МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2023

Работа выполнена в отделе биокинетики НИИ физико-химической биологии имени А.Н. Белозерского МГУ имени М.В. Ломоносова

Научный руководитель – ***Швядас Витаутас-Юозапас Каятоно,*** *доктор химических наук, профессор*

Официальные оппоненты – ***Мирошников Константин Анатольевич***, *доктор химических наук, чл.-корр. РАН, главный научный сотрудник, заведующий лабораторией молекулярной биоинженерии Института биоорганической химии РАН*
Шайтан Константин Вольдемарович, *доктор физико-математических наук, профессор, профессор кафедры биоинженерии биологического факультета МГУ имени М.В. Ломоносова*
Попинако Анна Владимировна, *кандидат биологических наук, научный сотрудник группы молекулярного моделирования Федерального исследовательского центра «Фундаментальные основы биотехнологии» РАН*

Защита диссертации состоится 22 июня 2023 года в 15:30 на заседании диссертационного совета МГУ.015.10 Московского государственного университета имени М.В. Ломоносова по адресу: 119234, Москва, Ленинские горы, д. 1, стр. 73, Факультет биоинженерии и биоинформатики, ауд. 221.
E-mail: dissovet@belozersky.msu.ru
С диссертацией можно ознакомиться в отделе диссертаций Научной библиотеки МГУ имени М.В. Ломоносова (Москва, Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/015.10/2557.

Автореферат разослан «___» мая 2023 г.

Ученый секретарь диссертационного совета, кандидат химических наук — Шаповалова И.В.

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.

ВВЕДЕНИЕ

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. До развития методов биоинформатики и молекулярного моделирования поиск таких участков осуществлялся только экспериментальными методами (Carter et al., 1986; Packer et al., 2015). К сожалению, получение информации о функциях белка с помощью таких 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 (De Juan et al., 2013; Chagoyen et al., 2016). 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 (Bai et al., 2015). 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 утеряны 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. 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.

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 (Valdar, 2002), специфических (Mirny, Gelfand, 2002; Lichtarge et al., 2016) и коррелирующих (Göbel et al., 1994; De Juan et al., 2013) позиций множественных выравниваний аминокислотных последовательностей.

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 (Konagurthu et al., 2006), ParMATT (Shegay et al., 2019), mTM-align (Dong et al., 2018), Matt (Menke et al., 2008) 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 (DeLano, 2002), VMD (Humphrey et al., 1996), ProDy (Bakan et al., 2011). 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 изучения конформаций одного белка, то есть альтернативных положений его структуры, для определения наиболее подвижных частей. Также существуют методы (PSSweb [Gaillard et al., 2016]; visualCMAT [Suplatov D. et al., 2018]), которые используются для визуализации статистики, рассчитанной по множественному (структурно-опосредованному) выравниванию последовательностей и методы (2StrucCompare [Drew et al., 2019] и FATCAT [Li et al., 2020]), 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 (Nilmeier, et al. 2017). 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.

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. 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.

Объект и предмет исследования

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

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.

Теоретическая и практическая значимость работы

Выявленные с использованием разработанного метода 3D-специфические паттерны, как показали результаты исследования, 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.

Методология и методы исследования

Для выявления и анализа структурных паттернов в процессе исследования были разработаны методы и подходы, использующие алгоритмы машинного обучения: DBSCAN (Ester M. et al., 1996), OPTICS (Ankerst et al., 1999), HDBSCAN (McInnes et al., 2017) 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 (Hrabe T. et al., 2016). Для получения множественного выравнивания структур гомологов использовали веб-сервер Mustguseal (Suplatov D.A. et al., 2018) и программу ParMATT (Shegay et al., 2019).

Степень достоверности

Разработанные методы выявления и анализа структурных паттернов были апробированы на конкретных примерах белков и суперсемейств белков 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. 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.

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.

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.

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).

Апробация работы

Результаты исследования были представлены на 5-и конференциях: «Moscow Conference on Computational Molecular Biology» (MCCMB'19 и MCCMB'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).

