МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2022, том 56, № 4, с. 663–670

БИОИНФОРМАТИКА

УДК 575.112

АНАЛИЗ МНОЖЕСТВЕННЫХ ВЫРАВНИВАНИЙ БЕЛКОВ С ИСПОЛЬЗОВАНИЕМ 3D-СТРУКТУРНОЙ ИНФОРМАЦИИ ПО ОРИЕНТАЦИИ БОКОВЫХ ЦЕПЕЙ АМИНОКИСЛОТ

© 2022 г. Д. С. Тимонина[a], Д. А. Суплатов[b, *]

[a]*Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, 119234 Россия*

[b]*Научно-исследовательский институт физико-химической биологии им. А.Н. Белозерского, Московский государственный университет им. М.В. Ломоносова, Москва, 119234 Россия*

**e-mail: d.a.suplatov@belozersky.msu.ru*

Поступила в редакцию 31.01.2022 г.
После доработки 25.02.2022 г.
Принята к публикации 02.03.2022 г.

Множественное выравнивание аминокислотных последовательностей гомологичных белков – ключевой инструмент современной биоинформатики и эволюционного анализа. Различия в пространственной ориентации боковых цепей аминокислот могут предопределять существенное функциональное разнообразие представителей одного суперсемейства, однако это обстоятельство никак не учитывают при построении выравниваний и последующем сравнительном анализе. Прежде всего, это связано с недостатками соответствующих алгоритмов, которые опираются на биохимическое сходство "алфавита" аминокислотных замен и либо вообще не используют информацию о 3D-структурной организации белков, либо ограничиваются сравнением остова (атомов основной цепи). Впервые разработано программное обеспечение для систематического исследования специфической ориентации боковых цепей аминокислот в эквивалентных позициях структур гомологов. Программа предназначена для использования в качестве вспомогательного средства при анализе множественных выравниваний аминокислотных последовательностей белков. Новый метод, основанный на алгоритме машинного обучения HDBSCAN, позволяет выявить статистически значимые различия в положении боковых цепей аминокислот в каждой позиции множественного выравнивания и классифицировать их на подсемейства. Метод апробирован на широкой выборке данных. Полученные результаты позволяют говорить о феномене специфической ориентации боковых цепей аминокислот как о достаточно распространенном явлении, требующем дальнейшего изучения и заслуживающем внимания при сравнительном анализе функционально разнообразных суперсемейств белков. Разработанное программное обеспечение находится в свободном доступе по адресу: https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations.

Ключевые слова: множественное выравнивание, биоинформатический анализ, суперсемейство белков, боковая цепь, специфическая позиция, машинное обучение

DOI: 10.31857/S0026898422040139

Сравнительный биоинформатический анализ гомологов, реализующих различные свойства в рамках общей 3D-структурной организации, признан эффективным инструментом изучения структурно-функциональных взаимосвязей в суперсемействах белков и механизмов их действия [1–4]. Один из примеров такого анализа заключается в построении множественного выравнивания гомологов, обладающих разными свойствами, с последующим исследованием аминокислотных замен в столбцах выравнивания и их корреляции с функциональным разнообразием [5, 6]. В большинстве подобных работ используется выравнивание аминокислотных последовательностей белков. Этот подход ведет свою историю с 1960-х годов, когда Эмиль Цукеркандль и Лайнус Полинг впервые высказали гипотезу о "выравнивании" аминокислотных последовательностей и их "побуквенном" сравнении для изучения эволюционных изменений в белках на молекулярном уровне [7]. Использование информации о пространственном расположении аминокислотных остатков позволит улучшить качество выравнивания и интерпретации функциональных особенностей гомологов [4]. Однако до недавнего времени сравнение белков на уровне 3D-структур имело ограниченное применение в практике [8, 9]. Это было связано, во-первых, с относительно небольшим

числом расшифрованных трехмерных структур белков в сравнении с числом известных последовательностей и, во-вторых, с существенно большей вычислительной трудоемкостью 3D-выравнивания.

