

**МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
имени М.В.ЛОМОНОСОВА**

---

**ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ**

## **Что способствует нуклеации?**

Проект выполнили студенты 4-го курса  
**Смородина Ева и Власов Антон**  
в рамках курса по структурной  
биоинформатике

**Москва**

**2020-2021 гг.**

## Введение

Одним из ключевых свойств живых систем является способность структур их молекулярных компонентов самоорганизовываться с высокой точностью и эффективностью. Процесс сворачивания белка или фолдинга из одномерной структуры в трехмерную - универсальный и фундаментальный пример такой самоорганизации. Основным результатом фолдинга является пространственное сближение функциональных групп, которое позволяет макромолекулам выполнять белкам самые разнообразные функции. Правильное сворачивание белка необходимо для его специфичности и стабильности. Соответственно, изучение процессов фолдинга необходимо как для фундаментального понимания устройства и функционирования живых систем, так и для применения выявленных принципов в практических целях.

Левинталь рассчитал, что существует астрономическое число конформаций белка, возникающих из принципов комбинаторики. Перебор по всем из них занял бы "бесконечное" время, при этом большинство белков сворачивается в течение миллисекунд или даже микросекунд, что говорит о необходимости существования какого-либо пути, который бы значительно сокращал данный перебор. Логичным предположением ученых была гипотеза о существовании серии промежуточных состояний, т.е. интермедиатов, которые образует белок, прежде, чем перейти из полностью развернутого состояния в полностью свернутое. И хотя для некоторых белков такие промежуточные состояния были показаны, про многие коротких белков (длиной от 60 до 100 остатков) утверждают, что для них существует только 2 состояния: полностью свернутое и полностью развернутое, т.е. не происходит накопления интермедиатов в состоянии равновесия или на пути фолдинга. Вместо этого для них речь идёт о единственном переходном состоянии с повышенной энергией, достижение которого является лимитирующей стадией фолдинга. Множество исследований предполагают, что фундаментальный механизм фолдинга включает в себя образование так называемого ядра фолдинга, которое формирует небольшое число остатков, после чего происходит быстрая конденсация оставшейся структуры. Взаимодействие остатков ядра фолдинга друг с другом называют процессом нуклеации [1, 2].

Экспериментально остатки, образующие переходное состояние, то есть ядро нуклеации, можно определить путём сравнения так называемых  $\phi$ -значений. С помощью направленного мутагенеза точно заменяют некоторые остатки, которые важны для стабилизации нативной структуры. После этого измеряют  $\Delta\Delta G_{N \rightarrow D}$  и  $\Delta\Delta G_{\# \rightarrow D}$ , то есть изменения свободной энергии фолдинга и образования переходного состояния, соответственно, по сравнению с исходным белком дикого типа.  $\Delta G_{\# \rightarrow D}$  определяют из измеренных кинетических констант реакции фолдинга. Для данного метода необходимо наличие способа отличать денатурированное состояние от нативного, например, с помощью флуоресценции нативной формы исследуемого белка. Эта необходимость накладывает существенные ограничения на описываемый метод. Числом  $\phi$  для фолдинга называют отношение  $\phi = \Delta\Delta G_{\# \rightarrow D} / \Delta\Delta G_{N \rightarrow D}$ . Число  $\phi$  близкое к нулю означает, что рассматриваемый остаток (или группа остатков в случае

множественного мутагенеза) в момент образования переходного состояния находится в денатурированной части белка. Если  $\phi$  близко к 1, то это означает, что мутация влияет на переходное состояние так же, как она влияет на полностью свернутое состояние, что говорит о том, что рассматриваемый остаток в момент переходного состояния находится в свернутом виде, то есть входит в ядро нуклеации. Затем проводят верификацию найденных позиций уже через эксперименты с парными мутациями и глицин-аланиновое сканирование [2].

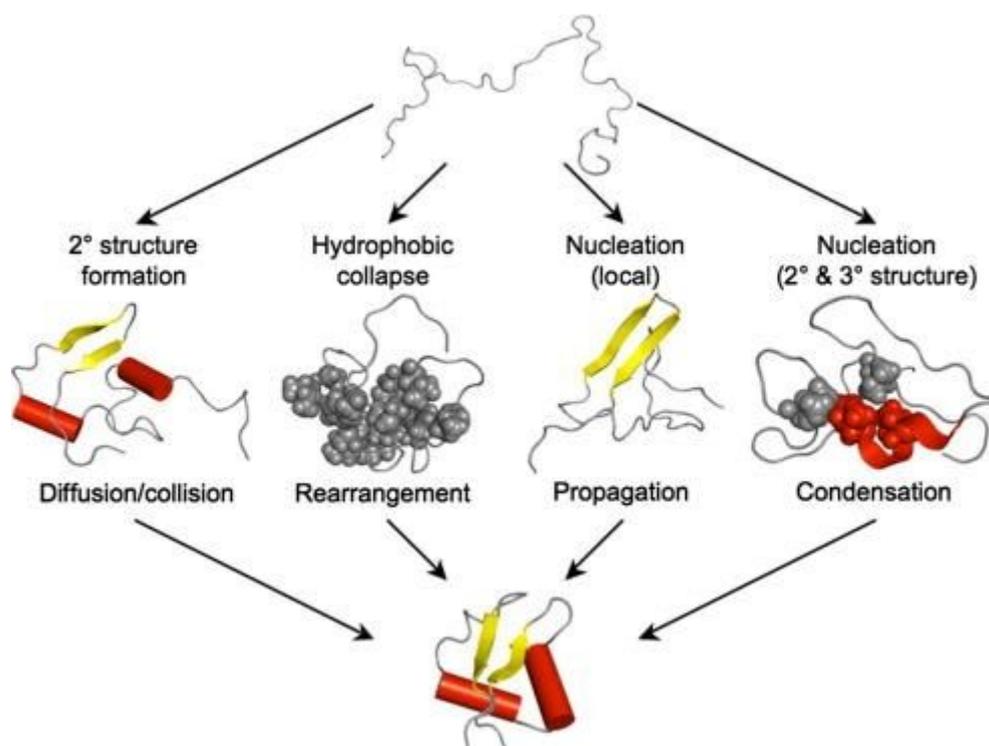
Другим методом определения остатков ядра нуклеации является водородно-дейтериевый обмен. Измеряют кинетику замещения водорода в NH-группах остова белка. При этом некоторые остатки меньше остальных обмениваются водородом с растворителями в ходе фолдинга и оказываются раньше всех защищенными от растворителя. Было показано, что описанные остатки также обладают высокими  $\phi$ -значениями, что позволяет применять данный метод для определения остатков, входящих в ядро нуклеации белка [3]. Далее будем называть остатки в структурах, определённые описанными выше методами, сайтами или ядрами нуклеации.

Целью данной работы стало изучить закономерности, связанные с ядрами нуклеации вторичной структуры и их окружения, а именно определить, какие остатки и почему наиболее склонны входить в состав ядер нуклеации, а также выяснить, есть ли взаимосвязь между остатками в самом ядре нуклеации и соседствующих с ними.

