

Практикум 8.

Задание 1.

В рамках первого задания я работала со структурой эндонуклеазы FokI из *Planomicrobium okeanoikoites* (PDB: 1FOK). Этот фермент связывает определенную последовательность ДНК и вносит неспецифический разрыв на некотором расстоянии от неё. ДНК – субстрат сложного строения, поэтому для такого рода фермента можно ожидать наличия нескольких доменов, ответственных за осуществление связывания ДНК и внесения разрыва.

Сначала попробуем предсказать наличие доменов методом SplitValue (индекс разделенности). Эта функция применяется для предсказания остатков, находящихся на границах между доменами, в рамках алгоритма DOMAK, опубликованного в 1995 году. Для расчета SplitValue для каждого остатка структуру условно делят на две части: от N-конца до рассматриваемого остатка и от следующего остатка до С-конца – и подсчитывают количество контактирующих пар остатков, которые образуются внутри каждой из двух подструктур и между ними. Таким образом, алгоритм основывается на определении структурного домена, согласно которому остатки в составе одного домена образуют много взаимодействий между собой и практически не имеют контактов с остатками других доменов. Формула SplitValue имеет следующий вид:

$$SplitValue = \frac{intA * intB}{extAB^2}$$

Где intA, intB – контактирующих пар остатков внутри подструктур A (от N-конца до рассматриваемого остатка) и B (от рассматриваемого остатка до С-конца), extAB – число таких контактов, где один остаток принадлежит A, а другой – B. Эта метрика максимизируется, когда внутри предполагаемых доменов A и B контактирующих пар остатков образуется много, и в то же время между доменами – достаточно мало. Я реализовала эту функцию для рассматриваемой структуры и получила значения индекса разделенности для каждого остатка, которые отражены на графике на рисунке 1.

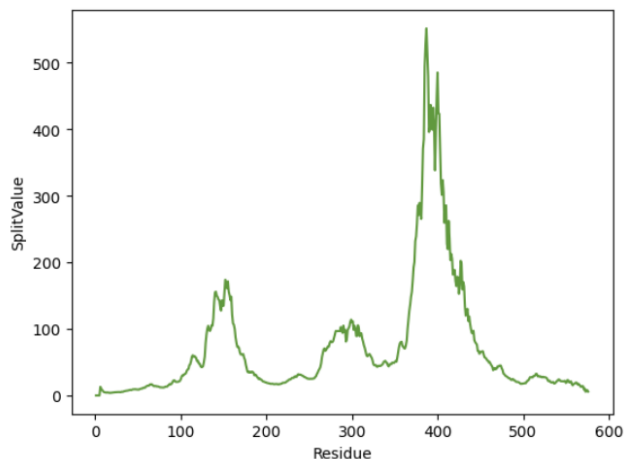


Рисунок 1. Значения функции SplitValue в зависимости от номера остатка. Три пика соответствуют остаткам 151, 298 и 386

На графике можно выделить три пика: 151, 298 и 386 остатки, причем третий более убедителен, но я решила принять во внимание все три. Пики меньшей высоты в непосредственной близости от 151, 298 и 386 я не рассматриваю отдельно, так как «расширенные» пики, вероятно, соответствуют протяженным линкерам между доменами, и в таком случае все позиции этого пика находится на одной и той же границе. Условно я провожу эту границу по остатку с максимальным значением SplitValue. Результат представлен на рисунке 2, биологическое значение найденных доменов будет описано далее.

