тромбина, для именования тяжелой и легкой цепей в PDB используется 31 различных вариант написания. Более того, во многих структурах присутствуют лиганды, в том числе белковой природы, которые, к тому же, включены в состав цепей белка. Такое разнообразие сложно поддается автоматизации, так как требует ручного учета всех выпадающих вариантов и визуального контроля. Тем не менее, в результате у нас получилось выделить более 460 индивидуальных структур человеческого тромбина, что делает его одним из самых изученных белков современности.

2. Кластеризация структур тромбина

На первый взгляд, может показаться, что выбор метрики для задачи сравнения структур одного и того же белка должен быть тривиальным. Однако при ближайшем рассмотрении оказывается, что, используя метрики, разработанные для сравне-

ния заведомо удаленных структур [44, 45], мы рискуем потерять очень много информации о локальных особенностях. Например, об изменении положений радикалов отдельных аминокислот в активном центре фермента. Аналогичная проблема возникнет, если использовать только C_{α} RMSD.

Если же мы хотим иметь максимальную чувствительность, логично использовать полноатомное RMSD. Однако то, что на первый взгляд выглядит как один белок, по факту является чрезвычайно гетерогенным ансамблем структур, отличающихся по длине из-за различий в количестве неразрешенных остатков (в том числе из-за наличия сильно подвижных петель). Но так же существенно различаются и последовательности: присутствуют как единичные мутанты по остаткам активного центра [131], так и более крупные делеционные [132] и даже химерные мутанты [133].

В ходе предварительного тестирования мы обнаружили, что оптимальный результат дает расширение набора атомов до всех тяжелых. В случае использования RMSD по меньшим наборам атомов (например, только остову), была отмечена тенденция группировки структур по авторам или методикам кристаллизации, нежели чем по функциональным особенностям структур.

В качестве алгоритма кластеризации был выбран Affinity Propagation. В первом приближении, это непараметрический алгоритм кластеризации. Его достоинствами также является

возможность работы с заранее вычисленной матрицей сходства и высокая скорость работы [94]. Единственный существенный параметр — *preferences* вычисляется, по умолчанию, как медиана от всех значений матрицы сходства, и таким образом влияет на шанс точки стать центром кластера, что, в свою очередь, косвенным образом влияет на количество кластеров.

Мы применили две поправки, изменяющие значение по умолчанию. Во-первых, поправка на среднее количество совпавших атомов (или длину выравнивания, Выражение 6) повышает вероятность того, что центрами кластеров будут структуры с максимально полной последовательностью. Поправка на разрешение (Выражение 7) способствует тому, что центрами кластеров будут становиться структуры с лучшим разрешением. Применение обеих поправок позволило уменьшить количество кластеров с 55 до 20 (за счет группировки структур в быстрой форме), а также получить репрезентативные структуры с разрешением, как правило, не превышающим 2 Å (за исключением структур протромбина, где значения заведомо хуже). Так, например, центром кластера промежуточных структур стала 3VXE_HL, обладающая лучшим разрешением в 1.25 Å своем кластере и одним из лучших во всем наборе.

В результате, с учетом объединения структур в быстрой форме, получено 11 кластеров, покрывающих все пространство (Рис. 5) биологических состояний тромбина: от профермента, до активной E:Na⁺ формы. В рамках данной работы мы ограничи-

лись только одной, наиболее явной особенностью, отличающей промежуточные структуры. Однако более глубокий анализ, возможно, позволит обнаружить и другие детали. Например, особенности субстрат-связывающих регионов, так как ранее было продемонстрировано, что связывание лигандов может сдвигать конформационное равновесие, переводя тромбин в активную форму при помощи аллостерической модуляции [134].

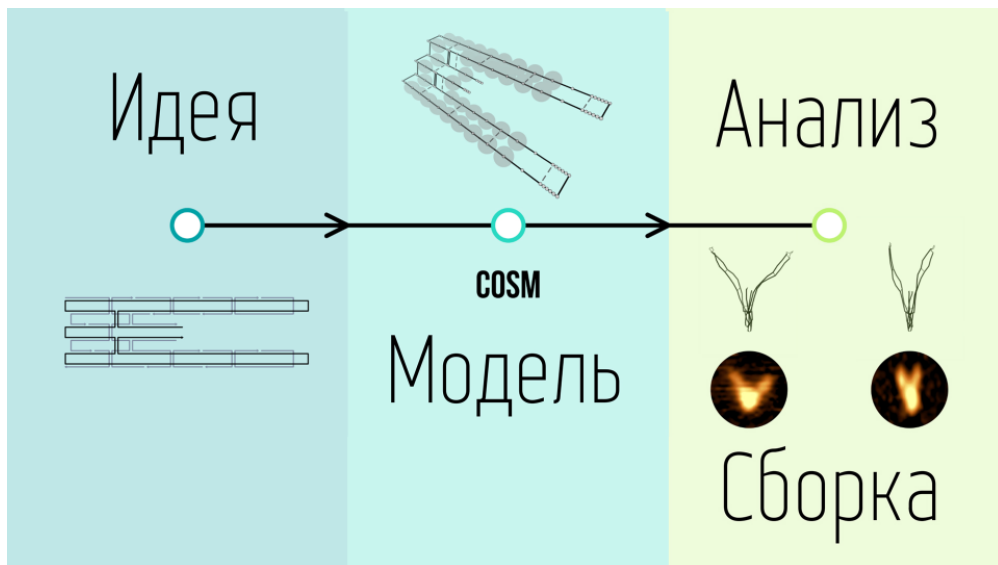


Рисунок 30. Схема работы и анализа сервиса для моделирования ДНК-оригами структур COSM [135]. В выдаче сервиса находятся результаты кластеризации при помощи инструмента AffBio. Экспериментальные изображения получены при помощи атомно-силовой спектроскопии.

Отсутствие необходимости априорного задания количества кластеров и гибкость управления характером кластеризации делают Affinity Propagation чрезвычайно удобным инстру-

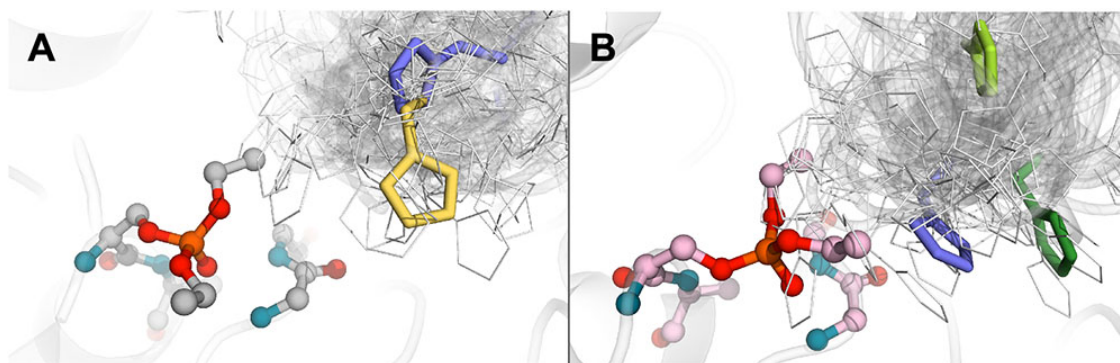


Рисунок 31. Кластеризация структур петли в составе бутирилхолинэстеразы человека. Цветом обозначены положения каталитических гистидинов в центрах кластеров для мутантов Clone_14 (слева) и Clone_15 (справа). В дальнейшем отобранные структуры подвергались КМ/ММ моделированию. [108].

