

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

НИИ ФИЗИКО-ХИМИЧЕСКОЙ БИОЛОГИИ

имени А.Н. БЕЛОЗЕРСКОГО

На правах рукописи

Тимонина Дарья Сергеевна

**Биоинформатический анализ суперсемейств белков на уровне
3D-структурной организации с использованием методов
машинного обучения**

1.5.8 - математическая биология, биоинформатика

ДИССЕРТАЦИЯ

на соискание ученой степени

кандидата биологических наук

Научный руководитель:
д.х.н., профессор Швядас Витаутас-Юозас Каятоно

Москва – 2023

Оглавление

1. Список сокращений.....	5
2. Введение.....	6
2.1. Актуальность темы исследований.....	6
2.2. Степень разработанности темы исследования.....	9
2.3. Цель и задачи работы.....	11
2.4. Объект и предмет исследования.....	12
2.5. Научная новизна.....	12
2.6. Теоретическая и практическая значимость работы.....	12
2.7. Методология и методы исследования.....	13
2.8. Степень достоверности.....	13
2.9. Личный вклад автора.....	14
2.10. Положения, выносимые на защиту.....	14
2.11. Публикации по теме работы.....	15
2.12. Апробация работы.....	16
2.13. Структура и объем диссертации.....	16
3. Обзор литературы.....	17
3.1. Понятия суперсемейства, семейства, подсемейства.....	17
3.2. Современный биоинформатический анализ суперсемейств белков.....	18
3.2.1. Анализ суперсемейств белков на уровне аминокислотных последовательностей....	18
3.2.1.1. Консервативные позиции суперсемейства белков.....	19
3.2.1.2. Специфические позиции подсемейства белков/позиции, определяющие специфичность.....	21
3.2.1.3. Коррелирующие позиции суперсемейства белков.....	25
3.2.2. Инструменты для анализа и сравнения гомологичных белков, использующие структурные данные.....	29
3.2.2.1. Построение множественного структурного выравнивания гомологичных белков	29
3.2.2.2. Пакеты молекулярной визуализации и анализа.....	35
3.2.2.3. Методы для анализа выравнивания структур гомологичных белков.....	38
3.2.2.4. Консервативные структурные паттерны суперсемейства белков: 3D-мотивы....	42
3.2.2.4.1. Понятие 3D-мотива.....	42
3.2.2.4.2. Методы выявления 3D-мотивов.....	43
3.3. Методы машинного обучения.....	48
3.3.1. Обучение с учителем.....	48
3.3.1.1. Постановка задачи.....	48
3.3.1.2. Алгоритмы машинного обучения с учителем.....	49
3.3.2. Обучение без учителя.....	49

3.3.2.1.	Постановка задачи	49
3.3.2.2.	Алгоритм кластеризации k-средних	50
3.3.2.3.	Алгоритм кластеризации DBSCAN.....	50
3.3.2.4.	Алгоритм кластеризации OPTICS.....	51
3.3.2.5.	Алгоритм кластеризации HDBSCAN	54
4.	Материалы и методы	57
4.1.	Построение множественного структурного выравнивания белков.....	57
4.2.	Визуализация структур белков	57
4.3.	Библиотеки, использованные в программном обеспечении для поиска 3D-специфических паттернов суперсемейства	57
4.4.	Оценка качества кластеризации участков основной и боковых цепей белков суперсемейства с использованием метрики силуэт	58
4.5.	База данных конформационного разнообразия белков PDBFlex. Создание выборки для расчета статистики с целью определения функционально-значимых 3D-специфических паттернов	58
4.6.	Z-оценка статистической значимости и соответствующая P-оценка	60
4.7.	Создание выборок для апробации нового подхода к анализу 3D-специфичности в структурах белков суперсемейства	61
4.8.	Применение метода Zebra2 для получения специфических позиций подсемейства	63
4.9.	Adjusted Rand Index – мера сходства двух кластеризаций, использованная для сравнения результатов разделения белков на подсемейства	63
4.10.	Извлечение дисульфидных мостиков из структур белков базы данных PDB для получения 3D-мотивов.....	64
4.11.	Апробация и расчет специфичности и чувствительности статистического критерия определения возможности вставки данного 3D-мотива в структуру белка	65
5.	Результаты и обсуждение	66
5.1.	Новый подход к анализу 3D-специфичности в структурах суперсемейства белков	67
5.1.1.	Поиск 3D-специфических паттернов в основной и боковых цепях белков суперсемейства.....	67
5.1.1.1.	Структура алгоритма поиска функционально-значимых 3D-специфических паттернов	67
5.1.1.2.	Подробное описание алгоритма поиска функционально-значимых 3D-специфических паттернов	69
5.1.1.2.1.	Выявление «общих» участков и участков «вариабельности» основной цепи суперсемейства белков	69
5.1.1.2.2.	Разделение участков основных и боковых цепей белков суперсемейства на пространственно-эквивалентные кластеры.....	72
5.1.1.2.3.	Оценка специфичности для 3D-специфического паттерна суперсемейства белков	77
5.1.1.2.4.	Статистическая модель для определения функционально значимых 3D-специфических паттернов суперсемейства белков.....	79

5.1.2.	Разработка программного обеспечения для поиска 3D-специфических паттернов суперсемейства.....	81
5.1.2.1.	Программное обеспечение для поиска 3D-специфических паттернов суперсемейства белков в основной цепи	81
5.1.2.1.1.	Описание входных данных	81
5.1.2.1.2.	Описание настраиваемых параметров	84
5.1.2.1.3.	Описание вывода.....	88
5.1.2.2.	Программное обеспечение для поиска 3D-специфических паттернов суперсемейства белков в боковой цепи	95
5.1.2.2.1.	Описание входных данных	95
5.1.2.2.2.	Описание настраиваемых параметров	95
5.1.2.2.3.	Описание вывода.....	96
5.2.	Апробация нового подхода на широкой выборке суперсемейств белков.....	99
5.2.1.	3D-специфические паттерны, отвечающие за различие в свойствах между ферментами, принадлежащими различным функциональным подсемействам	101
5.2.2.	3D-специфические паттерны, отвечающие за различия свойств конформеров одного белка	114
5.2.3.	Обобщение результатов исследования 3D-специфических паттернов суперсемейства белков	120
5.2.4.	Сравнение результатов применения метода выявления 3D-специфических паттернов и метода выявления специфических позиций подсемейства на выборке суперсемейств белков	126
5.2.5.	Сравнение результатов применения метода выявления 3D-специфических паттернов и метода выявления коррелирующих позиций на выборке суперсемейств белков	132
5.3	3D-мотивы. Статистическая модель оценки структурной гибкости основной цепи 3D-мотивов дисульфидных мостиков для определения возможности вставки данного 3D-мотива в структуру белка.....	136
5.3.1	Получение 3D-мотивов дисульфидных мостиков	136
5.3.3.	Статистическая модель оценки структурной гибкости основной цепи 3D-мотивов для определения возможности вставки данного 3D-мотива в структуру белка на примере 3D-мотивов дисульфидных мостиков.....	137
5.3.4.	Апробация статистической модели оценки структурной гибкости основной цепи 3D-мотивов дисульфидных мостиков.....	138
6.	Заключение	139
7.	Основные результаты и выводы	141
8.	Список литературы	143

1. Список сокращений

(H)DBSCAN – (Hierarchical) Density-Based Spatial Clustering of Applications with Noise/(Иерархическая) основанная на плотности пространственная кластеризация для приложений с шумами (метод кластеризации);

OPTICS – Ordering Points to Identify the Clustering Structure/Упорядочение точек для обнаружения кластерной структуры (метод кластеризации);

ООП – Объектно-Ориентированное Программирование;

RMSD – Root Mean Square Deviation/Среднеквадратичное отклонение;

СУОЦ – Специфический для подсемейств Участок Основной Цепи (3D-специфический паттерн, найденный в основной цепи);

СОБЦ – Специфическая для подсемейств Ориентация Боковой Цепи (3D-специфический паттерн, найденный в боковой цепи);

СПП – Специфическая Позиция семейства/Подсемейства / Subfamily-Specific Position;

PDB – Protein Data Bank/Банк структур белков (база данных);

ARI – Adjusted Rand Index.

2. Введение

2.1. Актуальность темы исследований

Определение элементов структуры белков/ферментов (участков основной цепи, отдельных аминокислотных остатков, ориентации боковых радикалов), имеющих значение для проявления их функциональных свойств, например, каталитической активности, субстратной специфичности и других – важная задача биоинформатики. Информация о таких структурных фрагментах белковой молекулы может помочь понять, как ферменты выполняют свои функции, кроме того, это может также помочь целенаправленно выбрать позиции для мутаций, чтобы при замене, удалении или вставке аминокислотных остатков получить белок с измененными свойствами. До развития методов биоинформатики и молекулярного моделирования выбор таких участков осуществлялся в значительной степени случайно. Последовательное введение циклов случайных мутаций при проведении направленной эволюции [1] и отбор клонов, содержащих мутанты с улучшенными свойствами, показали, что таким путем можно получать продуценты белков с улучшенными свойствами, однако процесс является трудоемким и необходимо искать более рациональные пути. После определения трёхмерных кристаллических структур белков стала доступна информация о структурной организации активных центров ряда ферментов и появилась возможность сайт-направленного мутагенеза без четкого понимания роли отдельных аминокислотных остатков в механизме действия [2]. Получение, экспериментальное изучение и отбор мутантов с искомыми свойствами также было весьма трудоемким и длительным процессом. Получение информации о взаимосвязи структуры и функции белков с помощью экспериментальных методов затратно по стоимости и времени, а также требует высокого развития навыков «мокрой» биологии. В связи с этим

в последние годы все большее внимание привлекают методы компьютерной («сухой») биологии. В частности, для выявления функционально важных элементов структуры белка используются методы сравнительного биоинформатического анализа гомологичных белков. До недавних пор наиболее популярным был анализ множественных выравниваний аминокислотных последовательностей суперсемейств белков без учета структурной информации [3,4]. В то же время становится доступно все больше информации о структурной организации белков: количество белковых структур в базе данных PDB составляет сотни тысяч, активно внедряются новые методы определения структуры [5]. Наряду с этим непрерывно увеличиваются вычислительные мощности компьютеров, становится возможным проводить не только выравнивания аминокислотных последовательностей больших суперсемейств белков, но и множественные выравнивания их структур. Проводя анализ структурных выравниваний, можно выявлять функционально важные фрагменты структуры, в частности, фундаментальный и практический интерес представляют *структурные паттерны суперсемейства белков* (или просто *структурные паттерны*) – характеристическое, повторяющееся в белках суперсемейства относительное расположение элементов структуры (отдельных аминокислотных остатков, петель, фрагментов вторичной структуры и других), которое может быть ответственно за субстратную специфичность, каталитическую активность, термостабильность и другие важные свойства и функции. Такой анализ множественных выравниваний структур белков имеет преимущества перед анализом выравниваний аминокислотных последовательностей, так как структура более консервативна, чем последовательность, и те паттерны, которые могут быть утрачены при эволюции последовательности, сохраняются в структуре.

В диссертационной работе проведено исследование структурных паттернов суперсемейства белков. Предложен новый подход, позволяющий выявлять структурные паттерны суперсемейства белков, схожие внутри подсемейств белков, но различающиеся между ними и отвечающие за функциональное разнообразие белков суперсемейства. Такие паттерны мы предлагаем называть *3D-специфическими паттернами суперсемейства* или просто *3D-специфическими паттернами*. 3D-специфические паттерны могут представлять как участки основной цепи белков, так и отдельные аминокислотные остатки и ориентацию их боковых радикалов. Примеры таких 3D-специфических паттернов изображены на рисунке 1. Предварительного деления суперсемейства белков на группы белков с близкими свойствами (подсемейства) данный подход не требует и предлагает автоматическое деление, свое для каждого 3D-специфического паттерна.

Также в данной работе рассмотрены такие структурные паттерны суперсемейства, как 3D-мотивы – структурные паттерны суперсемейства белков, общие для всех белков суперсемейства и отвечающие за общность их свойств и функций. На примере 3D-мотивов дисульфидных мостиков предложен метод статистической оценки структурной гибкости основной цепи 3D-мотива, для определения возможности вставки данного 3D-мотива в структуру белка.

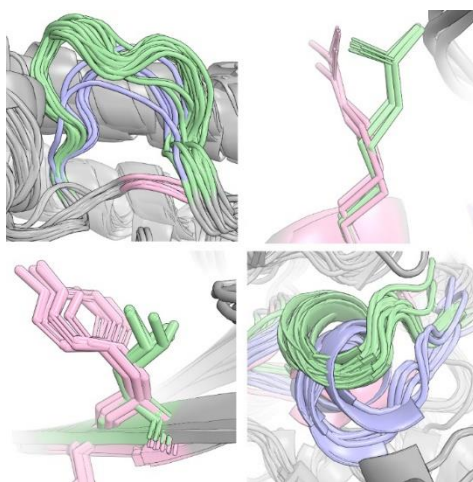


Рисунок 1. Примеры 3D-специфических паттернов.

2.2. Степень разработанности темы исследования

Для сравнительного анализа белков, входящих в состав суперсемейства, до недавних пор чаще всего использовался анализ множественных выравниваний аминокислотных последовательностей гомологичных белков. В частности, разработаны методы для выявления консервативных [6], специфических [7,8] и коррелирующих [4,9] позиций множественных выравниваний аминокислотных последовательностей. Консервативные позиции – позиции множественного выравнивания последовательностей гомологичных белков, аминокислотные остатки в которых ответственны за общность свойств и функций белков суперсемейства. Специфические позиции подсемейства/семейства (СПП) – позиции множественного выравнивания последовательностей белков, которые консервативны внутри подсемейств, но различаются между ними. Данные позиции отвечают за различие свойств и функций белков суперсемейства [3,10–13]. Коррелирующие позиции – это столбцы множественного выравнивания, вариативность аминокислотных остатков в которых взаимосвязана, то есть мутация аминокислотного остатка в одном столбце коррелирует с мутацией аминокислотного остатка в другом/других. Аминокислотные остатки, принадлежащие коррелирующим позициям, также важны для структуры и функции белка [14].

На данный момент, помимо методов, анализирующих множественные выравнивания аминокислотных последовательностей, существуют различные методы, которые помогают выполнять сравнительный анализ как структур белков в составе суперсемейства, так и различных конформаций одного белка. Например, существует класс методов, позволяющих выравнивать множество структур белков (MUSTANG [15], ParMATT [16], mTM-align [17], Matt [18], MultiProt [19], PROMALS3D [20], MAMMOTH-mult [21], Caretta [22] и другие). Полученные с помощью данных методов выравнивания белковых

структур могут использоваться как вспомогательные данные для визуального экспертного анализа, так и в качестве входных данных для других методов. Методы анализа наборов структур белков реализованы в пакетах молекулярной визуализации и анализа PyMOL [23], VMD [24], ProDy [25]. Такие программы позволяют визуализировать наборы структур белков, анализировать результаты молекулярно-динамического моделирования, считать различные метрики, в том числе расстояния и углы между атомами, среднеквадратичное отклонение (RMSD) между структурами макромолекул и их отдельными элементами. Эти методы в сочетании с визуальным экспертным анализом часто применяются для анализа конформаций одного белка, то есть альтернативных положений его структуры, для определения наиболее подвижных частей. Также существуют методы (PSSweb [26], visualCMAT [27]), которые используются для визуализации статистики, рассчитанной по множественному (структурно-опосредованному) выравниванию последовательностей и методы (2StrucCompare [28] и FATCAT [29]), которые позволяют проводить сравнительный анализ структур лишь двух гомологов.

Ни одна из приведенных выше групп методов не позволяет автоматически, без визуального экспертного анализа, выявлять элементы структур гомологичных белков, схожие внутри подсемейств и отличающихся между ними и отвечающие за функциональное разнообразие белков суперсемейства. Методы, выявляющие структурные паттерны суперсемейства белков, существуют, однако на данный момент применение информации, получаемой с их помощью, ограничено. Такие методы выявляют только консервативные структурные паттерны суперсемейства, то есть присутствующие во всех белках суперсемейства и отвечающие за общее свойство белков всего суперсемейства, так называемые 3D-мотивы [30–43]. Биоинформатический анализ гомологов, обладающих различными свойствами в пределах одного суперсемейства, до сегодняшнего времени применялся в основном на уровне аминокислотной последовательности (например, методы,

выявляющие СПП), в то время как методы, автоматически выявляющие 3D-специфические паттерны, практически отсутствуют.

2.3. Цель и задачи работы

Целью исследований была разработка нового подхода для выявления и анализа структурных паттернов суперсемейства белков. Для достижения поставленной цели были сформулированы следующие **задачи**:

1. Разработать метод выявления 3D-специфических паттернов (участков основной цепи, отдельных аминокислотных остатков, ориентации боковых радикалов) в суперсемействах белков с описанием теоретического алгоритма, представляющего последовательность шагов.
2. Разработать *S*-оценку специфичности для ранжирования выявленных в данном суперсемействе белков 3D-специфических паттернов.
3. Создать статистическую модель для отделения функционально значимых 3D-специфических паттернов от случайных колебаний белковой структуры.
4. Имплементировать разработанный метод определения 3D-специфических паттернов в виде программного кода и разработать соответствующее программное обеспечение.
5. Апробировать новый подход на широкой выборке суперсемейств белков, определить 3D-специфические паттерны и провести анализ их влияния на проявление различных функциональных свойств в гомологичных белках с использованием литературных данных.
6. Выявить 3D-мотивы дисульфидных мостиков.
7. Разработать и апробировать метод статистической оценки структурной гибкости основной цепи 3D-мотива, для определения возможности

вставки данного 3D-мотива в структуру белка на примере 3D-мотивов дисульфидных мостиков.

2.4. Объект и предмет исследования

Объектом исследования являются структурные паттерны суперсемейств белков. Предметом исследования являются 3D-специфические паттерны и 3D-мотивы.

2.5. Научная новизна

Разработан новый подход для сравнительного анализа структур гомологичных белков, обладающих различными функциональными свойствами, позволяющий определить специфические элементы структуры, называемые нами 3D-специфическими паттернами суперсемейства, которые определяют различия свойств в белках суперсемейства. Понятие 3D-специфических паттернов, а также предложенные методы их выявления и исследования являются авторскими и новыми. Предложена методология белкового дизайна в результате вставки выбранного 3D-мотива в структуру белка на примере 3D-мотивов дисульфидных мостиков, основанная на оценке гибкости основной цепи при выборе места вставки.

2.6. Теоретическая и практическая значимость работы

Выявленные с использованием разработанного метода 3D-специфические паттерны, как показали результаты исследования (см. главу 5.2), ответственны за различия в свойствах изученных нами ферментов, что помогает выявлять взаимосвязь структуры и функции рассматриваемых белков/ферментов. 3D-специфические паттерны могут быть целевыми позициями для мутаций, так как замена одного паттерна на другой в структуре белка может привести к изменению свойств. Это делает их поиск и изучение роли важной частью новых подходов к дизайну белков и биокатализаторов с улучшенными свойствами, а также поиску новых лекарств.

Разработанная методология белкового дизайна в результате вставки 3D-мотивов в структуру белка на примере 3D-мотивов дисульфидных мостиков может быть использована для получения стабилизированных препаратов белков и ферментов с измененными функциональными свойствами.

2.7. Методология и методы исследования

Для выявления и анализа структурных паттернов были разработаны методы и подходы, использующие алгоритмы машинного обучения (DBSCAN [44], OPTICS [45], HDBSCAN [46]) и методы математической статистики. Алгоритм выявления 3D-специфических паттернов был имплементирован на языке программирования Python 3 с использованием принципов объектно-ориентированного программирования (ООП). Изучаемые структуры белков были получены из базы данных PDB. Составление выборок для расчета статистики осуществляли с использованием базы данных PDBFlex [47]. Для получения множественного выравнивания структур гомологов использовали веб-сервер Mustguseal [48] и программу ParMATТ [16].

2.8. Степень достоверности

Разработанные методы выявления и анализа структурных паттернов были апробированы на конкретных примерах белков и суперсемейств белков (см. главу 5.2 и главу 5.3.4) и показали свою состоятельность. Выявленные нами 3D-специфические паттерны, как показывают опубликованные экспериментальные данные других научных групп, соответствуют важным для функций и свойств участкам структуры ферментов и отвечают 1) за различие в свойствах (таких как каталитическая активность, субстратная специфичность) между ферментами, принадлежащими различным подсемействам, 2) за различные функционально-значимые геометрические положения участка структуры фермента (см. главу 5.2). В методике

исследования были использованы апробированные и широко используемые алгоритмы машинного обучения и приемы математической статистики. Литературный обзор и обсуждение результатов основаны на анализе всей доступной литературы по теме. Результаты диссертационного исследования опубликованы в рецензируемых научных журналах и обсуждены на профильных научных конференциях.

2.9. Личный вклад автора

Личный вклад автора заключается в: 1) анализе литературных источников; 2) разработке новых методов выявления и анализа структурных паттернов; 3) имплементации разработанных методов в качестве программного кода; 4) апробации разработанных методов; 5) анализе полученных результатов; 6) подготовке научных статей и представлении результатов на научных конференциях.

2.10. Положения, выносимые на защиту

- Разработан новый метод и соответствующее программное обеспечение для сравнительного анализа структур белков суперсемейства, основанный на выявлении 3D-специфических паттернов - элементов структуры белков/ферментов (участков основной цепи, отдельных аминокислотных остатков, ориентации боковых радикалов), которые схожи внутри подсемейств белков, но различаются между ними и позволяют разделить суперсемейства на функционально обособленные подсемейства.
- Разработана S-оценка специфичности и статистическая модель для ранжирования выявленных 3D-специфических паттернов, а также отделения функционально-значимых 3D-специфических паттернов от результатов теплового колебания структуры белка.
- Предположено и при анализе литературных данных о функциональных свойствах изученных ферментов показано, что 3D-специфические

паттерны представляют важные для механизма действия элементы структуры ферментов и отвечают за различие свойств (таких как субстратная специфичность, каталитическая активность) ферментов, принадлежащих к различным функциональным подсемействам, а также конформеров одного фермента благодаря пространственной ориентации ключевых аминокислотных остатков и участков основной цепи.

- Предложена методология белкового дизайна в результате вставки 3D-мотивов в структуру белка на примере 3D-мотивов дисульфидных мостиков с целью получения стабилизированных препаратов белков и ферментов с измененными функциональными свойствами.

2.11. Публикации по теме работы

По материалам работы опубликованы 4 статьи в рецензируемых журналах, индексируемых в наукометрических базах данных Web of Science и/или Scopus (3 статьи в международных журналах и 1 статья в российском журнале из списка ВАК)¹:

- **Timonina D.**, Sharapova Y., Švedas V., Suplatov D. Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and conformational plasticity in protein superfamilies // *Computational and Structural Biotechnology Journal*. – 2021. – Т. 19. – С. 1302-1311 (0.63/0.45).
- **Тимонина Д.С.**, Суплатов Д.А. Анализ множественных выравниваний белков с использованием 3D-структурной информации по ориентации боковых цепей аминокислот // *Молекулярная биология*. – 2022. – Т. 56. – №. 4. – С. 663–670 (0.38/0.3).

¹ В скобках приведен объем публикации в печатных листах и вклад автора в печатных листах

- Suplatov D., **Timonina D.**, Sharapova Y., Švedas V. Yosshi: a web-server for disulfide engineering by bioinformatic analysis of diverse protein families // *Nucleic acids research*. – 2019. – Т. 47. – №. W1. – С. W308-W314 (0.44/0.2).
- Suplatov D., Sharapova Y., **Timonina D.**, Kopylov K., Švedas V. The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families // *Journal of Bioinformatics and Computational Biology*. – 2018. – Т. 16. – №. 02. – С. 1840005 (0.94/0.1).

2.12. Апробация работы

Результаты исследования были представлены на 5-и конференциях: «Moscow Conference on Computational Molecular Biology» (МССМВ'19 и МССМВ'21, Москва, Россия, 2019 и 2021 гг.), Международных научных конференциях студентов, аспирантов и молодых ученых «Ломоносов-2019» и «Ломоносов-2021» (Москва, Россия, 2019 и 2021 гг.), The 44th FEBS Congress (Краков, Польша, 2019).

2.13. Структура и объем диссертации

Диссертационная работа состоит из следующих разделов: оглавление, список сокращений, введение, обзор литературы, методы, результаты и обсуждение, заключение, основные результаты и выводы, список литературы. Работа изложена на 155 страницах, содержит 54 иллюстрации, 7 таблиц и цитирует 156 литературных источников.

3. Обзор литературы

3.1. Понятия суперсемейства, семейства, подсемейства

Белки – это линейные полимеры, состоящие из аминокислотных остатков, имеющих разные физико-химические свойства. Функция и свойства белка полностью определяются его последовательностью и структурой. В ходе эволюционного развития белков от общего предка, в результате изменения последовательности и структуры белка, некоторые свойства белков (общая укладка структуры, механизм реакции) могут сохраняться, в то время как другие (например, субстратная специфичность, каталитическая активность) могут изменяться, что приводит к функциональному разнообразию гомологичных белков. Чем более удалены белки друг от друга эволюционно, тем сильнее они различаются по последовательности и структуре, а следовательно, и функционально. Причем структура изменяется медленнее, чем последовательность, то есть более консервативна. В соответствии со степенью эволюционного родства белки объединяют в группы различного размера [49]:

- *Суперсемейство* – это множество белков с возможно небольшим сходством последовательности, но чья структура, функции и свойства предполагают наличие общего предка. Предполагается, что ферменты одного суперсемейства могут быть удалены от общего предка и иметь как различный тип катализируемой химической реакции, так и различную субстратную специфичность.
- *Семейство* – это множество белков, более близких эволюционно по сравнению с суперсемейством и имеющих значительное сходство последовательности, обычно имеющих общий механизм реакции, но различную субстратную специфичность. Белки, входящие в одно семейство, обычно имеют меньшее функциональное разнообразие, нежели белки, входящие в одно суперсемейство.

- *Подсемейство* – множество эволюционно близких белков, близких по последовательности, структуре и функции.

Деление суперсемейства на семейства и подсемейства субъективно и зависит от нашего понимания построения иерархии эволюционно родственных белков в зависимости от свойств и функций.

Сравнительный анализ последовательностей и структур белков суперсемейства помогает выявлять элементы последовательности и структуры белка, ответственных за ту или иную функцию или свойство.

3.2. Современный биоинформатический анализ суперсемейств белков

3.2.1. Анализ суперсемейств белков на уровне аминокислотных последовательностей

Экспериментальное выявление элементов белка (отдельных аминокислот, участков основной цепи, участков боковой цепи), имеющих значение для его свойств и функции, связано со значительными затратами времени и ресурсов, также требует высоких навыков экспериментальной работы. Поэтому для выявления таких функционально важных элементов белка используются биоинформатические методы, позволяющие сравнивать гомологичные белки, в том числе используется анализ множественных выравниваний аминокислотных последовательностей белков суперсемейства. В частности, функционально важными аминокислотными остатками могут быть консервативные [6], специфические [7,8] и коррелирующие [4,9] позиции множественного выравнивания суперсемейства, речь о которых идет в последующих главах 3.2.1.1, 3.2.1.2 и 3.2.1.3. У консервативных позиций суперсемейств белков есть трехмерный аналог – 3D-мотивы, речь о которых идет в главе 3.2.2.4. А специфические позиции подсемейств белков послужили прототипом 3D-специфических паттернов – понятия, которое вводится и является ключевым в этой работе.

3.2.1.1. Консервативные позиции суперсемейства белков

Консервативные позиции – это такие столбцы множественного выравнивания последовательностей белков суперсемейства, которые содержат аминокислотные остатки одного типа или близкие по свойствам различные аминокислотные остатки, и часто являются важными для функции и структуры белков всего суперсемейства. Как правило, такие позиции встречаются в функциональных сайтах белков [50,51]. Выявление консервативных позиций в множественном выравнивании суперсемейства белков – сложная биоинформатическая задача, которая решается различными подходами, и на сегодняшний день существует много оценок консервативности позиции [6].

Формула для вычисления консервативности колонки множественного выравнивания аминокислотных последовательностей белков, предложенная Сандером и Шнайдером, основана на энтропии Шеннона [52]:

$$V_{Schneider} = - \sum_{i=1}^K p_i \ln p_i \times \frac{1}{\ln K},$$

где $K = 20$ — количество канонических аминокислот, N – количество белков в выравнивании, $p_i = \frac{n_i}{N}$ – частота i -ой аминокислоты в колонке. $V_{Schneider}$ принимает значения в диапазоне от 0 до 1, причем 0 принимает в случае, если колонка выравнивания содержит только идентичные аминокислоты, а значение 1 в случае, когда в колонке присутствуют все типы аминокислотных остатков в одинаковых пропорциях.

Следующий шаг в развитии методов подсчета меры консервативности колонки множественного выравнивания – оценка, разработанная Карлин и Брочери (Karlin и Broschieri), которая учитывает химическую вариабельность аминокислотных остатков в колонке множественного выравнивания:

$$V_{Karlin} = \sum_{i=1}^N \sum_{j>i}^N M(S_i(x)S_j(x)) \times \frac{2}{N(N-1)},$$

где $M(a, b) = \frac{m(a,b)}{\sqrt{m(a,a)m(b,b)}}$ – мера схожести аминокислотных остатков, $m(a,b)$ — соответствующее значение матрицы аминокислотных замен BLOSUM62 [53] для аминокислотных остатков a и b , $S_i(x)$ – аминокислота в последовательности под номером i в колонке x , N – количество последовательностей в выравнивании. Этот подход имеет ряд недостатков, в том числе, отсутствие учета делеций, попытки исправить которые были предложены в работах [54,55].

Более новые оценки консервативности учитывают вес последовательности в выравнивании. Например, оценка веса последовательности в выравнивании, предложенная в работе [56], выглядит следующим образом:

$$w_i = \frac{1}{N-1} \sum_{j \neq i}^N d(s_i, s_j),$$

где d – расстояние между последовательностями, например, процент различающихся аминокислот, N – количество последовательностей в выравнивании. Существуют и другие способы оценки веса последовательности [57,58].

Оценки, использующие вес последовательности, дают возможность уменьшить влияние однотипных последовательностей при расчете консервативности. Примером такой оценки может быть измененная формула Сандера и Шнайдера [52]:

$$C_{Sander} = \lambda \sum_{i=1}^N \sum_{j>i}^N d(i, j) m(s_i, s_j),$$

где λ – нормирующий коэффициент, $d(i, j)$ – расстояние между последовательностями, равное 100% минус процент аминокислотной идентичности между последовательностями, $m(s_i, s_j)$ — соответствующее

значение матрицы аминокислотных замен BLOSUM62 [53] для аминокислотных остатков S_i и S_j .

3.2.1.2. Специфические позиции подсемейства белков/позиции, определяющие специфичность

Если консервативные позиции – это такие позиции множественного выравнивания последовательностей суперсемейства белков, аминокислотные остатки в которых отвечают за общие свойства всех белков суперсемейства, то специфические позиции подсемейства/семейства (СПП), также известные как позиции, определяющие специфичность – это позиции множественного выравнивания последовательностей суперсемейства белков, консервативные внутри подсемейств, но различающиеся между ними (см. рисунок 2).

Понятие СПП было введено в конце 1990-х [7,8]. Такие позиции являются, как правило, детерминантами функционального разнообразия, поэтому могут помочь понять, как ферменты выполняют свои функции, а также могут быть выбраны в процессе разработки лекарств в качестве точек мутаций для экспериментов по белковой инженерии как ключевые остатки, участвующие в селективном связывании лигандов [3,10]. Именно СПП послужили прототипом 3D-специфических паттернов, представленных в этой работе.

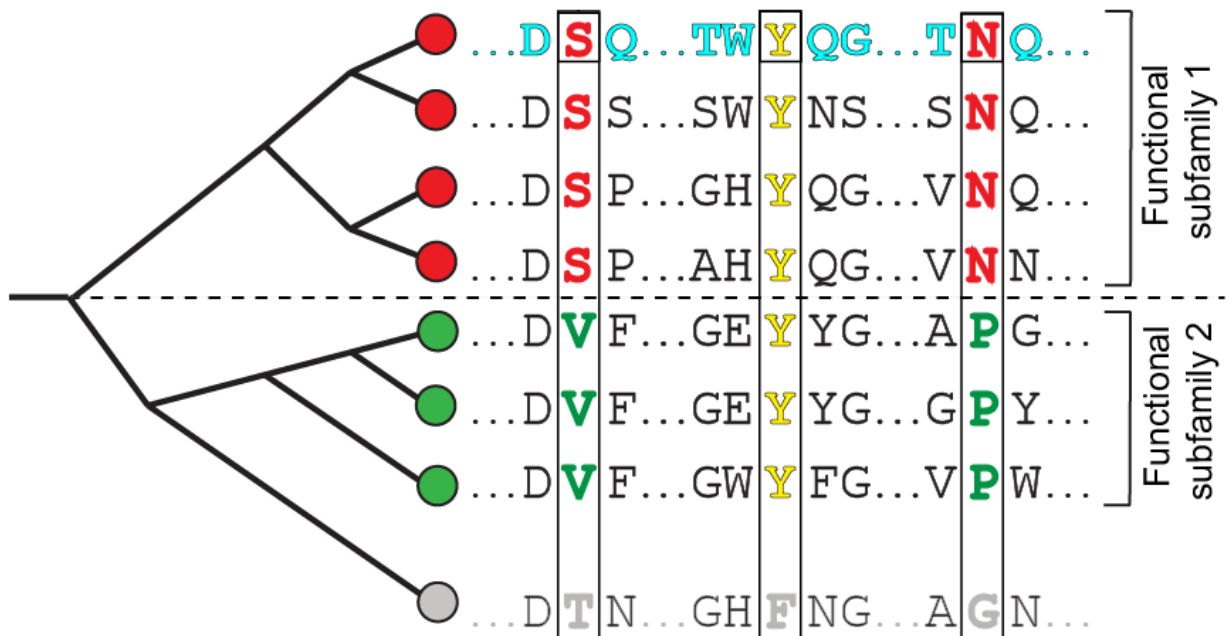


Рисунок 2. Множественное выравнивание аминокислотных последовательностей суперсемейства белков. Примеры консервативной и специфических позиций выделены цветом. Желтым цветом на рисунке выделена консервативная позиция суперсемейства белков, сочетанием красного и зеленого цветов выделены специфические позиции. Рисунок взят из [59].

Существует множество различных методов для выявления таких специфических позиций подсемейства. Одним из самых старых методов выявления специфических позиций является метод эволюционного следа, который был опубликован в работах [60–63]. На первом этапе данного алгоритма по множественному выравниванию последовательностей белков строится филогенетическое дерево. Далее белки, в соответствии с получившимся деревом, разбиваются на группы с разным уровнем сходства. На следующем этапе выбираются позиции, консервативные внутри групп и различающиеся между ними. Развитие этот метод получил в работе [64]. В работах [65–67] было показано, что СПП, выявленные таким образом, являются функционально значимыми остатками для исследованных суперсемейств белков.

Следующая группа подходов поиска СПП основывается на предположении, что аминокислоты, консервативные среди ферментов, входящих в одно подсемейство, и различающиеся между ферментами,

входящими в различные подсемейства, скорее всего, важны для специфического распознавания субстрата этих ферментов. В таких работах ищутся позиции в выравнивании, вариабельность аминокислотных остатков в которых коррелирует с разбиением на подсемейства. В работе [68] в качестве меры корреляции использована относительная энтропия позиции i относительно подсемейства s :

$$RE_i^s = \sum_x P_{i,x}^s \log \frac{P_{i,x}^s}{P_{i,x}^{\bar{s}}},$$

где $P_{i,x}^s$ и $P_{i,x}^{\bar{s}}$ — профильное значение (пропорционально частоте) аминокислоты x в позиции i в подвыравниваниях, соответствующих подсемействам s и не- s , соответственно. Значимость данной позиции выравнивания определяется как сумма RE_i^s по всем подсемействам. Этот метод успешно предсказывает детерминанты специфичности [68]. В работе [13] в качестве меры корреляции используется *взаимная информация*:

$$MI_i = \sum_{\substack{x=1,\dots,20 \\ y=1,\dots,Y}} f_i(x,y) \log \frac{f_i(x,y)}{f_i(x)f(y)},$$

где i — позиция выравнивания, Y — количество подсемейств, $f_i(x,y)$ — частота аминокислоты x в подсемействе y , $f_i(x)$ — частота аминокислоты x во всем выравнивании, $f(y)$ — доля белков в подсемействе y . В работах [69] и [70] показана эффективность данного подхода для определения функционально важных аминокислотных остатков белков суперсемейств. В работе [11] было показано, что эти две меры корреляции на реальных выравниваниях дают практически одинаковое ранжирование позиций.

Работа [13] получила свое продолжение в методе [12], где вместо $f(x,y)$, используется следующий аналог $\tilde{f}(x,y)$, учитывающий физико-химические свойства аминокислотных остатков (аминокислотные остатки с близкими физико-химическими свойствами считаются близкими):

$$\tilde{f}(x, y) = \frac{n(x, y) + \kappa(\sum_{z=1}^{20} n(z, y)m(z \rightarrow x))/\sqrt{n(y)}}{n(y) + \kappa\sqrt{n(y)}}$$

где $m(z \rightarrow x)$ – вероятность замены аминокислоты z на x , $n(y)$ – число белков в подсемействе, $n(z, y)$ – число появлений аминокислоты z в подсемействе y .

Еще один метод выявления СПП, показавший себя конкурентноспособным с предыдущими, представлен в работе GroupSim [71]. Подход заключается в следующем: сначала ищутся те колонки множественного выравнивания, аминокислотные остатки в которых расположены недалеко от лигандов и каталитических сайтов. Далее из найденных колонок исключаются колонки, которые:

- являются консервативными,
- не являются консервативными, но имеют сходное распределение аминокислот внутри подсемейств,
- если колонка не является консервативной ни в одном подсемействе.

Получившееся множество колонок являются претендентами на СПП.

Существует ряд инструментов, которые могут находить СПП и не требуют предопределения подсемейств [72–74]. Например, алгоритм Zebra [73] сначала автоматически разбивает белки по подсемействам на основе анализа аминокислотных последовательностей. Далее СПП ищутся на основе как данных о структуре, так и физико-химических свойств остатков, консервативных в подсемействах белков. Функция оценки специфичности в этой работе вводится следующим образом:

$$S_i = \frac{[\sum_G \sum_{AB} M(A, B) \times q_i(AB, G)] \times [\sum_G \sum_A q_i(A, G) \times \log \frac{q_i(A, G)}{q_i(A)}]}{n_G \sum_G \log \frac{N}{N_G}},$$

где A и B обозначают тип аминокислоты в колонке множественного выравнивания (в том числе и делеция); $q_i(AB)$ обозначает частоту пары AB в

колонке i , которая рассчитывается как число пар AB , разделенное на общее количество пар к колонке i ; $q_i(AB, G)$ – частота пары AB в подсемействе G колонки i ; $q_i(A)$ и $q_i(A, G)$ – частоты аминокислоты типа A соответственно в колонке i и в подсемействе G этой колонки; n_G обозначает общее количество подсемейств; N_G – число белков в подсемействе; значение $M(AB)$ соответствует оценке взаимозаменяемости аминокислот типов A и B . В работе [75] показана эффективность данного подхода для определения функционально важных аминокислотных остатков белков суперсемейств.

3.2.1.3. Коррелирующие позиции суперсемейства белков

Коррелирующие позиции суперсемейства белков (коррелирующие аминокислотные остатки) – это такие столбцы множественного выравнивания последовательностей, вариативность аминокислотных остатков в которых взаимосвязана, то есть мутация аминокислотного остатка в одном столбце коррелирует с мутацией аминокислотного остатка в другом/других [4,9] (см. рисунок 3).