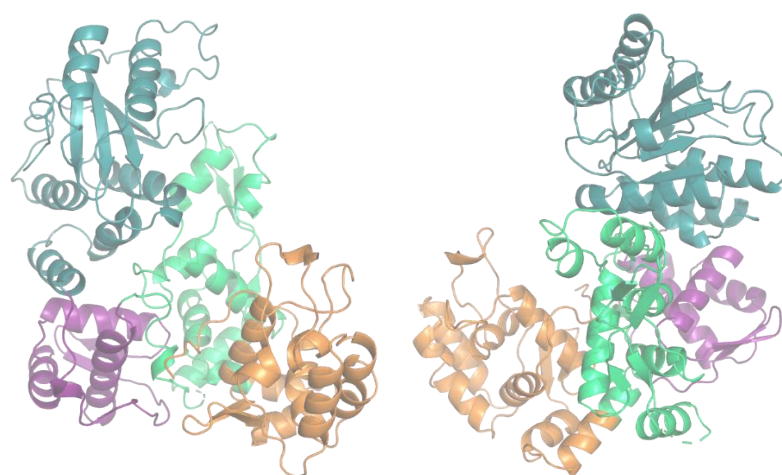


Рисунок 2. Доменная организация эндонуклеазы FokI, согласно SplitValue. Желтым показаны остатки 1-150, зеленым – 151-297, фиолетовым – 298-385, бирюзовым – 386-576. [Ссылка на сессию](#)

Как по графику SplitValue, так и по приведенному выше рисунку можно предположить, что точность выделения С-концевого домена выше, чем N-концевых: он отделен длинным неупорядоченным линкером и достаточно обособлен пространственно, чтобы можно было говорить о минимизации контактов с остальными доменами. Также ему соответствует более выдающийся пик SplitValue. N-концевым доменам соответствуют более сглаженные пики. Для первого пика его максимум лежит на альфа-спирали, соединяющей данный домен с центральным. Это уже предполагает меньшую структурную обособленность. Как видно на рисунке, в структуре оранжевого N-концевого домена есть короткий линкер, и, вероятно, граница по нему была бы более ожидаема, но с двух сторон от него находятся две альфа-спирали, расстояние между которыми допускает наличие контактов между боковыми радикалами составляющих их аминокислотных остатков. Вероятно, это является причиной именно такого разделения, предложенного алгоритмом. Граница между зеленым и фиолетовым доменами сомнений не вызывает, так как проходит через неупорядоченный участок и разделяет не контактирующие между собой элементы укладки.

Задание 2.

Разные базы данных предлагают различные варианты выделения структурных и функциональных доменов, однако каждая классификация включает ДНК-связывающие N-концевые и рестрикционные С-концевые. Варианты выделения этих доменов представлены в таблице 1.

Таблица 1. Границы доменов FokI, согласно разным базам данных

База данных	ДНК-связывающий домен	Рестрикционный домен
CATH	1-289, 299-385	398-575
SCOP	12-129, 156-265, 298-366	416-519
SCOPe	1-140, 141-278, 284-383	384-576
Pfam	3-138	384-573
*InterPro	3-136, 145-279, 293-359	384-571

Как видно из таблицы, предположение о том, что N-концевой домен на самом деле должен быть короче, подтверждается данными SCOP, SCOPe и Pfam. При этом в CATH, SCOP и SCOPe предложено выделять несколько ДНК-связывающих доменов, тогда как в Pfam – только один (выделен еще один домен 293-360 без определенной функции). На рисунках 3-6 визуализировано разделение на домены по версии этих четырех источников. Я акцентировала внимание на доменах, для которых определена функция в структуре. Также для каждого разбиения на домены представлена структура со связываемой молекулой ДНК.

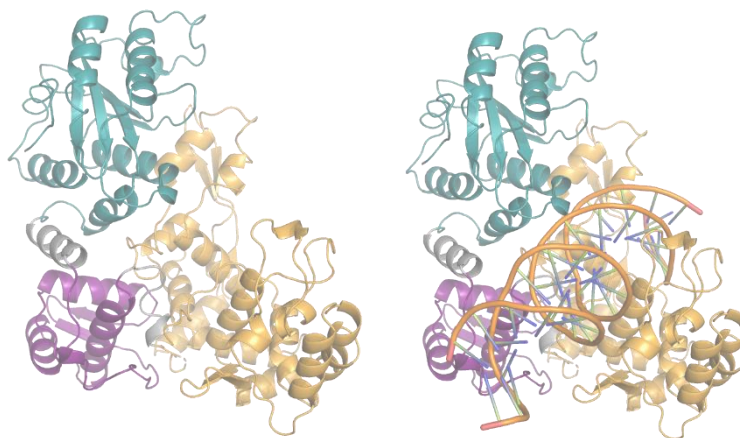


Рисунок 3. Разделение на домены, согласно CATH. Оранжевым и фиолетовым показаны ДНК-связывающие домены, бирюзовым – рестрикционный, серым – участки, не входящие в домены

