

# ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

## Лекция №4

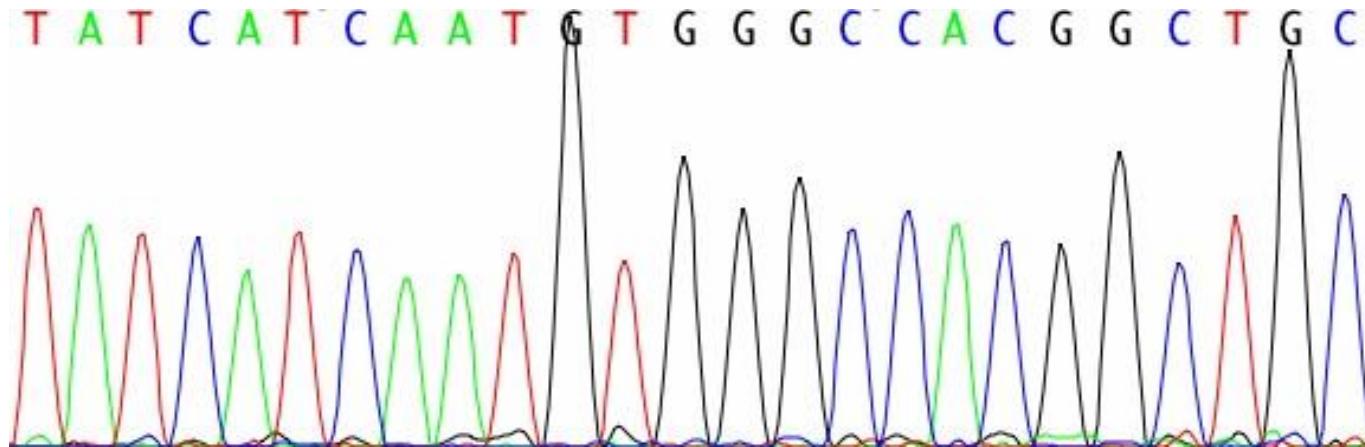
Сравнение последовательностей: матрицы сходства, динамическое программирование, локальное и глобальное выравнивание.

Матрицы замен.

Новоселецкий Валерий Николаевич  
к.ф.-м.н., доц. каф. биоинженерии  
[valery.novoseletsky@yandex.ru](mailto:valery.novoseletsky@yandex.ru)

Сайт курса <https://intbio.org/bioinf2018-2019/>

# Вы только что секвенировали новую последовательность ДНК



Что вы можете 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 остатками
- Выявить консервативные и переменные области
- Предположить эволюционные взаимосвязи

## Выравнивание последовательностей – основной инструмент биоинформатики

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 -
```

# Точечная матрица сходства (dotplot)

Рассмотрим две последовательности:

- DOROTHYNODGKIN
- 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)

Последовательность с повторами  
(ABRACADABRACADABRA):

	A	B	R	A	C	A	D	A	B	R	A	C	A	D	A	B	R	A
A	A			A		A		A			A	A		A			A	
B		B							B								B	
R			R							R							R	
A	A			A		A		A			A	A		A			A	
C					C							C						
A	A			A		A		A			A	A		A			A	
D						D							D					
A	A			A		A		A			A	A		A			A	
B		B							B								B	
R			R							R							R	
A	A			A		A		A			A	A		A			A	
C						C							C					
A	A			A		A		A			A	A		A			A	
D							D							D				
A	A			A		A		A			A	A		A			A	
B		B							B								B	
R			R							R							R	
A	A			A		A		A			A	A		A			A	

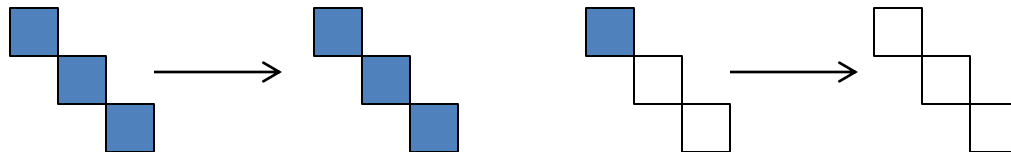
Палиндромная последовательность  
(MAX I STAY AWAY AT SIX AM):

	M	A	X	I	S	T	A	Y	A	W	A	Y	A	T	S	I	X	A	M
M	M																		M
A		A					A		A		A		A					A	
X			X															X	
I				I												I			
S					S											S			
T						T										T			
A		A					A		A		A		A					A	
Y								Y					Y						
A		A							A		A		A					A	
W										W									
A		A							A		A		A					A	
Y										Y						Y			
A		A								A		A		A				A	
T							T									T			
S								S								S			
I									I								I		
X			X															X	
A		A								A		A		A				A	
M	M																		M

# Точечная матрица сходства (dotplot)

Для реальных последовательностей (белок PAX-6 мыши (код по БД 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.



