ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №7

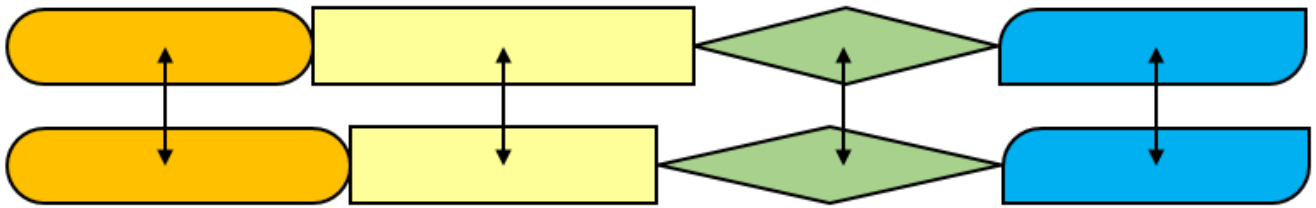
Значимость выравнивания. Экспресс-сравнение последовательностей (BLAST). Построение и визуализация профилей множественных выравниваний.

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

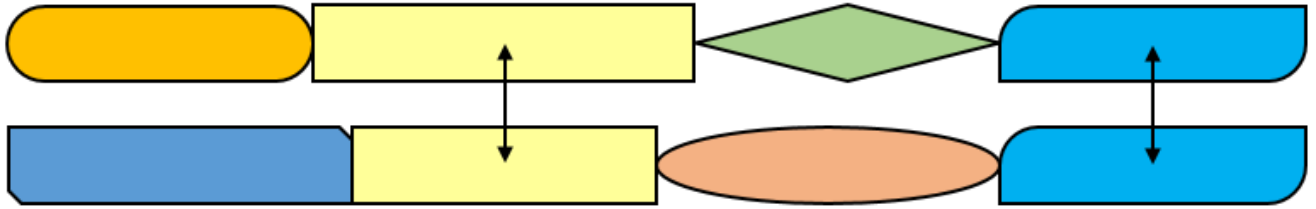
Сайт курса http://intbio.org/bioinf2018

Расчет выравнивания двух последовательностей

- Алгоритм Нидлмана-Вунша (1970) для глобального выравнивания
- Алгоритм Смита-Уотермана (1981) для локального выравнивания



Global Alignment



Local Alignment

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

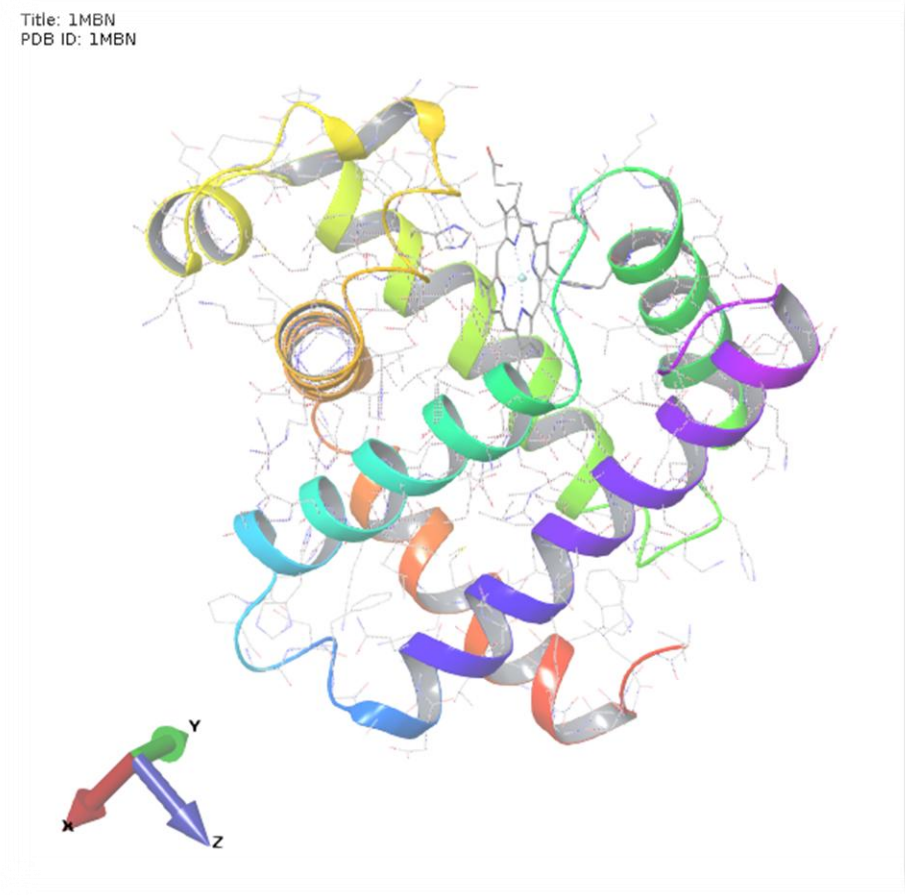
Кашалот

Люпин

Последний общий предок жил около 1 млрд лет назад!