Существует множество теорий нуклеации, однако только четыре из них наиболее распространены. Их можно считать классическими поскольку подавляющее большинство работ, связанных с нуклеацией или фолдингом, так или иначе на них опирается или ссылается (рисунок 1) [4]:

- framework (diffusion/collision) model – первыми формируются локальные элементы вторичной структуры, затем они диффундируют вместе, сталкиваются
- hydrophobic collapse model – белок быстро сворачивается вокруг своих гидрофобных боковых цепей, а затем перестраивается из ограниченной конформации этого промежуточного продукта «расплавленная глобула»
- nucleation propagation model - локальные взаимодействия образуют небольшое количество естественной вторичной структуры, которая действует как ядро для распространения наружу дополнительной собственной структуры
- nucleation-condensation model – наличие метастабильного ядра, которое не может запускать сворачивание, пока не накопится достаточное количество стабилизирующих дальнедействующих взаимодействий; как только это происходит, нативная структура конденсируется так быстро, что ядро еще не полностью сформировано в переходном состоянии

Рассмотрим их более подробно.



**Рисунок 1.** Четыре классических механизма нуклеации. Слева направо: diffusion/collision, hydrophobic collapse, nucleation propagation и nucleation-condensation [4].

Теория гидрофобного коллапса (hydrophobic collapse) утверждает, что зарождающийся полипептид формирует исходную вторичную структуру, создавая локализованные области преимущественно гидрофобных остатков. Иными словами выходящий из рибосомы полипептид начинает взаимодействовать с водой, создавая термодинамическое давление (гидрофобный эффект) на гидрофобные области. Это давление способствует агрегации этих остатков в гидрофобное ядро. После этого полярные остатки начинают торчать наружу, что благоприятно для взаимодействия с водой. Таким образом, получается третичная структура белка, в которой обращенная к растворителю поверхность состоит из гидрофильных остатков, а внутренняя – из гидрофобных [5].

Эта теория широко известна. Гидрофобный коллапс в глобулярных белках происходит в наносекундном временном диапазоне. Этот процесс сильно ускоряет уменьшение молекул воды в междоменных областях белка [6]. Было выяснено, что данный механизм можно разделить на два подпроцесса: непосредственно гидрофобный коллапс и образование первичных элементов третичной структуры [7]. Для образования гидрофобного ядра требуется, чтобы поверхностные остатки свернувшейся части белка контактировали как с полярными (гидрофильными) молекулами растворителя, так и с неполярными (гидрофобными) остатками внутри ядра. В связи с этим, аминокислоты агрегата должны быть амфифильными. Так, например,  $\alpha$ -спираль может иметь неполярные остатки в  $i + 3$ ,  $i + 4$  положениях, что позволяет ей проявлять гидрофобные свойства с одной стороны и гидрофильные – с другой [8].

В рамках теории нуклеации-конденсации (nucleation-condensation) считается, что реакция сворачивания инициируется образованием ядра, которое имеет высокую стабильность из-за наличия некоторого количества правильных взаимодействий, образующих вторичную структуру. В дальнейшем это ядро может служить шаблоном для быстрой конденсации остальных частей белка вокруг него. Таким образом снижается количество конформаций, которые должны быть отобраны в реакции нуклеации, которая в противном случае была бы астрономически большой [9].

На сегодняшний день теория нуклеации-конденсации очень популярна. Несмотря на то, что существенные доказательства этого механизма были получены (наличие четко выраженного единственного ядра сворачивания, уменьшение диаметра молекул в ходе фолдинга, одновременное сворачивание и наращивание контактов вторичной структуры) [9], охарактеризовать начальные события, ведущие к образованию ядра сворачивания все еще остается сложной задачей. Утверждается, что остатки, составляющие ядро сворачивания, имеют тенденцию взаимодействовать в денатурированном состоянии временным образом, а не одновременно, тем самым формируя неполные и искаженные версии ядра сворачивания. Только когда все контакты между этими ключевыми остатками в конечном итоге сформированы, белок может достичь переходных состояний (одного или нескольких, в литературе встречаются различные данные, но чаще таких состояний может быть больше, чем одно) и продолжить сворачивание [10].

Теория фреймворка, также известная как теория диффузии/коллизии, (framework, diffusion/collision) предполагает наличие в белковой молекуле нескольких фрагментов (микродоменов), каждый из которых достаточно мал для быстрого (за время, сильно меньшее, чем заявлено в парадоксе Левинтала, большее в источнике не указано) поиска всех конформационных альтернатив. Поскольку в разных белках. Это условие делает вторичную структуру микродоменов нестабильной. Следовательно, несколько (два или более) микродоменов должны столкнуться друг с другом и слиться в новую единую вторичную структуру для обретения стабильности [11, 12].

Возможно, это теория наименее распространена и известна, однако есть ряд работ, свидетельствующих о том, что она имеет право на существование. С момента создания теории (1976) были подтверждены многие ее аспекты: наличие вторичной структуры в небольших пептидах и фрагментах белка, зависимость скоростей сворачивания от вязкости растворителя и важность нативных путей взаимодействия микродомен-микродомен (например, ранние межспиральные взаимодействия) в складывании [13]. Было показано, что модель фреймворка согласуется с теорией нуклеации-конденсации, т. к. высокую скорость сворачивания белка можно объяснить каталитическим эффектом образования кластеров остатков, которые имеют наибольшую склонность для раннего образования регулярных вторичных структура [14].

Согласно теории распространения нуклеации (nucleation propagation) сначала в локализованной области полипептидной цепи формируется короткий участок

вторичной структуры. Этот процесс происходит с затратами свободной энергии и по сути является нуклеацией. После этого образовавшийся тип вторичной структуры распространяется в стороны от сайта нуклеации. Оба процесса влекут за собой энтальпийно-энтропийную компенсацию, которая возникает из-за потери конформационной энтропии при образовании вторичной структуры с сопутствующим выигрышем в благоприятных взаимодействиях [15].

Данная теория является спорной, т. к. существуют работы как подтверждающие, так и опровергающие ее. Например, Чжоу и Фасман [16] показали, что нуклеация вторичной структуры X начинается при наличии в кластере четырех из шести аминокислот с сильной склонностью к образованию X. После этого запущенная нуклеация распространяется в обоих направлениях почти независимо от склонности к образованию X соседних аминокислотных остатков, пока не завершится последовательностью “выключателей” с очень низкой склонностью к X. Однако Тумадже и Джонсон [17], повторив тот же эксперимент, что и Чжоу и Фасман, но с другими последовательностями, не получили никаких данных свидетельствующих о том, что за нуклеацией аминокислот со склонностью к образованию  $\alpha$ -спиралей не следует нуклеация соседних участков.

Также существуют и другие модели механизма нуклеации. Например, одна модель утверждает, что нуклеация  $\alpha$ -спирали начинается с N-конце и затем распространяется по полипептиду к C-концу [18]. Кроме того существует мнение, что разные белки сворачиваются разными путями и, следовательно, имеют разные или многостадийные механизмы нуклеации, которые могут включать перечисленные выше теории [19].

## **Что способствует нуклеации?**

Несмотря на многочисленные исследования в области нуклеации белков, зависимость между процессом нуклеацией и типом аминокислот в окрестностях точки ее начала не была изучены. Литературные данные ограничиваются только информацией о предпочтительности нахождения той или иной аминокислоты в одном из типов вторичных структур.

