ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №5

Значимость выравнивания. Экспресс-сравнение последовательностей (BLAST). Построение и визуализация профилей множественных выравниваний.

Новоселецкий Валерий Николаевич
к.ф.-м.н., доц. каф. биоинженерии
valery.novoseletsky@yandex.ru

Сайт курса http://intbio.org/bioinf2018

Надо ли обсудить результаты КР на лекции?

Принято ответов - 202

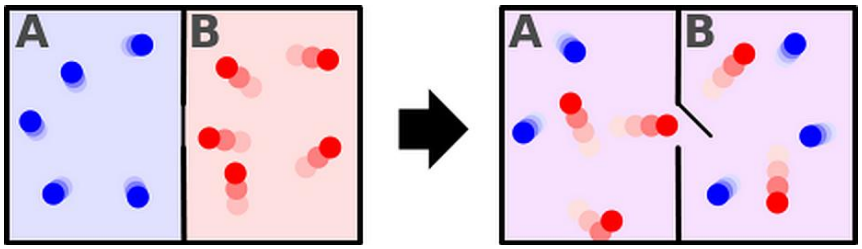
Ответов на 10 баллов – 6

«Всё и так понятно» - 25

Не ходят на лекции - 4

Задача №1

Как изменится термодинамическая энтропия системы при смешении 1 моль Ar и 2 моль Ne при н.у.?



155 / 202 correct responses

Используем понятие мольной доли Ni = ni/(n1+n2+…)

Тогда изменение энтропии при смешении двух идеальных газов

$$\Delta S = -(n1+n2)\, R\, (N1 \ln(N1) + N2 \ln(N2)) =$$

$$= -3\, R\, (\tfrac{1}{3} \ln(\tfrac{1}{3}) + \tfrac{2}{3} \ln(\tfrac{2}{3})) = -R\, (\ln(\tfrac{1}{3}) + 2 \ln(\tfrac{2}{3})) = R\, (3 \ln 3 - 2 \ln 2) > \mathbf{0}$$

Смешение газов необратимо!

Рассмотрение смешения двух объемов идентичного газа приводит к парадоксу Гиббса

Задача №2

Сколько бит необходимо для кодирования каждой буквы в 20-буквенном (аминокислоты) алфавите? Ответ укажите с точностью до сотых долей бита.

$$H = -\sum_{a \in \{s\}} p(a) \log_2 p(a)$$

При условии равномерной встречаемости букв $H = log(2)(20) = 4{,}32$

Задача №3

В поисках бозона Хиггса Большой адронный коллайдер производит 15 Пбайт информации в год. Какой способ передачи этих данных в Нью-Йорк займет меньше времени?

Хранилище такого объема помещается в стандартный контейнер, который может быть перевезен самолетом.

Задача №4

Используя понятие контрольной суммы и Международный идентификационный код ценной бумаги, восстановите последнюю цифру подлинного номера банковской карты 4287 0001 1385 061x

Последняя цифра 0. Справились почти все.

Задача №5

Запишите вторые 10 нуклеотидов в последовательности геномной ДНК JLA области из хлоропласта *Nicotiana tabacum*, о которой шла речь на лекции №3.

Правильный ответ `ggagagaaaa` или `GGAGAGAAAA` – см. Genbank.
Справились почти все.

Задача №6

Постройте разумное выравнивание и определите расстояние по Левенштейну между словами БИОИНФОРМАТИКА и КОНФОРМАЦИЯ

```
БИОИНФОРМАТИКА
-КО-НФОРМАЦИ-Я
```

Справились почти все. 29445 - ?!

Задача №7

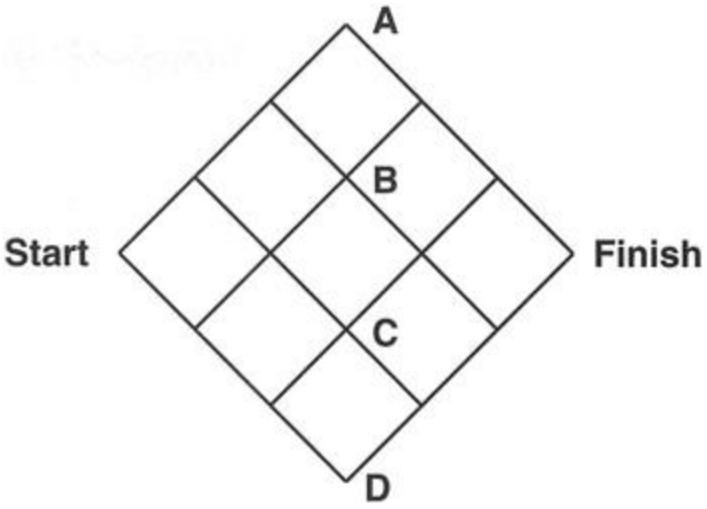
Какая аминокислотная замена (N<->D или K<->R) более вероятна по данным матрицы замен BLOSUM62?

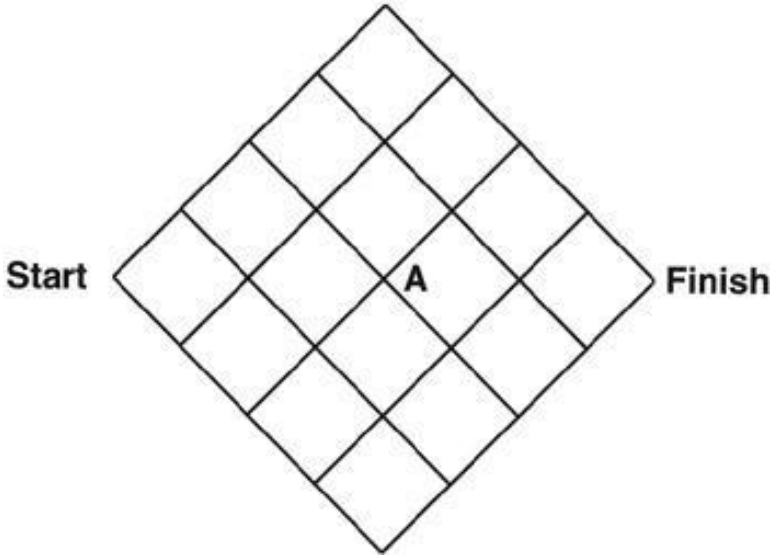
Правильный ответ K<->R (2 балла против 1 балла за N<->D)

Справились почти все.