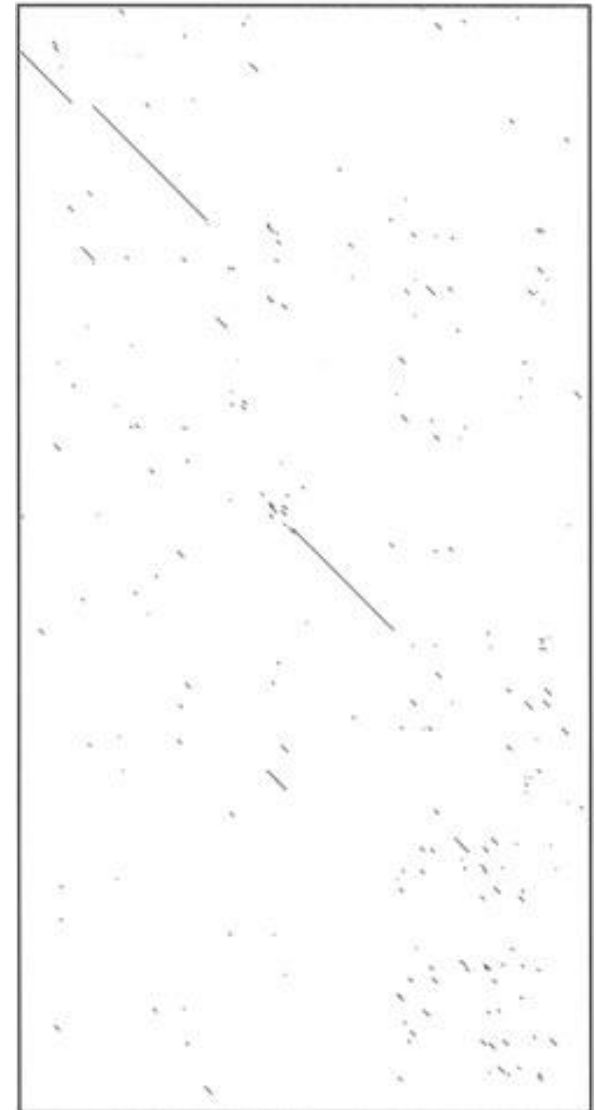
Интерактивное построение точечных матриц:

<http://www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html>

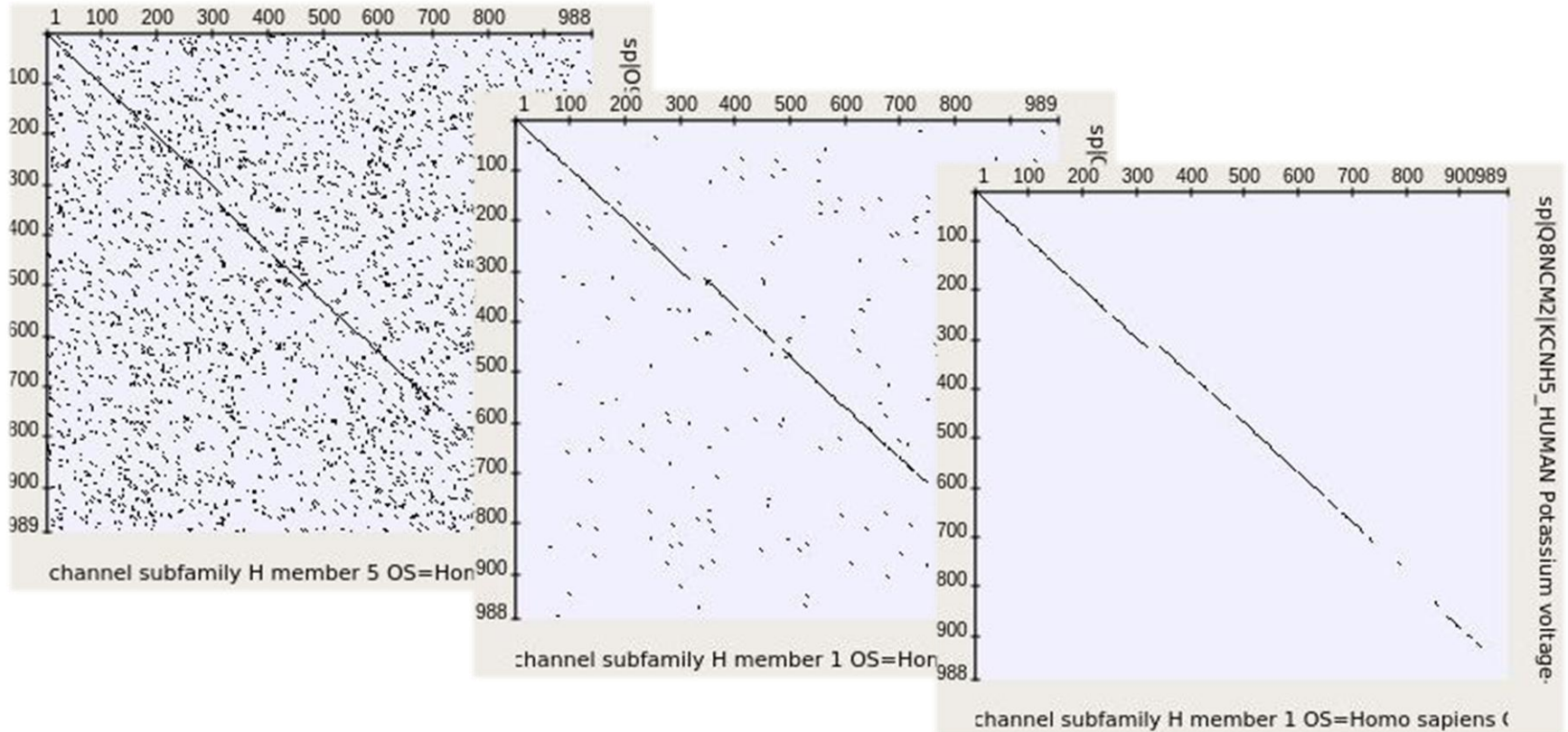
<http://myhits.isb-sib.ch/cgi-bin/dotlet>



mouse PAX-6 / Drosophila eyeless



# Точечная матрица сходства (dotplot)



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

Free bioinformatics software:

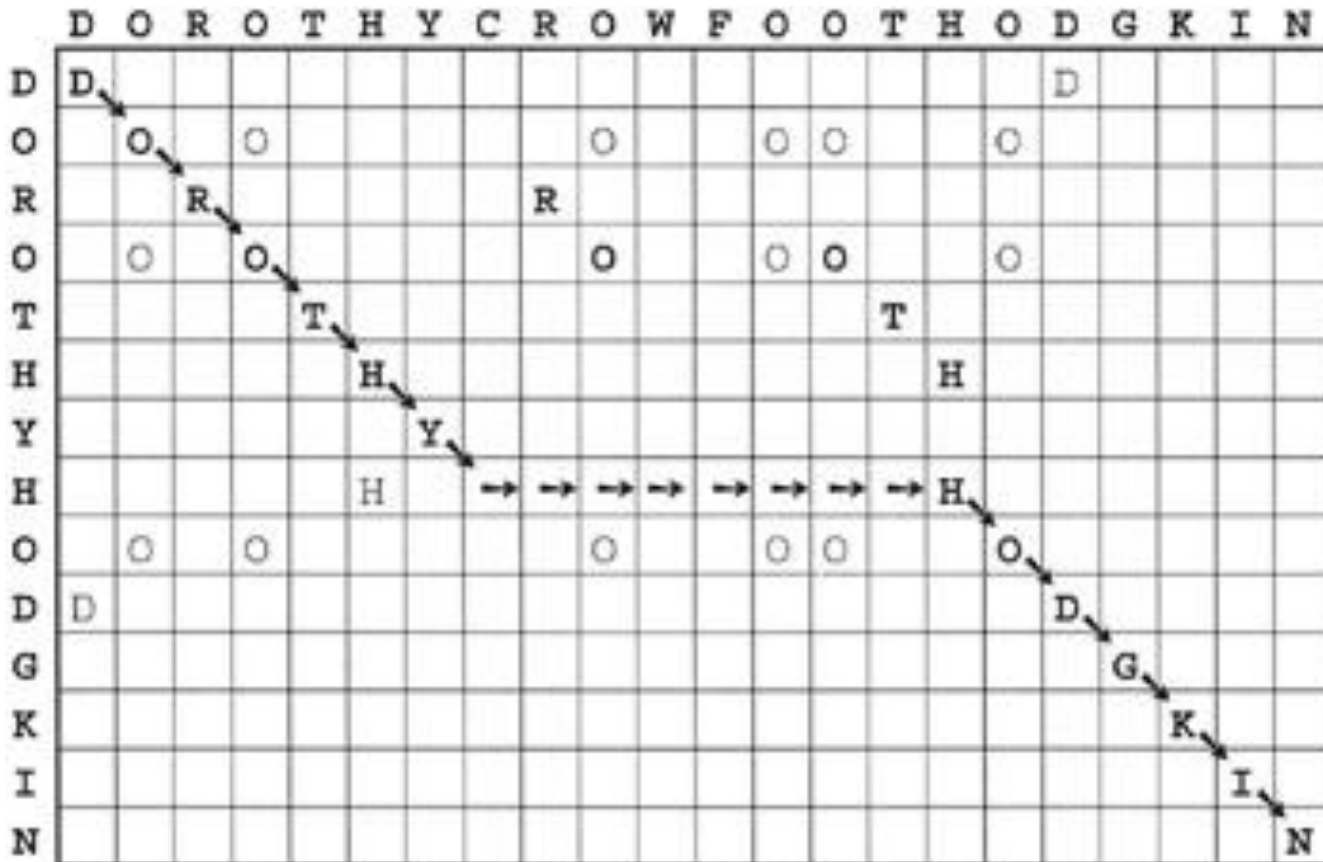
