Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет математики

**Струихина Ксения Александровна**

# Математические аспекты выделения достоверных блоков в выравниваниях аминокислотных последовательностей

Курсовая работа студентки 3 курса

образовательной программы бакалавриата «Математика»

Научный руководитель:

Кандидат физико-математических наук,
в.н.с. МГУ им. М.В.Ломоносова
Алексеевский Андрей Владимирович

Москва 2022

# 1 Введение

*Аминокислотная последовательность* — строка из двадцати букв латинского алфавита, соответствующих двадцати основным аминокислотам. Это общепринятый способ описания первичной структуры белка, поскольку белки состоят из 20 разных аминокислот. *Гомологичными* называют белки, произошедшие от общего предка. Последовательности аминокислот гомологичных белков отличаются друг от друга небольшим количеством замен букв, вставок и удалений (делеций).

*Выравнивание аминокислотных последовательностей* — способ отобразить эволюцию нескольких аминокислотных последовательностей от общего предка так, чтобы буквы, произошедшие от одного и того же остатка предка, оказались в одной колонке. Добиваются этого на основании сходства участков аминокислотных последовательностей. При этом учитывается частота замен разных аминокислот друг на друга в правильных выравниваниях. При выравнивании аминокислотных последовательностей также используются знаки разрыва (гэпы) между буквами для расположения схожих позиций друг под другом [1].

Столбец выравнивания, в котором находится одна буква для всех последовательностей, называется *консервативным*. Аминокислоты отвечают за структуру белков, поэтому у родственных последовательностей имеются консервативные колонки, которые отвечают за схожие функции. Для анализа эволюции родственных белков важно знать участки, в которых консервативные позиции делают соответствующие участки отличными от случайного сопоставления последовательностей. Такие участки назовём *достоверными*, идентификация их помогает определять участки, влияющие на функции белка. Белки обычно состоят из одной или нескольких функциональных областей, которые называются *доменами*. Различные комбинации доменов порождают разнообразные белки, встречающиеся в природе. Таким образом, идентификация доменов, которые встречаются внутри белков, может дать представление об их функциях.


Рис. 1: Пример выравнивания четырёх аминокислотных последовательностей.

# 2 Постановка задач

Для определения консервативных участков будем рассматривать консервативные столбцы. Определим, какое количество консервативных столбцов в участке достаточно, чтобы считать данный участок достоверным.

*Блоком* назовём участок выравнивания аминокислотных последовательностей.

Для начала рассмотрим простую модель, в которой все аминокислоты имеют одинаковую вероятность появления в аминокислотной последовательности.

**Задача 1.** Найти вероятность получить по крайней мере $m$ консервативных позиций в блоке при выравнивании $k$ случайных последовательностей длины $n$ без гэпов в предположении равновероятности аминокислот.

**Решение.** Пусть $a$ – число аминокислот ($a = 20$). Число выравниваний с $s$ консервативными позициями, где $n \geq s \geq m$:

$$\binom{n}{s} \cdot a^s \cdot (a^k - a)^{n-s} = \binom{n}{s} \cdot a^s \cdot a^{n-s} \cdot (a^{k-1} - 1)^{n-s} = \binom{n}{s} \cdot a^n \cdot (a^{k-1} - 1)^{n-s}.$$

Просуммируем полученное выражение по всем $s$, где $n \geq s \geq m$ и получим число выравниваний с хотя бы $s$ консервативными позициями:

$$\sum_{s=m}^{n} \binom{n}{s} \cdot a^n \cdot (a^{k-1} - 1)^{n-s} = a^n \cdot \sum_{s=m}^{n} \binom{n}{s} \cdot (a^{k-1} - 1)^{n-s}.$$

Тогда вероятность получить по крайней мере $m$ консервативных позиций в блоке при выравнивании $k$ случайных последовательностей длины $n$ без гэпов в предположении равновероятности аминокислот:

$$P = \frac{a^n \cdot \sum_{s=m}^{n} \binom{n}{s} \cdot (a^{k-1} - 1)^{n-s}}{a^{kn}} = \frac{\sum_{s=m}^{n} \binom{n}{s} \cdot (a^{k-1} - 1)^{n-s}}{a^{(k-1)n}} = \frac{\sum_{s=m}^{n} \binom{n}{s} \cdot (\lambda - 1)^{n-s}}{\lambda^n},$$

где $\lambda = a^{k-1}$.

Однако каждая аминокислота имеют свою вероятность появления в аминокислотной последовательности.