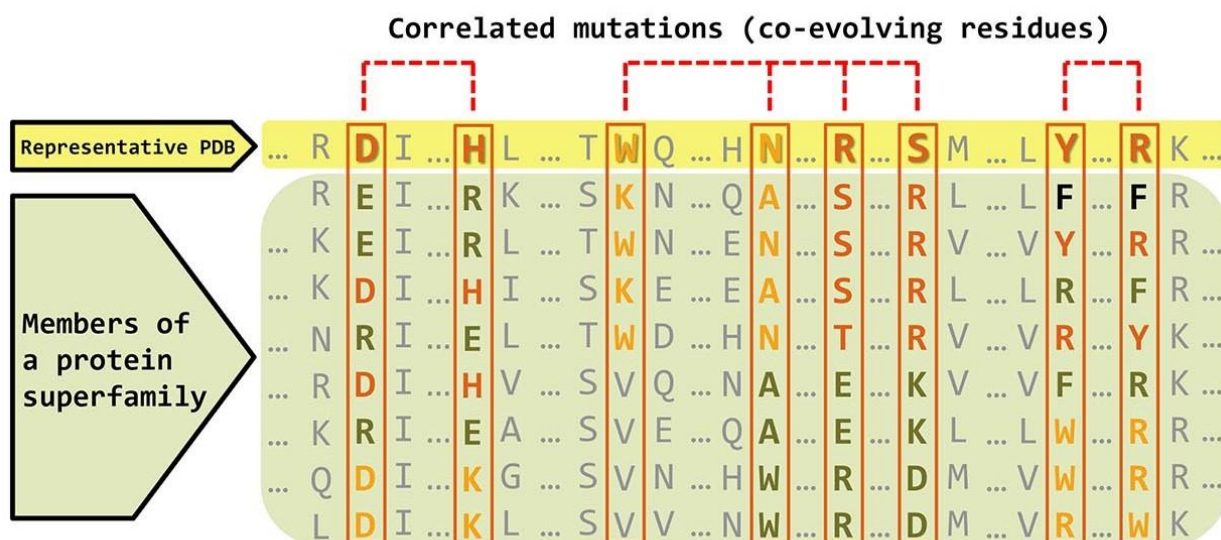


Рисунок 3. Множественное выравнивание аминокислотных последовательностей суперсемейства белков. Цветом выделены коррелирующие позиции. Рисунок взят из [76].

Коррелирующие позиции могут возникать по нескольким причинам, перечисленным ниже [14]:

- Структурные и функциональные причины. То есть корреляция может возникать из-за одних и тех же сил естественного отбора, действующего на обе позиции, чтобы сохранить структуру и функцию белкового домена [77]. Такие коррелирующие позиции представляют наибольший интерес. Их выявление может помочь, например, в предсказании структуры белка по последовательности [78,79] и может быть использована для функциональной аннотации межсубъединичных интерфейсов [80,81].
- Филогенетические причины. То есть корреляция может возникать из-за наличия у белков общего предка и никак не связана со структурой и функцией белка. Пример возникновения таких коррелирующих позиций представлен на рисунке 4.
- Стохастические причины. То есть корреляция может возникать случайно.

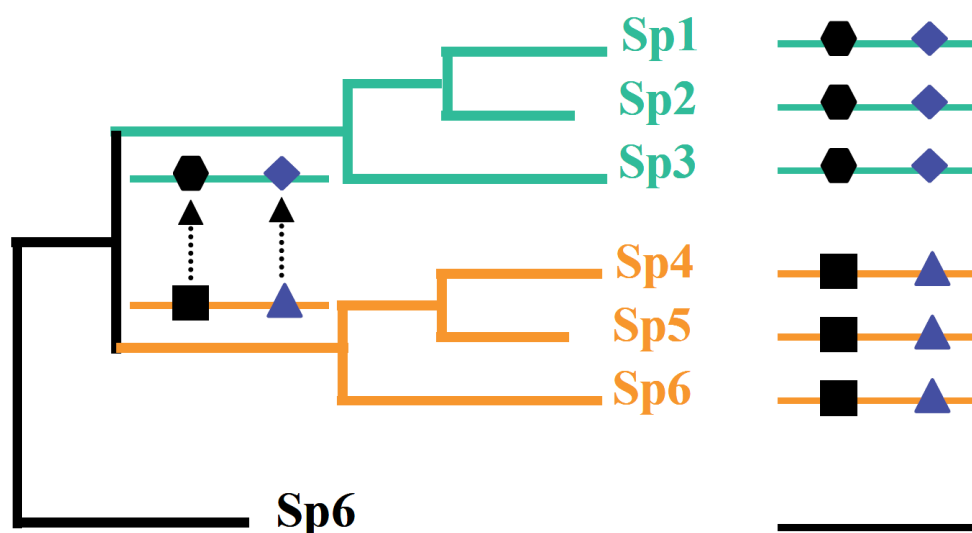


Рисунок 4. Пример возникновения коррелирующих позиций из-за филогенетических причин. Рисунок взят из [14].

Существует много методов выявления коррелирующих позиций [9,82–88]. Одной из известных метрик для выявления коррелирующих позиций в множественном выравнивании аминокислотных последовательностей суперсемейства является взаимная информация (MI). Для двух колонок i, j множественного выравнивания последовательностей белков взаимная информация рассчитывается по формуле [89]:

$$MI(i, j) = \sum_x \sum_y P(x_i, y_j) \log \left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right),$$

$$P(x_i) = \sum_y P(x_i, y_j),$$

где $P(x_i, y_j)$ – это вероятность того, что аминокислота x находится в колонке i и аминокислота y находится в колонке j .

Для исключения из результатов тех коррелирующих остатков, которые коррелируют из-за филогенетических или стохастических причин, в работе [82] была предложена измененная формула взаимной информации MI_C , которая использует следующий факт: пары коррелирующих позиций, которые коррелирует из-за филогенетических или стохастических причин, как правило, коррелируют не только друг с другом, но и с другими позициями. Метод [82] основывается на следующем предположении: если две пары позиций множественного выравнивания последовательностей имеют большое значение MI и они имеют схожие паттерны корреляции с другими позициями в выравнивания, то высокое значение MI обусловлено не структурными или функциональными, а филогенетическими причинами. Сходство паттернов корреляции позиций i, j вычисляется по формуле:

$$CPS(i, j) = \frac{1}{n-2} \sum_{m \neq i, j} MI(i, m)MI(j, m).$$

После нормализации получаем:

$$NCPS(i, j) = \frac{CPS(i, j)}{\sqrt{\frac{1}{n(n-1)} \sum_{i, j} CPS(i, j)}}$$

В итоге в работе [82] оценка корреляции позиций множественного выравнивания рассчитывается по формуле:

$$MI_C(i, j) = MI(i, j) - NCPS(i, j).$$

В работе [83] используется схожая оценка MI_P :

$$MI_P(i, j) = MI(i, j) - APS(i, j),$$

$$APC(i, j) = \frac{MI(i, \bar{x})MI(j, \bar{x})}{\overline{MI}},$$

где $MI(i, \bar{x}) = \frac{1}{n-1} \sum MI(i, x)$ – среднее значение взаимной информации колонки i , $\overline{MI} = \frac{2}{n(n-1)} \sum MI(i, j)$ – среднее значение взаимной информации всего выравнивания, $APC(i, j)$ – оценка взаимной информации двух позиций выравнивания, связанной с филогенетическими и стохастическими причинами. Оценка MI_P основывается на предположении, что пары позиций, коррелирующих из-за структурных и функциональных причин, редки и значение MI пары позиций, коррелирующих из-за функциональных и структурных причин, превосходит значение MI со всеми другими позициями, с которыми они коррелируют из-за филогенетических или стохастических причин.

Веб-сервер visualCMAT [27] – веб-сервер, предназначенный для визуализации коррелирующих позиций. VisualCMAT позволяет визуализировать в программе PyMOL [23] найденные с помощью оценок MI_C и MI_P коррелирующие позиции в структуре референсного белка. Пользователь сам может выбрать, какую оценку (MI_C или MI_P) использовать. По значениям MI_C и MI_P для данной пары позиций вычисляются Z -оценки:

$$Z_C(i, j) = \frac{MI_C(i, j) - \mu_C}{\sigma_C},$$

$$Z_P(i, j) = \frac{MI_P(i, j) - \mu_P}{\sigma_P},$$

где $\mu_C(\mu_P)$ – выборочное среднее, $\sigma_C(\sigma_P)$ – выборочное стандартное отклонение. По умолчанию пары позиций с $Z_C(i, j) < 3,5$ или $Z_P(i, j) < 3,5$ далее не рассматриваются, остальные считаются коррелирующими. На вход веб-серверу подается множественное выравнивание аминокислотных последовательностей суперсемейства белков и PDB-файл со структурой референсного белка из этого суперсемейства.

Результат работы веб-сервера – файл rymol-сессии, содержащий несколько «слоев»:

- Первый слой содержит структуру референсного белка, основная цепь которого окрашена градиентом (от серого к красному) в зависимости от значения максимальной Z -оценки для данного аминокислотного остатка.
- Второй слой содержит структуру референсного белка, коррелирующие аминокислотные остатки соединены пунктирными линиями. Пунктирные линии окрашены разными цветами в соответствии с величиной Z -оценки и расстояния между аминокислотными остатками.
- Третий слой содержит структуру референсного белка. Каждый $C\alpha$ -атом референсного белка имеет радиус, пропорциональный сумме всех Z -оценок, соответствующих всем парам коррелирующих позиций, в которые включен данный аминокислотный остаток.
- Четвертый слой содержит структуру референсного белка, на которой отмечены все сайты связывания, найденные с помощью алгоритма `frocket` [90]. Найденные сайты связывания делятся на три группы. Сайт связывания относится к одной из трех групп в зависимости от того, содержит ли он коррелирующие позиции и на каком расстоянии друг от друга найденные коррелирующие позиции находятся. Найденные сайты связывания ранжируются в соответствии со значением суммы Z -оценок коррелирующих позиций, входящих в них.

На первом и третьем слоях рассматриваются Z -оценки только близко расположенных аминокислотных остатков (расстояние не больше 5 Å).

3.2.2. Инструменты для анализа и сравнения гомологичных белков, использующие структурные данные

3.2.2.1. Построение множественного структурного выравнивания гомологичных белков

Сравнительный биоинформатический анализ гомологичных белков является важным шагом при изучении их структуры и функции. Как было показано в предыдущей главе, для этих целей может быть использовано множественное выравнивание последовательностей (например, оно

используется для нахождения консервативных и специфических позиций как функционально важных остатков). Но так как сходство белков по последовательности в результате эволюции может быть утрачено, сравнительный биоинформатический анализ гомологичных белков с использованием множественного выравнивания последовательностей может быть практически невозможен. В связи с этим необходимо анализировать множественные *структурные* выравнивания суперсемейств. Задача множественного структурного выравнивания заключается в пространственном наложении структур с минимизацией метрики качества выравнивания, например, такой как среднеквадратичное отклонение RMSD (см. рисунок 5). Но даже когда существует соглашение о том, какую метрику нужно оптимизировать, множественное выравнивание структур белков, поиск оптимального выравнивания является сложной задачей с вычислительной точки зрения. Существует много алгоритмов построения множественного структурного выравнивания: MUSTANG [15], ParMATT [16], mTM-align [17], Matt [18], MultiProt [19], PROMALS3D [20], MAMMOTH-mult [21], Caretta [22] и другие.



Рисунок 5. Пример множественного структурного выравнивания суперсемейства белков, полученного с помощью программы ParMATT. Рисунок взят из [91].

Matt [18] – один из самых известных алгоритмов построения множественного структурного выравнивания. Идея алгоритма заключается в следующем. Пусть количество структур белков, предназначенных для выравнивания, равно N . На первом этапе алгоритма количество групп уже выровненных белков равняется количеству белков, то есть N . Далее итеративно (количество итераций $N-1$) две группы выровненных структур объединяются в одну выровненную группу, уменьшая количество групп на одну. Итерации заканчиваются, когда остается всего одна группа. Для простоты объясним принцип выравнивания двух групп на примере выравнивания двух структур (то есть, когда каждая группа состоит из одной структуры). Сначала Matt рассматривает фрагменты из 5–9 соседних аминокислотных остатков. Пары фрагментов рассматриваются одинаковой длины, по одному от каждой структуры. Для каждой пары фрагментов рассчитывается P -оценка на основе минимального значения RMSD, достижимого путем трехмерного выравнивания одного фрагмента с другим (рассматриваются только C_α атомы). Далее Matt, сохраняя последовательность следования аминокислотных остатков вдоль основной цепи, собирает выровненные фрагменты в единое выравнивание двух структур: используя динамическое программирование, Matt создает все более и более длинные наборы выровненных фрагментов. Matt принимает решение, следует ли объединять два набора выровненных фрагментов вместе, на основе оценки, равной сумме оценок выравниваний отдельных выровненных фрагментов и штрафа. На каждом шаге итерации Matt объединяет те две группы структур, выравнивание которых имеет наилучшую оценку. Если остается всего одна группа – Matt переходит к финальному этапу построения множественного структурного выравнивания, оптимизирующему RMSD полученного выравнивания. Результатом работы программы Matt является как множественное структурное выравнивание белков, так и структурно-опосредованное выравнивание аминокислотных последовательностей.

Алгоритм ParMATT [16] представляет параллельную реализацию алгоритма Matt. ParMATT может работать быстрее, чем Matt на одном многоядерном процессоре, и обеспечивает значительное ускорение при выполнении программы на системах с распределенной памятью, то есть вычислительных кластерах и суперкомпьютерах, на которых размещены независимые по памяти вычислительные узлы. Наиболее требовательные к вычислительным ресурсам этапы алгоритма Matt – начальное построение попарных выравниваний между всеми входными структурами и дальнейшее итеративное выполнение множественного выравнивания – были распараллелены с использованием MPI и pthreads, а завершающий этап алгоритма был оптимизирован за счет OpenMP. ParMATT может значительно ускорить трудоемкий процесс построения множественного структурного выравнивания большого набора гомологичных белков.

Еще один алгоритм множественного структурного выравнивания гомологичных белков – mTM-align [17]. Для многих метрик этот алгоритм показывает результаты лучше, чем алгоритм Matt на наборах данных NOMSTRAD [92], SABmark_sup, SABmark_tw1 [93] и SISY-multiple [94]. Алгоритм состоит в следующем. На первом этапе с помощью алгоритма TM-align [95] строятся парные структурные выравнивания всех белков. Если белков N , то количество выравниваний $\frac{N(N-1)}{2}$. Для этого находится максимум следующей оценки:

$$TM-score = \max \frac{1}{L} \sum_{i=1}^N \frac{1}{1+(d_i/d_0)^2},$$

где d_i — расстояние между i -й парой атомов C_α двух структур, L — длина одного из белков, N — количество пар выравненных аминокислотных остатков, d_0 — числовой коэффициент. Далее с помощью алгоритма UPGMA строится филогенетическое дерево. Матрица расстояний для построения филогенетического дерева рассчитывается с использованием получившихся $TM-score$. Далее множественное структурное выравнивание строится согласно

этому филогенетическому дереву, то есть выравнивания сливаются от листьев дерева к корню. Это слияние, то есть выравнивание двух уже готовых выравниваний, строится с помощью алгоритма Нидлмана-Вунша [96]. Схема работы алгоритма представлена на рисунке 6.

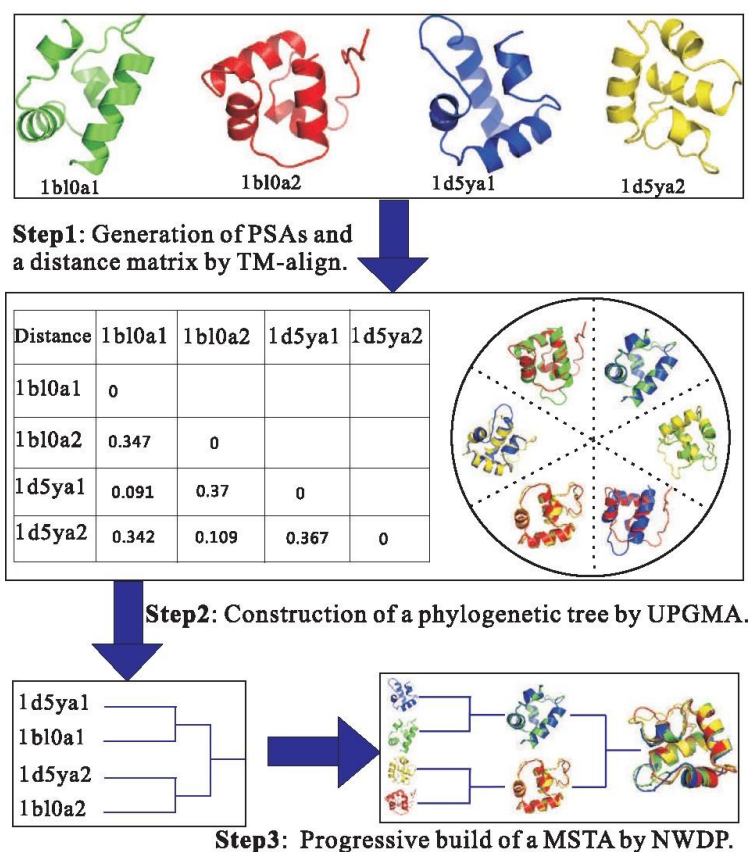


Рисунок 6. Схема работы алгоритма множественного структурного выравнивания mTM-align [17]. На первом этапе с помощью алгоритма TM-align строятся парные структурные выравнивания всех белков. Далее с помощью алгоритма UPGMA строится филогенетическое дерево с использованием получившихся TM-score. Далее множественное структурное выравнивание строится согласно этому филогенетическому дереву, то есть с помощью алгоритма Нидлмана-Вунша [96] выравнивания сливаются от листьев дерева к корню. Рисунок взят из [17].

Веб-сервер Mustguseal [48] – имплементированный в качестве веб-сервера алгоритм, позволяющий создавать большие выравнивания функционально разнообразных суперсемейств белков. В отличие от предыдущих алгоритмов, Mustguseal не требует в качестве входных данных все структуры белков интересующего пользователя суперсемейства. Для работы алгоритма достаточно одной структуры интересующего пользователя

белка – далее веб-сервер сам находит близкие по последовательности и структуре белки и выравнивает их.

Алгоритм веб-сервера Mustguseal состоит в следующем: на первом этапе с помощью алгоритма SSM [97] по сходству структур находятся эволюционно далекие родственники (сначала находятся близкие структуры, а потом из множества найденных белков удаляются близкие по последовательности). Ожидается, что найденные белки принадлежат разным семействам. Выравнивание найденных структур белков строится с помощью программы Matt. Далее для каждого найденного белка с помощью алгоритма поиска по последовательности BLAST [98,99] ищутся эволюционно близкие родственники – члены соответствующего семейства. Затем для каждого семейства с помощью метода MAFFT [100] строится множественное выравнивание найденных аминокислотных последовательностей. На последнем этапе на основе информации, полученной из выравнивания структур, строится общее выравнивание всех найденных аминокислотных последовательностей. Результатом работы метода является как данное выравнивание аминокислотных последовательностей, так и выравнивание найденных на первом этапе алгоритма структур. Схема работы алгоритма представлена на рисунке 7.

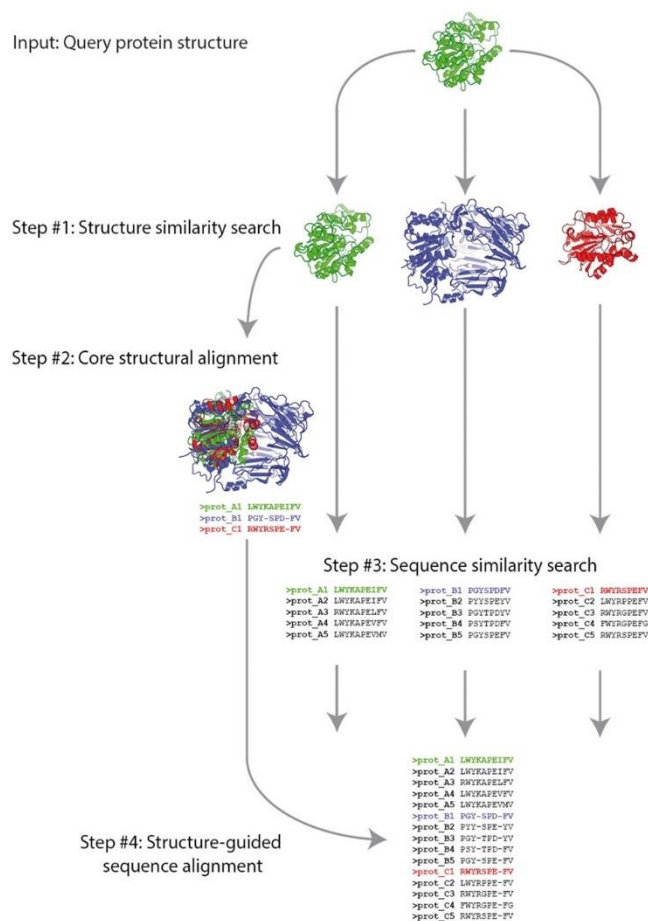


Рисунок 7. Схема работы веб-сервера Mustguseal [48]. См. пояснение в тексте. Рисунок взят из [48].

3.2.2.2. Пакеты молекулярной визуализации и анализа

Пакеты молекулярной визуализации и анализа – это такие программы, которые позволяют визуализировать интересующие пользователя молекулы и наборы молекул в удобном виде, в том числе позволяют визуализировать множественные структурные выравнивания белков, полученные с помощью методов, описанных в предыдущей главе. Такие программы в сочетании с визуальным экспертным анализом обычно используются для глобального сравнения укладок, анализа молекулярной динамики, нахождения наиболее подвижных элементов структуры белка. Пакеты молекулярной визуализации и анализа могут быть использованы для анализа суперсемейств белков и выявления функционально важных элементов структур, отвечающих за сходство и различие функций белков суперсемейства. Но, что вызывает трудность, выявление потенциально важных для функции участков структур

никак не автоматизировано в этих пакетах и полностью ложится на визуальный экспертный анализ.

Самая известная и широко используемая программа для визуального отображения белков с целью их дальнейшего экспертного анализа – PyMOL [23]. PyMOL – это пакет для молекулярной визуализации с открытым исходным кодом. Программа PyMOL представляет широкие возможности трехмерного изображения статичных структур как биополимеров (белков, нуклеиновых кислот), так и малых молекул. В одной сессии возможно одновременное отображение сотен различных структур, которые возможно накладывать друг на друга, считать RMSD между любыми двумя группами атомов, выполнять выравнивание структур как попарно, так и многих к одной. PyMOL предоставляет широкий спектр возможностей для визуализации: предлагает различные режимы отображения молекул (палочки, шарики, вторичная структура (Cartoon), поверхность), позволяет выделять и окрашивать нужным цветом интересующий пользователя набор атомов. Также имеются возможности редактирования: замена аминокислотных остатков белка, добавление, перемещение и удаление атомов и распространенных функциональных групп, таких как -ОН, -NH₂, -COOH, -C₆H₅ и т.п. Поскольку ядро PyMOL написано на языке Python, очень просто осуществляется автоматизация пакетной обработки молекул – возможен импорт пакета pymol в скрипт Python и вызов из него API-функций, которые подробно документированы на официальной PyMOLWiki [101]. Для построения изображений публикационного качества имеется встроенный движок рендеринга с использованием трассировки лучей.

Еще один популярный пакет молекулярной визуализации с открытым исходным кодом – VMD [24]. VMD предоставляет схожие с PyMOL возможности, но в отличие от PyMOL, эта программа в первую очередь предназначена для просмотра и анализа траекторий молекулярной динамики белка, а также подготовки структур для нее. Программа имеет очень широкий

спектр поддержки форматов файлов молекулярных структур: стандартные форматы файлов структур белка (PDB, mmCIF), форматы траекторий всех применяемых в научной среде пакетов молекулярной динамики (Amber, CHARMM, NAMD, Gromacs и др.), выходные файлы программ квантовой химии (Gamess, Gaussian, VASP и т.д.). С точки зрения функции анализа, VMD также обладает широкими возможностями, в частности, позволяет строить графики значений различных параметров молекулы вдоль траектории молекулярной динамики: расстояний и углов между атомами, RMSD между частью структуры молекулы на одном референсном кадре молекулярной динамики и остальными; подсчет динамически образуемых химических связей, водородных связей, солевых мостиков; автоматизированный анализ структур и траекторий из окна программы с помощью дополнительных внешних программ (Plumed, APBS, PropKa). Молекулы, как и в программе PyMOL, могут отображаться в различных стилях, в зависимости от выбранного пользователем метода представления и способа окраски для определенного пользователем подмножества атомов. Для возможностей автоматизации VMD имеет встроенный скриптовый язык на основе Tcl. Для построения финального изображения в высоком качестве (для публикации) программа имеет движок трассировки лучей Tachyon, поддерживающий ускорение рендеринга с использованием графических процессоров (CUDA, OptiX).

Пакет программ ProDy [25] создан для анализа молекулярной динамики, позволяет количественно охарактеризовать структурную вариативность в наборе структур белков, в частности позволяет проводить анализ нормальных мод. В рамках работы [25], разработан плагин *NMWiz* для программы VMD, позволяющий визуализировать полученные расчеты.

3.2.2.3. Методы для анализа выравнивания структур гомологичных белков

Существуют биоинформатические методы, которые позволяют анализировать выравнивания структур гомологичных белков с целью нахождения сходств и различий между структурами. Но такие методы либо только позволяют сравнивать структуры белков целиком, то есть не делают акцент на локальных сходствах и различиях структур, либо служат для визуализации статистики, рассчитанной по множественному структурно-опосредованному выравниванию последовательностей, либо находят локальные сходства и различия структур только двух гомологов. Все эти методы могут быть использованы для нахождения локальных структурных отличий, отвечающих за функциональное разнообразие белков суперсемейства, но требуют дополнительного визуального экспертного анализа. И в настоящий момент не существует метода, автоматически выявляющего 3D-специфические паттерны суперсемейства белков. В этой главе подробно рассмотрим существующие на данный момент методы анализа выравнивания структур гомологичных белков.

DALI [102] – это известный веб-сервер, который сравнивает структуру интересующего пользователя белка со структурами из PDB и в качестве результата работы выдает наиболее похожие на данную структуры (на основе специально введенной Z -оценки), а также соответствующие выравнивания структур и последовательностей. Сходство структур двух белков и соответствующую Z -оценку DALI определяет на основе сходства паттернов контактов между аминокислотными остатками, а именно меры сходства DALI (на основе которой считается Z -оценка), выглядящей следующим образом:

$$S(A, B) = \sum_{i=1}^L \sum_{j=1}^L \left(\theta - \frac{|a(i, j) - b(i, j)|}{r(i, j)} \right) w(r(i, j)),$$

где A и B – белки, L – количество пар совмещенных при выравнивании остатков (остальные не рассматриваются), $a(i, j)$ и $b(i, j)$ – расстояния между C_α атомами остатков в структурах белков A и B , соответственно, $r(i, j)$ – среднее величин $a(i, j)$ и $b(i, j)$, θ – константа, $w(r(i, j))$ – сверточная функция. Помимо этого, DALI считает Z -оценки для всех пар найденных белков, на основе которых проводит анализ: строит дендрограмму и матрицу структурного сходства всех найденных белков, в удобном для пользователя виде отображая анализ ее содержимого. Таким образом веб-сервер DALI анализирует, насколько похожи структуры белков в целом (с помощью введенной Z -оценки) и никак не анализирует сходства и различия структур локально.

Сервер PSSweb [26] строит множественное выравнивание последовательностей и структур интересующих пользователя белков. Далее строятся графики среднеквадратичной флуктуации декартовых координат, В-фактора и стандартного отклонения углов ϕ , ψ , ω , χ_1 , χ_2 , χ_3 и χ_4 (по оси x откладывается номер позиции выравнивания последовательностей, по оси y – значение соответствующей величины). Также строятся карта Рамачандрана (ϕ/ψ) [103], χ_1/χ_2 и другие графики, отражающие распределение двугранных углов для данного множества белков. Также веб-сервер строит кластеризацию белков на основе двугранных углов и соответствующую дендрограмму. Результаты работы этого сервера (см. рисунок 8) – графики для различных параметров – требуют дополнительной интерпретации и автоматически не выявляют важные элементы белков.

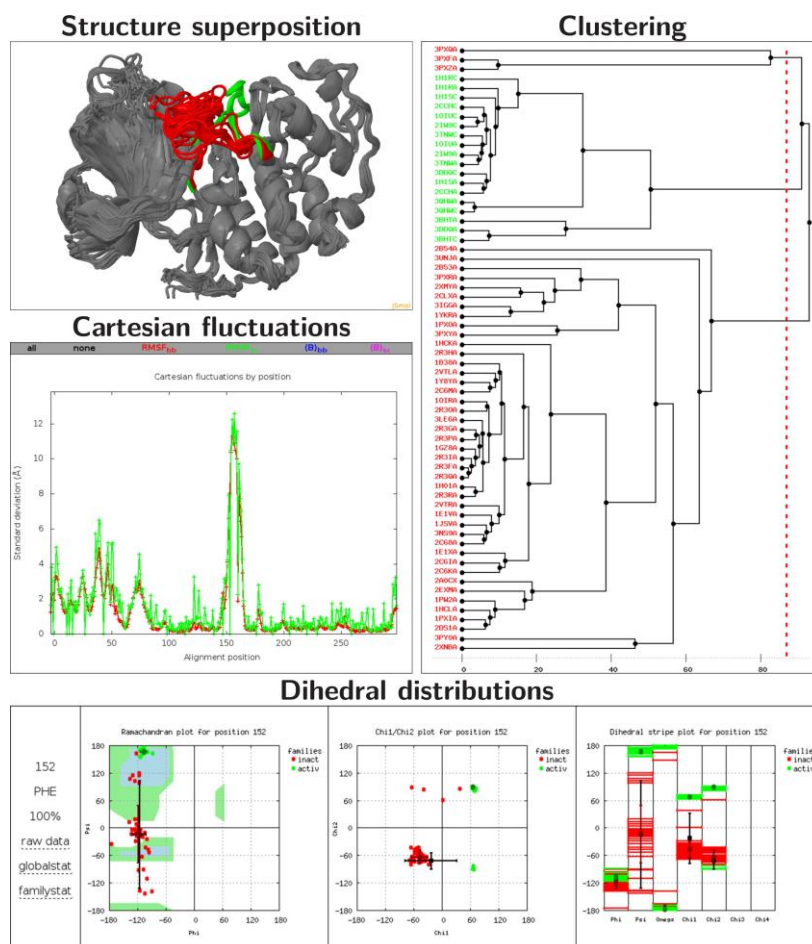


Рисунок 8. Результат работы веб-сервера PSSweb [26]: множественное выравнивание структур белков; график среднеквадратичной флуктуации декартовых координат вдоль позиций выравнивания; графики распределения двухгранных углов; кластеризация белков и соответствующая дендрограмма. Рисунок взят из [26].

Следующие два метода анализируют структурное выравнивание двух гомологичных белков или двух структур одного белка.

2StrucCompare [28] – веб-сервер, который сравнивает две структуры одного и того же белка или двух гомологов и выявляет различия между ними. Между каждой парой выровненных аминокислотных остатков интересующих пользователя структур белков рассчитываются:

- евклидово расстояние между C_{α} атомами;
- для одинаковых пар рассчитывается максимальное межатомное расстояние одинаковых атомов боковой цепи, определяемое после выравнивания основной цепи;

- число различающихся контактов (то есть для каждого из двух аминокислотных остатков находится множество аминокислотных остатков в радиусе 4 Å. А далее ищется мощность разности этих множеств).

Каждая из этих величин градиентным цветом наносится на трехмерные структуры белков (см. рисунок 9). Также с помощью четырех методов (DSSP, STRIDE, P-SEA и STICKS) рассчитываются вторичные структуры каждого из белков и показывается, в каких парах выравненных остатков вторичная структура различается. Веб-сервер 2StrucCompare может помочь в нахождении структурных различий белков, но может выполнять сравнение только двух гомологов и, таким образом, имеет ограниченную возможность для изучения больших суперсемейств.

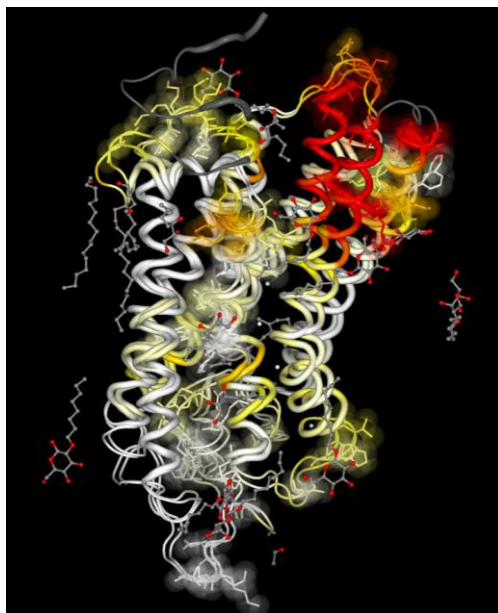


Рисунок 9. Пример результата работы веб-сервера 2StrucCompare [28]. Градиентным цветом показано евклидово расстояние между $C\alpha$ атомами двух выравненных структур гомологичных белков.

Как и сервер 2StrucCompare, веб-сервер FATCAT [29] работает только с двумя белковыми структурами. FATCAT позволяет «гибко» выравнять две структуры, то есть для минимизации RMSD FATCAT допускает повороты и переносы элементов выравняемых белков. Пользователю веб-сервера показывается само выравнивание, *P*-значение полученного выравнивания,

значение метрики RMSD, количество выравненных позиций и количество поворотов/переносов. Веб-сервер предоставляет различные иллюстрации структурных сходств и различий выравненных белков. FATCAT также осуществляет поиск по базе данных и находит белки, близкие к данному, по последовательности и структуре.

Наконец, машинное обучение появляется в структурной биоинформатике как мощный класс подходов к изучению растущего количества 3D-данных. Метод Caretta [22], помимо множественного структурного выравнивания, извлекает из уже выравненных белков признаки (каждый белок представляется в виде вектора фиксированной длины), основанные на координатах атомов, двугранных углах основной цепи, плоских углах, энергии водородных связей, поверхности, доступной растворителю, и др. Эти векторы признаков можно использовать в дальнейшем в алгоритмах машинного обучения (как с учителем, так и без учителя), например, для быстрого поиска по структурному сходству, классификации структур.

Тем не менее, несмотря на недавний прогресс в разработке передовых алгоритмов изучения белков на трехмерном уровне (то есть структур белков), в настоящее время не существует инструмента, направленного на систематический анализ специфических локальных трехмерных структурных различий между функционально разнообразными семействами белков.

3.2.2.4. Консервативные структурные паттерны суперсемейства белков: 3D-мотивы

3.2.2.4.1. Понятие 3D-мотива

Одним из способов сравнительного анализа структур гомологичных белков для понимания их свойств и функций является нахождение и анализ так называемых 3D-мотивов [30] – структурных аналогов консервативных позиций множественного выравнивания последовательностей (см. главу 3.2.1.1). 3D-мотив – это такая закономерность в локальной структуре белка,

которая присутствует во всех (или почти во всех) белках суперсемейства и отвечает за общность свойств и функций белков суперсемейства. Другими словами, 3D-мотивы – это характеристическое относительное положение элементов структуры белка, например, относительное расположение отдельных аминокислотных остатков, петель или фрагментов вторичной структуры, связанное с функцией и присутствующее во всех белках суперсемейства. Главным отличием 3D-мотивов от 3D-специфических паттернов является то, что 3D-мотивы – это ключевые атомы/остатки/участки цепи (основной или боковой), ответственные за *общую* функцию или *общее* для выбранных белков свойство.

Было разработано много методов для определения, выявления и анализа 3D-мотивов. Подходы различаются по типу и количеству входных данных, по тому, как 3D-мотивы описываются и сопоставляются со структурой белка и как они соотносятся с его функцией. Структуры белков можно сравнивать с известными трехмерными мотивами, ассоциированными с конкретными каталитическими свойствами, таким образом можно аннотировать белки с неизвестной функцией, а также можно искать новые 3D-мотивы в ранее изученных белках. 3D-мотив обычно описывается набором точек в пространстве, принадлежащих элементам структуры белка. Чаще всего в качестве таких точек используют координаты $C\alpha$ атомов аминокислотных остатков, однако иногда используют псевдоатомы (точки, характеризующие среднее положение атомов боковой цепи) или даже все атомы. Иногда учитываются и физико-химические особенности аминокислотных остатков.

3.2.2.4.2. Методы выявления 3D-мотивов

Ранние представления о 3D-мотивах основывались на наблюдениях, а не на алгоритмах. Наиболее изученным мотивом является каталитическая триада Ser-His-Asp (см. рисунок 10), впервые обнаруженная в сериновых протеазах [104,105], а затем и в других гидролазах (эстеразах и липазах). Каталитическая триада была детально изучена в работе [106].

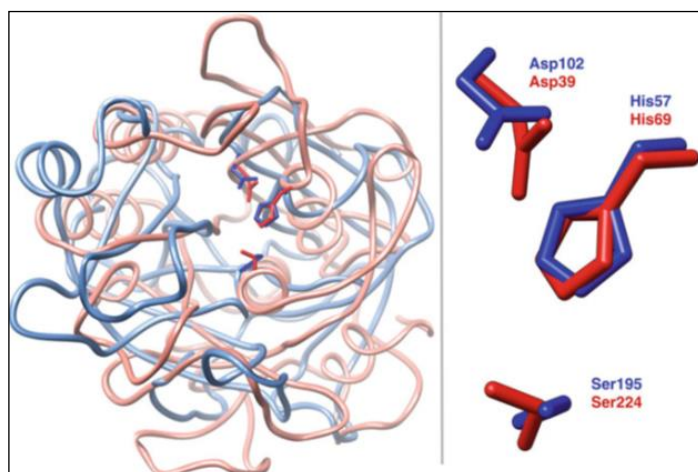


Рисунок 10. Сериновые протеазы: трипсин (PDB 1sgt, голубой) и протеиназа К (PDB 2pkc, красный), и их каталитические триады [30]. Рисунок взят из [30].

Позднее было разработаны алгоритмы для идентификации данного 3D-мотива в структурах белков, в том числе [30–43,107]. Рассмотрим некоторые из них.

Известный метод SPASM [37] был разработан для нахождения данного 3D-мотива в структурах белков. На вход алгоритму подается 3D-мотив пользователя. Каждая аминокислота представляется своим C_{α} -атомом и псевдоатомом (атомом, расположенном в центре тяжести атомов боковой цепи). Далее в каждой PDB-структуре из базы данных (множество структур базы данных – подмножество базы данных PDB) перебираются всевозможные комбинации аминокислотных остатков, такие что их типы соответствуют типам аминокислотных остатков в искомом 3D-мотиве (возможны замены на близкие по свойствам аминокислотные остатки). Полученные комбинации аминокислотных остатков в белке сравниваются с 3D-мотивом с помощью метрики RMSD. Если значение метрики не превосходит пороговое значение, то такая комбинация аминокислотных остатков считается соответствующей данному 3D-мотиву.

Схожий метод TESS [38] использует информацию о трех атомах аминокислоты (для каждой аминокислоты свои 3 атома), а информация об аминокислотных остатках белка хранится в специальной хэш-таблице для быстрого поиска мотива.

Одним из более новых алгоритмов поиска данного 3D-мотива является алгоритм [39]. Он учитывает только C_{α} -атомы, а также тип аминокислоты (возможны замены). Происходит перебор по аминокислотным остаткам рассматриваемого белка следующим образом: если в сфере заданного радиуса с центром в данном аминокислотном остатке присутствуют типы аминокислотных остатков данного мотива, то далее для всех возможных допустимых комбинаций вычисляется RMSD с искомым мотивом. Если полученное RMSD меньше порогового, то такая комбинация аминокислотных остатков считается соответствующей данному 3D-мотиву.

Похожим образом работает алгоритм BALLAST [40]. Он также перебирает все аминокислотные остатки белка, в котором ищет заданный пользователем мотив. Если в ϵ -окрестности данного аминокислотного остатка присутствуют аминокислотные остатки белка, удовлетворяющие определенным критериям (их количество равно количеству аминокислотных остатков в 3D-мотиве, их попарные расстояния соответствуют попарным расстояниям между аминокислотными остатками 3D-мотива, их типы соответствуют типам аминокислотных остатков 3D-мотива, метрика RMSD между данными аминокислотными остатками и аминокислотными остатками 3D-мотива не превышает порогового значения), то такой набор аминокислотных остатков белка считается соответствующим 3D-мотиву.