Взрывное развитие новых методов расшифровки 3D-структур белков и появление новых доступных вычислительных технологий позволило по-новому взглянуть на проблему сравнительного анализа [8, 9]. В последние годы все более активно развиваются новые решения в биоинформатике, направленные на внедрение 3D-данных в повседневную практику [3, 10]. Так, в недавно опубликованных нами работах [11, 12] сделан акцент на методах анализа 3D-структурной информации по эволюционно родственным белкам с разными свойствами. Это обусловлено тем, что рассмотрение не только аминокислотных последовательностей, но и возрастающего объема трехмерных моделей при изучении больших суперсемейств, содержащих эволюционно удаленные гомологи, позволяет существенно повысить качество биоинформатического исследования. Нами разработан комплекс подходов для построения и анализа множественных выравниваний аминокислотных последовательностей с использованием в явном виде структурной информации по представителям функционально разнообразных семейств – метод Mustguseal для построения структурно опосредованных множественных выравниваний суперсемейств белков [11], а также платформа методов Zebra2, pocketZebra, Yosshi и visualCMAT для их анализа [12]. Эти программы позволяют изучать структурно-функциональные взаимосвязи в суперсемействах и улучшать дизайн ферментов и прототипов лекарственных средств. Появление Mustguseal и аналогичных ему алгоритмов (например, MAFFT-DASH [13]) позволило уйти от “классических” множественных выравниваний аминокислотных последовательностей, ориентировавшихся только на “сходство алфавита”, в сторону потенциально более точного сравнения, использующего в качестве дополнительного критерия 3D-структурную информацию о репрезентативных белках. Активно развиваются алгоритмы построения и анализа истинно трехмерных выравниваний белков. Это метод parMatt для ускоренного построения множественных 3D-выравниваний больших выборок белков с использованием суперкомпьютерных технологий [8], метод gaMatt для улучшения качества множественных 3D-выравниваний белков на основе генетического алгоритма и оптимизации направляющего дерева [14]. Одновременно с распространением трехмерных выравниваний появляются и методы для их анализа. Нами недавно предложен первый такой метод Zebra3D – пионерный алгоритм для анализа 3D-выравниваний суперсемейств белков (то есть, именно координат в трехмерном пространстве, а не их “текстовое” отображение), позволяющий находить и классифицировать функционально значимые специфически организованные участки основной цепи в структурах гомологов [4].

Несмотря на возрастающее разнообразие биоинформатических алгоритмов сравнительного анализа белков, использование 3D-структурной информации в повседневной практике по-прежнему ограничено сравнением остова, то есть только атомов основной цепи гомологов. Уже хорошо известно, что отличия в ориентации боковых цепей аминокислот в эквивалентных позициях в структурах гомологов могут предопределять их функциональное разнообразие [4, 15]. Тем не менее ориентация боковых цепей аминокислот пока никак не учитывается при построении выравниваний и последующем сравнительном анализе. В этой работе мы впервые вводим понятие “специфическая ориентация боковых цепей аминокислот”. Оно характеризует такие позиции (“колонки”) во множественном 3D-структурном выравнивании гомологов, в которых атомы основной цепи разных белков расположены в пространстве эквивалентно друг относительно друга, а боковые радикалы аминокислот ориентированы не единообразно и не хаотично, а распределены по нескольким часто встречающимся состояниям, что позволяет классифицировать их на группы (подсемейства). Подобный термин введен нами по аналогии со специфическими позициями подсемейств – такими колонками множественного выравнивания последовательностей, в которых аминокислоты консервативны внутри функциональных подсемейств, но различаются между ними, что может говорить о роли соответствующей позиции в формировании функционального разнообразия [5, 16]. Единственный описанный в литературе аналог специфических позиций на 3D-структурном уровне – “специфические участки подсемейств”, поиск которых реализован в алгоритме Zebra3D [4]. Такие участки могут быть ассоциированы с функциональным разнообразием свойств, а также указывать на функционально значимую конформационную пластичность в суперсемействе. Насколько нам известно, систематических исследований специфической организации боковых радикалов в выравнивании гомологов до настоящего момента не проводилось. В представленной работе описаны первые шаги в изучении этого феномена. В рамках исследования разработано новое программное обеспечение, объединяющее методы машинного обучения и статистического анализа для поиска и классификации специфиче-