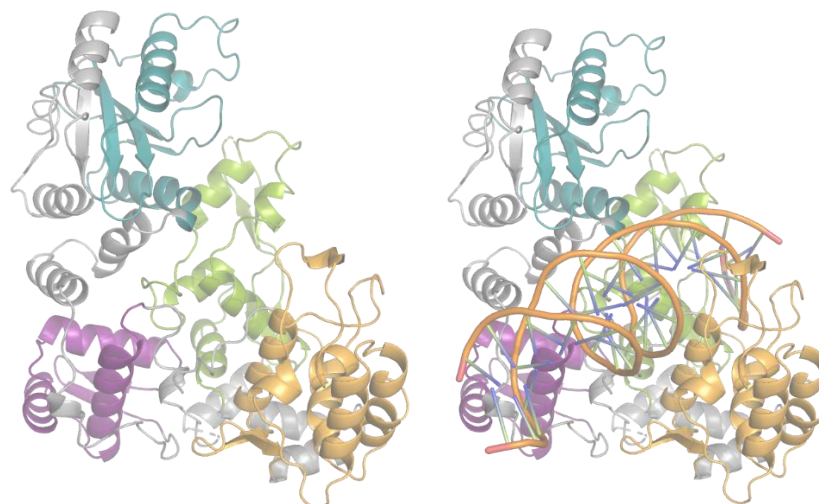


Рисунок 4. Разделение на домены, согласно SCOP. Оранжевым, зеленым и фиолетовым показаны ДНК-связывающие домены, бирюзовым – рестрикционный, серым – участки, не входящие в домены

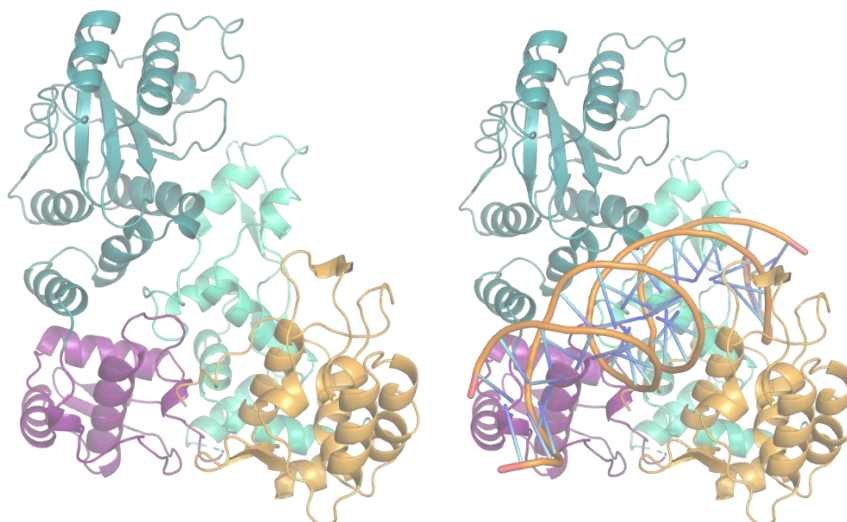


Рисунок 5. Разделение на домены, согласно SCOPe. Оранжевым, зеленым и фиолетовым показаны ДНК-связывающие домены, бирюзовым – рестрикционный, серым – участки, не входящие в домены