Что общего между ними сейчас?

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?

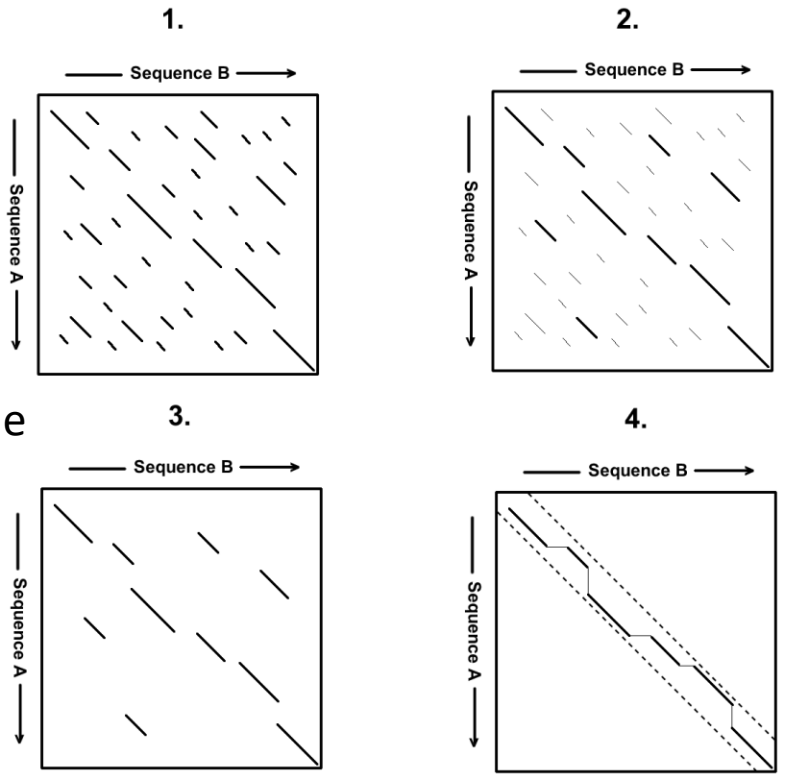
https://impossiblefoods.com/

Экспресс-методы сравнения последовательностей. FASTA

FASTA (Lipman and Pearson, 1985, Pearson and Lipman, 1988) предназначен для сравнения новых последовательностей с уже содержащимися в базах данных.

Алгоритм:

- Поиск идентичных *к*-буквенных слов (k = 2 для белков, 4-6 для нуклеотидов, 1 для коротких последовательностей) между последовательностями;
- Расширение точек идентичности до областей сходства с использованием матриц замен (**BLOSUM50** для белков, **единичная** для нуклеотидов);
- Переоценка выявленных областей сходства с помощью целевой матрицы замен, определение 10 областей с максимальной оценкой;
- Перебор возможных вариантов построения выравнивания с использованием этих областей и выбор оптимального с учётом оценок областей и штрафов за вставки;



FASTA и FASTQ форматы

Исходно – формат представления последовательностей для программы FASTA.

```
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens…
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLR…
```

FASTQ - обобщение формата FASTA.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Качество прочтения ***Q(sanger)* = - 10 log(10)(*e*)**,
где *e* – вероятность **ошибочного** прочтения

94 градации в формате **ASCII**:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOP
        QRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

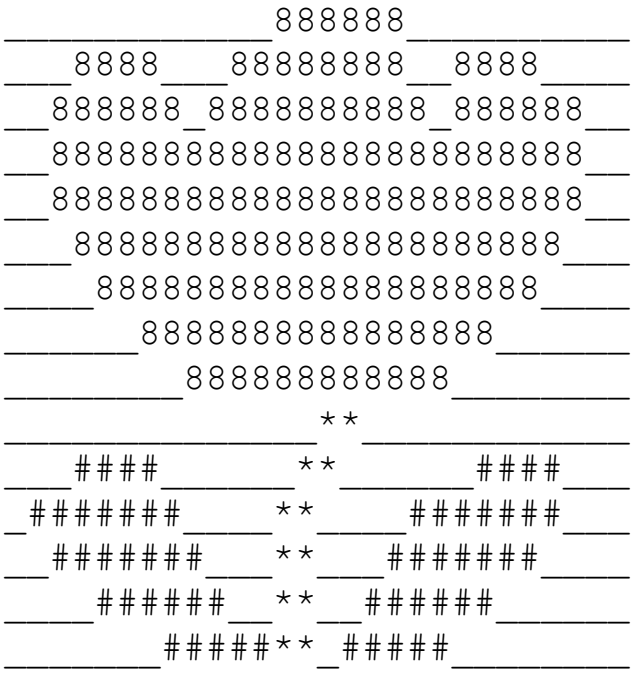
American **S**tandard **C**ode for **I**nformation **I**nterchange (1967 - 2007)

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

128 символов (256 = 2^8 в расширенной версии), так что для кодирования необходим **1 байт на 1 символ** (команду или знак).