ских ориентаций боковых цепей аминокислот. С его помощью проанализирована большая выборка реальных биологических данных.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Алгоритм анализа данных. Новый алгоритм основан на представлении боковых радикалов аминокислот в виде векторов, что позволяет адекватно сравнивать не только идентичные, но и функционально схожие боковые радикалы (например, аспартата и глутамата). Реализация представления боковых цепей аминокислот в виде векторов выполнена в соответствии с ранее описанным протоколом [17]. Например, аминокислота аланин представлена в виде вектора от атома CA в направлении CB, аминокислота глутамин – от координат атома CG в направлении точки, находящейся посредине между атомами NE2 и OE1, триптофан – от координат атома CD1 в направлении точки, находящейся посредине между атомами CZ2 и CZ3, и так далее (имена атомов приведены в соответствии со стандартной схемой формата Protein Data Bank [18]).

На вход новому алгоритму подается выравнивание аминокислотных последовательностей и 3D-структур белков в форматах FASTA и PDB, которые далее анализируют в три этапа: (1) идентификация “общих/консенсусных” и “вариабельных” позиций в 3D-выравнивании гомологов, (2) кластеризация боковых цепей тех аминокислот, которые принадлежат “общим/консенсусным” позициям, и (3) статистический анализ результатов. Первый этап с точки зрения программной реализации идентичен первому этапу алгоритма Zebra3D [4], а второй этап – качественно противоположен. Задачей Zebra3D был поиск участков основной цепи, которые специфически различались в структурах гомологов. Задача нового алгоритма, наоборот, заключается в поиске таких участков выравнивания, в которых основная цепь в гомологах устроена эквивалентно (в случае выравнивания последовательностей для описания подобных состояний колонки выравнивания применяется термин “консенсусный”, или “консервативный”), но которые различаются на уровне ориентации боковых цепей. Каждую “общую/консенсусную” позицию в 3D-выравнивании гомологов представляли в виде набора векторов, специфическую организацию которых анализировали с использованием машинного обучения. Рассчитывали все парные расстояния между векторами. Полученную матрицу расстояний использовали для кластеризации с помощью алгоритма машинного обучения HDBSCAN из библиотеки “hdbscan” [19]. На третьем этапе алгоритма отобранные специфические ориентации боковых цепей ранжировали в порядке убывания S-score (оценка специфичности) и Z-score (оценка статистической значимости), при этом ранжирование по любой из этих двух оценок всегда идентичны. Наиболее визуально заметные и статистически значимые позиции ранжировались первыми. Оценки S-score и Z-score вычисляли в соответствии с протоколами, которые ранее описаны и протестированы на примере алгоритма Zebra3D [4]. Для вычисления оценки статистической значимости (Z-score) использована модель, описанная ниже.

Модель для оценки статистической значимости. Для статистической оценки специфической ориентации боковых цепей аминокислот разработана модель, позволяющая исключить из последующего рассмотрения такие различия в ориентации боковых цепей в PDB-структурах гомологов, которые возникли скорее как результат случайных/тепловых колебаний и, следовательно, не ассоциированы с функцией. Как и в алгоритме Zebra3D, в основу статистической модели для нового алгоритма заложено допущение, что средний уровень конформационной пластичности боковых цепей аминокислот в достаточно широкой выборке структур белков случаен и вряд ли имеет прямое отношение к функции. Использовано 100 наборов данных из базы PDBFlex [20], каждый из которых содержал не менее 20 структур, которые представляли собой различные PDB-записи одного и того же белка (например, 325 PDB-структур p38α MAP-киназы человека). Использованные наборы содержали 26–515 записей PDB на набор (среднее значение – 59). Все структуры в каждом наборе были использованы для построения 3D-структурного выравнивания с использованием parMatt [8]. Выравнивания, в которых на основании экспертного визуального анализа были обнаружены существенные различия между PDB-записями, позволяющими предположить их принадлежность к разным функциональным состояниям одного белка (такие как смещение доменов, очевидная реорганизация протяженных участков структуры в пространстве и др.), исключали из последующего рассмотрения. Оставшиеся выравнивания анализировали с помощью нового алгоритма, описанного выше. Результатом работы алгоритма был ранжированный список позиций выравнивания, в которых была обнаружена специфическая ориентация боковых цепей аминокислот. Медианную специфичность рассматривали как характеристику “случайной” конформационной пластичности, не ассоциированной с функцией, по аналогии с алгоритмом Zebra3D [4]. На основании полученных данных рассчитывали величины σ и μ (среднего и дисперсии нормального распределения, описы-