Чоу и Фасман изучали количество вхождений аминокислот в тот или иной тип вторичной структуре. Исходя из собранных данных они для каждой аминокислоты рассчитали конформационные параметры с учетом относительной частоты данной аминокислоты в белке, ее присутствия в имеющемся типе вторичной структуры и доли остатков, встречающихся в этом типе структуры (таблица 1). Этот конформационный параметр ( $P_\alpha$ ,  $P_\beta$ ), по сути, является мерой предпочтения данной аминокислоты находиться в  $\alpha$ -спирали или  $\beta$ -листе.  $P_\alpha$  и  $P_\beta$  предположительно содержат информацию о физико-химических параметрах, определяющих стабильность белка, таких как гидрофобность, должным образом взвешенная с учетом их относительной важности. Предполагалось, что эти параметры могут быть полезны для прогнозирования вторичной структуры белка на основе аминокислотная последовательность [20]. Весь диапазон получившихся значений был разбит на 6

категорий и вместе с ним аминокислоты были подразделены на 6 групп в зависимости от того, насколько высокие значения предпочтительности они имеют: сильные и слабые инициаторы, безразличные b сильные и слабые терминаторы. Границы диапазонов условны и могут меняться (в таблице 1 указаны те диапазоны, которые изначально использовались). Можно отметить, что Glu, Met, Ala и Leu наиболее сильные формователи (инициаторы)  $\alpha$ -спирали, за ними по силе идут Lys, Phe, Gln, Trp, Ile и Val. В  $\beta$ -листе другое распределение: самыми сильными инициаторами являются Val, Ile и Tyr а затем – Phe, Trp, Leu, Cys, Thr, Gln и Met (таблица 1).

**Таблица 1.** Предпочтительность аминокислот, находящихся в  $\alpha$ -спирале и  $\beta$ -листе.

Предпочтительность обозначается как  $P_x$  - где  $x = \alpha$  соответствует спирали, а  $x = \beta$  – листу. фигурными скобками отмечены группы остатков, разделенные по степени их предпочтительности к тому или иному типу вторичной структуры:  $H_\alpha/H_\beta$  – сильные инициаторы вторичной структуры,  $h_\alpha/h_\beta$  – инициаторы средней силы,  $l_\alpha/l_\beta$  – слабые инициаторы,  $i_\alpha/i_\beta$  – аминокислоты, безразличные к типу вторичной структуры,  $B_\alpha/B_\beta$  – терминаторы вторичной структуры,  $V_\alpha/V_\beta$  – сильные терминаторы [21]. Инициаторами вторичной структуры являются те аминокислоты, которые предположительно вносят положительный вклад в образование вторичной структуры определенного типа, т. к. часто в ней встречаются (имеют высокие значения предпочтительности).

Безразличные к образованию вторичной структуры остатки никак появляются в спиральях и тяжах примерно одинаково (имеют средние значения предпочтительности). Аминокислоты-терминаторы являются завершающими остатками вторичной структуры, т. е. после них данный тип вторичной структуры прерывается (имеют низкие значения предпочтительности)

Остатки $\alpha$ -спирали	$P_\alpha$	Значимость для $\alpha$ -спирали	Остатки $\beta$ -листа	$P_\beta$	Значимость для $\beta$ -листа
Glu	1.53	$H_\alpha$	Met	1.67	$H_\beta$
Ala	1.45	$H_\alpha$	Val	1.65	$H_\beta$
Leu	1.34	$H_\alpha$	Ile	1.60	$H_\beta$
His	1.24	$h_\alpha$	Cys	1.30	$h_\beta$
Met	1.20	$h_\alpha$	Tyr	1.29	$h_\beta$
Gln	1.17	$h_\alpha$	Phe	1.28	$h_\beta$
Trp	1.14	$h_\alpha$	Gln	1.23	$h_\beta$
Val	1.14	$h_\alpha$	Leu	1.22	$h_\beta$
Phe	1.12	$h_\alpha$	Thr	1.20	$h_\beta$
Lys	1.07	$l_\alpha$	Trp	1.19	$h_\beta$
Ile	1.00	$l_\alpha$	Ala	0.97	$l_\beta$

Asp	0.98	$i_{\alpha}$	Arg	0.90	$i_{\beta}$
Thr	0.82	$i_{\alpha}$	Gly	0.81	$i_{\beta}$
Ser	0.79	$i_{\alpha}$	Asp	0.80	$i_{\beta}$
Arg	0.79	$i_{\alpha}$	Lys	0.74	$b_{\beta}$
Cys	0.77	$i_{\alpha}$	Ser	0.72	$b_{\beta}$
Asn	0.73	$b_{\alpha}$	His	0.71	$b_{\beta}$
Tyr	0.61	$b_{\alpha}$	Asn	0.65	$b_{\beta}$
Pro	0.59	$B_{\alpha}$	Pro	0.62	$b_{\beta}$
Gly	0.53	$B_{\alpha}$	Glu	0.26	$B_{\beta}$

Вычислив конформационные параметры (таблица 1), Чоу и Фасман сформулировали набор эмпирических правил для предсказания вторичной структуры белка. Разработка этих эмпирических правил проводилась исходя из основных соображений о структуре белка. Эти правила, когда примененные Чоу и Фасманом, обеспечили точность предсказания на 70-80%.

Правила никогда не разрабатывались как компьютерный алгоритм и, следовательно, не имеют такого строгого определения, которое требуется для компьютерного алгоритма. Это привело к большому разнообразию реализаций, которые имеют столь же широкий диапазон точности [20, 21]. Также, представленные в таблице 2 правила были сформированы в 1974 году и с того момента многое изменилось в области структурной биологии, однако за прошедшие 45 лет ничего нового в данной сфере не было сказано.

**Таблица 2.** Эмпирические правила нуклеации  $\alpha$ -спиралей и  $\beta$ -листов статистического алгоритма Чоу-Фасмана для предсказания вторичной структуры белков [21]