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.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В работе предложены методы выявления и анализа структурных паттернов суперсемейств белков. Для выявления 3D-специфических паттернов суперсемейства белков был разработан метод (Timonina et al., 2021; Тимонина, Суплатов, 2022), позволяющий выявлять такие части белковых структур, которые схожи (в смысле метрики RMSD) внутри подсемейств, но различаются между ними. Разработанный метод не требует предварительного разделения суперсемейства белков на подсемейства и 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 является результатом случайных колебаний белковых структур. Вводится специальная статистическая модель, позволяющая отделять функционально значимые 3D-специфические паттерны от случайных колебаний белковой структуры. Данный метод имплементирован в качестве программного кода, написанного на языке Python 3. Для 3D-

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

Описание алгоритма поиска 3D-специфических паттернов

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).

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 1). Необходимо учитывать не только содержание гэпов, но и содержание пространственно-смещенных остатков, так как программы для структурного выравнивания (Matt/parMATT) могут помещать в один столбец структурно-опосредованного выравнивания пространственно-смещенные, то есть структурно-невыравненные остатки (расположенные далеко друг от друга в пространстве). Для отделения структурно-невыравненных остатков от выравненных был разработан подход, описанный в абзаце ниже, который позволяет автоматически выбирать пороговое значение (для определения смещены ли аминокислотные остатки друг относительно друга) специально для каждого выравнивания.

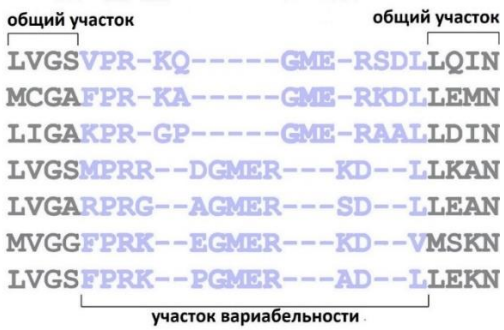

Рисунок 1. Структурно-опосредованное выравнивание последовательностей суперсемейства белков. Все столбцы выравнивания делятся на «общие» и «вариабельности» в зависимости от содержания в них гэпов и структурно-невыравненных аминокислотных остатков. Столбец считается «общим», если содержит суммарно не более 5% гэпов и пространственно-невыравненных 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, такого, в котором число гэпов составляет не более 5%. 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 (из рассчитанных попарных значений в данном столбце выравнивания). Полученные значения 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», чтобы автоматически обнаружить «локоть» графика (аналогично работе Syakur M.A. et al., 2018). Такая точка перегиба указывает на наиболее значительные изменение восходящего тренда метрики RMSD. Ордината точки перегиба берется в качестве порогового значения для различения хорошо и плохо выравненных аминокислотных остатков.

Далее это пороговое значение используется для отделения 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.

На втором этапе алгоритма участки структурного разнообразия основной и боковых цепей подаются на вход методу кластеризации, чтобы разделить эти участки локальной структуры на кластеры, т. е. подсемейства. Таким образом устанавливается являются ли эти участки 3D-специфическими паттернами. В случае выявления 3D-специфических паттернов в **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). Если соответствующие отрезки имеют разную длину (то есть разное количество аминокислотных остатков), меньший из них сопоставляется с 10^3 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 выявления 3D-специфических паттернов в **боковой** цепи рассматриваются «общие» позиции структурно-опосредованного выравнивания (то есть позиции, входящие в «общие» участки). Каждому аминокислотному остатку, принадлежащему «общей» позиции, ставится в соответствие вектор (аналогично работе Nadzirin N. et al., 2012, см. рисунок 2). 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. Полученные матрицы расстояний (в случае основной и боковых цепей) далее подаются на вход методу кластеризации машинного обучения. Предлагается применять алгоритм кластеризации HDBSCAN (McInnes et al., 2017). Также могут быть использованы два альтернативных алгоритма кластеризации: OPTICS (Ankerst et al., 1999) и DBSCAN (Ester M. et al., 1996) или другие.