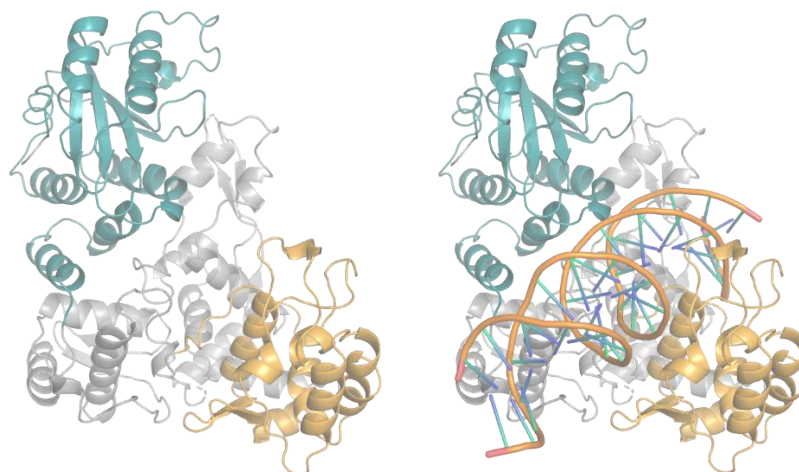


Рисунок 6. Разделение на домены, согласно Pfam. Оранжевым показан ДНК-связывающий домен, бирюзовым – рестрикционный, серым – участки, не входящие в домены

Обобщая наблюдения, можно отметить следующее:

- Во всех базах данных, кроме SCOP, С-концевой рестрикционный домен выделен практически одинаково, и его граница приблизительно совпадает с предсказанной по SplitValue
- Центральный домен присутствует не во всех базах данных, но его граница очень четко предсказывается SplitValue, что неожиданно, учитывая низкий и широкий пик на графике
- Выделение С-концевого домена неоднозначно, SplitValue предсказывает его неточно
- Данные SCOPe по значениям ближе всего к полученным из SplitValue

ДНК-связывающий домен FokI распознает фрагмент 5'-GGATG-3', рестрикционный – вносит разрывы на расстоянии 9 и 13 пар оснований от данного фрагмента. Согласно статье [1], распознающий ДНК домен состоит из трех субдоменов D1, D2, D3 (рисунок 7). 3'-конец специфичной последовательности распознается доменом D1 (главным образом спираль α_6), 5' – субдоменом D2 (спираль α_5). Оба домена образуют характерные для рестриктаз НТН-мотивы. Он есть и в домене D3, но его основная роль – связь распознающего ДНК домена с рестрикционным.

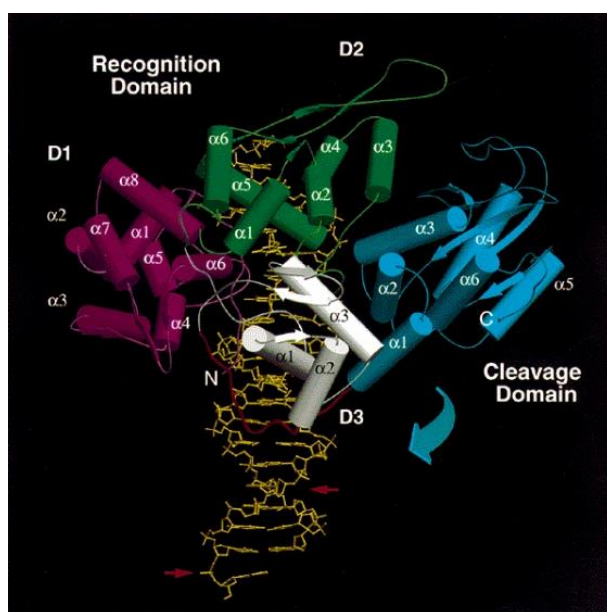
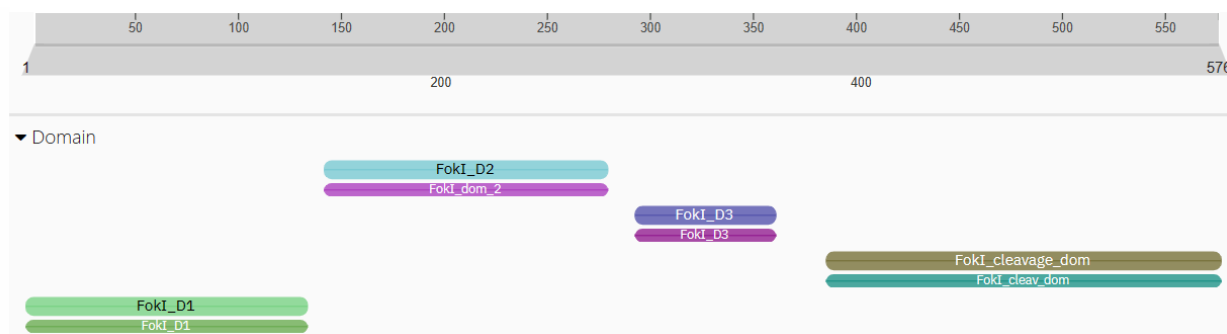