ментом для анализа структур биомолекул. Поэтому мы использовали его как основу для инструмента AffBio, разработанного нами для параллельных окружений. Этот инструмент оказался незаменимым при анализе конформационных ландшафтов биомолекул в масштабе от больших ДНК-оригами конструкций, до небольших антимикробных пептидов и даже отдельных петель в составе белка [108, 135, 136]. В частности, использование AffBio позволило выделить отдельные состояния из траекторий крупнозернистого моделирования ДНК-оригами объектов в силовом поле COSM, идентичные экспериментальным (Рис. 30) [135], а в случае бутирилхолинэстеразы человека за счет кластеризации удалось уменьшить количество состояний (Рис. 31), для которых было необходимо производить вычислительно затратные КМ/ММ расчеты [108]. Исходный код пакета AffBio доступен по ссылке: <https://github.com/aozalevsky/affbio/>

3. Ингибирующая водородная связь

Тромбин, как было упомянуто ранее, является одним из самых изученных белков современности. На протяжении более чем 30 лет различные группы получали структуры в различных условиях, в том числе, прицельно, в условиях, призванных имитировать медленную форму: отсутствие ионов натрия [24], избыток катиона калия [137] или даже специальные мутантные формы [132]. Неоднократно отмечался и сдвиг каталитического серина, но это явление объяснялось в рамках обычной погрешности реконструкции пространственных структур по данным РСА [28]. Возможно, отсутствие атомов водорода в этих структурах и не позволило сделать предположение о потенциальной водородной связи каталитического серина с остовом Ser214.

4. КМ/ММ моделирование

Описание нового ингибирующего состояния невозможно без изучения влияния связывания катиона натрия на представленность данного состояния. Несмотря на то, что сам факт аллоsterической модуляции в тромбине неоднократно показан при помощи экспериментальных методов [28, 138], атомистические детали данных трансформаций до сих пор ясны не до конца. И одной из самых больших загадок является механизм передачи аллоsterического сигнала из натрий-связывающего сайта к активному центру фермента. Согласно гипотезе профессора Ди Цера (Di Cera) [21], роль такого механизма может выполнять сети

молекул воды. Ион натрия координирует несколько молекул воды в натрий-связывающем сайте, которые, в свою очередь, через цепочку водородных связей оказывают влияние на положение гамма-кислорода каталитического серина, расположенного на удалении порядка 16 Å (Рис. 32). Данная гипотеза была сформулирована на основе наблюдения о том, что в большом количестве кристаллических структур данные молекулы воды расположены в очень близких позициях [21].

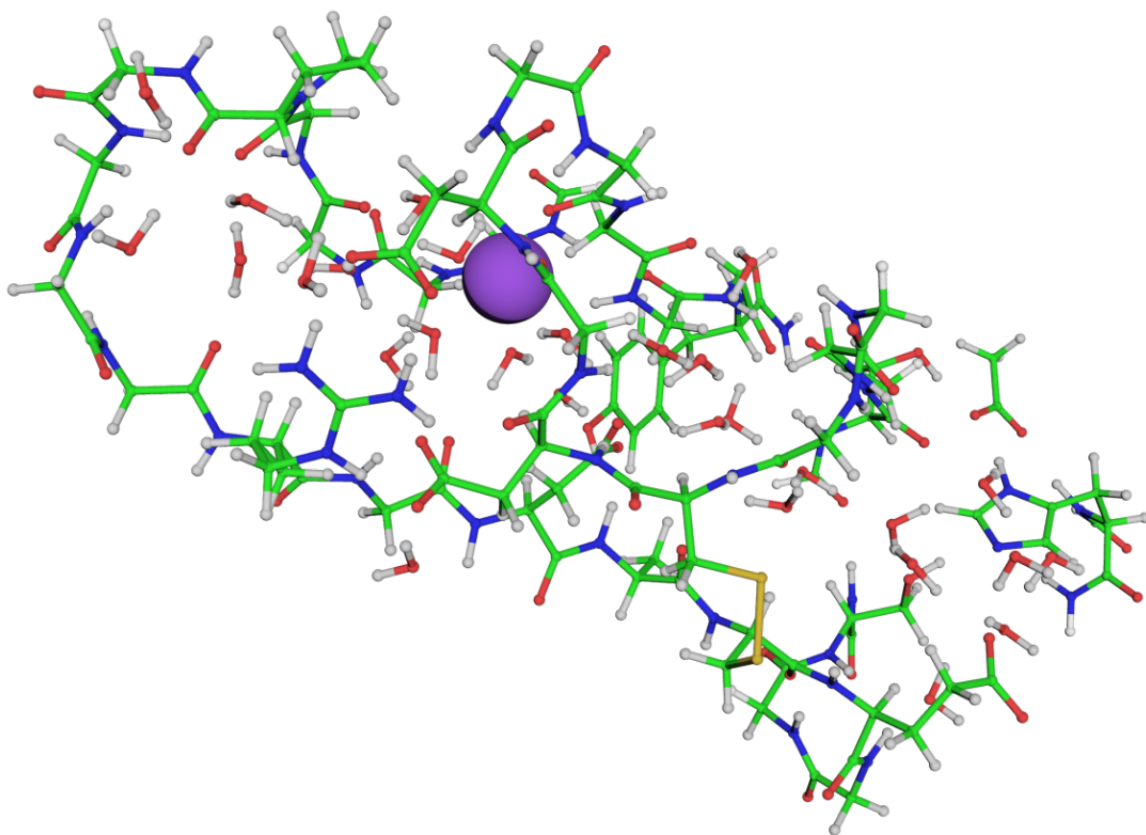


Рисунок 32. Сеть молекул воды в квантовой системе соединяет натрий-связывающий сайт активным центром.

Проверка подобной гипотезы при помощи экспериментальных методов является чрезвычайно сложной, так как невозможно устранить данную сеть и ее координаторов (среди которых присутствуют и остовные атомы) без полного разрушения субстрат-связывающего кармана. В тоже время, интересный эффект наблюдался при изучении ферментативной кинетики тромбина в растворителях с тяжелой (дейтериевой) водой. Присутствие тяжелой воды негативно сказывалось на скорости накопления продукта при реакции с субстратом, ассоциированным с быстрой формой фермента. В тоже время, при взаимодействии с субстратом, характерным для медленной формы, наоборот, наблюдалось ускорение реакции [130]. Интересно, что центр одной из промежуточных групп — 3VXE_NL, который натолкнул нас на идею о возможном существовании альтернативной водородной связи, также получен в рамках эксперимента по нейтронному рассеянию при pD 5.0 [139].

В случае, когда экспериментальная проверка может быть затруднительной, вычислительные методы способны оказать неоценимую помощь в проверке гипотезы. Однако в данном случае необходимо, чтобы выбранный вычислительный подход корректно описывал не просто геометрию белка, а еще и особенности координации молекул воды ионом металла, а также поведение молекул воды. Если, с точки зрения качества описания геометрии биомолекул, даже на длинных временах моделирования современные силовые поля уже достигли высокого

уровня точности [97, 140], то описание таких событий, как координация воды и катионов металлов в неполяризуемых полях, все еще далеко от совершенства [120]. Определенную сложность представляет и описание воды в изолированном окружении белка, где свойства среды существенно отличаются от нормальной фазы, как, например, в случае рибосомного туннеля [141].

С учетом описанных сложностей, гибридное КМ/ММ моделирование — один из немногих инструментов, с помощью которого можно описать системы подобного типа. Будучи корректно примененными, они могут раскрыть удивительно детальную картину того, что происходит в активном центре фермента [60]. Использование же методов повышения эффективности сканирования позволяет получить больше информации о поведении системы и количественно оценить величины соответствующих барьеров, а также сделать это за более короткое время, что особенно актуально для квантово-химических расчетов, которые являются бутылочным горлышком. Например, ранее мы успешно применили Well-Tempered метадинамику для изучения геометрии и барьера реакции бутирилхолинэстеразы человека. Это фермент с аналогичной каталитической триадой [108], состоящей из серина, гистидина и глутамата (в тромбине — аспартат) и имеющей сходную организацию активного центра — в том числе оксианионный сайт для координации карбонильного кислорода. Применение аналогичной методологии к

тромбину позволило нам обнаружить наличие локального минимума в состоянии с ингибирующей водородной связью в системе без иона. Причем величина барьера для перехода из этого состояния в основное каталитическое составляет ~ 0.8 ккал/моль.

