

Постановка задачи

Задание: Нужно считать рекурсивно все файлы из каталога. Каждый файл содержит множественное выравнивание. По набору множественных выравниваний построить матрицу BLOSUM X, где X - процент идентичных карт.

Входные данные: На вход поступает путь к каталогу и целочисленное значение X.

Выходные данные: Матрица BLOSUM X.

Работа с файлами

Функция `walk` модуля `os` предназначена для обхода каталога. Она возвращает генератор с помощью которого можно генерить кортежи из трёх элементов:

```
>>> for root, dirs, files in os.walk(path):  
...     print(root, dirs, files)  
"/test" ["test1", "test2"] [".DS_Store", "test.c"]  
"/test/test1" [] [".DS_Store", "test"]  
"/test/test2" [] [".DS_Store", "teat.py"]
```

Вычисление матрицы BLOSUM

1. Выбрать блоки множественного выравнивания с заданным порогом схожести X:

A	A	B	C	D	---	B	B	C	D	A
D	A	B	C	D	-A-	B	B	C	B	B
B	B	B	C	D	B	A-	B	C	C	A
A	A	A	C	D	C-	D	C	B	C	D
C	C	B	A	D	B-	D	B	B	D	C
A	A	A	C	A	---	B	B	C	C	C

2. Подсчет пар перехода одного символа в другой для каждого столбца.

Пример

Построим для столбца *AACABA*:

ij	n_i	n_j	c_{ij}^k
AA	4	4	$\frac{(4)(4-1)}{2} = 6$
AB	4	1	$4 * 1 = 4$
AC	4	1	$4 * 1 = 4$
BB	1	1	$\frac{(1)(1-1)}{2} = 0$
BC	1	1	$1 * 1 = 1$
CC	1	1	$\frac{(1)(1-1)}{2} = 0$

3. Далее суммируем для каждого столбца полученные значения:

$$c_{ij} = \sum_k c_{ij}^k$$

$$T = \sum_{i \leq j} c_{ij}$$

$$q_{ij} = \frac{c_{ij}}{T}$$

Пример вычисления для выравнивания с картинки:

$$q_{AB} = \frac{4 + 8 + 0 + 0 + 0 + 0 + 0}{7 \frac{6*5}{2}} = \frac{12}{105}$$

4. Вычисление вероятностей:

$$p_i = q_{ii} + \frac{\sum_{i \neq j} q_{ij}}{2}$$

Ожидаемая частота для каждой пары:

$$e_{ii} = p_i^2$$

$$e_{ij} = 2p_i p_j$$

5. Тогда элемент матрицы BLOSUM:

$$s_{ij} = 2 \log_2 \frac{q_{ij}}{e_{ij}}$$