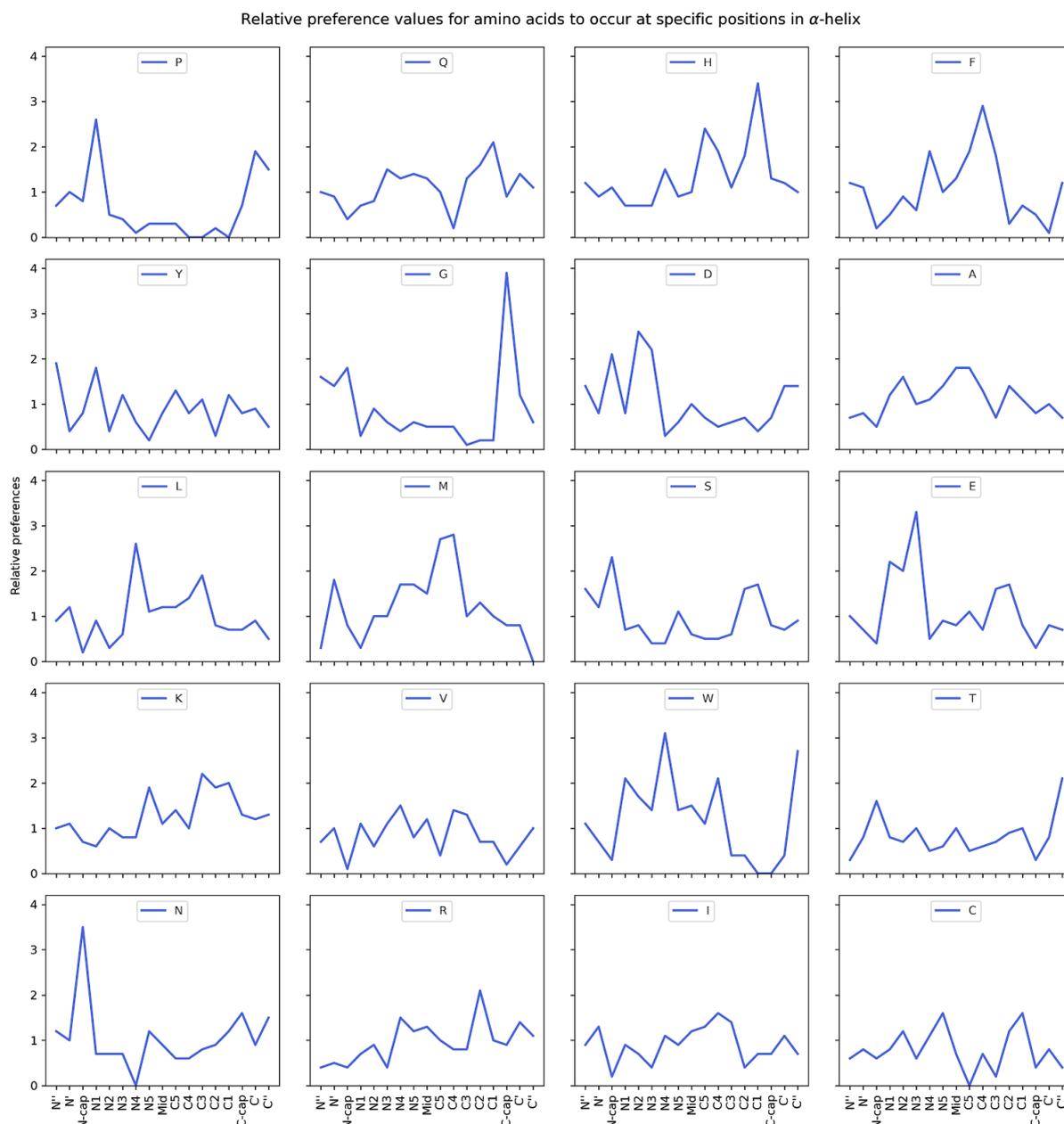
<b><math>\alpha</math>-спираль</b>	<b><math>\beta</math>-лист</b>
Нуклеация	
Для нуклеации необходима последовательность четырех из шести аминокислот с хорошей предпочтительностью к $\alpha$ -спирале ( $H_{\alpha}$ или $h_{\alpha}$ , таблица 1)	Для нуклеации необходим либо кластер из трех либо последовательность трех из пяти аминокислот с хорошей предпочтительностью к $\beta$ -листу ( $H_{\beta}$ или $h_{\beta}$ , таблица 1)
Распространение вторичной структуры	
Спиральный сегмент расширяется в обоих направлениях до тех пор, пока в последовательности не появятся наборы тетрапептидных прерывателей	Листовой сегмент расширяется в обоих направлениях до тех пор, пока в последовательности не появятся наборы тетрапептидных прерывателей $\beta$ -листов

$\alpha$ -спиралей (аминокислоты, чья предпочтительность $P_\alpha < 1.00$ , таблица 1)	(аминокислоты, чья предпочтительность $P_\beta < 1.00$ , таблица 1)
Pro не может находиться во внутренней спирали или на C-концевом спиральном конце, но может находиться в пределах последних трех остатков на N-концевом конце	Любой сегмент с $P_\beta > 1.05$ или с $P_\beta > P_\alpha$ предсказывается как лист (таблица 1)
Любой сегмент с $P_\alpha > 1.03$ или с $P_\alpha > P_\beta$ предсказывается как лист (таблица 1)	

Помимо предпочтительности аминокислот к типам вторичной структуры для спиралей также известна зависимость типов остатков от приближенности позиции к N- или C-концу (рисунки 3, 4). Видно, что распределение положительно и отрицательно заряженных аминокислот в первом и последнем витке ассиметрично. Помимо этого заметно предпочтение Asn находиться в позиции N-cap, а Pro – в N-cap + 1. Кроме того Gly явно превосходит все остальные остатки по расположению в C-cap, а гидрофобные аминокислоты склонны находиться в положениях N-cap + 4 и C-cap - 4 [22].



**Рисунок 3.** Предпочтительность аминокислот располагаться в определенных местах  $\alpha$ -спирали. В качестве позиций рассматривались N- и C-конец (N-cap, C-cap), до 5 остатков внутрь спирали от концов (N1, N2, N3, N4, N5, C1, C2, C3, C4, C5), центр (Mid), как позиции, расположенные дальше 5 остатков от концов, и по два остатка за пределами N- и C-конца (N', N'', C', C'') [21, 22]



**Рисунок 4.** Зависимость типа остатка от положения в  $\alpha$ -спирали. По вертикали отображена предпочтительность аминокислоты находиться в позиции, отмеченной по горизонтали. Чем больше предпочтительность, тем чаще аминокислота встречается в данном месте [22]

Для самостоятельного изучения зависимости между аминокислотными остатками в области сайта нуклеации и образующимся типом вторичной структуры мы собрали датасет из 182 записей для 20 различных белков

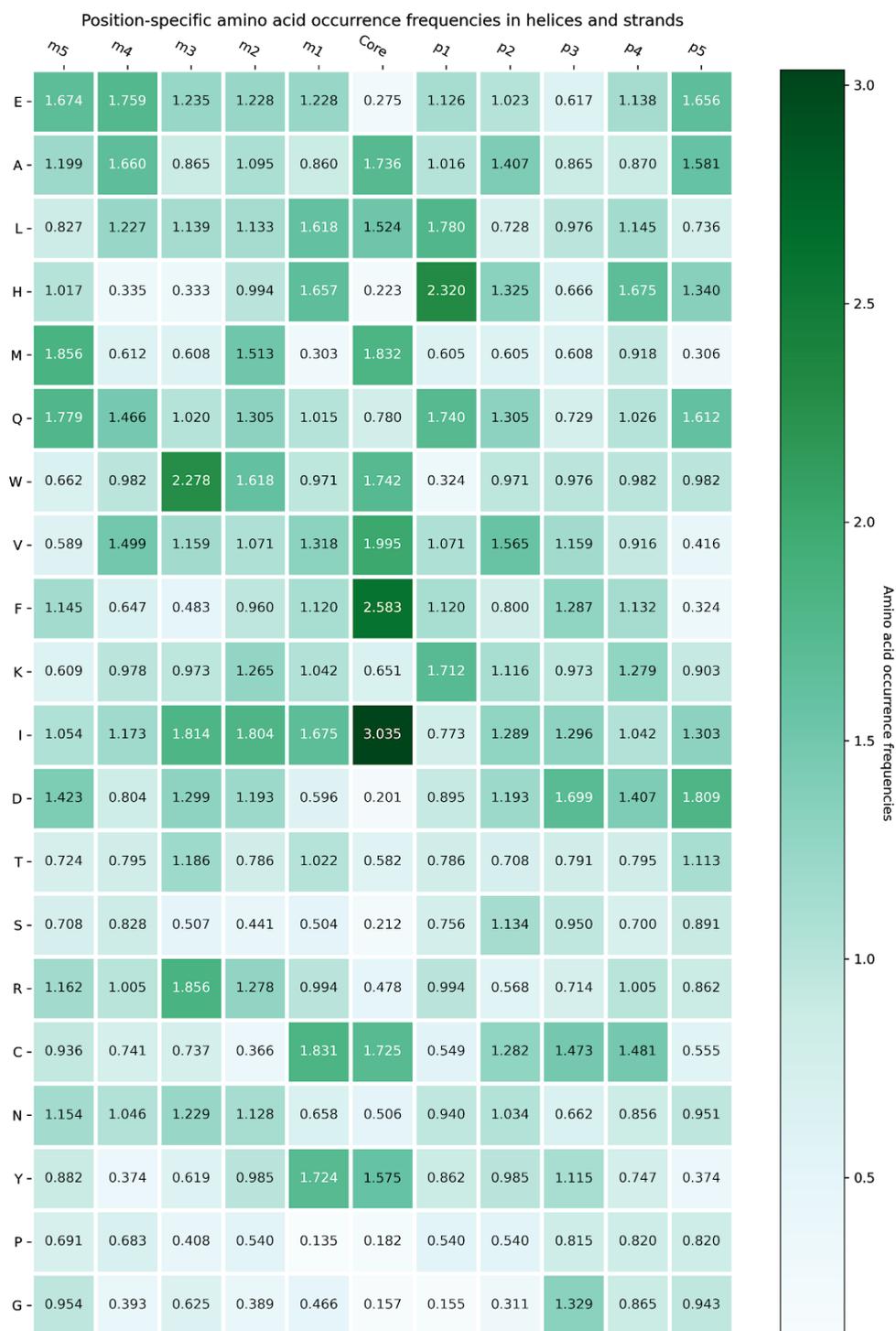
(<https://docs.google.com/spreadsheets/d/1WEG6w71ooNv1r64dmS1WwoibGaGhS6-4O3QCVSVVPGQ/edit?usp=sharing>, лист Cores).

