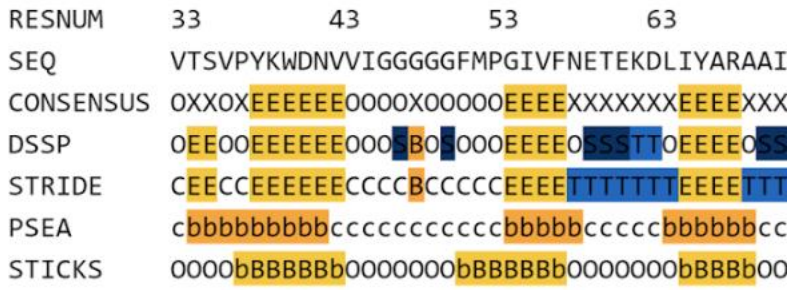
Задание 1. Работа с разметкой вторичной структуры в ручном режиме

Разметка вторичной структуры — это важная задача в структурной биологии: элементы вторичной структуры часто ведут себя как монолитные элементы и являются затравками для фолдинга. Также элементы вторичных структур могут определять функциональность белка. Существует множество алгоритмов разметки вторичной структуры, различающихся преимущественно восприятием того, какими свойствами элемент вторичной структуры определяется: паттерном водородных связей, расстояниями между остатками, углами поворота остатков, торсионными углами. Чаще всего пользуются, тем не менее, алгоритмом DSSP.
В связи с этим интересно сравнить качество разметки вторичной структуры разными алгоритмами на конкретном примере. Я взяла структуру 2cn3 (из практикума по валидации) рассматривала только цепь B. Структура эта получена для комплекса мутанта фермента ксилоглюканазы Xgh74A-D70A. Нормальная ксилоглюканаза из целлюлосомы бактерии Clostridium thermocellum катализирует гидролиз ксилоглюкана, являющегося важным компонентом клеточной стенки растений. Мутант по каталитическому остатку Asp70 неактивен.
На сервисе 2struc я получила разметку от 4 алгоритмов: DSSP, STRIDE, PSEA и STICKS. Далее идет сравнение на 3 участках структуры.

<u>Остатки 58-64.</u>



DSSP аннотирует меткой “T” (петлевой участок с остовной водородной связью) только остатки, между которыми эта связь есть. DSSP аннотирует остовные водородные связи в петлевых областях очень аккуратно (не излишне - сравнить со STRIDE - он тоже видит водородную связь, но распространяет ее влияние на достаточно длинный участок) благодаря возможности разметить что-то как “S” - изгиб (наименьший приоритет). У STRIDE такой альтернативы, видимо, нет.
PSEA и STICKS из-за особенностей работы (не учитывают водородные связи, а только углы поворота и расстояния между остатками (в случае PSEA) или торсионные углы (в случае STICKS)) не умеют находить именно бета-листы, а умеют только бета-полоски. PSEA, кстати, по-видимому, склонен видеть бета-листы более длинными как раз из-за того, что считает не длины водородных связей, а расстояния между остатками. Наверное, PSEA тут более правильно разметил.