Рассмотренные выше алгоритмы поиска данного 3D-мотива в структуре белка работают аналогично, то есть перебирают подмножества аминокислотных остатков белка и, используя метрику RMSD, сравнивают 3D-мотив с данным подмножеством аминокислотных остатков белка, а пороговое значения RMSD для определения того, соответствует ли данное множество аминокислотных остатков белка данному 3D-мотиву, выбирается жестко и заранее без использования статистики. Различия же заметны по скорости работы и количеству атомов в аминокислотах, которые эти алгоритмы учитывают, а также в дополнительных критериях отбора.

Еще одним классом методов работы с 3D-мотивами являются методы, использующие графовый подход. Вершины такого графа – аминокислотные остатки белка, которые соединены ребром, если, к примеру, расстояние между ними меньше порогового значения. Подграфы такого графа соответствуют мотиву. Изоморфные подграфы соответствуют одному мотиву. Например, графовый подход используется в работе *SPRITE and ASSAM* [33]. В этом методе каждому аминокислотному остатку ставится в соответствие вектор из двух псевдоатомов. Далее каждый белок представляется в виде графа (вершина графа – аминокислотный остаток). Веса ребер в графе соответствуют расстоянию между соответствующими векторами. Каждый узел графа также имеет свои метки, а именно принадлежность остатка к элементу вторичной структуры, расстояние до известного сайта связывания, доступность для растворителя и другие параметры. Далее с помощью различных графовых алгоритмов в случае программы *ASSAM* происходит поиск 3D-мотива пользователя среди белков в базе данных *PDB*. В случае программы *SPRITE*, напротив, происходит поиск известных мотивов в заданном пользователем *PDB*-файле. Схема работы *SPRITE and ASSAM* представлена на рисунке 11.

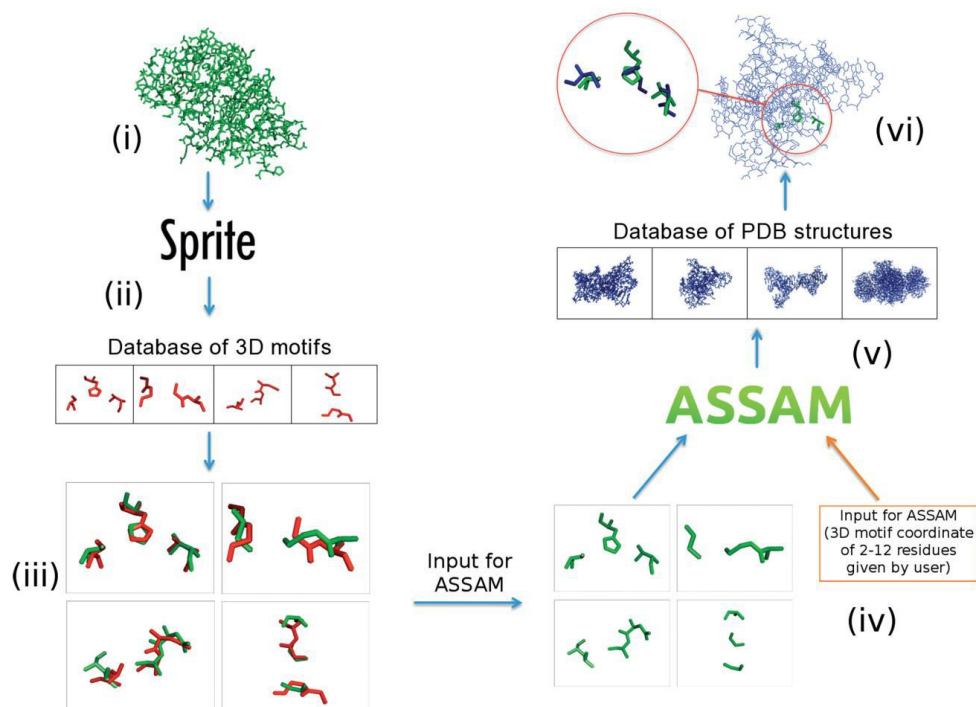


Рисунок 11. Схема работы сервера SPRITE (слева) и ASSAM (справа) [33]. В качестве входных данных SPRITE принимает структуру белка в виде PDB-файла и сравнивает ее с базой данных трехмерных мотивов. Результатом работы SPRITE является список аминокислотных остатков в структуре белка-запроса, который соответствует мотиву из базы данных. Входными данными для ASSAM могут быть либо 3D-мотив, заданные пользователем, либо результат работы SPRITE. ASSAM сравнивает этот 3D-мотив с репрезентативными структурами в PDB. Результатом работы ASSAM является список структур PDB, которые содержат данный 3D-мотив. Рисунок взят из [33].

Определение потенциально важных для функции белка 3D-мотивов – отдельная задача биоинформатики. В этом вопросе могут быть использованы различные подходы, например, в качестве мотива могут быть рассмотрены консервативные аминокислоты, каталитическая триада белка, все остатки, взаимодействующие с лигандом. В работе по поиску новых 3D-мотивов – Gremlin [31,42] – происходит поиск белок-лигандных контактов. Строится граф, ребра которого соответствуют взаимодействию атомов белка и лиганда. Далее ищется методом *gspan* максимально частый подграф [108]. Найденные подграфы считаются 3D-мотивами.

3.3. Методы машинного обучения

Машинное обучение (МО) – это класс математических алгоритмов, которые способны автоматически обнаружить в данных скрытые и ранее неизвестные закономерности, а также самостоятельно приобретать свойства, необходимые для распознавания этих закономерностей. Машинное обучение является частью науки искусственного интеллекта (ИИ). Алгоритмы машинного обучения строят модель на основе выборочных данных и делают прогнозы или принимают решения без явного прямого программирования этого [109]. Алгоритмы машинного обучения используются в самых разных областях, таких как медицина, распознавание речи, компьютерное зрение и биоинформатика, где сложно или невозможно разработать обычные алгоритмы для решения поставленных задач [110].

3.3.1. Обучение с учителем

3.3.1.1. Постановка задачи

Пусть X – это пространство объектов, Y – множество допустимых ответов. Пусть существует неизвестная целевая зависимость – отображение $y^*: X \rightarrow Y$, значения которой известны на объектах обучающей выборки $\{(x_1, y_1), \dots, (x_N, y_N)\}$ из N элементов, где каждый $x_i \in X$ – это вектор признаков, а каждый $y_i \in Y$ – это метка/класс. Задача алгоритма машинного обучения с учителем состоит в том, чтобы найти алгоритм $g: X \rightarrow Y$, который приближал бы неизвестную целевую зависимость, то есть который дает достаточно точные ответы как на обучающей выборке, так и на данных, выходящих за пределы имеющейся обучающей выборки. Функция g – это алгоритм, способный для любого возможного входного объекта из пространства X выдать достаточно точное значение метки. Для измерения точности ответов вводится оценочный функционал качества.

Множество возможных ответов Y может быть:

- бесконечно (ответы являются действительными числами или векторами), такая задача называется задачей регрессии и аппроксимации;
- конечно (такая задача называется задачей классификации).

3.3.1.2. Алгоритмы машинного обучения с учителем

Существует множество алгоритмов машинного обучения с учителем. В данной работе использовались только алгоритмы машинного обучения без учителя, поэтому в этой главе остановимся на простом перечислении алгоритмов машинного обучения с учителем. Вот некоторых из них:

- метод ближайших соседей [111];
- линейная регрессия [112];
- логистическая регрессия [113];
- машина опорных векторов [114];
- решающее дерево [115];
- нейронные сети [116].

3.3.2. Обучение без учителя

3.3.3. Постановка задачи

Машинное обучение без учителя решает разные типы задач, например: задачи кластеризации, поиска аномалий, сокращения размерности, визуализации данных. В данной главе остановимся на задаче кластеризации, так как именно ее мы решали в рамках этой работы. Пусть X – множество объектов, и на множестве X задана функция расстояния между объектами $\rho(x, x')$. Пусть дан конечный набор объектов $X_m = \{x_1, \dots, x_m\} \subset X$. Необходимо разбить набор объектов на кластеры (подмножества), то есть каждому объекту $x_i \in X_m$ сопоставить метку кластера, таким образом, чтобы объекты внутри каждого кластера были близки относительно метрики ρ , а объекты из разных кластеров значительно различались.

3.3.3.1. Алгоритм кластеризации k -средних

Алгоритм кластеризации k -средних [117] — это метод кластеризации, стремящийся минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

Кластеризация k -средних разделяет набор n точек (x_1, x_2, \dots, x_n) на k ($\leq n$) подмножеств $S = \{S_1, \dots, S_k\}$, таким образом, чтобы минимизировать сумму квадратов расстояний от каждой точки кластера до его центра, то есть находит:

$$V = \underset{S}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2,$$

где

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x,$$

k — число кластеров, S_i — полученные кластеры, $i = 1, 2, \dots, k$, а μ_i — центр масс всех точек x из кластера S_i .

Основные недостатки этого метода кластеризации заключаются в том, что:

- от пользователя требуется задание числа кластеров k ;
- метод не определяет выбросы, то есть все точки набора относит к одному из кластеров;
- результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен;
- не гарантируется достижение глобального минимума суммарного квадратичного отклонения V , а только одного из локальных минимумов.

3.3.3.2. Алгоритм кластеризации DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) — это алгоритм кластеризации данных, разработанный в 1996 году [44]. Это алгоритм кластеризации, основанный на плотности данных. На вход алгоритму подается набор точек в пространстве. Алгоритм DBSCAN

группирует вместе точки, лежащие рядом друг с другом, и отмечает как точки выброса те точки, ближайшие соседи которых далеко.

Более подробно алгоритм кластеризации DBSCAN состоит в следующем. Пусть дан набор точек в пространстве и два параметра, заданные пользователем: $minPts$ и ε . Будем называть точку *базовой*, если в ε -окрестности этой точки $N_\varepsilon(P)$ больше, чем $minPts$ точек. Выполняются следующие шаги:

- В цикле обрабатываются все точки набора, не отнесенные еще ни к одному кластеру и не помеченные как выбросы.
- Если рассматриваемая точка не базовая, то такая точка помечается как выброс и алгоритм переходит к рассмотрению следующей точки.
- Если точка базовая, то сама точка и все ее соседи, не отнесенные еще ни к одному кластеру, помечаются как точки, входящие в новообразованный кластер. Все соседи базовых точек, отнесенных к данному кластеру, также относятся алгоритмом к этому кластеру.

Алгоритм DBSCAN является одним из самых распространенных алгоритмов кластеризации. Можно выделить следующие преимущества данного алгоритма:

- от пользователя не требуется задания количества кластеров;
- метод позволяет находить кластеры произвольной формы;
- метод относит часть точек к выбросам.

Из недостатков данного алгоритма можно выделить необходимость задания двух параметров $minPts$ и ε , а задание последнего может быть проблемой особенно в тех случаях, когда данные имеют большую разницу в плотности.

3.3.3.3. Алгоритм кластеризации OPTICS

Ordering points to identify the clustering structure (OPTICS) – это алгоритм кластеризации данных на основе их плотности [45]. Основная идея

алгоритма OPTICS похожа на DBSCAN, но у него нет одного из основных недостатков DBSCAN – проблемы обнаружения кластеров с различной плотностью. Для этого кластеризуемые точки линейно упорядочиваются таким образом, чтобы точки, лежащие близко в пространстве, были соседями.

Как и DBSCAN, OPTICS требует на вход два параметра:

- параметр ε равен радиусу окрестности точки, принимаемой во внимание;
- параметр $minPts$ описывает минимальное количество точек, необходимых для формирования кластера.

Также, как и в алгоритме DBSCAN, будем называть точку *базовой*, если в ε -окрестности этой точки $N_\varepsilon(P)$ больше, чем $minPts$ точек. Каждой точке ставится в соответствие базовое расстояние, описывающее расстояние до ближайшей $minPts$ -ой точки:

$$\begin{aligned}
 & cordist_{\varepsilon, minPts}(P) \\
 = & \begin{cases} \text{Неопределено, если } |N_\varepsilon(P)| < minPts, \\ \text{Расстояние до ближайшей } minPts \text{-ой точки в } N_\varepsilon(P), \text{ иначе.} \end{cases}
 \end{aligned}$$

Расстояние достижимости от точки P до точки O равно расстоянию между точками O и P или базовому расстоянию P , в зависимости от того, что больше:

$$\begin{aligned}
 & reachability - dist_{\varepsilon, minPts}(P) \\
 = & \begin{cases} \text{Неопределено, если } |N_\varepsilon(p)| < MinPts, \\ \max(coredist_{\varepsilon, minPts}(P), dist(P, O)), \text{ иначе.} \end{cases}
 \end{aligned}$$

Идея шагов алгоритма OPTICS состоит в следующем:

- Вычисляются базовые расстояния для всех точек.
- Далее в цикле обрабатываются все точки набора, для каждой из которых вычисляется расстояние достижимости: каждая следующая точка, выбранная для обработки, имеет наименьшее расстояние достижимости

относительно данной и ставится в очередь следующей. Таким образом точки из одного кластера (то есть лежащие близко в пространстве) оказываются близкими в данном отношении порядка. В результате этого шага точки набора помещаются в упорядоченный список и помечаются расстоянием достижимости.

- График, представляющий упорядоченный список, полученный на предыдущем шаге, показан на рисунке 12. По оси x отложены точки набора в соответствующем порядке, а по оси y отложено расстояние достижимости. Поскольку точки, принадлежащие одному кластеру, имеют небольшое расстояние достижимости до ближайшего соседа – представителя своего кластера, кластеры выглядят, как «долины» на графике достижимости. Чем глубже «долина», тем плотнее кластер.
- Далее находятся локальные максимумы и «долины» в графике. Каждая «долина» соответствует одному кластеру.



Рисунок 12. При кластеризации с помощью метода OPTICS точки линейно упорядочиваются таким образом, чтобы точки, лежащие близко в пространстве, были соседями. На графике по оси x отмечены точки в построенном порядке, по оси y отложено расстояние достижимости. Кластеры на графике выглядят, как «долины». Рисунок взят из [118].

И базовое расстояние, и расстояние достижимости могут быть не определены для точки. При достаточно большом значении параметра ϵ такая проблема никогда не возникнет, но в этом случае время выполнения алгоритма значительно увеличивается. Параметр ϵ , строго говоря, не нужен, и его можно принять равным максимально возможному значению.

Преимущества алгоритма OPTICS перед алгоритмом DBSCAN:

- параметр ε является необязательным, его можно положить равным бесконечности;
- позволяет обнаруживать кластеры с различной плотностью.

3.3.3.4. Алгоритм кластеризации HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [46] – это алгоритм кластеризации, улучшающий алгоритм DBSCAN таким образом, что позволяет находить кластеры с различной плотностью и не требует задания параметра ε . Входными параметрами алгоритма HDBSCAN являются: $min_samples$ и $min_cluster_size$.

На первом этапе алгоритма определяется новая метрика на пространстве объектов – *расстояние взаимной досягаемости*:

$$d_{mreach}(a, b) = \max\{cordist_{minSamples}(a), cordist_{minSamples}(b), d(a, b)\},$$

где $cordist_{minSamples}(a)$ – это базовое расстояние точки a , определяемое как в методе OPTICS, $d(a, b)$ – это исходное расстояние между точками a и b . Таким образом, точки, лежащие в областях с большей плотностью (то есть точки с маленьким базовым расстоянием), остаются на одном и том же расстоянии друг от друга, а точки, находящиеся в более разреженных областях, отодвигаются от любой другой точки на базовое расстояние. Такое преобразование пространства понижает плотность разреженных областей, что делает «шум» менее заметным для кластеризации, в то время как плотные области, претендующие на то, чтобы стать кластерами, свою плотность сохраняют.

Далее с помощью алгоритма Прима строится минимальное остовное дерево (см. рисунок 13), ребра которого помечены расстоянием взаимной досягаемости между вершинами, на основе которого строится дендрограмма (см. рисунок 14).

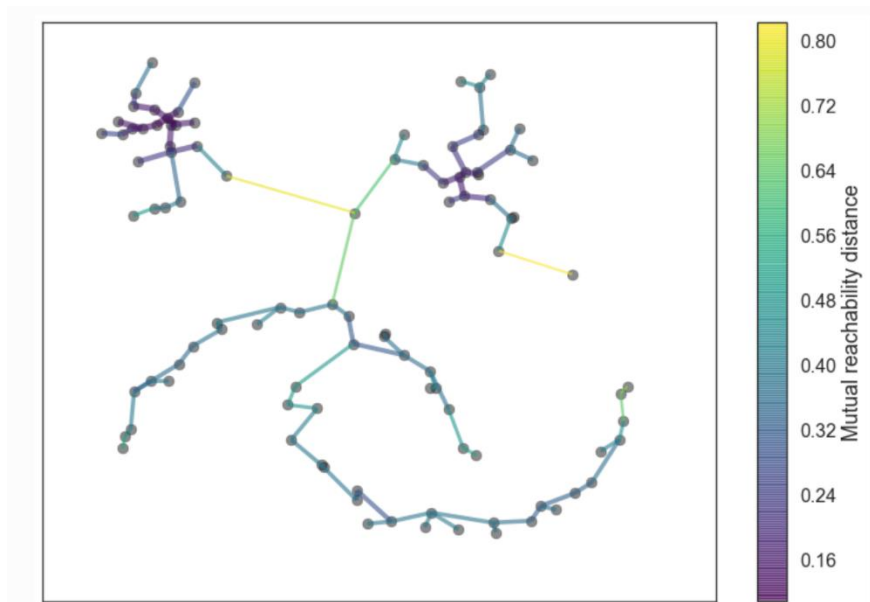


Рисунок 13. Этап алгоритма кластеризации HDBSCAN – построение минимального остовного дерева. Метки ребер – расстояния взаимной досягаемости между вершинами. Рисунок взят из [119].

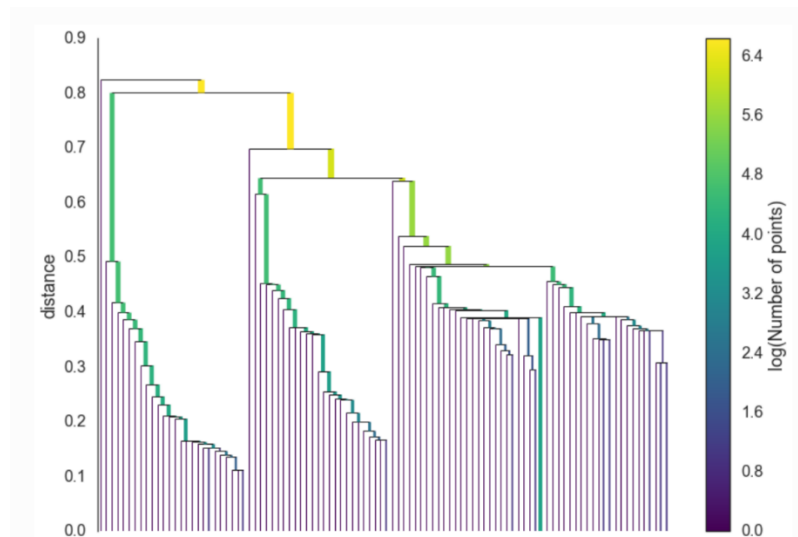


Рисунок 14. Этап алгоритма кластеризации HDBSCAN – построение дендрограммы на основе минимального остовного дерева. Рисунок взят из [119].

Претенденты на возможные кластеры получаются следующим образом: проходя сверху вниз по дендрограмме, каждый кластер разделяется на два меньших по размеру, если оба кластера имеют размер не меньший, чем значение параметра *min_cluster_size*. Если же один из получившихся кластеров имеет размер, меньший, чем *min_cluster_size*, то исходный кластер не делится на два, а просто рассматривается как «теряющий точки» кластер. Возможные претенденты на кластеры изображены на рисунке 15. В ответ же попадают

наиболее стабильные кластеры, то есть кластеры, имеющие наибольшую площадь на графике (см. рисунок 15).

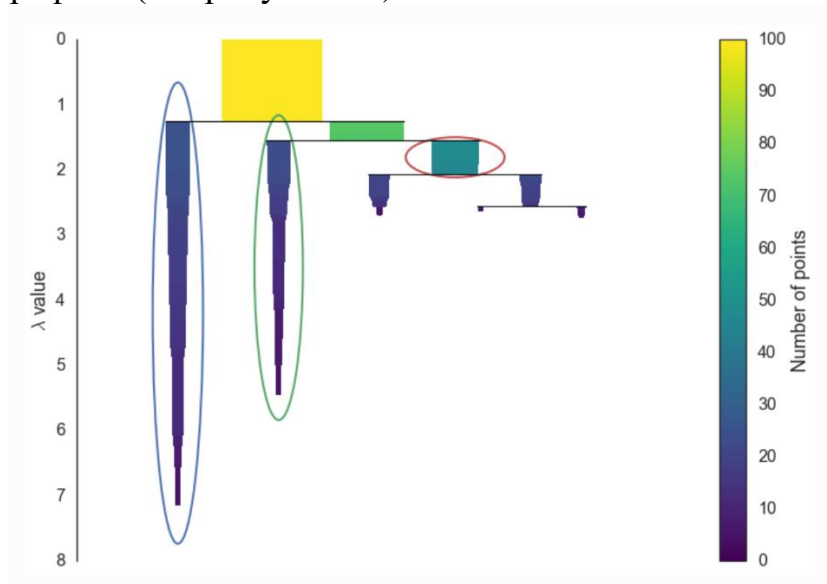


Рисунок 15. Закрашенные области – полученные с помощью анализа дендрограммы претенденты на кластеры. В качестве результата кластеризации выбираются те кластеры, которые имеют наибольшую площадь на графике. Рисунок взят из [119].

Преимущества алгоритма HDBSCAN перед алгоритмом DBSCAN:

- не требуется задание параметра ϵ ;
- метод позволяет обнаруживать кластеры с различной плотностью.

4. Материалы и методы

4.1. Построение множественного структурного выравнивания белков

Множественные структурные выравнивания суперсемейств белков были получены с помощью веб-сервера Mustguseal [48] и программы ParMAT [16].

4.2. Визуализация структур белков

Для визуализации структур белков была использована программа PyMOL версии 2.5.4 [23].

4.3. Библиотеки, использованные в программном обеспечении для поиска 3D-специфических паттернов суперсемейства

Для написания программного обеспечения, реализующего поиск 3D-специфических паттернов суперсемейства, был использован язык Python3. Программа написана с использованием принципов объектно-ориентированного программирования (ООП). Используемые библиотеки описаны в таблице 1.

Таблица 1. Описание использованных в программах для нахождения 3D-специфических паттернов библиотек.

Название библиотеки	Версия	Назначение
biopython	1.74	Используется для выравнивания последовательностей белков, а именно для сопоставления структурно-опосредованного выравнивания белков с 3D-структурным выравниванием.
numpy	1.17.2	Используется для работы с матрицами.
scikit-learn	0.21.3	Из этой библиотеки используются реализованные в ней алгоритмы кластеризации DBSCAN и OPTICS.
scipy	1.3.1	Используется для расчета статистики.
hdbscan	0.8.23	Реализует алгоритм кластеризации HDBSCAN.

4.4. Оценка качества кластеризации участков основной и боковых цепей белков суперсемейства с использованием метрики силуэт

Для оценки качества кластеризации участков основной и боковых цепей белков суперсемейства была использована метрика силуэт [120]. Пусть при кластеризации множества были получены кластеры $C_i, i=1, \dots, N$. Тогда для каждой точки k множества метрика силуэт рассчитывается по формуле:

$$sil(k) = \frac{b(k) - a(k)}{\max(a(k), b(k))},$$

где

$$a(k) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, j \neq k} d(k, j),$$

$$b(k) = \min_{n, n \neq k} \frac{1}{|C_n|} \sum_{j \in C_n} d(k, j).$$

Здесь C_k – кластер, которому принадлежит точка k , $a(k)$ – среднее расстояние от точки k до других точек в этом кластере, $b(k)$ – среднее расстояние от точки k до точек в соседнем кластере, $d(k, j)$ – расстояние между точками k и j .

Метрика качества кластеризации силуэт для кластеризации данного множества вычисляется как среднее значение метрик силуэт всех точек множества и показывает, насколько близки точки внутри одного кластера по отношению к точкам из соседнего кластера.

4.5. База данных конформационного разнообразия белков PDFFlex. Создание выборки для расчета статистики с целью определения функционально-значимых 3D-специфических паттернов

PDFFlex – база данных конформационного разнообразия белков, которая исследует гибкость белковых структур путем анализа структурных различий между различными PDB-структурами и цепями в асимметричных субъединицах одного и того же белка [47]. Для создания выборки для расчета

статистики из базы данных PDBFlex было выбрано 100 случайных наборов, содержащих не менее 20 PDB-структур в каждом. Каждый набор представляет собой различные примеры структурных флуктуаций одного белка. Окончательно выбранные наборы содержат от 26 до 515 PDB-файлов, со средним значением 59 файлов в наборе. Каждый набор содержит снимки различных конформаций структуры одного и того же белка (например, 325 PDB-файлов соответствуют человеческой p38a MAP-киназе). Каждое такое множество подавалось на вход программе структурного выравнивания parMATT [16]. Наборы, для которых очевидны глобальные структурные перестройки (например, перемещение доменов) далее не рассматривались (отсеивались). Таким образом было получено 76 наборов, каждый из которых содержит разные структуры одного белка. Для поиска 3D-специфических паттернов к каждому такому набору были применены программы, описанные в главе 5.1.2 со значением настраиваемых параметров, представленными в таблице 2.

Таблица 2. Значения настраиваемых параметров программ для нахождения 3D-специфических паттернов. См. описание в тексте.

Название параметра	cpu_threads	method	min_size_of_subfamily (в случае выявления СУОЦов)	max_content_of_gaps	max_content_of_mismatch	max_ssr_length (в случае СУОЦов)
Значение	все	hdbscan	10% от общего числа белков, но не меньше 2	5	5	10% средней длины белков супер-семейства

Название параметра	max_outliers	exclude_ncterm (в случае СУОЦов)	min_samples (в случае СОБЦов)	min_cluster_size (в случае СОБЦов)	number_of_result_resids (в случае СОБЦов)
Значение	40	5	None	10% от общего числа белков, но не меньше 2	50

4.6. Z-оценка статистической значимости и соответствующая P-оценка

Z-оценка [121] – это мера относительного разброса наблюдаемого значения, которая показывает, сколько стандартных отклонений составляет его разброс относительно среднего значения. Если известны среднее значение μ и стандартное отклонение σ генеральной совокупности, то Z-оценка вычисляется по формуле

$$Z = \frac{x - \mu}{\sigma}.$$

Абсолютное значение Z-оценки представляет собой расстояние между x и средним значением генеральной совокупности в единицах стандартного отклонения. Значение Z-оценки является отрицательным, когда x ниже среднего значения, и положительным, когда выше.

Для вычисления Z-оценки по этой формуле требуется знание среднего значения совокупности и стандартного отклонения совокупности. Однако получение истинного среднего значения и стандартного отклонения совокупности часто бывает невозможно. Поэтому, когда среднее значение и стандартное отклонение неизвестны, как в нашем случае, Z-оценка получается с использованием среднего значения и стандартного отклонения случайной выборки. В этих случаях Z-оценка определяется как:

$$Z = \frac{x - \bar{x}}{S},$$

где \bar{x} – выборочное среднее, S – стандартное отклонение случайной выборки.

P-оценка – это статистическая достоверность, вероятность случайной встречи полученных результатов. Значения, близкие к 0, говорят о высокой значимости результатов. Математически P-оценка рассчитывается как площадь области под кривой распределения вероятностей, находящаяся правее наблюдаемого значения (см. рисунок 16).

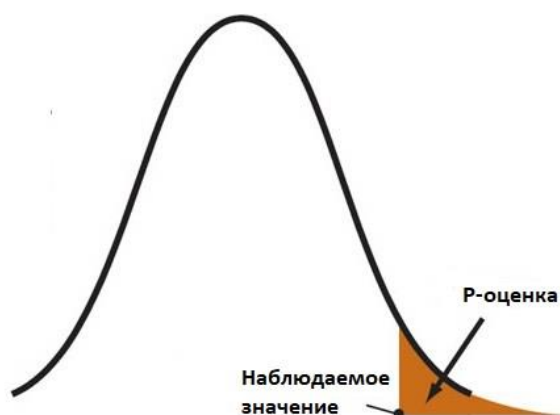


Рисунок 16. *P*-оценка рассчитывается как площадь области под кривой распределения вероятностей, находящаяся правее наблюдаемого значения.

4.7. Создание выборок для апробации нового подхода к анализу 3D-специфичности в структурах белков суперсемейства

С целью создания выборок для апробации нового подхода к анализу 3D-специфичности в структурах белков суперсемейства, описанного в главе 5.1, были взяты 10 репрезентативных суперсемейств белков, каждое из которых содержит известные из литературы участки *основной* цепи белков, отвечающих за каталитическое разнообразие ферментов суперсемейства, различие в субстратной специфичности ферментов суперсемейства и за конформационную вариабельность, обеспечивающую вклад динамики глобулы белка в катализ. PDB-код представителя каждого из этих суперсемейств был подан на вход веб-серверу Mustguseal для поиска избыточного набора из не более 32 структур гомологов в базе данных PDB [48]. Результат работы Mustguseal – множественные структурные выравнивания суперсемейств белков, включающие в себя эти белки-представители.

Для создания выборки для апробации в случае 3D-специфических паттернов в *боковой* цепи делалось следующее. Из базы CSA [122] были случайно выбраны 195 ферментов, относящихся к разным суперсемействам. Для каждого фермента брали аннотацию каталитических остатков из CSA и выбирали соответствующую PDB запись с максимально высоким разрешением, закристаллизованную вместе с лигандом, расположенным в радиусе 5 Å от каталитических остатков. После чего выделяли все остатки в

структуре фермента, расположенные на расстоянии 5 Å от выбранного лиганда. Проаннотированные в результате описанной процедуры остатки считали функционально важными на том основании, что они расположены в непосредственной близости от лиганда, а значит, с высокой вероятностью принимают непосредственное участие в его связывании и/или превращении. Каждый репрезентативный фермент был использован в качестве ввода для алгоритма Mustguseal для поиска избыточного набора из не более 32 структур гомологов в базе данных PDB и построения соответствующего структурного выравнивания.

К полученным структурным выравниваниям суперсемейств (в случае основной и боковой цепей) для выявления 3D-специфических паттернов были применены программы, описанные в главе 5.1.2 со значениями параметров, представленными в таблице 3.

Далее в случае боковой цепи отбирались такие 3D-специфические паттерны, для которых соответствующие позиции в структурно-опосредованном выравнивании являются консервативными (то есть содержат остатки одного типа), а соответствующие аминокислотные остатки репрезентативного белка являются функционально значимыми. Среди выбранных таким образом 3D-специфических паттернов с помощью литературных источников находились ответственные за каталитическое разнообразие ферментов суперсемейства, различие в субстратной специфичности ферментов суперсемейства и за конформационную вариабельность, обеспечивающую вклад динамики глобулы белка в катализ. Рассматривались именно консервативные позиции, так как они вызывают наибольший интерес. Ведь вклад таких позиций в структурное и функциональное разнообразие белков суперсемейства невозможно предсказать только по выравниванию аминокислотных последовательностей.

Таблица 3. Значения настраиваемых параметров программ для нахождения 3D-специфических паттернов. См. описание в тексте.

Название параметра	cpu_threads	method	min_size_of_subfamily (в случае СУОЦов)	max_content_of_gaps	max_content_of_mismatch	max_ssr_length (в случае СУОЦов)
Значение	все	hdbscan	10% от общего числа белков, но не меньше 2	5	5	-

Название параметра	max_outliers	exclude_ncterm (в случае СУОЦов)	min_samples (в случае СОБЦов)	min_cluster_size (в случае СОБЦов)	number_of_result_resids (в случае СОБЦов)
Значение	-	-	None	10% от общего числа белков, но не меньше 2	50

4.8. Применение метода Zebra2 для получения специфических позиций подсемейства

Для разделения суперсемейства белков на подсемейства по сходству последовательностей и расчета специфических позиций подсемейства был применен метод Zebra2 [123]. Среди полученных кластеризаций нами рассматривалась наиболее статистически значимая кластеризация. В качестве порогового значения *P*-оценки для отделения специфических позиций подсемейства рассматривался *global statistical significance threshold*.

4.9. Adjusted Rand Index – мера сходства двух кластеризаций, использованная для сравнения результатов разделения белков на подсемейства

Для сравнения кластеризации, полученной для данного 3D-специфического паттерна с кластеризацией, полученной с помощью метода Zebra2 [123], использовалась мера сходства двух кластеризаций *Adjusted Rand Index* [124,125].

Пусть даны две кластеризации множества элементов мощности n , то есть два разделения данного множества на s и r наборов: $X = \{X_1, \dots, X_s\}$ и $Y = \{Y_1, \dots, Y_r\}$. Пусть $n_{i,j}$ – мощность множества пересечения множеств X_i и Y_j .

$X \setminus Y$	Y_1	Y_2	\dots	Y_s	sums
X_1	n_{11}	n_{12}	\dots	n_{1s}	a_1
X_2	n_{21}	n_{22}	\dots	n_{2s}	a_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
X_r	n_{r1}	n_{r2}	\dots	n_{rs}	a_r
sums	b_1	b_2	\dots	b_s	

Тогда *Adjusted Rand Index* (ARI) вычисляется по формуле:

$$ARI = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}$$

Значения ARI, близкие к единице, говорят о высоком сходстве сравниваемых кластеризаций. Значения ARI, близкие к нулю, говорят о случайном разделении множества элементов на кластеры.

4.10. Извлечение дисульфидных мостиков из структур белков базы данных PDB для получения 3D-мотивов

Из базы данных PDB были загружены все структуры с разрешением не более 2.5 Å. С помощью программы CD-HIT [126] загруженные структуры были профильтрованы по сходству аминокислотных последовательностей с порогом 95%. В полученных структурах были проанализированы все пары цистеинов, находящиеся на расстоянии не более 2.5 Å. В результате анализа выяснилось, что расстояния между такими парами цистеинов имеют нормальное распределение со средним $\mu = 2.05$ Å и стандартным отклонением $\sigma = 0.06$ Å, что согласуется с работами [127,128]. Поэтому в качестве дисульфидных связей в выбранных структурах белков рассматривались такие пары цистеинов, расстояние между которыми попадает в интервал

(1.88 Å; 2.22 Å) (что соответствует $\mu \pm 3\sigma$). Таким образом было получено 16956 дисульфидных мостиков. По полученным дисульфидным мостикам была рассчитана статистика расстояний между C_α и C_β атомами цистеинов, входящих в один дисульфидный мостик. Согласно этой статистике $dist(C_\alpha, C'_\alpha) < 7.70 \text{ \AA}$ (что соответствует $\mu + 3\sigma$, $\mu = 5.52 \text{ \AA}$, $\sigma = 0.73 \text{ \AA}$), $dist(C_\beta, C'_\beta) < 4.48 \text{ \AA}$ (что соответствует $\mu + 3\sigma$, $\mu = 3.85 \text{ \AA}$, $\sigma = 0.21 \text{ \AA}$).

4.11. Апробация и расчет специфичности и чувствительности статистического критерия определения возможности вставки данного 3D-мотива в структуру белка

Для апробации разработанной статистической модели в качестве истинных дисульфидных мостиков был рассмотрен большой набор пар цистеинов, найденных в структурах белков базы данных PDB с разрешением не более 2.5 Å, таких что расстояние между цистеинами не превышает 2.5 Å. В качестве пар аминокислотных остатков, которые не могут образовывать дисульфидный мостик, выбирались пары цистеинов с $dist(C_\alpha, C'_\alpha) > 7.70 \text{ \AA}$, $dist(C_\beta, C'_\beta) > 4.48 \text{ \AA}$ (см. главу 4.10).

Специфичность и чувствительность были рассчитаны по формулам:

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP}$$

где TP – количество истинных дисульфидных связей, которые соответствуют хотя бы одному 3D-мотиву (согласно статистическому критерию), FN – количество истинных дисульфидных связей, которые не соответствуют ни одному 3D-мотиву (согласно статистическому критерию), TN – количество цистеинов, не образующих дисульфидный мостик, которые не соответствуют ни одному 3D-мотиву (согласно статистическому критерию), FP – количество цистеинов, не образующих дисульфидный мостик, которые соответствуют хотя бы одному 3D-мотиву (согласно статистическому критерию).

5. Результаты и обсуждение

В работе предложены методы выявления и анализа структурных паттернов суперсемейств белков, таких как 3D-специфические паттерны и 3D-мотивы.

Для поиска 3D-специфических паттернов суперсемейства белков, то есть структурных паттернов суперсемейства, схожих внутри подсемейств белков, но различающихся между ними и отвечающих за функциональное разнообразие белков суперсемейства, был разработан метод [129,130], описанный в главе 5.1, позволяющий выявлять такие части белковых структур (как в основной, так и в боковых цепях), которые схожи (в смысле метрики RMSD) внутри подсемейств, но различаются между ними. Такие 3D-специфические паттерны являются структурным аналогом специфических позиций подсемейства (см. главу 3.2.1.2). Разработанный метод не требует предварительного деления суперсемейства белков на подсемейства, а делает это сам (причем деление суперсемейства на подсемейства может отличаться в зависимости от рассматриваемого 3D-специфического паттерна), а также ранжирует полученные 3D-специфические паттерны в зависимости от значения специально введенной S -оценки. Чем выше S -оценка, тем ниже вероятность того, что данный 3D-специфический паттерн – результат случайных колебаний белковых структур. В главе 5.1.1.2.4 вводится специальная статистическая модель, позволяющая отделять функционально значимые 3D-специфические паттерны от случайных колебаний белковой структуры. В главе 5.2 показано, что найденные 3D-специфические паттерны отвечают за функциональное разнообразие белков суперсемейства, а именно могут быть ответственны за субстратную специфичность, каталитическую активность и вариабельность конформаций подвижных структур центров связывания лигандов. Данный метод имплементирован в качестве программного кода, написанного на языке Python 3.

Для 3D-мотивов в главе 5.3 предложена статистическая модель оценки структурной гибкости основной цепи 3D-мотивов для определения

возможности вставки данного 3D-мотива в структуру белка на примере 3D-мотивов дисульфидных мостиков.

5.1. Новый подход к анализу 3D-специфичности в структурах суперсемейства белков

5.1.1. Поиск 3D-специфических паттернов в основной и боковых цепях белков суперсемейства

5.1.1.1. Структура алгоритма поиска функционально-значимых 3D-специфических паттернов

В главе 5.1.1 описывается метод (теоретический алгоритм) выявления и ранжирования 3D-специфических паттернов, а также статистическая модель, разработанная для отделения функционально значимых 3D-специфических паттернов от случайных колебаний белковой структуры.

Входными данными для алгоритма являются:

1. множественное структурное выравнивание суперсемейства белков;
2. соответствующее структурно-опосредованное множественное выравнивание аминокислотных последовательностей белков суперсемейства (т. е. представление множественного структурного выравнивания в виде выравнивания последовательностей), так как эта информация не может быть однозначно восстановлена только из наложения структур.

Метод выявления 3D-специфических паттернов состоит из следующих этапов, подробно описанных в последующих главах:

1. На первом этапе алгоритм классифицирует столбцы структурно-опосредованного выравнивания последовательностей белков суперсемейства на две категории: «общие» столбцы и столбцы «вариабельности». «Общие» столбцы – это столбцы, аминокислотные остатки в которых входят в фрагменты основной цепи, которые являются близкими в смысле метрики RMSD для всех белков суперсемейства. Столбцы «вариабельности» – это столбцы, которые

входят в участки «вариабельности» основной цепи, то есть в такие участки основной цепи, в которых присутствует 3D-структурное разнообразие.