Рисунок 7. Доменная организация FokI, согласно публикации [1]. ДНК-связывающие субдомены показаны малиновым, белым и зеленым, рестрикционный – голубым

Описанному разбиению на домены наилучшим образом соответствуют результат SplitValue и данные SCOP и SCOPe, участвующие в связывании ДНК спирали отнесены к правильным доменам. CATH объединяет D1 и D2 в один домен (обе ключевые для связывания ДНК спирали в одном домене), Pfam выделяет только один ДНК-связывающий домен (D1). Рестрикционный домен выделен приблизительно верно везде, так как несмотря на разные варианты границ, каталитические остатки Asp450, Asp467 и Lys469 входят в его состав. Кстати, положение каталитических остатков не в кармане связывания ДНК, а в удаленном от места связывания домене, довольно необычно, как полагают авторы статьи, в которой данная структура была описана. Между двумя функциональными доменами существует сложная сеть преимущественно электростатических внутрибелковых взаимодействий, опосредованных линкерным фрагментом и субдоменами D2 и D3. При активации фермента эти взаимодействия замещаются взаимодействиями с ДНК.

Задание 3.

Далее я запустила поиск InterPro по последовательности белка, и результат получился вполне красивым и вполне ожидаемым: найдены и определены все 4 домена:



Координаты найденных доменов указаны выше в таблице 1, они приблизительно соответствуют содержащимся в других базах данных и найденным с помощью SplitValue. Функции доменов в составе белка, очевидно, соответствуют описанным InterPro. На рисунке 8 полученное разделение на домены показано на структуре.

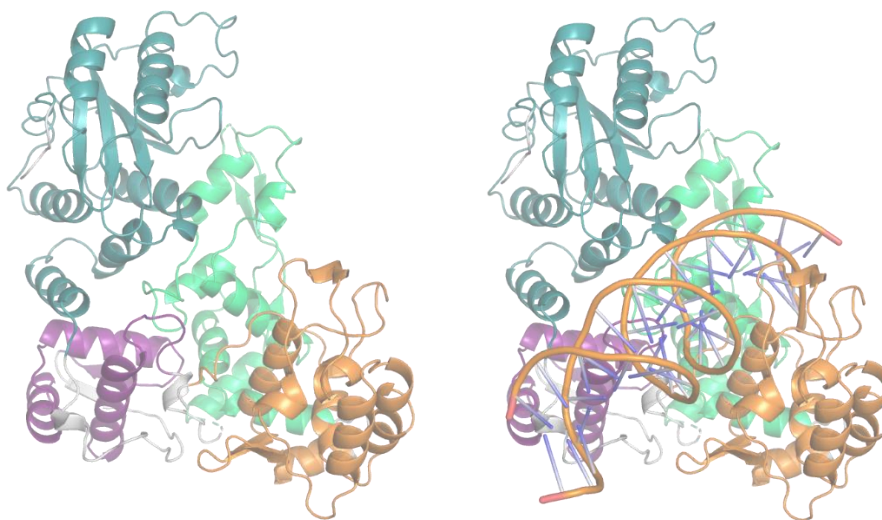
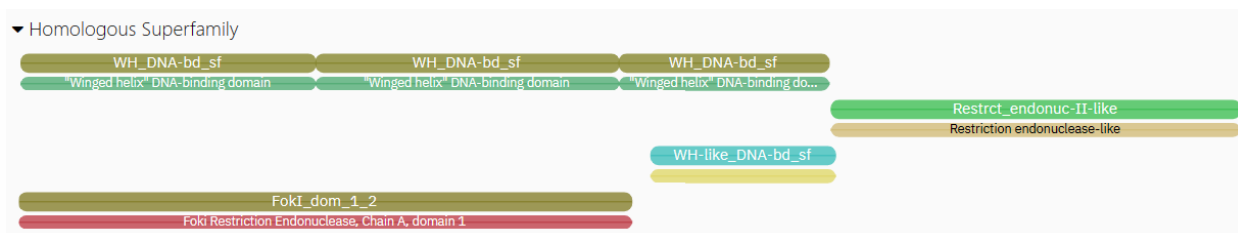


