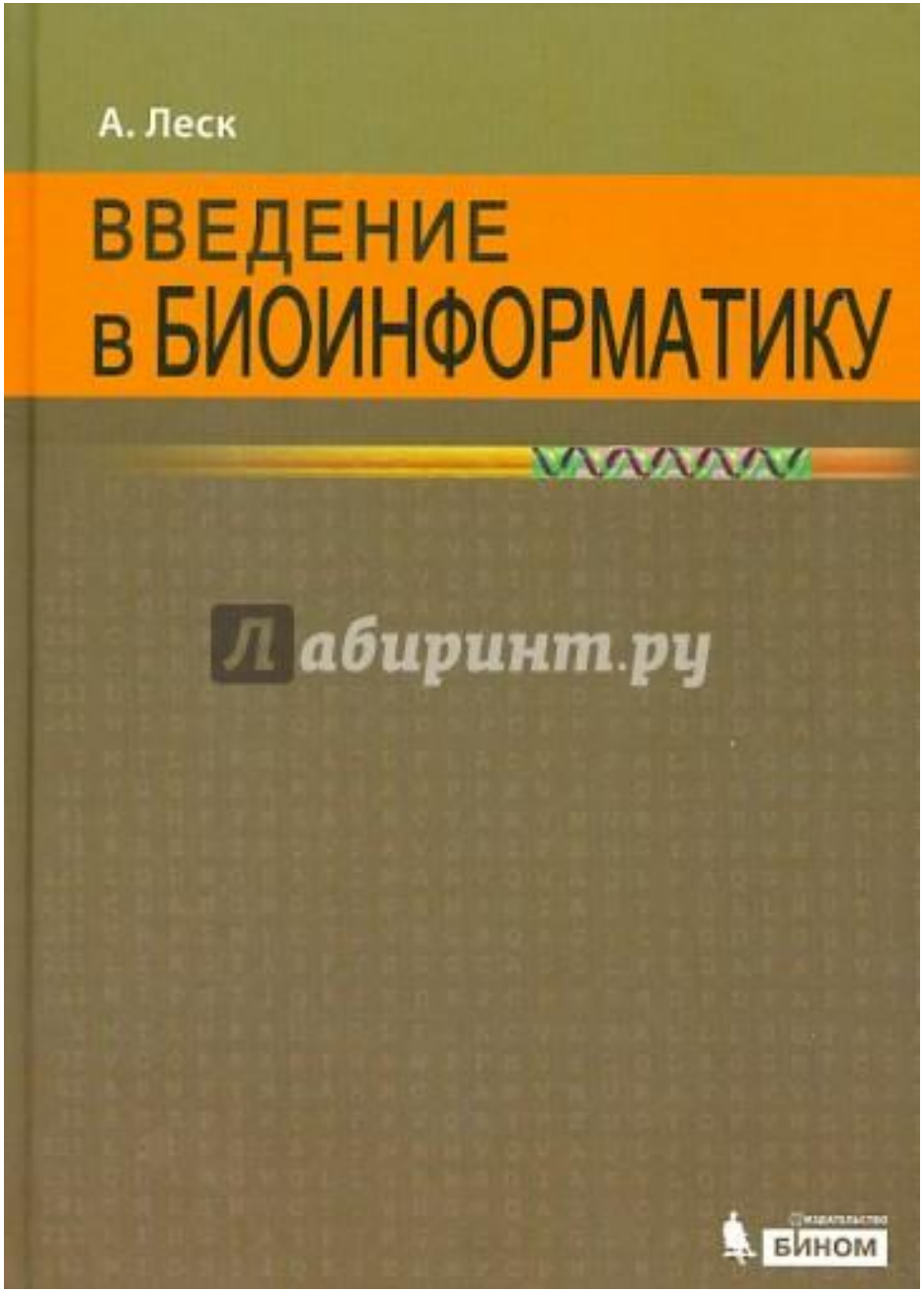
ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №5

Сравнение последовательностей: матрицы сходства, динамическое программирование, локальное и глобальное выравнивание. Матрицы замен.

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

Сайт курса https://intbio.org/bioinf2020-2021/



Артур Леск

EN = 5

Зарождение биоинформатики

Маргарет Дэйхоф
(1925 – 1983)

Инициировала создание Atlas of Protein Sequence and Structure (1965) (65 последовательностей) -> PIR -> Uniprot

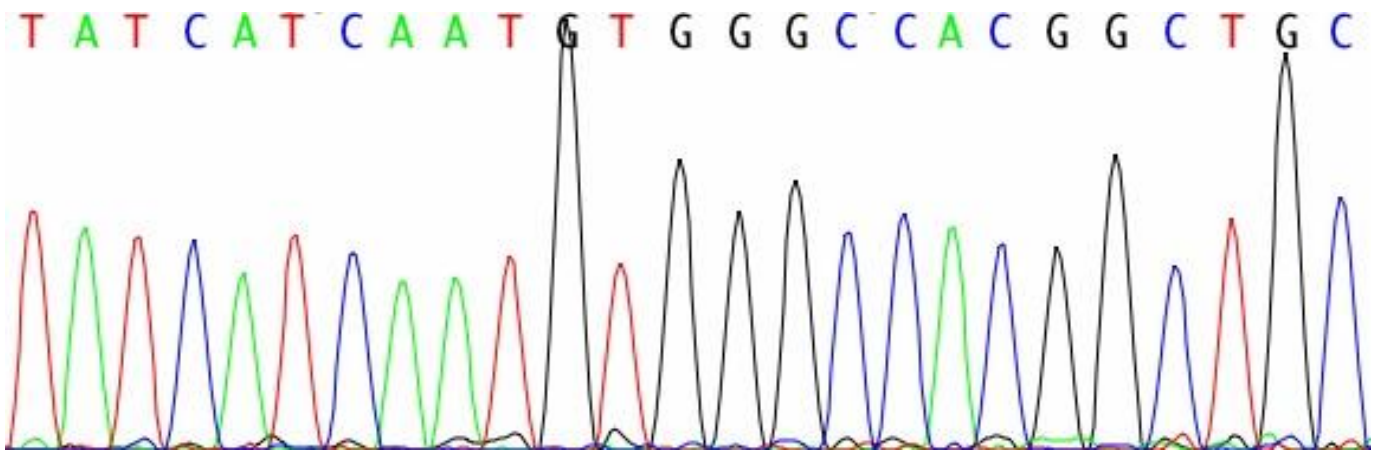
Первой начала использование компьютеров для сравнения последовательностей

Предложила однобуквенный код для аминокислот

Впервые провела реконструкцию эволюционного дерева исходя из выравнивания последовательностей (1966)

David Lipman, (ex)director of the NCBI: “She was the mother and father of bioinformatics".

Мы только что секвенировали новую последовательность ДНК



Что мы можем о ней узнать?

- Относится ли она к уже известному гену?
- Как она соотносится с другими известными генами: имеют ли они общего предшественника или сходство функций получено путем конвергентной эволюции?
- Какая белковая последовательность соответствует этой последовательности ДНК и что это может быть за белок?
- …

Сравнение последовательностей

Цели:

- Соизмерить сходство последовательностей и установить соответствие между нуклеотидами или аминокислотными остатками
- Выявить консервативные и вариабельные области
- Предположить эволюционные взаимосвязи
- …

Выравнивание последовательностей – основной инструмент биоинформатики

Сравнение последовательностей

Рассмотрим последовательности `gctgaacg` и `ctataatc`.

Их можно выровнять:

неинформативно

```
- - - - - - - g c t g a a c g
c t a t a a t c - - - - - - - -
```

без пропусков

```
g c t g a a c g
c t a t a a t c
```

с пропусками так

```
g c t g a - a - - c g
- - c t - a t a a t c
```

или так

```
g c t g - a a - c g
- c t a t a a t c -
```

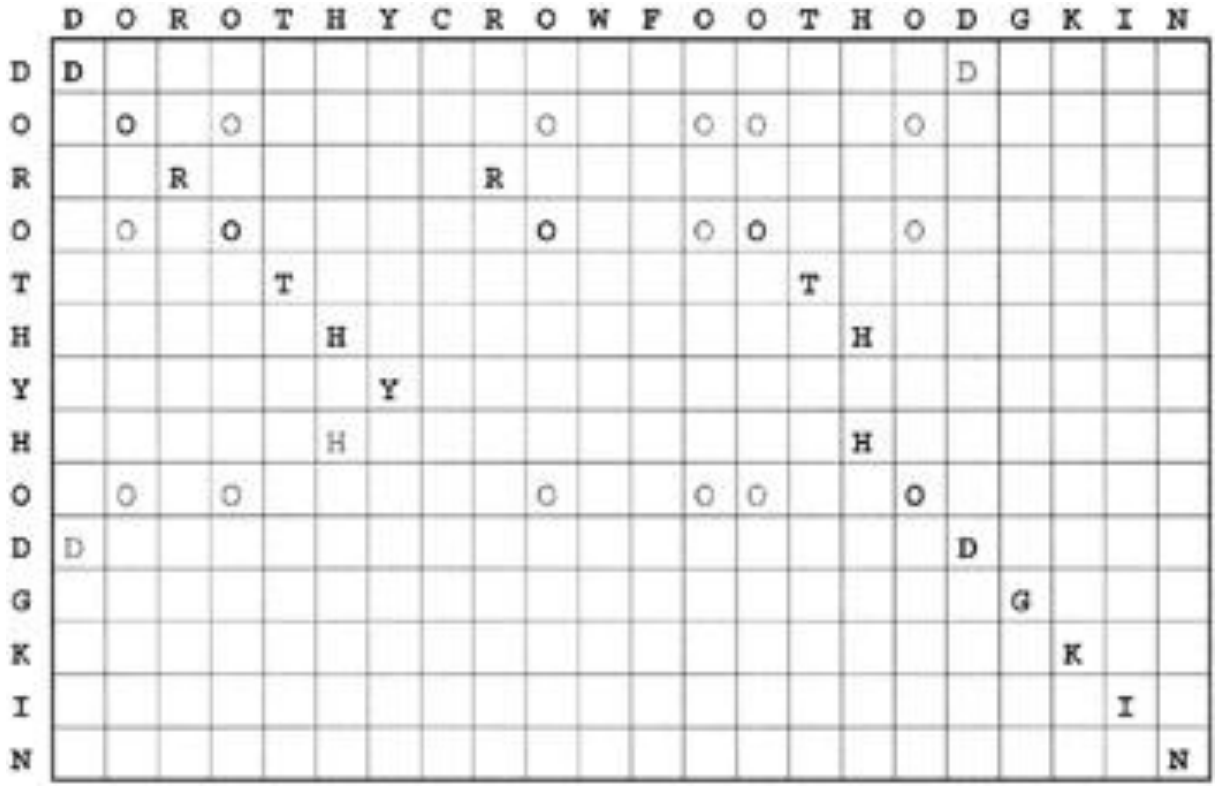
Сколько существует выравниваний для последовательностей длины *m* и *n*?