Каждая запись состоит из PDB ID структуры, в которой изучалась зависимость, цепь, в которой располагаются интересующие нас остатки, тип (спираль или тяж) и идентификатор вторичной структуры, сайт нуклеации, аминокислоты, расположенные на расстоянии 5 и меньше остатков от сайта нуклеации, и ссылка на источник, из которого бралась информация об остатках ядра нуклеации. Представленные в таблице остатки нуклеации были получены исследователями в результате применения методов изучения фолдинга, описанных во введении, то есть путём анализа  $\phi$ -значений или водородно-дейтериевого обмена. Расположение аминокислоты по отношению к сайту нуклеации мы обозначали как  $p_i$  или  $m_i$ , где  $p$  и  $m$  – это плюс и минус  $i$  остатков от сайта соответственно ( $p_1, p_2, p_3, p_4, p_5, m_1, m_2, m_3, m_4, m_5$ ). Аминокислоты  $p_1, p_2, p_3, p_4, p_5, m_1, m_2, m_3, m_4, m_5$  считаются “соседями” сайта нуклеации.

Вначале были подсчитаны частоты различных остатков в позициях  $m_5 - p_5$  и ядрах нуклеации. Частота каждого остатка на позиции была нормирована на частоту данного остатка в совокупности рассмотренных последовательностей. По полученным данным были построены тепловые карты (рисунки 5-7). На них заметны остатки, которые чаще, чем обычно, встречаются в ядрах элементов вторичной структуры и их соседях. Среди остатков, формирующих ядра нуклеации и бета-тяжей, и альфа-спиралей, чаще, чем в 2 раза встречаются изолейцин и фенилаланин. Дополнительно можно выделить метионин, лейцин, аланин, цистеин и триптофан в спиральях и валин с тирозином в листах. Все перечисленные остатки можно отнести или к классу гидрофобных, или амфифильных. Что касается окружение ядер нуклеации, каких-либо четких закономерностей не прослеживается: нормированные частоты меняются хаотично. Конечно, заметно, что рядом с ядрами нуклеации альфа-спиралей не присутствуют пролины и глицины, это ожидаемо, так как данные остатки негативно сказываются на геометрии альфа-спирали. Особенно выделяется гистидин в +1 положении, который встречается там в 4 раза чаще, чем вообще в последовательностях, однако, странно, что похожего эффекта не наблюдается в -1 позиции. Гистидина вообще заметно больше к С-концу от ядра нуклеации спирали, чем к N-концу. Объяснение этому не найдено, однако, не исключена возможность случайности этого наблюдения.

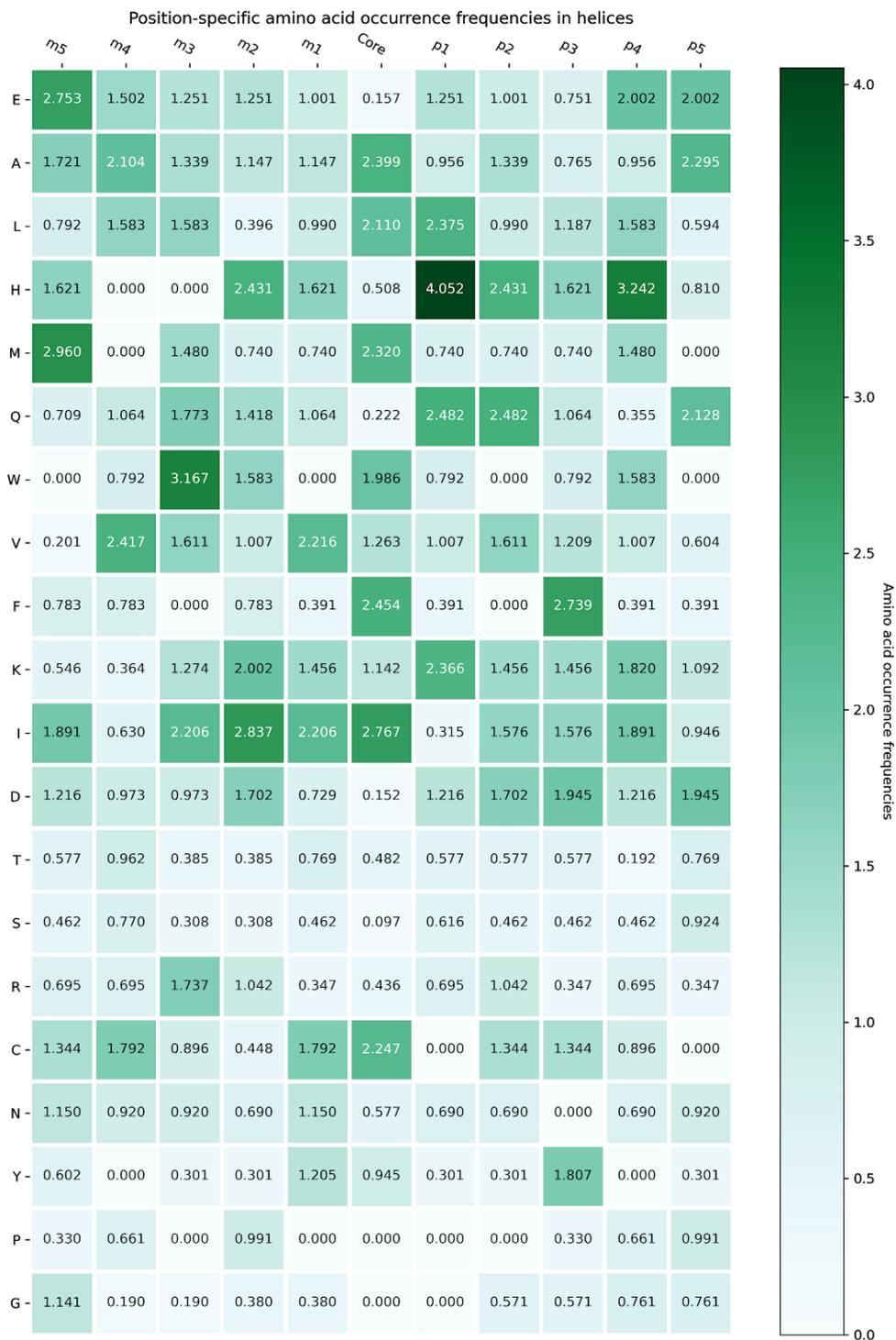
Для наборов последовательностей по позициям  $-5...-1$  и  $+1...+5$ , т. е.  $m_5, m_4, m_3, m_2, m_1$  и  $p_1, p_2, p_3, p_4, p_5$ , были построены лого последовательностей (рисунок 8) методом Кульбака-Лейблера. Такой способ позволяет выделить как самые популярные, так и наименее распространенные аминокислоты в профиле выравнивания [23]. Из представленных лого какой-либо очевидной закономерности выделить нельзя и сигнала фолдинга точно не прослеживается. Возможно, для нуклеации важны не конкретные остатки вокруг ядра, а их класс, поэтому, были построены столбчатые диаграммы (рисунок 9) с распределением частот групп остатков (полярные, заряженные и гидрофобные) ядер нуклеации и соседних позиций. Заметно, что частоты групп остатков, образующих ядра нуклеации, заметно отличаются как от