вающего случайную модель), которые используют при каждом анализе для расчета оценок статистической значимости (Z-score).

Имплементация программы для ЭВМ. Новый алгоритм имплементирован на языке Python3. Соответствующая программа свободно распространяется через ссылку на публичный репозиторий github (https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations). Установка новой программы, требования к формату входных файлов, а также формат и легенда выходных файлов аналогичны таковым для ранее разработанной нами программы Zebra3D, ссылка на документацию к которой приведена в соответствующей публикации [4]. Результатом работы новой программы был список позиций 3D-выравнивания гомологов, в которых была зарегистрирована специфическая ориентация боковых цепей аминокислот, ранжированных по убыванию статистической значимости. По умолчанию выводится не более 10 лучших позиций. Для каждой позиции приводятся следующие характеристики: оценка специфичности, оценка статистической значимости, число кластеров ориентаций боковых радикалов (потенциальных функциональных подсемейств). Также в результате работы программы формируется сессия в формате PyMol (PSE), в которой подсемейства показаны с использованием цветовой легенды.

Апробация и тестирование алгоритма. Для апробации алгоритма была составлена оригинальная выборка на основе базы Catalytic Site Atlas (CSA), содержащей аннотацию каталитически важных аминокислотных остатков в PDB-записях ферментов [21]. Из базы CSA случайно выбрали 195 ферментов, относящихся к разным семействам. 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 этого 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, и, значит, с высокой вероятностью принимают непосредственное участие в его связывании и/или превращении. Подобное допущение можно считать предсказанием, однако его широко используют в мировой биоинформатической практике для тестирования и апробации алгоритмов на большой выборке реальных биологических данных (например, см. [22]) и потому вполне допустимо и в данной работе. Таким образом, в автоматическом режиме была собрана выборка репрезентативных ферментов из различных семейств с аннотацией потенциальных каталитических и субстратсвязывающих остатков в их структурах. Каждый репрезентативный фермент использовали в качестве входных данных для алгоритма Mustguseal для поиска неизбыточного набора из не более 32 структур гомологов в базе PDB со сходством не менее 70% по элементам вторичной структуры и построения соответствующего 3D-выравнивания [11].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

С использованием нового алгоритма исследованы 3D-выравнивания 195 семейств ферментов. Получен список позиций, в которых зарегистрировали статистически значимую специфическую ориентацию боковых цепей аминокислот, из которых 448 позиций содержали каталитически важные и субстратсвязывающие остатки активного центра, проаннотированные нами в структурах репрезентативных ферментов. Для более детального анализа отобрали только такие позиции (“колонки” множественного 3D-структурного выравнивания), которые консервативны по аминокислотной последовательности (т.е. содержали остатки одного типа), однако характеризовались специфической организацией боковых цепей в соответствии с оценкой нового алгоритма. Обнаружили девять таких позиций в восьми различных семействах ферментов. Подробный анализ литературы подтвердил важность каждой идентифицированной позиции для функционирования фермента (табл. 1). Например, Arg409 в структуре тирозиновой протеинфосфатазы из *Yersinia* меняет ориентацию бокового радикала в процессе каталитического акта, что ранее объяснялось образованием новых водородных связей в результате взаимодействия с субстратом [23]. Идентификация этой позиции нашим алгоритмом объясняется тем, что PDB-структуры разных гомологов, случайно выбранных и использованных при построении 3D-выравнивания, были закристаллизованы в разных состояниях (с аналогом субстрата и без него) и на разных стадиях каталитического процесса. Исчерпывающий анализ остальных 439 обнаруженных позиций в структурах белков затруднен отсутствием достаточной информации о роли ориентации ротамеров в механизме их действия. Использованная в работе выборка семейств ферментов, соответствующие 3D-выравнивания и описание результатов их анализа с использованием новой программы предоставляются всем заинтересованным читателям журнала по запросу автору для переписки.