Задача №8

При условии, что движение возможно только слева направо, рассчитайте, сколько путей от старта к финишу проходят через точку B (см. рис.). Ответ: 9.





Задача №9

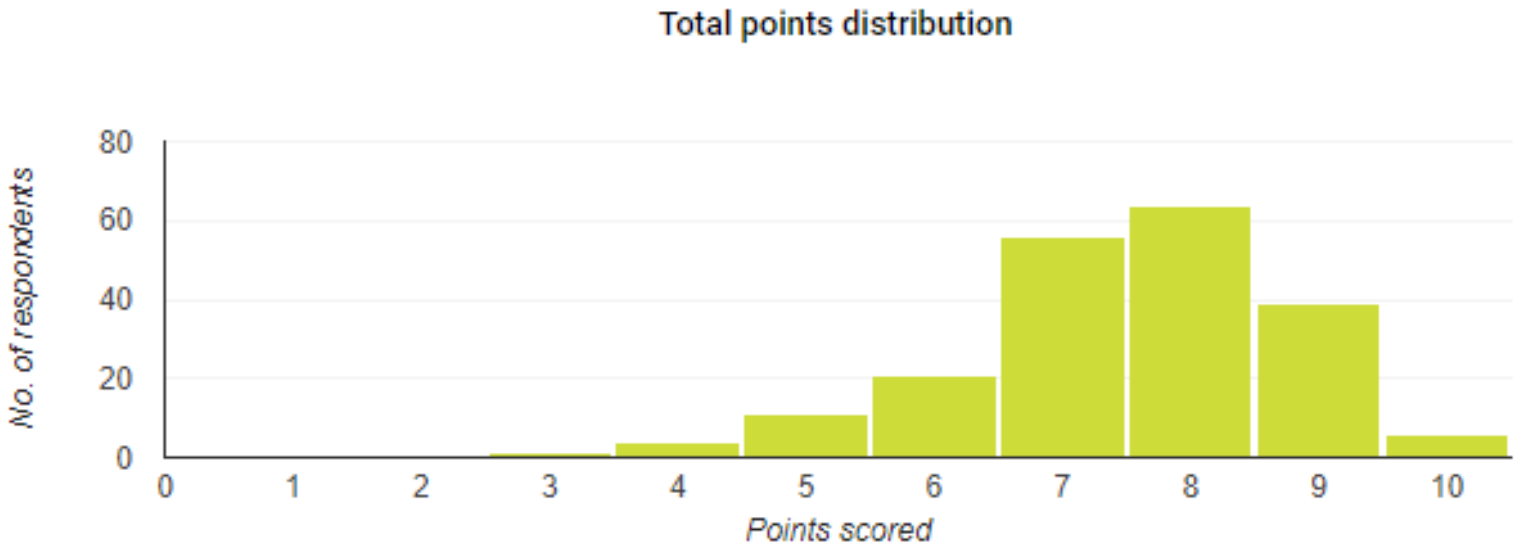
При условии, что движение возможно только слева направо, рассчитайте, сколько всего существует путей от старта к финишу (см. рис.). Воспользуйтесь тем, что любой путь состоит из 8 отрезков (по 4 восходящих и нисходящих), а уникальность пути определяется последовательностью этих отрезков. Ответ 70.

Задача №10

В силу избыточности генетического кода, одна и та же аминокислотная последовательность может быть закодирована различными нуклеотидными последовательностями. Каково максимальное число различных последовательностей ДНК, кодирующих некую последовательность из 10 аминокислот?

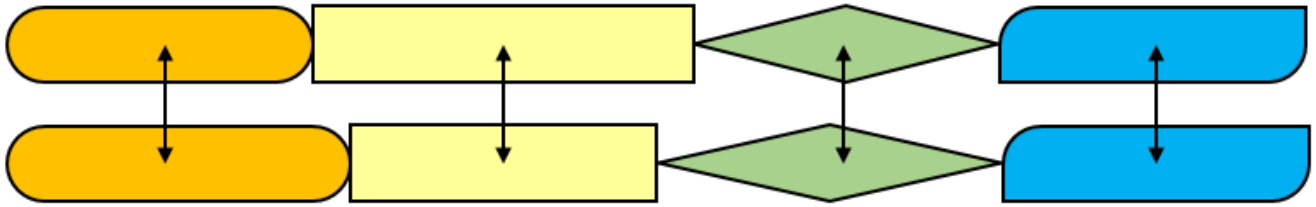
6^10 = 60466176

Справились почти все.

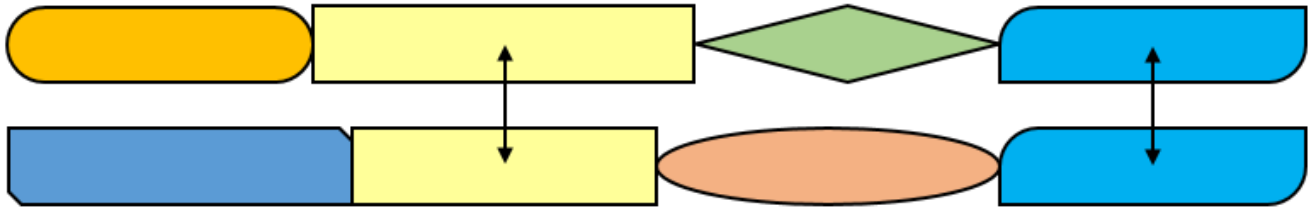


Расчет выравнивания двух последовательностей. Алгоритм Смита-Уотермана (1981)

– модификация алгоритма Нидлмана-Вунша с аффинным штрафом за вставку и обнулением отрицательных значений в матрице – позволяет выявлять локальные выравнивания.