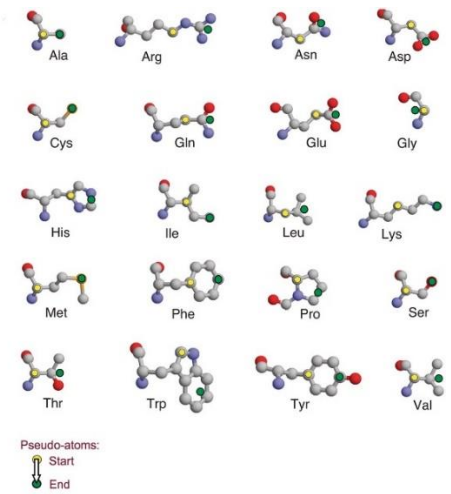

Рисунок 2. Представление боковой цепи в виде вектора, используемое в данной работе, аналогично работе Nadzirin N. et al., 2012. 20 типов аминокислот, для каждой аминокислоты показано расположение псевдоатомов (желтые и зеленые кружки), используемое для представления боковых цепей в виде векторов (начало вектора – желтый кружок, конец вектора – зеленый кружок). Рисунок взят из работы Nadzirin N. et al., 2012.

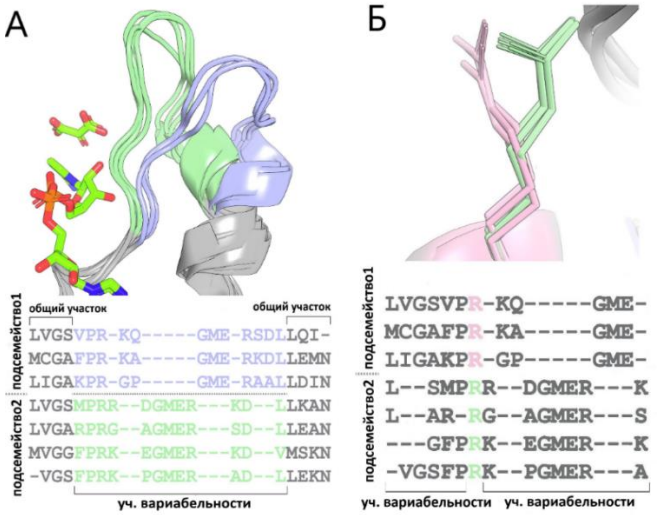

Рисунок 3. Структурное выравнивание (включая кристаллографические лиганды) и соответствующее структурно-опосредованное выравнивание последовательностей. На рисунке А изображен 3D-специфический паттерн, найденный в основной цепи суперсемейства белков (СУОЦ); на рисунке Б изображен 3D-специфический паттерн, найденный в боковой цепи суперсемейства белков (СОБЦ). См. описание в тексте.

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 3А), а в случае боковой цепи 3D-специфический паттерн будем называть *специфической для подсемейств ориентацией боковой цепи* (СОБЦ, рисунок 3Б). Окончательно выбранные 3D-специфические паттерны данного суперсемейства белков ранжируются в порядке убывания *S*-оценки (оценки специфичности) и *Z*-оценки (оценки статистической значимости). СУОЦы и СОБЦы ранжируются отдельно.

Оценка специфичности для 3D-специфического паттерна суперсемейства белков

Оценка специфичности – *S*-оценка – показывает, насколько для данного 3D-специфического паттерна участки цепи белка 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 является результатом 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 (Rousseeuw P.J., 1987) для данного 3D-специфического паттерна (рассчитывается по полученной кластеризации), D – диаметр данного 3D-специфического паттерна, то есть наибольшее расстояние между любыми двумя подсемействами/кластерами этого 3D-специфического паттерна (выбросы не учитываются), Sh_{min} и Sh_{max}, D_{min} и D_{max} – нормировочные коэффициенты, которые выбираются среди всех найденных 3D-специфических паттернов в данном суперсемействе белков и всех 3D-специфических паттернов, найденных в наборах, отобранных из базы данных PDBFlex (см. следующую главу). Бóльшие значения S-оценки 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.

Статистическая модель для отделения функционально-значимых 3D-специфических паттернов суперсемейства белков

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 PDBFlex (Hrabe T. et al., 2016) было отобрано 76 наборов, каждый из которых содержит разные структуры одного белка. К каждому из этих наборов был применен описанный в предыдущей главе метод для получения 3D-специфических паттернов. Далее рассматриваются все получившиеся для данного набора структур 3D-специфические паттерны данного типа (СУОЦы или СОБЦы), для каждого рассчитывается *S*-оценка. Для каждого набора 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». Нами рассматривается именно медианное значение *S*-оценки как характеристика «случайной» пластичности. Мы не рассматриваем максимальное значение S-оценки, так как самые большие и наиболее заметные колебания структуры могут соответствовать функционально-значимым конформационным перестройкам. Предполагая, что *S*-оценка имеет стандартное нормальное распределение, соответствующие значения σ и μ, рассчитанные по S-оценкам «случайных» 3D-специфических паттернов, далее используются для расчета *Z*-оценки статистической значимости и соответствующего значения *P*-оценки найденного в рассматриваемом суперсемействе белков 3D-специфического паттерна. 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.