2. На втором этапе алгоритма участки структурного разнообразия основной и боковых цепей подаются на вход методу кластеризации машинного обучения, чтобы разделить эти участки локальной структуры на кластеры, т. е. подсемейства. В случае основной цепи на вход методу кластеризации подаются участки «вариабельности» основной цепи, а именно расстояния (RMSD) между основными цепями гомологов. В случае боковой цепи на вход методу кластеризации подаются координаты атомов боковой цепи аминокислотных остатков, входящих в «общие» столбцы выравнивания. Набор получившихся кластеров для каждого участка «вариабельности» (в случае основной цепи) и для каждого «общего» столбца (в случае боковой цепи) будем считать 3D-специфическим паттерном или *СУОЦ/СОБЦ* (случай основной цепи/случай боковой цепи). Получившееся деление суперсемейства на кластеры для каждого 3D-специфического паттерна – это деление суперсемейства на подсемейства. Для каждого 3D-специфического паттерна это деление на подсемейства может отличаться.
3. На третьем этапе рассчитывается *S-оценка* специфичности для каждого 3D-специфического паттерна и соответствующее ей ранжирование 3D-специфических паттернов. *S-оценка* показывает, насколько близки (в смысле метрики RMSD) участки цепи (основной или боковой) для представителей одного подсемейства и далеки для представителей разных подсемейств. То есть наиболее визуально заметные 3D-специфические паттерны, которые пространственно согласованы в пределах кластера/подсемейства, но далеки друг от друга между подсемействами, получают более высокие значения *S-оценки* и

занимают первое место в ранжировании для облегчения их экспертного анализа.

4. Для определения функционально-значимых 3D-специфических паттернов и отделения их от тех 3D-специфических паттернов, которые никак не влияют на функцию и возникли из-за случайных колебаний белковой структуры, была разработана статистическая модель, которая определяет, какие значения S -оценки (для данного суперсемейства белков) значимы для функционального разнообразия белков суперсемейства. По S -оценке рассчитывается Z -оценка. Наиболее высокие значения Z -оценки присваиваются участкам, в которых 3D-структурное разнообразие существенно отличается от среднего уровня случайных колебаний белковых структур.

5.1.1.2. Подробное описание алгоритма поиска функционально-значимых 3D-специфических паттернов

5.1.1.2.1. Выявление «общих» участков и участков «вариабельности» основной цепи суперсемейства белков

На первом этапе алгоритма выбираются «общие» участки основной цепи суперсемейства белков как столбцы структурно-опосредованного выравнивания аминокислотных последовательностей, содержащие (суммарно) небольшое количество гэпов и пространственно-смещенных (структурно-невыравненных) остатков (по умолчанию суммарное количество гэпов и пространственно-смещенных остатков не более, чем 5% от общего числа белков в суперсемействе). Промежутки между «общими» участками основной цепи будем называть промежутками «вариабельности» (см. рисунок 17). Необходимо учитывать не только содержание гэпов, но и содержание пространственно-смещенных остатков, так как программы для структурного выравнивания (Matt/parMATT [16,18]) могут помещать в один столбец структурно-опосредованного выравнивания пространственно-смещенные, то есть структурно-невыравненные остатки (расположенные далеко друг от друга в пространстве). Например, такое может произойти, если в

соответствующем столбце расположены аминокислоты с одинаковым названием и/или сам участок структурно-опосредованного выравнивания не содержит гэпов. Обнаружение структурно-смещенных остатков (пространственно удаленных друг от друга) в одном столбце структурно-опосредованного выравнивания – непростая задача, поскольку необходимо учитывать структурное разнообразие/пластичность/колебания белков. Обычно данная задача решается с помощью применения жесткого порогового значения, например 5 Å [18]. В нашем случае подход с таким жестким пороговым значением приводил к плохим результатам (например, приводил к значительно большей длине участков «вариабельности» и более низкому месту в ранжировании известных из литературы функционально важных участков основной цепи белков суперсемейств). Поэтому, был разработан альтернативный подход, описанный в абзаце ниже, который позволяет автоматически выбирать пороговое значение (для определения смещены ли аминокислотные остатки друг относительно друга) специально для каждого выравнивания.

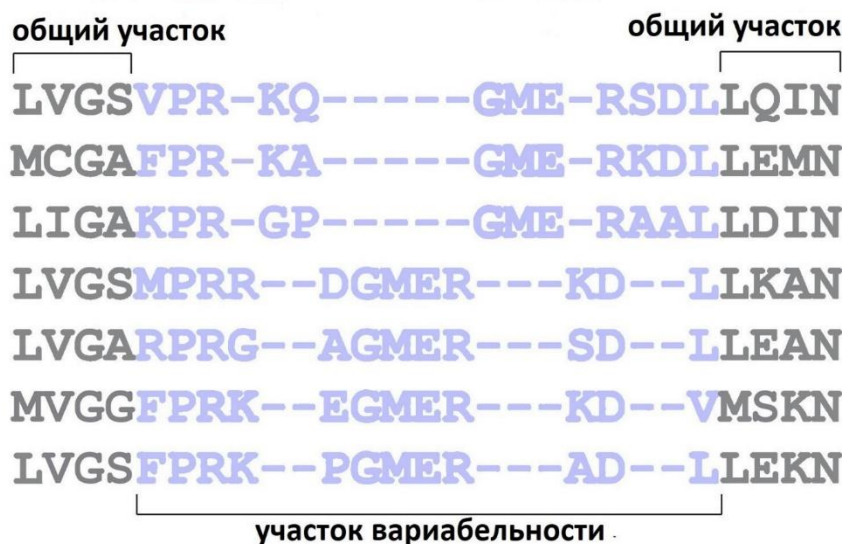


Рисунок 17. Структурно-опосредованное выравнивание последовательностей суперсемейства белков. Все столбцы выравнивания делятся на «общие» и «вариабельности» в зависимости от содержания в них гэпов и структурно-невыверенных аминокислотных остатков. Столбец считается «общим», если содержит суммарно не более 5% гэпов и пространственно-невыверенных аминокислотных остатков. Иначе столбец считается столбцом «вариабельности». «Общие» столбцы образуют «общие» участки. Столбцы «вариабельности» образуют участки «вариабельности».

Для поиска порогового значения сначала рассчитываются попарные значения RMSD между аминокислотными остатками для каждого столбца, содержащего не более 5% гэпов. На этом этапе в каждой позиции рассматриваются только тяжелые атомы основной цепи (C, C_α, N и O), а типы аминокислот и атомы боковой цепи игнорируются. Далее в каждом столбце выбирается наибольшее значение RMSD (т.е. «диаметр» столбца). Полученные значения диаметров сортируются по возрастанию и наносятся на ось ординат, а соответствующие порядковые номера столбцов наносятся на ось абсцисс. Далее к получившемуся графику применяется эвристический метод «локтя» (или «колена»), чтобы автоматически обнаружить локоть (колени) графика (аналогично работе [131]). Такая точка перегиба указывает на наиболее значительное изменение восходящего тренда метрики RMSD. Существование такого колена гарантируется условием, что входное выравнивание белков содержит как хорошо выравненные участки с малыми диаметрами, так и плохо/не выравненные участки с большим диаметром. Ордината точки перегиба берется в качестве порогового значения для различения хорошо и плохо выравненных аминокислотных остатков.

Далее это пороговое значение используется для расчета процента структурно-невыравненных остатков в каждом столбце следующим образом: если наибольшее значение попарных RMSD в столбце выше порогового значения, то аминокислотный остаток с наибольшей суммой всех попарных значений RMSD с другими остатками рассматривается как структурно-невыравненный и исключается из дальнейшего рассмотрения. Этот процесс повторяется до тех пор, пока все попарные значения RMSD между оставшимися остатками не будут ниже порогового значения. Такие аминокислотные остатки считаются структурно-выравненными. Наконец, столбцы в выравнивании последовательностей, содержащие суммарно не более 5% структурно-невыравненных остатков и гэпов, считаются «общими» и образуют «общие» участки основной цепи.

5.1.1.2.2. Разделение участков основных и боковых цепей белков суперсемейства на пространственно-эквивалентные кластеры

На втором этапе алгоритма участки структурного разнообразия основной и боковых цепей подаются на вход методу кластеризации машинного обучения, чтобы разделить эти участки локальной структуры на кластеры, т. е. подсемейства. Таким образом выявляется, являются ли эти участки 3D-специфическими паттернами.

В случае основной цепи рассматриваются участки «вариабельности». Вначале для каждого участка «вариабельности» рассчитывается матрица расстояний. Матрица расстояний рассчитывается следующим образом: между участками основной цепи попарно рассчитываются значения RMSD для всех белков суперсемейства (для каждого аминокислотного остатка рассматриваются только тяжелые атомы основной цепи – С, С_α, N и O, и метрика RMSD между участками основной цепи рассчитывается по ним). Если соответствующие отрезки имеют разную длину (то есть разное количество аминокислотных остатков), меньший из них сопоставляется с 10³ случайно выбранными подфрагментами той же длины внутри большего, а соответствующие значения усредняются. Таким образом получаем матрицу расстояний для каждого участка «вариабельности» основной цепи.

В случае боковой цепи рассматриваются «общие» позиции структурно-опосредованного выравнивания. Каждому аминокислотному остатку, принадлежащему «общей» позиции, ставится в соответствие вектор, как показано в таблице 4 и на рисунке 18 (аналогично работе [33]). Для каждого «общего» столбца выравнивания рассчитываются все парные расстояния между аминокислотными остатками, входящими в этот столбец (расстояния между векторами). Таким образом, получаем матрицу расстояний для каждой «общей» позиции выравнивания.

Таблица 4. 20 типов аминокислот, для каждой аминокислоты написаны координаты псевдоатомов (координаты начала и конца векторов), используемые для представления боковых цепей аминокислотных остатков в виде вектора.

Название аминокислоты	Начало вектора	Конец вектора
Аланин (Ala)	coordinates(CA)	coordinates(CB)
Аргинин (Arg)	coordinates(CD)	(Coordinates(NH1) + Coordinates(NH2))/2
Аспарагин (Asn)	coordinates(CB)	(coordinates(OD1)+ coordinates(ND2))/2
Аспарагиновая кислота (Asp)	coordinates(CB)	(coordinates(OD1)+ coordinates(OD2))/2
Цистеин (Cys)	coordinates(CA)	coordinates(SG)
Глутамин (Gln)	coordinates(CG)	(coordinates(NE2)+ coordinates(OE1))/2
Глутаминовая кислота (Glu)	coordinates(CG)	(coordinates(OE1)+ coordinates(OE2))/2
Глицин (Gly)	coordinates(CA)	(coordinates(N)+ coordinates(C))/2
Гистидин (His)	coordinates(CG)	(coordinates(CE1)+ coordinates(NE2))/2
Изолейцин (Ile)	coordinates(CB)	coordinates(CD1)
Лейцин (Leu)	coordinates(CB)	(coordinates(CD1)+ coordinates(CD2))/2
Лизин (Lys)	coordinates(CG)	coordinates(NZ)
Метионин (Met)	coordinates(CA)	coordinates(SD)
Фенилаланин (Phe)	coordinates(CG)	coordinates(CZ)
Пролин (Pro)	coordinates(CA)	(coordinates(CG)+ coordinates(CD))/2
Серин (Ser)	coordinates(CA)	coordinates(OG)
Треонин (Thr)	coordinates(CA)	(coordinates(CG2)+ coordinates(OG1))/2
Триптофан (Trp)	coordinates(CD1)	(coordinates(CZ2)+ coordinates(CZ3))/2
Тирозин (Tyr)	coordinates(CG)	coordinates(CZ)
Валин (Val)	coordinates(CA)	(coordinates(CG1)+ coordinates(CG2))/2

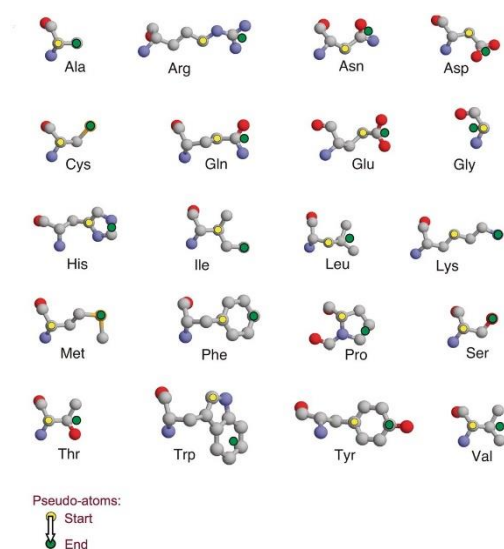


Рисунок 18. Представление боковой цепи в виде вектора, используемое в данной работе, аналогично работе [33]. 20 типов аминокислот, для каждой аминокислоты показано расположение псевдоатомов (желтые и зеленые кружки), используемое для представления боковых цепей в виде векторов (начало вектора – желтый кружок, конец вектора – зеленый кружок). Рисунок взят из [33].

Полученные матрицы расстояний (в случае основной и боковой цепей) далее подаются на вход методу кластеризации машинного обучения. Предлагается применять алгоритм кластеризации HDBSCAN [46]. Также могут быть использованы два альтернативных алгоритма (алгоритмы кластеризации OPTICS [45] и DBSCAN [44]) и другие.

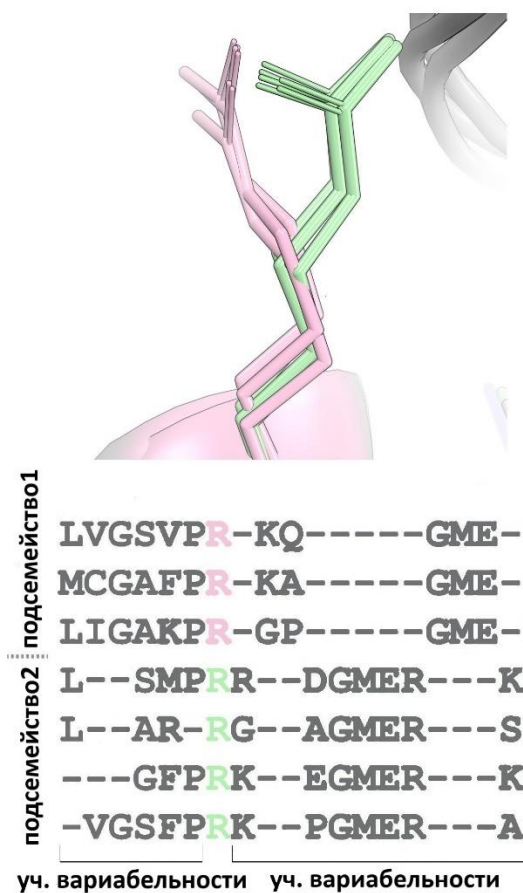


Рисунок 20. 3D-специфический паттерн, найденный в боковой цепи суперсемейства белков (СОБЦ); структурное выравнивание и соответствующее структурно-опосредованное выравнивание последовательностей.

Полученные наборы кластеров в случае, если выявлено два или более кластера, представляют собой 3D-специфические паттерны. В случае основной цепи 3D-специфический паттерн будем называть *специфическим для подсемейств участком основной цепи* (СУОЦ, рисунок 19), а в случае боковой цепи 3D-специфический паттерн будем называть *специфической для подсемейств ориентацией боковой цепи* (СОБЦ, рисунок 20).

Окончательно выбранные полученные 3D-специфические паттерны данного суперсемейства белков ранжируются в порядке убывания *S*-оценки (оценка специфичности) и *Z*-оценки (оценка статистической значимости) (ранжирования по любой из этих двух оценок эквивалентны). СУОЦы и СОБЦы ранжируются отдельно.

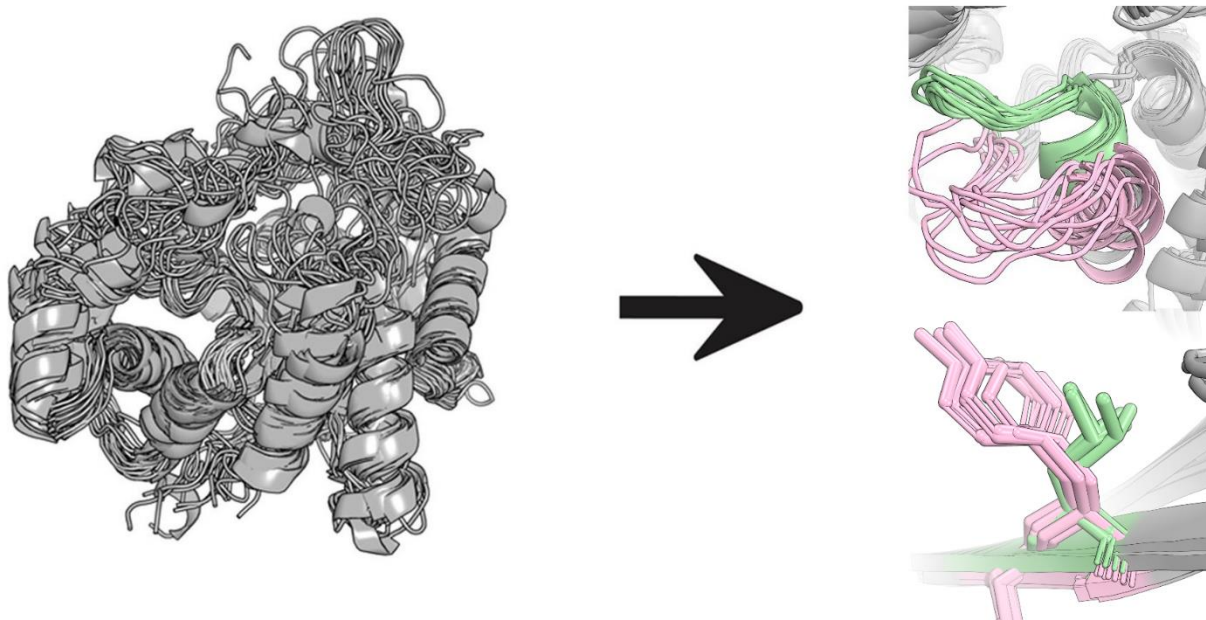


Рисунок 21. Результат поиска 3D-специфических паттернов в множественном структурном выравнивании суперсемейства белков.

5.1.1.2.3. Оценка специфичности для 3D-специфического паттерна суперсемейства белков

Оценка специфичности – S -оценка – показывает, насколько для данного 3D-специфического паттерна участки цепи (основной или боковой) внутри одного подсемейства расположены компактно по отношению к участкам из соседних подсемейств и далеко от других подсемейств. Чем выше S -оценка, тем ниже вероятность того, что данный 3D-специфический паттерн – результат случайных колебаний белковых структур. S -оценка для каждого 3D-специфического паттерна вычисляется по формуле:

$$S = Sh^{std} \times D^{std},$$

$$Sh^{std} = \frac{Sh - Sh_{min}}{Sh_{max} - Sh_{min}}$$

$$D^{std} = \frac{D - D_{min}}{D_{max} - D_{min}},$$

где Sh – значение метрики силуэт (см. главу 4.4) для данного 3D-специфического паттерна, то есть показатель того, насколько близки (в смысле

метрики RMSD) участки цепи (основной в случае СУОЦ или боковой в случае СОБЦ) внутри одного подсемейства/кластера по отношению к соседним [120]. Явно не учитывает, насколько далеко друг от друга находятся подсемейства/кластеры. Диаметр D для данного 3D-специфического паттерна – это наибольшее расстояние между любыми двумя подсемействами/кластерами этого 3D-специфического паттерна (выбросы не учитываются). Расстояние между двумя кластерами рассчитывается здесь как среднее расстояние между их элементами.

Поскольку две метрики (D и Sh) изначально рассчитываются по разным шкалам (т. е. метрика силуэт может принимать значения в диапазоне $[-1; 1]$, а диаметр измеряется в ангстремах), исходные значения Sh и D нормируются согласно приведенным выше формулам (получаются значения в диапазоне $[0; 1]$). Соответствующие коэффициенты для стандартизации (т.е. Sh_{min} и Sh_{max} , D_{min} и D_{max} – минимальное и максимальное значения метрики силуэт и диаметра) выбираются по всем 3D-специфическим паттернам соответствующего типа (СУОЦ или СОБЦ – в зависимости от того, для 3D-специфический паттерна какого типа мы вычисляем S -оценку), полученным для данного суперсемейства белков, и по всем 3D-специфическим паттернам данного типа, найденных в наборах, отобранных из базы данных PDBFlex. Получение таких наборов и нахождение 3D-специфических паттернов в них объяснено ниже в главе 5.1.1.2.4. Окончательно рассчитанная оценка специфичности S -оценка принимает значения в диапазоне $[0; 1]$. Большие значения указывают на такие 3D-специфические паттерны, которые содержат наиболее компактные и пространственно-удаленные друг от друга подсемейства/кластеры, то есть наиболее визуально заметные 3D-специфические паттерны ранжируются первыми.

5.1.1.2.4. Статистическая модель для определения функционально значимых 3D-специфических паттернов суперсемейства белков

Целью статистического анализа в предсказательной биоинформатике является различение значимых и незначительных результатов с учетом данного конкретного контекста.

В нашем случае необходимо разработать такую универсальную модель, которая сможет определить функционально-значимые 3D-специфические паттерны и отделить их по значению S -оценки от тех 3D-специфических паттернов, которые никак не влияют на функцию и возникли из-за случайных колебаний белковой структуры. Это сложная задача, так как несмотря на недавний прогресс в изучение суперсемейств белков и структурной пластичности, наше понимание взаимосвязи между структурой и функцией остается неполным. На основании информации, содержащейся в базе данных PDBeFlex, и данных, полученных с помощью молекулярного моделирования [132], можно сделать вывод, что большая часть белковых структур хотя бы в некоторой степени колеблется. В большинстве случаев нет никаких доказательств того, что эта конформационная пластичность напрямую связана с функцией белка. Те случаи, когда такие доказательства были получены из экспериментов и моделирования, обычно соответствуют структурным перестройкам с очень большой амплитудой, т.е. разницей в RMSD между конформациями выше средней [26,47,133]. Таким образом, универсальная статистическая модель была разработана на основе предположения, что средний уровень конформационной пластичности участка белковой структуры вряд ли будет иметь прямое отношение к функции.

Из базы данных PDBeFlex было отобрано 76 наборов, каждый из которых содержит разные структуры одного белка. К каждому из этих наборов был применен описанный в предыдущих главах метод для получения 3D-специфических паттернов (см. подробности в главе 4.5). Далее

рассматриваются все получившиеся для данного набора структур СУОЦы/СОБЦы (в зависимости от того, для какого типа 3D-специфического паттерна в интересующем суперсемействе мы хотим рассчитать Z -оценку), для каждого рассчитывается S -оценка. Нормирующие коэффициенты Sh_{min} и Sh_{max} , D_{min} и D_{max} в данном случае рассчитываются как максимумы и минимумы метрики силуэт и диаметра среди СУОЦов/СОБЦов, полученных во всех выбранных наборах структур базы данных PDBFlex и среди всех СУОЦов/СОБЦов, найденных в рассматриваемом суперсемействе белков. Далее в каждом наборе базы данных PDBFlex выбирается один 3D-специфический паттерн с медианным значением S -оценки. Такой выбранный 3D-специфический паттерн будем рассматривать как «случайный», т. е. как результат случайных колебаний структуры белка. Все такие 3D-специфические паттерны с медианным значением S -оценки считаем «случайными». Количество «случайных» 3D-специфических паттернов равно количеству наборов и равно 76. Мы не рассматриваем максимальное значение S -оценки, так как самые большие и наиболее заметные колебания структуры могут соответствовать функционально-значимым конформационным перестройкам (примером может послужить движение «активационной петли», которое зафиксировано в различных PDB-структурах человеческой р38а MAP-киназы [133]). По полученным медианным значениям S -оценки (то есть по S -оценкам «случайных» 3D-специфических паттернов; считаем, что S -оценка имеет стандартное нормальное распределение) рассчитываются соответствующие значения σ и μ , которые далее используются для расчета Z -оценки статистической значимости и соответствующего значения P -оценки найденного в рассматриваемом суперсемействе белков 3D-специфического паттерна (см. главу 4.6). То есть каждому найденному в данном суперсемействе белков 3D-специфическому паттерну ставится в соответствие Z -оценка, показывающая, насколько сильно данный 3D-специфический паттерн отличается от случайного.

5.1.2. Разработка программного обеспечения для поиска 3D-специфических паттернов суперсемейства

Описанный в главе 5.1.1 теоретический алгоритм был реализован в качестве программ на языке Python3 с использованием принципов ООП, речь о которых идет в следующих главах. Используемые в них библиотеки описаны в главе 4.3. Предложенные программы служат для выявления 3D-специфических паттернов в интересующих пользователя суперсемействах белков. Данные программы можно скачать по ссылкам [134,135].

5.1.2.1. Программное обеспечение для поиска 3D-специфических паттернов суперсемейства белков в основной цепи

5.1.2.1.1. Описание входных данных

Нами разработано программное обеспечение для нахождения 3D-специфических паттернов в основной цепи белков суперсемейства. Входными данными для программы являются (1) множественное структурное выравнивание белков суперсемейства, представленное в виде папки с отдельными PDB-файлами, соответствующими выравненным белкам. Каждый PDB-файл должен содержать одну цепь белка, (2) FASTA-файл с представлением множественного структурного выравнивания в виде выравнивания последовательностей, то есть со структурно-опосредованным выравниванием. Такое структурное выравнивание белков можно получить с помощью программы parMATТ [16] или с помощью веб-сервера Mustguseal [48]. Соответствующие входные данные должны выглядеть в файловом менеджере так, как показано на рисунке 22 и на рисунке 23.

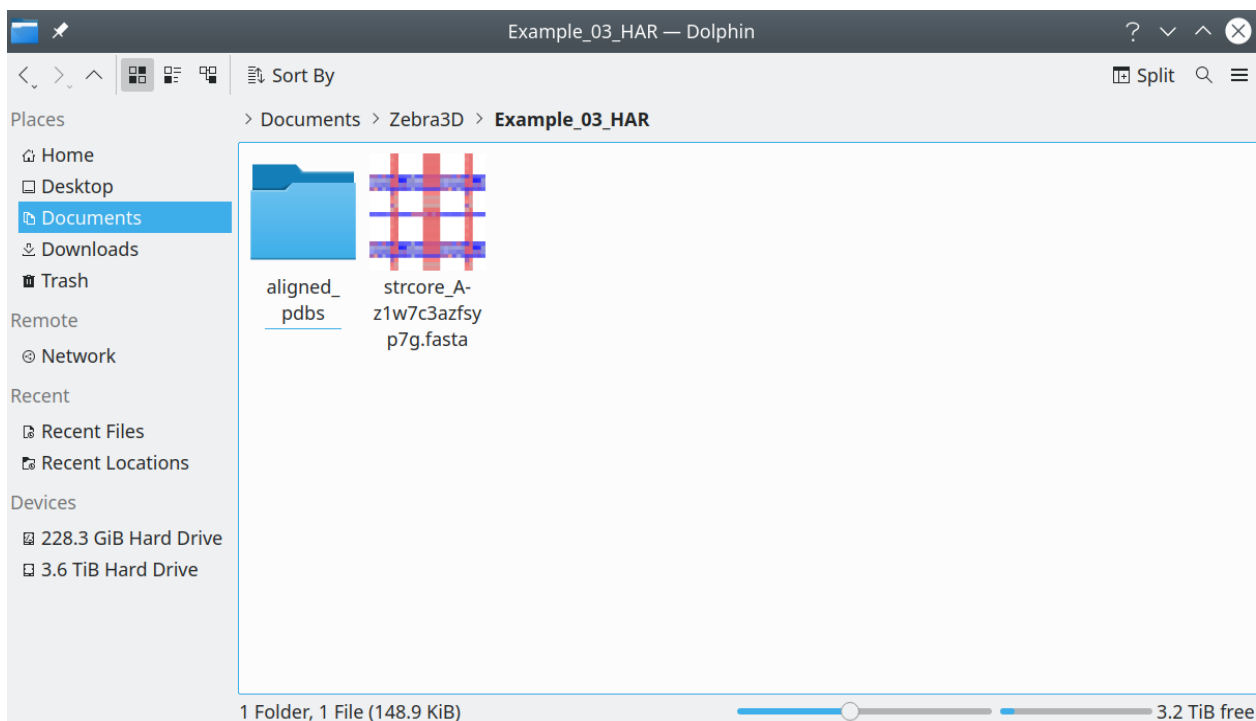


Рисунок 22. Необходимыми входными данными для программы являются 1) множественное структурное выравнивание белков суперсемейства, представленное в виде папки с отдельными PDB-файлами, соответствующими выравненным структурам белков, 2) FASTA-файл с представлением множественного структурного выравнивания в виде выравнивания последовательностей, то есть со структурно-опосредованным выравниванием.

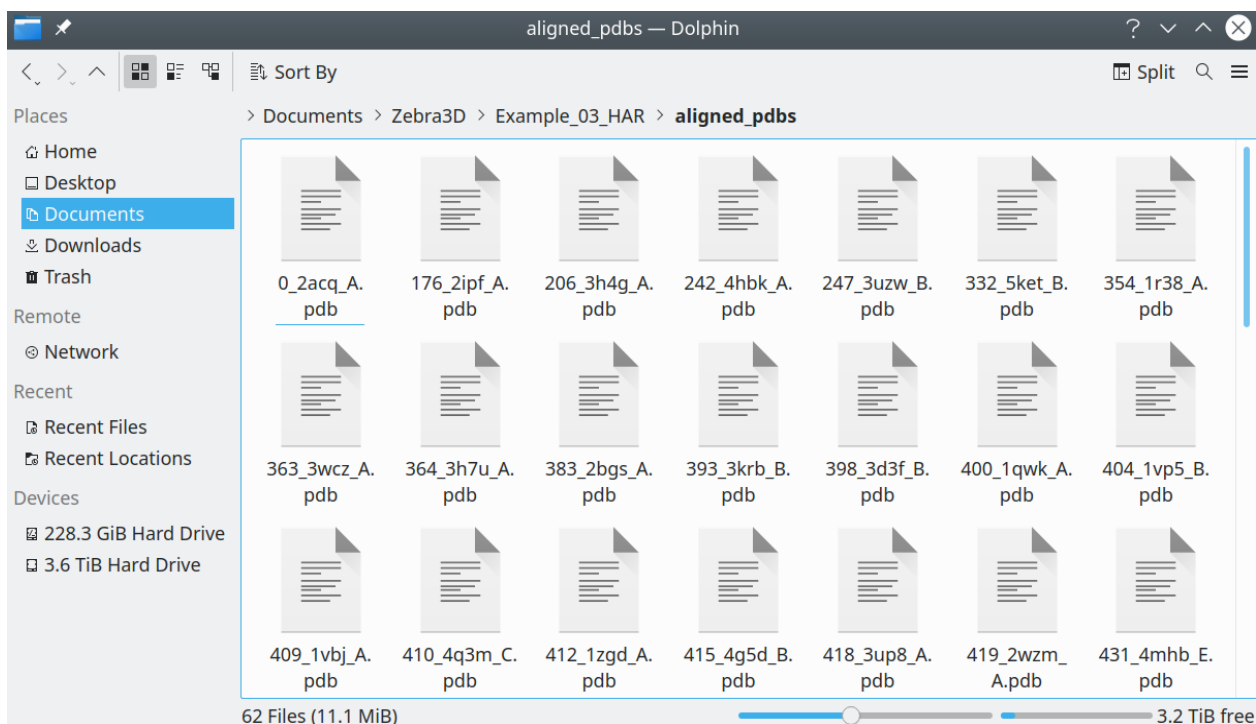


Рисунок 23. Необходимыми входными данными для программы является множественное структурное выравнивание белков суперсемейства, представленное в виде папки с отдельными PDB-файлами, соответствующими выравненным структурам белков.

Структуры белков в PDB-файлах должны быть выравнены друг с другом. Если открыть все сразу в программе PyMOL [23], окно программы должно показать биологически значимое структурное выравнивание (см. рисунок 24). Файл со структурно-опосредованным выравниванием должен быть текстовым файлом, содержащим выравнивание последовательностей белков в формате FASTA (см. рисунок 25).

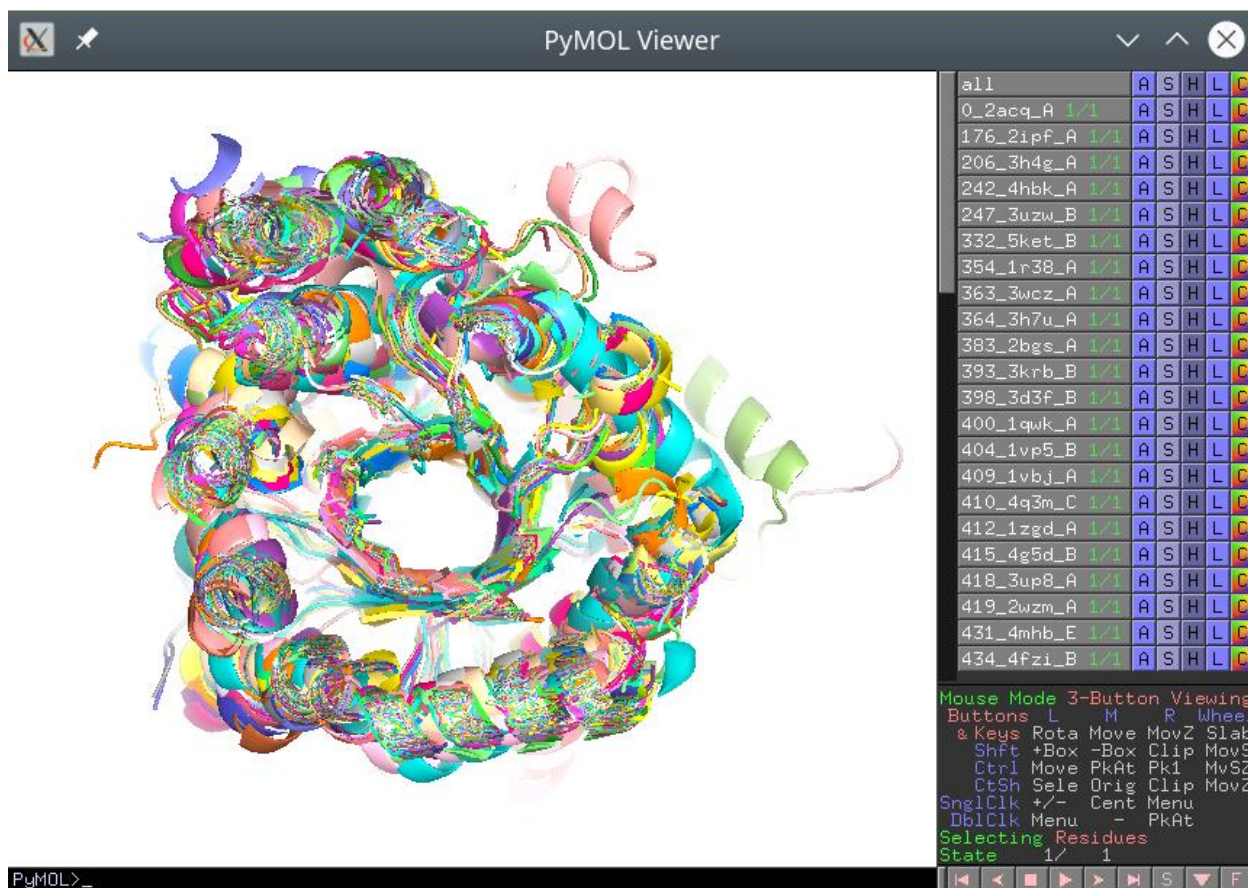


Рисунок 24. Выравненные структуры белков суперсемейства, открытые в программе PyMOL [23].

```

strcore_A-z1w7c3azf5yp7g.fasta [----] 0 L:[ 1+ 0 1/2605] *(0 /152456b) 0062 0x03E [*][X]
>0_2acq_A
-----A-----
-----SRL-LLN-NG-----
-----K-----A-----
-----MPILGLGTWKS-----P-----P-----GQVT-----
-----EAVKVAID-V-----GYRHIDC-----AHVY-----
-----Q-----NENEVGVAI-Q-EK-L-----
RE--QVV-KRE--E--L-FIVSKL-WC-----T-----
Y-HEKG-----L-----V
KGACQKTL5-DL--KLDY-----LDLYLIHWP--TGFKPG-----K--EF-F
-PLDESGNV-----V-P-----S-----DT-----
Y-D-----E-GL-V-----KAIGIS-NF-----ILD-----TWAAMEE-L-
PGL--K--Y-K-----P-AVNQI-----ECHP--Y-----LT
QE-----KLIQYQC-S--K--GI-VVTAYS-PL-GS--PDR--PW--AK
PE-DP-----S-----
-----LL-----
-----E-----
-----D-----
-----P-R-I-K--AI-A-AKH-N-K-----T-----
-----T--AQV-LIR-F-PMQR--N-----L-VVIPKSV-TP--E
R-IAENF--K-V--FD--F--E-L-SSQ--DMTT--LL-S-----YN-----R-
-N-----W--RV-CAL-----L-S-----C-T-----SHKD--Y-----
1Help 2Save 3Mark 4Replac 5Copy 6love 7Search 8Delete 9PullDn 10Quit

```

Рисунок 25. FASTA-файл, содержащий структурно-опосредованное выравнивание, открытый в текстовом редакторе.

5.1.2.1.2. Описание настраиваемых параметров

Программа позволяет настраивать параметры алгоритма с помощью ключей. Часть этих параметров являются обязательными. Их названия и описания приведены ниже.

Обязательные параметры:

1. **aligned_pdbs** – путь к папке с выравненными структурами белков суперсемейства в виде отдельных PDB-файлов (каждый файл должен содержать одну цепь белка). Пример: aligned_pdbs=./aligned_pdbs.
2. **aligned_fasta** – путь к FASTA-файлу со структурно-опосредованным выравниванием аминокислотных последовательностей белков суперсемейства. Пример: aligned_fasta=./alignment.fasta.
3. **output** – путь к папке для сохранения файлов с результатами работы программы. Пример: output=./results.

Необязательные параметры:

1. **cpu_threads** – количество используемых ядер процессора. По умолчанию используются все ядра процессора. Пример: cpu_threads=10.

2. **method** – метод кластеризации (hdbscan/optics/dbscan). Программа позволяет выбирать метод кластеризации. По умолчанию используется метод hdbscan. Пример: method=optics.
3. **eps** – обязательный параметр в случае использования метода кластеризации DBSCAN (параметр eps настраивается в случае method=dbscan). Представляет собой радиус окрестности элемента при кластеризации (более подробно см. 3.3.3.2). Пример: eps=1.
4. **min_size_of_subfamily** – параметр, регулирующий размер кластера/подсемейства в каждом СУОЦе. Реализация этого параметра зависит от выбранного метода кластеризации. Когда используется алгоритм кластеризации HDBSCAN, значение этого параметра передается параметру *min_cluster_size* этого метода кластеризации и устанавливает минимальный размер кластера (см. главу 3.3.3.4). При использовании методов OPTICS или DBSCAN значение передается параметру *minPts* (см. главу 3.3.3.2 и главу 3.3.3.3). По умолчанию значение этого параметра принимается равным 10% от общего количества белков, но не меньше 2. Пример: min_size_of_subfamily=3.
5. **max_content_of_gaps** – параметр, определяющий допустимое содержание (в процентах) гэпов в столбце множественного структурно-опосредованного выравнивания для выбора столбцов, в которых рассчитываются попарные значения RMSD между аминокислотными остатками для построения графика и применения метода «локтя» (см. главу 5.1.1.2.1). По умолчанию значение этого параметра принимается равным 5. Пример: max_content_of_gaps=5.
6. **max_content_of_mismatch**. Определяет допустимое суммарное содержание структурно-невыровненных аминокислотных остатков и гэпов в столбце выравнивания в процентах от общего количества белков суперсемейства для отнесения этого столбца к «общему» участку (см. главу 5.1.1.2.1). По умолчанию значение этого параметра принимается равным 5. Пример: max_content_of_mismatch=5.