фона, так и от частот окружения: среди них значительно преобладают гидрофобные (p-value критерия хи-квадрат менее 0,001 и для тяжей, и для спиралей). Для альфа-спиралей изменения доли гидрофобных остатков в зависимости от позиции оказались более заметными, однако, в полученных данных не было выявлено статистически значимой закономерности. Небольшое волнообразное поведение графика для альфа-спиралей можно объяснить чередованием гидрофобных и гидрофильных остатков в её структуре.

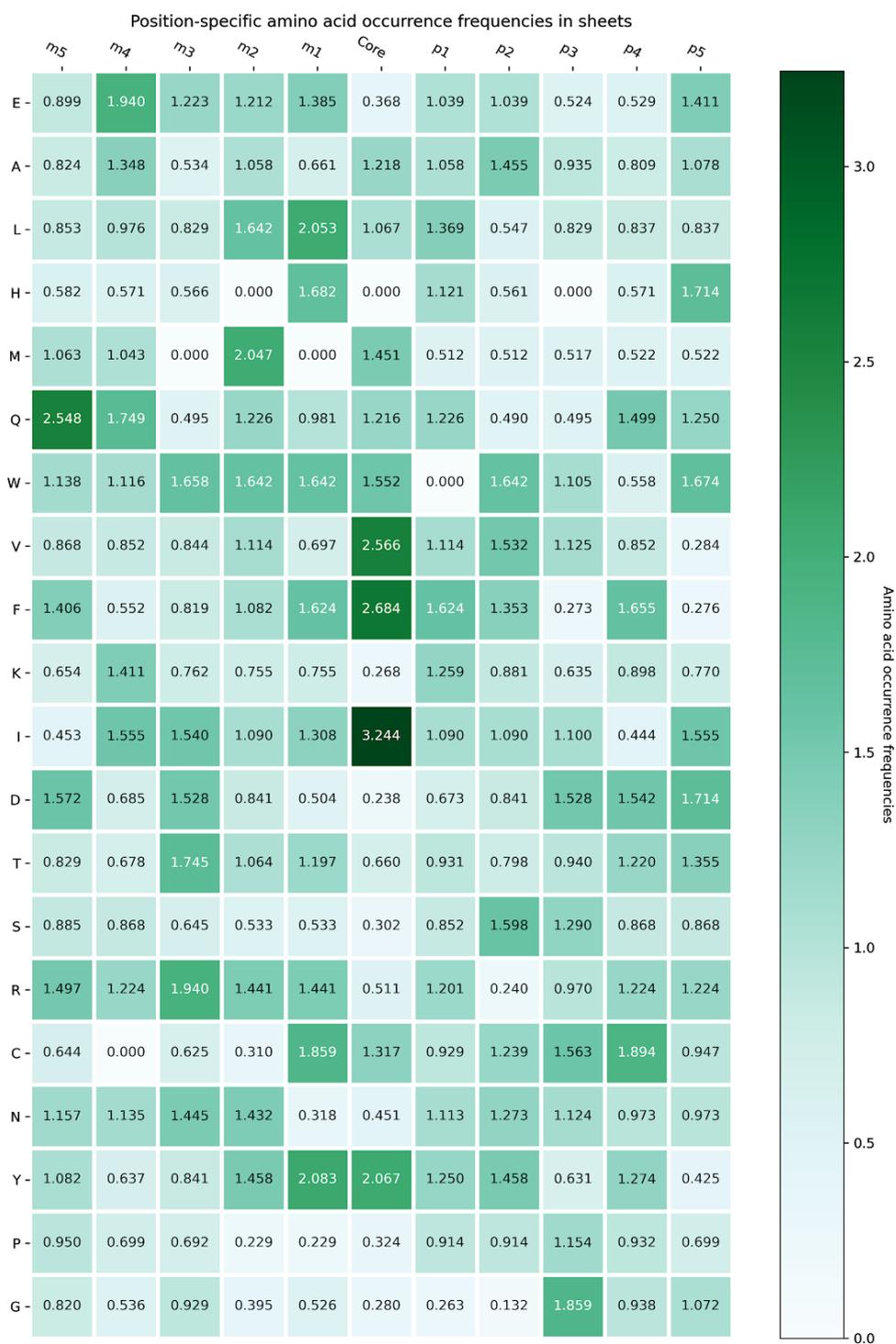


**Рисунок 5.** Частоты встречаемости аминокислот в окрестности 5 остатков от предполагаемого сайта нуклеации спиралей и тяжей. По вертикали расположены

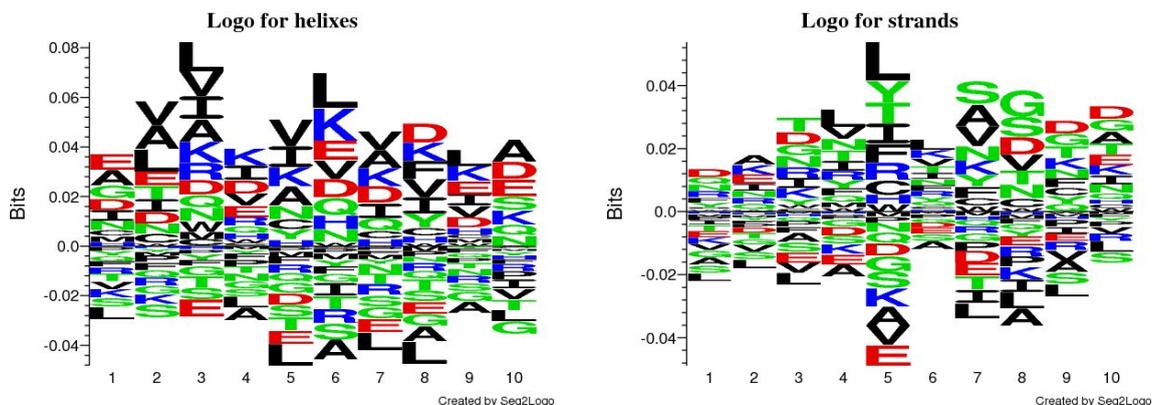
однобуквенные названия аминокислот, по горизонтали – позиция аминокислоты, где  $\rho_i$  – плюс  $i$ ,  $m_i$  – минус  $i$  остатков от сайта нуклеации. Нормировки количества аминокислот в каждой ячейки heat map осуществлялась по числу записей в датасете.