Global Alignment



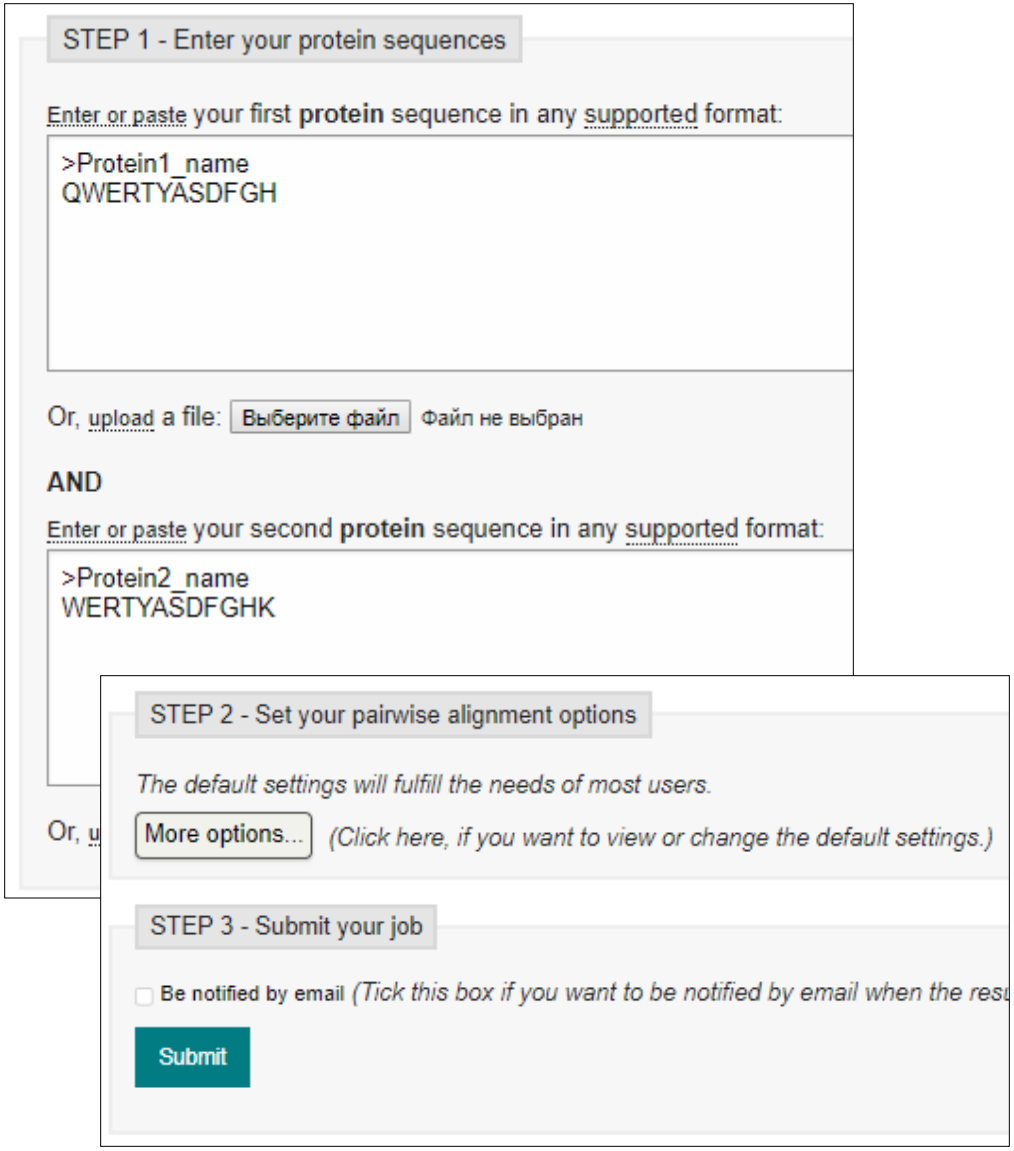
Local Alignment

Расчет выравнивания двух последовательностей. Алгоритм Смита-Уотермана

	"-"	g	c	t	g	a	a	c	g
"-"	0	0	0	0	0	0	0	0	0
c	0	0	0,5	0	0	0	0	0,5	0
t	0	0	0	1	0	0	0	0	0
a	0	0	0	0	0	0,5	0,5	0	0
t	0	0	0	0,5	0	0	0	0	0
a	0	0	0	0	0	0,5	0,5	0	0
a	0	0	0	0	0	0,5	1	0	0
t	0	0	0	0,5	0	0	0	0	0
c	0	0	0,5	0	0	0	0	0,5	0

match	mut	gap
0,5	-1	-1

Пример выравнивания двух последовательностей. EMBOSS Needle




```
# Aligned_sequences: 2
# 1: Protein1_name
# 2: Protein2_name
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 13
# Identity: 11/13 (84.6%)
# Similarity: 11/13 (84.6%)
# Gaps: 2/13 (15.4%)
# Score: 67.0
#
#=================================
Protein1_name 1 QWERTYASDFGH- 12
                  ||||||||||
Protein2_name 1 -WERTYASDFGHK 12
```

Значимость выравнивания

Насколько значимо полученное выравнивание?

Имеет ли оно биологический смысл или образовалось случайно?

```
SEQUENCE   1 VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSH   50
             |||.|||||||||||.|||.||||||||||||||.||||||||||||||
SEQUENCE   1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSH   50

SEQUENCE  51 GSAQVKAHGKKVADGLTLAVGHLDDLPGALSDLSNLHAHKLRVDPVNFKL  100
             ||||||.|||||||.||.||.|:||:|.|||.||:|||||||||||||||
SEQUENCE  51 GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKL  100

SEQUENCE 101 LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR  141
             ||||||.|||.|||.:||||||||||||||:||||||||||
SEQUENCE 101 LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR  141

# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 141
# Identity:     123/141 (87.2%)
# Similarity:   128/141 (90.8%)
# Gaps:           0/141 ( 0.0%)
```