7. **mismatch_threshold** – значение порога для отделения пространственно-выравненных остатков от невыравненных в столбце структурно-опосредованного выравнивания, в ангстремах. По умолчанию это значение выбирается автоматически по входному выравниванию с использованием эвристики «метод локтя» (см. главу 5.1.1.2.1). Пример: mismatch_threshold=5.
8. **ref** – название референсного белка, то есть белка, который наиболее интересен пользователю. Нумерация остатков в файлах результатов производится в соответствии с PDB-структурой этого белка. По умолчанию референсным белком считается первый белок в структурно-опосредованном выравнивании в FASTA-файле. Пример: ref=0_1bvt_A.
9. **max_ssr_length** – максимальная длина СУОЦа. Более длинные СУОЦы отклоняются (т.е. такие СУОЦы не ранжируются в списке результатов). Длина СУОЦа рассчитывается как максимальное значение длин кластеров, входящих в этот СУОЦ. Длина каждого кластера рассчитывается как среднее значение длин фрагментов основной цепи, входящих в этот кластер. Длина фрагмента основной цепи белка измеряется в количестве аминокислотных остатков, входящих в этот фрагмент. По умолчанию максимальная длина СУОЦа не устанавливается. Пример: max_ssr_length=20.
10. **max_outliers** – процент выбросов, полученных в результате кластеризации от общего количества белков. СУОЦы с бóльшим процентом выбросов отклоняются. По умолчанию значение этого параметра равно 100. Пример: max_outliers=40.
11. **exclude_ncterm**. СУОЦы, которые включают N- и C-концевые участки (первые или последние exclude_ncterm остатков любой PDB-структуры) отклоняются. N- и C-концевые участки структур, депонированных в PDB, могут быть неточными из-за ограничений экспериментальных методов и могут содержать очень подвижные фрагменты, не имеющие значение для функции. Поэтому включение этого фильтра может

улучшить ранжирование функционально важных СУОЦов. Например, если установлено значение `exclude_ncterm=5`, СУОЦы, содержащие первые пять или последние пять остатков любой PDB-структуры, будут отклонены (т.е. такие СУОЦы не ранжируются в списке результатов). Важно понимать, что этот фильтр постобработки применяется только к белкам, входящим в одно из подсемейств. В частности, СУОЦ не будет отклонен, если N- или C-концевой фрагмент принадлежит белку-выбросу. По умолчанию в анализ включаются N- и C-концевые участки, чтобы сохранить как можно больше данных для дальнейшего экспертного анализа. Пример: `exclude_ncterm=5`.

12. **ssr_start**. Программа позволяет проанализировать только один участок (любой на выбор пользователя) основной цепи как потенциальный СУОЦ, если указать номера первого и последнего аминокислотного остатка этого участка (нумерация производится в соответствии с PDB-структурой референсного белка). Данный параметр обозначает номер первого аминокислотного остатка. Пример: `ssr_start=50`.
13. **ssr_end**. Программа позволяет проанализировать только один участок основной цепи как потенциальный СУОЦ, если указать номера первого и последнего аминокислотного остатка этого участка (нумерация производится в соответствии с PDB-структурой референсного белка). Данный параметр обозначает номер последнего аминокислотного остатка. Параметры `ssr_start` и `ssr_end` нужно использовать вместе. Пример: `ssr_end=60`.
14. **compile_pymol_pse** – параметр, который позволяет отключить компиляцию PyMOL-сессий с 3D-аннотацией результатов, так как этот процесс при большом количестве белков может занять много времени. При необходимости компиляцию можно сделать вручную с помощью команды `pymol -qc RESULTS.py` или `pymol -qc ssr_rank.py`. По умолчанию PyMOL-сессии компилируются. Пример: `compile_pymol_pse=false`.

5.1.2.1.3. Описание вывода

Вывод программы содержит два типа результатов: список самих СУОЦов и для каждого СУОЦа его описание (в том числе классификацию белков по подсемействам). Каждый участок основной цепи оценивается независимо, поэтому подсемейство, к которому относится данный белок, может варьироваться для каждого СУОЦа. СУОЦы автоматически ранжируются в соответствии с их *S*-оценками специфичности и *Z*-оценками статистической значимости. Наиболее визуально заметные СУОЦы, которые содержат наиболее компактные и пространственно-удаленные друг от друга подсемейства/кластеры, ранжируются первыми для облегчения их экспертного анализа. Сопутствующая *Z*-оценка статистической значимости показывает, значительно ли найденный 3D-специфический паттерн отличается от случайных колебаний в структуре белка.

В случае успешного выполнения программы создаются текстовые файлы и файлы PyMOL-сессий с интуитивно понятным визуальным представлением результатов (см. рисунок 26):

- Файлы под названием RESULTS.xxx содержат краткое описание результатов;
- Файлы под названием ssr_rank.xxx содержат подробное описание каждого СУОЦа.

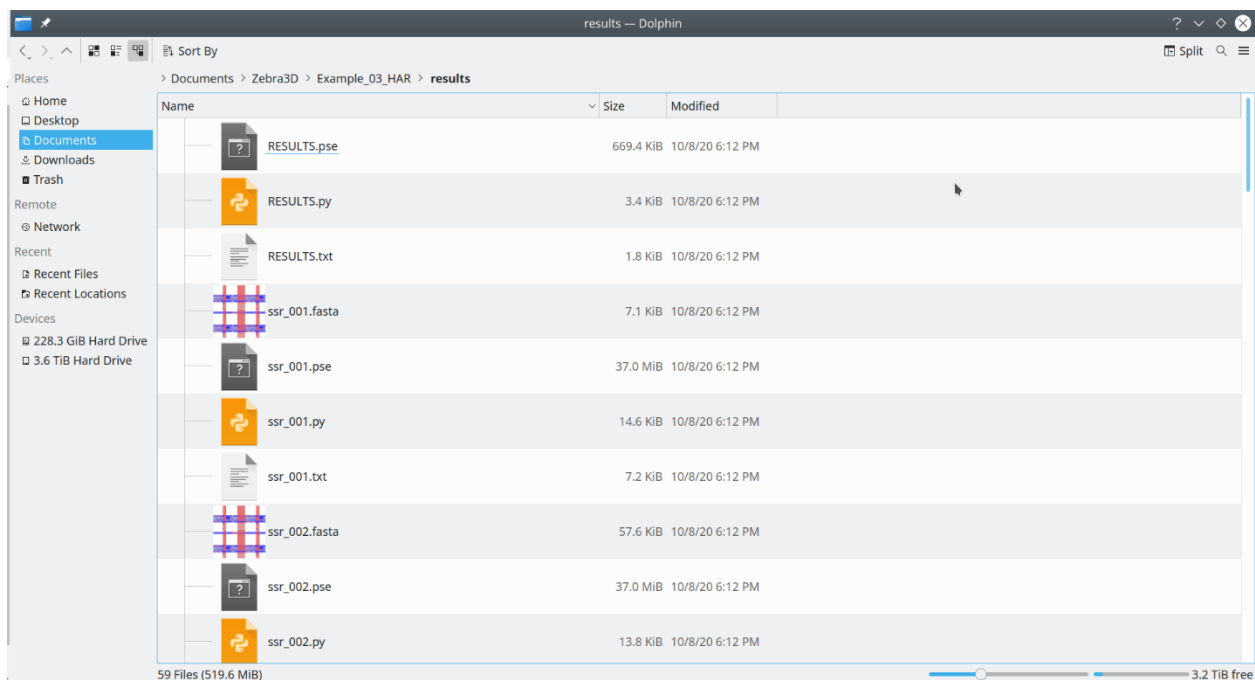


Рисунок 26. Пример множества файлов – результата работы программы для поиска 3D-специфических паттернов в основной цепи суперсемейства белков. Файлы под названием RESULTS.xxx содержат краткое описание результатов, файлы под названием ssr_rank.xxx содержат подробное описание каждого СУОЦа.

Перейдем к более подробному описанию содержания файлов – результатов работы программы.

Файл RESULTS.txt

Файл RESULTS.txt содержит следующие данные (см. рисунок 27):

- дата и время выполнения программы;
- значение всех настраиваемых параметров;
- общее количество белков в предоставленном пользователем множественном выравнивании структур;
- общее количество найденных СУОЦов.

О каждом СУОЦе в файле содержится следующая информация:

- ранг данного СУОЦа;
- количество выявленных подсемейств/кластеров;
- количество выбросов (фрагментов локальной структуры с уникальной пространственной ориентацией);

- *S*-оценка специфичности данного СУОЦа;
- *Z*-оценка статистической значимости и соответствующая *P*-оценка;
- нумерация остатков (т. е. границ данного СУОЦа) в PDB-структуре референсного белка.

```

RESULTS.txt
1 Zebra3D. Version 1.1
2 Started at 2020-10-29 21:02:46
3
4 INPUT:
5 Path to aligned protein 3D-structures in PDB format is /home/sda/Desktop/Zebra3D-1.1/Example_03_HAR/aligned_pdb
6 Path to sequence representation of 3D-structural alignment in FASTA format is /home/sda/Desktop/Zebra3D-1.1/Example_03_HAR/strcore_A-z1w7c3azf5yp7g.fasta
7 Path to output folder is /home/sda/Desktop/Zebra3D-1.1/Example_03_HAR/results
8 Number of analyzed proteins is 62
9 The reference protein is 0_2acq_A
10 Cluster analysis method is hdbscan
11 Max content of gaps in alignment column is 5.0%
12 Max amount of outliers in SSR is 100.0%
13 Max average length within subfamily in SSR is 9999 amino acids
14 Max content of mismatch in alignment column is 5.0%
15 Min size of a subfamily in SSR is 6
16 Dismiss SSRs that assign N-/C-terminal regions (first and last 5 residues of any PDB entry) to subfamilies
17 Compile PyMol PSE session files YES
18
19 Number of SSR = 12
20
21
22 SSR Rank   # of subfamilies   # of outliers   S-score   Z-score   P-value   SSR boundaries (residue numbering as in reference protein PDB)
23 1           3                   9               0.474    10.107   2.577535e-24  112-135
24 2           3                   6               0.274    5.396    3.401075e-08  47-49
25 3           3                   8               0.23     4.359    6.530818e-06  188-192
26 4           2                   5               0.207    3.817    6.764508e-05  146-153
27 5           3                   2               0.207    3.81     6.945198e-05  80-87
28 6           3                   13              0.164    2.796    2.589901e-03  170-178
29 7           3                   21              0.144    2.333    9.831204e-03  58-71
30 8           4                   14              0.102    1.339    9.031731e-02  137-137
31 9           3                   27              0.091    1.081    1.397889e-01  20-26
32 10          2                   14              0.054    0.227    4.101275e-01  36-36
33 11          2                   28              0.019    -0.613   7.299960e-01  200-200
34 12          2                   31              0.01     -0.824   7.950727e-01  98-99

```

Рисунок 27. Пример файла RESULTS.txt – результата работы программы. Файл RESULTS.txt содержит краткую информацию о всех найденных СУОЦах: ранг, количество выявленных подсемейств, количество выбросов, *S*-оценку, *Z*-оценку, *P*-оценку и границы данного СУОЦа.

Участки «вариабельности» основной цепи, в которых метод кластеризации выявил как минимум два кластера, представляют собой основной результат работы программы и представляют собой СУОЦы. Но кроме того, в выводе программы также присутствуют участки основной цепи, в которых выявлен только один кластер/подсемейство и уникальные конфигурации (выбросы). Если такие особые участки будут найдены в данном суперсемействе, им будут присвоены нулевые («N/A») *S*-оценка, *Z*-оценка, *P*-оценка, и они будут ранжированы последними отдельно от «канонических» СУОЦов (т. е. участков основной цепи с 2+ подсемействами/кластерами, рисунок 28)

SSR Rank	# of subfamilies numbering as in reference protein PDB)	# of outliers	S-score	Z-score	P-value	SSR boundaries (residue)
1	3	5	0.671	4.496	3.469928e-06	159-167
2	2	15	0.644	4.271	9.749059e-06	323-330
3	4	7	0.473	2.853	2.164103e-03	61-76
4	2	5	0.462	2.768	2.817501e-03	343-352
5	2	2	0.413	2.36	9.140093e-03	127-128
6	3	7	0.36	1.924	2.718948e-02	207-216
7	2	1	0.255	1.055	1.457848e-01	338-340
8	2	5	0.237	0.904	1.830151e-01	198-202
9	2	4	0.228	0.827	2.040079e-01	138-143
10	3	7	0.196	0.562	2.869728e-01	379-386
11	2	12	0.192	0.532	2.972298e-01	225-257
12	2	3	0.175	0.388	4.490017e-01	413-414
13	2	1	0.128	0.007	7.972144e-01	204-205
14	2	3	0.123	-0.04	5.158519e-01	113-120
15	1	3	0.092	-0.298	6.172727e-01	374-372
16	1	3	N/A	N/A	N/A	131-132
17	1	17	N/A	N/A	N/A	174-193
18	1	17	N/A	N/A	N/A	262-301
19	1	17	N/A	N/A	N/A	355-364

Рисунок 28. Пример файла RESULTS.txt, содержащего помимо «канонических» СУОЦов участки основной цепи, в которых выявлен только один кластер/подсемейство и уникальные конфигурации (выбросы). Такие участкам присваиваются нулевые («N/A») S-оценка, Z-оценка, P-оценка, и они ранжируются последними.

Файл RESULTS.pse

Файл RESULTS.pse представляет собой PyMOL-сессию, содержащую все идентифицированные СУОЦы, отмеченные в структуре референсного белка. СУОЦы названы/пронумерованы в соответствии с их рангом и окрашены пропорционально Z-оценке по градиенту от красного к серому, где интенсивный красный цвет соответствует наиболее значимым СУОЦам с высокими значениями Z-оценки, занимающим первые места, а серый цвет соответствует наименее значимым СУОЦам с низкими значениями Z-оценки (см. рисунок 29).

Файл RESULTS.py

RESULTS.py – это файл, содержащий код, который подается на вход программе PyMOL для создания файла RESULTS.pse. При желании код можно отредактировать и выполнить вручную с помощью команды: `rumol -qc RESULTS.py`.

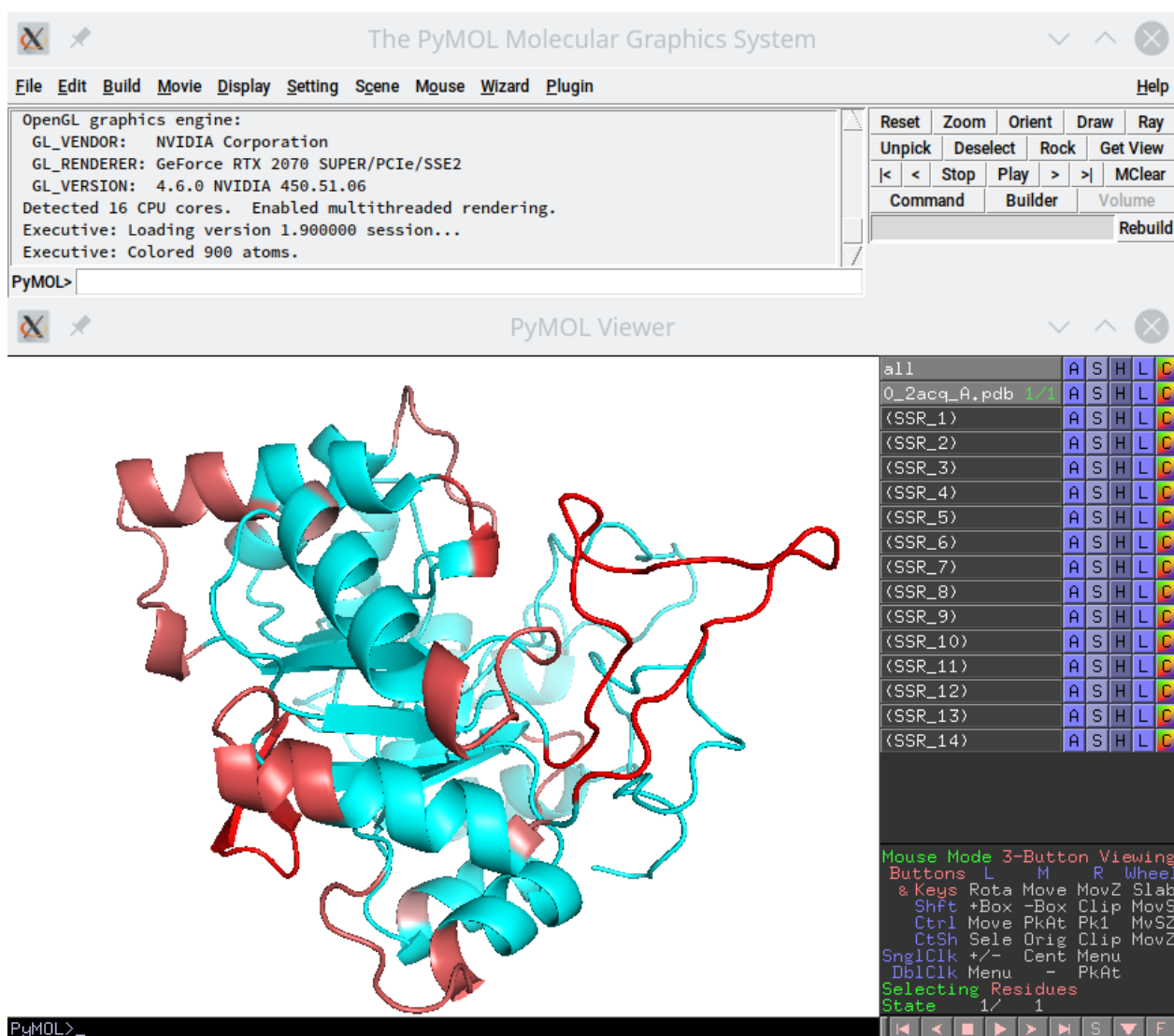


Рисунок 29. Пример PyMOL-сессии RESULTS.pse. Все найденные в суперсемействе СУОЦы отмечены в структуре референсного белка, названы/пронумерованы в соответствии с их рангом и окрашены пропорционально Z-оценке по градиенту от красного к серому, где интенсивный красный цвет соответствует наиболее значимым СУОЦам с высокими значениями Z-оценки, занимающим первые места, а серый цвет соответствует наименее значимым СУОЦам с низкими значениями Z-оценки.

Файл `ssr_rank.txt`

Характеристики каждого СУОЦа подробно описаны в серии специальных файлов `ssr_rank.txt` (см. рисунок 30). Файлы названы/пронумерованы в соответствии с рангом соответствующего СУОЦа. Каждый файл `ssr_rank.txt` представляет собой текстовый файл, содержащий следующие данные:

- Z -оценку статистической значимости, соответствующую P -оценку и S -оценку специфичности данного СУОЦа;
- ненормированное и нормированное значения силуэта и диаметра (Sh^{std} , Sh , D^{std} , D), которые использовались для расчета S -оценки (см. главу 5.1.1.2.3);
- классификацию белков суперсемейства по подсемействам – т.е. номер подсемейства/кластера, в которое был отнесен каждый белок из входного выравнивания;
- нумерацию остатков, входящих в состав данного СУОЦа, для каждого белка (как в PDB-структуре данного белка).

```

ssr_001.txt
1 Z-score = 10.108
2 P-value = 2.535267e-24
3 S-score = 0.394
4 Silhouette_score-raw = 0.54
5 Silhouette_score-std = 0.474
6 Diameter-raw = 16.414
7 Diameter-std = 0.832
8 Subfamily ID Reference Protein
9 1 * 0_2acq_A.pdb [112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129,
10 1 176_2lpf_A.pdb [119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136,
11 1 206_3h4g_A.pdb [115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132,
12 1 242_4hbk_A.pdb [112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129,
13 1 247_3uzw_B.pdb [122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139,
14 1 332_5ket_B.pdb [116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133,
15 1 354_1r38_A.pdb [116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133,
16 1 363_3wcz_A.pdb [122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134]
17 1 364_3h7u_A.pdb [116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133,
18 1 383_2bgs_A.pdb [123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140,
19 1 393_3krb_B.pdb [106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123,
20 1 400_1qwk_A.pdb [116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129]
21 1 412_1zgd_A.pdb [122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139,
22 1 435_5az0_A.pdb [139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151]
23 1 447_4tjr_A.pdb [133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150,
24 1 482_3vxx_A.pdb [127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137]
25 2 398_3d3f_B.pdb [114, 115, 116, 117]
26 2 404_1vp5_B.pdb [114, 115, 116]
27 2 409_1vbj_A.pdb [112, 113, 114, 115]
28 2 410_4q3m_C.pdb [100, 109, 110, 111]
29 2 418_3up8_A.pdb [107, 108, 109, 110, 111]
30 2 419_2vzm_A.pdb [118, 119, 120, 121, 122]
31 2 431_4mhb_E.pdb [123, 124, 125]
32 2 434_4fzi_B.pdb [113, 114, 115, 116]
33 2 439_1a80_A.pdb [110, 111, 112, 113, 114, 115]
34 2 442_3o0k_D.pdb [129, 130, 131, 132, 133]
35 2 457_3wbx_A.pdb [114, 115, 116, 117, 118, 119]
36 2 450_4mxx_A.pdb [110, 111, 112, 113, 114, 115]

```

Рисунок 30. Пример файла `ssr_rank.txt`, содержащего полную информацию о данном СУОЦе: Z -оценку, P -оценку, S -оценку, ненормированное и нормированное значения силуэта и диаметра, классификацию белков по подсемействам и нумерацию остатков. Файл назван/пронумерован в соответствии с рангом данного СУОЦа.

Файл `ssr_rank.pse`

Файл `ssr_rank.pse` представляет собой PyMOL-сессию с 3D-аннотацией конкретного СУОЦа, представленной в удобном для визуального экспертного анализа виде (см. рисунок 31). В этом файле представлены PDB-структуры всех белков суперсемейства, окрашенные в серый цвет. Соответствующие

фрагменты белковых структур, принадлежащие к данному СУОЦу, окрашены в цвет в соответствии с подсемейством (зеленый, голубой, розовый, желтый и т. д.), к которому отнесен данный белок (то есть представители различных кластеров окрашены в разные цвета). Выбросы, то есть уникальные по своей геометрии фрагменты основной цепи, принадлежащие к данному СУОЦу, окрашены в синий цвет. Все белки суперсемейства перечислены на панели объектов PyMOL и названы/пронумерованы в соответствии с номером подсемейства, к которому относятся: Subfam1_, Subfam2_, Subfam3_ и т. д. Outlier_. Референсный белок имеет дополнительный префикс Ref_.

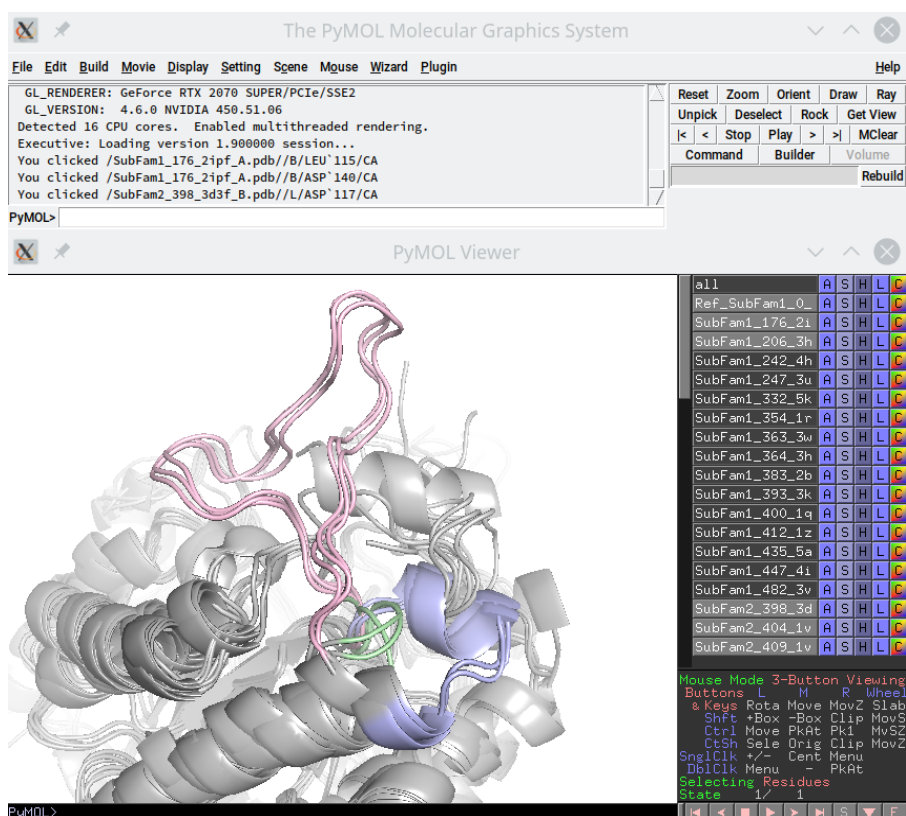


Рисунок 31. Пример файла `ssr_rank.pse`. Цветами выделен СУОЦ. Соответствующие фрагменты белковых структур, принадлежащие к данному СУОЦу, окрашены в цвет в соответствии с подсемейством (зеленый, голубой, розовый, желтый и т. д.), к которому отнесен данный белок.

Файл `ssr_rank.py`

Файл `ssr_rank.py` — это файл, содержащий скрипт, который подается на вход программе PyMOL для создания PyMOL-сессии `ssr_rank.pse`. При

желании скрипт можно отредактировать и выполнить вручную: `runol -qc ssr_rank.py`.

Файл `ssr_rank.fasta`

Наконец, файл `ssr_rank.fasta` представляет собой текстовый файл с множественным выравниванием фрагментов аминокислотных последовательностей белков суперсемейства, соответствующих данному СУОЦу.

5.1.2.1.4. Программное обеспечение для поиска 3D-специфических паттернов суперсемейства белков в боковой цепи

Описанное в этой главе программное обеспечение реализует алгоритм, описанный в главе 5.1.1, и позволяет находить 3D-специфические паттерны в боковых цепях в интересующем пользователя суперсемействе белков.

5.1.2.1.5. Описание входных данных

Входными данными для данной программы, как и для программы, описанной в главе 5.1.2.1, являются (1) множественное структурное выравнивание белков суперсемейства, представленное в виде папки с отдельными PDB-файлами, соответствующими выравненным белкам, (2) FASTA-файл с представлением множественного структурного выравнивания в виде выравнивания последовательностей, то есть со структурно-опосредованным выравниванием. Такое структурное выравнивание белков можно получить с помощью программы `ragMAT` [16] или с помощью веб-сервера `Mustguseal` [48].

5.1.2.1.6. Описание настраиваемых параметров

Программа позволяет настраивать параметры алгоритма с помощью ключей. Следующие параметры по названию, смыслу и описанию полностью совпадают с такими же параметрами из главы 5.1.2.1.2: **`aligned_pdbs`**,

aligned_fasta, output, cpu_threads, method, eps, max_content_of_gaps, max_content_of_mismatch, mismatch_threshold, compile_pymol_pse.

Помимо перечисленных, программа позволяет настраивать следующие параметры:

1. **minpts**. В случае, когда `method=dbscan/optics`, значение этого параметра передается параметру *minPts* используемого алгоритма кластеризации (см. главу 3.3.3.2 и главу 3.3.3.3). В случае `method=hdbscan` значение этого параметра передается параметру алгоритма *min_samples* (см. главу 3.3.3.4) и по умолчанию равно `None`, в остальных случаях принимается равным 10% от общего количества белков, но не меньше 2. Пример: `minpts=3`.
2. **min_cluster_size** – параметр, регулирующий размер кластера/подсемейства в каждом СОБЦе. Данный параметр настраивается в случае `method=hdbscan` и соответствует одноименному параметру алгоритма кластеризации HDBSCAN (см. главу 3.3.3.4). Значение по умолчанию принимается равным 10% от общего количества белков, но не меньше 2. Пример: `min_cluster_size=3`.
3. **number_of_result_resids** – параметр, регулирующий количество СОБЦов, показываемых пользователю программы в файлах вывода программы. Если программа нашла больше СОБЦов, чем `number_of_result_resids`, пользователю будет показано `number_of_result_resids` СОБЦов, имеющих наивысшие Z-оценки. Значение по умолчанию равно 10. Пример: `number_of_result_resids=3`.

5.1.2.1.7. Описание вывода

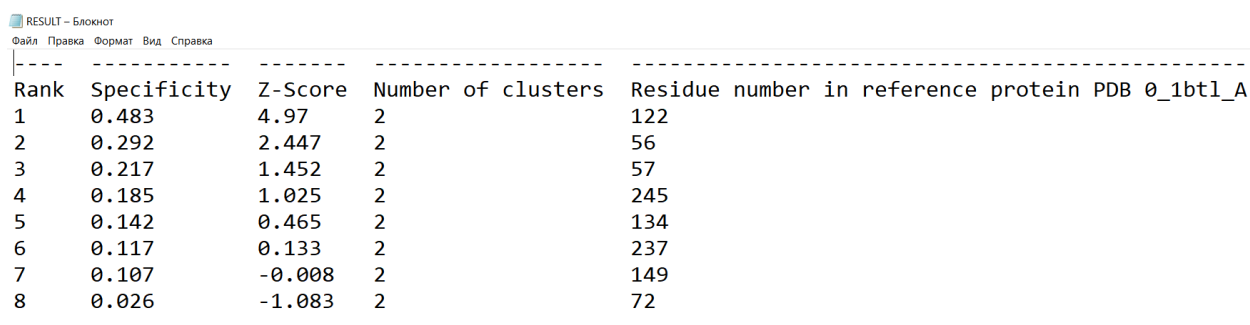
Результатом работы программы являются файлы, описывающие найденные СОБЦы.

Файл RESULT.txt

Файл RESULT.txt содержит общую информацию о найденных программой СОБЦах (см. рисунок 32). О каждом СОБЦе в файле содержится следующая информация:

- ранг данного СОБЦа;
- количество выявленных подсемейств/кластеров;
- *S*-оценка специфичности данного СОБЦа;
- *Z*-оценка статистической значимости;
- номер остатка, соответствующий данному СОБЦу, в PDB-структуре референсного белка.

Как и в случае программы, описанной в главе 5.1.2.1, в список результатов включаются не только «канонические» СОБЦы, но и такие результаты кластеризации, в которых выявлен только один кластер/подсемейство и уникальные конфигурации (выбросы). Если такие особые боковые цепи аминокислотных остатков будут найдены в данном суперсемействе, им будут присвоены нулевые («N/A») *S*-оценка, *Z*-оценка, *P*-оценка и они будут ранжированы последними отдельно от «каноничных» СОБЦов (т. е. участков боковой цепи с 2+ подсемействами/кластерами).

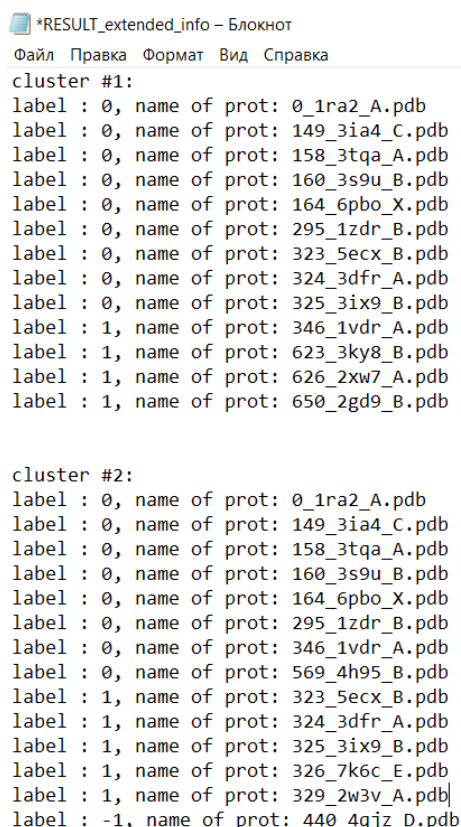


Rank	Specificity	Z-Score	Number of clusters	Residue number in reference protein PDB 0_1btl_A
1	0.483	4.97	2	122
2	0.292	2.447	2	56
3	0.217	1.452	2	57
4	0.185	1.025	2	245
5	0.142	0.465	2	134
6	0.117	0.133	2	237
7	0.107	-0.008	2	149
8	0.026	-1.083	2	72

Рисунок 32. Пример файла RESULT.txt. Файл RESULT.txt содержит общую информацию о каждом из найденных программой СОБЦе: ранг, *S*-оценку, *Z*-оценку, количество выявленных кластеров, номер остатка в PDB-структуре референсного белка.

Файл RESULT extended info.txt

В данном файле содержится более подробная информация о найденных СОБЦах, а именно, для каждого СОБЦа указывается, какие белки суперсемейства к каким кластерам/подсемействам отнесены (см. рисунок 33).



```
*RESULT_extended_info - Блокнот
Файл  Правка  Формат  Вид  Справка
cluster #1:
label : 0, name of prot: 0_1ra2_A.pdb
label : 0, name of prot: 149_3ia4_C.pdb
label : 0, name of prot: 158_3tqa_A.pdb
label : 0, name of prot: 160_3s9u_B.pdb
label : 0, name of prot: 164_6pbo_X.pdb
label : 0, name of prot: 295_1zdr_B.pdb
label : 0, name of prot: 323_5ecx_B.pdb
label : 0, name of prot: 324_3dfr_A.pdb
label : 0, name of prot: 325_3ix9_B.pdb
label : 1, name of prot: 346_1vdr_A.pdb
label : 1, name of prot: 623_3ky8_B.pdb
label : 1, name of prot: 626_2xw7_A.pdb
label : 1, name of prot: 650_2gd9_B.pdb

cluster #2:
label : 0, name of prot: 0_1ra2_A.pdb
label : 0, name of prot: 149_3ia4_C.pdb
label : 0, name of prot: 158_3tqa_A.pdb
label : 0, name of prot: 160_3s9u_B.pdb
label : 0, name of prot: 164_6pbo_X.pdb
label : 0, name of prot: 295_1zdr_B.pdb
label : 0, name of prot: 346_1vdr_A.pdb
label : 0, name of prot: 569_4h95_B.pdb
label : 1, name of prot: 323_5ecx_B.pdb
label : 1, name of prot: 324_3dfr_A.pdb
label : 1, name of prot: 325_3ix9_B.pdb
label : 1, name of prot: 326_7k6c_E.pdb
label : 1, name of prot: 329_2w3v_A.pdb
label : -1, name of prot: 440_4qjz_D.pdb
```

Рисунок 33. Пример файла RESULT_extended_info.txt. Для каждого СОБЦа указывается, какие белки суперсемейства к каким кластерам/подсемействам отнесены.

Файл RESULT.pse

Файл RESULT.pse представляет собой PyMOL-сессию с 3D-аннотацией всех найденных СОБЦов, представленной в удобном для визуального экспертного анализа виде (см. рисунок 34). В этом файле представлены PDB-структуры всех белков суперсемейства, окрашенные в серый цвет. Соответствующие фрагменты белковых структур, принадлежащие СОБЦам, окрашены в цвет в соответствии с подсемейством (зеленый, голубой, розовый, желтый и т. д.), к которому отнесен данный белок для данного СОБЦа. Выбросы, то есть уникальные по своей геометрии фрагменты основной цепи,

окрашены в синий цвет. Референсный белок на панели объектов PyMOL имеет дополнительный префикс Ref_.

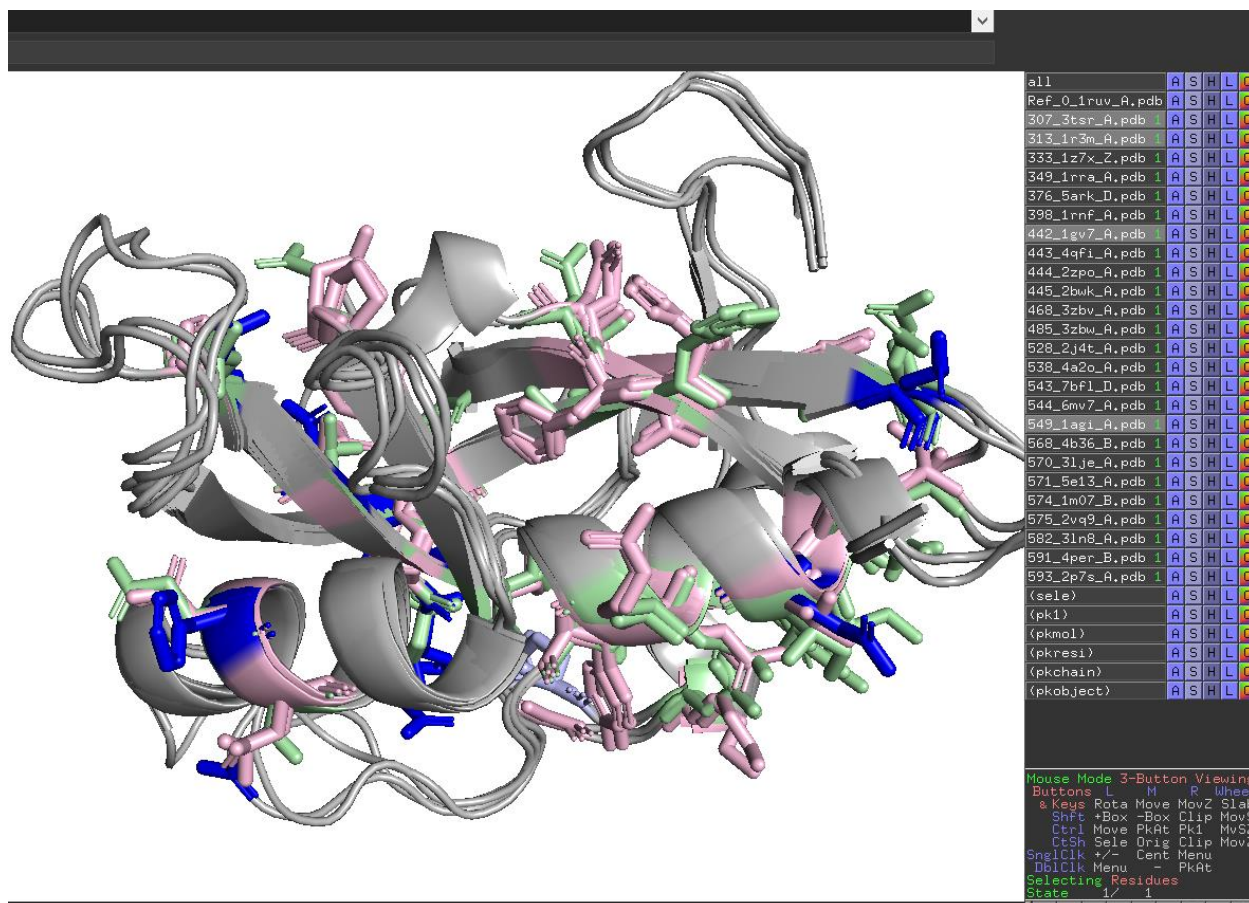


Рисунок 34. Пример файла RESULT.pse. Цветами выделены СОБЦы. Соответствующие фрагменты белковых структур, принадлежащие к данному СОБЦу, окрашены в цвет в соответствии с подсемейством (зеленый, голубой, розовый, желтый и т. д.), к которому отнесен данный белок для данного СОБЦа.

Файл RESULT.py

Файл RESULT.py – это файл, содержащий скрипт, который подается на вход программе PyMOL для создания PyMOL-сессии RESULT.pse. При желании скрипт можно отредактировать и выполнить вручную: `pymol -qc RESULT.py`.

5.2. Апробация нового подхода на широкой выборке суперсемейств белков

Разработанный нами метод выявления 3D-специфических паттернов (см. раздел 5.1.1) был апробирован на суперсемействах белков. Создание выборок для апробации описано в главе 4.7. На данных выборках были

запущены программы, описанные в разделе 5.1.2. Параметры, с которыми были запущены программы, описаны в главе 4.7. Таким образом были найдены 3D-специфические паттерны в данных суперсемействах. Анализ полученных 3D-специфических паттернов был проведен нами с целью проверки их роли и соответствия каким-либо функционально значимым элементам структуры белков суперсемейства, определенным независимо и известным из литературы. Выяснилось, что участки структуры ферментов, соответствующие найденным 3D-специфическим паттернам, играют важную роль в механизме действия, в том числе в связывании субстрата и его доставке в активный центр, определяют субстратную специфичность и каталитическую активность.

3D-специфические паттерны могут отвечать:

- За различие в свойствах между ферментами, принадлежащими различным функциональным подсемействам. То есть геометрия участков структуры, представляющих 3D-специфические паттерны, отличается у представителей различных функциональных подсемейств и отвечает за различие в функциональных свойствах между подсемействами. Разделение ферментов суперсемейства на кластеры/подсемейства для такого 3D-специфического паттерна соответствует классификации ферментов по функциональным подсемействам.
- За различные положения участка структуры фермента, важные для его функциональных свойств и отличающиеся в различных функционально-значимых конформациях *одного* (в этом отличие от предыдущего пункта) фермента (т.е. такие 3D-специфические паттерны являются *детерминантами конформационной вариабельности* центров связывания лигандов). То есть геометрия участков структуры, представляющих 3D-специфические паттерны, отличается в различных конформациях фермента, например, в присутствии или отсутствии

связанного лиганда, в активной и неактивной форме фермента или может зависеть от химической природы связанного лиганда. Разделение ферментов суперсемейства на кластеры/подсемейства для такого 3D-специфического паттерна соответствует разделению на группы PDB-файлов ферментов с различными структурными положениями данного участка.