5. Моделирование комплекса с субстратом

Для того, чтобы оценить влияние водородной связи на взаимодействие с субстратом, мы произвели моделирование субстрат-ферментного комплекса. Анализ фрагмента ландшафта, соединяющего области N-O и O-O (Рис. 24), демонстрирует, что если фермент-субстратный комплекс образовался с конформацией фермента, имеющей водородную связь с остовом Ser214, то для выхода из этого состояния необходимо преодолеть барьер 0.8 ккал/моль. Очевидным образом, это будет вносить вклад в общую кинетику процесса.

Отдельно стоит отметить наличие третьего состояния — вытесненного или Expelled (Рис. 24). Данное состояние аналогично кристаллическим структурам с негидролизруемыми ингибиторами тромбина [142] и, по всей видимости, является следствием того, что в формализме классической молекулярной механики невозможны разрывы и образования связей. Таким образом, до определенной степени субстрат выполняет роль негидролизуемого аналога. В то же время это означает, что все три характерных состояния, присутствующих в данном модели-

ровании, имеют однозначное биологическое соответствие.

6. Предсказание субстратной специфичности

Так как было известно, что переход из быстрой в медленную форму тромбина сдвигает специфичность, то мы решили проверить, обладает ли обнаруженная ингибирующая водородная связь подобным эффектом. Для этого мы предсказали профиль субстратной специфичности при помощи массового параллельного пептидного докинга (см. Рис28-29). Результаты предсказаний для быстрой формы, на основе библиотеки трипептидов, качественно соотносятся с экспериментальными данными — в положении P2 чаще всего встречаются аминокислоты GPST, в положении P1 самая частая аминокислота — R, в положении P1' — AGS [69]. В дальнейшем, точность предсказания можно увеличить за счет использования большего количества индивидуальных конформаций белка, а также с помощью перехода от библиотеки трипептидов к тетрапептидам, которые представляют текущий технический лимит для массивного параллельного докинга [78].

Предсказание для формы с ингибирующей водородной связью скорее демонстрирует полное пропадание реакционно пригодных конфигураций субстрата. Причиной для этого, по всей видимости, являются несколько факторов: во-первых, образование водородной связи Ser195 OG — Ser214 O изменяет микроокружение Ser214, которые участвует в распознавании и свя-

зывании субстратов [143]; во-вторых, происходит смещение гамма-кислорода каталитического серина 195, что также изменяет ландшафт субстрат-связывающего кармана.

7. Уникальность тромбина

Сводя воедино наблюдения, полученные из данных моделирования, можно заключить, что наличие возможности для формирования альтернативной водородной связи в активном центре тромбина существенно меняет поведение фермента. При этом катион натрия стабилизирует фермент в быстрой форме через сеть водородных связей, опосредованных цепочкой молекул воды. Что же делает тромбин таким уникальным? Почему для большинства других представителей семейства химотрипсиновых протеаз не наблюдается такая зависимость?

Возможная причина — наличие остатка Ala190 в субстрат-связывающем кармане S1. Большинство других представителей данного семейства имеют в данном положении серин, а замена на аланин произошла для аккомодации заряженных аминокислот — аргинина и лизина в P1 положении [144, 145]. Из-за отсутствия полярного радикала, аланин, в отличие от серина, не способен самостоятельно поддерживать цепочку водородных связей. Таким образом, появление натрий-связывающего сайта может быть попыткой нейтрализовать дезорганизирующий эффект аланина. В то же время, это создает и еще один регуляторный элемент, который может использоваться для тонкой регуляции

гемостаза. Эта гипотеза требует дальнейшей вычислительной и экспериментальной проверки.

ВЫВОДЫ

1. Новая реализация непараметрического алгоритма кластеризации Affinity Propagation позволяет выявлять биологически релевантные кластеры.

2. Сравнение кластеров, полученных из 400 записей тромбина в банке структур PDB, выявило возможность существования ранее не описанной ингибирующей водородной связи между O γ каталитического Ser195 и остовным кислородом Ser214.

3. С помощью нового пакета для гибридного КМ/ММ моделирования продемонстрировано влияние наличия иона Na⁺ на образование ингибирующей водородной связи. Дополнительный энергетический барьер для переключения ингибирующей водородной связи в присутствии пептидного субстрата составляет ~0.8 ккал/моль, что соответствует экспериментальным данным о замедлении реакции в 3-10 раз.

4. Новый метод ранжирования аффинности библиотек пептидов к белковым мишеням позволяет идентифицировать новые биоактивные пептиды.

5. Образование ингибирующей водородной связи изменяет субстратную специфичность тромбина.

ЗАКЛЮЧЕНИЕ

С помощью арсенала вычислительных методов — от квантово-химических расчетов, с повышенной эффективностью сканирования фазового пространства, до массивного параллельного докинга комбинаторных пептидных библиотек, обнаружена и продемонстрирована роль новой ингибирующей водородной связи в активном центре тромбина. Учитывая влияние катиона натрия на ее появление и величины энергетических барьеров, находящиеся в согласии с экспериментальными данными, можно предположить, что образование данной водородной связи является одним из ключевых событий при переключении белка из быстрой в медленную форму.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Krem, M.M. Di Cera, E. (2001) Molecular markers of serine protease evolution, *EMBO J.*, **20**, 3036–3045.
2. Wells, C.M. Di Cera, E. (1992) Thrombin is a Na(+)-activated enzyme, *Biochemistry*, **31**, 11721–11730.
3. Di Cera, E. (2008) Thrombin, *Mol. Aspects Med.*, **29**, 203–254.
4. Bradford, H.N. Krishnaswamy, S. (2012) Meizothrombin is an unexpectedly zymogen-like variant of thrombin, *J. Biol. Chem.*, **287**, 30414–30425.
5. Chen, Z., Pelc, L.A., Di Cera, E. (2010) Crystal structure of prethrombin-1, *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 19278–19283.
6. Whelihan, M.F., Zachary, V., Orfeo, T., Mann, K.G. (2012) Prothrombin activation in blood coagulation: the erythrocyte contribution to thrombin generation, *Blood*, **120**, 3837–3845.
7. Wikipedia (2019), Coagulation — Wikipedia, The Free Encyclopedia, URL: <https://en.wikipedia.org/wiki/Coagulation>, [Online; accessed 10-October-2019].
8. Hedstrom, L. (2002) Serine protease mechanism and specificity, *Chem. Rev.*, **102**, 4501–4524.

9. Agback, P. Agback, T. (2018) Direct evidence of a low barrier hydrogen bond in the catalytic triad of a Serine protease, *Sci Rep*, **8**, 10078.
10. Anderson, V.E., Ruszczycky, M.W., Harris, M.E. (2006) Activation of oxygen nucleophiles in enzyme catalysis, *Chem. Rev.*, **106**, 3236–3251.
11. Hoffman, R., Benz, E., Furie, B., Shattil, S. (2009) *Hematology: basic principles and practice*, Churchill Livingstone/Elsevier.
12. Wen, D.Z., Dittman, W.A., Ye, R.D., Deaven, L.L., Majerus, P.W., Sadler, J.E. (1987) Human thrombomodulin: complete cDNA sequence and chromosome localization of the gene, *Biochemistry*, **26**, 4350–4357.
13. Rydel, T.J., Ravichandran, K.G., Tulinsky, A., Bode, W., Huber, R., Roitsch, C., Fenton, J.W. (1990) The structure of a complex of recombinant hirudin and human alpha-thrombin, *Science*, **249**, 277–280.
14. Martorell, L., Martinez-Gonzalez, J., Rodriguez, C., Gentile, M., Calvayrac, O., Badimon, L. (2008) Thrombin and protease-activated receptors (PARs) in atherothrombosis, *Thromb. Haemost.*, **99**, 305–315.
15. Fernandez, P.V., Quintana, I., Cerezo, A.S., Caramelo, J.J., Pol-Fachin, L., Verli, H., Estevez, J.M., Ciancia, M. (2013)