Разработка программного обеспечения для поиска 3D-специфических паттернов суперсемейства

Описанный теоретический алгоритм был реализован в качестве программ на языке Python3 с использованием принципов ООП. Предложенные программы служат для выявления 3D-специфических паттернов в интересующих пользователя суперсемействах белков. Данные программы можно скачать по ссылкам:

- http://biokinet.cmm.msu.ru/zebra3d – программа для выявления СУОЦов.
- https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations– программа для выявления СОБЦов.

На вход программам подается множественное структурное выравнивание белков суперсемейства, представленное в виде папки с отдельными PDB-файлами, а также 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 parMATT/Mustguseal. В качестве результата работы программ получаем ранжированный список найденных 3D-специфических паттернов, представленный в удобном для дальнейшего экспертного анализа виде, и соответствующие 3D-специфическим паттернам Z-оценки.

Апробация нового подхода на широкой выборке суперсемейств белков

Разработанный нами метод выявления 3D-специфических паттернов был апробирован на суперсемействах белков. 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* фермента (в этом отличие от предыдущего пункта). 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.

Ниже 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*. 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; Z-оценка = 2.37; см. рисунок 4А) представляет собой ранее описанную в литературе (Deng X. et al., 2010; Lee J. et al., 2007) 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.

- *Альдо-кеторедуктазы*. В исследовании альдо-кеторедуктаз был идентифицирован СУОЦ №1 (из 14; *Z*-оценка = 10.11; см. рисунок 4Б), представляющий подвижный участок, расположенный наверху канонической (α/β)8-цилиндрической структуры и соответствующий петле A (Campbell et al., 2013), которая существенно различается между гомологами с различной субстратной специфичностью и участвует в связывании субстрата. Полученные нами кластеры соответствуют ферментам с различной субстратной специфичностью.
- *Гомологи полиэфиргидролазы из Pseudomonas aestusnigri*. При анализе множественного структурного выравнивания гомологов полиэфиргидролазы из *Pseudomonas aestusnigri* нами были идентифицированы 3D-специфические паттерны СУОЦы № 5, 8 (из 13; *Z*-оценка = 3.84 и *Z*-оценка = 1.66; см. рисунок 4В) – структурные фрагменты субстрат-связывающего участка активного центра (Bollinger A. et al., 2020). Автоматически полученная для СУОЦа № 5 кластеризация суперсемейства (для СУОЦа №8 кластеризация аналогичная) соответствует разделению ферментов по субстратной специфичности. Таким образом получены два кластера, первому из которых принадлежат ПЭТ-гидролазы и близкородственные промискуитетные кутиназы, второму – эстеразы, не обладающие ПЭТ-гидролазной активностью и превращающие другие сложные эфиры.
- *Металло-зависимые гидролазы*. Найденный нами при множественном структурном выравнивании металло-зависимых гидролаз СУОЦ №7 (из 22; *Z*-оценка = 3.32; см. рисунок 4Г) представляет собой петлю, участвующую в распознавании субстрата. Мутации этого участка (укорачивание) в гуаниндеаминазе человека привели к тому, что полученный вариант фермента был более активным в отношении аммелида (меньшего по размеру субстрата) и менее активным в отношении гуанина (большего по размеру субстрата), чем фермент дикого типа, так как карман для связывания субстрата уменьшился (Murphy et al., 2009). 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 (бóльшим по размеру субстратам).