Все это делает поиск и изучение 3D-специфических паттернов важным элементом новых подходов к дизайну улучшенных биокатализаторов и поиску новых лекарств. В следующих главах 5.2.1, 5.2.2 подробно описаны все исследованные нами суперсемейства белков, найденные в них функционально значимые 3D-специфические паттерны и полученное распределение белков суперсемейства по подсемействам.

5.2.1. 3D-специфические паттерны, отвечающие за различие в свойствах между ферментами, принадлежащими различным функциональным подсемействам

В этой главе приведены примеры проведенного анализа по поиску взаимосвязи и роли определенных нами 3D-специфических паттернов в проявлении различных функциональных свойств ферментов, принадлежащих различным функциональным подсемействам.

- *Суперсемейство пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α -цилиндра.* Ферменты этого суперсемейства присутствуют в большинстве организмов и катализируют декарбоксилирование различных субстратов, необходимых для биосинтеза полиаминов и лизина. Было рассмотрено множественное структурное выравнивание пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α -цилиндра. Идентифицированный СУОЦ №8 (из 23 определенных 3D-специфических паттернов – нумерация приведена в соответствии с ранжированием) представляет собой ранее описанную в литературе [136,137] 3_{10} -спираль, которая расположена на одной

стороне полости связывания субстрата и принимает альтернативные ориентации в подсемействах, соответствующих ферментам с различной субстратной специфичностью. Автоматически полученная для данного 3D-специфического паттерна кластеризация суперсемейства соответствует разделению ферментов по субстратной специфичности. Первый из трех полученных кластеров ферментов, соответствует орнитин-декарбоксилазам, второй – диаминопимелат-декарбоксилазам. Орнитин-декарбоксилазы связывают относительно короткий субстрат (L-орнитин). Гомологичные диаминопимелат-декарбоксилазы могут связывать субстраты большего размера благодаря смещению 3_{10} -спирали, что приводит к высвобождению дополнительного пространства в участке связывания субстрата (см. рисунок 35). Размеры связывающей полости между кофактором пиридоксальфосфатом и СУОЦ №8, по-видимому, служат ключевым структурным фактором для различения предпочтений субстратов у этих гомологов [136,137]. Найденный 3D-специфический паттерн изображен на рисунке 36.

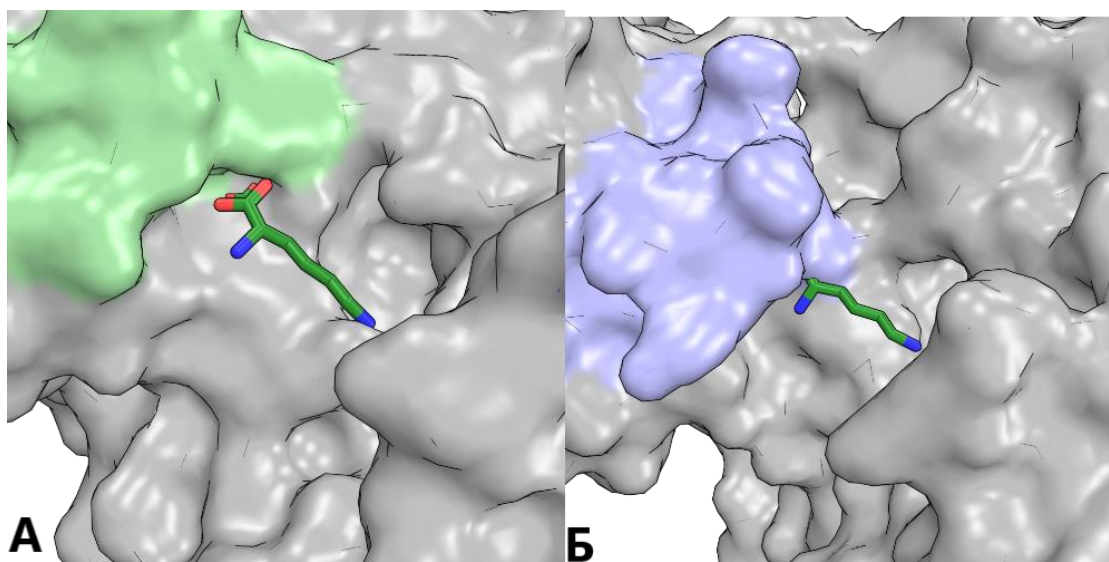


Рисунок 35. Представитель диаминопимелат-декарбоксилаз (PDB 6knh), рисунок А. Представитель орнитин-декарбоксилаз (PDB 7odc), рисунок Б. Цветом на обоих рисунках выделены 3_{10} -спираль и субстрат (лизин – заменяет L-орнитин). Орнитин-декарбоксилазы связывают относительно короткий субстрат (L-орнитин). Гомологичные диаминопимелат-декарбоксилазы могут связывать субстраты большего размера – благодаря смещению 3_{10} -спирали, что приводит к высвобождению дополнительного пространства в участке связывания субстрата.

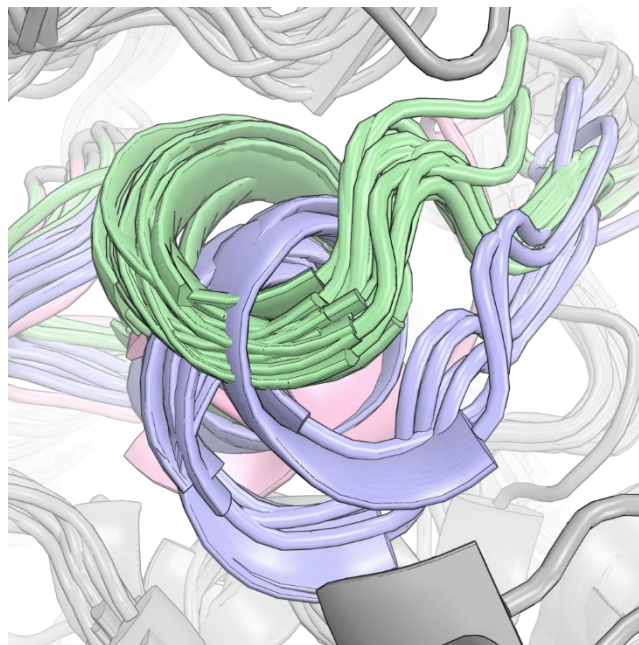


Рисунок 36. Множественное структурное выравнивание суперсемейства пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α -цилиндра. Цветом выделен СУОЦ №8 (из 23), который представляет 3_{10} -спираль, которая расположена на одной стороне полости связывания субстрата и принимает альтернативные ориентации в подсемействах, соответствующих ферментам с различной субстратной специфичностью. Полученные при кластеризации подсемейства ферментов обозначены разными цветами: «голубой» кластер соответствует орнитин-декарбоксилазам, «зеленый» - диаминопимелат-декарбоксилазам. Представители различных подсемейств связывают субстраты разных размеров. На этом рисунке, как и на рисунках 37-51, для ясности изображены несколько представителей суперсемейства.

- *Альдо-кеторедуктазы.* Альдо-кеторедуктазы (AKR) составляют большое и разнообразное семейство оксидоредуктазных ферментов и встречаются почти у каждого вида. Наиболее изученными представителями этого семейства являются АКР млекопитающих, участвующие в метаболизме гормонов: стероидов и простагландинов. В исследовании альдо-кеторедуктаз был идентифицирован СУОЦ №1 (из 14). Он представляет собой подвижный участок, расположенный наверху канонической $(\alpha/\beta)_8$ -цилиндрической структуры и соответствующий петле А [138], которая значительно различается между гомологами с различной субстратной специфичностью и участвует в связывании субстрата. Введение соответствующей гибкой петли из альдозоредуктазы человека (окрашена в розовый цвет на

рисунке 37) вместо соответствующего фрагмента (который эквивалентен участку, выделенным зеленым цветом на рисунке 37) в структуре гипертермостабильной алкогольдегидрогеназы D из *Pyrococcus furiosus* было одним из ключевых шагов в создании химеры, которая проявляла субстратную специфичность донорного фермента и унаследовала термостабильность родительского фермента [138]. Полученные нами кластеры соответствуют ферментам с различной субстратной специфичностью. Найденный 3D-специфический паттерн изображен на рисунке 37.

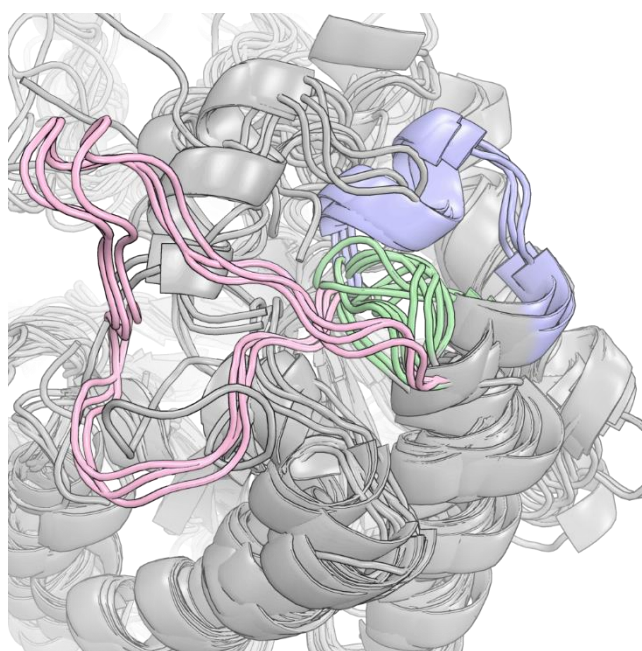


Рисунок 37. Множественное структурное выравнивание альдо-кеторедуктаз. Цветом выделен СУОЦ №1 (из 14), который представляет собой подвижный участок, соответствующий петле А, которая значительно различается между гомологами с разной субстратной специфичностью и участвует в связывании субстрата. Введение соответствующей гибкой петли из альдозоредуктазы человека (окрашена в розовый цвет на рисунке) вместо соответствующего фрагмента в структуре гипертермостабильной алкогольдегидрогеназы D из *Pyrococcus furiosus* (который эквивалентен участку, выделенным зеленым цветом на рисунке) было одним из ключевых шагов в создании химеры, которая проявляла субстратную специфичность донорного фермента и унаследовала термостабильность родительского фермента [138].

- *Гомологи полиэфиргидролазы из Pseudomonas aestusnigri.* Полиэфиргидролаза из *Pseudomonas aestusnigri* относится к типу Па ПЭТ-гидролаз и гидролизует полиэтилентерефталат. При анализе множественного структурного выравнивания гомологов

полиэфиргидролазы из *Pseudomonas aestusnigri* нами были идентифицированы СУОЦы № 5, 8 (из 13) – структурные фрагменты субстрат-связывающего участка активного центра [139]. Автоматически полученная для СУОЦа № 5 кластеризация суперсемейства соответствует разделению ферментов по субстратной специфичности. Было получено два кластера, первому из которых принадлежат ПЭТ-гидролазы и близкородственные промискуитетные кутиназы, а второму – эстеразы, которые не обладают ПЭТ-гидролазной активностью и превращают другие сложные эфиры. Кластеризация белков по подсемействам суперсемейства для СУОЦа №8 дала результаты, аналогичные тем, которые получились для СУОЦа №5. Найденные 3D-специфические паттерны представлены на рисунке 38.

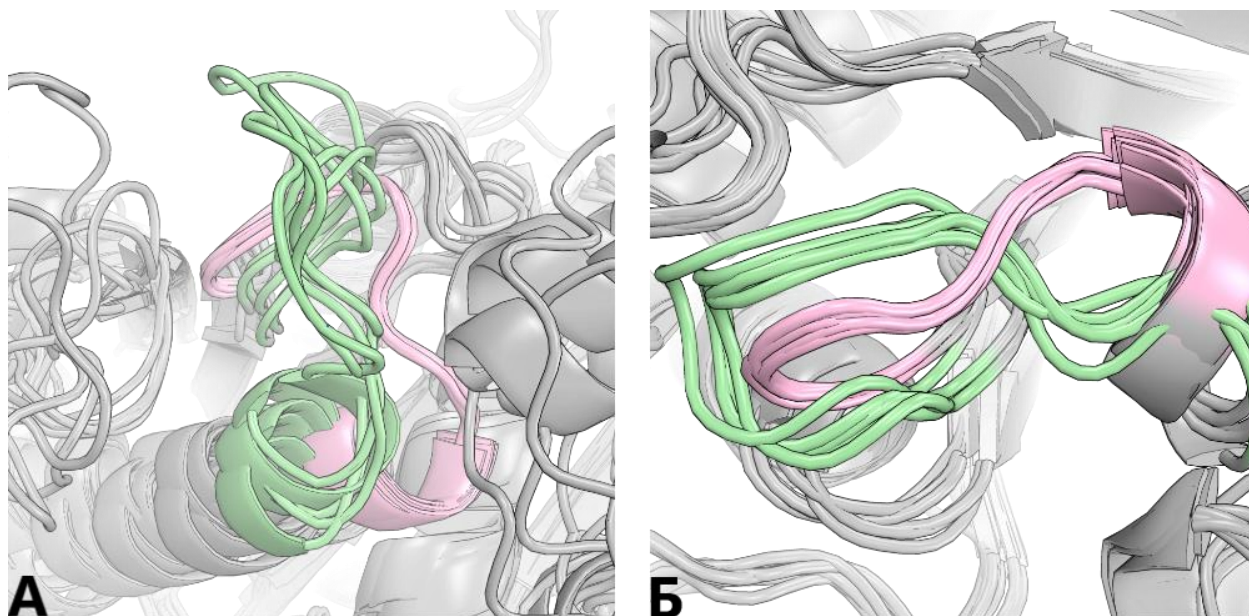


Рисунок 38. Множественное структурное выравнивание гомологов полиэфиргидролазы из *Pseudomonas aestusnigri*. На рисунке А цветом выделен СУОЦ № 5 (из 13). На рисунке Б цветом выделен СУОЦ № 8 (из 13). На обоих рисунках полученные при кластеризации подсемейства обозначены разными цветами. Полученный «розовый» кластер соответствует ПЭТ-гидролазам и близкородственным промискуитетным кутиназам, а «зеленый» кластер соответствует подсемейству, в который входят ферменты, являющиеся гидролазами других сложных эфиров.

- *Металло-зависимые гидролазы.* Нами были исследовано множественное структурное выравнивание металло-зависимых гидролаз. Найденный

СУОЦ №7 (из 22) представляет собой петлю, распознающую субстрат. Мутации этого участка (укорачивание) в гуаниндеаминазе человека привели к тому, что полученный фермент более активен в отношении аммида (меньшего по размеру субстрата) и менее активен в отношении гуанина (большего по размеру субстрата), чем фермент дикого типа, так как карман для связывания субстрата уменьшился. Относительное изменение специфичности составляет $2,5 \cdot 10^6$ раз [140]. Полученные для данного 3D-специфического паттерна кластеры соответствуют металло-зависимым гидролазам с различной субстратной специфичностью. К первому кластеру относятся ферменты с субстратной специфичностью к моноциклическим азотсодержащим гетероциклам (меньшим по размеру субстратам), а ко второму – к бициклическим (большим по размеру субстратам). Найденный 3D-специфический паттерн изображен на рисунке 39.

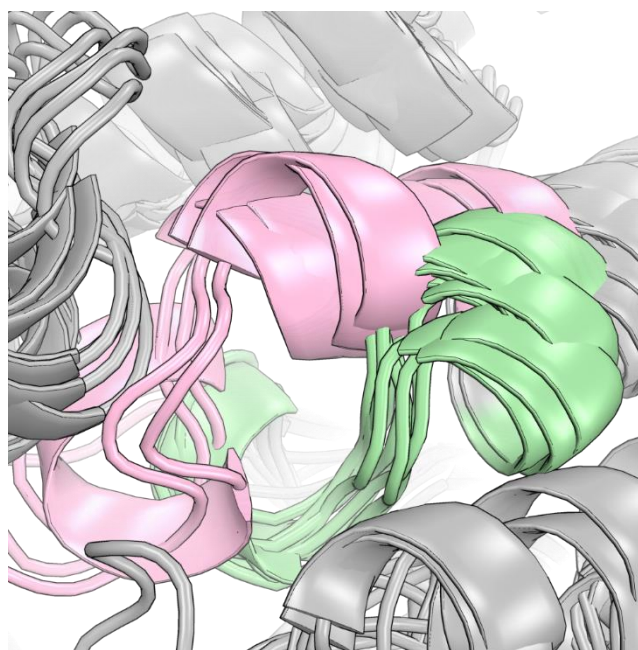


Рисунок 39. Множественное структурное выравнивание металло-зависимых гидролаз. Цветом выделен СУОЦ №7 (из 22). На рисунке полученные при кластеризации подсемейства обозначены разными цветами. К подсемейству, обозначенному розовым цветом, относятся ферменты с субстратной специфичностью к моноциклическим азотсодержащим гетероциклам (меньшим по размеру субстратам), а к подсемейству, обозначенному зеленым цветом, относятся ферменты с субстратной специфичностью к бициклическим азотсодержащим гетероциклам (большим по размеру субстратам).

- *Суперсемейство нейраминидаз GH34.* Нами было изучено множественное структурное выравнивание нейраминидаз GH34 – суперсемейства ферментов, входящих в состав оболочек вируса гриппа А и В. Нейраминидазы специфически отщепляют остаток сиаловой кислоты от полисахаридов мембраны эритроцита, тем самым разрушая рецепторы к вирусу на клетках организма-хозяина. Нейраминидазы N1 и N2 штаммов вируса гриппа А, циркулирующих в человеческой популяции в настоящий момент, принадлежат к двум разным филогенетическим группам. Первая группа включает подтипы N1, N4, N5 и N8, вторая – N2, N3, N6, N7 и N9 (все эти подтипы относятся к нейраминидазам гриппа А). Для нейраминидаз вируса гриппа из литературы известны две функционально важные полости: «полость-430» и «полость-150», обозначаемые так в соответствии с нумерацией петель, образующих эти участки («петля-430» и «петля-150»). На сегодняшний день «полость-430» изучена достаточно поверхностно, хотя предыдущие исследования показали, что «полость-430» может быть одним из промежуточных сайтов связывания субстрата на пути к каталитическому центру, что делает ее многообещающей мишенью для создания новых лекарств с целью подавления активности нейраминидаз N1-N9 [141–143]. «Петля-150» в кристаллографических структурах апо-форм нейраминидаз первой филогенетической группы ориентирована наружу от активного центра и образует дополнительную «полость-150». В присутствии лиганда в активном центре «петля-150» взаимодействует с ним и принимает конформацию, аналогичную той, что наблюдается у ферментов второй группы, с лигандом или без. Отсюда наиболее вероятно следует, что лиганд (субстрат/ингибитор) связывается с нейраминидазой в открытой конформации «петли-150», после чего происходит медленная конформационная перестройка – петля переходит в закрытую конформацию, образуя взаимодействия с

молекулой лиганда. Было предложено, таким образом, использовать «полость-150» для дизайна новых ингибиторов, селективных к нейраминидазам первой группы [144–147]. Нами в множественном структурном выравнивании суперсемейства нейраминидаз GH34 были найдены СУОЦы, соответствующие «петле-430» и «петле-150». СУОЦ №3 (из 19) соответствует «петле-430». Автоматически полученная кластеризация нейраминидаз GH3 для СУОЦа №3 соответствует разделению ферментов по каталитической активности. Нейраминидазы вируса гриппа N1-N9 были разделены на три подсемейства в соответствии с их филогенетической классификацией. Нейраминидаза-подобные белки (гомологи нейраминидаз) N10 и N11 летучих мышей, которые лишены нейраминидазной активности [148], а также гомологи из менее патогенного вируса гриппа В, были отнесены к другим двум подсемействам; в обоих случаях полость-430 либо отсутствовала, либо была представлена в редуцированном виде. СУОЦ №10 соответствует «петле-150». Кластеризация белков по подсемействам нейраминидаз вируса гриппа для СУОЦа №10 дала результаты, аналогичные тем, которые получились для СУОЦа №3. Найденные 3D-специфические паттерны изображены на рисунке 40.

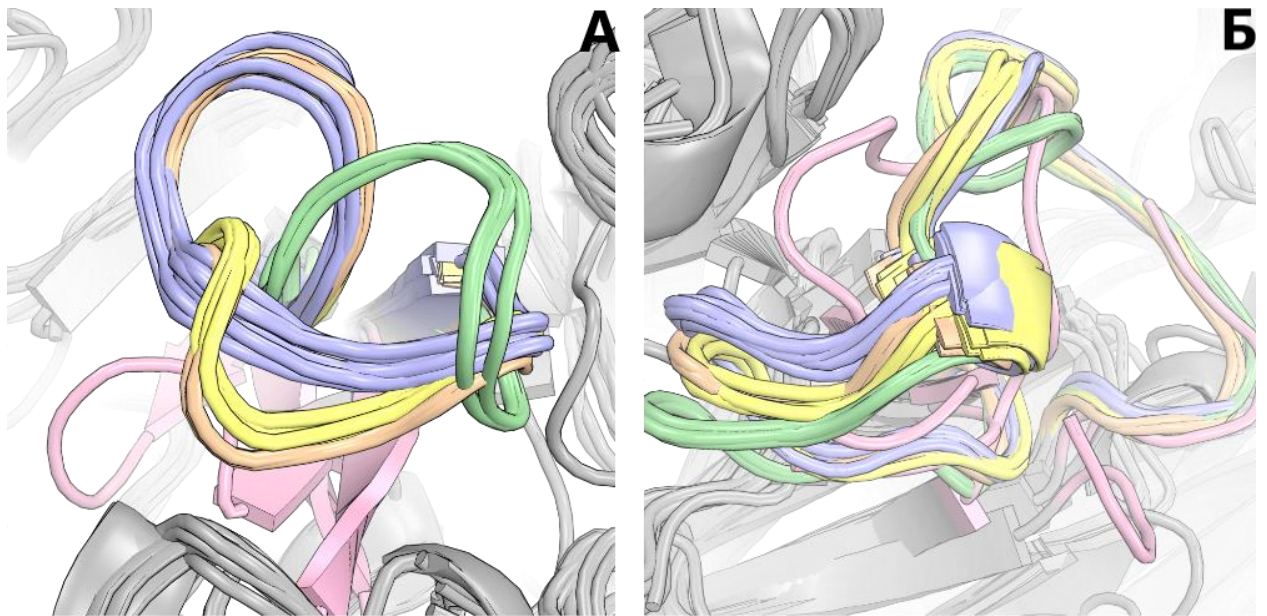


Рисунок 40. Множественное структурное выравнивание суперсемейства нейраминидаз GN34. На рисунке А цветом выделен СУОЦ № 3 (из 19), соответствующий «петле-430». На рисунке Б цветом выделен СУОЦ № 10 (из 19), соответствующий «петле-150». На рисунках полученные при кластеризации подсемейства обозначены разными цветами. На обоих рисунках полученные «розовый» и «зеленый» кластеры соответствуют нейраминидаза-подобным белкам N10 и N11 летучих мышей, которые лишены нейраминидазной активности, а также гомологи из менее патогенного вируса гриппа В. «Синий» кластер соответствует первой филогенетической группе нейраминидаз гриппа А. «Желтый» и «оранжевый» кластеры соответствуют второй филогенетической группе нейраминидаз гриппа А.

- *Суперсемейство α/β -гидролаз.* При исследовании множественного структурного выравнивания α/β -гидролаз были идентифицированы СУОЦ № 3 и СУОЦ № 5 (из 14), первый из которых включает фрагмент петли L9 и α 4-спирали, а второй – петлю L14 в структуре люциферазы из *Renilla reniformis*. Эти участки пространственно различаются в гомологах из суперсемейства α/β -гидролаз с различной каталитической активностью. Недавно проведенные исследования выявили решающую роль этих подвижных областей в ферментативном катализе. Они непосредственно влияют на размер входного отверстия туннеля (открытие/закрытие туннеля, см. рисунок 41), который соединяет скрытый активный центр с окружающим растворителем и участвуют в связывании субстрата/продукта [149]. Введение случайных вставок/делений в этих двух областях в структуру термостабильного

малоактивного предка галоалкандегалогеназы и люциферазы из *Renilla* (код PDB 6G75) путем направленной эволюции привело к 100-кратному увеличению каталитической активности (k_{cat}/K_M). Дальнейшая вставка области L9- $\alpha 4$ из высокоактивной современной люциферазы RLuc8 в этот малоактивный предковый фермент позволила создать химеру, характеризующуюся 7000-кратным увеличением каталитической эффективности [149]. По-видимому, изменение размера отверстия входного туннеля влияет на доступность активного центра для связывания субстрата, и таким образом, влияет на каталитическую активность. Получившиеся кластеры соответствуют α/β -гидролазам с различной каталитической активностью и отличаются размерами и формой входного отверстия туннеля. Найденные 3D-специфические паттерны изображены на рисунке 42.

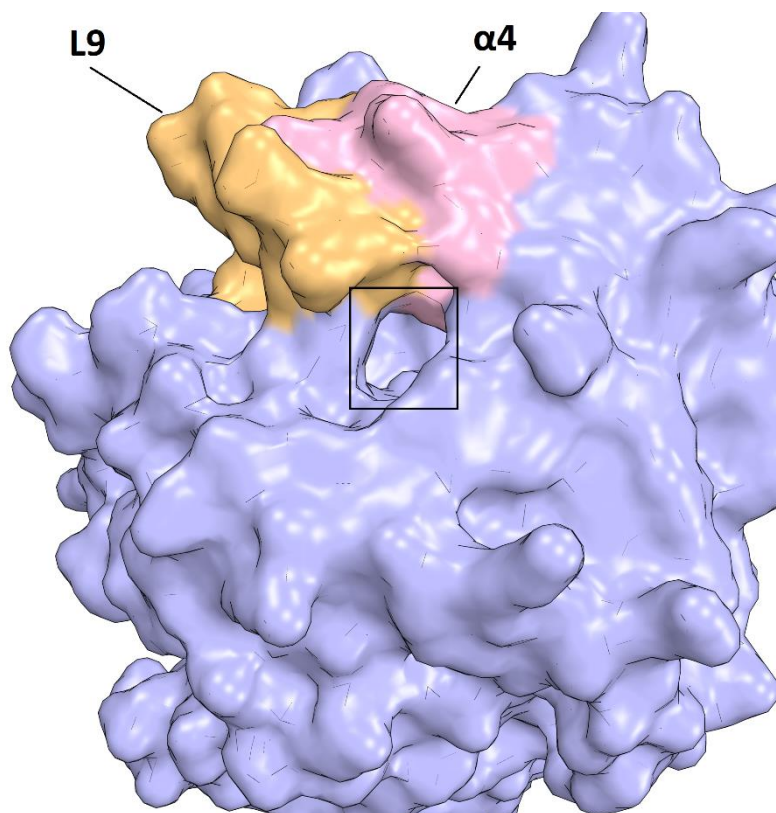


Рисунок 41. Общий предок дегалогеназ галоалканов и люциферазы из *Renilla* (PDB 6g75). Петля L9 и $\alpha 4$ -спираль (выделены цветом) непосредственно влияют на размер входного отверстия туннеля (обведен на рисунке в квадрат), который соединяет скрытый активный центр с окружающим растворителем и участвуют в связывании субстрата/продукта [145]. По-видимому, изменение размера отверстия входного туннеля влияет на доступность активного центра для связывания субстрата и, таким образом, влияет на каталитическую активность.

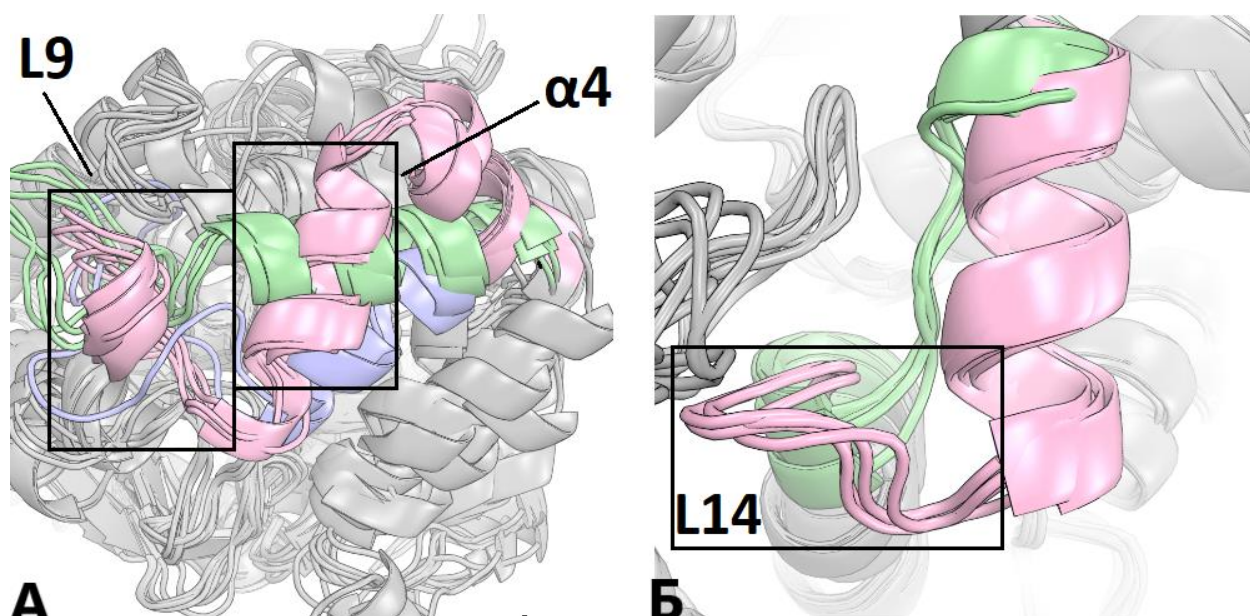


Рисунок 42. Множественное структурное выравнивание суперсемейства α/β -гидролаз. На рисунке А цветом выделен СУОЦ № 3 (из 14), включающий фрагмент петли L9 и $\alpha 4$ -спирали (эти фрагменты структуры на рисунке показаны на ферментах, принадлежащих розовому кластеру). На рисунке Б цветом выделен СУОЦ № 5 (из 14), включающий петлю L14 (петля L14 на рисунке показана на ферментах, принадлежащих розовому кластеру). Получившиеся кластеры соответствуют α/β -гидролазам с различной каталитической активностью: розовый кластер – дегалогеназы (на обоих рисунках), зеленый – эстеразы и хлоропероксидазы, а также ферменты с неустановленными свойствами (на обоих рисунках). Представители полученных кластеров отличаются размером и формой входного отверстия туннеля (на рисунке А).

- *Гомологи б-пирувоилтетрагидроптеринсинтазы крысы.* Представители этого суперсемейства осуществляют метаболизм птериновых кофакторов в организме. При анализе множественного структурного выравнивания гомологов б-пирувоилтетрагидроптеринсинтазы крысы был выявлен СОБЦ № 8 (из 12). СОБЦ № 8 соответствует аминокислотному остатку Glu107 б-пирувоилтетрагидроптеринсинтазы крысы, участвующему в связывании субстрата. Glu107 формирует солевой мостик с протонированной аминогруппой кольца птерина (субстрата), таким образом закрепляя птерин [150]. Автоматическая кластеризация ферментов суперсемейства, полученная для СОБЦа №8, соответствует разделению ферментов по каталитической активности. Получившиеся

кластеры соответствуют дигидронеоптеринальдолозам («зеленый» кластер на рисунке 43) и 6-пирувоилтетрагидро(био)птеринсинтазам («розовый» кластер на рисунке 43). Дигидронеоптеринальдолозы разрывают С-С связь субстрата, 6-пирувоилтетрагидро(био)птеринсинтазы разрывают С-О связь субстрата.

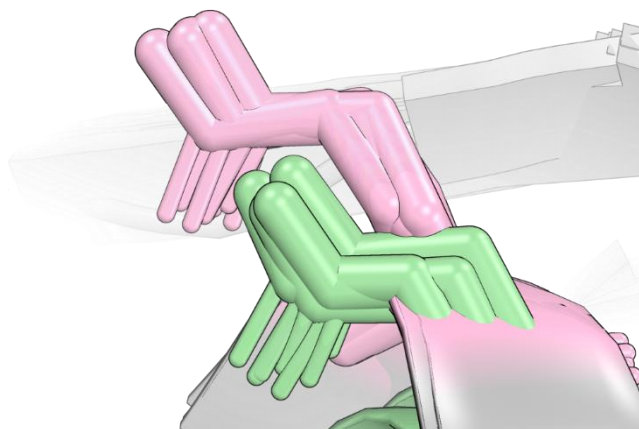


Рисунок 43. Множественное структурное выравнивание гомологов 6-пирувоилтетрагидроптерин синтазы крысы. Цветом выделен СОБЦ № 8 (из 12), соответствующий аминокислотному остатку Glu107 6-пирувоилтетрагидроптеринсинтазы крысы, участвующему в связывании субстрата. «Розовый» кластер соответствует 6-пирувоилтетрагидро(био)птеринсинтазам, «зеленый» кластер – дигидронеоптеринальдолозам.

- *Суперсемейство металло-бета-лактамаз.* Металло-бета-лактамазы – ферменты, встречающиеся у многих бактерий и являющиеся основной причиной резистентности бактерий к β -лактамным антибиотикам. Нами были рассмотрено множественное структурное выравнивание металло-бета-лактамаз, в которых были идентифицированы СУОЦы № 3, 4, 5 (из 13). СУОЦ №3 соответствует функционально важной петле L3 активного центра, СУОЦ № 4, 5 – фрагменты функционально важной петли L10 активного центра. Активные центры металло-бета-лактамаз ограничены двумя петлями, конформация одной из которых, петли L3,

как было показано с помощью мутации данной петли, определяет скорость протонирования ключевых промежуточных продуктов реакции [151]. Автоматическая кластеризация ферментов суперсемейства металло-бета-лактамаз, полученная для СУОЦа №3, дала следующие результаты. Полученные кластеры соответствуют различным классам/типам металло-бета-лактамаз: металло-бета-лактамазы подкласса В3 принадлежат отдельному кластеру, а подкласса В1 – двум другим. Подклассы В1 и В3 отличаются структурой активного центра [152]. Кластеризация белков по подсемействам для СУОЦов №4, 5 дала результаты, аналогичные тем, которые получились для СУОЦа №3. Найденные 3D-специфические паттерны изображены на рисунке 44.

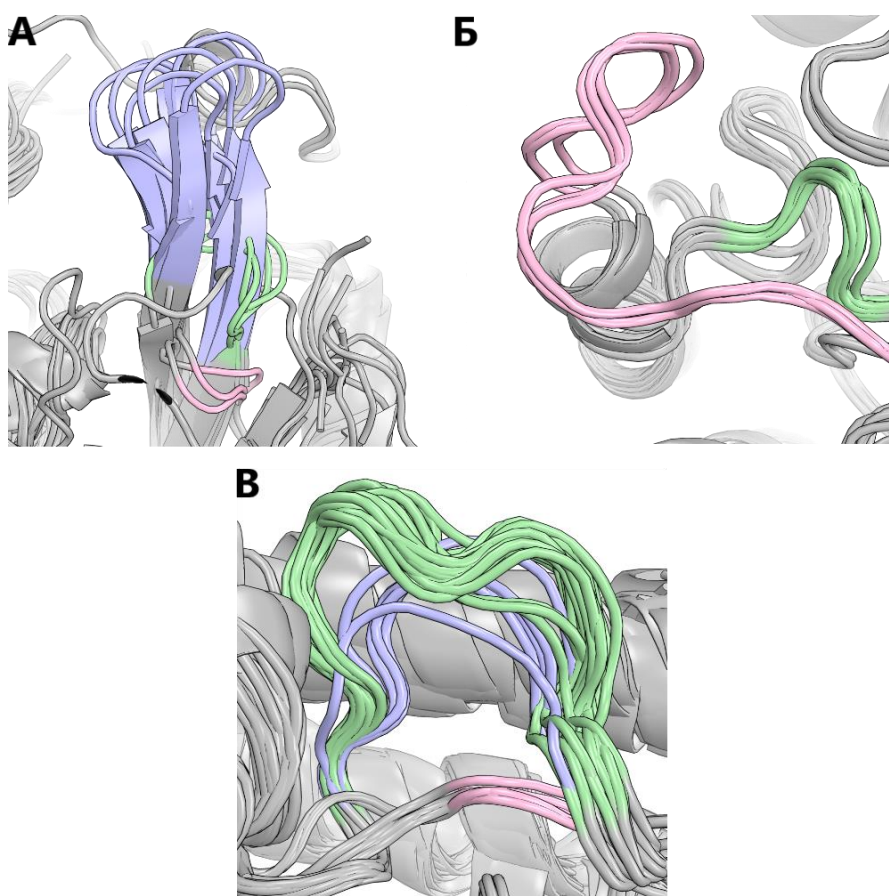


Рисунок 44. Множественное структурное выравнивание металло-бета-лактамаз. На рисунке А цветом выделен СУОЦ № 3 (из 13), соответствующий функционально важной петле L3 активного центра. На рисунке Б цветом выделен СУОЦ № 4 (из 13), соответствующий части функционально важной петли L10 активного центра. На рисунке В цветом выделен СУОЦ № 5 (из 13), соответствующий части функционально важной петли L10 активного центра. На рисунках полученные при кластеризации подсемейства

обозначены разными цветами. Полученный «розовый» кластер (на всех трех рисунках) соответствует металло-бета-лактамазам подкласса В3, два другие кластера соответствуют металло-бета-лактамазам подкласса В1.

5.2.2. 3D-специфические паттерны, отвечающие за различия свойств конформеров одного белка

В этой главе приведен список примеров суперсемейств белков, в которых найденные 3D-специфические паттерны отвечают за различные положения участка структуры фермента, важные для его свойств и отличающиеся в различных функционально-значимых конформациях одного фермента.

- *Суперсемейство киназы фосфорилаз.* Представители суперсемейства киназы фосфорилаз катализируют перенос фосфатной группы от АТФ на аминокислотные остатки белковых субстратов. При анализе множественного структурного выравнивания суперсемейства киназы фосфорилаз был идентифицирован СУОЦ №3 (из 11). СУОЦ №3 соответствует активационной петле р38 α MAP-киназы человека, находящейся возле активного центра и содержащей два остатка тирозина, которые в ответ на различные провоспалительные и стрессовые стимулы организма фосфорилируются вышестоящей в MAP-киназном каскаде MAPK с помощью АТФ, что вызывает переход петли в каталитически активную конформацию DFG-in. То есть присоединение фосфата из молекулы АТФ к остаткам Thr180 и Tyr182 активационной петли приводит к конформационной перестройке белка, что позволяет осуществить реакцию переноса фосфата АТФ на субстрат-белок. р38 α MAP-киназа человека в состоянии DFG-in может связывать молекулу АТФ, поскольку в результате фосфорилирования аминокислотных остатков активационной петли уменьшается расстояние между доменами фермента. Благодаря этому расстояние между γ -фосфатом АТФ (связан в N-домене) и гидроксильной группой белка-субстрата (связан в C-домене) уменьшается до необходимого, и реакция становится возможной (см. рисунок 45). Фосфорилирование и

связывание АТФ индуцируют полностью активную конформацию р38α MAPK [133]. Полученные кластеры соответствуют PDB-файлам, отвечающим за различные структурные положения активационной петли. Найденный 3D-специфический паттерн изображен на рисунке 46.