Кодировка ASCII заменена на Unicode (в настоящее время содержит 137 439 символов, охватывающих 146 современных и исторических шрифтов, математические символы, эмотиконы и т.д.)



ASCII art

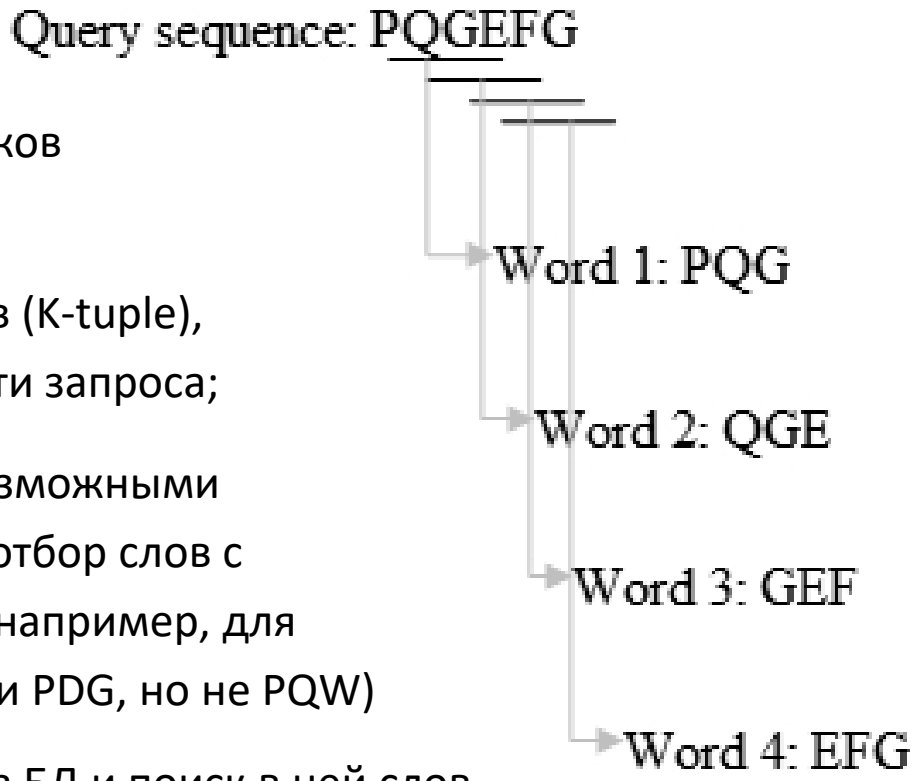
```
**********************************************
***                                        ***
***              FoldX 4 (c)               ***
***                                        ***
***      code by the FoldX Consortium      ***
***                                        ***
***      Jesper Borg, Frederic Rousseau    ***
***     Joost Schymkowitz, Luis Serrano    ***
***     Peter Vanhee, Erik Verschueren     ***
***      Lies Baeten, Javier Delgado       ***
***        and Francois Stricher           ***
*** and any other of the 9! permutations   ***
***   based on an original concept by      ***
***   Raphael Guerois and Luis Serrano     ***
**********************************************
```

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.

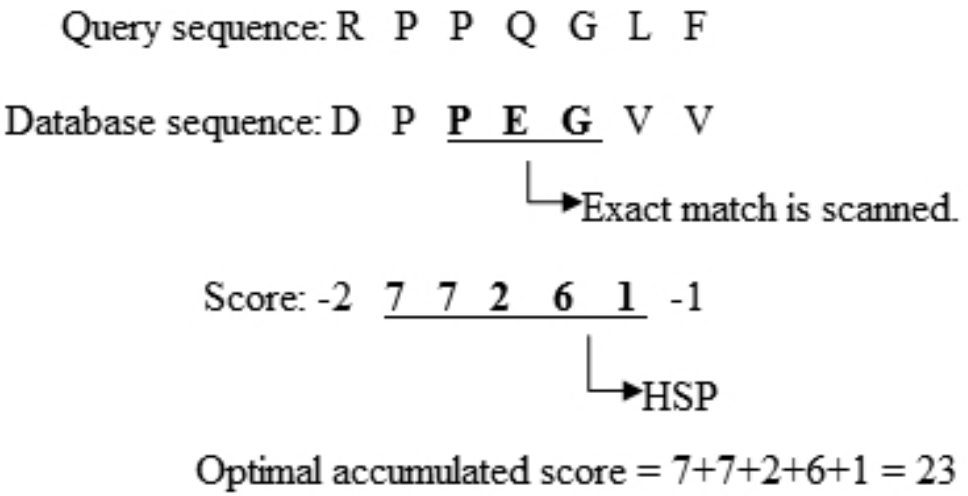
Алгоритм:

