Поиск доменов в белках

Менторы: Юрий Вяткин (yuri@nprog.ru), Анастасия Бакулина (bakulina@gmail.com)

Задача: предсказать пространственные домены белков, если известна их трехмерная структура, но не известно разбиение этой структуры по доменам.

Результат проекта: новый инструмент для биоинформатиков, который будет использоваться для классификации белков, предсказания их свойств, проектирования новых белков.

Белки в клетке формируют структуры на различных уровнях. *Первичной структурой* белка называют последовательность формирующих его аминокислот, прочитываемую с С-конца к N-концу. Например, белок Cas9 имеет следующую первичную структуру:

`KKYSIGLAIGTNSVGWAYITDEYKVPSKK...SITGLYETRIDLSQLG` (всего 1368 аминокислот)

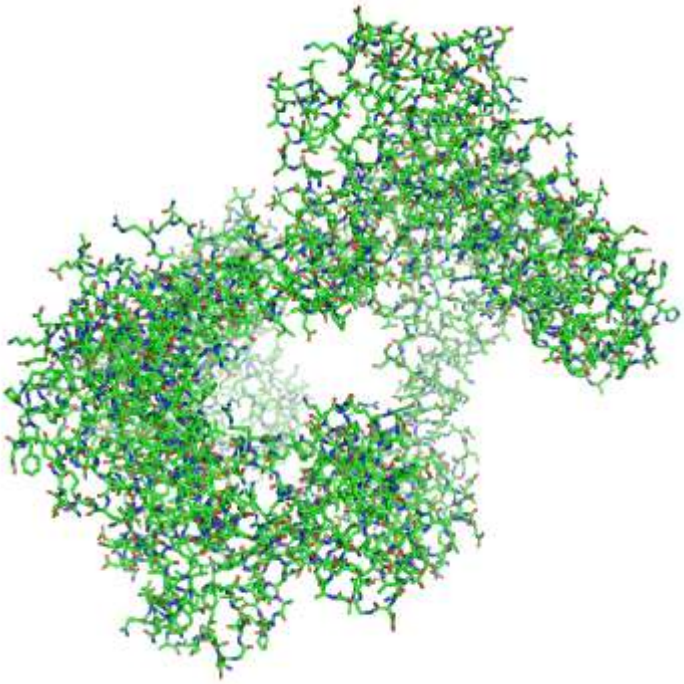
где каждый символ кодирует одну из 20-ти аминокислот.

Вторичная структура белка определяется наличием определенных структур в аминокислотной цепи: альфа-спиралей и бета-листов. *Третичная (пространственная) структура* белка определяется взаимным расположением аминокислот в пространстве. Как правило, третичная структура белка выявляется методами рентгеноструктурного анализа и записывается в формате *PDB*. Например, третичная структура белка Cas9 имеет следующее описание:

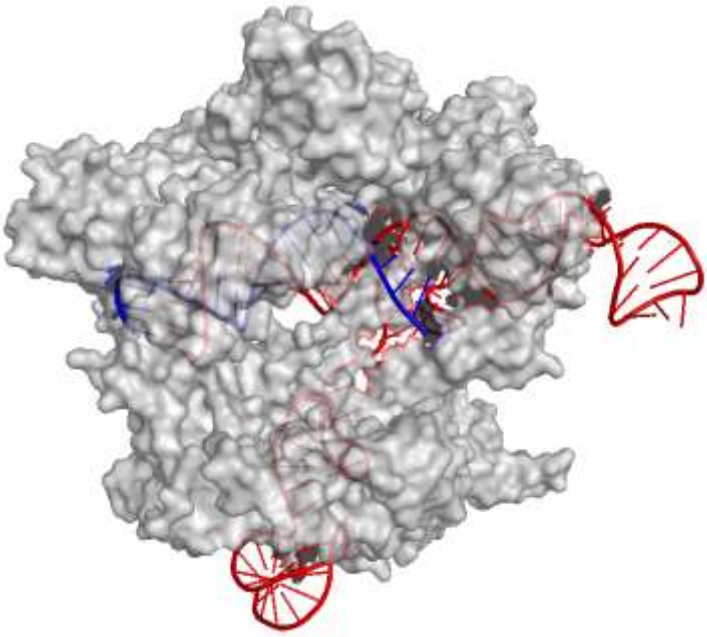
```
...
ATOM      1  N   LYS A   3      -6.376  49.345 -16.720  1.00120.08   N
ATOM      2  CA  LYS A   3      -7.578  49.543 -17.522  1.00119.89   C
ATOM      3  C   LYS A   3      -7.455  50.780 -18.406  1.00117.41   C
ATOM      4  O   LYS A   3      -6.842  51.775 -18.020  1.00116.81   O
ATOM      5  CB  LYS A   3      -8.810  49.660 -16.622  1.00119.93   C
ATOM      6  N   LYS A   4      -8.043  50.709 -19.596  1.00117.22   N
ATOM      7  CA  LYS A   4      -8.001  51.818 -20.541  1.00114.26   C
ATOM      8  C   LYS A   4      -9.306  52.608 -20.520  1.00111.91   C
ATOM      9  O   LYS A   4     -10.369  52.075 -20.841  1.00115.64   O
ATOM     10  CB  LYS A   4      -7.717  51.303 -21.955  1.00113.93   C
...
```