Точечная матрица сходства (dotplot) (1970)

Снова две последовательности:

- DOROTHYHODGKIN
- DOROTHYCROWFOOTHODGKIN

Матрица сходства для них будет выглядеть так:



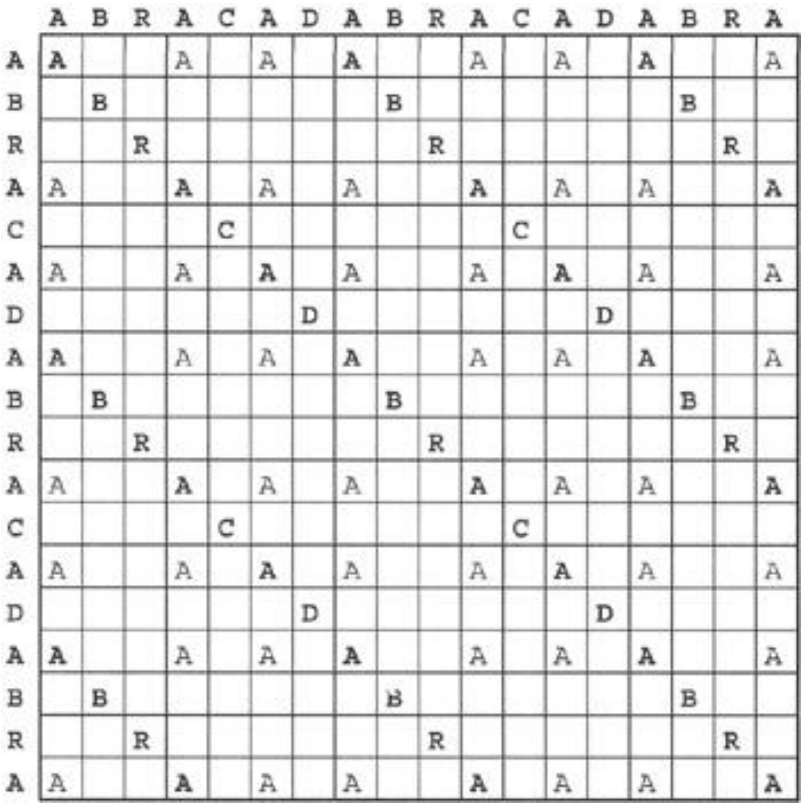
(1910 - 1994)

Холестерин (1937)
Пенициллин (1945)
Витамин B12 (1954)
Инсулин (1969)

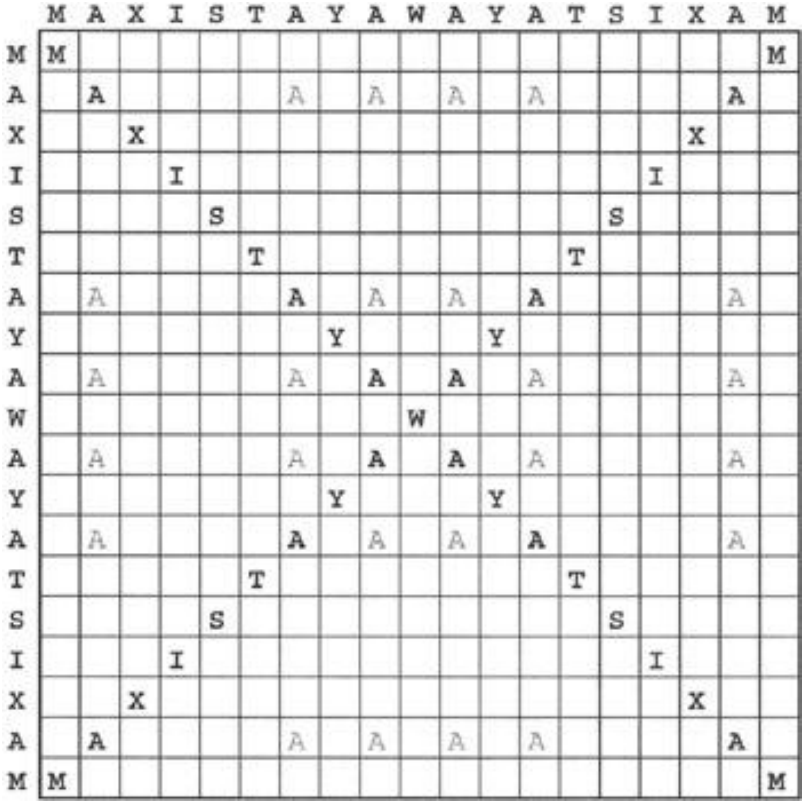
Нобелевский лауреат 1964 года

Точечная матрица сходства (dotplot) (1970)

Последовательность с повторами (ABRACADABRACADABRA):



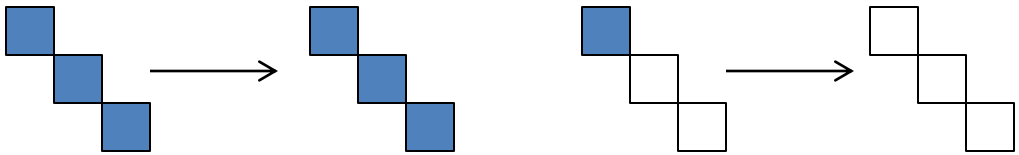
Палиндромная последовательность (MAX I STAY AWAY AT SIX AM):

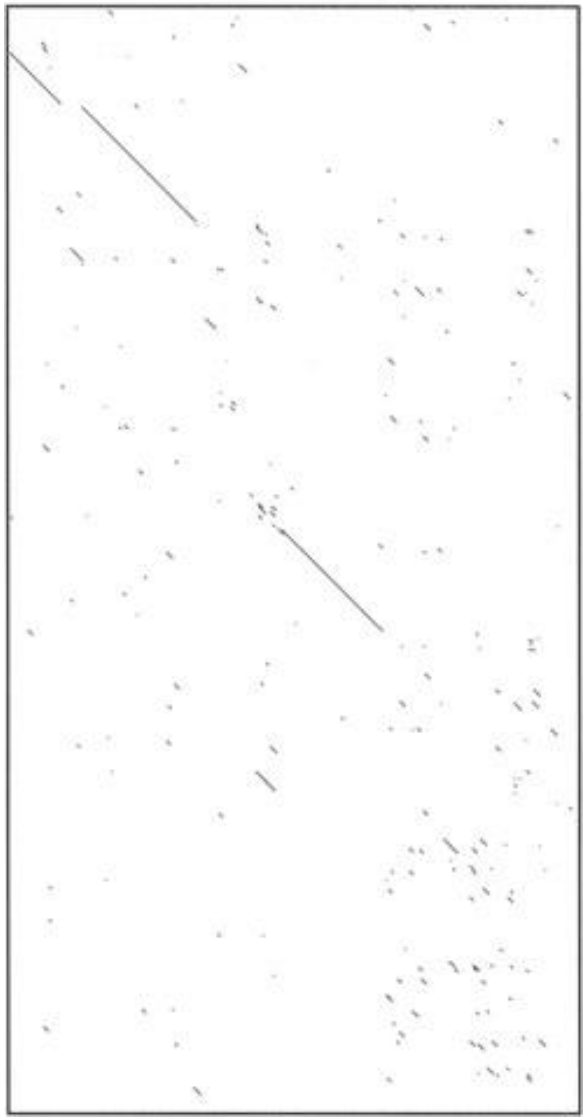


Точечная матрица сходства (dotplot) (1970)

Для реальных последовательностей (белок PAX-6 мыши (Uniprot P63015) и белок eyeless плодовой мушки (O18381)) матрица сходства имеет более сложный вид. Хорошо заметны три протяженных участка сходства.

Для сокращения шумов применяется фильтрование результатов. Например, точки не показываются до тех пор, пока в окне заданной ширины не окажется заданное число совпадений.

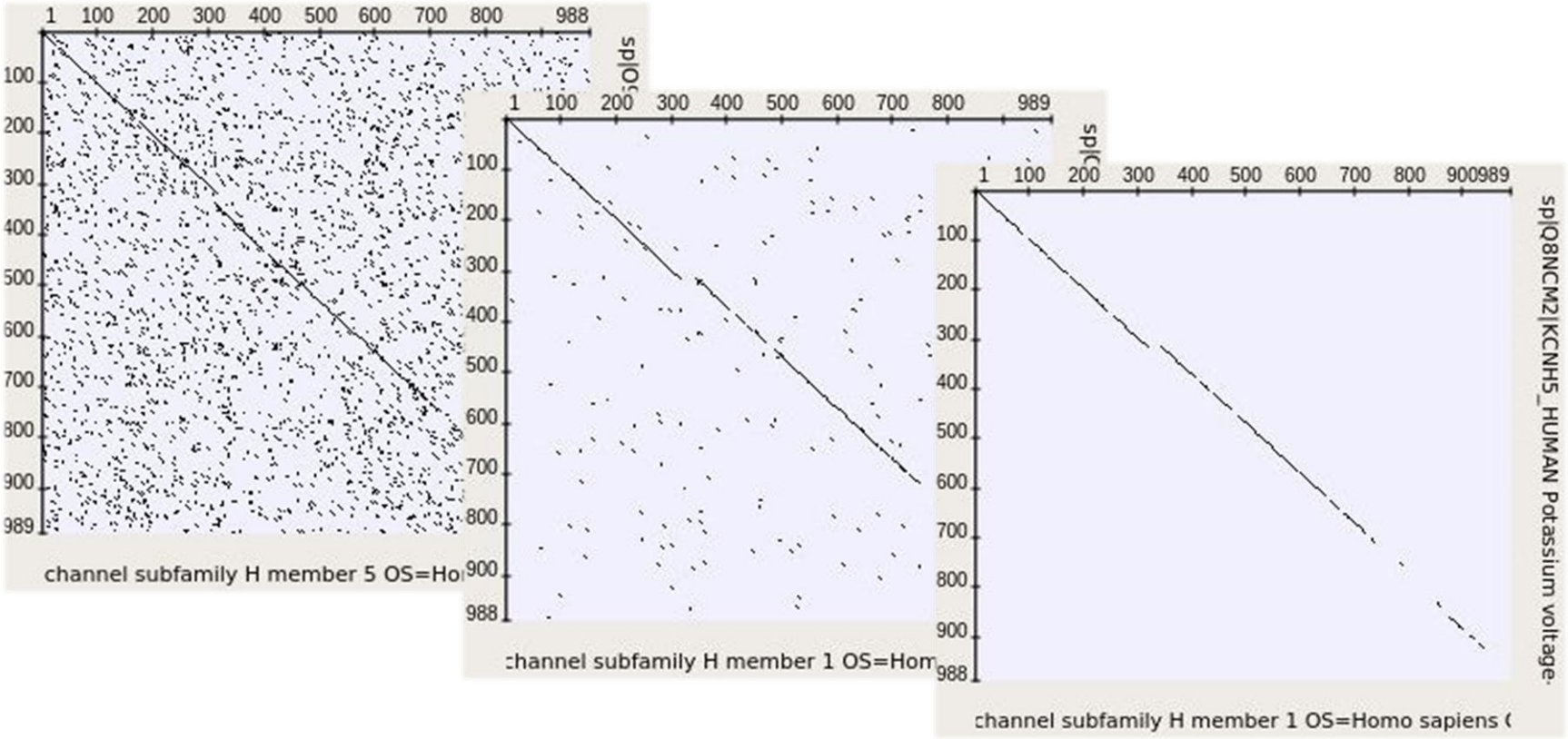




Интерактивное построение точечных матриц:

http://www.bioinformatics.nl/emboss-explorer/ -> dotpath

Точечная матрица сходства (dotplot) (1970)



Матрицы сходства последовательностей каналов KCNH1 и KCNH5 человека при длине отображаемых повторов 2, 3 и 4.