Таблица 1: Распределение частот аминокислотных остатков

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Ala(A) | 0.0906 | Gln(Q) | 0.0381 | Leu(L) | 0.0988 | Ser(S) | 0.0679 |
| Arg(R) | 0.0582 | Glu(E) | 0.0623 | Thr(T) | 0.0555 | Asn(N) | 0.0379 |
| Lys(K) | 0.0493 | Gly(G) | 0.0727 | Met(M) | 0.0234 | Trp(W) | 0.0130 |
| Asp(D) | 0.0546 | His(H) | 0.0222 | Phe(F) | 0.0388 | Tyr(Y) | 0.0288 |
| Cys(C) | 0.0128 | Ile(I) | 0.0554 | Pro(P) | 0.0497 | Val(V) | 0.0687 |

**Задача 2.** Найти вероятность получить по крайней мере $m$ консервативных позиций в блоке при выравнивании $k$ случайных последовательностей длины $n$ без гэпов в предположении различных вероятностей для аминокислот.

**Решение.** Пусть нам даны вероятности $p_1, p_2, ..., p_{20}$ аминокислот. Вероятность получить один столбец в выравнивании, составленный только из $i$-той аминокислоты

равна $p_i^k$. Тогда вероятность получить $s$ консервативных столбцов у $k$ последовательностей равна

$$(p_1^k + p_2^k + ... + p_{20}^k)^s,$$

поскольку каждый моном при раскрытии скобок в этом выражении отвечает за выбор аминокислоты на соответствующее место.

Вероятность получить неконсервативный столбец равна $1-(p_1^k+p_2^k+...+p_{20}^k)$ как дополнение до полной вероятности.

Вероятность получить выравнивание в предположении различных вероятностей для аминокислот с $s$ консервативными позициями, где $n \geq s \geq m$:

$$\binom{n}{s} \cdot (p_1^k + p_2^k + ... + p_{20}^k)^s \cdot (1 - (p_1^k + p_2^k + ... + p_{20}^k))^{n-s}$$

Просуммируем полученное выражение по всем $s$, где $n \geq s \geq m$ и получим вероятность выравнивания с хотя бы $s$ консервативными позициями:

$$\sum_{s=m}^{n} \binom{n}{s} \cdot (p_1^k + p_2^k + ... + p_{20}^k)^s \cdot (1 - (p_1^k + p_2^k + ... + p_{20}^k))^{n-s} = \sum_{s=m}^{n} \binom{n}{s} p^s \cdot (1-p)^{n-s},$$

где $p = p_1^k + p_2^k + ... + p_{20}^k$.

На основании формулы, полученной в предыдущей задаче, подберём параметры для выявления достоверных блоков. Так как для небольшого числа последовательностей появление консервативного столбца может быть случайностью, то будем рассматривать выравнивание хотя бы шести последовательностей.

Участки, влияющие на функцию белка, содержат аминокислоты, стоящие рядом. Поэтому консервативные столбцы на небольшом расстоянии говорят о более вероятной достоверности фрагмента выравнивания. В работе будем рассматривать блоки длины 10.

Поскольку один консервативный столбец может быть подогнан алгоритмом выравнивания даже для неродственных позиций, то найдём вероятность получить хотя бы два консервативных столбца в блоке длины 10 при выравнивании аминокислотных последовательностей.

**Задача 3.** Вероятность получить хотя бы две консервативные позиции в блоке длины 10 при выравнивании случайных последовательностей для разных вероятностей аминокислот.

**Решение.** Для начала мы найдём дополнение: вероятность получить не более одной консервативной позиции в блоке длины 10.

$$(1 - (p_1^k + p_2^k + ... + p_{20}^k))^{10} + 10 \cdot (1 - (p_1^k + p_2^k + ... + p_{20}^k))^9 \cdot (p_1^k + p_2^k + ... + p_{20}^k))$$

Тогда искомая вероятность равна

$$1 - (1 - (p_1^k + p_2^k + ... + p_{20}^k))^{10} - 10 \cdot (1 - (p_1^k + p_2^k + ... + p_{20}^k))^9 \cdot (p_1^k + p_2^k + ... + p_{20}^k)).$$

Для $k = 6$ эта вероятность меньше $2 \cdot 10^{-10}$, поэтому при наличии хотя бы двух консервативных столбцов в окне длины 10 будем считать данный участок достоверным.

# 3 Методы

Для выделения достоверных участков была написана программа. Её работа основывается на следующих шагах:

- Для работы были взяты последовательности из базы данных *Pfam* [2] в формате STOCKHOLM.
- Из последовательностей были отобраны выравнивания хотя бы 6 последовательностей длины не меньше 30.
- При наличии в выравнивании более, чем сорока последовательностей, выбирались тридцать из них.
- Каждое из выравниваний было разбито на блоки без гэпов.
- Для каждого блока были найдены достоверные подблоки – участки длины 10, содержащие хотя бы два консервативных столбца.
- Достоверные участки блока определялись как объединение достоверных подблоков.