<http://ugene.net>

Unipro **UGENE**

# Точечные матрицы и выравнивание последовательностей

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

DOROTHY-----HODGKIN  
 DOROTHY CROWFOOTHODGKIN





# Точечные матрицы и выравнивание последовательностей

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

DORO-----THYHOD---GKIN  
 DOROTHYHCROWFOOTHODGKIN



# Мера сходства последовательностей

## Выравнивание

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

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

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

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

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

**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)

Лауреат премии  
Хэмминга 2006 года

# Мера сходства последовательностей

Выравнивание

Расстояние по  
Левенштейну

- - - - - g c t g a a c g =  
c t a t a a t c - - - - -

g c t g a a c g =  
c t a t a a t c

g c t g a - a - - c g =  
- - c t - a t a a t c

g c t g - a a - c g =  
- c t a t a a t c -

# Мера сходства последовательностей

Выравнивание

Расстояние по  
Левенштейну

- - - - - g c t g a a c g = 15  
c t a t a a t c - - - - -

g c t g a a c g = 6  
c t a t a a t c

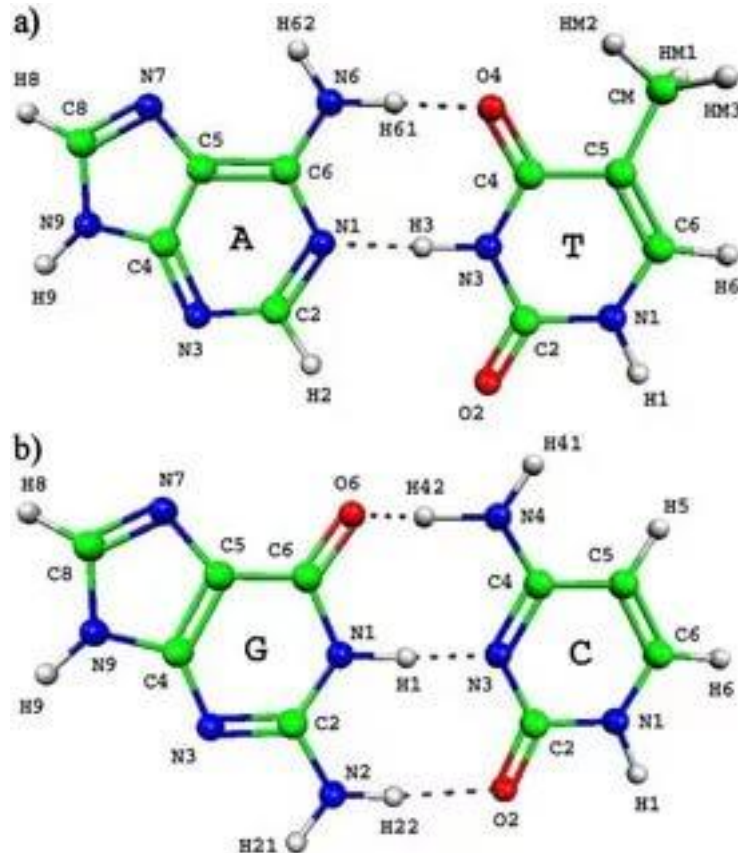
g c t g a - a - - c g = 11  
- - c t - a t a a t c

g c t g - a a - c g = 5  