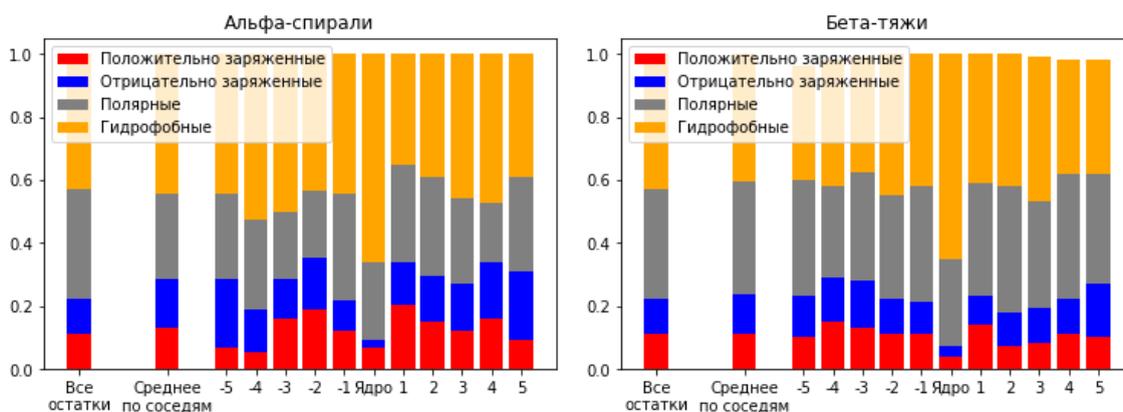
**Рисунок 6.** Частоты встречаемости аминокислот в окрестности 5 остатков от предполагаемого сайта нуклеации спиралей. По вертикали расположены однобуквенные названия аминокислот, по горизонтали – позиция аминокислоты, где  $\rho_i$  – плюс  $i$ ,  $m_i$  – минус  $i$  остатков от сайта нуклеации. Нормировки количества аминокислот в каждой ячейки heat map осуществлялась по числу записей в датасете.



**Рисунок 7.** Частоты встречаемости аминокислот в окрестности 5 остатков от предполагаемого сайта нуклеации тяжей. По вертикали расположены однобуквенные названия аминокислот, по горизонтали – позиция аминокислоты, где  $p_i$  – плюс  $i$ ,  $m_i$  – минус  $i$  остатков от сайта нуклеации. Нормировки количества аминокислот в каждой ячейки heat map осуществлялась по числу записей в датасете.



**Рисунок 8.** Лого для соседей ядер нуклеации альфа-спиралей (слева) и бета-тяжей (справа). Нумерация: 1-5 соответствует позициям -5 – -1, 6-10: +1 – +5. Сами остатки ядер нуклеации в лого не представлены.



**Рисунок 9.** Столчатые диаграммы, показывающие доли типов остатков в зависимости от их позиции относительно ядра нуклеации для альфа-спиралей (слева) и бета-тяжей (справа). “Среднее по соседям” показывает усредненные доли остатков среди участков -5 ... +5. Столбец “все остатки” демонстрирует долю остатков разных типов в рассмотренных последовательностях. “Ядро” показывает доли остатков соответствующих типов среди остатков ядра нуклеации.

Далее мы рассматривали только закономерности, связанные с остатками ядра фолдинга, поскольку для соседствующих остатков их выявить не удалось. Чтобы отделить остатки, которые чаще, чем остальные встречаются во вторичных структурах, от тех, которые для рассматриваемых вторичных структур чаще всего оказываются в ядре нуклеации, данные для частот встречаемости различных остатков в ядрах нуклеации из тепловых карт 5-7 были нормированы на распространённость остатков в зависимости от вторичной структуры из статьи [21] и представлены в таблице 3. В целом, в представленной таблице отношения частот встречаемости все еще соответствует выводам, сделанным о наиболее часто встречающихся остатках в ядре нуклеации из тепловых карт. Однако можно ещё заметить, что чаще, чем обычно в ядре нуклеации спирали появляется цистеин, а в случае с бета-тяжами еще одним более распространённым остатком становится триптофан. Ещё в таблице видно, что глутамат в ядрах нуклеации листа встречается чаще, чем обычно в бета-листах. Однако его нормированная частота из тепловой карты 9 составляет только 0,368 и

является значительно меньшей, чем у соседних остатков, поэтому маловероятно, что данный остаток играет важную роль в образовании ядра фолдинга. Выявленные остатки лишь частично согласуются с делением аминокислот на “слабые” и “сильные” инициаторы и терминаторы вторичной структуры из статьи [21]. Скорее, больше наблюдается факт того, что гидрофобные и/или амфифильные остатки чаще встречаются в ядрах нуклеации, в то время как полярные и гидрофильные - реже. В принципе, этот вывод можно было сделать и из рисунка 9.

**Таблица 3.** Отношение частоты остатка в ядре нуклеации для определенной вторичной структуры к частоте остатков в рассматриваемой вторичной структуре. Данные о частотах остатков во вторичных структурах взяты из таблицы 1.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
<b>Ядро нуклеации бета-тяжа</b>	0,7	1,0	0,3	1,4	2,1	0,3	0,0	2,0	0,4	0,9	0,9	0,7	0,5	1,0	0,6	0,4	0,5	1,6	1,3	1,6
<b>Ядро нуклеации спирали</b>	1,7	2,9	0,2	0,1	2,2	0,0	0,4	2,8	1,1	1,6	1,9	0,8	0,0	0,2	0,6	0,1	0,6	1,1	1,7	1,5

Поскольку стало заметно, что некоторые часто встречающиеся в ядрах остатки совпадают как для тяжей, так и для спиралей, мы расширили список остатков ядер нуклеации с помощью базы данных Start2Fold [24]. Была получена выборка “ранних” остатков, то есть остатков, которые раньше всех оказываются скрытыми от растворителя в процессе фолдинга. Данные об этих остатках были получены с помощью водородно-дейтериевого обмена, а как было сказано во введении, выявленные таким образом “ранние” остатки также обычно обладают значениями  $\phi$ , близкими к 1. К сожалению, для ядер нуклеации в базе данных не указаны типы вторичной структуры, к которым они относятся. Были подсчитаны доли “ранних” частоты остатков и нормированы на доли остатков в рассмотренных последовательностях (таблица 4). Также в ней представлены данные для ядер из тепловой карты 5 для сравнения. Заметно, что хотя бы в плане бинарности “больше-меньше” наблюдается сходимость между двумя выборками.

**Таблица 4.** Отношение частоты остатка в ядре нуклеации к частоте этого остатка в рассмотренных последовательностях.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
<b>Ядра нуклеации Start2Fold</b>	1,1	1,3	0,6	0,8	1,4	0,5	1,0	1,5	1,1	1,3	1,4	0,5	0,6	0,9	0,7	0,7	0,8	1,6	1,6	1,4
<b>Ядра нуклеации из таблицы</b>	1,7	1,7	0,2	0,3	2,6	0,2	0,2	3,0	0,7	1,5	1,8	0,5	0,2	0,8	0,5	0,2	0,6	2,0	1,7	1,6

Тем не менее, видно, что разброс величин для ядер нуклеации из таблицы с изначальной выборкой больше, чем для данных из базы данных. Например, гистидин в выборке из таблицы встречался гораздо реже, чем в базе данных. Это можно объяснить тем, что в базе данных выборка больше: 57 записей против 20, поэтому значения в таблице из 20 белков более экстремальные, так и тем, что в выборках может быть разное соотношение бета-тяжей и альфа-спиралей.