где числа вида -6.376, 49.345, -16.720 - X,Y,Z-координаты атома азота (N) в аминокислоте LYS (тоже, что и символ K в первичной последовательности).

Для визуализации пространственных структур существует множество программных решений. Например, пространственную структуру белка Cas9 можно визуализировать с помощью программы PyMol так:

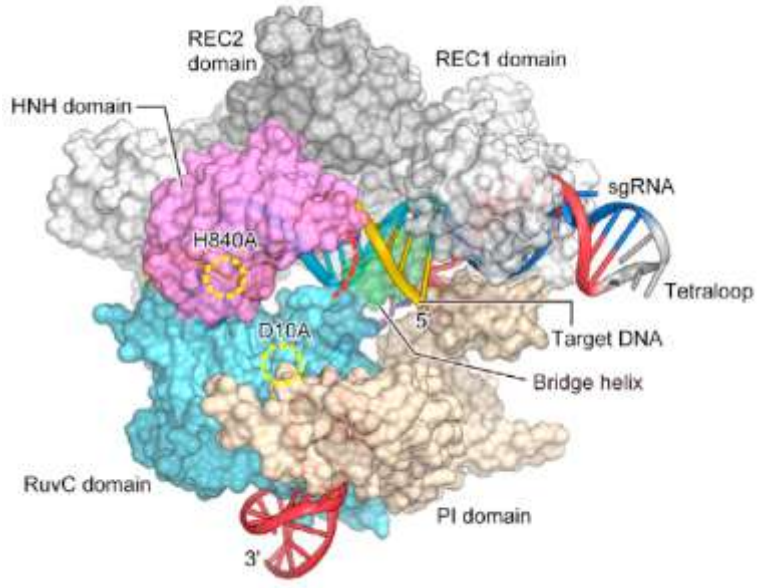


или так



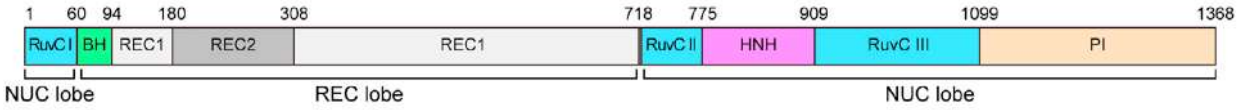
(на рисунке спиралями также показаны молекулы ДНК и РНК, связанные белком Cas9)

С точки зрения функциональной организации, белки состоят из доменов, каждый из которых несет свою функциональность и располагается пространственно отдельно от других доменов. Так, с помощью тщательного анализа, в белке Cas9 были выделены следующие домены: RuvC I, BH, REC1, REC2, RuvC II, HNH, RuvC III, PI. Они визуализируются на пространственной структуре белка следующим образом:



(ср. с предыдущим изображением; из Nishimasu et al. Cell. 2014 Feb 27;156(5):935-49)

Эти домены также отображаются на первичную последовательность белка Cas9:



(из Nishimasu et al. Cell. 2014 Feb 27;156(5):935-49)

Существует *выборка белков*, для которых известна пространственная (и, следовательно, первичная) структура в формате PDB, а также разбиение этих белков по доменам, с указанием координат доменов в аминокислотной последовательности. Все необходимые инструменты визуализации будут предоставлены.

Для решения задачи полезно привлекать методы *машинного обучения*.

Задача заключается в предсказании пространственных доменов белка, если известна его трехмерная структура, но не известно разбиение структуры по доменам. Выходными данными может быть список доменов белка, их координаты на первичной аминокислотной последовательности и пространственная визуализация.