Free bioinformatics software:

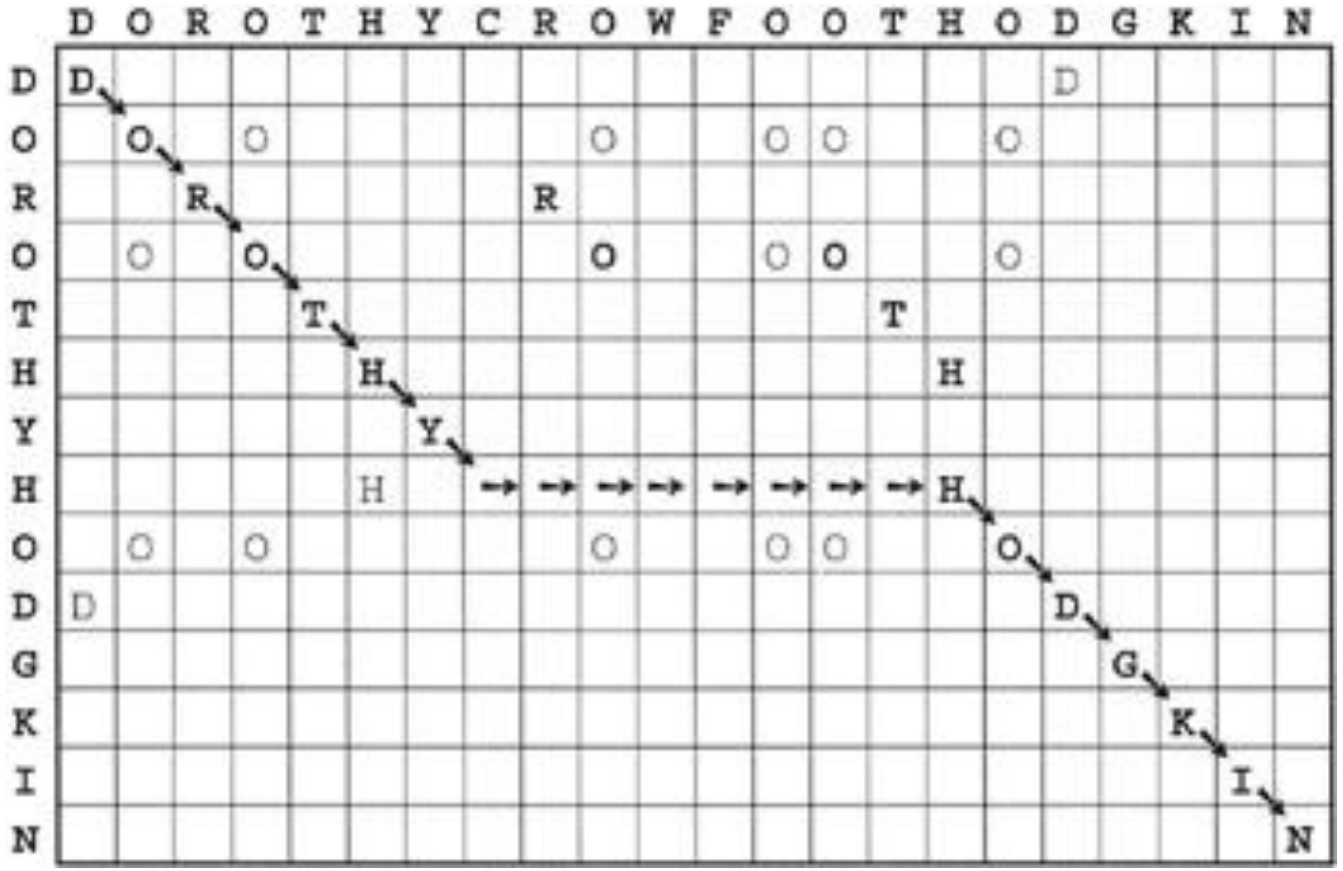
http://ugene.net

Unipro UGENE

Точечные матрицы и выравнивание последовательностей

Любой путь по точечной матрице из левого верхнего угла в правый нижний **дает возможное выравнивание:**

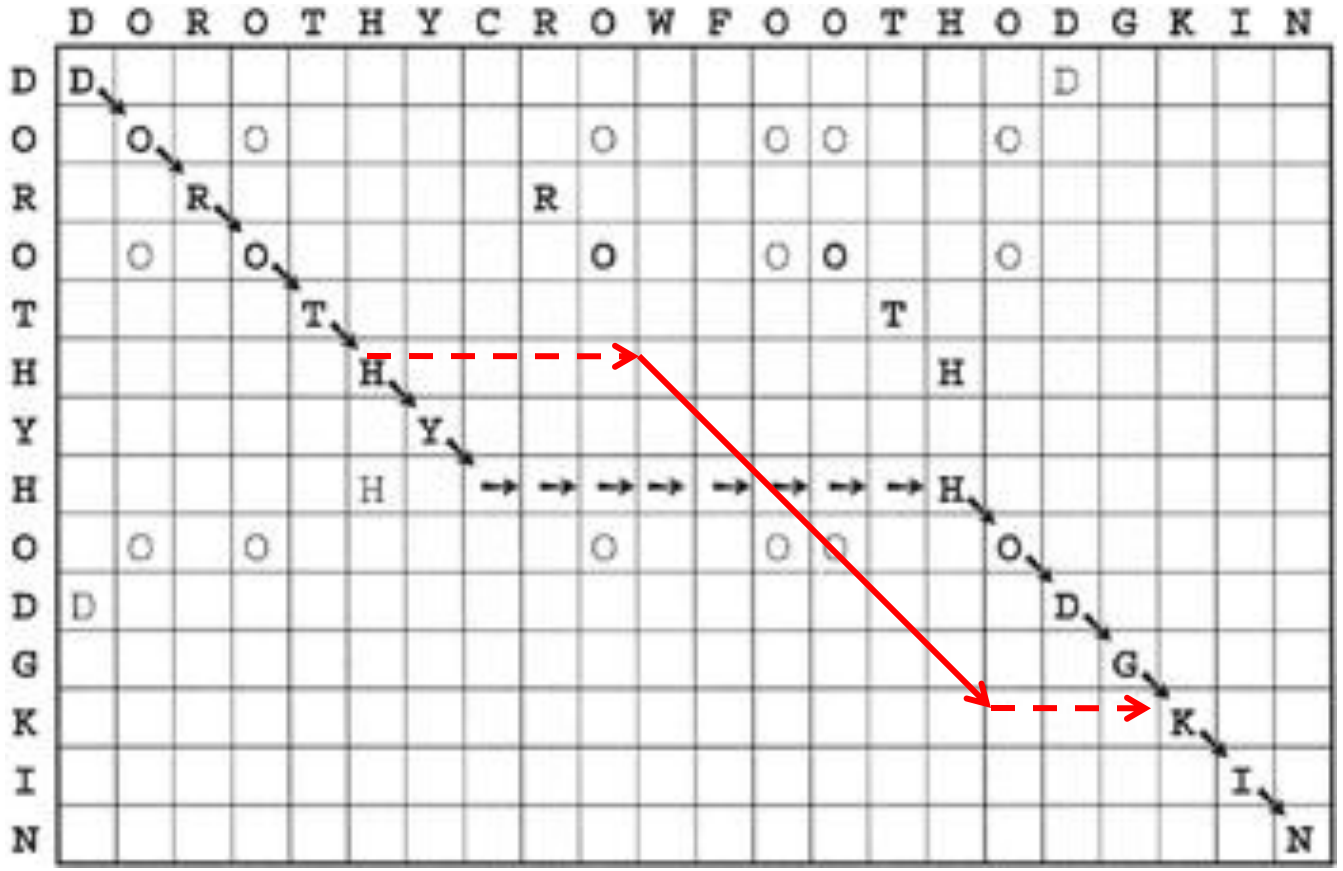
```
DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN
```



Точечные матрицы и выравнивание последовательностей

Любой путь по точечной матрице из левого верхнего угла в правый нижний **дает возможное выравнивание:**

```
DORO-----THYHOD---GKIN
DOROTHYCROWFOOTHODGKIN
```



Как выбрать оптимальное выравнивание из всех возможных? **Нужна мера сходства!**

Мера сходства последовательностей

Расстояние по Хэммингу (1950) – количество несовпадающих позиций между последовательностями одинаковой длины;

Ричард Хэмминг
(1915 - 1998)

Лауреат премии Тьюринга 1968 года
EN = 4

Расстояние по Левенштейну («редакционное расстояние») (1965) – минимальное число операций редактирования, необходимых для превращения одной строки в другую (длины строк могут не совпадать).

Владимир Левенштейн
(1935 - 2017)

Мехмат МГУ (1958)

Медаль Хэмминга 2006 года
EN = 3

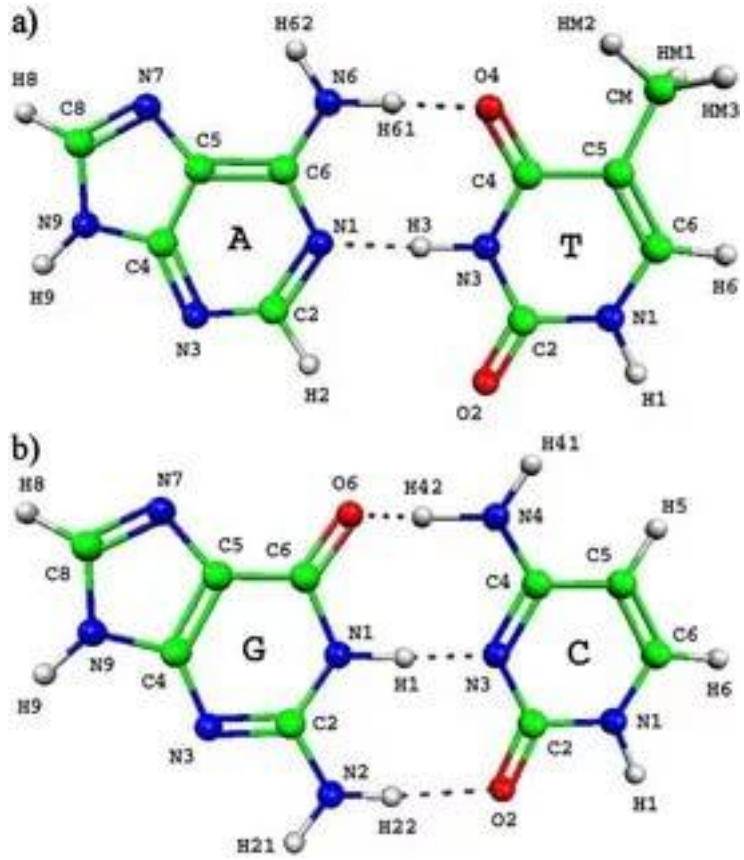
Мера сходства последовательностей

Выравнивание		Расстояние по Левенштейну
- - - - - - - g c t g a a c g c t a t a a t c - - - - - - - -	=	15
g c t g a a c g c t a t a a t c	=	6
g c t g a - a - - c g - - c t - a t a a t c	=	11
g c t g - a a - c g - c t a t a a t c -	=	5

Мера сходства последовательностей

Однако для более аккуратной оценки нужно иметь в виду, что:

- Некоторые замены происходят вероятнее других (особ. аминокислотные);
- Совместная делеция нескольких остатков более вероятна, чем независимая;
- ...