- 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;

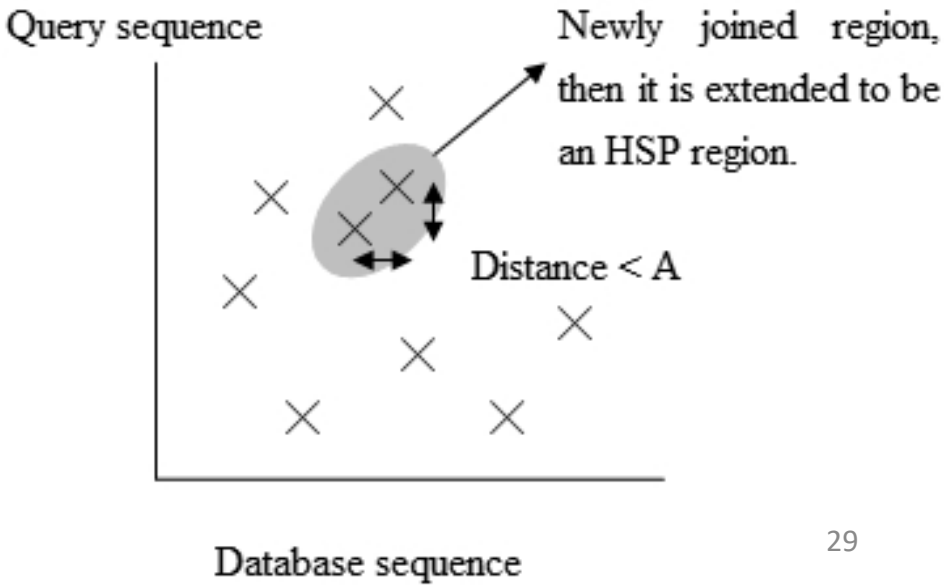


Экспресс-методы сравнения последовательностей. BLAST

- 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)(?)

Альфред Реньи
(1921 – 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) ≈ 0,577215…

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln n\right) = \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+\ldots+\frac{1}{n}-\ln n\right)$$

(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, Altschul, 1990?) и для максимальной оценки S_n,m сегментной пары (**HSP**), образованной последовательностями длиной n и m:

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

Нормализация

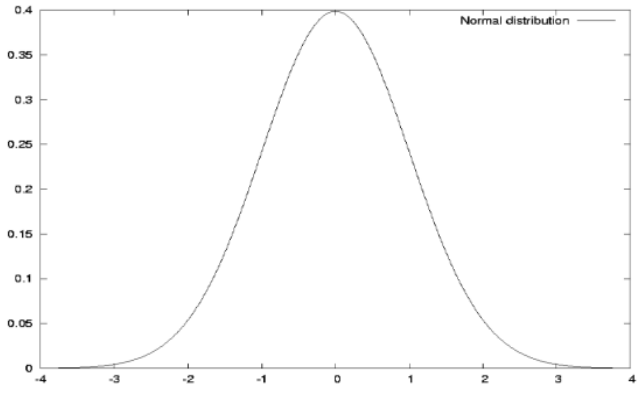
$$\tilde{S}_{n,m}=S_{n,m}-\frac{\ln(Knm)}{\lambda}$$

BLAST. Значимость выравнивания

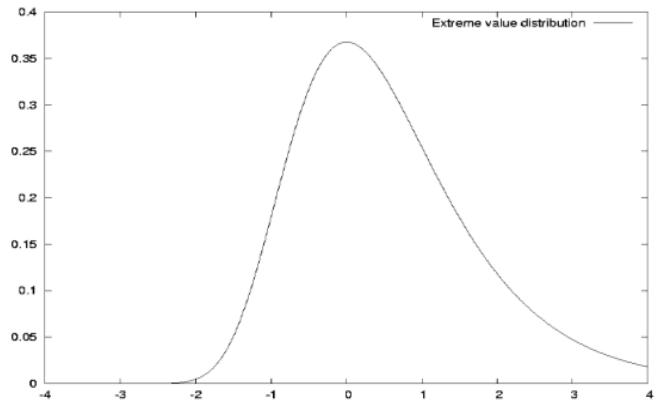
Распределение нормализованных максимальных оценок $\check{S}\_n,m$ подчиняется распределению Гумбеля (распределению экстремальных значений, Gumbel, 1937; Гнеденко, 1943), для которого

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

где S – случайная величина в интервале (-∞; +∞) (но последнее приближение действительно только для больших *S*, когда *Kmn*exp(- λS)* мало!)