База данных *Pfam* представляет собой коллекцию белковых семейств, каждое из которых представлено выравниванием нескольких последовательностей. Эти последовательности определяют домены, которые дают представление о функции белка. Последовательности из этих выравниваний были выбраны для работы, поскольку для сохранения функции белка у родственных последовательностей необходима консервативность или функциональная консервативность (замена одной аминокислоты на другую с сохранением функции) позиций.

# 4 Результаты и обсуждение

Таблица 2: Результат подсчёта программой достоверных блоков для части данных

| **id** | **Число посл-тей в выравн.** | **Длина посл.** | **Сумма длин блоков без гэпов** | **Сумма длин дост. блоков** | **Кол-во дост. блоков** | **% дост. блоков** |
|---|---|---|---|---|---|---|
| 1-cysPrx_C | 40 | 55 | 0 | 0 | 0 | 0 |
| 12TM_1 | 7 | 504 | 0 | 0 | 0 | 0 |
| 14-3-3 | 30 | 301 | 69 | 18 | 1 | 6 |
| 17kDa_Anti | 6 | 122 | 87 | 65 | 2 | 53 |
| Hacid_dh | 30 | 363 | 240 | 0 | 0 | 0 |
| Hacid_dh_C | 30 | 258 | 0 | 0 | 0 | 0 |
| oxoacid_dh | 30 | 299 | 63 | 17 | 1 | 6 |
| oxogl_dehyd | 30 | 44 | 40 | 15 | 1 | 34 |
| ph_phosp | 30 | 344 | 30 | 0 | 0 | 0 |
| thiour_desulf | 30 | 250 | 0 | 0 | 0 | 0 |
| 23ISL | 16 | 169 | 29 | 21 | 1 | 12 |
| 23S_rRNA | 30 | 127 | 33 | 0 | 0 | 0 |
| 2CSK_N | 30 | 247 | 0 | 0 | 0 | 0 |
| 2C_adapt | 18 | 40 | 0 | 0 | 0 | 0 |
| 2EXR | 30 | 291 | 0 | 0 | 0 | 0 |
| 2S_thioredx | 30 | 198 | 39 | 0 | 0 | 0 |
| phosphodiest | 13 | 126 | 0 | 0 | 0 | 0 |
| 2HCT | 30 | 518 | 275 | 72 | 3 | 14 |
| FeII_Oxy | 30 | 242 | 0 | 0 | 0 | 0 |
| FeII_Oxy_2 | 30 | 465 | 0 | 0 | 0 | 0 |
| FeII_Oxy_3 | 30 | 251 | 0 | 0 | 0 | 0 |
| FeII_Oxy_4 | 16 | 119 | 0 | 0 | 0 | 0 |
| FeII_Oxy_5 | 31 | 131 | 30 | 0 | 0 | 0 |
| FeII_Oxy_6 | 42 | 330 | 0 | 0 | 0 | 0 |
| Fe_Oxy_2 | 30 | 288 | 0 | 0 | 0 | 0 |
| 2TM | 30 | 146 | 0 | 0 | 0 | 0 |
| RNA_ligase2 | 30 | 396 | 0 | 0 | 0 | 0 |
| 3-alpha | 30 | 46 | 0 | 0 | 0 | 0 |
| dmu-9_3 | 19 | 138 | 31 | 0 | 0 | 0 |
| 3-PAP | 23 | 152 | 0 | 0 | 0 | 0 |
| 3A | 8 | 286 | 0 | 0 | 0 | 0 |
| 3Beta_HSD | 8 | 301 | 65 | 35 | 2 | 12 |
| 3D | 41 | 83 | 0 | 0 | 0 | 0 |
| 3H | 30 | 111 | 34 | 0 | 0 | 0 |
| 3HBOH | 13 | 748 | 463 | 378 | 10 | 51 |
| 3HCDH | 30 | 109 | 0 | 0 | 0 | 0 |
| 3HCDHN | 30 | 209 | 0 | 0 | 0 | 0 |

Для более подробного анализа приведём Таблицу 2 вывода программы. Как мы видим в таблице, более половины последовательностей не содержат блоков без гэпов. Это значит, что не нашлось тридцати подряд идущих позиций, в которых нет гэпов ни у одной из последовательностей.

У тех пятнадцати последовательностей, у которых нашлись блоки без гэпов, восемь выравниваний имеют достоверные блоки, при этом количество блоков почти везде, кроме самого длинного выравнивания, не больше трёх. Можно предположить, что именно эти блоки отвечают за функции семейства белков.

Кроме того, программа была запущена на всём файле и построены гистограммы распределения процента достоверных блоков.