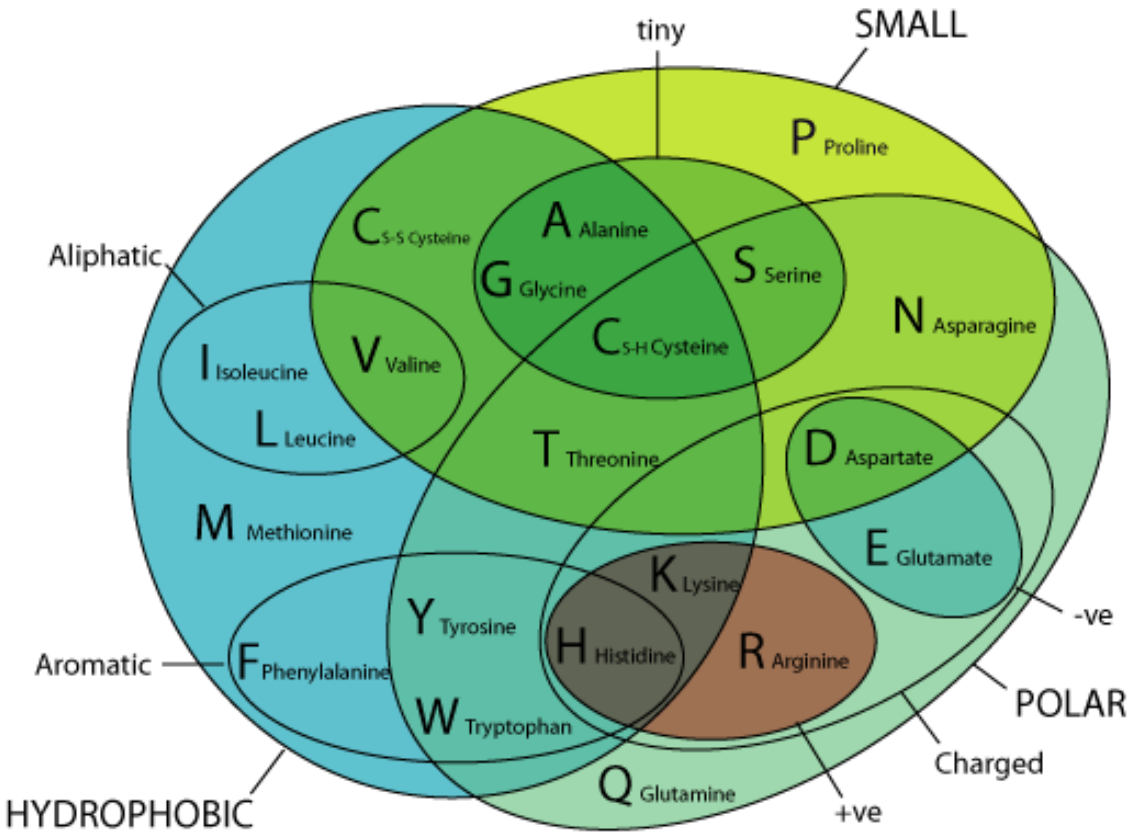
Гипотетическая **матрица** нуклеотидных **замен**, учитывающая, что транзиции встречаются чаще трансверсий:

	a	g	t	c
a	10	0	-5	-5
g	0	10	-5	-5
t	-5	-5	10	0
c	-5	-5	0	10

Составление матриц замен

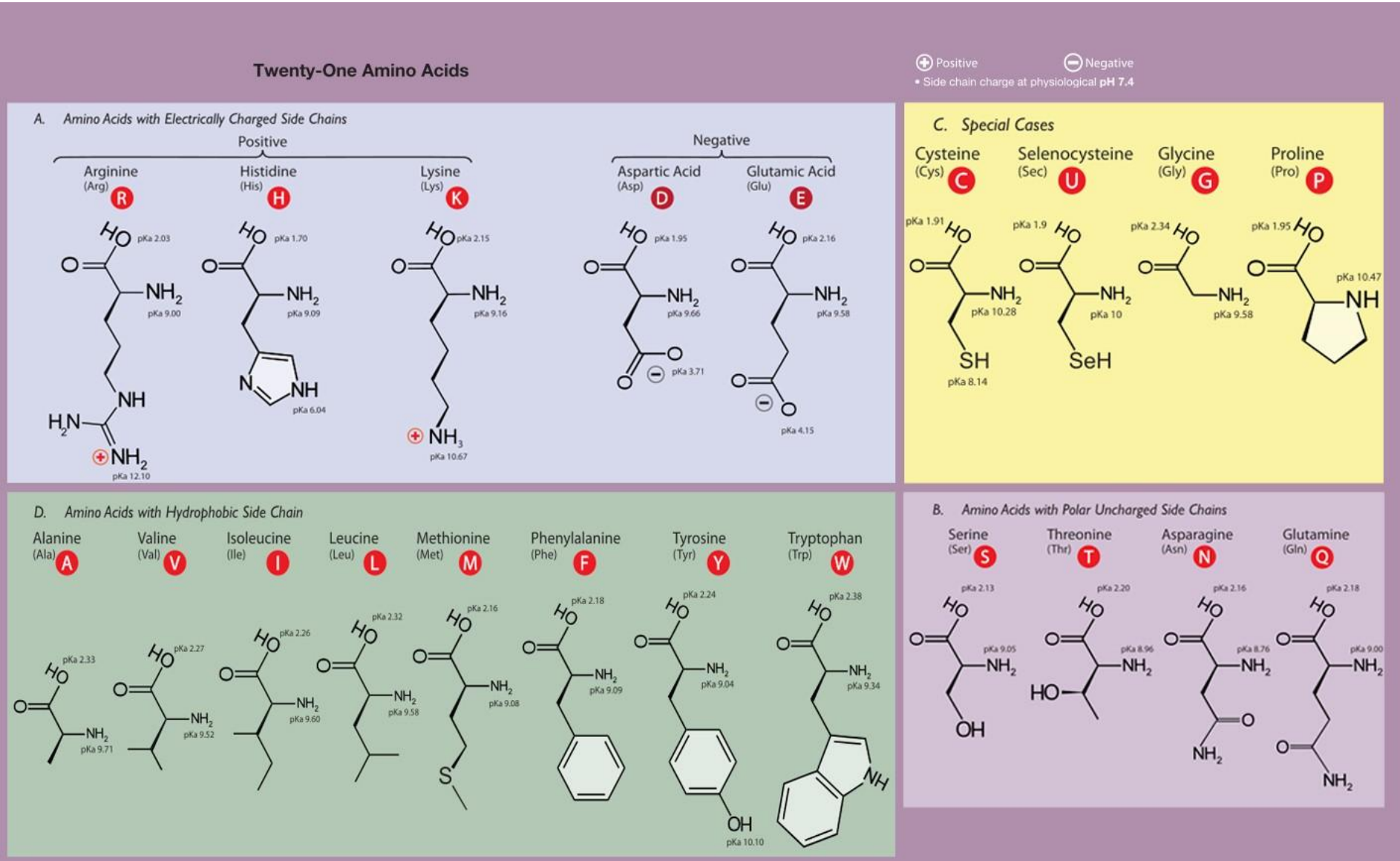
0) Простейший вариант: единичная матрица. Позволяет выявлять идентичные или очень похожие последовательности; для аминокислотных последовательностей не используется.

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$



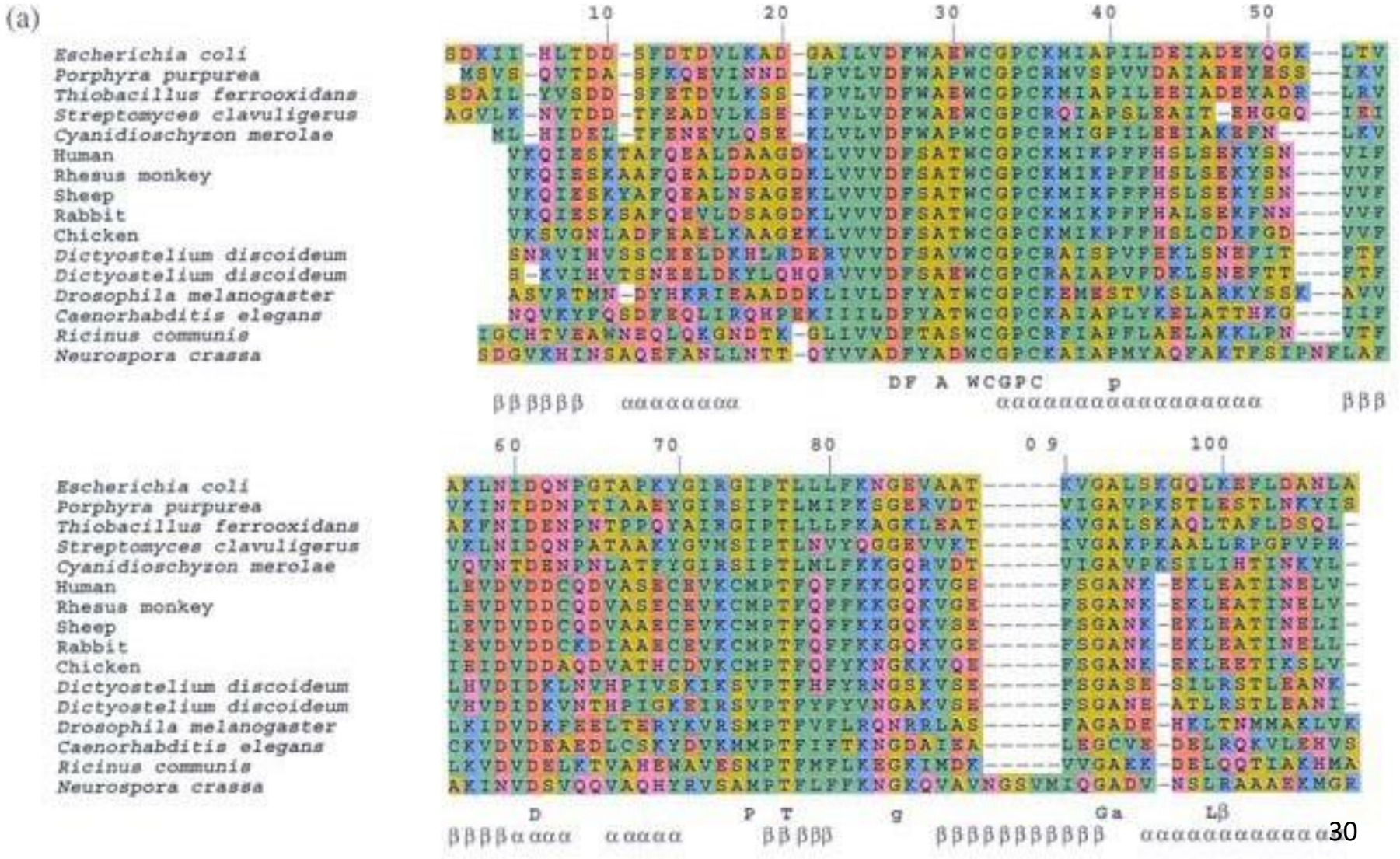
1) «Рациональный подход»: классифицируем аминокислоты на группы, а затем прибавляем 1 к оценке выравнивания за мутацию внутри группы или вычитаем 1 за мутацию между группами.