- *Суперсемейство нейраминидаз GH34*. В множественном структурном выравнивании суперсемейства нейраминидаз GH34 были найдены СУОЦы №3, 10 (из 19; *Z*-оценка = 6.00 и *Z*-оценка = 2.32), соответствующие функционально важным «петле-430» (Le et al., 2010; Tran et al., 2013; Nilov D.K. et al., 2022) и «петле-150» (Russell et al., 2006; Amaro et al., 2011; Wu et al., 2013), соответственно. Автоматически полученная кластеризация нейраминидаз GH3 для СУОЦа №3 (кластеризация для СУОЦа №10 аналогичная) 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 (Wu et al., 2014), а также гомологи из менее патогенного вируса гриппа B, были отнесены к другим двум подсемействам.

- *Суперсемейство α/β-гидролаз*. При исследовании множественного структурного выравнивания α/β-гидролаз были идентифицированы СУОЦы № 3, 5 (из 14; *Z*-оценка = 3.64 и *Z*-оценка = 2.53), первый из которых включает фрагмент петли L9 и α4-спирали, второй – петлю L14 в структуре люциферазы из *Renilla reniformis*. Эти участки пространственно различаются в гомологах из суперсемейства α/β-гидролаз с различной каталитической активностью, непосредственно влияют на размер входного отверстия туннеля, который соединяет скрытый активный центр с окружающим растворителем, и участвуют в связывании субстрата/продукта (Schenkmayerova A. et al., 2021). Получившиеся кластеры соответствуют α/β-гидролазам с различной каталитической активностью и отличаются размерами и формой входного отверстия туннеля.
- *Гомологи 6-пирувоилтетрагидроптеринсинтазы крысы.* При анализе множественного структурного выравнивания гомологов 6-пирувоилтетрагидроптеринсинтазы крысы был выявлен СОБЦ № 8 (из 12; *Z*-оценка = 0.009). СОБЦ № 8 соответствует аминокислотному остатку Glu107 6-пирувоилтетрагидроптеринсинтазы крысы, участвующему в связывании субстрата. Glu107 формирует солевой мостик с протонированной аминогруппой кольца птерина (субстрата), таким образом закрепляя птерин (Ploom et al., 1999). Автоматическая кластеризация ферментов суперсемейства, полученная для СОБЦа №8, соответствует разделению ферментов по каталитической активности:

получившиеся кластеры соответствуют дигидронеоптеринальдолазам и 6-пирувоилтетрагидро(био)птеринсинтазам.

- *Суперсемейство металло-бета-лактамаз.* При множественном структурном выравнивании металло-бета-лактамаз нами были идентифицированы СУОЦы № 3, 4, 5 (из 13; *Z*-оценка = 8.13, *Z*-оценка = 7.75 и *Z*-оценка = 6.42). СУОЦ №3 соответствует функционально важной петле L3 активного центра, СУОЦы № 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 (Palacios et al., 2019). Автоматическая кластеризация ферментов суперсемейства металло-бета-лактамаз, полученная для СУОЦа №3 (аналогично для СУОЦов №4, 5) соответствуют разделению суперсемейства на различные классы/типы металло-бета-лактамаз.

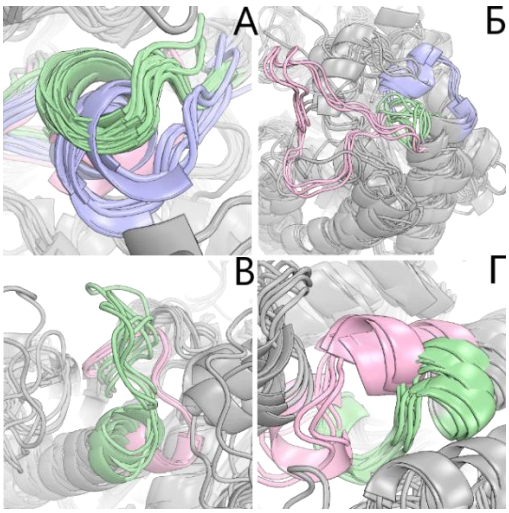

Рисунок 4. Найденные в суперсемействах 3D-специфические паттерны. На рисунке А изображен СУОЦ №8, найденный в суперсемействе пиридоксаль-зависимых ферментов из группы декарбоксилаз основных аминокислот с укладкой типа β/α-цилиндра; на рисунке Б изображен СУОЦ №1, найденный в суперсемействе альдо-кеторедуктаз; на рисунке В изображен СУОЦ №5, найденный в гомологах полиэфиргидролазы из *Pseudomonas aestusnigri*; на рисунке Г изображен СУОЦ №7, найденный в суперсемействе металло-зависимых гидролаз; см. описание в тексте. 3D-специфические паттерны выделены цветом и окрашены в соответствии с автоматически полученной кластеризацией белков суперсемейства по подсемействам. На каждом рисунке для ясности изображено несколько представителей данного суперсемейства.