- Anticoagulant activity of a unique sulfated pyranosic (1->3)-I2-L-arabinan through direct interaction with thrombin, *J. Biol. Chem.*, **288**, 223–233.
16. Dang, Q.D., Sabetta, M., Di Cera, E. (1997) Selective loss of fibrinogen clotting in a loop-less thrombin, *J. Biol. Chem.*, **272**, 19649–19651.
 17. Gianni, S., Ivarsson, Y., Bah, A., Bush-Pelc, L.A., Di Cera, E. (2007) Mechanism of Na(+) binding to thrombin resolved by ultra-rapid kinetics, *Biophys. Chem.*, **131**, 111–114.
 18. Vogt, A.D., Bah, A., Di Cera, E. (2010) Evidence of the E*-E equilibrium from rapid kinetics of Na⁺ binding to activated protein C and factor Xa, *J Phys Chem B*, **114**, 16125–16130.
 19. Lechtenberg, B.C., Freund, S.M., Huntington, J.A. (2012) An ensemble view of thrombin allostery, *Biol. Chem.*, **393**, 889–898.
 20. Bode, W. (2006) Structure and interaction modes of thrombin, *Blood Cells Mol. Dis.*, **36**, 122–130.
 21. Pineda, A.O., Carrell, C.J., Bush, L.A., Prasad, S., Caccia, S., Chen, Z.W., Mathews, F.S., Di Cera, E. (2004) Molecular dissection of Na⁺ binding to thrombin, *J. Biol. Chem.*, **279**, 31842–31853.
 22. Zhang, E. Tulinsky, A. (1997) The molecular environment of the Na⁺ binding site of thrombin, *Biophys. Chem.*, **63**, 185–200.

23. Carter, W.J., Myles, T., Gibbs, C.S., Leung, L.L., Huntington, J.A. (2004) Crystal structure of anticoagulant thrombin variant E217K provides insights into thrombin allostery, *J. Biol. Chem.*, **279**, 26387–26394.
24. Johnson, D.J., Adams, T.E., Li, W., Huntington, J.A. (2005) Crystal structure of wild-type human thrombin in the Na⁺-free state, *Biochem. J.*, **392**, 21–28.
25. Pineda, A.O., Chen, Z.W., Bah, A., Garvey, L.C., Mathews, F.S., Di Cera, E. (2006) Crystal structure of thrombin in a self-inhibited conformation, *J. Biol. Chem.*, **281**, 32922–32928.
26. De Filippis, V., De Dea, E., Lucatello, F., Frasson, R. (2005) Effect of Na⁺ binding on the conformation, stability and molecular recognition properties of thrombin, *Biochem. J.*, **390**, 485–492.
27. Pozzi, N., Zerbetto, M., Acquasaliente, L., Tescari, S., Frezzato, D., Polimeno, A., Gohara, D.W., Di Cera, E., De Filippis, V. (2016) Loop Electrostatics Asymmetry Modulates the Preexisting Conformational Equilibrium in Thrombin, *Biochemistry*, **55**, 3984–3994.
28. Huntington, J.A. (2008) How Na⁺ activates thrombin—a review of the functional and structural data, *Biol. Chem.*, **389**, 1025–1035.
29. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.

- (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.*, **112**, 535–542.
30. Chen, D. Dorling, A. (2009) Critical roles for thrombin in acute and chronic inflammation, *J. Thromb. Haemost.*, **7 Suppl 1**, 122–126.
31. Imokawa, Y. Brookes, J.P. (2003) Selective activation of thrombin is a critical determinant for vertebrate lens regeneration, *Curr. Biol.*, **13**, 877–881.
32. Choi, C.I., Yoon, H., Drucker, K.L., Langley, M.R., Kleppe, L., Scarisbrick, I.A. (2018) The Thrombin Receptor Restricts Subventricular Zone Neural Stem Cell Expansion and Differentiation, *Sci Rep*, **8**, 9360.
33. Lee, C.J. Ansell, J.E. (2011) Direct thrombin inhibitors, *Br J Clin Pharmacol*, **72**, 581–592.
34. Zavyalova, E.G., Legatova, V.A., Alieva, R.S., Zalevsky, A.O., Tashlitsky, V.N., Arutyunyan, A.M., Kopylov, A.M. (2019) Putative Mechanisms Underlying High Inhibitory Activities of Bimodular DNA Aptamers to Thrombin, *Biomolecules*, **9**, 41.
35. Tung, C.H., Gerszten, R.E., Jaffer, F.A., Weissleder, R. (2002) A novel near-infrared fluorescence sensor for detection of thrombin activation in blood, *Chembiochem*, **3**, 207–211.

36. Jiang, C., Zhao, T., Li, S., Gao, N., Xu, Q.H. (2013) Highly sensitive two-photon sensing of thrombin in serum using aptamers and silver nanoparticles, *ACS Appl Mater Interfaces*, **5**, 10853–10857.
37. Markou, M. Singh, S. (2003) Novelty Detection: A Review - Part 1: Statistical Approaches, *Signal Processing*, **83**, 2003.
38. Arkhangel'Skii, A., Pontryagin, L., O'Shea, D., Fedorchuk, V. (2011) *General Topology I: Basic Concepts and Constructions Dimension Theory*, Springer London, Limited.
39. Wallin, S., Farwer, J., Bastolla, U. (2003) Testing similarity measures with continuous and discrete protein models, *Proteins*, **50**, 144–157.
40. Kufareva, I. Abagyan, R. (2012) Methods of protein structure comparison, *Methods Mol. Biol.*, **857**, 231–257.
41. Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors, *Acta Crystallographica Section A*, **32**, 922–923.
42. Coutsiias, E.A., Seok, C., Dill, K.A. (2004) Using quaternions to calculate RMSD, *J Comput Chem*, **25**, 1849–1857.
43. Schwarzer, F. Lotan, I. (2003) Approximation of protein structure for fast similarity measures, В *Proceedings of the seventh annual international conference on Research in*

computational molecular biology, RECOMB '03, 267–276, ACM, New York, NY, USA.

44. Zemla, A., Venclovas, C., Moult, J., Fidelis, K. (1999) Processing and analysis of CASP3 protein structure predictions, *Proteins*, **Suppl 3**, 22–29.
45. Zhang, Y. Skolnick, J. (2004) Scoring function for automated assessment of protein structure template quality, *Proteins*, **57**, 702–710.
46. Siew, N., Elofsson, A., Rychlewski, L., Fischer, D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality, *Bioinformatics*, **16**, 776–785.
47. Bowman, G.R., Huang, X., Pande, V.S. (2009) Using generalized ensemble simulations and Markov state models to identify conformational states, *Methods*, **49**, 197–201.
48. Ye, Y. Godzik, A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching, *Nucleic Acids Res.*, **32**, W582–585.
49. Langfelder, P., Zhang, B., Horvath, S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R, *Bioinformatics*, **24**, 719–720.
50. Mallapragada, P.K., Jin, R., Jain, A. (2010) Non-parametric mixture models for clustering, в *Proceedings of the 2010 joint*

IAPR international conference on Structural, syntactic, and statistical pattern recognition, SSPR&SPR'10, 334–343, Springer-Verlag, Berlin, Heidelberg.

51. Li, J., Ray, S., Lindsay, B.G. (2007) A Nonparametric Statistical Approach to Clustering via Mode Identification, *J. Mach. Learn. Res.*, **8**, 1687–1723.
52. Zalevsky, A.O., Reshetnikov, R.V., Golovin, A.V. (2019) New QM/MM Implementation of the MOPAC2012 in the GROMACS, *B Supercomputing* (Voevodin V Sobolev S, eds.), 279–288, Springer International Publishing, Cham.
53. Warshel, A. Levitt, M. (1976) Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme, *J. Mol. Biol.*, **103**, 227–249.
54. Dapprich, S., Komáromi, I., Suzie, B., Morokuma, K., Frisch, M.J. (1999) A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives¹Dedicated to Professor Keiji Morokuma in celebration of his 65th birthday.¹, *Journal of Molecular Structure: THEOCHEM*, **461-462**, 1 – 21.
55. Abraham, M.J., van der Spoel, D., Lindahl, E., Hess, B. (2015) *GROMACS User Manual version 5.0.7*.