Значимость выравнивания

```
SEQUENCE   1 -VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHL    49
              .|:|.:..||...|.:..|::..|.....|.:.:..|...   |.|..|
SEQUENCE   1 GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAK---DLFSFL    47

SEQUENCE  50 KTEAEM-KASEDLKKHGVTV----------LTALGAILKKKGHHEAELKP    88
             |..:|: :.:.:|:.|...|          |...|.::.     :|.||.
SEQUENCE  48 KGTSEVPQNNPELQAHAGKVFKLVYEAAIQLEVTGVVVT-----DATLKN    92

SEQUENCE  89 LAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELF   138
             |...|.:| .:...:...:.|||:.:..    ..||.....:|.|..:.
SEQUENCE  93 LGSVHVSK-GVADAHFPVVKEAILKTIKE----VVGAKWSEELNSAWTIA   137

SEQUENCE 139 RKDIAAKY-KELGYQG   153
             ..::|... ||:....
SEQUENCE 138 YDELAIVIKKEMDDAA   153

# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 166
# Identity:      35/166 (21.1%)
# Similarity:    62/166 (37.3%)
# Gaps:          26/166 (15.7%)
```

Значимость выравнивания

Миоглобин кашалота (1mbn, 1969)

Леггемоглобин люпина (1gdj, 1995) (ИК РАН)

Значимость выравнивания

миоглобин кашалота и леггемоглобин люпина

Идентичность 18%*, но родство подтверждается сходством структуры и функции

What about Impossible Burger?

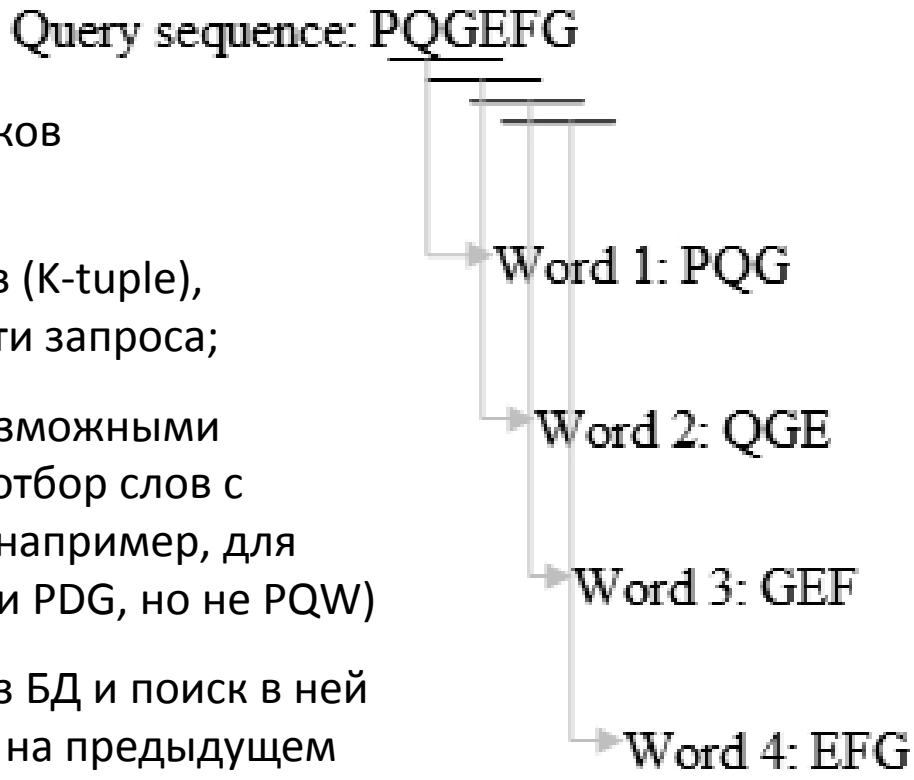
https://impossiblefoods.com/

Экспресс-методы сравнения последовательностей. BLAST

BLAST - **B**asic **L**ocal **A**lignment **S**earch **T**ool (Altschul et al.,1990) предназначен для сравнения новых последовательностей с уже содержащимися в базах данных.

Алгоритм:

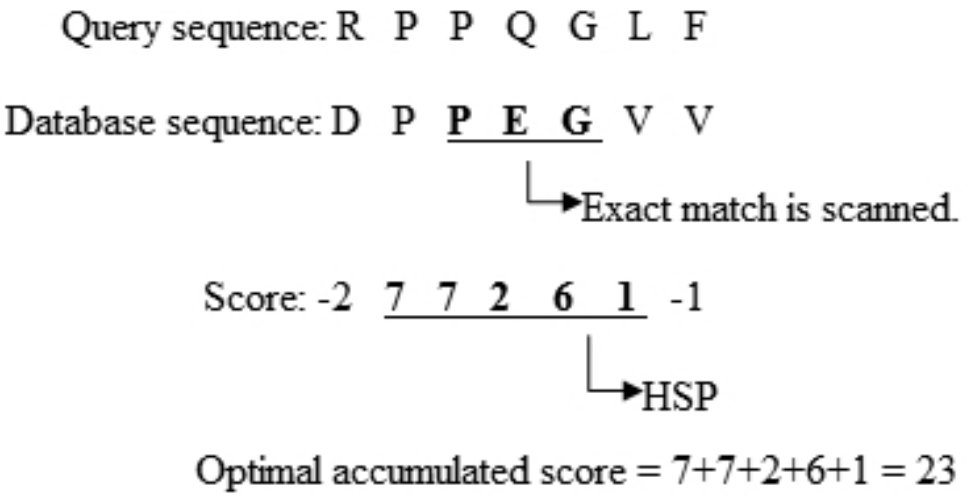
- Удаление малоинформативных участков последовательности (повторы и т.п.);
- Составление списка *к*-буквенных слов (K-tuple), присутствующих в последовательности запроса;
- Сопоставление этих слов со всеми возможными словами длины *к* и оценка сходства; отбор слов с оценкой, превышающей пороговую (например, для слова PQG сходными будут PNG, PEG и PDG, но не PQW)
- Сканирование последовательности из БД и поиск в ней слов с высокой оценкой, полученных на предыдущем шаге;