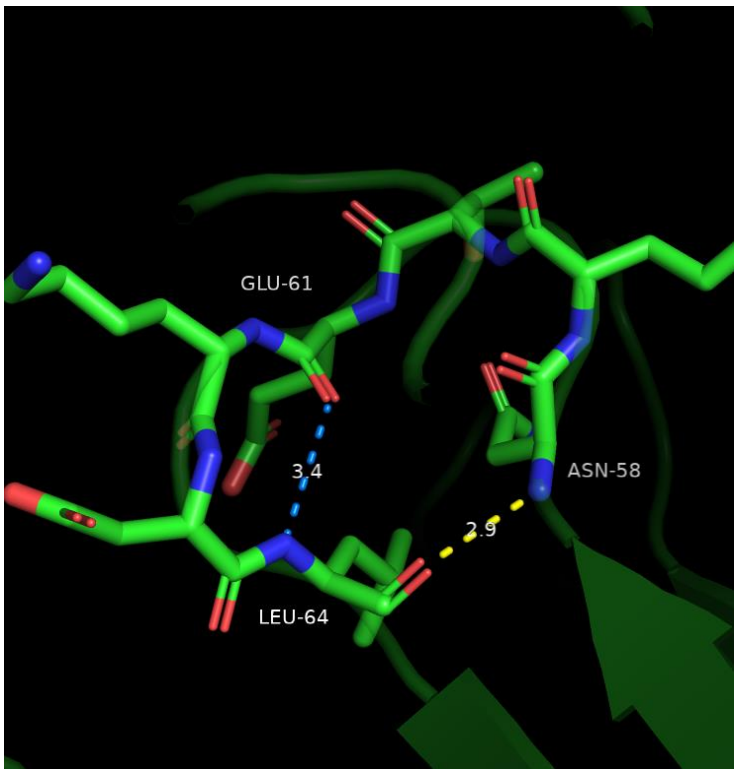


Рис.1. Остатки 58-64

207-218

```
RESNUM      203        213        203        213
SEQ         PNPGTYIYDPNFDYTKDIIG PNPGTYIYDPNFDYTKDIIG
CONSENSUS   OOOOOOOOOOOOOOOOOOOO OOOOOOXXOXXXXXXOOXOO
DSSP        OOOOOOOOOOOHHHOOOOO OOOOOOBSOTTSGGGOSBOO
STRIDE      OOOOOOOOOOOOOOOOOOOO CCCCCCBTTTTTTTTCBCC
PSEA        OOOOOOOOOOOOOOOOOOOO CCCCCCCCCCCCCCCCCCCC
STICKS      OOOEEEEEOOOOOOOOOEEE OOObBBBbOOOOOOOOObBB
```

DSSP находит 3-спиральный участок, стабилизированный 2 водородными связями, но значения углов N-O-C слишком критические (105 и 116 градусов), чтобы считать, что эти водородные связи есть.
STICKS находит единичную бета-полоску, которая таковой не является. Это может объясняться особенностями работы алгоритма: в нем вторичная структура определяется не наличием водородных связей, а локальным распределением двугранных углов и длин связей.
DSSP и STRIDE заявляют об изолированном бета-мостике между остатками 207 и 218, но его там нет. PSEA же проаннотировал так, как есть.

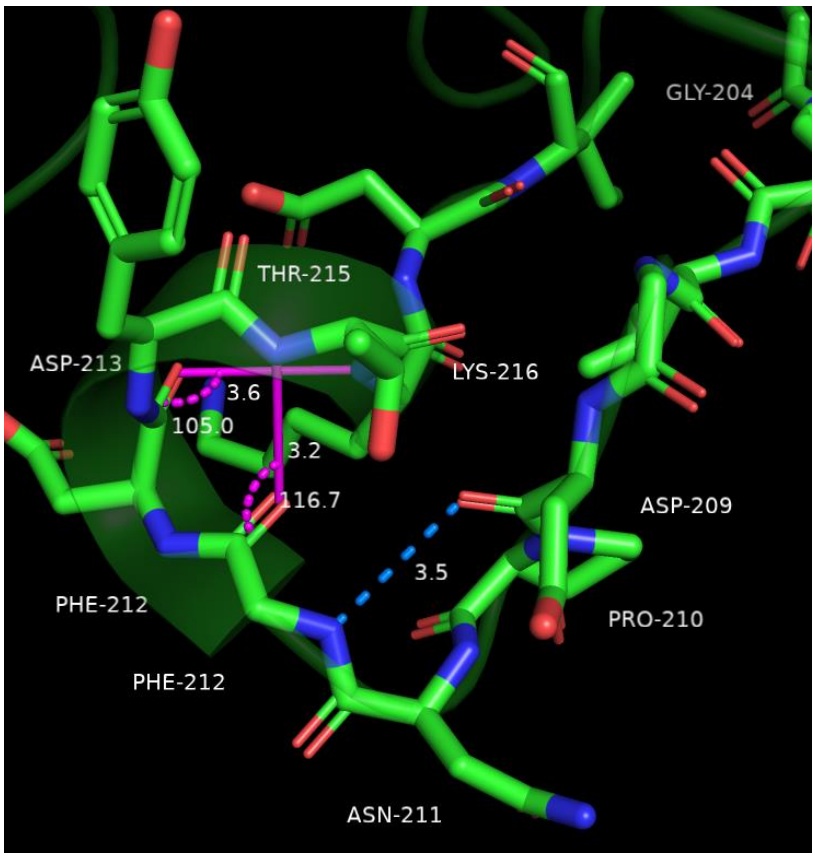


Рис.2. Остатки 207-218

<u>387-401</u>

В данном случае DSSP сработал хорошо - он увидел 3_{10}-спираль, и она там действительно есть, и углы C-O-N нормальные (не показаны). STICKS тоже узнал 3_{10}-спираль - тут считающиеся им углы поворота между остатками играют решающую роль. STRIDE не нашел ничего - только петли с водородными связями.
PSEA, так хорошо показавший себя в предыдущих 2 случаях, тут заинтересовал: нашел одиночный бета-лист. С одной стороны, это может казаться ошибкой - парного листа среди остатков 402-407 он не находит, но с другой стороны именно этот одиночный лист наводит на мысль, что бета лист действительно может быть - напомню, структура была не очень хорошего качества (по результатам практикума 6). К сожалению, 3_{10}-спираль он не нашел.