- c t a t a a t c -

# 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

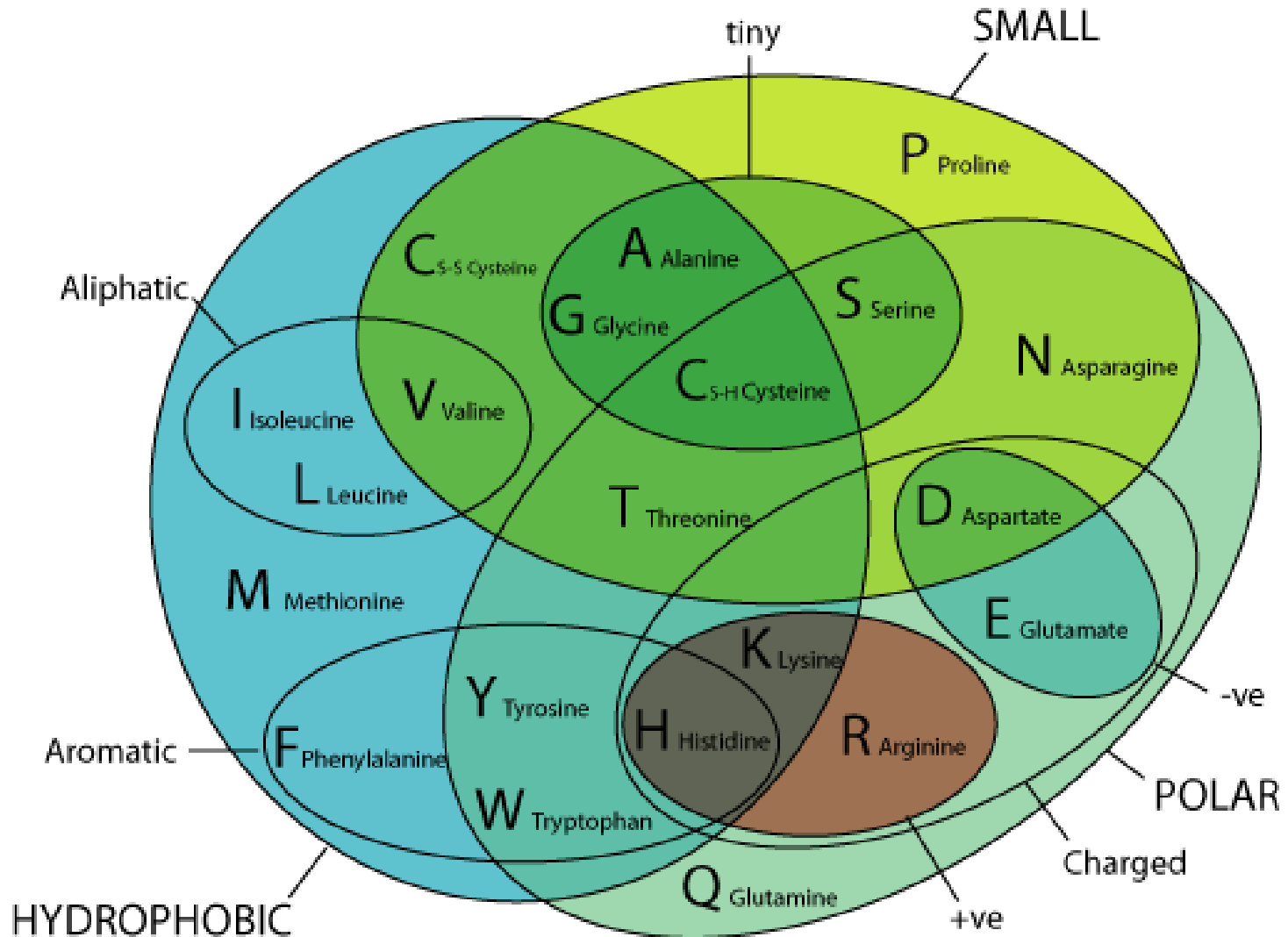
# Составление матриц замен

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

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

# Составление матриц замен

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.



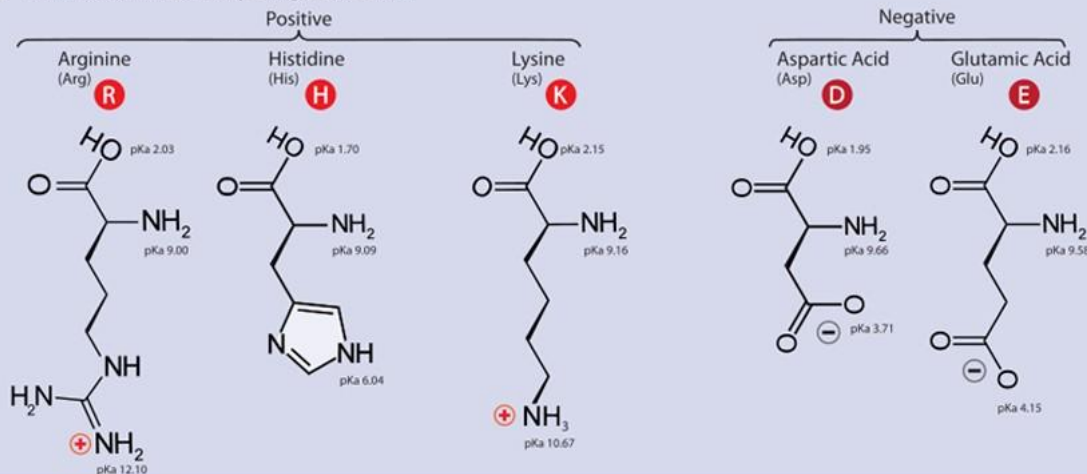


# Структура аминокислот

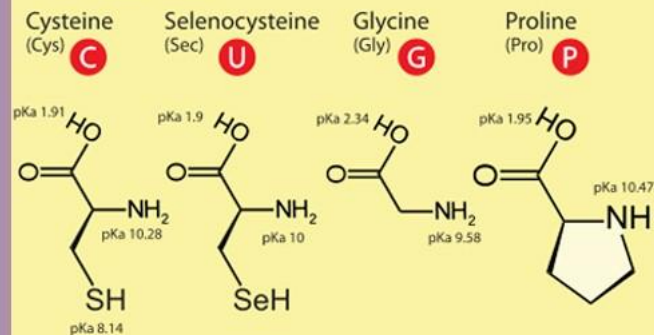