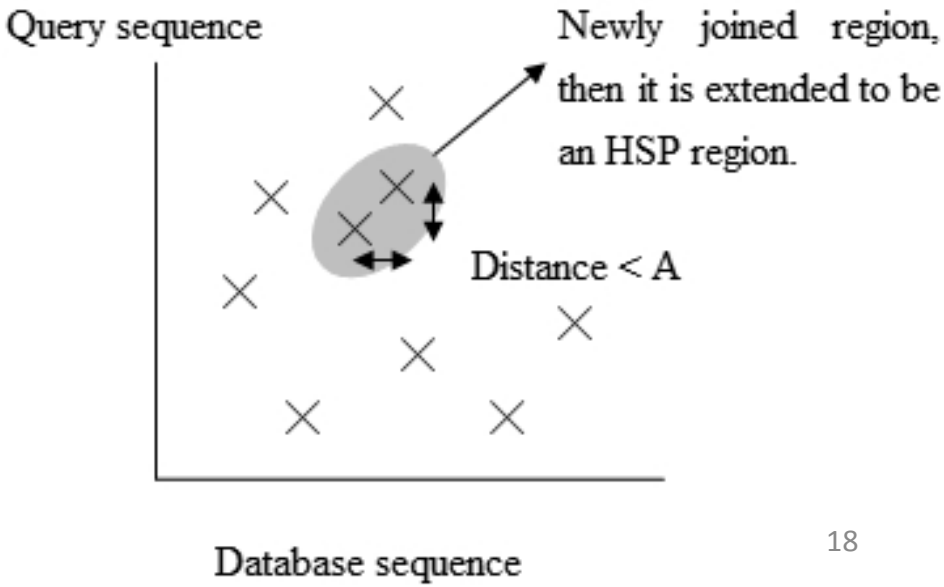
Сколько всего 3х-буквенных слов? Откуда берутся оценки? В чем смысл такого “словаря”?

Экспресс-методы сравнения последовательностей. BLAST

- Расширение локальных выравниваний в обе стороны до тех пор, пока суммарная оценка выравнивания не начинает уменьшаться (построение сегментных пар (high-scoring segment pair, HSP));



- Объединение сегментных пар, лежащих на удалении меньше A;
- Составление списка сегментных областей с высокой оценкой;
- **Расчет статистической значимости этих оценок.**



Бросание кубика

Рассмотрим бросание кубика *n* раз.

Каково математическое ожидание

числа N_l повторов длины l?

$$E(N_l) = (n - l + 1)(1 - p)\,p^l \cong n(1 - p)\,p^l,$$
$$где \;\; p - вероятность\; выпадения\; грани$$

В случае наиболее длинного повтора $N = 1$. Тогда для мат. ожидания длины L этого повтора справедливо

$$1 \cong n(1 - p)\,p^{E(L)} \Rightarrow E(L) = \log_{1/p} n(1 - p)$$

(Erdős, Rényi, 1970?)

Бросание кубика

Точечная матрица – двумерное обобщение бросания кубика с заменой вероятности выпадения грани на вероятность совпадения букв. Для последовательностей длиной n и m имеем

$$E(N) = (n-l+1)(m-l+1)(1-p)\,p^l \cong mn(1-p)\,p^l,$$
$$где\ \ p - вероятность\ совпадения$$

Более точное рассмотрение дает для мат. ожидания длины двумерного повтора выражение

$$E(L) = \log_{1/p}(mn) + \log_{1/p}(1-p) + \gamma \log_{1/p} e - 1/2$$

где «гамма» (постоянная Эйлера) ≈ 0,577 215…

(1735 – Эйлер – 5 знаков, … 1973 – Уотерман – 4879 знаков, …)

Бросание кубика

Переходя к натуральному логарифму и вводя обозначения

$$\log_{1/p}(1-p)+\gamma\log_{1/p}e-1/2\equiv\log_{1/p}K,\qquad \ln 1/p\equiv\lambda$$

Получаем

$$E(L)=\ln(Kmn)/\lambda$$

Аналогичное справедливо (Karlin and Altschul, 1990?) и для максимальной оценки S_n,m сегментной пары (HSP), образованной последовательностями длиной n и m,

$$S_{n,m}\propto\frac{\ln(Kmn)}{\lambda}$$

BLAST. Значимость выравнивания

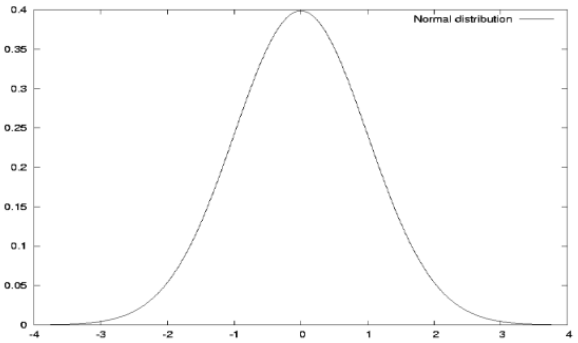
$$\tilde{S}_{n,m} = S_{n,m} - \frac{\ln(Knm)}{\lambda}$$

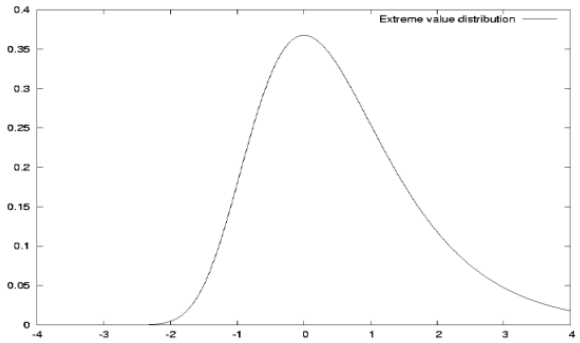
Распределение нормализованных максимальных оценок HSP подчиняется распределению Гумбеля (распределению экстремальных значений, Gumbel, 1937; Гнеденко, 1943), для которого

$$P\left(\tilde{S}_{n,m} > S\right) \approx 1 - \exp\left(-K\,m\,n\,e^{-\lambda S}\right) \approx K\,m\,n\,e^{-\lambda S}$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