Ниже приведены примеры суперсемейств белков, в которых найденные 3D-специфические паттерны отвечают за различные положения участка структуры, важные для свойств одного фермента и отличающиеся в различных его функционально-значимых конформациях.

- *Суперсемейство киназ фосфорилаз.* При анализе множественного структурного выравнивания суперсемейства киназ фосфорилаз был идентифицирован СУОЦ №3 (из 11; *Z*-оценка = 9.90; см. рисунок 5А). СУОЦ №3 соответствует активационной

петле p38α MAP-киназы человека, находящейся возле активного центра и содержащей два остатка тирозина, которые в ответ на различные провоспалительные и стрессовые стимулы организма фосфорилируются, что вызывает переход петли в каталитически активную конформацию DFG-in (Suplatov D. et al., 2019). Полученные кластеры соответствуют PDB-файлам, отвечающим за различные структурные положения активационной петли.

- *Суперсемейство протеин-тирозин-фосфатаз.* В исследовании множественного структурного выравнивания суперсемейства протеин-тирозин-фосфатаз был выявлен СОБЦ № 2 (из 5; *Z*-оценка = 2.04; см. рисунок 5Б). 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 (Hoff et al., 1999). Полученные кластеры соответствуют PDB-файлам, отвечающим за различные структурные положения боковой цепи Arg409.
- *Суперсемейство гистидин-киназ-подобных АТФаз.* В исследовании множественного структурного выравнивания суперсемейства гистидин-киназ-подобных АТФаз был выявлен СУОЦ № 3 (из 9; *Z*-оценка = 2.52; см. рисунок 5В). СУОЦ № 3 включает в себя подвижный сегмент крышки и α-спираль3 Hsp90 человека. В зависимости от связанного лиганда α-спираль3 принимает три конформации: конформацию непрерывной спирали, «loop-in» и «loop-out» (Amaral M. et al., 2017). Полученные кластеры соответствуют PDB-файлам с различными положениями сегмента крышки и α-спирали3.
- *Гомологи рибонуклеазы А из Bos taurus.* В исследовании множественного структурного выравнивания гомологов рибонуклеазы А из *Bos taurus* был выявлен СОБЦ № 2 (из 29; *Z*-оценка = 1.44; см. рисунок 5Г), соответствующий His119. Остаток His119 участвует в связывании РНК, в механизме катализе и 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 pH среды в кислую сторону (Berisio et al., 1999). Полученные два кластера соответствуют гош- и транс- конформации остатка.
- *Гомологи лактатдегидрогеназы из Thermus thermophilus.* В исследовании множественного структурного выравнивания гомологов лактатдегидрогеназы из *Thermus thermophilus* был выявлен СУОЦ № 2 (из 21; *Z*-оценка = 3.83), 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 (Coquelle N. et al., 2007). Полученные кластеры соответствуют PDB-файлам с различными структурными положениями подвижного участка.