Таблица 1. Позиции множественного 3D-структурного выравнивания гомологов со специфической ориентацией боковых цепей аминокислот, которые соответствуют консервативным каталитически важным и субстратсвязывающим остаткам активного центра, а также критериям статистической значимости

№	Название	PDB	Остаток	Роль
1	Тирозиновая протеинфосфатаза из *Yersinia enterocolitica*	1YTW	ARG409	Участвует в связывании субстрата и каталитическом превращении [23]
2	6-пирувоилтетрагидро-птеринсинтаза крысы	1B66	GLU107	Участвует в связывании субстрата [24]
3	Аргининкиназа из *Limulus polyphemus*	1BG0	LEU187	Участвует в связывании субстрата
			TRP221	Участвует в связывании субстрата
4	S-аденозил-L-гомоцис-теингидролаза из печени крысы	1B3R	THR157	Участвует в связывании субстрата [25]
5	UDP-глюкозо-6-дегидрогеназа из *Streptococcus pyogenes*	1DLI	GLU141	Непосредственно участвует в катализе [26]
6	Рибонуклеаза А крупного рогатого скота	1RUV	HIS119	Непосредственно участвует в катализе [27]
7	Нитрогеназа из *Azotobacter vinelandii*	1N2C	CYS62	Непосредственно участвует в катализе [28]
8	5'-дезокси-5'-метилтио-аденозинфосфорилаза человека	1CG6	MET196	Участвует в связывании субстрата [29]

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В представленной статье сделана попытка систематически проанализировать множественные выравнивания 3D-структур гомологичных белков на уровне пространственной ориентации боковых цепей аминокислот. Сравнительный биоинформатический анализ, как правило, ограничен уровнем аминокислотных последовательностей. В последнее время получили развитие новые подходы, ориентированные на использование 3D-структурной информации относительно основной цепи белков при построении множественных выравниваний [4, 8, 11, 13, 14]. Сравнительный анализ ориентаций боковых радикалов в эквивалентных позициях родственных белков проводят редко – как правило, в режиме визуального экспертного анализа, – и ориентирован он на отдельные частные случаи. В этом плане наша работа – первое систематическое исследование подобного рода. Нами предложен новый алгоритм биоинформатического анализа 3D-структурного выравнивания белков суперсемейства. В результате работы алгоритма получен список позиций (“колонок” 3D-выравнивания), в которых выявлена специфическая ориентация боковых цепей аминокислот, ранжированных по убыванию статистической значимости. Апробация на большой выборке данных продемонстрировала способность нового алгоритма детектировать функционально значимые позиции. Так, в структурах 195 семейств ферментов обнаружено 448 таких позиций, которые содержат каталитически важные и субстратсвязывающие остатки активного центра с боковыми радикалами в специфической ориентации. Недостаток данных о роли сходств и различий конфигурации боковых цепей аминокислот в формировании функционального разнообразия среди гомологов не позволяет нам провести формальную количественную оценку эффективности алгоритма и дать исчерпывающую характеристику каждому идентифицированному остатку.