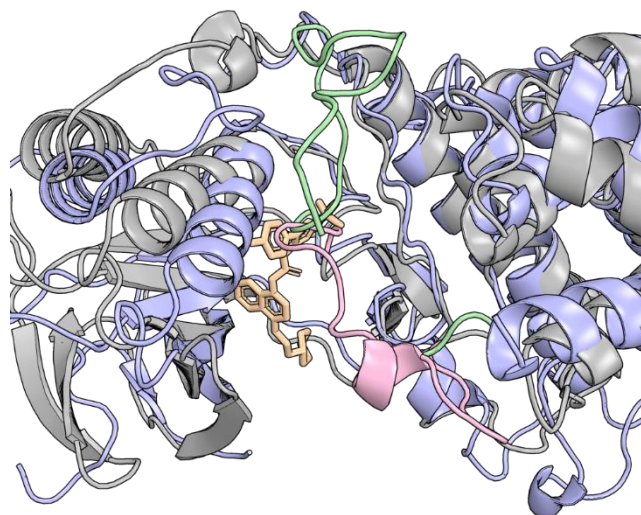


Рисунок 45. Представители суперсемейства киназы-фосфорилаз (серым цветом окрашена структура PDB 1r3c – неактивная форма, голубым цветом окрашена структура PDB 3gr9 – активная форма). Фермент в активной форме может связывать молекулу АТФ, поскольку в результате фосфорилирования аминокислотных остатков активационной петли (выделена цветом) уменьшается расстояние между доменами фермента. Благодаря этому расстояние между γ -фосфатом АТФ (связан в N-домене) и гидроксильной группой белка-субстрата (связан в С-домене) уменьшается до необходимого, и реакция становится возможной.

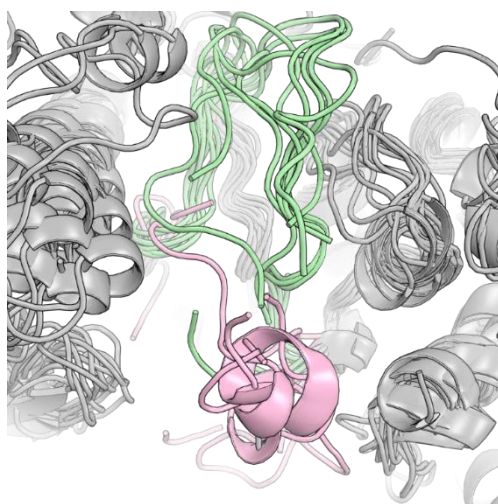


Рисунок 46. Множественное структурное выравнивание суперсемейства киназы фосфорилаз. Цветом выделен СУОЦ № 3 (из 11), соответствующий активационной петле,

находящейся возле активного центра. У активационной петли известны два положения: DFG-in и DFG-out. Полученный «розовый» кластер соответствует положению DFG-out, «зеленый» кластер – DFG-in.

- *Суперсемейство протеин-тирозин-фосфатаз.* В исследовании множественного структурного выравнивания суперсемейства протеин-тирозин-фосфатаз был выявлен СОБЦ № 2 (из 5). СОБЦ № 2 (из 5) соответствует аминокислотному остатку Arg409 тирозиновой протеинфосфатазы из *Yersinia*, находящемуся в активном центре. Ориентация Arg409 меняется в зависимости от связывания фосфата в активном центре [153]. При связывании фосфата (Arg409 образует с фосфатом две водородные связи и ионную связь), боковой радикал Arg409 поворачивается, и таким образом, обеспечивается удержание субстрата (фосфорилированного белка) в активном центре и правильное положение атомов фосфора для отщепления фосфата нуклеофилом Cys403 от белка-субстрата (см. рисунок 47). Другие анионы, такие как вольфрамат или сульфат, приводят к аналогичным конформационным изменениям. Полученные кластеры соответствуют PDB-файлам, отвечающим за различные структурные положения боковой цепи Arg409. Найденный 3D-специфический паттерн изображен на рисунке 48.

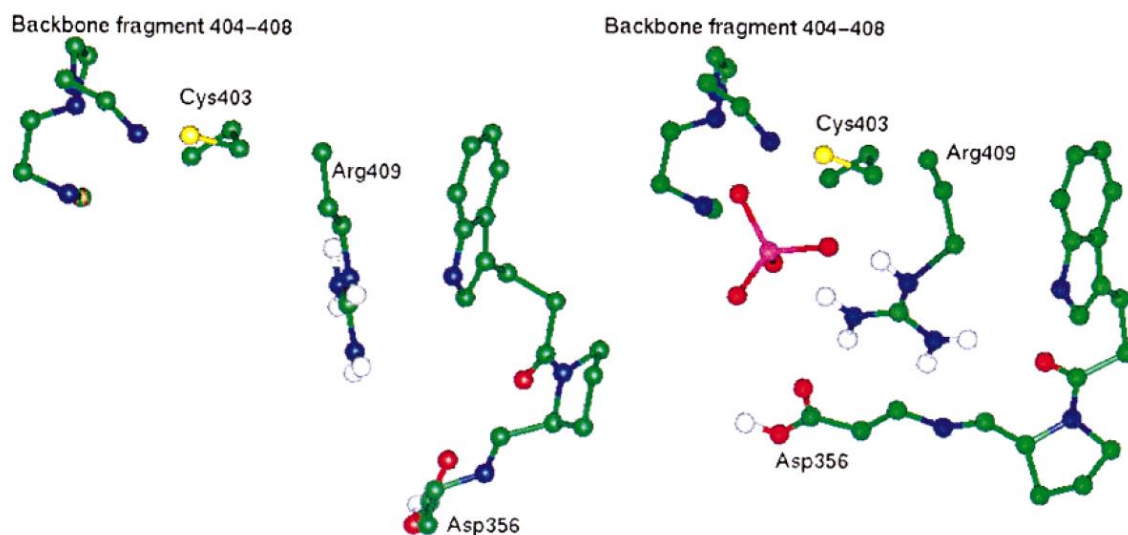


Рисунок 47. Ориентация Arg409 тирозиновой протеинфосфатазы из *Yersinia* меняется в зависимости от связывания фосфата в активном центре. При связывании фосфата (Arg409 образует с фосфатом две водородные связи и ионную связь), боковой радикал Arg409

поворачивается, и таким образом, обеспечивается удержание субстрата (фосфорилированного белка) в активном центре и правильное положение атомов фосфора для отщепления фосфата нуклеофилом Cys403 от белка-субстрата. Рисунок взят из [153].

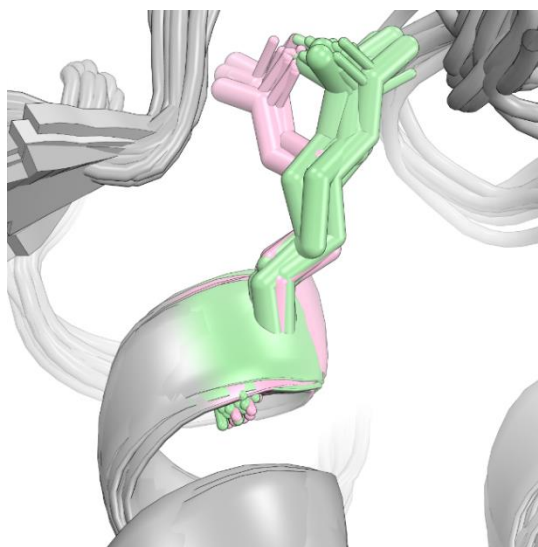


Рисунок 48. Множественное структурное выравнивание суперсемейства протеин-тирозин-фосфатаз. Цветом выделен СОБЦ № 2 (из 5), соответствующий Arg409, ориентация которого меняется в зависимости от связывания фосфата в активном центре. Найденные кластеры соответствуют различным положениям Arg409.

- *Суперсемейство гистидин-киназ-подобных АТФаз.* Представитель этого суперсемейства – Hsp90 человека – это белок-шаперон, который участвует в фолдинге белков, стабилизирует белки при тепловом шоке и способствует деградации белка. В исследовании множественного структурного выравнивания суперсемейства гистидин-киназ-подобных АТФаз был выявлен СУОЦ № 3 (из 9). СУОЦ № 3 включает в себя подвижный сегмент крышки (аминокислотные остатки 107–141) и α -спираль3 (аминокислотные остатки 101–123) Hsp90 человека. В зависимости от связанного лиганда α -спираль3 принимает три конформации: конформацию непрерывной спирали, «loop-in» и «loop-out». Механизм, каким образом гибкость Hsp90 влияет на связывание малых молекул, не совсем ясен. Исследования показали [154], что распознавание и связывание подвижного сегмента крышки с различными субстратами и ингибиторами, вероятно, включает как механизмы индуцированного соответствия, так и конформационной

селекции. Полученные кластеры соответствуют PDB-файлам с различными положениями сегмента крышки и α -спирали3 [154]. Найденный 3D-специфический паттерн изображен на рисунке 49.

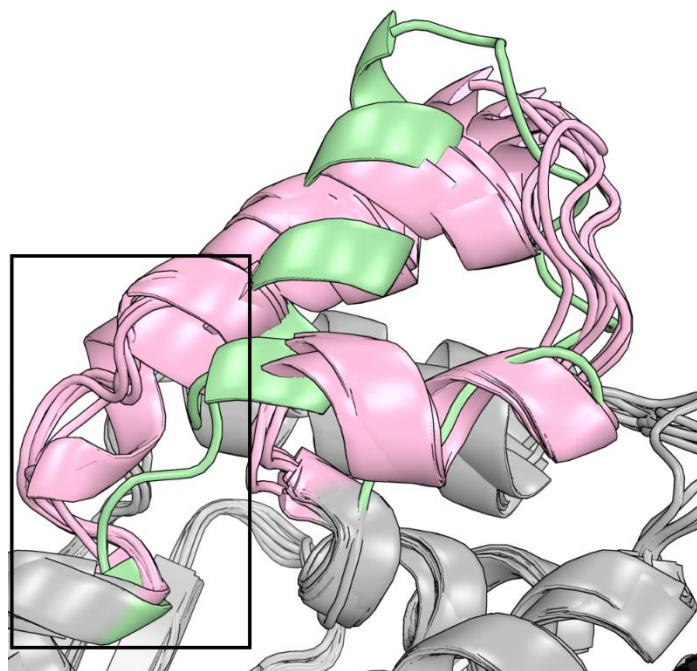


Рисунок 49. Множественное структурное выравнивание суперсемейства гистидин-киназ-подобных АТФаз. Цветом выделен СУОЦ № 3 (из 9), соответствующий подвижному сегменту крышки и включающий в себя α -спираль3. Выделенный рамкой участок соответствует α -спирали3, которая принимает три конформации: конформацию непрерывной спирали, «loop-in» и «loop-out». «Розовый» кластер соответствует конформации непрерывной спирали и «loop-out». «Зеленый» кластер соответствует конформации «loop-in».

- *Гомологи лактатдегидрогеназы из Thermus thermophilus.* Лактатдегидрогеназа – фермент, принимающий участие в реакциях гликолиза. В исследовании множественного структурного выравнивания гомологов лактатдегидрогеназы из *Thermus thermophilus* был выявлен СУОЦ № 2 (из 21), соответствующий подвижному участку, который включает остатки, участвующие в катализе и связывании, и геометрия которого отличается в апо-форме и в тройном комплексе с субстратом и кофактором [155]. Полученные кластеры соответствуют PDB-файлам с различными структурными положениями подвижного участка. Найденный 3D-специфический паттерн изображен на рисунке 50.

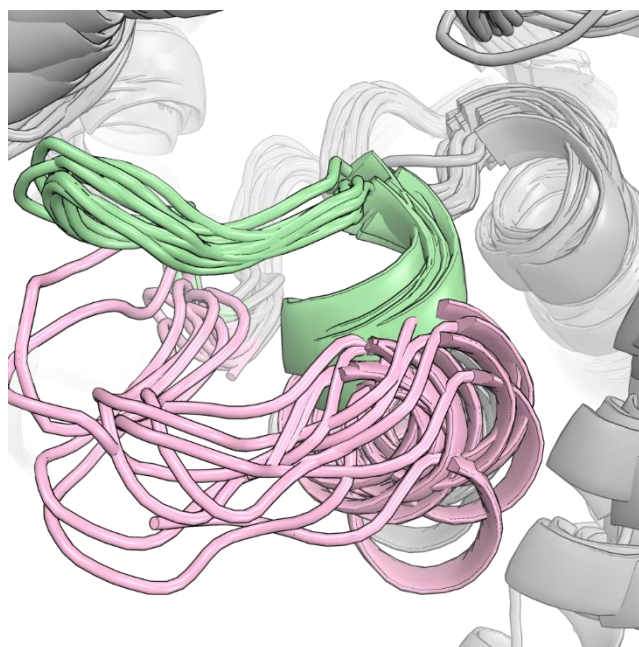


Рисунок 50. Множественное структурное выравнивание гомологов лактатдегидрогеназы из *Thermus thermophilus*. Цветом выделен СУОЦ № 2 (из 21), соответствующий подвижному участку, который включает остатки, участвующие в катализе и связывании, и геометрия которого отличается в апо-форме и в тройном комплексе с субстратом и кофактором. Полученные кластеры соответствуют PDB-файлам с различными структурными положениями подвижного участка.

- *Гомологи рибонуклеазы А из Bos taurus.* Рибонуклеаза А связывает и гидролизует РНК. В исследовании множественного структурного выравнивания гомологов рибонуклеазы А из *Bos taurus* был выявлен СОБЦ № 2 (из 29), соответствующий His119. Остаток His119 участвует в связывании РНК и в катализе реакции. Может быть в двух конформациях – гош-конформации и транс-конформации. Гош-конформация наблюдается в щелочной среде и при отсутствии связанного субстрата. Транс-конформация наблюдается всегда при связывании субстрата (РНК) или при сдвиге рН среды в кислую сторону [156]. Полученные два кластера соответствуют гош- и транс-конформации. Найденный 3D-специфический паттерн изображен на рисунке 51.

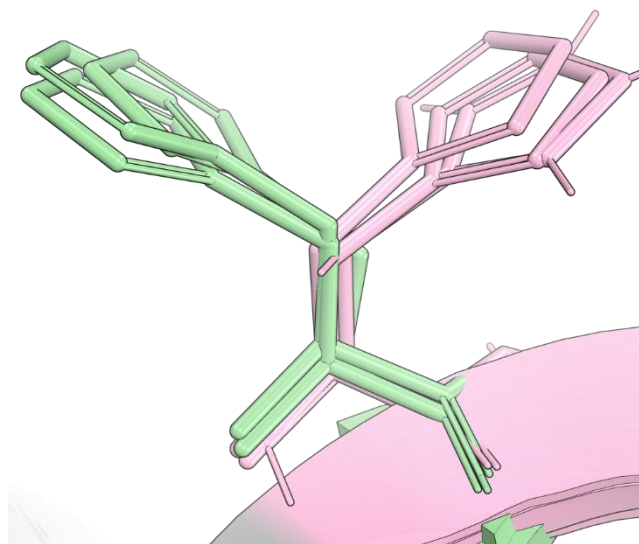


Рисунок 51. Множественное структурное выравнивание гомологов рибонуклеазы А из *Bos taurus*. Цветом выделен СОБЦ № 2 (из 29), соответствующий HIS119, который может быть в двух конформациях – гош-конформации и транс-конформации. Гош-конформация наблюдается в щелочной среде и при отсутствии связанного субстрата. Транс-конформация наблюдается всегда при связывании субстрата (РНК) или при сдвиге рН среды в кислую сторону. Полученный «розовый» кластер соответствует гош-конформации, «зеленый» – транс-конформации.

5.2.3. Обобщение результатов исследования 3D-специфических паттернов суперсемейства белков

Разработанный нами метод выявления 3D-специфических паттернов был апробирован на 13 суперсемействах белков, в которых было найдено 18 функционально-важных 3D-специфических паттернов. Из них 13 отвечают за различие в свойствах между ферментами, принадлежащими различным функциональным подсемействам, 5 отвечают за различные положения участка структуры фермента, важные для его свойств и отличающиеся в различных функционально-значимых конформациях одного фермента. Найденные функционально-значимые 3D-специфические паттерны имеют высокий ранг и маленькое значение P -оценки, что подтверждает состоятельность предложенной оценки специфичности S -оценки и предложенной статистической модели для различения функционально-важных 3D-специфических паттернов от результатов случайных тепловых колебаний белковой структуры. Нахождение 3D-специфических паттернов в

суперсемействах белков – важный этап сравнительного анализа структур гомологичных белков, позволяющий находить те элементы структур белков, которые ответственны за их свойства и функции, а также различия в гомологичных белках. В таблице 5 приведен общий список найденных 3D-специфических паттернов.

Таблица 5. В столбце «PDB» указаны PDB-коды и идентификаторы цепочки, которые были отправлены в качестве запроса в веб-сервер Mustguseal для автоматического сбора и выравнивания избыточного набора трехмерных структур гомологов (см главу 4.7), то есть для получения соответствующего множественного структурного выравнивания суперсемейства. В столбце «Кол-во PDB-файлов в выравни.» указано общее количество окончательно выбранных PDB-файлов в таком выравнивании. «Кол-во СУОЦов/СОБЦов» указывает общее количество найденных 3D-специфических паттернов соответствующего типа. «Участок/аминокислота», «Ранг», «Z-оценка» и «P-оценка» указывают первый и последний аминокислотные остатки в СУОЦе/номер аминокислоты в СОБЦе (в соответствии с нумерацией в PDB-файле), его ранг, Z-оценку статистической значимости и соответствующую P-оценку. В столбцах «Общее количество СУОЦов/СОБЦов» и «Ранг» в случае СУОЦов в скобках указано значение с учетом N-/C-концевых участков. Первые 13 3D-специфических паттернов отвечают за различие в свойствах между ферментами, принадлежащими различным подсемействам, следующие 5 отвечают за различные положения элемента структуры фермента, важные для его свойств и отличающиеся в различных функционально-значимых конформациях одного фермента.

#	Запрос		Кол-во PDB-файлов в выравни.	Кол-во СУОЦов/СОБЦов	СУОЦ/СОБЦ				Интерпретация классификации по подсемействам	
	Название	PDB			Участок/аминокислота	Ранг	Z-оценка	P-оценка		Роль для функции
1	Орнитин-декарбоксилаза из <i>Trypanosoma brucei</i>	1F3T:A	31	23 (21)	322-336	8 (6)	2.37	8.8e-03	3 ₁₀ -спираль, которая расположена на одной стороне полости связывания субстрата и принимает альтернативные ориентации, соответствующие ферментам с различной субстратной специфичностью.	Пиридоксаль-зависимые ферменты из группы декарбоксилаз основных аминокислот с укладкой типа β/α-цилиндра с различной субстратной специфичностью.
2	Альдокеторедуктаза человека	2ACQ:A	62	14 (12)	112-135	1 (1)	10.11	2.5e-24	Подвижный участок, участвующий в связывании субстрата.	Альдокеторедуктазы с различной субстратной специфичностью.
3	Полиэфиргидролаза из	6SBN:A	20	13 (11)	127-134	5 (3)	3.84	6.1e-05	Субстрат-связывающий элемент активного сайта.	ПЭТ-гидролазы и близко-родственные промискуитетные

	<i>Pseudomonas aestusnigri</i>									кутиназы против ферментов, которые не являются ПЭТ-гидролазами.
					97-103	8 (6)	1.66	4.8e-02	Субстрат-связывающий элемент активного сайта.	ПЭТ-гидролазы и близко-родственные промискуитетные кутиназы против ферментов, которые не являются ПЭТ-гидролазами.
4	Гуаниндеаминаза человека	2UZ9:A	23	22 (20)	215-222	7 (5)	3.32	4.5e-04	Элемент, распознающий субстрат.	Металло-зависимые гидролазы с различной субстратной специфичностью.
5	H1N1 нейраминидаза	3B7E:A	29	19 (17)	427-440	3 (1)	6.00	9.6e-10	«Петля-430», подвижность которой приводит к формированию «полости-430».	Нейроминидазы из штаммов/типов гриппа с различной патогенностью.
					136-156	10 (8)	2.32	1.0e-02	«Петля-150», подвижность которой приводит к формированию «полости-150».	Нейроминидазы из штаммов/типов гриппа с различной патогенностью.
6	Общий предок дегалогеназ галоалканов и люцифер	6G75:A	41	14 (13)	146-176	3 (3)	3.64	1.4e-04	Участок, который включает фрагмент петли L9 и α 4-спирали, и участвующий в связывании субстрата.	α/β -гидролазы с различной каталитической активностью.
					222-239	5 (4)	2.53	5.7e-03	Участок, включающий L14 петлю,	α/β -гидролазы с различной

	азы из <i>Renilla</i>								участвующую в связывании субстрата.	каталитической активностью.
7	6-пирувоилтетрагидроптеринсинтаза крысы	1B66:A	19	12	GLU107	8	0.009	0.5	Участвует в связывании субстрата.	Дигидронеоптеринальдолазы против пирувоилтетрагидро(био)птерин-синтаз.
8	Цинк-зависимая металло-бета-лактамаза из <i>Bacillus cereus</i>	1BVT:A	40	13 (12)	30-39	3 (3)	8.13	2.2e-16	Функционально важная L3-петля активного сайта.	Различные классы/типы металло-бета-лактамаз.
					181-184	4 (4)	7.75	4.5e-15	Часть функционально важной L10-петли активного сайта.	Различные классы/типы металло-бета-лактамаз.
					170-179	5 (5)	6.42	6.8e-11	Часть функционально важной L10-петли активного сайта.	Различные классы/типы металло-бета-лактамаз.
9	p38α MAP-киназа человека	1R3C:A	61	11 (10)	169-185	3 (2)	9.90	2.1e-23	Активационная петля.	PDB-файлы, отвечающие за различные структурные положения активационной петли.
10	Тирозиновая протеинфосфатаза из <i>Yersinia</i>	1YTW:A	32	5	ARG409	2	2.04	2.1e-02	Аминокислотный остаток, находящийся в активном центре, ориентация которого меняется в зависимости от связывания фосфата.	PDB-файлы, отвечающие за различные структурные положения ARG409.

11	HSP90 человека	1YET:A	19	9 (7)	107-136	3 (2)	2.52	5.8e-03	Подвижный сегмент крышки и α -спираль ³ .	PDB-файлы с различными структурными положениями сегмента крышки и α -спираль ³ .
12	Лактатдегидрогеназа из <i>Thermus thermophilus</i>	2V7P:C	60	21 (21)	100-112	2 (2)	3.83	6.4e-05	Подвижный участок, который включает остатки, участвующие в катализе и связывании, и геометрия которого отличается в апо-форме и в тройном комплексе с субстратом и кофактором	PDB-файлы с различными структурными положениями подвижного участка
13	Рибонуклеаза А из <i>Bos taurus</i>	1RUV:A	26	29	HIS119	2	1.44	7.2e-02	Остаток HIS119 участвует в связывании РНК и в катализе реакции. Может быть в двух конформациях – гош-конформации и транс-конформации. Гош-конформация наблюдается в основной среде и при отсутствии связанного субстрата. Транс-конформация наблюдается всегда при связывании субстрата (РНК) или при сдвиге рН среды в кислую сторону.	PDB-файлы, отвечающие за гош-конформацию и транс-конформацию.

5.2.4. Сравнение результатов применения метода выявления 3D-специфических паттернов и метода выявления специфических позиций подсемейства на выборке суперсемейств белков

3D-специфические паттерны являются трехмерным аналогом специфических позиций подсемейства/позиций, определяющих специфичность (см. главу 3.2.1.2). Для выявления специфических позиций подсемейства существует много методов (см. главу 3.2.1.2). В этой главе сравниваются результаты применения метода выявления 3D-специфических паттернов и метода выявления специфических позиций подсемейства Zebra2 [123]. Метод Zebra2 предоставляет автоматическое деление суперсемейства белков на подсемейства по сходству последовательностей и выявляет специфические позиции подсемейства.

Для сравнения результатов, полученных с помощью метода Zebra2, и представленного в этой диссертационной работе метода выявления 3D-специфических паттернов, метод Zebra2 был применен (см. главу 4.8) к суперсемействам белков, описанных в главах 5.2.1, 5.2.2. Получившаяся с помощью метода Zebra2 кластеризации белков для каждого суперсемейства сравнивалась с кластеризацией, полученной для функционально значимого 3D-специфического паттерна, найденного в данном суперсемействе с помощью *Adjusted Rand Index* (см. главу 4.9). В таблице 6 представлены результаты работы метода Zebra2, а именно: (1) полученные с помощью этого метода специфические позиции подсемейства для каждого суперсемейства и (2) *Adjusted Rand Index* (ARI) между кластеризацией, полученной с помощью метода Zebra2 и кластеризацией, полученной для функционально значимого 3D-специфического паттерна, найденного в данном суперсемействе.

Таблица 6. Результат применения метода Zebra2 [123] к суперсемействам белков. Номера специфических позиций подсемейства, выявленных методом Zebra2, приведены в соответствии с PDB-структурой референсного белка и указаны в столбце «Выявленные специфические позиции». В колонке «ARI с учетом выбросов» приведено значение метрики ARI, рассчитанное между кластеризацией, полученной с помощью метода Zebra2, и кластеризацией, полученной для функционально значимого 3D-специфического паттерна, найденного в данном суперсемействе. Выбросы при расчете учитывались. В колонке «ARI без учета выбросов» при расчете ARI белки, отнесенные к выбросам хотя бы в одной из кластеризаций, не учитывались. Над жирной чертой в таблице приведены суперсемейства белков, в которых найдены 3D-специфические паттерны, отвечающие за различие в свойствах между ферментами, принадлежащими различным подсемействам, а под жирной чертой – отвечающие за различные положения участка структуры в одном ферменте.

Название суперсемейства белков	PDB референсного белка	Выявленные специфические позиции	ARI с учетом выбросов	ARI без учета выбросов
Суперсемейство пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α -цилиндра	1F3T:A	316, 112, 90, 70, 198, 132, 196, 113, 101, 386, 390, 405, 356, 273, 320, 315, 284, 353, 109, 134, 234, 382, 238, 216, 293, 114, 406, 98, 141, 215, 60, 156, 338, 319, 64, 95, 143, 279, 222, 43, 263, 147, 89, 55, 383, 195, 292, 73, 63, 62, 290, 133, 289, 172, 278, 255, 87, 354, 380, 207, 317, 110, 229, 287, 314, 181, 407, 119, 30, 146, 366, 285, 219, 103, 369, 111, 370, 375, 45, 142, 218, 233, 189, 155, 93, 145, 150, 180, 251, 152, 288, 188, 28	0.78	0.91
Альдо-кеторедуктазы	2ACQ:A	183, 77, 159, 262, 157, 108, 18, 43, 110, 260, 209, 106, 184, 16, 104, 185, 207, 179, 51, 161, 249, 190, 211, 95, 212, 178, 88, 101, 182, 42, 79, 44, 186, 105, 160, 12, 257, 180, 80, 20, 17, 154, 272, 91, 45, 39, 55, 210, 40, 92, 73	0.65	0.64
Гомологи полиэфиргидролазы из <i>Pseudomonas aestusnigri</i>	6SBN:A	217, 174, 173, 169, 116	0.81	1
Металло-зависимые гидролазы	2UZ9:A	167, 238, 143, 208, 278, 142, 169, 140, 86, 212, 141, 334, 163, 386, 218, 144, 276, 211, 198, 102, 442, 165, 372, 171, 388, 428, 104, 154, 80, 287, 168, 207, 161, 194, 432, 244, 13, 336, 190, 157, 166, 396, 275, 402, 298, 271, 282, 300, 224, 170, 137, 29, 209, 172, 261, 15, 346, 31, 90, 138, 365, 441, 229, 363, 227, 164, 77, 299, 235, 87, 11, 15	0.97	1

		0, 329, 119, 220, 294, 324, 311, 280, 378, 376, 159, 374, 214, 45, 85, 447, 233, 149, 328, 47, 111, 266		
Суперсемейство α/β -гидролаз	6G75:A	118, 53, 117, 51, 285, 52, 57, 79, 119, 201, 77, 219, 142, 55, 218, 282, 98, 123, 140, 259, 289, 95, 203, 61, 275, 283, 141, 217, 125, 27, 197, 178,	0.86	0.80
Гомологи б-пирувоил тетрагидроптерин синтазы крысы	1B66:A	50, 133, 105, 134, 48, 72, 24, 137, 23, 106, 61, 21, 92, 60, 114, 141, 126, 22, 143, 108, 63, 87, 112, 19, 62, 15, 129, 91, 75, 83, 69, 111, 118, 20, 70	0.95	1
Суперсемейство металло-бета-лактамаз	1BVT:A	168, 166, 152, 41, 180, 91, 28, 57, 167, 107, 209, 44, 206, 29, 45, 97, 66, 207, 156, 43, 208, 165, 48, 189, 144, 200, 78, 146, 193, 30, 25, 185, 153, 218, 154, 103, 178, 70, 105, 87, 67	0.78	0.81
Суперсемейство киназы фосфорилаз	1R3C:A	188, 152, 153, 289, 211, 138, 288, 285, 204, 154, 222, 209, 299, 206, 297, 112, 88, 214, 24, 302, 282, 186, 274, 189, 127, 135, 128, 223, 239, 68, 142, 90, 193, 67, 231, 305, 137, 52, 156, 221, 187, 166, 281, 200, 235, 146, 134, 212, 219, 55, 116, 141, 216	0.55	0.64
Тирозиновая протеинфосфатаза из <i>Yersinia</i>	1YTW:A	407, 231, 286, 347, 418, 420, 414, 236, 277, 368, 299, 285, 441, 439, 416, 412, 252, 400, 348, 462, 373, 436, 422, 235, 338, 248, 426, , 263, 267, 276, 250, 445, 460, 444, 247, 434, 365, 300, 435, 461, 241, 417, 224	0.59	0.54

HSP90 человека	1YET:A	39, 100, 194, 36, 193, 38, 35, 37, 210, 200, 96, 102, 160, 220, 82, 78, 61, 158, 216, 202, 166, 168, 171, 28, 105, 88, 196, 156, 31, 68, 26, 101, 201, 62, 81, 74, 29, 71, 90, 189, 215, 69, 195, 185, 222, 72, 67, 73, 92, 205, 173, 77, 161, 142, 219, 212, 91, 169	0.67	0.69
Гомологи лактатдегидрогеназы из <i>Thermus Thermophilus</i>	2V7P:C	260, 143, 199, 123, 257, 52, 145, 154, 27, 34, 197, 247, 249, 248, 113, 163, 196, 97, 194, 153, 246, 120, 31, 41, 232, 200, 189, 170, 65, 271, 62, 318, 272, 74, 144, 228, 46, 270, 329, 201, 242, 237, 240, 169, 172, 146, 264, 88, 173, 202, 233, 147, 133, 24, 64, 128, 165, 98	0.57	0.49
Гомологи рибонуклеазы А из <i>Bos taurus</i> .	1RUV:A	47, 36, 13, 5, 51, 35, 7, 43, 120, 54, 79, 29, 105, 9, 42, 115, 106, 34, 111, 81, 10, 53, 49, 116, 83, 73, 25, 55	0.46	0.47

По полученным значениям ARI можем сделать вывод, что разделение суперсемейства белков на подсемейства, полученное с помощью метода Zebra2, похоже на разделение на подсемейства, полученное для 3D-специфических паттернов, отвечающих за различие в свойствах между ферментами, принадлежащими различным подсемействам (среднее значение ARI с учетом выбросов равно 0.82, без учета выбросов равно 0.88). То есть кластеризация белков по аминокислотным последовательностям совпадает с кластеризацией белков для найденных функционально важных 3D-специфических паттернов. В случае суперсемейств белков, в которых были найдены 3D-специфические паттерны, отвечающие за различные положения участка структуры в одном ферменте, кластеризации имеют значительные отличия (среднее значение ARI с учетом выбросов равно 0.568, без учета выбросов равно 0.566). Это связано с тем, что в кластеризации, полученной для такого 3D-специфического паттерна, в разные кластеры попадают PDB-структуры белков, отвечающие за различное положение данного подвижного участка структуры, что никак не связано с аминокислотной последовательностью (для одного и того же белка данный участок структуры может быть в разных положениях в зависимости от условий).

Выявленные с помощью метода Zebra2 специфические позиции подсемейства не входят в найденные описанные в предыдущей главе функционально важные 3D-специфические паттерны (за исключением найденных специфических позиций 218, 222 в суперсемействе металло-зависимых гидролаз и специфических позиций 30, 178 в суперсемействе металло-бета-лактамаз), так как:

- Позиции выравнивания, входящие в данный 3D-специфический паттерн, могут быть консервативными по аминокислотной последовательности среди белков суперсемейства. Такие позиции методом выявления специфических позиций не могут быть выявлены согласно определению специфических позиций (примером могут

послужить гомологи рибонуклеазы А из *Bos taurus* и найденный в этом суперсемействе СОБЦ. См. рисунок 51. Другой пример см. на рисунке 52).

- Полученные структурные и структурно-опосредованные выравнивания являются выравниваниями эволюционно удаленных гомологов. Последовательность менее консервативна, чем структура, поэтому последовательности данных белков могли сильно измениться в процессе эволюции и взаимосвязь последовательность-функция могла быть утеряна для данных позиций выравнивания.

В случае СОБЦов найденные функционально-значимые 3D-специфические паттерны были обозначены методом Zebra2 как *консервативные* позиции выравнивания. Это связано с тем, что методы выявления специфических позиций подсемейства никак не учитывают ориентацию боковых цепей аминокислотных остатков в пространстве, что, в отличие от них, делает метод выявления 3D-специфических паттернов.

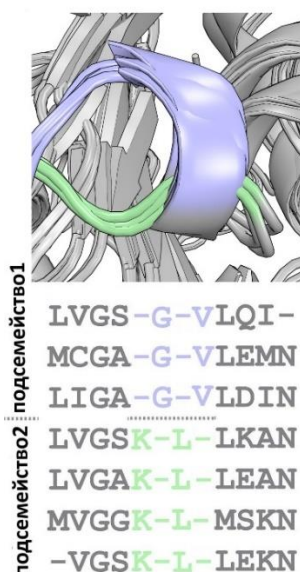


Рисунок 52. Структурное выравнивание и соответствующее структурно-опосредованное выравнивание последовательностей. На рисунке цветом выделен 3D-специфический паттерн. См. описание в тексте.

5.2.5. Сравнение результатов применения метода выявления 3D-специфических паттернов и метода выявления коррелирующих позиций на выборке суперсемейств белков

Помимо сравнения метода выявления 3D-специфических паттернов с методом выявления специфических позиций подсемейства, мы сравнили метод выявления 3D-специфических паттернов с методом выявления коррелирующих позиций. Для этих целей мы воспользовались веб-сервером visualСМАТ [27]. Для получения коррелирующей позиции мы подали на вход этому веб-серверу выравнивания суперсемейств белков, описанных в главах 5.2.1, 5.2.2. Результаты работы visualСМАТ представлены в таблице 7. Можно сделать вывод, что лишь небольшое количество найденных коррелирующих позиций входит в функционально значимые, описанные в 5.2.3 главе, 3D-специфические паттерны (в среднем в один 3D-специфический паттерн входит 1.5 коррелирующие позиции). Это можно объяснить теми же причинами, что и схожий результат в случае специфических позиций подсемейства (см. предыдущую главу 5.2.4).

Из результатов, полученных в этой главе и в главе 5.2.4 можно сделать следующий вывод: выявление и анализ 3D-специфических паттернов является важным дополнением к анализу выравниваний аминокислотных последовательностей гомологичных белков (то есть при биоинформатическом исследовании суперсемейства белков стоит использовать как анализ множественного выравнивания аминокислотных последовательностей, так и анализ множественного структурного выравнивания), так как (1) позволяет находить такие участки структуры белка, которые ответственны за функциональное разнообразие белков суперсемейства, которые не может найти ни метод выявления специфических позиций подсемейства, ни метод выявления коррелирующих позиций, (2) предоставляет кластеризацию белков, где различные кластеры отвечают различным положениям подвижного участка структуры.

Таблица 7. Результат применения веб-сервера visualCMAT [27] к суперсемействам белков. Номера коррелирующих позиций, выявленных методом visualCMAT, приведены в соответствии с PDB-структурой референсного белка и указаны в столбце «Выявленные коррелирующие позиции».

Название суперсемейства белков	PDB референсного белка	Выявленные коррелирующие позиции	Выявленные коррелирующие позиции, входящие в функционально важные 3D-специфические паттерны
Суперсемейство пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α -цилиндра	1F3T:A	22, 23, 24, 26, 29, 30, 66, 67, 71, 88, 91, 106, 109, 110, 112, 114, 115, 118, 128, 129, 138, 154, 155, 169, 177, 189, 195, 198, 201, 211, 233, 235, 236, 237, 239, 245, 246, 247, 248, 249, 250, 258, 263, 266, 274, 275, 277, 281, 290, 294, 313, 319, 323, 324, 328, 329, 333, 342, 343, 344, 345, 348, 351, 355, 357, 361, 362, 363, 364, 368, 369, 372, 387, 388, 389, 393, 396, 397, 400, 401	328, 329, 333, 324,323
Альдо-кеторедуктазы	2ACQ:A	3, 4, 7, 8, 264, 10, 11, 270, 18, 157, 159, 287, 290, 295, 43, 299, 45, 172, 176, 51, 55, 183, 187, 67, 76, 77, 82, 95, 96, 226, 230, 104, 232, 106, 105, 235, 108, 110, 244	
Гомологи полиэфиргидролазы из <i>Pseudomonas aestusnigri</i>	6SBN:A	293, 294, 233, 43, 301, 175, 50, 52, 281, 126, 155, 221, 158	
Металло-зависимые гидролазы	2UZ9:A	17, 19, 20, 21, 24, 40, 54, 55, 67, 77, 80, 87, 92, 95, 102, 107, 108, 109, 115, 120, 132, 137, 161, 163, 172, 181, 182, 186, 187, 188, 199, 202, 204, 206, 211, 212, 213, 214, 218, 221, 230, 231, 232, 243, 246, 250, 251, 252, 253, 255, 256, 259, 267, 268, 274, 278, 282, 287, 292, 300, 302, 304, 305, 310, 311, 319, 322, 325, 328, 329, 332, 334, 344, 346, 351, 352, 353, 354, 361, 372, 375, 376, 377, 380, 383, 393, 396, 402, 410, 411, 415, 416, 421, 422, 430, 442	221, 218

		385, 386, 387, 390, 393, 142, 284, 285, 84, 341, 214, 440, 442, 190, 449, 200, 457, 330, 459, 460, 458, 332, 463, 336, 83, 468, 85, 86, 87, 344, 345, 90, 470, 340, 89, 342, 91, 347, 218, 228, 359, 105, 111, 125, 120, 252, 381	440, 142
Суперсемейство α/β -гидролаз	6G75:A	259, 134, 12, 17, 20, 288, 33, 171, 303, 48, 51, 59, 62, 69, 73, 74, 201, 77, 206, 207, 208, 79, 210, 205, 84, 90, 98, 231, 104, 106, 235, 244, 117, 118, 251	171, 231, 235
Гомологи б-пирувоил тетрагидроптерин синтазы крысы	1B66:A	96, 35, 68, 133, 105, 43, 49, 50, 23, 24, 26, 94	
Суперсемейство металло-бета-лактамаз	1BVT:A	137, 16, 146, 148, 149, 152, 153, 28, 156, 157, 31, 32, 33, 34, 29, 164, 38, 166, 168, 42, 43, 175, 179, 180, 183, 56, 57, 191, 192, 205, 206, 86, 88, 222, 98, 111, 117	34, 38, 33, 31, 32, 183, 179, 175
Суперсемейство киназы фосфорилаз	1R3C:A	261, 262, 263, 265, 266, 138, 268, 269, 271, 149, 152, 153, 24, 154, 289, 38, 303, 305, 50, 51, 185, 188, 190, 319, 80, 211, 215, 346, 221, 222, 112, 241, 243, 244, 246, 247, 250, 251, 252, 25	185
Тирозиновая протеинфосфатаза из <i>Yersinia</i>	1YTW:A	446, 261, 266, 298, 234, 269, 270, 239, 271, 398, 243, 404, 308, 405, 411, 284, 414, 286	
HSP90 человека	1YET:A	130, 131, 141, 16, 23, 27, 163, 35, 165, 38, 40, 174, 175, 181, 182, 186, 189, 66, 67, 69, 72, 73, 200, 78, 83, 85, 86, 215, 88, 90, 219, 92, 220, 221, 222, 100, 101, 110, 112	110, 131, 112, 130

Гомологи лактатдегидрогеназы из <i>Thermus Thermophilus</i>	2V7P:C	131, 140, 141, 269, 271, 144, 143, 23, 154, 155, 27, 34, 163, 164, 37, 162, 296, 297, 170, 300, 173, 174, 301, 48, 314, 188, 319, 65, 193, 195, 196, 197, 198, 321, 200, 73, 70, 331, 79, 88, 220, 97, 98, 102, 107, 239, 240, 111, 114, 243, 246, 248, 249, 122, 252, 253	102, 107, 111
Гомологи рибонуклеазы А из <i>Bos taurus</i> .	1RUV:A	70, 71, 72, 11, 78, 48, 82, 116, 61	

5.3 3D-мотивы. Статистическая модель оценки структурной гибкости основной цепи 3D-мотивов дисульфидных мостиков для определения возможности вставки данного 3D-мотива в структуру белка

3D-специфические паттерны – это структурные паттерны суперсемейства белков, схожие внутри подсемейств белков, но различающиеся между ними и отвечающие за функциональное разнообразие белков суперсемейства. В отличие от них 3D-мотивы – это структурные паттерны суперсемейства белков, отвечающие за общность свойств и функций белков суперсемейства. Если 3D-специфические паттерны – это оригинальное понятие и методы их выявления и анализа являются новыми, то методы выявления и анализа 3D-мотивов существуют (см. главу 3.2.2.4). Часто для определения, соответствует ли данный 3D-мотив данному элементу структуры белка, используется жесткое эмпирически выбранное пороговое значение (например, 5 Å). Нами предлагается выбирать это пороговое значение с помощью специальной статистической модели, описанной в главе 5.3.3 на примере 3D-мотивов дисульфидных мостиков.