56. Vanatta, D.K., Shukla, D., Lawrenz, M., Pande, V.S. (2015) A network of molecular switches controls the activation of the two-component response regulator NtrC, *Nat Commun*, **6**, 7283.
57. Laio, A. Parrinello, M. (2002) Escaping free-energy minima, *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 12562–12566.
58. Quhe, R., Nava, M., Tiwary, P., Parrinello, M. (2015) Path Integral Metadynamics, *J Chem Theory Comput*, **11**, 1383–1388.
59. Bussi, G., Laio, A., Parrinello, M. (2006) Equilibrium free energies from nonequilibrium metadynamics, *Phys. Rev. Lett.*, **96**, 090601.
60. Zlobin, A.S., Zalevsky, A.O., Mokrushina, Y.A., Kartseva, O.V., Golovin, A.V., Smirnov, I.V. (2018) The Preferable Binding Pose of Canonical Butyrylcholinesterase Substrates Is Unproductive for Echothiophate, *Acta Naturae*, **10**, 121–124.
61. Stewart, J.J. (2013) Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters, *J Mol Model*, **19**, 1–32.
62. Gaus, M., Cui, Q., Elstner, M. (2012) DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB), *J Chem Theory Comput*, **7**, 931–948.

63. Grimme, S., Bannwarth, C., Shushkov, P. (2017) A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements ($Z = 1-86$), *J Chem Theory Comput*, **13**, 1989–2009.
64. Bannwarth, C., Ehlert, S., Grimme, S. (2019) GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions, *J Chem Theory Comput*, **15**, 1652–1671.
65. Pronk S., Pall S., S.R.L.P.B.P.A.R.S.M.R.S.J.C.K.P.M.v.d.S.D.H.B.L.E. (2013) GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, *Bioinformatics*, **29**, 845–854.
66. Kovatch, P., Costa, A., Giles, Z., Fluder, E., Cho, H.M., Mazurkova, S. (2015) Big Omics Data Experience, *SC Conf Proc*, **2015**.
67. Bevc, S., Konc, J., Stojan, J., Hodošek, M., Penca, M., Praprotnik, M., Janež, D. (2011) ENZO: a web tool for derivation and evaluation of kinetic models of enzyme catalyzed reactions, *PLoS ONE*, **6**, e22265.
68. Poreba, M., Salvesen, G.S., Drag, M. (2017) Synthesis of a HyCoSuL peptide substrate library to dissect protease substrate

- specificity, *Nat Protoc*, **12**, 2189–2214.
69. Gallwitz, M., Enoksson, M., Thorpe, M., Hellman, L. (2012) The extended cleavage specificity of human thrombin, *PLoS ONE*, **7**, e31756.
70. Yi, L., Gebhard, M.C., Li, Q., Taft, J.M., Georgiou, G., Iverson, B.L. (2013) Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries, *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 7229–7234.
71. Packer, M.S., Rees, H.A., Liu, D.R. (2017) Phage-assisted continuous evolution of proteases with altered substrate specificity, *Nat Commun*, **8**, 956.
72. duVerle, D.A. Mamitsuka, H. (2012) A review of statistical methods for prediction of proteolytic cleavage, *Brief. Bioinformatics*, **13**, 337–349.
73. Boyd, S.E., Pike, R.N., Rudy, G.B., Whisstock, J.C., Garcia de la Banda, M. (2005) PoPS: a computational tool for modeling and predicting protease specificity, *J Bioinform Comput Biol*, **3**, 551–585.
74. Chaudhury, S. Gray, J.J. (2009) Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance, *Structure*, **17**, 1636–1648.

75. Zalevsky, A.O., Zlobin, A.S., Gedzun, V.R., Reshetnikov, R.V., Lovat, M.L., Malyshev, A.V., Doronin, I.I., Babkin, G.A., Golovin, A.V. (2019) PeptoGrid-Rescoring Function for AutoDock Vina to Identify New Bioactive Molecules from Short Peptide Libraries, *Molecules*, **24**, 277.
76. Kruger, D.M., Glas, A., Bier, D., Pospiech, N., Wallraven, K., Dietrich, L., Ottmann, C., Koch, O., Hennig, S., Grossmann, T.N. (2017) Structure-Based Design of Non-natural Macrocyclic Peptides That Inhibit Protein-Protein Interactions, *J. Med. Chem.*, **60**, 8982–8988.
77. Cheng, T., Li, Q., Zhou, Z., Wang, Y., Bryant, S.H. (2012) Structure-based virtual screening for drug discovery: a problem-centric review, *AAPS J*, **14**, 133–141.
78. Rentzsch, R. Renard, B.Y. (2015) Docking small peptides remains a great challenge: an assessment using AutoDock Vina, *Brief. Bioinformatics*, **16**, 1045–1056.
79. Blaszczyk, M., Kurcinski, M., Kouza, M., Wieteska, L., Debinski, A., Kolinski, A., Kmiecik, S. (2016) Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking, *Methods*, **93**, 72–83.
80. Ciemny, M., Kurcinski, M., Kamel, K., Kolinski, A., Alam, N., Schueler-Furman, O., Kmiecik, S. (2018) Protein-peptide

docking: opportunities and challenges, *Drug Discov. Today*, **23**, 1530–1537.

81. Meng, X.Y., Zhang, H.X., Mezei, M., Cui, M. (2011) Molecular docking: a powerful approach for structure-based drug discovery, *Curr Comput Aided Drug Des*, **7**, 146–157.
82. Limongelli, V., Bonomi, M., Parrinello, M. (2013) Funnel metadynamics as accurate binding free-energy method, *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 6358–6363.
83. Thompson, D.C., Humblet, C., Joseph-McCarthy, D. (2008) Investigation of MM-PBSA rescoring of docking poses, *J Chem Inf Model*, **48**, 1081–1091.
84. Genheden, S. Ryde, U. (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities, *Expert Opin Drug Discov*, **10**, 449–461.
85. Li, H., Leung, K.S., Wong, M.H., Ballester, P.J. (2015) Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets, *Mol Inform*, **34**, 115–126.
86. Kurkinen, S.T., Niinivehmas, S., Ahinko, M., Latti, S., Pentikainen, O.T., Postila, P.A. (2018) Improving Docking Performance Using Negative Image-Based Rescoring, *Front Pharmacol*, **9**, 260.

87. Phatak, S.S., Stephan, C.C., Cavasotto, C.N. (2009) High-throughput and in silico screenings in drug discovery, *Expert Opin Drug Discov*, **4**, 947–959.
88. Efremov, R.G., Chugunov, A.O., Pyrkov, T.V., Priestle, J.P., Arseniev, A.S., Jacoby, E. (2007) Molecular lipophilicity in protein modeling and drug design, *Curr. Med. Chem.*, **14**, 393–415.
89. Fu, D.Y. Meiler, J. (2018) RosettaLigandEnsemble: A Small-Molecule Ensemble-Driven Docking Approach, *ACS Omega*, **3**, 3655–3664.
90. Rydel, T.J., Ravichandran, K.G., Tulinsky, A., Bode, W., Huber, R., Roitsch, C., Fenton, J.W. (1990) The structure of a complex of recombinant hirudin and human alpha-thrombin, *Science*, **249**, 277–280.
91. Bakan, A., Meireles, L.M., Bahar, I. (2011) ProDy: protein dynamics inferred from theory and experiments, *Bioinformatics*, **27**, 1575–1577.
92. Irback, A. Mohanty, S. (2006) PROFASI: A Monte Carlo simulation package for protein folding and aggregation, *J Comput Chem*, **27**, 1548–1555.
93. Tange, O. (2011) GNU Parallel - The Command-Line Power Tool, *;login: The USENIX Magazine*, **36**, 42–47.