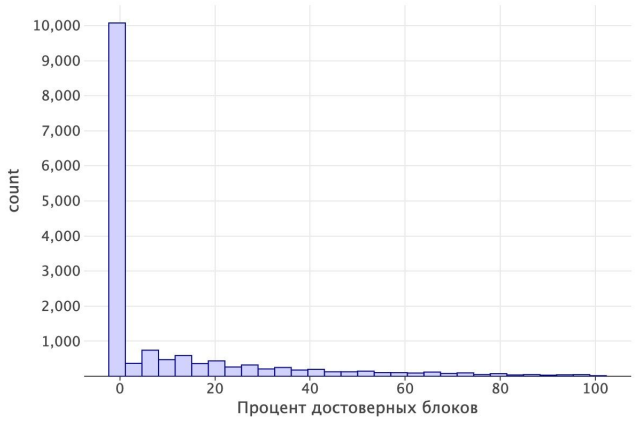


Рис. 2: Распределение процента достоверных блоков для всего файла.

Заметим, что во всём файле из 15825 выравниваний более десяти тысячь не имеют достоверных блоков. Причиной этому могут быть одна отличающаяся от других буква в столбце, которая ломает консервативный столбец или замена одной буквы на другую с сохранением функций. Кроме того, для выравниваний, в которых нет блоков длины хотя бы 30 без гэпов, мы не считали достоверные блоки. Поэтому отдельно рассмотрим выравнивания, где нашлись блоки без гэпов.

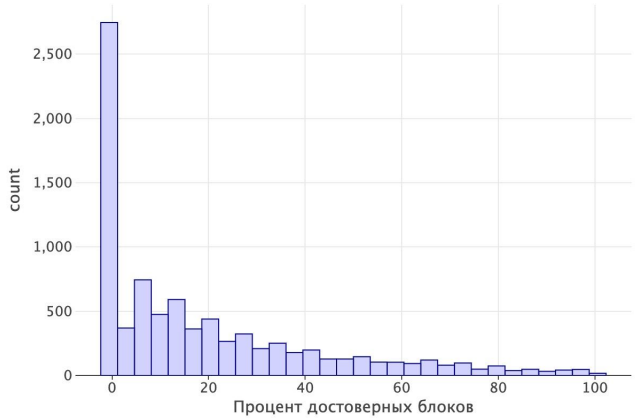


Рис. 3: Распределение процента достоверных блоков для выравниваний с ненулевой суммой длин блоков без гэпов.

Как мы видим на гистограмме, количество выравниваний без достоверных блоков

уменьшилось почти в 4 раза. Но их до сих пор больше половины от всех выравниваний, поэтому отдельно также рассмотрим выравнивания с ненулевым количеством достоверных блоков.

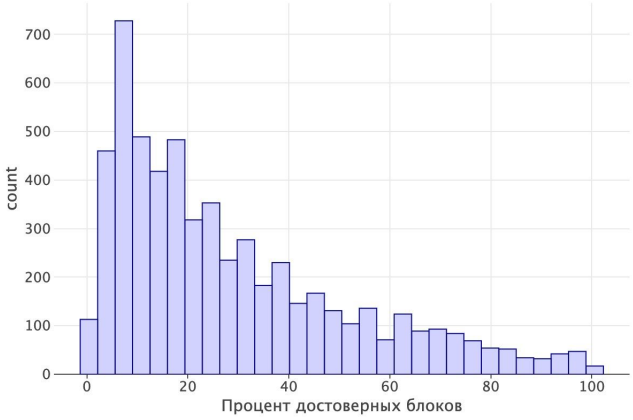


Рис. 4: Распределение процента достоверных блоков для выравниваний с ненулевым числом достоверных блоков.

Судя по гистограмме, около половины выравниваний с процентом достоверных блоков большим 20. Такое значение говорит о том, что в этих выравниваниях белки действительно гомологичны. Наконец рассмотрим, какое количество достоверных блоков у выравниваний с ненулевым процентом достоверных блоков.


Рис. 5: Распределение количества достоверных блоков для выравниваний с ненулевым числом достоверных блоков.

Большинство выравниваний имеют один достоверный блок, при этом длина этого блока у выравниваний занимает около 20% от всей длины выравнивания, судя по гистограмме выше. Вероятно, в этих блоках и находятся функциональные части белков.

# 5 Список литературы


1. Хаубольд Б., Вие Т. Введение в вычислительную биологию: эволюционный подход. – М.-Ижевск: НИЦ "Регулярная и хаотическая динамика". Ижевский институт компьютерных исследований, 2011. – 456 с.

2. Pfam: The protein families database in 2021: J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman

# 6 Приложение

Ссылка на ноутбук, в котором реализована программа, а также проведён подсчёт вероятности вероятности получения хотя бы двух консервативных столбцов в блоке длины 10.

https://colab.research.google.com/drive/1Zwy7MYRSy2MoPZcgN45EI192lTvCLAnz