$$\varphi(x) = e^{-x} \cdot e^{-e^{-x}}$$





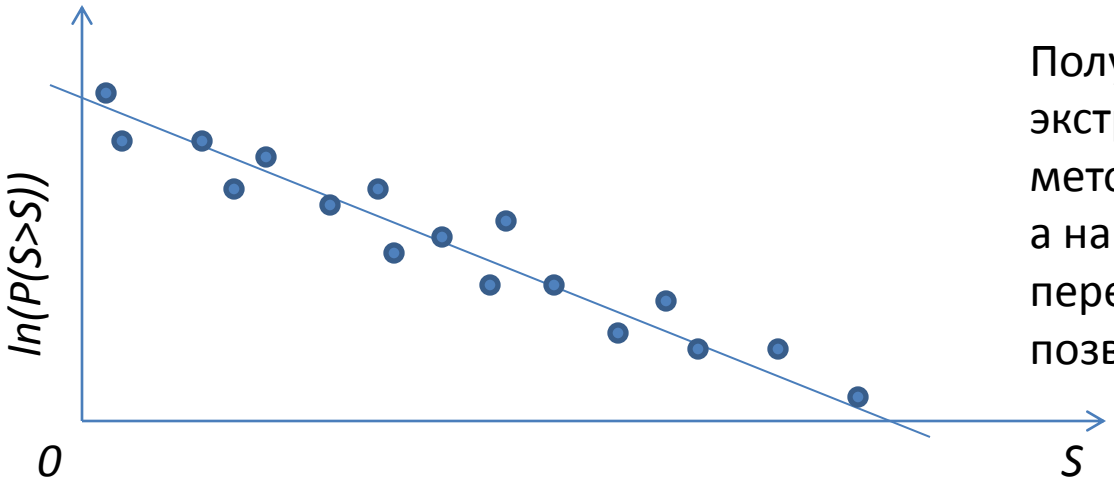
BLAST. Значимость выравнивания

$$P\left(\tilde{S}_{n,m} > S\right) \approx 1 - \exp\left(-K\,m\,n\,e^{-\lambda\,S}\right) \approx K\,m\,n\,e^{-\lambda\,S}$$

Но каковы в этом случае *K* и *λ*? Прологарифмируем найденное выражение:

$$ln\left(P(\tilde{S} > S)\right) = ln(Kmn) - \lambda S$$

Таким образом, мы можем построить много выравниваний для одной и той же последовательности запроса длиной *n* и базы данных суммарной длиной *m* (причем почти все эти выравнивания будут заведомо неправильными) и для каждого выравнивания поставить точку в координатах (S, ln(P(S>S))):



Получившиеся точки можно экстраполировать прямой по методу наименьших квадратов, а наклон этой прямой и её пересечение с осью ординат позволят определить *K* и *λ*.

BLAST. Значимость выравнивания

Подставляем полученные *K* и *λ* в формулу для расчета вероятности:

$$P\left(\tilde{S}_{n,m} > S\right) \approx 1 - \exp\left(-K\, m\, n\, e^{-\lambda S}\right) \approx K\, m\, n\, e^{-\lambda S}$$

$$P(\tilde{S}_{n,m} > S) \cdot m \equiv E - value$$

E < 0,02	высокая вероятность гомологии
0,02 < E < 1	гомология не очевидна
E > 1	сходство случайно

```
>SP:MYG_PHYCD P02185 Myoglobin OS=Physeter catodon OX=9755 GN=MB PE=1
SV=2
Length=154

Score = 104 bits (259),  Expect = 1e-29
Identities = 50/50 (100%), Positives = 50/50 (100%), Gaps = 0/50 (0%)

Query  1    LAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELF  50
            LAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELF
Sbjct  90   LAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELF  139
```

BLAST. Значимость выравнивания

$$P(\tilde{S}_{n,m} > S) \cdot m \equiv E - value$$

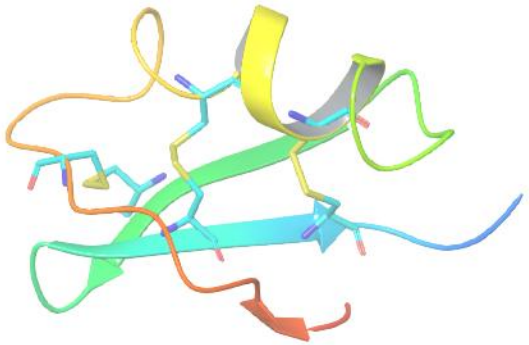
E < 0,02	высокая вероятность гомологии
0,02 < E < 1	гомология не очевидна
E > 1	сходство случайно

Для коротких последовательностей сходство может быть НЕ случайным даже при E > 1 !!

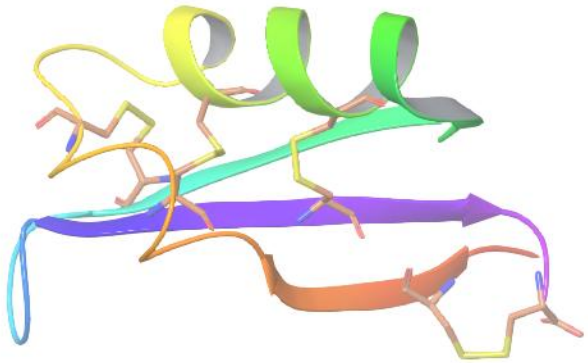
Пример: поиск в PDB по последовательности калиотоксина дает, среди прочего, неожиданный результат:

```
>lcl|PDB:1TI5_A mol:protein length:46 plant defensin Length=46

Score = 25.8 bits (74), Expect = 1.9
Identities = 12/31 (39%), Positives = 14/31 (45%), Gaps = 2/31 (6%)

Query  7 KCSGSPQCLKPCKDAGMRFGKC--MNRKCHC 35  калиотоксин
         KC     C   CK+ G   G C  M R C+C
Sbjct 12 KCLIDTTCAHSCKNRGYIGGNCKGMTRTCYC 42  дефензин
```

BLAST. Значимость выравнивания



калиотоксин



дефензин

```
3ODV_T    1 -GVEINV----KCSGSPQCLKPCKDAGMRFGKCM-N-RKCHCTPK-  38
1TI5_A    1 RTCMIKKEGWGKCLIDTTCAHSCKNRGYIGGNCKGMTRTCYCLVNC  46
                *:     **  .  * : **: *   *:*    * *:*  :
```

```
>lcl|PDB:1TI5_A mol:protein length:46 plant defensin Length=46

Score = 25.8 bits (74), Expect = 1.9
Identities = 12/31 (39%), Positives = 14/31 (45%), Gaps = 2/31 (6%)

Query  7 KCSGSPQCLKPCKDAGMRFGKC--MNRKCHC 35  калиотоксин
         KC     C   CK+ G   G C  M R C+C
Sbjct 12 KCLIDTTCAHSCKNRGYIGGNCKGMTRTCYC 42  дефензин
```

Множественное выравнивание последовательностей

Что полезного?