94. Frey, B.J. Dueck, D. (2007) Clustering by Passing Messages Between Data Points, *Science*, **315**, 972–976.
95. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825–2830.
96. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., Shaw, D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins*, **78**, 1950–1958.
97. Aliev, A.E., Kulke, M., Khaneja, H.S., Chudasama, V., Sheppard, T.D., Lanigan, R.M. (2014) Motional timescale predictions by molecular dynamics simulations: case study using proline and hydroxyproline sidechain dynamics, *Proteins*, **82**, 195–215.
98. Best, R.B. Hummer, G. (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides, *J Phys Chem B*, **113**, 9004–9015.
99. Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E. (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation, *J Chem Theory Comput*, **4**, 435–447.

100. Westerlund, A.M., Harpole, T.J., Blau, C., Delemotte, L. (2018) Inference of Calmodulin's Ca^{2+} -Dependent Free Energy Landscapes via Gaussian Mixture Model Validation, *J Chem Theory Comput*, **14**, 63–71.
101. Smirnov, I.V., Golovin, A.V., Chatziefthimiou, S.D., Stepanova, A.V., Peng, Y., Zolotareva, O.I., Belogurov, A.A., Kurkova, I.N., Ponomarenko, N.A., Wilmanns, M., Blackburn, G.M., Gabibov, A.G., Lerner, R.A. (2016) Robotic QM/MM-driven maturation of antibody combining sites, *Sci Adv*, **2**, e1501695.
102. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. (1983) Comparison of simple potential functions for simulating liquid water, *The Journal of Chemical Physics*, **79**, 926–935.
103. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., Shaw, D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field, *Proteins*, **78**, 1950–1958.
104. Stewart, J.J.P. (1989) Optimization of parameters for semiempirical methods I. Method, *Journal of Computational Chemistry*, **10**, 209–220.
105. Nachon, F., Asojo, O.A., Borgstahl, G.E., Masson, P., Lockridge, O. (2005) Role of water in aging of human butyrylcholinesterase

- inhibited by echothiophate: the crystal structure suggests two alternative mechanisms of aging, *Biochemistry*, **44**, 1154–1162.
106. Nachon, F., Carletti, E., Wandhammer, M., Nicolet, Y., Schopfer, L.M., Masson, P., Lockridge, O. (2011) X-ray crystallographic snapshots of reaction intermediates in the G117H mutant of human butyrylcholinesterase, a nerve agent target engineered into a catalytic bioscavenger, *Biochem. J.*, **434**, 73–82.
107. Trott, O. Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, *J Comput Chem*, **31**, 455–461.
108. Zlobin, A., Mokrushina, Y., Terekhov, S., Zalevsky, A., Bobik, T., Stepanova, A., Aliseychik, M., Kartseva, O., Panteleev, S., Golovin, A., Belogurov, A., Gabibov, A., Smirnov, I. (2018) QM/MM Description of Newly Selected Catalytic Bioscavengers Against Organophosphorus Compounds Revealed Reactivation Stimulus Mediated by Histidine Residue in the Acyl-Binding Loop, *Front Pharmacol*, **9**, 834.
109. Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, *J Comput Chem*, **30**, 2785–2791.
110. Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E., Hutchison, G.R. (2012) Avogadro: an advanced

- semantic chemical editor, visualization, and analysis platform, *J Cheminform*, **4**, 17.
111. Alhossary, A., Handoko, S.D., Mu, Y., Kwoh, C.K. (2015) Fast, accurate, and reliable molecular docking with QuickVina 2, *Bioinformatics*, **31**, 2214–2216.
112. Schrödinger, LLC (2015) The PyMOL Molecular Graphics System, Version 1.8, pymol.
113. Hunter, J.D. (2007) Matplotlib: A 2D graphics environment, *Computing In Science & Engineering*, **9**, 90–95.
114. Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., Richardson, D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography, *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12–21.
115. Joosten, R.P., Salzemann, J., Bloch, V., Stockinger, H., Berglund, A.C., Blanchet, C., Bongcam-Rudloff, E., Combet, C., Da Costa, A.L., Deleage, G., Diarena, M., Fabbretti, R., Fettahi, G., Flegel, V., Gisel, A., Kasam, V., Kervinen, T., Korpelainen, E., Mattila, K., Pagni, M., Reichstadt, M., Breton, V., Tickle, I.J., Vriend, G. (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB, *J Appl Crystallogr*, **42**, 376–384.
116. Arunan, E., Desiraju, G.R., Klein, R.A., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D.C., Crabtree, R.H., Dannenberg,

- J.J., Hobza, P., Kjaergaard, H.G., Legon, A.C., Mennucci, B., Nesbitt, D.J. (2011) Definition of the hydrogen bond (IUPAC Recommendations 2011), *Pure and Applied Chemistry*, **83**, 1637 – 1641.
117. Ahmed, H.U., Blakeley, M.P., Cianci, M., Cruickshank, D.W., Hubbard, J.A., Helliwell, J.R. (2007) The determination of protonation states in proteins, *Acta Crystallogr. D Biol. Crystallogr.*, **63**, 906–922.
118. Smirnov, I., Carletti, E., Kurkova, I., Nachon, F., Nicolet, Y., Mitkevich, V.A., Debat, H., Avasse, B., Belogurov, A.A., Kuznetsov, N., Reshetnyak, A., Masson, P., Tonevitsky, A.G., Ponomarenko, N., Makarov, A.A., Friboulet, A., Tramontano, A., Gabibov, A. (2011) Reactibodies generated by kinetic selection couple chemical reactivity with favorable protein dynamics, *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 15954–15959.
119. Kromann, J.C., Steinmann, C., Jensen, J.H. (2018) Improving solvation energy predictions using the SMD solvation method and semiempirical electronic structure methods, *J Chem Phys*, **149**, 104102.
120. Bucher, D., Guidoni, L., Carloni, P., Rothlisberger, U. (2010) Coordination numbers of K(+) and Na(+) Ions inside the selectivity filter of the KcsA potassium channel: insights from first principles molecular dynamics, *Biophys. J.*, **98**, L47–49.

121. Tribello, G.A., Bonomi, M., Branduardi, D., Camilloni, C., Bussi, G. (2014) PLUMED 2: New feathers for an old bird, *Computer Physics Communications*, **185**, 604 – 613.
122. Bonomi, M., Bussi, G., Camilloni, C., Tribello, G.A., Bana?, P., Barducci, A., Bernetti, M., Bolhuis, P.G., Bottaro, S., Branduardi, D., Capelli, R., Carloni, P., Ceriotti, M., Cesari, A., Chen, H., Chen, W., Colizzi, F., De, S., De La Pierre, M., Donadio, D., Drobot, V., Ensing, B., Ferguson, A.L., Filizola, M., Fraser, J.S., Fu, H., Gasparotto, P., Gervasio, F.L., Giberti, F., Gil-Ley, A., Giorgino, T., Heller, G.T., Hocky, G.M., Iannuzzi, M., Invernizzi, M., Jelfs, K.E., Jussupow, A., Kirilin, E., Laio, A., Limongelli, V., Lindorff-Larsen, K., Lohr, T., Marinelli, F., Martin-Samos, L., Masetti, M., Meyer, R., Michaelides, A., Molteni, C., Morishita, T., Nava, M., Paissoni, C., Papaleo, E., Parrinello, M., Pfaendtner, J., Piaggi, P., Piccini, G., Pietropaolo, A., Pietrucci, F., Pipolo, S., Provati, D., Quigley, D., Raiteri, P., Raniolo, S., Rydzewski, J., Salvalaglio, M., Sosso, G.C., Spiwok, V., ?poner, J., Swenson, D.W.H., Tiwary, P., Valsson, O., Vendruscolo, M., Voth, G.A., White, A. (2019) Promoting transparency and reproducibility in enhanced molecular simulations, *Nat. Methods*, **16**, 670–673.
123. Pozzi, N., Barranco-Medina, S., Chen, Z., Di Cera, E. (2012) Exposure of R169 controls protein C activation and autoactivation, *Blood*, **120**, 664–670.
124. Korb, O., Stutzle, T., Exner, T.E. (2009) Empirical scoring