Структура аминокислот



Составление матриц замен

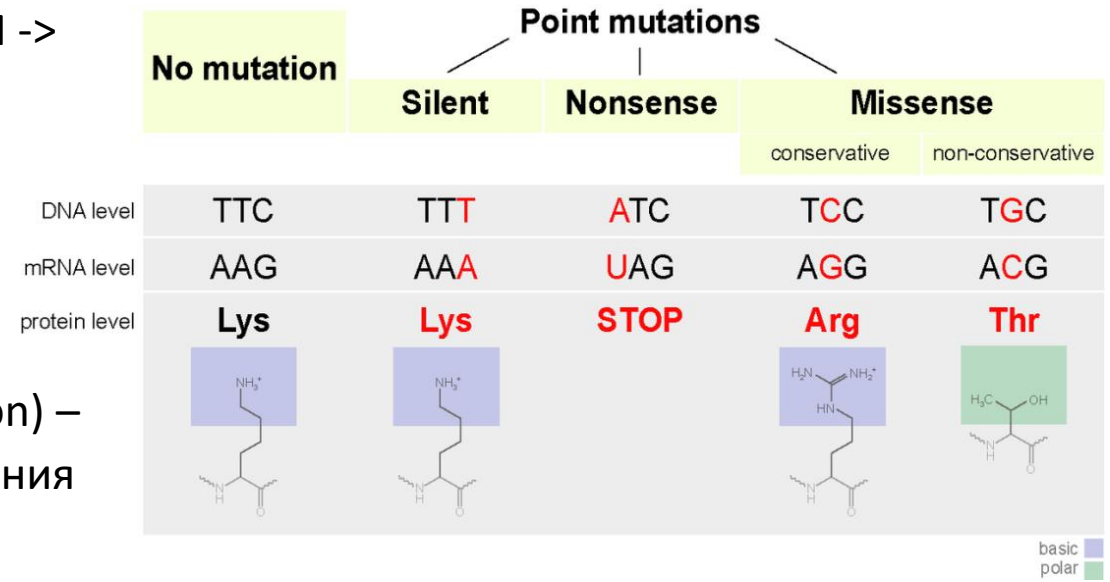
2) «Эмпирический подход»: собираем статистические данные о встречаемости аминокислотных замен в выравниваниях известных белков



Составление матриц вероятностей замен. PAM

Accepted Point Mutation (APM -> PAM) – замена одной аминокислоты на другую, закрепившаяся в процессе эволюции.

PAM (Percent Accepted Mutation) – единица измерения расхождения последовательностей в выравнивании.



PAM*N* – набор матриц вероятностей замен, аппроксимированных для выравниваний, содержащих *N* замен на 100 остатков. Исходно матрица PAM1 нормирована из расчета не более 1 мутации на 100 остатков; остальные получаются **матричным умножением**.

Матрица вероятностей замен PAM1

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	9867	3	10	17	2	21	2	6	2	4	6	9	22	8	2	35	32	18	0	2
C	1	9973	0	0	0	0	1	1	0	0	0	0	1	0	1	5	1	2	0	3
D	6	0	9859	53	0	6	4	1	3	0	0	42	1	6	0	5	3	1	0	0
E	10	0	56	9865	0	4	2	3	4	1	1	7	3	35	0	4	2	2	0	1
F	1	0	0	0	9946	1	2	8	0	6	4	1	0	0	1	2	1	0	3	28
G	21	1	11	7	1	9935	1	0	2	1	1	12	3	3	1	21	3	5	0	0
H	1	1	3	1	2	0	9912	0	1	1	0	18	3	20	8	1	1	1	1	4
I	2	2	1	2	7	0	0	9872	2	9	12	3	0	1	2	1	7	33	0	1
K	2	0	6	7	0	2	2	4	9926	1	20	25	3	12	37	8	11	1	0	1
L	3	0	0	1	13	1	4	22	2	9947	45	3	3	6	1	1	3	15	4	2
M	1	0	0	0	1	0	0	5	4	8	9874	0	0	2	1	1	2	4	0	0
N	4	0	36	6	1	6	21	3	13	1	0	9822	2	4	1	20	9	1	1	4
P	13	1	1	3	1	2	5	1	2	2	1	2	9926	8	5	12	4	2	0	0
Q	3	0	5	27	0	1	23	1	6	3	4	4	6	9876	9	2	2	1	0	0
R	1	1	0	0	1	0	10	3	19	1	4	1	4	10	9913	6	1	1	8	0
S	28	11	7	6	3	16	2	2	7	1	4	34	17	4	11	9840	38	2	5	2
T	22	1	4	2	1	2	1	11	8	2	6	13	5	3	2	32	9871	9	0	2
V	13	3	1	2	1	3	3	57	1	11	17	1	3	2	2	2	10	9901	0	2
W	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	0	0	9976	1
Y	1	3	0	1	21	0	4	1	0	1	0	3	0	0	0	1	1	1	2	9945

(p · 10 000)

Составление оценочных матриц для PAM

$$\text{Цена мутации } i \leftrightarrow j = 10 \lg \frac{\text{Наблюдаемая частота мутаций } i \leftrightarrow j}{\text{Частота мутаций } i \leftrightarrow j\text{, ожидаемая из частоты встречаемости а.о. i и j}}$$

(упрощенно)

71 **дерево**, 1572 замены -> PAM1 -> **PAM250**

(M.O.Dayhoff, 1978)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Составление матриц замен

$f_i = \frac{N_i}{N}$ Встречаемость аминокислотного остатка типа *i* в белках

$p_{i,j}$ Вероятность замены остатка типа *i* на остаток типа *j*

$f_i p_{i,j}$ Вероятность того, что пара остатков в выравнивании является следствием мутации, т.е. позиции выровнены правильно

$f_i f_j$ Вероятность того, что пара остатков в выравнивании образовалась случайно, т.е. позиции совпали случайно

$\frac{f_i p_{i,j}}{f_i f_j} = \frac{p_{i,j}}{f_j}$ Соотношение правдоподобности выравнивания позиции

$\prod_k \frac{p_{i_k,j_k}}{f_{j_k}}$ Соотношение правдоподобности выравнивания всех *k* позиций в белке

$S = \sum_k \log_2 \left(\frac{p_{i_k,j_k}}{f_{j_k}} \right)$ Оценка выравнивания (во избежание ошибок округления используют логарифмирование)

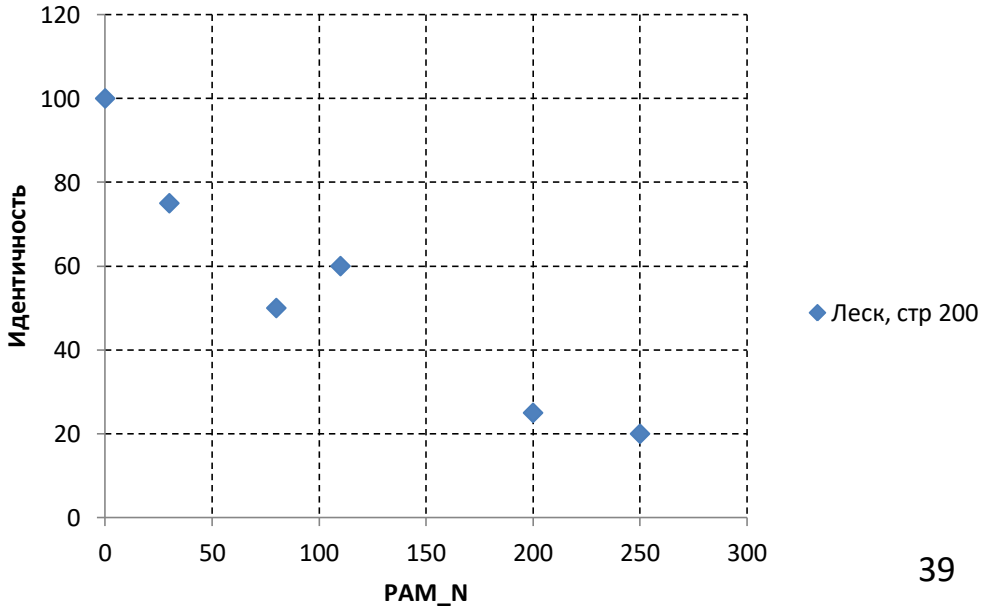
PAM. Что насчет консервативности?

PAM	0	1	30	80	110	200	250
~ % идентичности	100	99	75	50	60	25	20

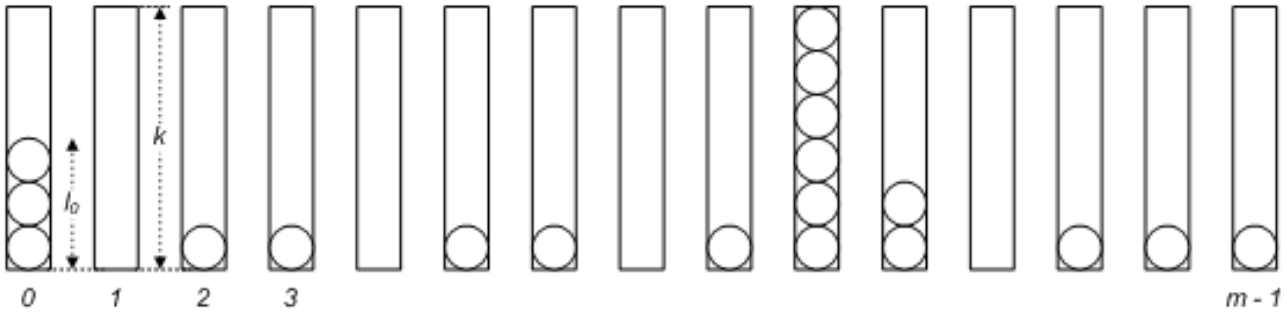
(А. Леск. Введение в биоинформатику (2009), стр. 200)

1) Как оценить консервативность, если матрицы PAM2…PAM250 получены чисто теоретически?

2) Верны ли числа в этой таблице?



PAM. Что насчет консервативности?



N – число лунок

K – общее число шаров

q – вероятность того, что наугад выбранная лунка окажется пустой

E(L) – математическое ожидание числа пустых лунок

$$q = \left(\frac{N-1}{N} \right)^K$$

$$E(L) = Nq = N\left(\frac{N-1}{N} \right)^K = N\left(1 - \frac{1}{N} \right)^K$$

PAM. Что насчет консервативности?

Т.о. если бы мутации различных остатков были равновероятны, то

$$E(L) = N(1-\frac{1}{N})^K; E(M) = N - N(1-\frac{1}{N})^K$$

N – длина последовательности

K – общее число мутаций

M – число позиций с мутациями

E(L) – математическое ожидание числа консервативных позиций

E(M) – математическое ожидание числа позиций с мутациями

PAM	**0**	**1**	**30**	**80**	**110**	**200**	**250**
~ % идентичности	100	99	74	45	33	13	8
	100	99	75	50	60	25	20

PAM. Что насчет консервативности?