Проведенное исследование позволяет говорить о том, что феномен специфической ориентации боковых цепей аминокислот не ограничивается частными случаями, а имеет все признаки достаточно распространенного явления, требующего дальнейшего изучения и достойного внимания при сравнительном анализе функционально разнообразных суперсемейств белков. Разработанный в рамках этой работы новый алгоритм и соответствующее программное обеспечение доступны в сети интернет. Программа может быть использована в качестве вспомогательного средства при анализе множественных выравниваний аминокислотных последовательностей гомологичных белков, для которых доступна 3D-структурная информация. Следует отдельно отметить, что потенциал применения этого подхода не ограничива-

ется экспериментально расшифрованными структурами, доступными в базе данных PDB. Последний год запомнился прогрессом в структурной биологии – появлением алгоритмов AlphaFold2 [30] и RoseTTAFold [31], способных с неожиданной точностью предсказывать 3D-структуру белка по его аминокислотной последовательности. Точность предсказания положения боковых цепей этими методами продолжает оставаться предметом дискуссий, однако это не мешает уже сейчас пытаться применять их для тех белков, 3D-структуры которых пока недоступны. В этом контексте свободное распространение нового программного обеспечения представляется нам важным шагом к коллективному решению проблемы изучения взаимосвязи структуры и функции в белках. Систематическое использование предложенной в этой статье программы в повседневной практике научных лабораторий позволит лучше понять взаимосвязь между ориентацией боковых цепей аминокислот в гомологах и механизмом их действия.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 18-29-13060).

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией данной статьи.

Вклад авторов: идея работы и планирование эксперимента (Д.А.С), разработка алгоритма анализа данных и соответствующего программного обеспечения, сбор и обработка данных (Д.С.Т), написание и редактирование рукописи (Д.С.Т., Д.А.С).

СПИСОК ЛИТЕРАТУРЫ

1. Chagoyen M., García-Martín J., Pazos F. (2016) Practical analysis of specificity-determining residues in protein families. *Brief. Bioinform.* **17**, 255–261. https://doi.org/10.1093/bib/bbv045
2. De Juan D., Pazos F., Valencia A. (2013) Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261. https://doi.org/10.1038/nrg3414
3. Marques S., Planas-Iglesias J., Damborsky J. (2021) Web-based tools for computational enzyme design. *Curr. Opin. Struct. Biol.* **69**, 19–34. https://doi.org/10.1016/j.sbi.2021.01.010
4. Timonina D., Sharapova Y., Švedas V., Suplatov D. (2021) Bioinformatic analysis of subfamily-specific regions in 3D-structures of homologs to study functional diversity and conformational plasticity in protein superfamilies. *Comput. Struct. Biotechnol. J.* **19**, 1302–1311. https://doi.org/10.1016/j.csbj.2021.02.005