- functions for advanced protein-ligand docking with PLANTS, *J Chem Inf Model*, **49**, 84–96.
125. Wang, Z., Sun, H., Yao, X., Li, D., Xu, L., Li, Y., Tian, S., Hou, T. (2016) Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power, *Phys Chem Chem Phys*, **18**, 12964–12975.
126. Froestl, W. (2010) Chemistry and pharmacology of GABAB receptor ligands, *Adv. Pharmacol.*, **58**, 19–62.
127. Nogami, K., Zhou, Q., Wakabayashi, H., Fay, P.J. (2005) Thrombin-catalyzed activation of factor VIII with His substituted for Arg372 at the P1 site, *Blood*, **105**, 4362–4368.
128. Riester, D., Wirsching, F., Salinas, G., Keller, M., Gebinoga, M., Kamphausen, S., Merkwirth, C., Goetz, R., Wiesenfeldt, M., Sturzebecher, J., Bode, W., Friedrich, R., Thurk, M., Schwienhorst, A. (2005) Thrombin inhibitors identified by computer-assisted multiparameter design, *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 8597–8602.
129. Lee, S.L., Alexander, R.S., Smallwood, A., Trievel, R., Mersinger, L., Weber, P.C., Kettner, C. (1997) New inhibitors of thrombin and other trypsin-like proteases: hydrogen bonding of an aromatic cyano group with a backbone amide of the P1 binding

- site replaces binding of a basic side chain, *Biochemistry*, **36**, 13180–13186.
130. Zhang, D. Kovach, I.M. (2005) Full and partial deuterium solvent isotope effect studies of alpha-thrombin-catalyzed reactions of natural substrates, *J. Am. Chem. Soc.*, **127**, 3760–3766.
131. Gandhi, P.S., Chen, Z., Di Cera, E. (2010) Crystal structure of thrombin bound to the uncleaved extracellular fragment of PAR1, *J. Biol. Chem.*, **285**, 15393–15398.
132. Bah, A., Carrell, C.J., Chen, Z., Gandhi, P.S., Di Cera, E. (2009) Stabilization of the E* form turns thrombin into an anticoagulant, *J. Biol. Chem.*, **284**, 20034–20040.
133. Pozzi, N., Chen, R., Chen, Z., Bah, A., Di Cera, E. (2011) Rigidification of the autolysis loop enhances Na(+) binding to thrombin, *Biophys. Chem.*, **159**, 6–13.
134. Handley, L.D., Fuglestad, B., Stearns, K., Tonelli, M., Fenwick, R.B., Markwick, P.R., Komives, E.A. (2017) NMR reveals a dynamic allosteric pathway in thrombin, *Sci Rep*, **7**, 39575.
135. Reshetnikov, R.V., Stolyarova, A.V., Zalevsky, A.O., Panteleev, D.Y., Pavlova, G.V., Klinov, D.V., Golovin, A.V., Protopopova, A.D. (2018) A coarse-grained model for DNA origami, *Nucleic Acids Res.*, **46**, 1102–1112.

136. Rogozhin, E., Zalevsky, A., Mikov, A., Smirnov, A., Egorov, T. (2018) Characterization of Hydroxyproline-Containing Hairpin-Like Antimicrobial Peptide EcAMP1-Hyp from Barnyard Grass (*Echinochloa crusgalli* L.) Seeds: Structural Identification and Comparative Analysis of Antifungal Activity, *Int J Mol Sci*, **19**, 3449.
137. Papaconstantinou, M.E., Carrell, C.J., Pineda, A.O., Bobofchak, K.M., Mathews, F.S., Flordellis, C.S., Maragoudakis, M.E., Tsopanoglou, N.E., Di Cera, E. (2005) Thrombin functions through its RGD sequence in a non-canonical conformation, *J. Biol. Chem.*, **280**, 29393–29396.
138. Lechtenberg, B.C., Johnson, D.J., Freund, S.M., Huntington, J.A. (2010) NMR resonance assignments of thrombin reveal the conformational and dynamic effects of ligation, *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 14087–14092.
139. Yamada, T., Kurihara, K., Ohnishi, Y., Tamada, T., Tomoyori, K., Masumi, K., Tanaka, I., Kuroki, R., Niimura, N. (2013) Neutron and X-ray crystallographic analysis of the human $\hat{I}\pm$ -thrombin-bivalirudin complex at pD 5.0: protonation states and hydration structure of the enzyme-product complex, *Biochim. Biophys. Acta*, **1834**, 1532–1538.
140. Beauchamp, K.A., Lin, Y.S., Das, R., Pande, V.S. (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark

- on 524 Diverse NMR Measurements, *J Chem Theory Comput*, **8**, 1409–1414.
141. Lucent, D., Snow, C.D., Aitken, C.E., Pande, V.S. (2010) Non-bulk-like solvent behavior in the ribosome exit tunnel, *PLoS Comput. Biol.*, **6**, e1000963.
142. Malley, M.F., Tabernero, L., Chang, C.Y., Ohringer, S.L., Roberts, D.G., Das, J., Sack, J.S. (1996) Crystallographic determination of the structures of human alpha-thrombin complexed with BMS-186282 and BMS-189090, *Protein Sci.*, **5**, 221–228.
143. Krem, M.M., Prasad, S., Di Cera, E. (2002) Ser(214) is crucial for substrate binding to serine proteases, *J. Biol. Chem.*, **277**, 40260–40264.
144. Katz, B.A., Mackman, R., Luong, C., Radika, K., Martelli, A., Sprengeler, P.A., Wang, J., Chan, H., Wong, L. (2000) Structural basis for selectivity of a small molecule, S1-binding, submicromolar inhibitor of urokinase-type plasminogen activator, *Chem. Biol.*, **7**, 299–312.
145. Sichler, K., Hopfner, K.P., Kopetzki, E., Huber, R., Bode, W., Brandstetter, H. (2002) The influence of residue 190 in the S1 site of trypsin-like serine proteases on substrate selectivity is universally conserved, *FEBS Lett.*, **530**, 220–224.