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



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

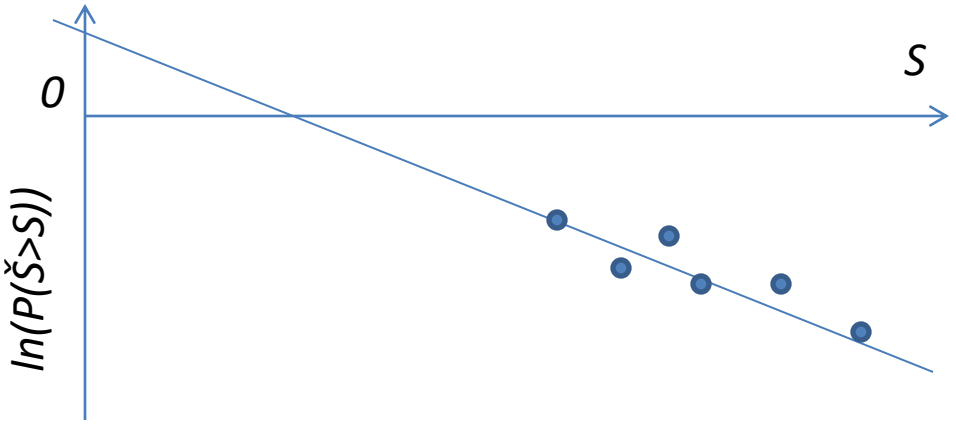
BLAST. Значимость выравнивания

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

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(Š>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

Для BLOSUM62 были получены значения $K = 0.040$ и $\lambda = 0.254$ (Altshul, Gish, 1996)(?).

Подставляем их в формулу для расчета вероятности:

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

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

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

Пример (миоглобин кашалота):

```
>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

Для BLOSUM62 были получены значения $K = 0.040$ и $\lambda = 0.254$ (Altshul, Gish, 1996)(?).

Подставляем их в формулу для расчета вероятности:

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

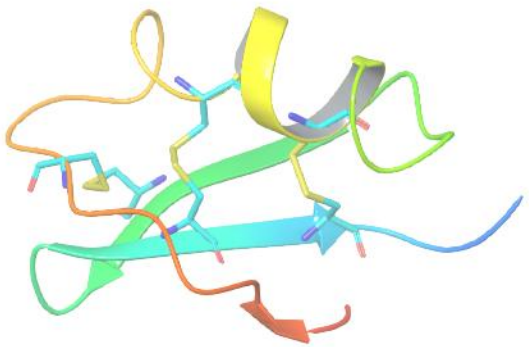
$$P(\tilde{S}_{n,m} > S) \equiv E - value$$

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

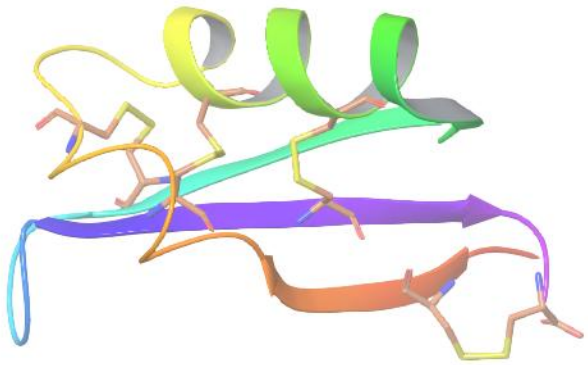
Другой пример (калиотоксин):