5.3.1 Получение 3D-мотивов дисульфидных мостиков

Для получения 3D-мотивов дисульфидных мостиков (то есть типовых уникальных геометрий дисульфидных мостиков) из структур белков базы данных PDB были извлечены дисульфидные мостики так, как описано в главе 4.10. Каждому дисульфидному мостику ставился в соответствие вектор $\{\cos\varphi_1, \sin\varphi_1, \dots, \cos\varphi_9, \sin\varphi_9\}$, где $\varphi_1, \dots, \varphi_9$ – это набор обычных и двухгранных углов, полностью описывающий геометрию данного S-S мостика: (C α -C β -S γ , C β -S γ -S' γ , S γ -S' γ -C' β , S' γ -C' β -C' α) – обычные и (C-C α -C β -S γ , C α -C β -S γ -S' γ , C β -S γ -S' γ -C' β , S γ -S' γ -C' β -C' α , S' γ -C' β -C' α -C') – двухгранные углы (см. рисунок 53). Получившиеся векторы подавались на вход методу кластеризации HDBSCAN [46]. В результате кластеризации было получено 273 кластера и 4748 выброса (выбросы составляют 28% от общего количества рассматриваемых дисульфидных связей). Каждый кластер/выброс представляет собой 3D-мотив, а дисульфидные связи, входящие в один

кластер, считаются нами геометрически эквивалентными. В качестве «центрального» S-S мостика данного 3D-мотива в каждом кластере выбиралась одна репрезентативная дисульфидная связь, в которой достигается минимум суммы расстояний до других членов данного кластера.

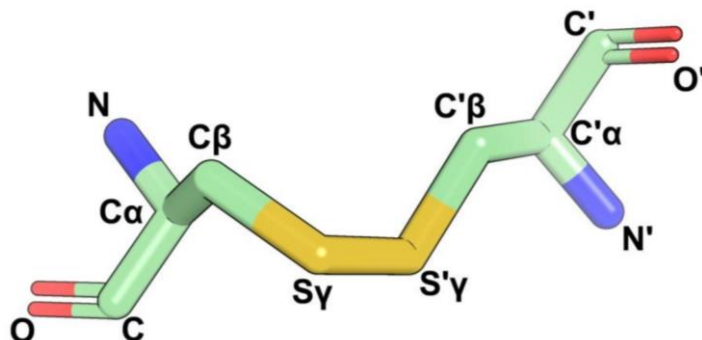


Рисунок 53. Каждому дисульфидному мостику ставится в соответствие вектор $\{\cos\varphi_1, \sin\varphi_1, \dots, \cos\varphi_9, \sin\varphi_9\}$, где $\varphi_1, \dots, \varphi_9$ – это набор обычных и двухгранных углов, полностью описывающий геометрию данного S-S мостика: $(C\alpha-C\beta-S\gamma, C\beta-S\gamma-S'\gamma, S\gamma-S'\gamma-C'\beta, S'\gamma-C'\beta-C'\alpha)$ – обычные и $(C-C\alpha-C\beta-S\gamma, C\alpha-C\beta-S\gamma-S'\gamma, C\beta-S\gamma-S'\gamma-C'\beta, S\gamma-S'\gamma-C'\beta-C'\alpha, S'\gamma-C'\beta-C'\alpha-C')$ – двухгранные углы.

5.3.3. Статистическая модель оценки структурной гибкости основной цепи 3D-мотивов для определения возможности вставки данного 3D-мотива в структуру белка на примере 3D-мотивов дисульфидных мостиков

Для определения, может ли данная пара аминокислотных остатков остатков при мутации на цистеины образовать дисульфидную связь, то есть соответствует ли данный элемент структуры данному 3D-мотиву дисульфидного мостика, нами была разработана специальная статистическая модель. Для этого в каждом из полученных кластеров были выполнены все попарные структурные выравнивания между «центральным» дисульфидным мостиком данного 3D-мотива и всеми другими членами кластера. Для каждого парного выравнивания были рассчитаны два значения RMSD: для одной и второй пары цистеинов, причем учитывались только атомы основной цепи. В каждом кластере было выбрано только наибольшее значение из рассчитанных в этом кластере RMSD. Полученные 273 независимых значения ($\mu = 0,16 \text{ \AA}$, $\sigma = 0,07 \text{ \AA}$) были использованы для создания статистической модели. В рамках этой статистической модели две позиции в структуре интересующего белка выбираются как целевые позиции для образования дисульфидной связи, если

они совпадают по крайней мере с одним 3D-мотивом по следующему критерию. Критерий заключается в следующем: после выравнивания атомов основной цепи выбранных для мутации двух аминокислотных остатков с атомами основной цепи данного 3D-мотива, оба значения RMSD между атомами основной цепи двух пар выравненных друг с другом аминокислотных остатков находятся в пределах $0,28 \text{ \AA}$ (на рисунке 54 изображен пример такого выравнивания, значения RMSD рассчитываются между атомами основной цепи аминокислотных остатков S_1 и E_1 , а также между S_2 и E_2 и должны быть меньше $0,28 \text{ \AA}$). Значение $0,28 \text{ \AA}$ соответствует P -значению = 0.05 нормального распределения с $\mu = 0,16 \text{ \AA}$ и $\sigma = 0,07 \text{ \AA}$. При невыполнении данного критерия пара позиций в структуре белка не рассматривается в качестве целевой для образования дисульфидной связи.

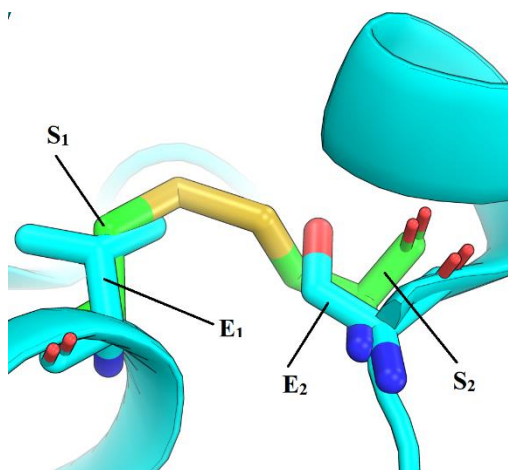


Рисунок 54. См. описание в тексте.

5.3.4. Апробация статистической модели оценки структурной гибкости основной цепи 3D-мотивов дисульфидных мостиков

Мы апробировали разработанную статистическую модель (см. главу 4.11). В результате чувствительность разработанного статистического критерия составила $Sensitivity = 97,59\%$, а специфичность $Specificity = 94.71\%$ (см. главу 4.11).

6. Заключение

Выявление элементов структуры белка (отдельных аминокислотных остатков, ориентации боковых радикалов, участков основной цепи), ответственных за разнообразие функций и свойств белков внутри суперсемейства – важнейшая задача современной биологии. Раньше поиск функционально важных аминокислотных остатков и участков основной цепи проводили либо с помощью визуального экспертного анализа, либо с помощью биоинформатического анализа множественного выравнивания аминокислотных последовательностей белков суперсемейства. В последнее время становится доступно все больше информации о структурной организации белков, а также увеличиваются вычислительные мощности компьютеров. Поэтому становится возможным проводить не только выравнивания аминокислотных последовательностей больших суперсемейств белков, но и множественные выравнивания их структур. Проводя анализ структурных выравниваний, можно выявлять функционально важные фрагменты структуры, такие как структурные паттерны суперсемейства белков – характеристическое, повторяющееся в белках суперсемейства относительное расположение элементов структуры, которое может быть ответственно за важные свойства и функции. К структурным паттернам суперсемейства белков относятся как известные из литературы 3D-мотивы, так и предложенные в этой работе 3D-специфические паттерны.

Нами разработан новый подход, позволяющий выявлять 3D-специфические паттерны – структурные паттерны суперсемейства белков, которые структурно схожи внутри подсемейств и различаются между ними, ответственны за свойства и функции белков и играют ключевую роль в функциональном разнообразии белков суперсемейства. Разработанный метод выявления 3D-специфических паттернов основан на использовании

множественного структурного выравнивания белков суперсемейства, а также широко известных приемов статистики и методов машинного обучения. Подход имплементирован в виде программного кода, написанного на языке Python3, и свободно доступен для скачивания и использования. Новый метод выявления 3D-специфических паттернов апробирован на примерах суперсемейств белков, для которых из литературных источников известны функционально-важные элементы структур. Апробация показала, что найденные разработанным нами методом 3D-специфические паттерны играют роль в проявлении различной каталитической активности, различной субстратной специфичности белков суперсемейства и являются характеристиками конформационной пластичности участков связывания лигандов. Это позволяет сделать вывод о том, что информация о 3D-специфических паттернах может быть использована в белковом дизайне при выборе целевых позиций для направленного изменения структуры и функции, а также при создании лекарственных препаратов на основе селективных модуляторов белков и ферментов.

7. Основные результаты и выводы

- Разработан новый метод сравнительного анализа структур белков суперсемейства, основанный на выявлении 3D-специфических паттернов - элементов структуры белков/ферментов (участков основной цепи, отдельных аминокислотных остатков, ориентации боковых радикалов), которые схожи внутри подсемейств белков, но различаются между ними и позволяют разделить суперсемейства на функционально обособленные подсемейства.
- Разработана *S*-оценка специфичности, позволяющая ранжировать выявленные 3D-специфические паттерны по их функциональной значимости в данном суперсемействе.
- Разработана статистическая модель для отделения функционально-значимых 3D-специфических паттернов от результатов теплового колебания структуры белка.
- Разработано свободно доступное программное обеспечение [134,135], позволяющее находить 3D-специфические паттерны в заданном пользователем суперсемействе белков.
- Предположено и при анализе литературных данных о функциональных свойствах изученных ферментов показано, что 3D-специфические паттерны представляют важные для механизма действия элементы структуры ферментов и отвечают за различие свойств (таких как субстратная специфичность, каталитическая активность) ферментов, принадлежащих к различным функциональным подсемействам, а также конформеров одного фермента благодаря пространственной ориентации ключевых аминокислотных остатков и участков основной цепи. 3D-специфические паттерны могут быть использованы при функциональной аннотации и рациональном дизайне белков.
- Результаты, полученные с помощью метода сравнительного анализа структур белков суперсемейства, основанного на выявлении 3D-

специфических паттернов, качественно дополняют результаты, полученные с помощью методов выявления коррелирующих позиций и специфических позиций подсемейства (методов сравнительного анализа аминокислотных последовательностей белков).

- Предложена методология белкового дизайна в результате вставки 3D-мотивов дисульфидных мостиков в структуру белков с целью получения стабилизированных препаратов с измененными функциональными свойствами.

8. Список литературы

1. Packer M.S., Liu D.R. Methods for the directed evolution of proteins // *Nature Reviews Genetics*. Nature Publishing Group, 2015. Vol. 16, № 7. P. 379–394.
2. Carter P. Site-directed mutagenesis. // *Biochemical Journal*. Portland Press Ltd, 1986. Vol. 237, № 1. P. 1.
3. Chagoyen M., García-Martín J.A., Pazos F. Practical analysis of specificity-determining residues in protein families // *Briefings in bioinformatics*. Oxford University Press, 2016. Vol. 17, № 2. P. 255–261.
4. De Juan D., Pazos F., Valencia A. Emerging methods in protein co-evolution // *Nature Reviews Genetics*. Nature Publishing Group, 2013. Vol. 14, № 4. P. 249–261.
5. Bai X.-C., McMullan G., Scheres S.H. How cryo-EM is revolutionizing structural biology // *Trends in biochemical sciences*. Elsevier, 2015. Vol. 40, № 1. P. 49–57.
6. Valdar W.S. Scoring residue conservation // *Proteins: structure, function, and bioinformatics*. Wiley Online Library, 2002. Vol. 48, № 2. P. 227–241.
7. Lichtarge O., Bourne H.R., Cohen F.E. An evolutionary trace method defines binding surfaces common to protein families // *Journal of molecular biology*. Elsevier, 1996. Vol. 257, № 2. P. 342–358.
8. Casari G., Sander C., Valencia A. A method to predict functional residues in proteins // *Nature structural biology*. Nature Publishing Group, 1995. Vol. 2, № 2. P. 171–178.
9. Göbel U. et al. Correlated mutations and residue contacts in proteins // *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 1994. Vol. 18, № 4. P. 309–317.
10. Suplatov D., Voevodin V., Švedas V. Robust enzyme design: Bioinformatic tools for improved protein stability // *Biotechnology journal*. Wiley Online Library, 2015. Vol. 10, № 3. P. 344–355.
11. Kalinina O.V. et al. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families // *Protein Science*. Wiley Online Library, 2004. Vol. 13, № 2. P. 443–456.
12. Kalinina O.V. et al. SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins //

- Nucleic Acids Research. Oxford University Press, 2004. Vol. 32, № suppl_2. P. W424–W428.
13. Mirny L.A., Gelfand M.S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors // Journal of molecular biology. Elsevier, 2002. Vol. 321, № 1. P. 7–20.
 14. Codoñer F.M., Fares M.A. Why should we care about molecular coevolution? // Evolutionary Bioinformatics. SAGE Publications Sage UK: London, England, 2008. Vol. 4. P. 117693430800400000.
 15. Konagurthu A.S. et al. MUSTANG: a multiple structural alignment algorithm // Proteins: Structure, Function, and Bioinformatics. Wiley Online Library, 2006. Vol. 64, № 3. P. 559–574.
 16. Shegay M.V. et al. parMATT: parallel multiple alignment of protein 3D-structures with translations and twists for distributed-memory systems // Bioinformatics. Oxford University Press, 2019. Vol. 35, № 21. P. 4456–4458.
 17. Dong R. et al. mTM-align: an algorithm for fast and accurate multiple protein structure alignment // Bioinformatics. Oxford University Press, 2018. Vol. 34, № 10. P. 1719–1725.
 18. Menke M., Berger B., Cowen L. Matt: local flexibility aids protein multiple structure alignment // PLoS computational biology. Public Library of Science San Francisco, USA, 2008. Vol. 4, № 1. P. e10.
 19. Shatsky M., Nussinov R., Wolfson H.J. A method for simultaneous alignment of multiple protein structures // Proteins: Structure, Function, and Bioinformatics. Wiley Online Library, 2004. Vol. 56, № 1. P. 143–156.
 20. Pei J., Kim B.-H., Grishin N.V. PROMALS3D: a tool for multiple protein sequence and structure alignments // Nucleic acids research. Oxford University Press, 2008. Vol. 36, № 7. P. 2295–2300.
 21. Lupyan D., Leo-Macias A., Ortiz A.R. A new progressive-iterative algorithm for multiple structure alignment // Bioinformatics. Oxford University Press, 2005. Vol. 21, № 15. P. 3255–3263.
 22. Akdel M. et al. Caretta—a multiple protein structure alignment and feature extraction suite // Computational and structural biotechnology journal. Elsevier, 2020. Vol. 18. P. 981–992.
 23. DeLano W.L. Pymol: An open-source molecular graphics tool // CCP4 Newsl. Protein Crystallogr. Citeseer, 2002. Vol. 40, № 1. P. 82–92.
 24. Humphrey W., Dalke A., Schulten K. VMD: visual molecular dynamics // Journal of molecular graphics. Elsevier, 1996. Vol. 14, № 1. P. 33–38.

25. Bakan A., Meireles L.M., Bahar I. ProDy: protein dynamics inferred from theory and experiments // *Bioinformatics*. Oxford University Press, 2011. Vol. 27, № 11. P. 1575–1577.
26. Gaillard T., Stote R.H., Dejaegere A. PSSweb: protein structural statistics web server // *Nucleic Acids Research*. Oxford University Press, 2016. Vol. 44, № W1. P. W401–W405.
27. Suplatov D. et al. The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families // *Journal of Bioinformatics and Computational Biology*. World Scientific, 2018. Vol. 16, № 02. P. 1840005.
28. Drew E.D., Janes R.W. 2StrucCompare: a webserver for visualizing small but noteworthy differences between protein tertiary structures through interrogation of the secondary structure content // *Nucleic acids research*. Oxford University Press, 2019. Vol. 47, № W1. P. W477–W481.
29. Li Z. et al. FATCAT 2.0: towards a better understanding of the structural diversity of proteins // *Nucleic Acids Research*. 2020. Vol. 48, № W1. P. W60–W64.
30. Nilmeier J.P. et al. 3D Motifs // *From Protein Structure to Function with Bioinformatics* / ed. J. Rigden D. Dordrecht: Springer Netherlands, 2017. P. 361–392.
31. Ribeiro V.S. et al. visGREMLIN: graph mining-based detection and visualization of conserved motifs at 3D protein-ligand interface at the atomic level // *BMC Bioinformatics*. 2020. Vol. 21, № 2. P. 80.
32. He W. et al. Lib ME—automatic extraction of 3D ligand-binding motifs for mechanistic analysis of protein–ligand recognition // *FEBS Open Bio*. Wiley Online Library, 2016. Vol. 6, № 12. P. 1331–1340.
33. Nadzirin N. et al. SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures // *Nucleic Acids Research*. 2012. Vol. 40, № W1. P. W380–W386.
34. Ivanisenko V.A. et al. PDBSite: a database of the 3D structure of protein functional sites // *Nucleic Acids Research*. 2005. Vol. 33, № suppl_1. P. D183–D187.
35. Nebel J.-C. Generation of 3D templates of active sites of proteins with rigid prosthetic groups // *Bioinformatics*. 2006. Vol. 22, № 10. P. 1183–1189.
36. Laskowski R.A., Watson J.D., Thornton J.M. Protein Function Prediction Using Local 3D Templates // *Journal of Molecular Biology*. 2005. Vol. 351, № 3. P. 614–626.

37. Kleywegt G.J. Recognition of spatial motifs in protein structures // Edited by J. Thornton // *Journal of Molecular Biology*. 1999. Vol. 285, № 4. P. 1887–1897.
38. Wallace A.C., Borkakoti N., Thornton J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites // *Protein science*. Wiley Online Library, 1997. Vol. 6, № 11. P. 2308–2323.
39. Kaiser F., Eisold A., Labudde D. A novel algorithm for enhanced structural motif matching in proteins // *Journal of Computational Biology*. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, 2015. Vol. 22, № 7. P. 698–713.
40. He L. et al. Ballast: a ball-based algorithm for structural motifs // *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2012. P. 79–93.
41. Suplatov D. et al. Yosshi: a web-server for disulfide engineering by bioinformatic analysis of diverse protein families // *Nucleic Acids Research*. 2019. Vol. 47, № W1. P. W308–W314.
42. Santana C.A. et al. Gremlin: A graph mining strategy to infer protein-ligand interaction patterns // *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2016. P. 28–35.
43. Nilmeier J.P. et al. Rapid catalytic template searching as an enzyme function prediction procedure // *PloS one*. Public Library of Science San Francisco, USA, 2013. Vol. 8, № 5. P. e62535.
44. Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. // *kdd*. 1996. Vol. 96, № 34. P. 226–231.
45. Ankerst M. et al. OPTICS: Ordering points to identify the clustering structure // *ACM Sigmod record*. ACM New York, NY, USA, 1999. Vol. 28, № 2. P. 49–60.
46. McInnes L., Healy J., Astels S. hdbscan: Hierarchical density based clustering. // *J. Open Source Softw*. 2017. Vol. 2, № 11. P. 205.
47. Hrabe T. et al. PDBFlex: exploring flexibility in protein structures // *Nucleic acids research*. Oxford University Press, 2016. Vol. 44, № D1. P. D423–D428.
48. Suplatov D.A. et al. Mustguseal: a server for multiple structure-guided sequence alignment of protein families // *Bioinformatics*. 2018. Vol. 34, № 9. P. 1583–1585.

49. Suplatov D. et al. Bioinformatic analysis of protein families to select function-related variable positions // *Understanding enzymes: Function, design, engineering, and analysis*. Pan Stanford Publishing Singapore, 2016. P. 351–385.
50. Zuckerkandl E., Pauling L. Evolutionary divergence and convergence in proteins // *Evolving genes and proteins*. Elsevier, 1965. P. 97–166.
51. Villar H.O., Kauvar L.M. Amino acid preferences at protein binding sites // *FEBS letters*. Wiley Online Library, 1994. Vol. 349, № 1. P. 125–130.
52. Sander C., Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment // *Proteins: Structure, Function, and Bioinformatics*. Wiley Online Library, 1991. Vol. 9, № 1. P. 56–68.
53. Eddy S.R. Where did the BLOSUM62 alignment score matrix come from? // *Nature biotechnology*. Nature Publishing Group, 2004. Vol. 22, № 8. P. 1035–1036.
54. Zvelebil M.J. et al. Prediction of protein secondary structure and active sites using the alignment of homologous sequences // *Journal of molecular biology*. Elsevier, 1987. Vol. 195, № 4. P. 957–961.
55. Armon A., Graur D., Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information // *Journal of molecular biology*. Elsevier, 2001. Vol. 307, № 1. P. 447–463.
56. Vingron M., Argos P. A fast and sensitive multiple sequence alignment algorithm // *Bioinformatics*. Oxford University Press, 1989. Vol. 5, № 2. P. 115–121.
57. Altschul S.F., Lipman D.J. Equal animals // *Nature*. Springer, 1990. Vol. 348, № 6301. P. 493–494.
58. May A.C. Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics // *Protein engineering*. Oxford University Press, 2001. Vol. 14, № 4. P. 209–217.
59. <http://biokinet.cmm.msu.ru/zebra2>.
60. Lichtarge O., Bourne H.R., Cohen F.E. An evolutionary trace method defines binding surfaces common to protein families // *Journal of molecular biology*. Elsevier, 1996. Vol. 257, № 2. P. 342–358.

61. Lichtarge O., Yamamoto K.R., Cohen F.E. Identification of functional surfaces of the zinc binding domains of intracellular receptors // *Journal of molecular biology*. Elsevier, 1997. Vol. 274, № 3. P. 325–337.
62. Lichtarge O. et al. Accurate and scalable identification of functional sites by evolutionary tracing // *Journal of structural and functional genomics*. Springer, 2003. Vol. 4, № 2. P. 159–166.
63. Reš I., Mihalek I., Lichtarge O. An evolution based classifier for prediction of protein interfaces without using protein structures // *Bioinformatics*. Oxford University Press, 2005. Vol. 21, № 10. P. 2496–2501.
64. Mihalek I., Reš I., Lichtarge O. A family of evolution–entropy hybrid methods for ranking protein residues by importance // *Journal of molecular biology*. Elsevier, 2004. Vol. 336, № 5. P. 1265–1282.
65. Sowa M.E. et al. Prediction and confirmation of a site critical for effector regulation of RGS domain activity // *Nature structural biology*. Nature Publishing Group, 2001. Vol. 8, № 3. P. 234–237.
66. Quan X.-J. et al. Evolution of neural precursor selection: functional divergence of proneural proteins. Oxford University Press for The Company of Biologists Limited, 2004.
67. Madabushi S. et al. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions // *Journal of Biological Chemistry*. ASBMB, 2004. Vol. 279, № 9. P. 8126–8132.
68. Hannenhalli S.S., Russell R.B. Analysis and prediction of functional sub-types from protein sequence alignments // *Journal of molecular biology*. Elsevier, 2000. Vol. 303, № 1. P. 61–76.
69. Li L., Shakhnovich E.I., Mirny L.A. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases // *Proceedings of the National Academy of Sciences*. National Acad Sciences, 2003. Vol. 100, № 8. P. 4463–4468.
70. Donald J.E., Shakhnovich E.I. Predicting specificity-determining residues in two large eukaryotic transcription factor families // *Nucleic acids research*. Oxford University Press, 2005. Vol. 33, № 14. P. 4455–4465.
71. Capra J.A., Singh M. Characterization and prediction of residues determining protein functional specificity // *Bioinformatics*. Oxford University Press, 2008. Vol. 24, № 13. P. 1473–1480.

72. Pei J. et al. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios // *Bioinformatics*. Oxford University Press, 2006. Vol. 22, № 2. P. 164–171.
73. Suplatov D. et al. Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity // *Journal of Biomolecular Structure and Dynamics*. Taylor & Francis, 2014. Vol. 32, № 1. P. 75–87.
74. Gaucher E.A. et al. Predicting functional divergence in protein evolution by site-specific rate shifts // *Trends in biochemical sciences*. Elsevier, 2002. Vol. 27, № 6. P. 315–321.
75. Suplatov D.A. et al. Bioinformatic analysis of alpha/beta-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities // *Protein Engineering, Design & Selection*. Oxford University Press, 2012. Vol. 25, № 11. P. 689–697.
76. <http://biokinet.cmm.msu.ru/visualcmat>.
77. Atchley W.R. et al. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis // *Molecular biology and evolution*. Oxford University Press, 2000. Vol. 17, № 1. P. 164–178.
78. Marks D.S., Hopf T.A., Sander C. Protein structure prediction from sequence variation // *Nature biotechnology*. Nature Publishing Group, 2012. Vol. 30, № 11. P. 1072–1080.
79. Marks D.S. et al. Protein 3D structure computed from evolutionary sequence variation // *PloS one*. Public Library of Science San Francisco, USA, 2011. Vol. 6, № 12. P. e28766.
80. Dos Santos R.N. et al. Dimeric interactions and complex formation using direct coevolutionary couplings // *Scientific reports*. Nature Publishing Group, 2015. Vol. 5, № 1. P. 1–10.
81. Malinverni D. et al. Large-scale conformational transitions and dimerization are encoded in the amino-acid sequences of Hsp70 chaperones // *PLoS computational biology*. Public Library of Science San Francisco, CA USA, 2015. Vol. 11, № 6. P. e1004262.
82. Lee B.-C., Kim D. A new method for revealing correlated mutations under the structural and functional constraints in proteins // *Bioinformatics*. Oxford University Press, 2009. Vol. 25, № 19. P. 2506–2513.
83. Dunn S.D., Wahl L.M., Gloor G.B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction // *Bioinformatics*. Oxford University Press, 2008. Vol. 24, № 3. P. 333–340.

84. Korber B.T. et al. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. // *Proceedings of the National Academy of Sciences*. National Acad Sciences, 1993. Vol. 90, № 15. P. 7176–7180.
85. Fares M.A., Travers S.A. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses // *Genetics*. Oxford University Press, 2006. Vol. 173, № 1. P. 9–23.
86. Weigt M. et al. Identification of direct residue contacts in protein–protein interaction by message passing // *Proceedings of the National Academy of Sciences*. National Acad Sciences, 2009. Vol. 106, № 1. P. 67–72.
87. Morcos F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families // *Proceedings of the National Academy of Sciences*. National Acad Sciences, 2011. Vol. 108, № 49. P. E1293–E1301.
88. Jones D.T. et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments // *Bioinformatics*. Oxford University Press, 2012. Vol. 28, № 2. P. 184–190.
89. Jeong C.-S., Kim D. Reliable and robust detection of coevolving protein residues // *Protein Engineering, Design & Selection*. Oxford University Press, 2012. Vol. 25, № 11. P. 705–713.
90. Schmidtke P. et al. Fpocket: online tools for protein ensemble pocket detection and tracking // *Nucleic acids research*. Oxford University Press, 2010. Vol. 38, № suppl_2. P. W582–W589.
91. <https://biokinet.belozersky.msu.ru/parMATT>.
92. Stebbings L.A., Mizuguchi K. HOMSTRAD: recent developments of the homologous protein structure alignment database // *Nucleic acids research*. Oxford University Press, 2004. Vol. 32, № suppl_1. P. D203–D207.
93. Van Walle I., Lasters I., Wyns L. SABmark—a benchmark for sequence alignment that covers the entire known fold space // *Bioinformatics*. Oxford University Press, 2005. Vol. 21, № 7. P. 1267–1268.
94. Berbalk C., Schwaiger C.S., Lackner P. Accuracy analysis of multiple structure alignments // *Protein Science*. Wiley Online Library, 2009. Vol. 18, № 10. P. 2027–2035.
95. Zhang Y., Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score // *Nucleic acids research*. Oxford University Press, 2005. Vol. 33, № 7. P. 2302–2309.

96. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins // *Journal of Molecular Biology*. 1970. Vol. 48, № 3. P. 443–453.
97. Krissinel E., Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions // *Acta Crystallographica Section D*. 2004. Vol. 60, № 12 Part 1. P. 2256–2268.
98. Altschul S.F. et al. Basic local alignment search tool // *Journal of Molecular Biology*. 1990. Vol. 215, № 3. P. 403–410.
99. Vouzis P.D., Sahinidis N.V. GPU-BLAST: using graphics processors to accelerate protein sequence alignment // *Bioinformatics*. 2011. Vol. 27, № 2. P. 182–188.
100. Katoh K., Standley D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability // *Molecular Biology and Evolution*. 2013. Vol. 30, № 4. P. 772–780.
101. PyMol Wiki https://pymolwiki.org/index.php/Main_Page.
102. Holm L., Laakso L.M. Dali server update // *Nucleic acids research*. Oxford University Press, 2016. Vol. 44, № W1. P. W351–W355.
103. Ramachandran G. t, Sasisekharan V. Conformation of polypeptides and proteins // *Advances in protein chemistry*. Elsevier, 1968. Vol. 23. P. 283–437.
104. BLOW D.M., BIRKTOFT J.J., HARTLEY B.S. Role of a Buried Acid Group in the Mechanism of Action of Chymotrypsin // *Nature*. 1969. Vol. 221, № 5178. P. 337–340.
105. WRIGHT C.S., ALDEN R.A., KRAUT J. Structure of Subtilisin BPN' at 2.5 Å Resolution // *Nature*. 1969. Vol. 221, № 5177. P. 235–242.
106. Wallace A.C., Laskowski R.A., Thornton J.M. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases // *Protein Science*. Wiley Online Library, 1996. Vol. 5, № 6. P. 1001–1013.
107. Barker J.A., Thornton J.M. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis // *Bioinformatics*. 2003. Vol. 19, № 13. P. 1644–1649.
108. Yan X., Han J. gspan: Graph-based substructure pattern mining // 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE, 2002. P. 721–724.

109. Koza J.R. et al. Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming // Artificial Intelligence in Design '96 / ed. Gero J.S., Sudweeks F. Dordrecht: Springer Netherlands, 1996. P. 151–170.
110. J. Hu et al. Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning // IEEE Transactions on Vehicular Technology. 2020. Vol. 69, № 12. P. 14413–14423.
111. Peterson L.E. K-nearest neighbor // Scholarpedia. 2009. Vol. 4, № 2. P. 1883.
112. Seber G.A., Lee A.J. Linear regression analysis. John Wiley & Sons, 2012.
113. Wright R.E. Logistic regression. American Psychological Association, 1995.
114. Hearst M.A. et al. Support vector machines // IEEE Intelligent Systems and their applications. IEEE, 1998. Vol. 13, № 4. P. 18–28.
115. Song Y.-Y., Ying L.U. Decision tree methods: applications for classification and prediction // Shanghai archives of psychiatry. Shanghai Mental Health Center, 2015. Vol. 27, № 2. P. 130.
116. Bishop C.M. Neural networks and their applications // Review of scientific instruments. American Institute of Physics, 1994. Vol. 65, № 6. P. 1803–1832.
117. Bock H.-H. Clustering methods: a history of k-means algorithms // Selected contributions in data analysis and classification. Springer, 2007. P. 161–172.
118. Wikipedia <https://ru.wikipedia.org/wiki/>.
119. <https://hdbscan.readthedocs.io/en/latest/index.html>.
120. Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65.
121. Abdi H. Z-scores // Encyclopedia of measurement and statistics. Sage Thousand Oaks, CA, 2007. Vol. 3. P. 1055–1058.
122. Porter C.T., Bartlett G.J., Thornton J.M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data // Nucleic acids research. Oxford University Press, 2004. Vol. 32, № suppl_1. P. D129–D133.
123. Suplatov D. et al. Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in

- diverse protein superfamilies // *Nucleic Acids Research*. 2020. Vol. 48, № W1. P. W65–W71.
124. Steinley D. Properties of the hubert-arable adjusted rand index. // *Psychological methods*. American Psychological Association, 2004. Vol. 9, № 3. P. 386.
125. Hubert L., Arabie P. Comparing partitions // *Journal of classification*. Springer, 1985. Vol. 2. P. 193–218.
126. Fu L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data // *Bioinformatics*. Oxford University Press, 2012. Vol. 28, № 23. P. 3150–3152.
127. Pijning A.E. et al. Identification of allosteric disulfides from labile bonds in X-ray structures // *Royal Society open science*. The Royal Society Publishing, 2018. Vol. 5, № 2. P. 171058.
128. Rubinstein R., Fiser A. Predicting disulfide bond connectivity in proteins by correlated mutations analysis // *Bioinformatics*. Oxford University Press, 2008. Vol. 24, № 4. P. 498–504.
129. Timonina D. et al. Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and conformational plasticity in protein superfamilies // *Computational and Structural Biotechnology Journal*. Elsevier, 2021. Vol. 19. P. 1302–1311.
130. Тимонина Д.С., Суплатов Д.А. Анализ множественных выравниваний белков с использованием 3D-структурной информации по ориентации боковых цепей аминокислот // *Молекулярная биология*. 2022. Vol. 56, № 4. P. 663–670.
131. Syakur M.A. et al. Integration k-means clustering method and elbow method for identification of the best customer profile cluster // *IOP conference series: materials science and engineering*. IOP Publishing, 2018. Vol. 336, № 1. P. 012017.
132. Suplatov D., Sharapova Y., Švedas V. EasyAmber: A comprehensive toolbox to automate the molecular dynamics simulation of proteins // *Journal of Bioinformatics and Computational Biology*. World Scientific, 2020. Vol. 18, № 06. P. 2040011.
133. Suplatov D. et al. Human p38 α mitogen-activated protein kinase in the Asp168-Phe169-Gly170-in (DFG-in) state can bind allosteric inhibitor Doramapimod // *Journal of Biomolecular Structure and Dynamics*. Taylor & Francis, 2019. Vol. 37, № 8. P. 2049–2060.
134. <http://biokinet.cmm.msu.ru/zebra3d>.

135. <https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations>.
136. Deng X. et al. Evolution of substrate specificity within a diverse family of β/α -barrel-fold basic amino acid decarboxylases: X-ray structure determination of enzymes with specificity for L-arginine and carboxynorspermidine // *Journal of Biological Chemistry*. ASBMB, 2010. Vol. 285, № 33. P. 25708–25719.
137. Lee J. et al. Phylogenetic diversity and the structural basis of substrate specificity in the β/α -barrel fold basic amino acid decarboxylases // *Journal of Biological Chemistry*. ASBMB, 2007. Vol. 282, № 37. P. 27115–27125.
138. Campbell E., Chuang S., Banta S. Modular exchange of substrate-binding loops alters both substrate and cofactor specificity in a member of the aldoketo reductase superfamily // *Protein Engineering, Design & Selection*. Oxford University Press, 2013. Vol. 26, № 3. P. 181–186.
139. Bollinger A. et al. A novel polyester hydrolase from the marine bacterium *Pseudomonas aestusnigri*—structural and functional insights // *Frontiers in microbiology*. Frontiers Media SA, 2020. Vol. 11. P. 114.
140. Murphy P.M. et al. Alteration of enzyme specificity by computational loop remodeling and design // *Proceedings of the National Academy of Sciences*. National Acad Sciences, 2009. Vol. 106, № 23. P. 9215–9220.
141. Tran D.-T.T., Le L.T., Truong T.N. Discover binding pathways using the sliding binding-box docking approach: application to binding pathways of oseltamivir to avian influenza H5N1 neuraminidase // *Journal of computer-aided molecular design*. Springer, 2013. Vol. 27, № 8. P. 689–695.
142. Le L. et al. Molecular dynamics simulations suggest that electrostatic funnel directs binding of Tamiflu to influenza N1 neuraminidases // *PLoS computational biology*. Public Library of Science San Francisco, USA, 2010. Vol. 6, № 9. P. e1000939.
143. Nilov D.K. et al. Search for Ligands Complementary to the 430-cavity of Influenza Virus Neuraminidase by Virtual Screening // *Supercomputing Frontiers and Innovations*. 2022. Vol. 9, № 2. P. 79–83.
144. Wu Y. et al. Induced opening of influenza virus neuraminidase N2 150-loop suggests an important role in inhibitor binding // *Scientific reports*. Nature Publishing Group, 2013. Vol. 3, № 1. P. 1–8.
145. Amaro R.E. et al. Mechanism of 150-cavity formation in influenza neuraminidase // *Nature communications*. Nature Publishing Group, 2011. Vol. 2, № 1. P. 1–7.

146. Russell R.J. et al. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design // Nature. Nature Publishing Group, 2006. Vol. 443, № 7107. P. 45–49.
147. Шарапова Я. Поиск новых путей регуляции функциональных свойств нейраминидазы NanA как ключевого фермента патогенеза *Streptococcus pneumoniae* с использованием методов компьютерной биологии: Кандидатская диссертация. МГУ имени М.В. Ломоносова, 2021.
148. Wu Y. et al. Bat-derived influenza-like viruses H17N10 and H18N11 // Trends in microbiology. Elsevier, 2014. Vol. 22, № 4. P. 183–191.
149. Schenkmyerova A. et al. Engineering the protein dynamics of an ancestral luciferase // Nature Communications. Nature Publishing Group, 2021. Vol. 12, № 1. P. 1–16.
150. Ploom T. et al. Crystallographic and kinetic investigations on the mechanism of 6-pyruvoyl tetrahydropterin synthase // Journal of molecular biology. Elsevier, 1999. Vol. 286, № 3. P. 851–860.
151. Palacios A.R. et al. The reaction mechanism of metallo- β -lactamases is tuned by the conformation of an active-site mobile loop // Antimicrobial agents and chemotherapy. Am Soc Microbiol, 2019. Vol. 63, № 1. P. e01754-18.
152. Bebrone C. Metallo- β -lactamases (classification, activity, genetic organization, structure, zinc coordination) and their superfamily // Biochemical pharmacology. Elsevier, 2007. Vol. 74, № 12. P. 1686–1701.
153. Hoff R.H. et al. Does positive charge at the active sites of phosphatases cause a change in mechanism? The effect of the conserved arginine on the transition state for phosphoryl transfer in the protein-tyrosine phosphatase from *Yersinia* // Journal of the American Chemical Society. ACS Publications, 1999. Vol. 121, № 41. P. 9514–9521.
154. Amaral M. et al. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding // Nature communications. Nature Publishing Group, 2017. Vol. 8, № 1. P. 1–14.
155. Coquelle N. et al. Activity, stability and structural studies of lactate dehydrogenases adapted to extreme thermal environments // Journal of molecular biology. Elsevier, 2007. Vol. 374, № 2. P. 547–562.
156. Berisio R. et al. Protein titration in the crystal state // Journal of molecular biology. Elsevier, 1999. Vol. 292, № 4. P. 845–854.