Т.о. если бы мутации различных остатков были **равновероятны**, то

$$E(L) = N(1-\frac{1}{N})^K; E(M) = N - N(1-\frac{1}{N})^K$$

N – длина последовательности

K – общее число мутаций

M – число позиций с мутациями

E(L) – математическое ожидание числа консервативных позиций

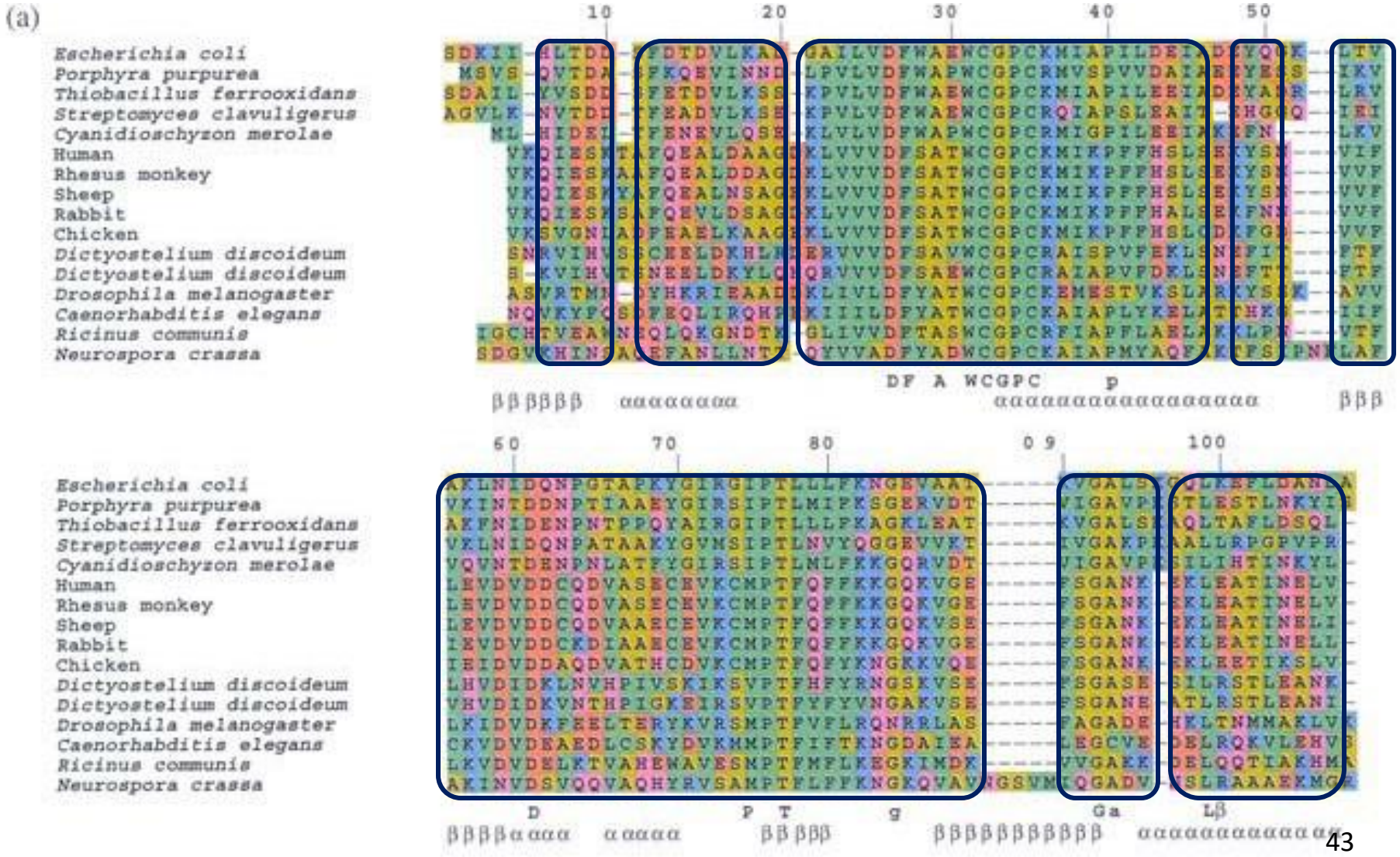
E(M) – математическое ожидание числа позиций с мутациями

PAM	**0**	**1**	**30**	**80**	**110**	**200**	**250**
~ % идентичности	100	99	**74**	**45**	**33**	**13**	**8**
	100	99	75	50	60	25	20

Доверяй, но проверяй! ☺

Составление матриц замен. BLOSUM

BLOck SUbstitution Matrix



Составление матриц замен. BLOSUM

559 семейств, 2106 блоков («conserved regions»), >15 млн замен + Кластеризация для уменьшения вклада очень похожих последовательностей -> BLOSUMXX

(S. Henikoff and J.G. Henikoff, 1992)

BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Другие матрицы замен

FASTA – поиск гомологов в базах данных

Matrix Name	Target Identity	Abbreviation
BLOSUM50	25%	BL50
BLASTP62	30%	BP62
BLOSUM80	40%	BL80
PAM250	20%	P250
PAM120	35%	P120
MDM40	65%	M40
MDM20	85%	M20
MDM10	90%	M10
VTML160	25%	VT160
VTML120	35%	VT120
VTML80	40%	VT80
VTML40	65%	VT40
VTML20	85%	VT20
VTML10	90%	VT10

-даже не гуглится ☹

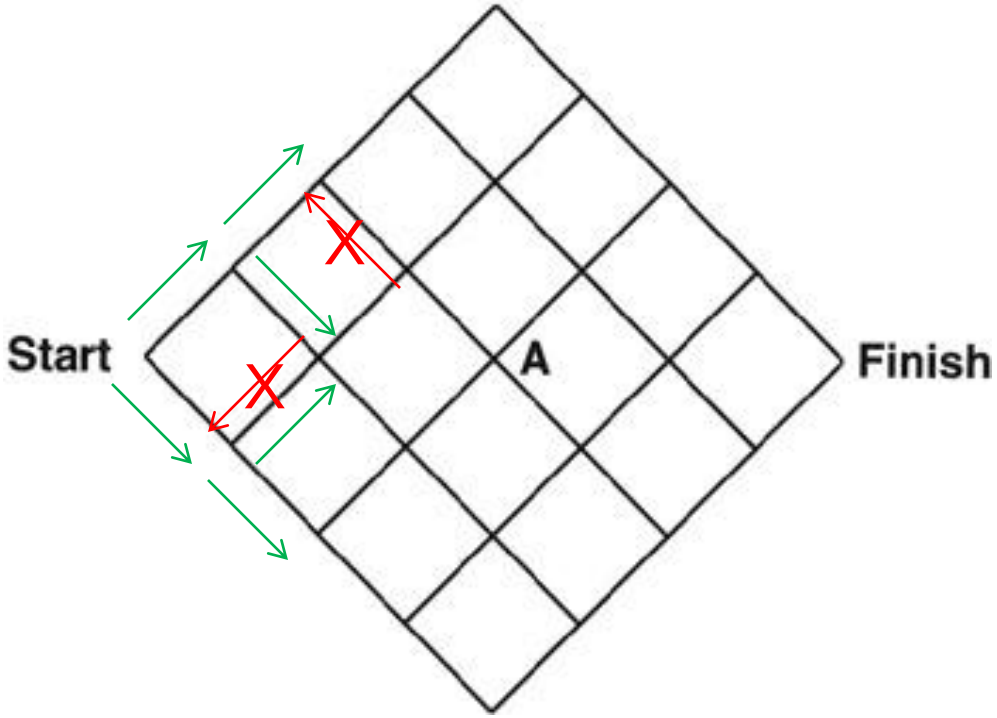
- аналоги PAM по данным на 1992 год

Получены в 2000 году путем построения
- теоретической модели дивергенции
последовательностей в процессе эволюции

Расчет выравнивания двух последовательностей

Сколько путей ведет от старта к финишу и проходит через точку A?

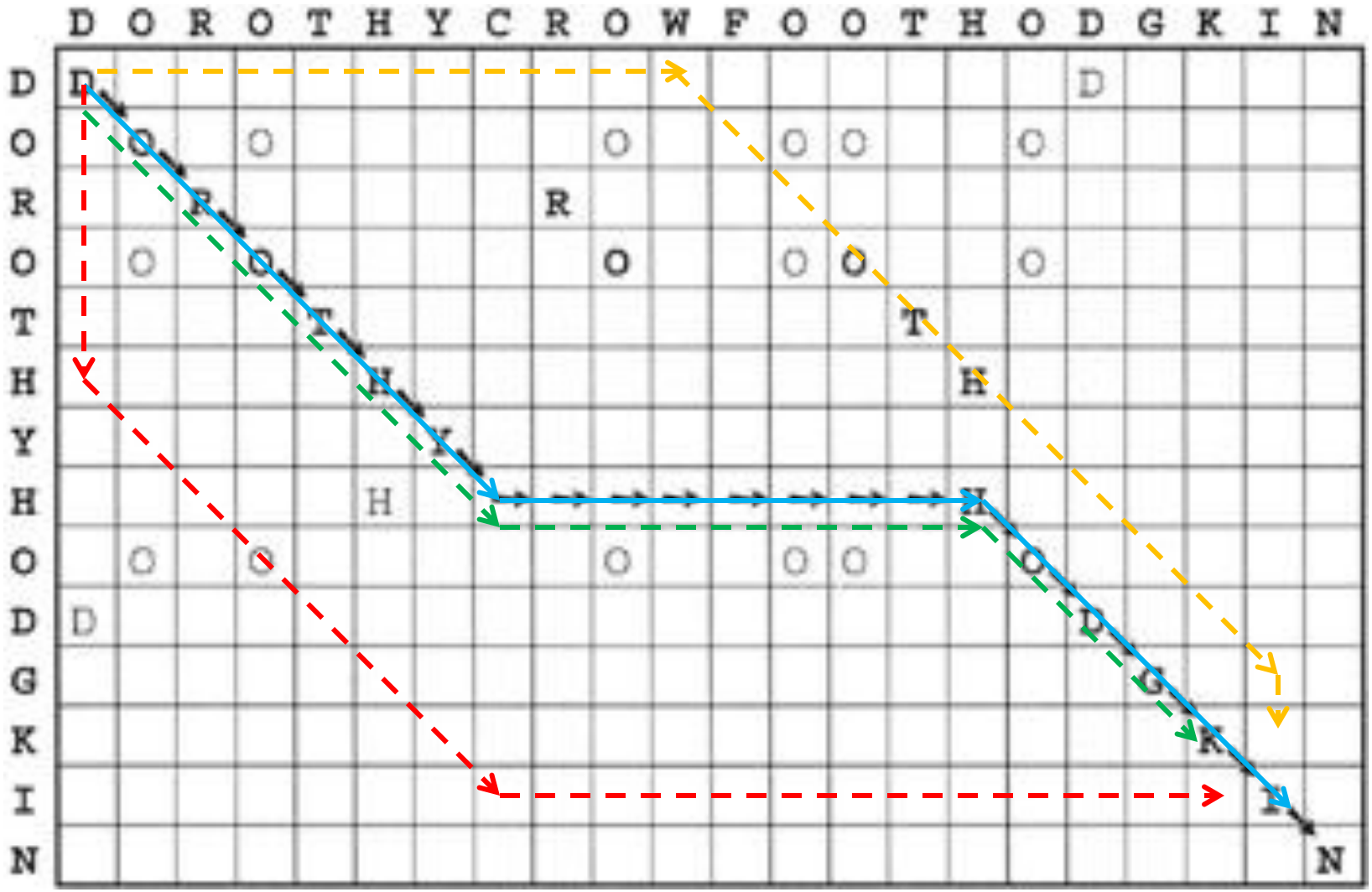
Сколько путей надо оценить, чтоб выбрать лучший?



Метод динамического программирования по нахождению оптимального пути в матрице основан на идее её систематического разделения на все более мелкие фрагменты.

Поиск глобального выравнивания

Оптимальный путь до точки (i, j) определяется оптимальными путями до точек (i-1, j), (i, j) и (i, j-1) и оценкой последнего перехода.



Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша (1970)

	"-"	g	c	t	g	a	a	c	g
"-"	0	-1	-2	-3	-4	-5	-6	-7	-8
c	-1								
t	-2								
a	-3								
t	-4								
a	-5								
a	-6								
t	-7								
c	-8								

match mut gap

2 -2 -1

Базис:

$$F_{i0} = d \cdot i$$

$$F_{0j} = d \cdot j$$

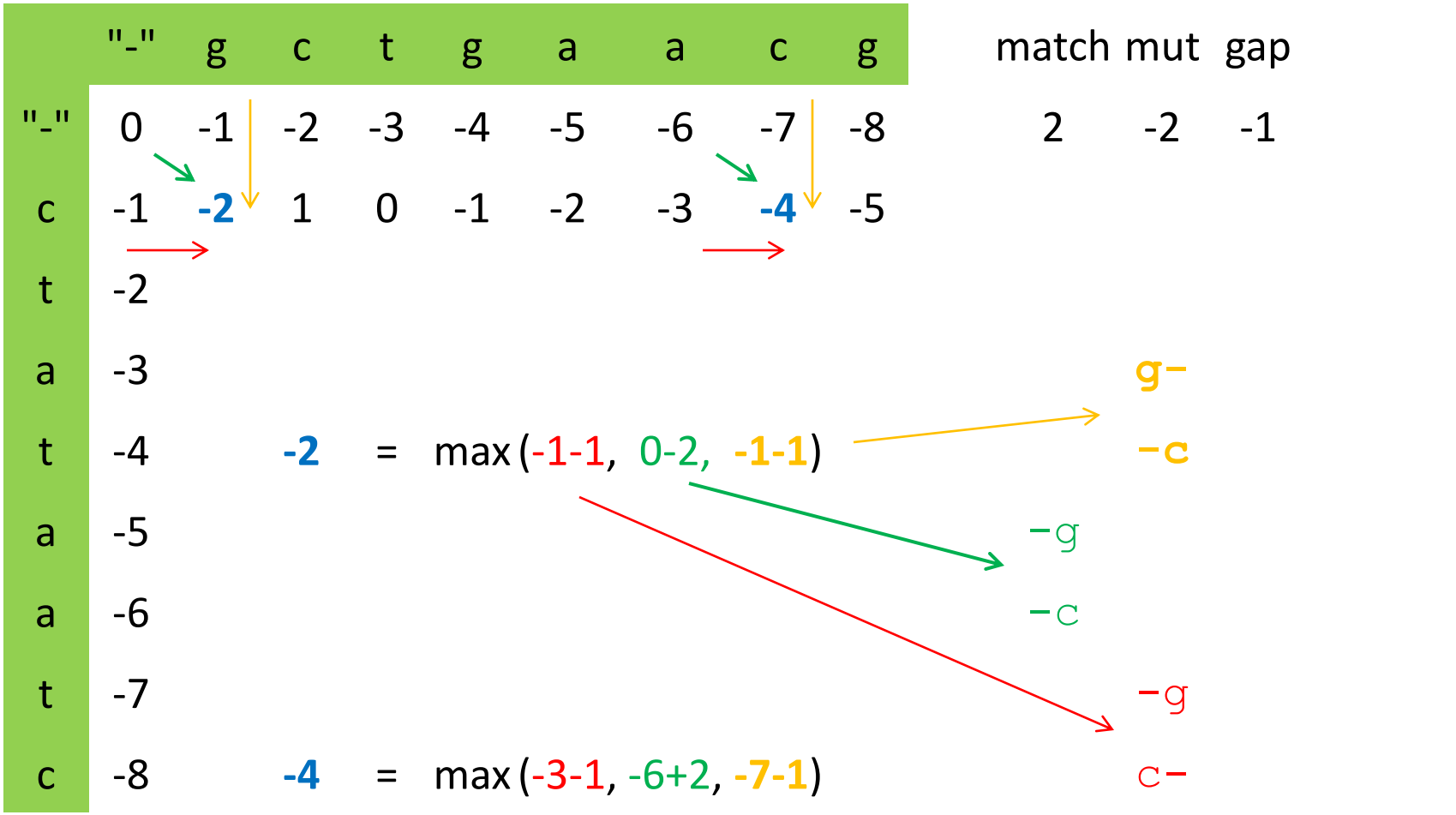
Рекурсия, основанная на принципе оптимальности:

$$F_{ij} = \max(F_{i-1,\,j-1} + S(A_i,\ B_j),\ F_{i,\,j-1} + d,\ F_{i-1,\,j} + d).$$

матрица замен

штраф за вставку

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша (1970)



Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша (1970)

	"-"	g	c	t	g	a	a	c	g
"-"	0	-1	-2	-3	-4	-5	-6	-7	-8
c	-1	-2	1	0	-1	-2	-3	-4	-5
t	-2	-3	0	3	2	1	0	-1	-2
a	-3	-4	-1	2	1	4	3	2	1
t	-4	-5	-2	1	0	3	2	1	0
a	-5	-6	-3	0	-1	2	5	4	3
a	-6	-7	-4	-1	-2	1	4	3	2
t	-7	-8	-5	-2	-3	0	3	2	1
c	-8	-9	-6	-3	-4	-1	2	5	4

match mut gap

2 -2 -1

g c t g a – a – – c g
– c t – a t a a t c –

L = 6

Построение пути «по максимумам»

Прочтение выравнивания

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша (1970)

	"-"	g	c	t	g	a	a	c	g
"-"	0	-1	-2	-3	-4	-5	-6	-7	-8
c	-1	-1	-1	-2	-3	-4	-5	-6	-7
t	-2	-2	-2	-1	-2	-3	-4	-5	-6
a	-3	-3	-3	-2	-2	-2	-3	-4	-5
t	-4	-4	-4	-3	-3	-3	-3	-4	-5
a	-5	-5	-5	-4	-4	-3	-3	-4	-5
a	-6	-6	-6	-5	-5	-4	-3	-4	-5
t	-7	-7	-7	-6	-6	-5	-4	-4	-5
c	-8	-8	-7	-7	-7	-6	-5	-4	-5

match mut gap

0 **-1** -1

```
g c t - g a a c g
- c t a t a a t c
```

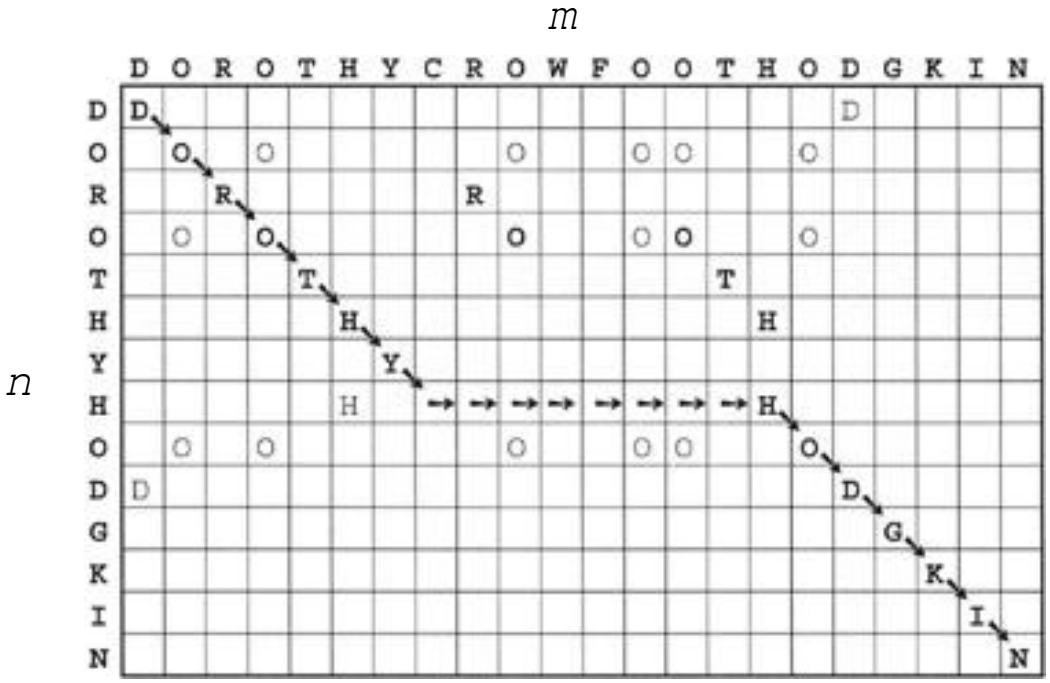
L = 5

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша (1970)

Варианты штрафов за вставку: постоянный, линейный, аффинный, …

Построение глобального выравнивания не годится для поиска по БД (сотни тысяч последовательностей), т.к. требует (*m x n*) времени и памяти для хранения матриц.

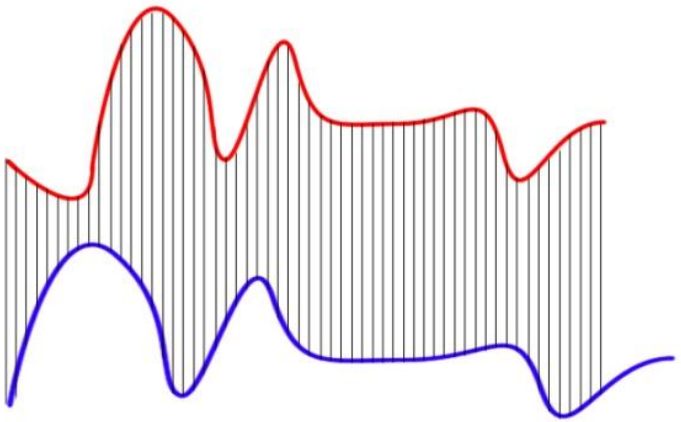
Сколько существует выравниваний для последовательностей длины *m* и *n*?



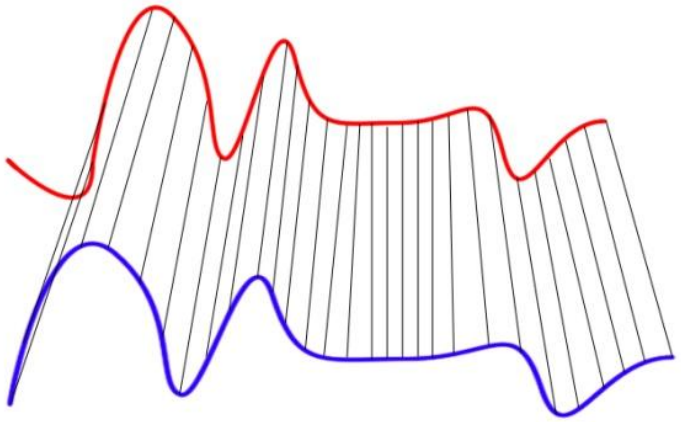
$$f(m, n) = f(m - 1, n) + f(m - 1, n - 1) + f(m, n - 1);$$
$$f(0, n) = f(m, 0) = 1$$

Другие приложения динамического программирования

«Распознавание слов устной речи методами динамического программирования»
(Винцюк Т.К., 1968)



Euclidean Matching

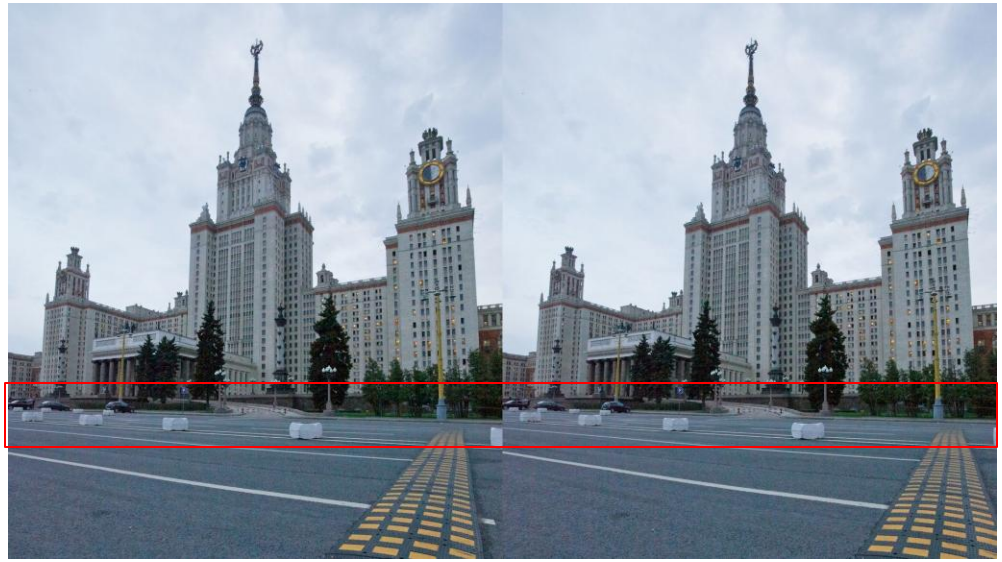


Dynamic Time Warping Matching

[Getting to know Siri:
Dynamic Time Warping
algorithm for speech recognition]

Другие приложения динамического программирования

Компьютерное стереозрение



Правый (= нижний) фрагмент растянут и смещен относительно левого (= верхнего) – типичная задача для выравнивания

The European Bioinformatics Institute (EMBL-EBI)

maintains the world's most comprehensive range of freely available and up-to-date molecular data resources

Tools & Data Resources

Search all tools & data resourc

Tools

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Multiple sequence alignment

InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.