ПРИЛОЖЕНИЕ

Таблица 5 . Описание кластеров

Представитель	Состав
5JDU_CD	5JDU_CD
1NU9_A	1NU9_A 1NU9_D
5EDM_A	5EDM_A
4NZQ_A	4NZQ_A 4003_A
3NXP_A	3NXP_A 4HZH_A 4HZH_B 5EDK_A
6C2W_A	6BJR_A 6C2W_A 6C2W_B
2AFQ_CD	2A0Q_CD 2AFQ_AB 2AFQ_CD 5JDU_AB
3BEI_AB	2GP9_AB 3BEI_AB 3GIC_AB 3JZ2_AB 3S7H_AB
3K65_AB	1HAG_E 3K65_AB 3SQE_E 3SQH_E 4H6T_A 4RN6_A 4RN6_B
4CH8_AB	1AWH_CD 1B7X_AB 1DE7_LH 1JOU_AB 1JOU_EF 1RD3_AB 1RD3_CD 1SFQ_AB 1SHH_AB 1SR5_CB 1TB6_HL 1THP_AB 1TQ7_AB 1UVS_HL 1XMN_CD 1XMN_EF 1XMN_GH 1Z8I_AB 2A0Q_AB 2HWL_AB 2HWL_CD 2OD3_AB 2PGQ_AB 2THF_AB 3B9F_HL 3BV9_AB 3JZ1_AB 3QDZ_AB 3QDZ_CD 3QGN_AB 3S7K_AB 3S7K_CD 4CH2_AB 4CH2_CD 4CH8_AB 4CH8_CD 4CH8_EF 4CH8_GH 4DII_HLD 4DT7_CD 4H6S_AB 4HFP_AB 4I7Y_HL 4RKJ_AB 5L6N_IHL 5NHU_DC
3VXE_HL	1NY2_12 1SFQ_DE 1TMT_HL 1TMU_HL 3VXE_HL 3VXF_HL
1NT1_A	1MU6_AB 1MU8_AB 1MUE_AB 1NT1_A 1SL3_A 1TA2_A 1TA6_A 1Z71_A 1ZGI_A 1ZGV_A 1ZRB_A 3C1K_A
1A4W_HL	1A4W_HL 1ABI_HL 1ABJ_HL 1BHX_ABEF 1BMM_HL 1BMN_HL 1DWB_HL 1DWC_HL 1DWD_HL 1DWE_HL 1E0F_AD 1E0F_BE 1E0F_CF 1FPC_HL 1FPH_HL 1HLT_JK 1JMO_HL 1O0D_LH 1OOK_AB 1PPB_HL 1QUR_HL 1SG8_DE 1SGI_AB 1THR_HL 1THS_LH 1TQ0_AB 1TQ0_CD 1VR1_HL 2A2X_LH 2A45_AB 2A45_ED 2HPQ_HLP 2PGB_AB 2PW8_HL 3BEF_AB 3BEF_DE 3E6P_HL 3EE0_AB 3HKJ_AB 3LU9_AB 3PMH_AB 4DT7_AB 4HFP_CD 4HTC_LH 4LZ4_CD 4MLF_AB 5NHU_LH 6EVV_HL 6GN7_HL
1BB0_AB	1A2C_HL 1A46_HL 1A5G_HL 1A61_HL 1AE8_HL 1AFE_HL 1AHT_HL 1AY6_HL 1B5G_HL 1BA8_AB 1BB0_AB 1C1U_HL 1C1V_HL 1C1W_HL 1C4U_12 1C4V_12 1C4Y_12 1C5L_HL 1C5N_HL 1C5O_HL 1CA8_AB 1D6W_A 1D9I_A 1DIT_HL 1DM4_AB 1GHV_HL 1GHW_HL 1GHX_HL 1GHY_LH 1GJ4_HL 1GJ5_HL 1HAH_HL 1HAI_HL 1HAO_HL 1HAP_HL 1HGT_HL 1HLT_HL 1HUT_HL 1HXE_HL 1HXF_HL 1JWT_A 1LHC_HL 1LHD_HL 1LHE_HL 1LHF_HL 1LHG_HL 1NRO_HL 1NRP_HL 1NRQ_HL 1NRR_HL 1NRS_HL 1O2G_HL 1O5G_HL 1SGI_DE 1T4U_HL 1T4V_HL 1TBZ_HL 1TMB_HL 1TWX_AB 1UMA_HL 1Z8J_AB 2HGT_HL 2HNT_CELF 3HAT_HL 4THN_HL 5GDS_HL 6CYM_AB 6CYM_CD 7KME_HL 8KME_12

Продолжение на следующей странице

Таблица 5 – Продолжение

Представитель	Состав
3EQ0_HL	1AD8_HL 1BBR_LHE 1D3D_AB 1D3P_AB 1D3Q_AB 1D3T_AB 1D4P_AB 1G30_AB 1G32_AB 1H8D_HL 1H8I_HL 1HDT_HL 1KTS_AB 1KTT_AB 1W7G_HL 2ANK_HL 2ANM_HL 2BDY_A 2FES_HL 2PKS_ACB 2R2M_AB 2ZC9_HL 2ZDA_HL 2ZDV_HL 2ZF0_HL 2ZFF_HL 2ZFP_HL 2ZGB_HL 2ZGX_HL 2ZHE_HL 2ZHQ_HL 2ZHW_HL 2ZI2_HL 2ZIQ_HL 2ZNK_HL 2ZO3_HL 3C27_AB 3D49_HL 3DHK_HL 3DT0_HL 3DUX_HL 3EGK_HL 3EQ0_HL 3F68_HL 3HKJ_DE 3PO1_ACB 3QLP_HL 3SHC_HL 3U8R_HL 3U8T_HL 3U9A_HL 3UTU_HL 4BOH_LHM 4LZ4_AB 5E8E_HL
5AFY_HL	1EB1_HL 1K21_HL 1MH0_A 1NM6_A 1NO9_HL 1NU7_AB 1NU7_EF 1NZQ_HL 1TOM_HL 1WAY_AB 1WBG_AB 1XM1_A 2B5T_AB 2B5T_CD 2BVR_HL 2FEQ_LH 2H9T_HL 3BF6_HL 3HKI_AB 3P70_AB 3P70_CD 3P70_GH 3QTV_HL 3QWC_HL 3R3G_AB 3RLW_HL 3RLY_HL 3RM0_HL 3RM2_HL 3RML_HL 3RMM_HL 3RMN_HL 3RMO_HL 3SHA_HL 3SI3_HL 3SI4_HL 3SV2_HL 3T5F_HL 3U98_HL 3UWJ_HL 4BOH_BA 4UFD_HL 5AFY_HL 5GIM_AB 5MM6_HL
4YES_AB	1BBR_JK 1BBR_MN 1DE7_JK 1DX5_AM 1DX5_CO 1EOJ_A 1EOL_A 1JOU_CD 1MH0_B 1NRN_HL 1SB1_HL 1YCP_JKM 1YCP_LH 2GDE_HL 2JH0_CD 2JH5_CD 2JH6_CD 3GIS_AB 3GIS_CD 3GIS_EF 3HKI_DE 3LDX_HL 3LU9_DE 3TU7_HL 4BAK_AB 4BAM_AB 4BAN_AB 4BAO_AB 4BAQ_AB 4DY7_AB 4DY7_ED 4E7R_MG 4LOY_HL 4LXB_HL 4RKO_AB 4YES_AB 5CMX_HL 6EO8_HL 6EO9_HL
5AFZ_HL	1A3B_HL 1A3E_HL 1A18_HL 1AIX_HL 1DOJ_A 1DX5_BN 1DX5_DP 1G37_A 1HBT_HL 1IHS_HL 1IHT_HL 1P8V_CB 1QBV_HL 1RIW_ACB 1SG8_AB 1SHH_DE 1XMN_AB 1YPK_HL 1YPM_HL 2BVS_HL 2BVX_HL 2BXT_HL 2BXU_HL 2C8W_AB 2C8X_AB 2C8Y_AB 2C8Z_AB 2C90_AB 2C93_AB 2HPP_HLP 2ZFQ_HL 2ZFR_HL 2ZG0_HL 2ZHF_HL 3B23_AB 3BIU_HL 3BIV_HL 3DD2_HL 3P17_HL 3P6Z_AB 3P6Z_GH 3P70_EF 3QTO_HL 3QX5_HL 3U69_HL 3U8O_HL 4AX9_HL 4DIH_HLD 4E05_HL 4E06_HL 4E7R_LH 4LZ1_HLD 4UD9_HL 4UDW_HL 4UE7_HL 4UEH_HL 4UFE_HL 4UFF_HL 4UFG_HL 5A2M_HL 5AF9_HL 5AFZ_HL 5AHG_HL 5DO4_HL 5EW1_HL 5EW2_HL 5JFD_HL 5JZY_HL 5LCE_HL 5LPD_HL 5MJT_HL 5MLS_HL 5NHU_BA 5TO3_AB 6EO6_HL 6EO7_HL 6FJT_HL 6GBW_HL
2CN0_HL	1K22_HL 1OYT_HL 1VZQ_HL 1YPE_HL 1YPG_HL 1YPJ_HL 1YPL_HL 2CF8_HL 2CF9_HL 2CN0_HL 2UUF_AB 2UUI_AB 2UUK_AB 2V3H_HL 2V3O_HL 3DA9_AB 4AYV_AB 4AYY_AB 4AZ2_AB 4BAH_AB
1BCU_HL	1AWF_HL 1AWH_AB 1BCU_HL 1QHR_AB 1QJ1_AB 1QJ6_AB 1QJ7_AB