```
>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
```

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

Множественное выравнивание последовательностей

Множественное выравнивание последовательностей

Цели:

- Построение филогенетических деревьев
- Выявление консервативных остатков и мотивов
- Построение профилей (визуализация)
- Итеративное выявление удаленной гомологии
- …

Алгоритмы:

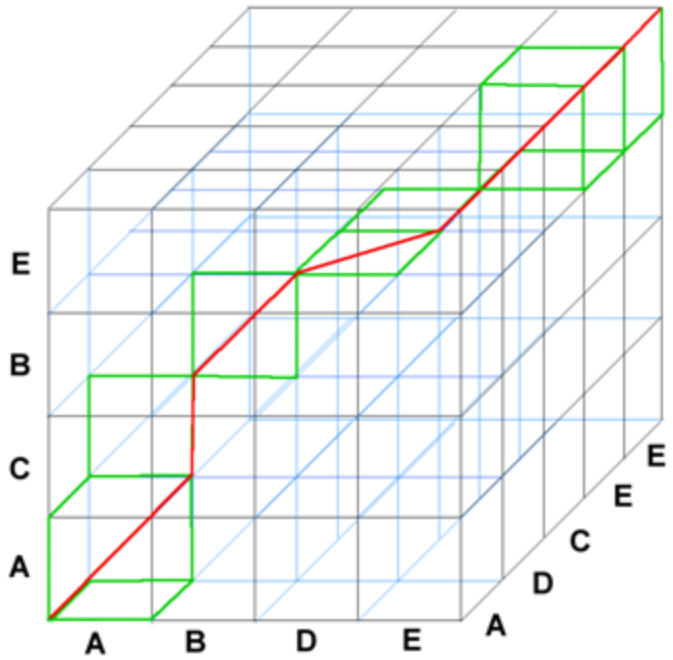
- **Динамическое программирование**
- Прогрессивное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?

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*, а также зависит от длины последовательностей. Время поиска может быть оценено как **O(L^N)**.

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



Множественное выравнивание последовательностей

Цели:

- Построение филогенетических деревьев
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- Итеративное выявление удаленной гомологии
- …

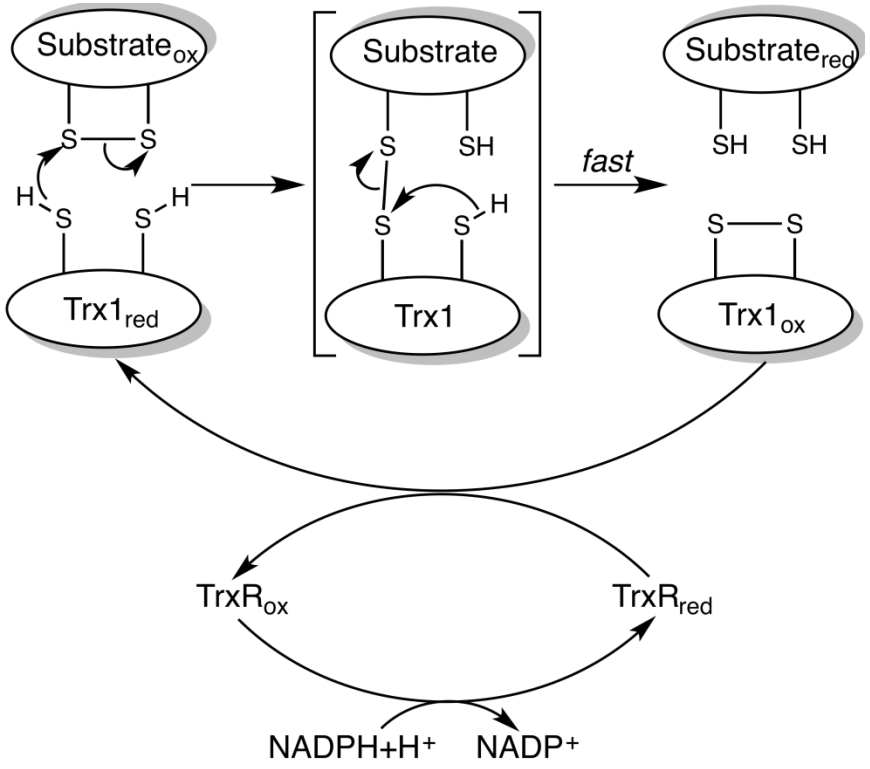
Алгоритмы:

- Динамическое программирование – не годится