- Итеративное выявление удаленной гомологии
- Выявление и консервативных остатков и мотивов
- Построение филогенетических деревьев
- …

Алгоритмы:

- Динамическое программирование
- Прогрессивное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?! (2017)

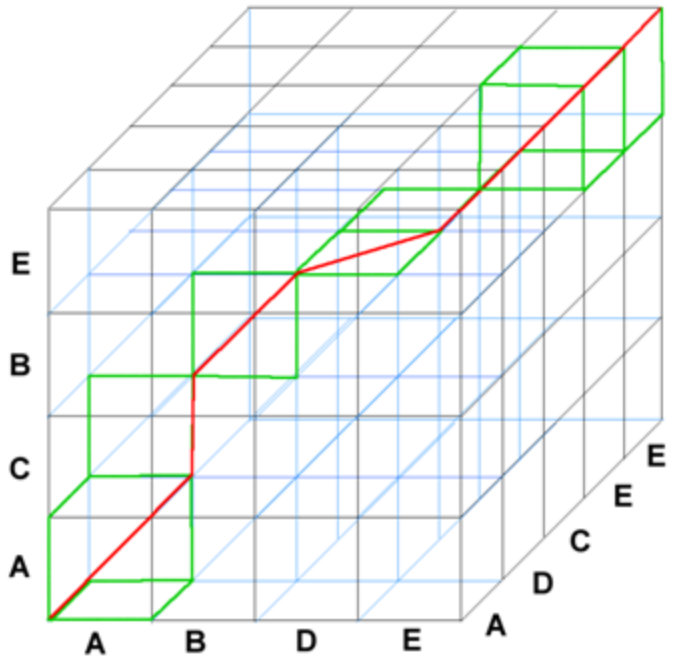
Визуализация – построение профилей

Динамическое программирование

- Прямой метод выполнения множественного выравнивания, обеспечивающий нахождение глобального оптимума.

- Для выравнивания *N* последовательностей требуется построение *N*-мерной матрицы. Таким образом, пространство поиска растет экспоненциально с ростом *N* и также зависит от длины последовательностей, а время поиска может быть оценено как O(L^N).