Рисунок 8. Разделение на домены, согласно InterPro. Оранжевым (D1), зеленым (D2) и фиолетовым (D3) показаны ДНК-связывающие домены, бирюзовым – рестрикционный, серым – участки, не входящие в домены

Теперь можно с уверенностью сказать, что на рисунке 2 оранжевый домен соответствует D1, зеленый – D2, фиолетовый – D3, бирюзовый – рестрикционному.



Также InterPro позволяет получить данные о семействах, к которым относятся найденные домены. Так, ДНК-связывающие домены относятся к суперсемейству «winged helix» – содержащих «крылатый» мотив спираль-поворот-спираль (HTH), где роль «крыльев» играют небольшие бета-листы или петли. Мотив содержит два «крыла» (W), три альфа-спирали (H) и три бета-листа (S), расположенные в порядке H1-S1-H2-H3-S2-W1-S3-W2. Одна из альфа-спиралей такого домена распознает специфическую последовательность ДНК [2].

Рестрикционный домен является родственным эндонуклеазам рестрикции II типа. Для этих доменов характерно наличие топологической структуры $\alpha/\beta/\alpha$. Эндонуклеазы рестрикции II типа являются сайт-специфичными и вносят специфичный двуцепочечный разрыв, оставляя фосфат на 5'-конце. Сайт рестрикции находится рядом с сайтом узнавания. Чаще это гомодимерные ферменты, сайт узнавания которых представляет собой палиндром [3].

Задание 4.

Я воспользовалась программой 2Struc для аннотации вторичной структуры моего белка двумя алгоритмами: DSSP и STRIDE. Как видно уже по начальному фрагменту, представленному ниже, в аннотациях довольно много мелких различий, несмотря на общую схожесть.

RESNUM	34	44	54	64	74	84	94												
SEQ	ISKVHNEVKNIKIPTLVKESKIQKELVAIMNQHDLIYTYKELVGTGTSIRSEAPCDAIIQATIADQGNK																		
CONSENSUS	1	X	HHHHHXXXXX	HHHH	000	HHHHHHHHHH	XXXXXXO	EEHHHHH	000XXXXXXO00X	HHHHH	XOXXXXX								
DSSP	1	S	HHHHH	IIIII	HHHH	000	HHHHHHHHHH	TTSSS	00	EEHHHHH	0000	SSGGG	0000	S	HHHHH	S	00	SS	0S
STRIDE	1	T	HHHHHHHHHHHHHHHH	CCC	HHHHHHHHHHH	C	TTTT	C	EEHHHHH	CCC	TTTTTTT	CCCC	HHHHH	CC	TTTTT				

Я рассмотрю подробно участок с 32 по 45 остаток, который представляет собой альфа-спираль по версии STRIDE и альфа-спираль с вставкой пи-спирали по версии DSSP. Определить, какова на самом деле вторичная структура участка, можно путем определения расстояния между аминокислотными остатками, атомы остова которых образуют водородные связи.