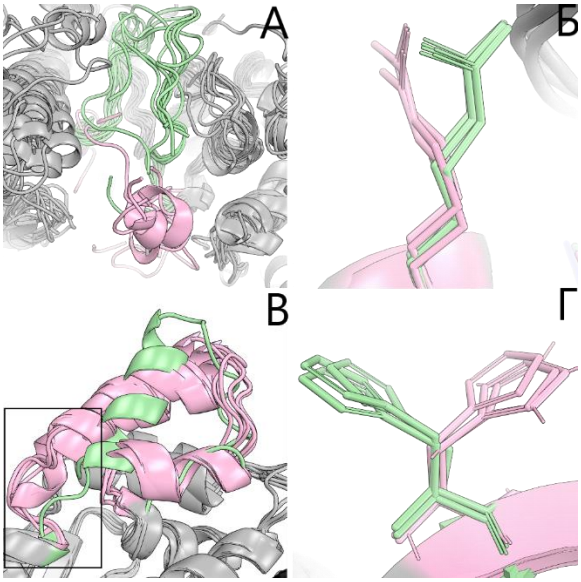

Рисунок 5. Найденные в суперсемействах 3D-специфические паттерны. На рисунке А изображен СУОЦ №3, найденный в суперсемействе киназы фосфорилаз; на рисунке Б изображен СОБЦ №2, найденный в суперсемействе протеин-тирозин-фосфатаз; на рисунке В изображен СУОЦ №3, найденный в суперсемействе гистидин-киназ-подобных АТФаз (выделенный рамкой участок соответствует α-спирали3); на рисунке Г изображен СОБЦ №2, найденный в гомологах рибонуклеазы А из *Bos taurus*; см. описание в тексте. 3D-специфические паттерны выделены цветом и окрашены в соответствии с автоматически полученной кластеризацией белков суперсемейства по подсемействам. На каждом рисунке для ясности изображено несколько представителей данного суперсемейства.

Сравнение результатов применения метода выявления 3D-специфических паттернов с методами выявления специфических позиций подсемейства и коррелирующих позиций на выборке суперсемейств белков

3D-специфические паттерны являются трехмерным аналогом специфических позиций подсемейства/позиций, определяющих специфичность. В этой главе сравниваются результаты применения метода выявления 3D-специфических паттернов с результатами метода выявления специфических позиций подсемейства Zebra2 (Suplatov D. et al., 2020), а также с результатами метода выявления коррелирующих позиций. Метод Zebra2 предоставляет автоматическое деление суперсемейства белков на подсемейства по сходству последовательностей и выявляет специфические позиции подсемейства. Веб-сервер visualCMAT (Suplatov D. et al., 2018) рассчитывает коррелирующие позиции.

Для сравнения результатов методов выявления специфических и коррелирующих позиций с представленным в диссертационной работе методом выявления 3D-специфических паттернов, в суперсемействах белков, описанных в предыдущей главе, были найдены коррелирующие и специфические позиции. Полученную с помощью метода Zebra2 кластеризацию белков для каждого суперсемейства сравнивали с кластеризацией,

полученной для функционально значимого описанного в предыдущей главе 3D-специфического паттерна, найденного в данном суперсемействе с помощью *Adjusted Rand Index* (Hubert, Arabie, 1985; Steinley, 2004). 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 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 равно 0.56). 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 специфические позиции подсемейства не входят (за некоторым исключением) в описанные в предыдущей главе найденные функционально важные 3D-специфические паттерны. И лишь незначительное количество коррелирующих позиций (в среднем 1.5 коррелирующие позиции на один 3D-специфический паттерн) входит в описанные в предыдущей главе 3D-специфические паттерны. Такой результат можно объяснить следующими причинами: 1) Позиции выравнивания, входящие в данный 3D-специфический паттерн, могут быть консервативными по аминокислотной последовательности среди белков суперсемейства (пример – найденный СОБЦ в суперсемействе протеин-тирозин-фосфатаз, другой пример см. на рисунке 6). Такие позиции методом выявления специфических/коррелирующих позиций не могут быть выявлены согласно определению специфических/коррелирующих позиций. 2) 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.

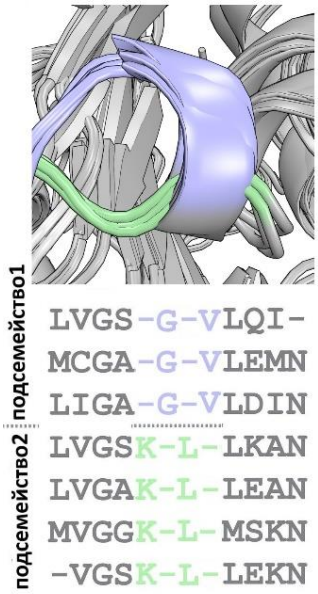

Рисунок 6. Структурное выравнивание и соответствующее структурно-опосредованное выравнивание последовательностей. На рисунке цветом выделен 3D-специфический паттерн. См. описание в тексте.

Из вышесказанного можно сделать 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) представляет кластеризацию белков, где различные кластеры отвечают различным положениям подвижного участка структуры.