Protein feature detection | Sequence motif recognition

BLAST [protein]

Fast local similarity search tool for protein sequence databases.

Sequence similarity search

Data resources

Ensembl

Genome browser, API and database, providing access to reference genome annotation

UniProt

A comprehensive resource for protein sequence and functional annotation.

PDBe

The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

Europe PMC

A database to search the worldwide life sciences literature

Browse by type

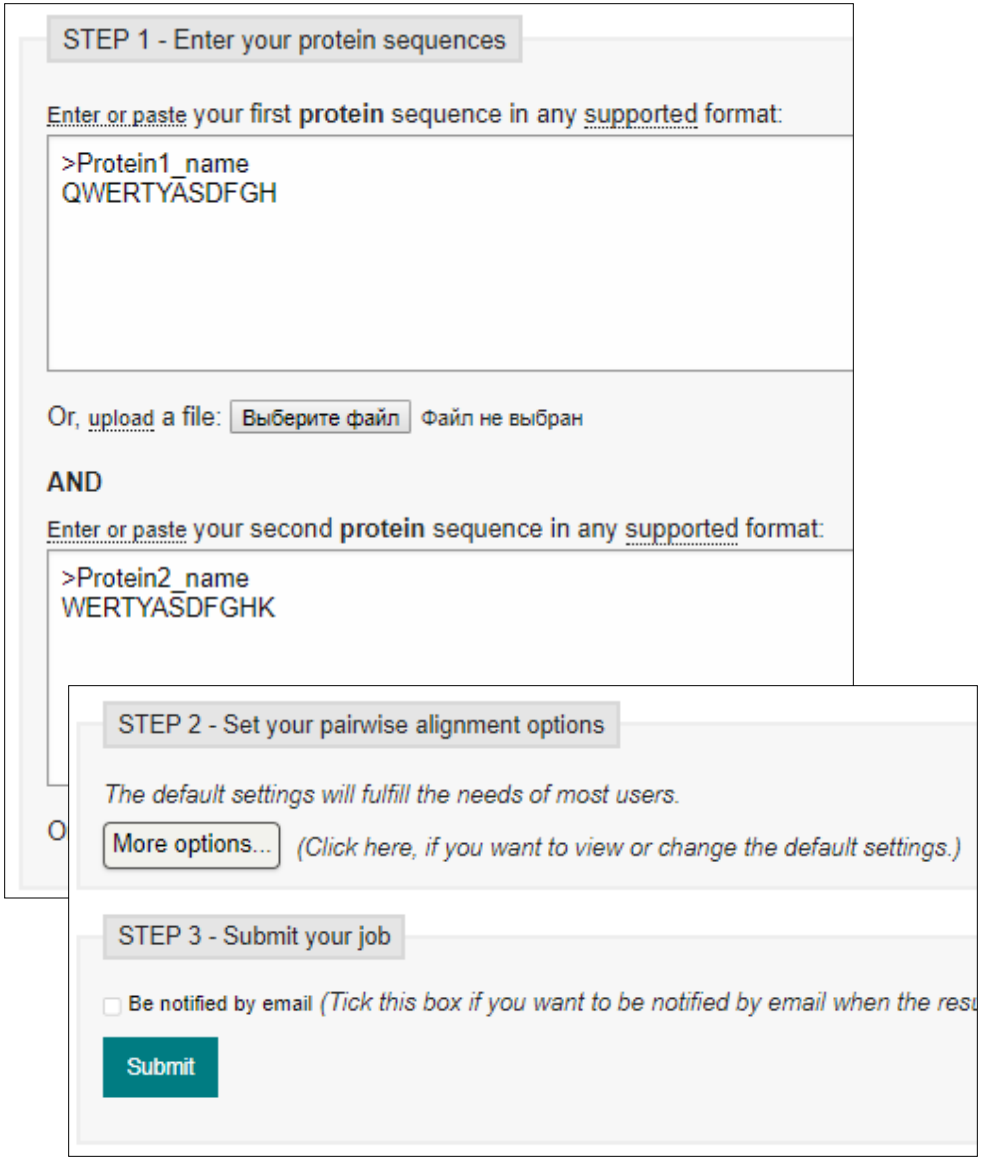
DNA & RNA	Gene Expression	Proteins
Structures	Systems	Chemical biology
Ontologies	Literature	Cross domain

Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources will work well together.

Browse EMBL-EBI web services

Пример выравнивания двух последовательностей. EMBOSS Needle




```
# Aligned_sequences: 2
# 1: Protein1_name
# 2: Protein2_name
# Matrix: BLOSUM62
# 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
```

Пример выравнивания двух последовательностей.

EMBOSS Needle

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Выберите файл Файл не выбран
3. To enter the sequence data manually, type here: gctgaacg

Select an input sequence. Use one of the following three fields:

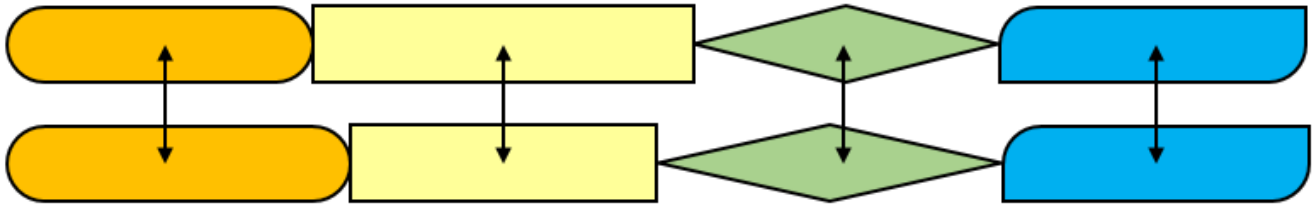
1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Выберите файл Файл не выбран
3. To enter the sequence data manually, type here: ctataatc

Matrix file. Use one of the following two fields:

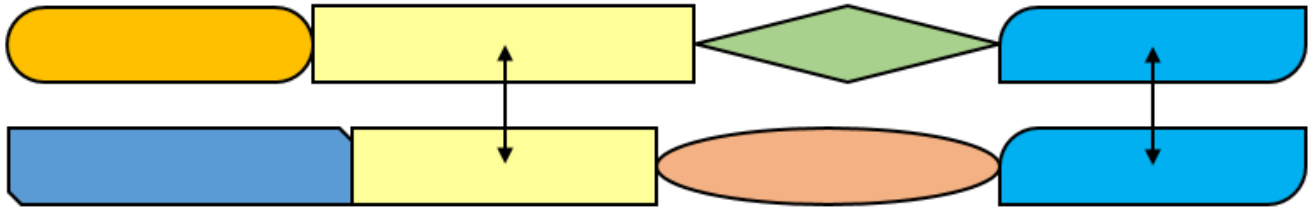
1. To access a standard EMBOSS data file, enter the name here:
   (default is EBLOSUM62 for protein, EDNAFULL for nucleic)
2. To upload a data file from your local computer, select it here: Выберите файл Файл не выбран

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 (1981)

	"-"	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 Water

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Выберите файл Файл не выбран
3. To enter the sequence data manually, type here: gctgaacg

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Выберите файл Файл не выбран
3. To enter the sequence data manually, type here: ctataatc

Matrix file. Use one of the following two fields:

1. To access a standard EMBOSS data file, enter the name here:
   (default is EBLOSUM62 for protein, EDNAFULL for nucleic)
2. To upload a data file from your local computer, select it here: Выберите файл Файл не выбран

Пример выравнивания двух последовательностей.

EMBOSS Water

Input section

Select an input sequence. Use one of the following three fields:

1. To access a sequence from a database, enter the USA here:
2. To upload a sequence from your local computer, select it here: Выберите файл Файл не выбран
3. To enter the sequence data manually, type he

```
# Aligned_sequences: 2
# 1:
# 2:
# Matrix: DNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 2
# Identity: 2/2 (100.0%)
# Similarity: 2/2 (100.0%)
# Gaps: 0/2 ( 0.0%)
# Score: 10.0
#
#
#=======================================
         2 ct 3
           ||
         1 ct 2
```

Select an input sequence. Use one of the following

1. To access a sequence from a database, enter
2. To upload a sequence from your local comp
3. To enter the sequence data manually, type he

Matrix file. Use one of the following two fields:

1. To access a standard EMBOSS data file, enter the name here:
   (default is EBLOSUM62 for protein, EDNAFULL for nucleic)
2. To upload a data file from your local computer, select it here: Выберите файл Файл не выбран

Другие матрицы замен. DNAFULL

	A	**T**	**G**	**C**	S	W	R	Y	K	M	B	V	H	D	N
A	**5**	**-4**	**-4**	**-4**	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2
T	**-4**	**5**	**-4**	**-4**	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2
G	**-4**	**-4**	**5**	**-4**	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2
C	**-4**	**-4**	**-4**	**5**	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-3	-1	-1
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-3	-1	-1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

ROSALIND

Locations

Rosalind is a platform for learning bioinformatics and programming through problem solving. Take a tour to get the hang of how Rosalind works.

If you don't know anything about programming, you can start at the Python Village. For a collection of exercises to accompany Bioinformatics Algorithms book, go to the Textbook Track. Otherwise you can try to storm the Bioinformatics Stronghold right now.

Python Village

If you are completely new to programming, try these initial problems to learn a few basics about the Python programming language. You'll get familiar with the operations needed to start solving bioinformatics challenges in the Stronghold.

Bioinformatics Stronghold

Discover the algorithms underlying a variety of bioinformatics topics: computational mass spectrometry, alignment, dynamic programming, genome assembly, genome rearrangements, phylogeny, probability, string algorithms and others.

Bioinformatics Armory

Ready-to-use software tools abound for bioinformatics analysis. Whereas in the Stronghold you implement algorithms on your own, in the Armory you solve similar problems by using existing tools.

http://rosalind.info