5. Fesko K., Suplatov D., Švedas V. (2018) Bioinformatic analysis of the fold type I PLP-dependent enzymes reveals determinants of reaction specificity in L-threonine aldolase from *Aeromonas jandaei*. *FEBS Open Bio.* **8**(6), 1013–1028. https://doi.org/10.1002/2211-5463.12441
6. Suplatov D., Sharapova Y., Geraseva E., Švedas V. (2020) Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies. *Nucleic Acids Res.* **48**, W65–W71. https://doi.org/10.1093/nar/gkaa276
7. Zuckerkandl E., Pauling L. (1965) Evolutionary divergence and convergence in proteins. In: *Evolving Genes and Proteins*. Eds Bryson V., Vogel H.J. New York: Academic Press, pp. 97–166. https://doi.org/10.1016/B978-1-4832-2734-4.50017-6
8. Shegay M., Suplatov D., Popova N., Švedas V., Voevodin V. (2019) parMATT: parallel multiple alignment of protein 3D-structures with translations and twists for distributed-memory systems. *Bioinformatics.* **35**(21), 4456–4458. https://doi.org/10.1093/bioinformatics/btz224
9. Suplatov D., Shegay M., Sharapova Y., Timokhin I., Popova N., Voevodin V., Švedas V. (2021) Co-designing HPC-systems by computing capabilities and management flexibility to accommodate bioinformatic workflows at different complexity levels. *J. Supercomput.* **77**, 12382–12398. https://doi.org/10.1007/s11227-021-03691-x
10. Sequeiros-Borja C.E., Surpeta B., Brezovsky J. (2021) Recent advances in user-friendly computational tools to engineer protein function. *Brief. Bioinform.* **22**(3), bbaa150. https://doi.org/10.1093/bib/bbaa150
11. Suplatov D., Kopylov K., Popova N., Voevodin V., Švedas V. (2018) Mustguseal: a server for multiple structure-guided sequence alignment of protein families. *Bioinformatics.* **34**(9), 1583–1585. https://doi.org/10.1093/bioinformatics/btx831
12. Suplatov D., Sharapova Y., Švedas V. (2021) Mustguseal and sister web-methods: a practical guide to bioinformatic analysis of protein superfamilies. In: *Multiple Sequence Alignment.* Ed. Katoh K. Humana Press, New York, pp. 179–200. https://doi.org/10.1007/978-1-0716-1036-7_12
13. Rozewicki J., Li S., Amada K., Standley D., Katoh K. (2019) Mafft-dash: integrated protein sequence and structural alignment. *Nucleic Acids Res.* **47**, W5–W10. https://doi.org/10.1093/nar/gkz342
14. Shegay M., Švedas V., Voevodin V., Suplatov D., Popova N. (2021) Guide tree optimization with genetic algorithm to improve multiple protein 3D-structure alignment. *Bioinformatics.* **38**(4), 985–989. https://doi.org/10.1093/bioinformatics/btab798
15. Li J., Koehl P. (2014) 3D representations of amino acids-applications to protein sequence comparison and classification. *Comput. Struct. Biotechnol. J.* **11**(18), 47–58. https://doi.org/10.1016/j.csbj.2014.09.001
16. Kalinina O., Mironov A., Gelfand M., Rakhmaninova A. (2004) Automated selection of positions determining

functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*. **13**(2), 443–456.
https://doi.org/10.1110/ps.03191704

17. Nadzirin N., Gardiner E., Willett P., Artymiuk P., Firdaus-Raih M. (2012) SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res*. **40**(W1), W380–W386.
https://doi.org/10.1093/nar/gks401
18. Burley S., Berman H., Bhikadiya C., Bi C., Chen L., Di Costanzo L., Christie C., Dalenberg K., Duarte J., Dutta S., Feng Z. (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res*. **47**(D1), D464–D474.
https://doi.org/10.1093/nar/gky1004
19. McInnes L., Healy J., Astels S. (2017) hdbscan: Hierarchical density based clustering. *J. Open Source Softw*. **2**(11), 205.
https://doi.org/10.21105/joss.00205
20. Hrabe T., Li Z., Sedova M., Rotkiewicz P., Jaroszewski L., Godzik A. (2016) PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res*. **44**(D1), D423–D428.
https://doi.org/10.1093/nar/gkv1316
21. Porter C., Bartlett G., Thornton J. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*. **32**, D129–D133.
https://doi.org/10.1093/nar/gkh028
22. Capra J., Singh M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*. **24**(13), 1473–1480.
https://doi.org/10.1093/bioinformatics/btn214
23. Hoff R., Wu L., Zhou B., Zhang Z., Hengge A. (1999) 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*. *J. Am. Chem. Soc*. **121**(41), 9514–9521.
https://doi.org/10.1021/ja992361o
24. Ploom T., Thöny B., Yim J., Lee S., Nar H., Leimbacher W., Richardson J., Huber R., Auerbach G. (1999) Crystallographic and kinetic investigations on the mechanism of 6-pyruvoyl tetrahydropterin synthase. *J. Mol. Biol*. **286**(3), 851–860.
https://doi.org/10.1006/jmbi.1998.2511
25. Kusakabe Y., Ishihara M., Umeda T., Kuroda D., Nakanishi M., Kitade Y., Gouda H., Nakamura K., Tanaka N. (2015) Structural insights into the reaction mechanism of S-adenosyl-*L*-homocysteine hydrolase. *Sci. Rep*. **5**, 16641.
https://doi.org/10.1038/srep16641
26. Ge X., Penney L., Van De Rijn I., Tanner M. (2004) Active site residues and mechanism of UDP-glucose dehydrogenase. *Eur. J. Biochem*. **271**(1), 14–22.
https://doi.org/10.1046/j.1432-1033.2003.03876.x
27. Kasireddy C., Ellis J., Bann J., Mitchell-Koch K. (2016) Tautomeric stabilities of 4-fluorohistidine shed new light on mechanistic experiments with labeled ribonuclease A. *Chem. Phys. Lett*. **666**, 58–61.
https://doi.org/10.1016/j.cplett.2016.10.072
28. Igarashi R., Seefeldt L. (2003) Nitrogen fixation: the mechanism of the Mo-dependent nitrogenase. *Crit. Rev. Biochem. Mol. Biol*. **38**(4), 351–384.
https://doi.org/10.1080/10409230391036766
29. Guan R., Tyler P., Evans G., Schramm V. (2013) Thermodynamic analysis of transition-state features in picomolar inhibitors of human 5'-methylthioadenosine phosphorylase. *Biochemistry*. **52**(46), 8313–8322.
https://doi.org/10.1021/bi401188w
30. Cramer P. (2021) AlphaFold2 and the future of structural biology. *Nat. Struct. Mol. Biol*. **28**(9), 704–705.
https://doi.org/10.1038/s41594-021-00650-1
31. Baek M., DiMaio F., Anishchenko I., Dauparas J., Ovchinnikov S., Lee G., Wang J., Cong Q., Kinch L.N., Schaeffer R.D., Millán C., Park H., Adams C., Glassman C.R., DeGiovanni A., Pereira J.H., Rodrigues A.V., van Dijk A.A., Ebrecht A.C., Opperman D.J., Sagmeister T., Buhlheller C., Pavkov-Keller T., Rathinaswamy M.K., Dalwadi U., Yip C.K., Burke J.E., Garcia K.C., Grishin N.V., Adams P.D., Read R.J., Baker D. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **373**(6557), 871–876.
https://doi.org/10.1126/science.abj8754

ANALYSIS OF MULTIPLE PROTEIN ALIGNMENTS USING 3D-STRUCTURAL INFORMATION ON THE ORIENTATION OF AMINO ACID SIDE-CHAINS

D. S. Timonina[1] and D. A. Suplatov[2, *]

[1] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, 119234 Russia
[2] Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, 119234 Russia
**e-mail: d.a.suplatov@belozersky.msu.ru*

Multiple alignment of amino acid sequences of homologous proteins is a key tool in state-of-the-art bioinformatics and evolutionary analysis. Differences in the spatial orientation of amino acid side-chains can predetermine a significant functional diversity in members of one superfamily; however, this is usually not taken into account in any way when constructing alignments and during subsequent comparative analysis. First of all, this is due to the limitation of existing algorithms, which are guided by the biochemical similarity of the

"alphabet" of amino acid substitutions and either do not use information about 3D-structural organization of proteins at all, or are limited to comparing the backbone only (i.e. atoms of the main-chain). In this work, for the first time we introduce the software for a systematic analysis of specific orientation of amino acid side-chains in equivalent positions of homologous protein structures. The program is intended to assist the analysis of protein multiple sequence alignments. The new algorithm, based on machine learning HDBSCAN technique, can identify statistically significant differences in the side-chain orientations and classify them into subfamilies at each position of multiple alignment. The method has been tested on a wide set of real biological data. The results obtained allow us to speak of the specific orientation of amino acid side-chains as a common phenomenon that requires further study and deserves attention in a comparative analysis of functionally diverse protein superfamilies. The developed software is freely available at https://github.com/TimoninaDaria/Subfamily-Specific-Sidechain-Orientations.

Keywords: multiple alignment, bioinformatics analysis, protein superfamily, side-chain, specific position, machine learning