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

Найденные в структурах белков базы данных PDB дисульфидные мостики были разделены на кластеры, и, таким образом, выявлены 3D-мотивы дисульфидных мостиков, их типовые уникальные геометрии. Был разработан критерий, позволяющий определить возможность вставки данного 3D-мотива в структуру белка: чтобы после выравнивания атомов основной цепи выбранных для мутации двух аминокислотных остатков с атомами основной цепи данного 3D-мотива оба значения RMSD между двумя парами выравненных друг с другом аминокислотных остатков находились в пределах $X = 0{,}28$ Å. На рисунке 7 изображен пример такого выравнивания, значения RMSD рассчитываются между атомами

основной цепи аминокислотных остатков S1 и E1, а также между S2 и E2. Оба значения RMSD должны быть меньше $X = 0{,}28$ Å. При невыполнении данного критерия пара позиций в структуре белка не рассматривается в качестве целевой для образования дисульфидной связи. Число $X = 0{,}28$ Å соответствует $P - значению = 0.05$ нормального распределения с $\mu = 0{,}16$ Å, $\sigma = 0{,}07$ Å и получено на основе следующей статистической модели: в каждом из найденных 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 наибольшее значение из рассчитанных в этом кластере RMSD. Полученные независимые значения были использованы для создания статистической модели. Чувствительность разработанного метода составила 98%, специфичность – 95% (Suplatov D. et al., 2019).

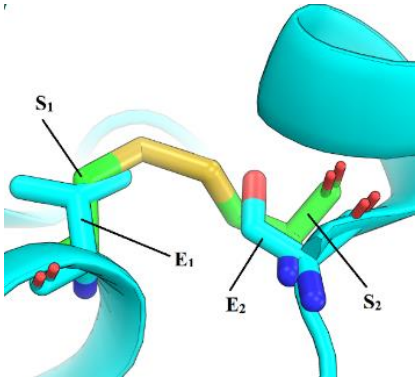

Рисунок 7. См. описание в тексте.

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 (http://biokinet.cmm.msu.ru/zebra3d, https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations), 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.

НАУЧНЫЕ СТАТЬИ ПО ТЕМЕ ДИССЕРТАЦИИ, ОПУБЛИКОВАННЫЕ В ЖУРНАЛАХ SCOPUS, WOS, RSCI[1]

1. **Timonina D.**, Sharapova Y., ŠvedasV., 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; DOI:10.1016/j.csbj.2021.02.005; SJR:6.39 (0.63/0.45).
2. **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; РИНЦ:1.045 (0.38/0.3).
3. 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; DOI: 10.1093/nar/gkz385; SJR:19.36 (0.44/0.2).

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

4. 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; DOI: 10.1142/S021972001840005X; SJR:1.08 (0.94/0.1).

ДРУГИЕ НАУЧНЫЕ РАБОТЫ ПО ТЕМЕ ДИССЕРТАЦИИ

1. **Тимонина Д.С.** Использование 3D-мотивов для компьютерной инженерии дисульфидных связей в белках с учетом структурной подвижности их основной цепи //*Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2019»*., 8-12 апреля 2019, Москва, Россия.

2. Sharapova Y., Suplatov D., **Timonina D.**, Schmalhausen E., Fesko K., Muronets V., Voevodin V., Švedas V. Assessing protein flexibility in computational enzymology: conformational sampling or 3D-motif analysis? //*44th FEBS Congress: From Molecules to Living Systems.* – 6-11 июля 2019 – Краков, Польша.
3. Suplatov D., **Timonina D.**, Sharapova Y., Švedas V. Yosshi: the bioinformatic approach to protein disulfide engineering. //*Moscow Conference on Computational Molecular Biology «MCCMB'19».* – 27-30 июля 2019 – Москва, Россия.
4. **Тимонина Д. С.** Биоинформатический анализ суперсемейств белков на уровне 3D-структуры с использованием методов машинного обучения //*Международная научная конференция студентов, аспирантов и молодых ученых «Ломоносов-2021».* – 12-23 апреля 2021 – Москва, Россия.
5. **Timonina D.**, Sharapova Y., Švedas V., Suplatov D. Bioinformatic analysis of local 3D-structure patterns in protein superfamilies. //*Moscow Conference on Computational Molecular Biology «MCCMB'21».* – Июль 2021 – Москва, Россия.