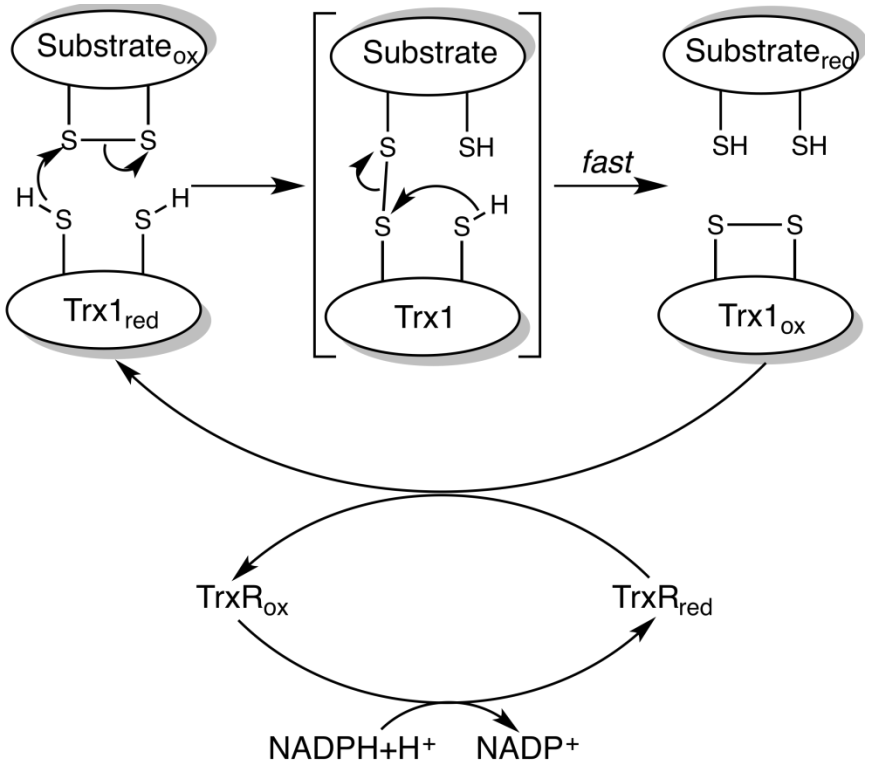
```
A-BD-E-
ACB--E-
A--DCEE
```



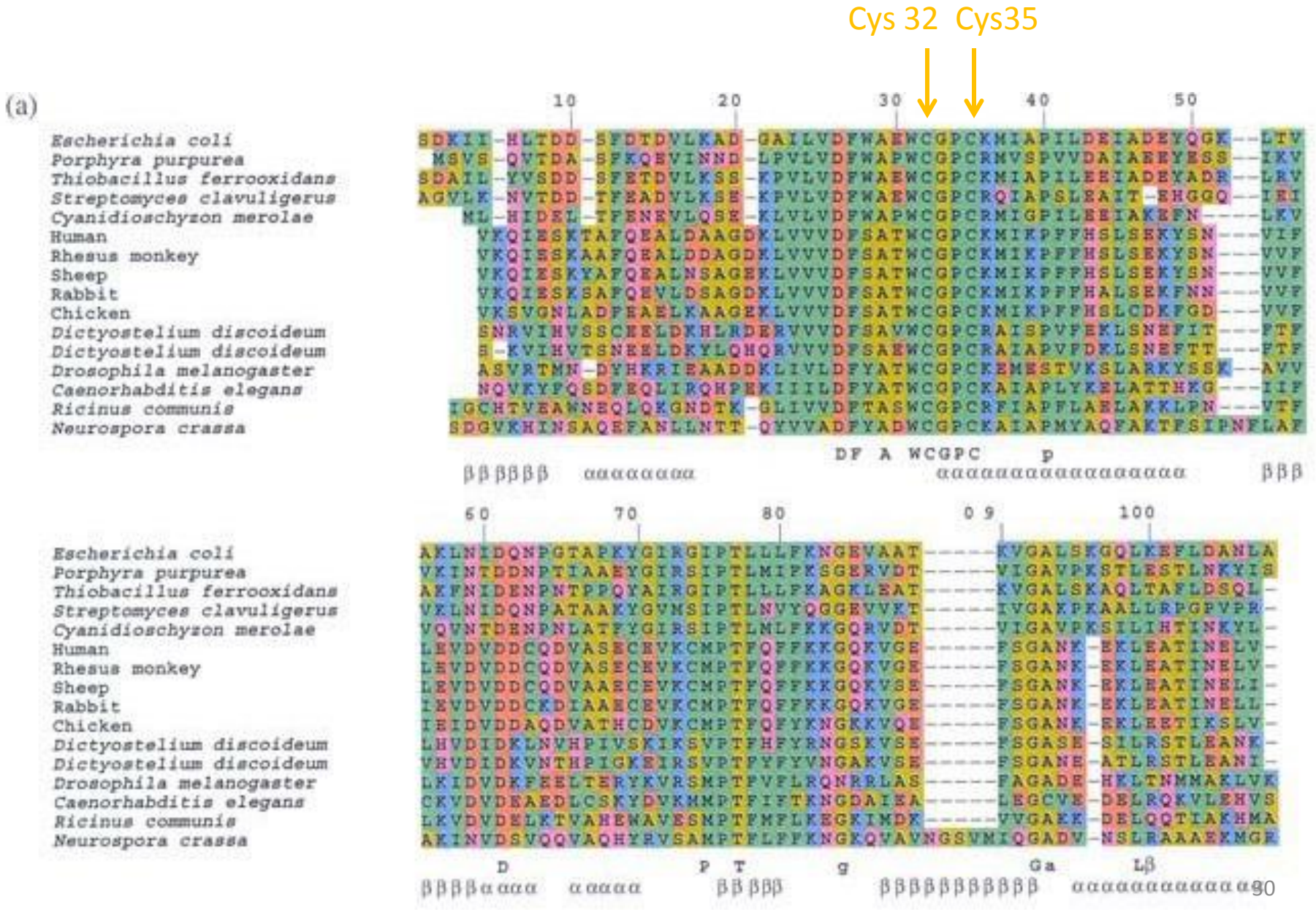
Построение и визуализация профилей

Тиоредоксины – семейство белков, отвечающих за восстановление дисульфидных связей в белках и встречающихся как в животном, так и в растительном мире.

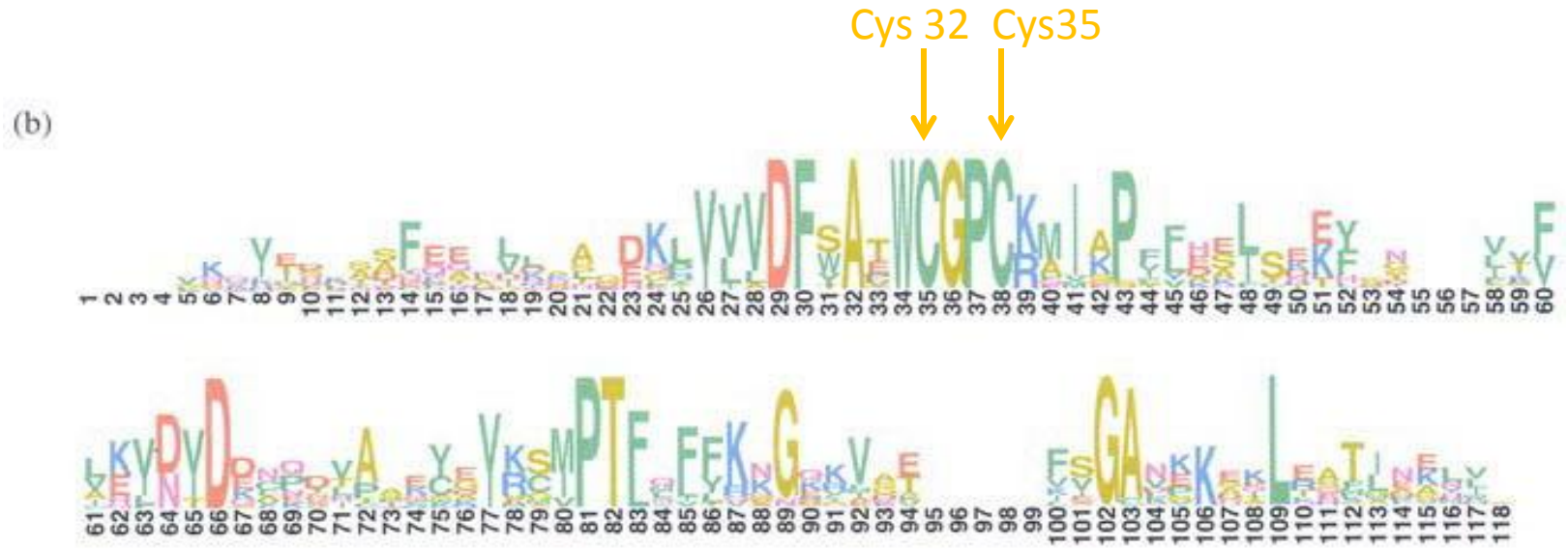
Выравнивание структур тиоредоксина человека и мушки *Drosophila melanogaster*.



Построение и визуализация профилей



Построение и визуализация профилей



s – число символов в алфавите (4, 20, …)
p(a, i) – вероятность появления буквы *a* в позиции *i*

Sequence logos: a new way to display consensus sequences (Schneider & Stephens, 1990)

$$H_i = -\sum_{\text{по всему алфавиту}} p(a,i)\log_2 p(a,i)$$

Шенноновская энтропия i-той позиции

$$R_i = \log_2 s - H_i$$

Информационная значимость i-той позиции

$$h(a,i) = p(a,i)R_i$$

высота символа в профиле