Таким образом, по результатам анализа частот остатков, чаще всего входящих в ядра нуклеации, можно утверждать, что их в основном образуют гидрофобные и амфифильные аминокислоты как в случае с бета-листами, так и альфа-спиральями. Полученным данным больше всего соответствует рассмотренная во введении теория гидрофобного коллапса. Однако нельзя отбрасывать и другие возможные теории, поскольку полярные и заряженные остатки также встречаются в ядрах нуклеации, а значит, должна быть причина, по которой они там появляются, которая описывается другими механизмами фолдинга. Более того, большинство ядер нуклеации из рассмотренных выборок было получено с помощью водородно-дейтериевого обмена на основании предположения о том, что остатки, которые раньше всех оказываются защищёнными от растворителя, и есть ядро фолдинга, что подтверждается высокими значениями  $\phi$  для них. Однако этим может и объясняться смещение выборки в сторону гидрофобных остатков: действительно, им “выгодно” раньше всех защититься от растворителя. Более непредвзятые результаты даёт непосредственно анализ значений  $\phi$ , но этот метод гораздо более трудоемкий и требует наличие для белка способа определения кинетики фолдинга, поэтому таких данных значительно меньше, чем данных водородного обмена.

## **Вывод**

Задача поиска закономерностей фолдинга выходит за рамки простой статистической оценки частот встречаемости аминокислотных остатков в ядре нуклеации и среди его соседей по последовательности. К тому же данных непосредственно о ядрах нуклеации не очень много ввиду сложности их достоверного определения. Вероятно, рассмотренные выборки содержат скрытые переменные, которые не позволяют в общем случае выделить особенные остатки среди ближайших к ядру нуклеации. Что касается самих ядер нуклеации, были найдены остатки, которые реже или чаще встречаются в них. Более распространённые остатки гидрофобные и амфифильные, что согласуется с теорией гидрофобного коллапса. Причём по полученным данным наиболее склонны к нахождению в ядре фолдинга остатки с крупными или разветвленными боковыми группами, такие как фенилаланин и изолейцин.

## Список литературы

1. Dobson C. M. Protein folding and misfolding // *Nature*. – 2003. – Т. 426. – №. 6968. – С. 884-890.
2. Fersht A. R. Nucleation mechanisms in protein folding // *Current opinion in structural biology*. – 1997. – Т. 7. – №. 1. – С. 3-9.
3. Li R., Woodward C. The hydrogen exchange core and protein folding // *Protein Science*. – 1999. – Т. 8. – №. 8. – С. 1571-1590.
4. Nickson, A. A., & Clarke, J. (2010). What lessons can be learned from studying the folding of homologous proteins? *Methods*, 52(1), 38–50. doi:10.1016/j.ymeth.2010.06.003
5. Voet, D., Voet, J. G., & Pratt, C. W. (2008). *Fundamentals of biochemistry: Life at the molecular level*. Hoboken, NJ: Wiley
6. Zhou, R. (2004). Hydrophobic Collapse in Multidomain Protein Folding. *Science*, 305(5690), 1605–1609. doi:10.1126/science.1101176
7. Lapidus, L. J., Yao, S., McGarrity, K. S., Hertzog, D. E., Tubman, E., & Bakajin, O. (2007). Protein Hydrophobic Collapse and Early Folding Steps Observed in a Microfluidic Mixer. *Biophysical Journal*, 93(1), 218–224. doi:10.1529/biophysj.106.103077
8. Sharadadevi, A., Sivakamasundari, C., & Nagaraj, R. (2005). Amphipathic  $\alpha$ -helices in proteins: Results from analysis of protein structures. *Proteins: Structure, Function, and Bioinformatics*, 59(4), 791–801. doi:10.1002/prot.20459
9. Nölting, B., & Agard, D. A. (2008). How general is the nucleation-condensation mechanism? *Proteins: Structure, Function, and Bioinformatics*, 73(3), 754–764. doi:10.1002/prot.22099
10. Kukic, P., Pustovalova, Y., Camilloni, C., Gianni, S., Korzhnev, D. M., & Vendruscolo, M. (2017). Structural Characterization of the Early Events in the Nucleation–Condensation Mechanism in a Protein Folding Process. *Journal of the American Chemical Society*, 139(20), 6899–6910. doi:10.1021/jacs.7b01540
11. Karplus, M., & Weaver, D. L. (1976). Protein-folding dynamics. *Nature*, 260(5550), 404–406. doi:10.1038/260404a0
12. Daggett, V., & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences*, 28(1), 18–25. doi:10.1016/s0968-0004(02)00012-9
13. Karplus, M., & Weaver, D. L. (2008). Protein folding dynamics: The diffusion-collision model and experimental data. *Protein Science*, 3(4), 650–668. doi:10.1002/pro.5560030413
14. Nölting, B., & Andert, K. (2000). Mechanism of protein folding. *Proteins: Structure, Function, and Genetics*, 41(3), 288–298. doi:10.1002/1097-0134(20001115)41:3<288::aid-prot20>3.0.co;2-c
15. Vorov, O. K., Livesay, D. R., & Jacobs, D. J. (2009). Helix/Coil Nucleation: A Local Response to Global Demands. *Biophysical Journal*, 97(11), 3000–3009. doi:10.1016/j.bpj.2009.09.013
16. Chou, P. Y., & Fasman, G. D. (2006). Prediction of the Secondary Structure of Proteins from their Amino Acid Sequence. *Advances in Enzymology and Related Areas of Molecular Biology*, 45–148. doi:10.1002/9780470122921.ch2

17. Toumadje, A., & Johnson, W. C. (1994). A CD study of the  $\alpha$ -helix nucleation hypothesis. *Biopolymers*, 34(7), 969–973. doi:10.1002/bip.360340715
18. De Jongh, H. H. J. (2000). The helix nucleation site and propensity of the synthetic mitochondrial presequence of ornithine carbamoyltransferase. *European Journal of Biochemistry*, 267(18), 5796–5804. doi:10.1046/j.1432-1327.2000.01654.x
19. Daggett, V., & Fersht, A. R. (2003). Is there a unifying mechanism for protein folding? *Trends in Biochemical Sciences*, 28(1), 18–25. doi:10.1016/s0968-0004(02)00012-9
20. Chou, P. Y., & Fasman, G. D. (1974). Conformational parameters for amino acids in helical,  $\beta$ -sheet, and random coil regions calculated from proteins. *Biochemistry*, 13(2), 211–222. doi:10.1021/bi00699a001
21. Prediction of Protein Structure and the Principles of Protein Conformation. GERALD D. FAsMANI, Ed. Plenum, New York, 1989
22. Richardson, J., & Richardson, D. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240(4859), 1648–1652. doi:10.1126/science.3381086
23. <http://www.cbs.dtu.dk/biotools/Seq2Logo-2.1/>  
Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Martin Christen Frolund Thomsen; Morten Nielsen, *Nucleic Acids Research* 2012; 40 (W1): W281-W287.
24. <https://www.bio2byte.be/start2fold/>  
R. Pancsa and M. Varadi, P. Tompa, W. Vranken: Start2Fold: a database of hydrogen/deuterium exchange data on protein folding and stability *Nucl.Acids Res.* (04 January 2016) 44 (D1): D429-D434