- Прогрессивное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?

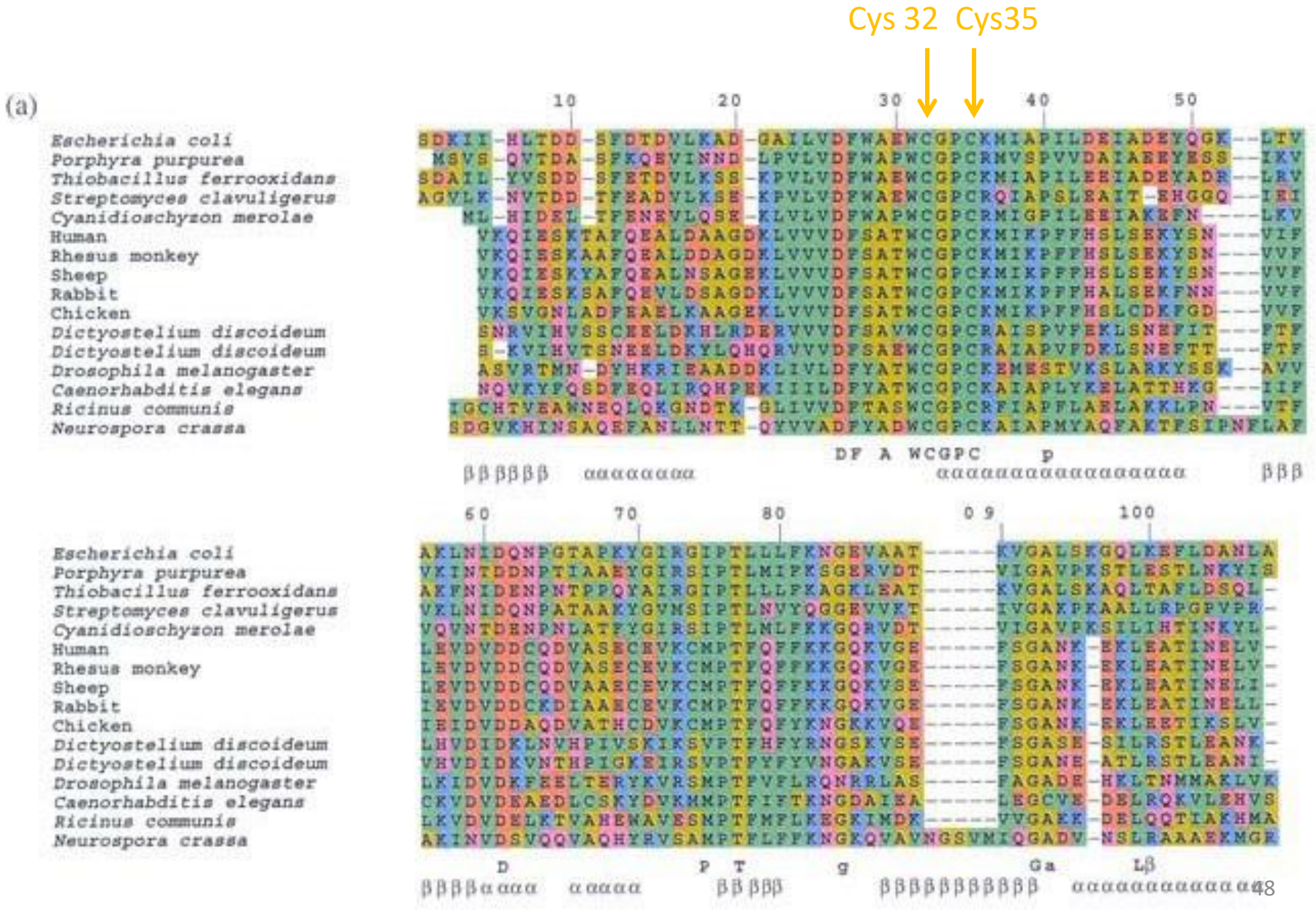
Построение и визуализация профилей

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

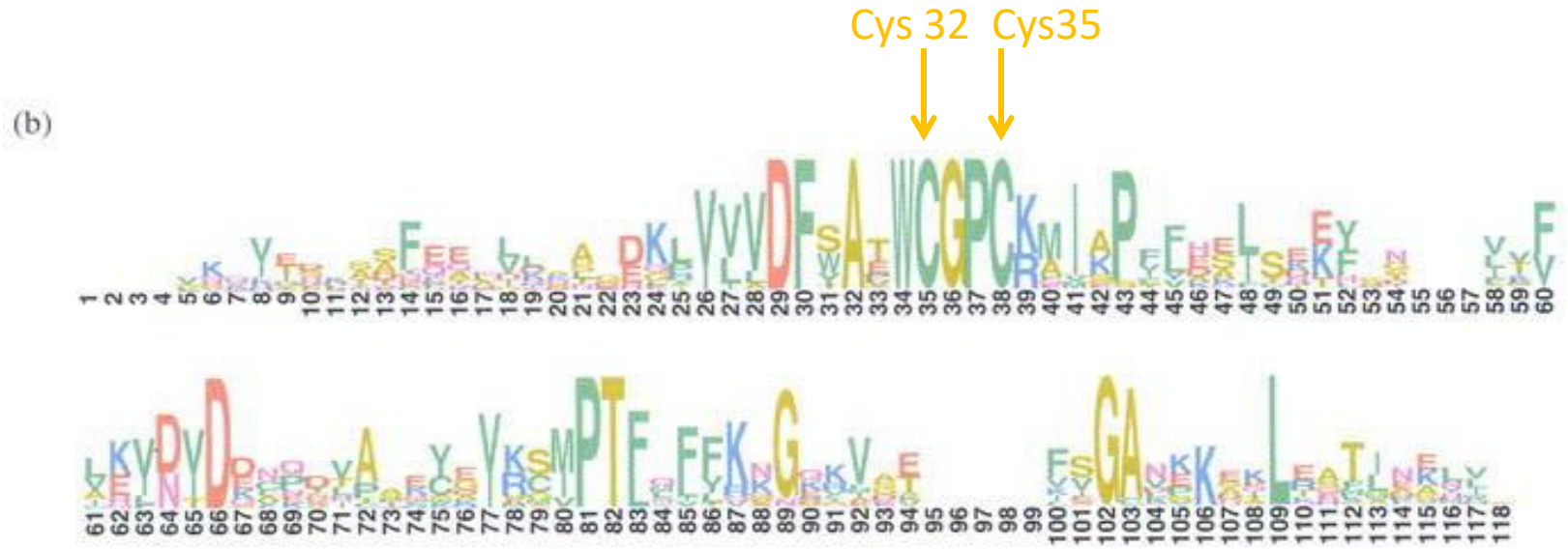
Выравнивание структур тиоредоксина человека и мушки *Drosophila melanogaster*.



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

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

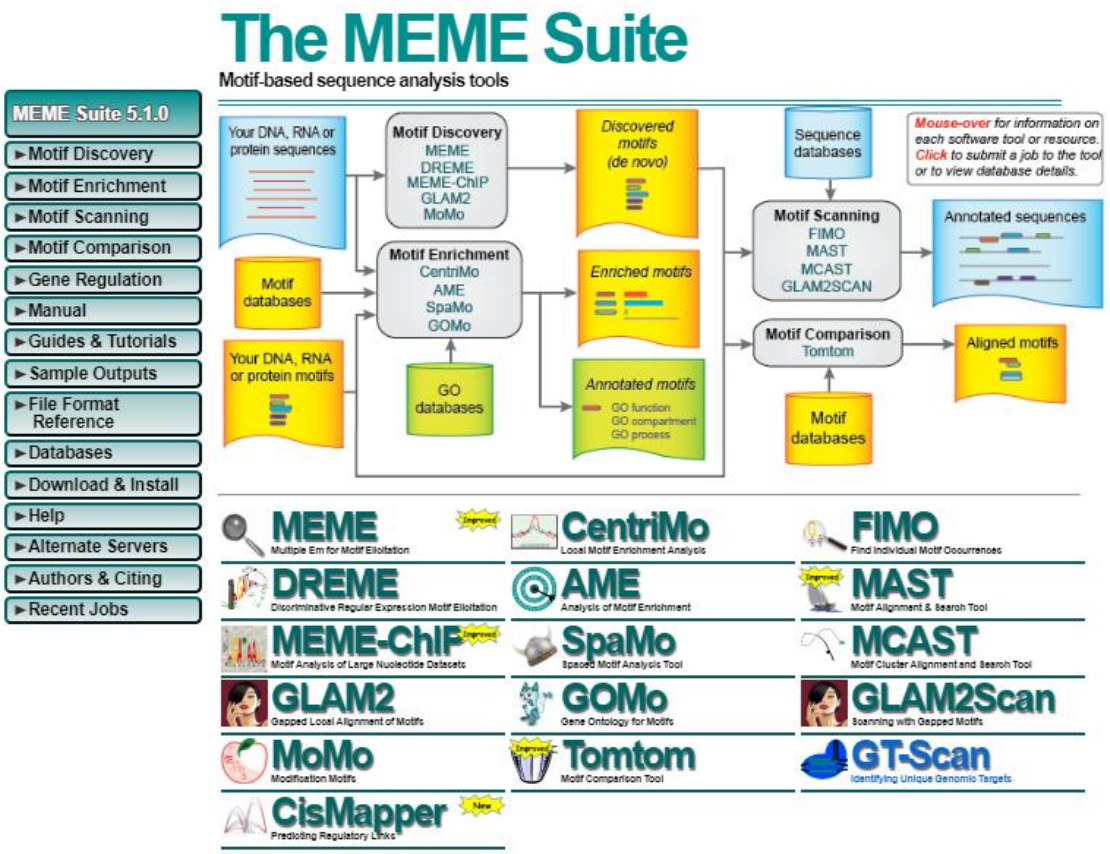
Высота символа в профиле

Консенсусная последовательность и мотивы

Консенсусная последовательность – последовательность, в каждой позиции которой находится символ, наиболее часто встречающийся в данной позиции выравнивания.

Мотив – набор символов, (предположительно) имеющий биологическое значение.

N{P}[ST]{P}



Множественное выравнивание последовательностей

Цели:

- Построение филогенетических деревьев
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

Алгоритмы:

- Динамическое программирование – не годится
- Прогрессивное выравнивание
- Скрытые марковские модели
- Квантовые компьютеры?

Position-Specific Iterative BLAST

Position-Specific Iterative BLAST (**PSI-BLAST**) (Altschul et al. 1997)

http://www.ebi.ac.uk/Tools/sss/psiblast/

Алгоритм:

- Просмотр БД и составление списка последовательностей, сходных с последовательностью запроса;
- Расчет множественного выравнивания и набора позиционно-специфичных матриц замен (PSSM) для него;
- Новый просмотр БД с использованием полученных PSSM вместо стандартных матриц;
- Повторение процедуры либо желаемое число раз, либо до достижения сходимости (состояния, когда новые последовательности не обнаруживаются).

Давид Липман

EN = 3

PSI-BLAST эффективен при поиске удалённых гомологов.

Множественное выравнивание последовательностей

Цели:

- **Построение филогенетических деревьев**
- **Выявление консервативных остатков и мотивов**
- **Построение профилей (визуализация)**
- **Итеративное выявление удаленной гомологии**
- …

Алгоритмы:

- Динамическое программирование – не годится
- **Прогрессивное выравнивание**
- **Скрытые марковские модели**
- Квантовые компьютеры?