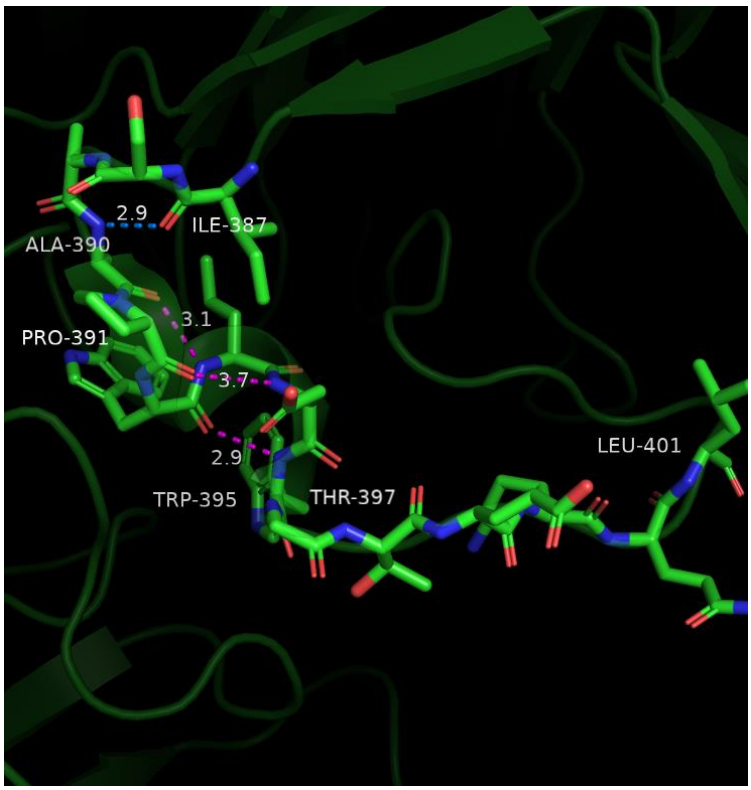


Рис.3. Остатки 387-401

Вывод: ни один из алгоритмов не показал идеальных результатов. Однако DSSP тут ведет себя лучше, потому что ошибка II рода хуже ошибки I рода. DSSP иногда находит структуру там, где ее нет, однако такие кейсы могут быть теоретически проверены вручную. Если же, как в случае PSEA, алгоритм вообще не найдет вторичной структуры, то отследить это будет уже труднее. Видимо, поэтому большинство исследователей пользуется DSSP по умолчанию (структура 2cn3 - не исключение).

Задание 2. Разметка в автоматическом режиме

Задача состояла в том, чтобы на небольшом массиве данных файлов pdb оценить склонность разных аминокислотных остатков входить в элементы вторичных структур. Для этого взяли pdb из списка, провели разметку вторичной структуры алгоритмом DSSP, реализованным в программе mkdssp на kodomo (написали цикл в bash). С помощью скрипта parse_dssp.py получили удобные таблицы, показывающие разметку каждого остатка по 3-меточной системе C – петли, E – листы, H – спирали (написали цикл в bash). Соединили файлы, посчитали все по формуле **Pik = (n_ik/n_i) / (N_k/N)**, получили таблицу из 20 строк, содержащую количества аминокислот такого типа, входящего в такую вторичную структуру (n_ik) и склонность аминокислот такого типа входить в такую вторичную структуру (P_ik) (написали скрипт на R).
Больше склонны входить **в спираль**: лейцин, аланин, глутамин, аргинин.
Более склонны входить **в бета-лист**: валин, изолейцин, фенилаланин, тирозин, цистеин.
Более склонны входить **в петли**: пролин, глутамат, глицин, аспартат.

Таблица 1. Склонность остатков входить в элементы вторичных структур (C – петли, E – листы, H - спирали)

aa	n_ik_C	n_ik_E	n_ik_H	P_ik_C	P_ik_E	P_ik_H
A	154	68	163	0,857	0,758	1,410

C	25	20	16	0,878	1,407	0,874
D	148	34	50	1,367	0,629	0,718
E	122	43	104	0,972	0,686	1,288
F	47	57	47	0,667	1,620	1,037
G	201	45	50	1,455	0,652	0,563
H	57	20	36	1,081	0,760	1,061
I	66	102	71	0,592	1,831	0,990
K	109	49	78	0,989	0,891	1,101
L	127	83	161	0,733	0,960	1,446
M	33	14	13	1,178	1,001	0,722
N	116	19	35	1,462	0,480	0,686
P	159	15	22	1,738	0,328	0,374
Q	62	23	59	0,922	0,685	1,365
R	82	50	87	0,802	0,980	1,323
S	142	56	62	1,170	0,924	0,794
T	102	59	50	1,036	1,200	0,789
V	100	141	77	0,674	1,903	0,807
W	21	16	18	0,818	1,248	1,090
Y	46	44	35	0,788	1,511	0,933