На рисунке 9 показана интересующая нас спираль с отмеченными водородными связями. Как можно заметить сразу, PyMol не определяет элемент как обычную альфа-спираль, и в ней действительно наблюдается изгиб, характерный для π -спиралей. Рассмотрев водородные связи, видим, что практически по всей длине сохраняется типичная для α -спирали структура 4_{13} (водородные связи между 32-36, 33-37, 34-38, 35-39, 36-40), однако на 39-41 остатках происходит нарушение структуры – вставка «лишнего» остатка в спираль, приводящая к увеличению диаметра витка и изгибу спирали. В результате, помимо водородной связи 36-40, образуется связь 36-41, что соответствует уже структуре π -спирали 5_{16} . Далее паттерн $i-i+5$ повторяется: водородная связь между 37 и 42 остатком. Затем структура спирали снова нарушается, так как на позиции 43 находится пролин, не способный образовывать водородные связи через азот, в результате чего атомы кислорода 38 и 39 остатков не образуют водородных связей. Затем восстанавливается

обычная структура α -спирали $i-i+4$: водородные связи 40-44, 41-45. Таким образом, рассматриваемый элемент вторичной структуры не является обычной α -спиралью, следовательно, аннотация его методом STRIDE неверна. DSSP предсказал π -спираль с 37 по 41 остаток, что не совсем верно, так как 40 и 41 остатки уже образуют водородные связи, характерные для обычной α -спирали, но сам тип структуры предсказан успешно.

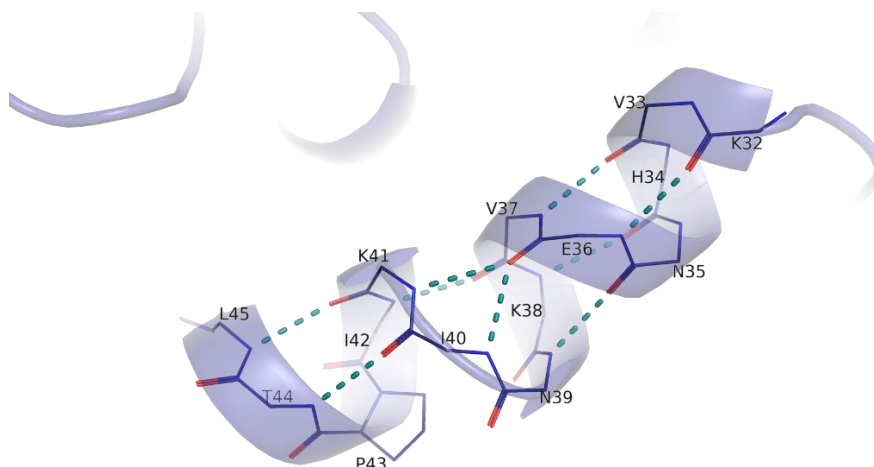


Рисунок 5. Спираль 32-45 с неопределенной структурой. [Ссылка на сессию](#)

Оба метода для определения вторичной структуры рассматривают водородные связи остова для каждого остатка, детектируют предпочтительный тип вторичной структуры для имеющегося набора углов и наращивают элементы структуры из отдельных остатков, для которых она определена. Отличие STRIDE заключается в том, что этот алгоритм для верификации результата использует данные экспертных аннотаций. Вероятно, это мешает ему находить нетипичные структуры, которых довольно мало в базе данных. Такой эффект наблюдался по всей длине выравнивания: STRIDE чаще делает выбор в пользу более типичных структур - α -спиралей и β -листов, вероятно, за счет того, что алгоритм обычно делает выбор в пользу часто встречающихся в базах данных элементов. DSSP свободен от таких предубеждений, поэтому набор его предсказаний более разнообразен и, вероятно, более точен, так как лучше описывает нюансы организации конкретной структуры.

Список литературы

1. Wah, D., Hirsch, J., Dorner, L. et al. Structure of the multimodular endonuclease FokI bound to DNA. *Nature* 388, 97–100 (1997)
2. Winged helix proteins. Gajiwala KS, Burley SK. *Curr. Opin. Struct. Biol.* 10, 110-6, (2000)
3. Type II restriction endonucleases: structure and mechanism. Pingoud A, Fuxreiter M, Pingoud V, Wende W. *Cell. Mol. Life Sci.* 62, 685-707, (2005).