## Twenty-One Amino Acids

⊕ Positive      ⊖ Negative  
• Side chain charge at physiological pH 7.4

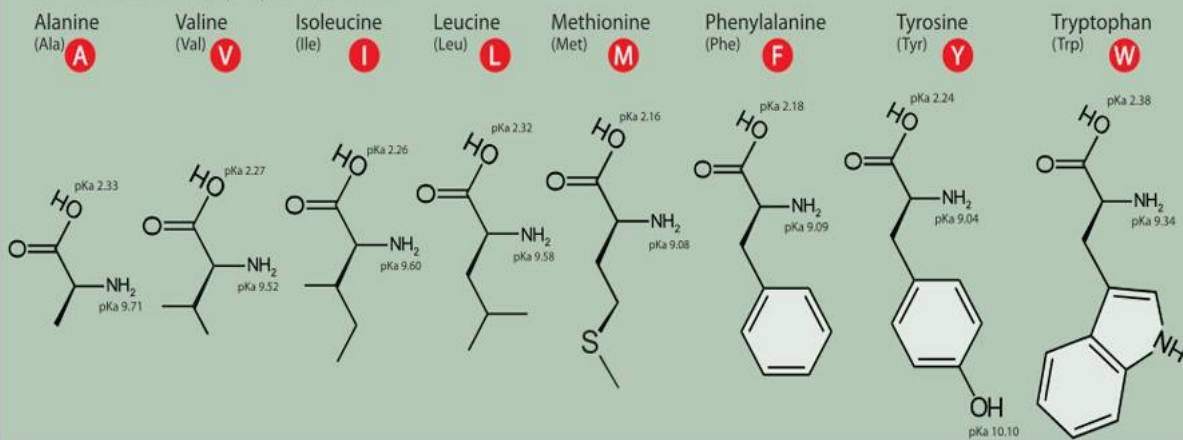
### A. Amino Acids with Electrically Charged Side Chains



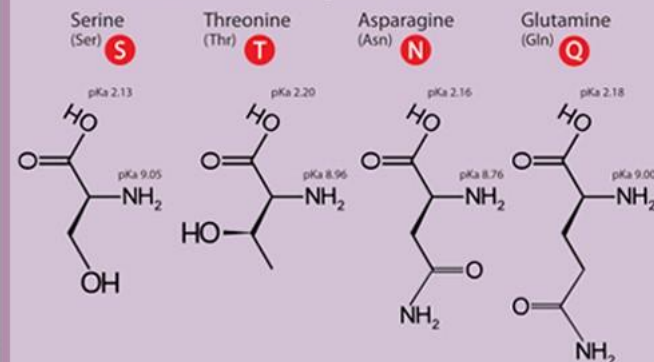
### C. Special Cases



### D. Amino Acids with Hydrophobic Side Chain

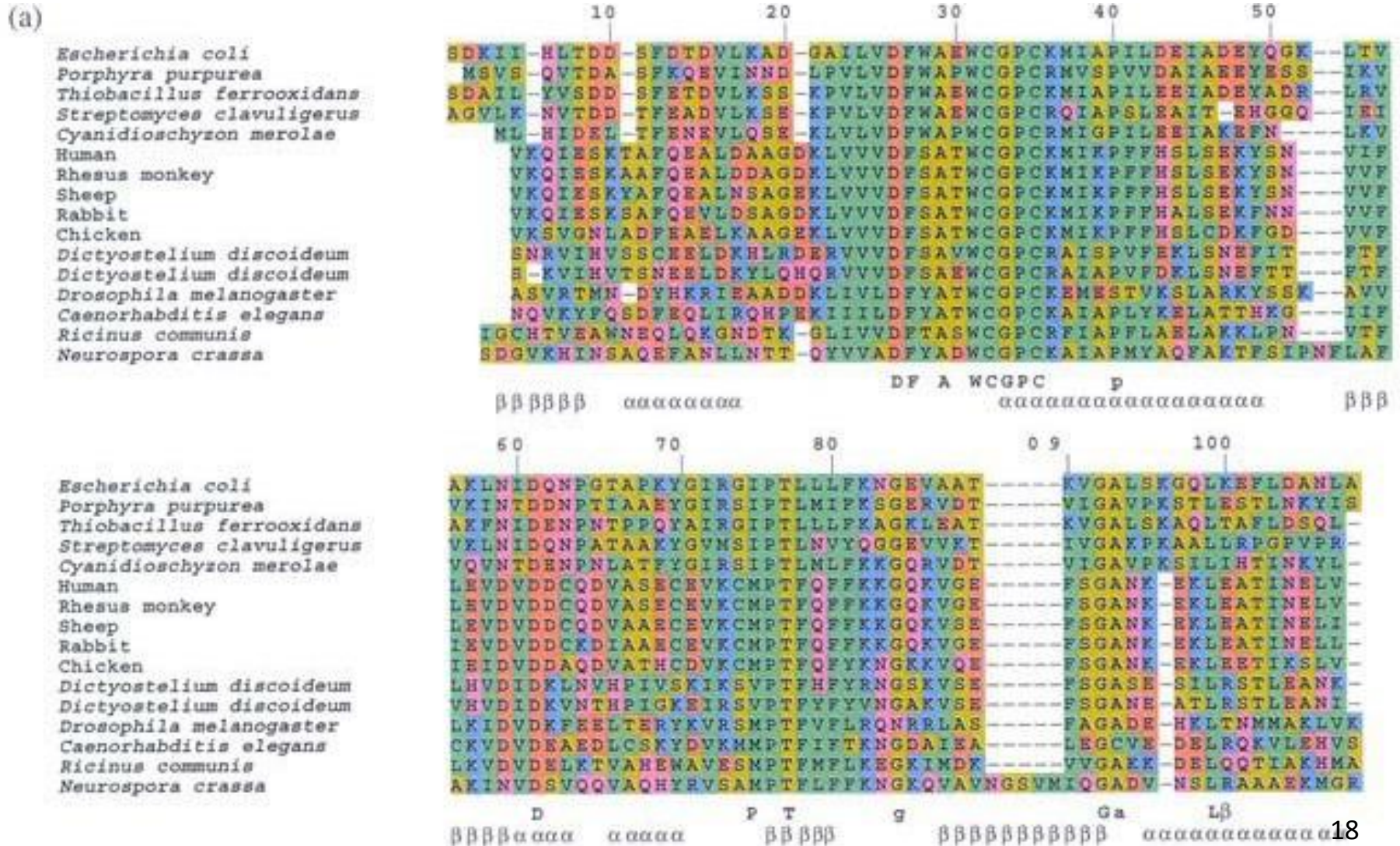


### B. Amino Acids with Polar Uncharged Side Chains



# Составление матриц замен

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



# Составление матриц замен



Маргарет Дэйхоф  
(1925 – 1983)

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

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**



David Lipman, director of the NCBI: “**She was the mother and father of bioinformatics**”.

# 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

$$S = \sum_k \log_2 \left( \frac{p_{i_k, j_k}}{f_{j_k}} \right)$$

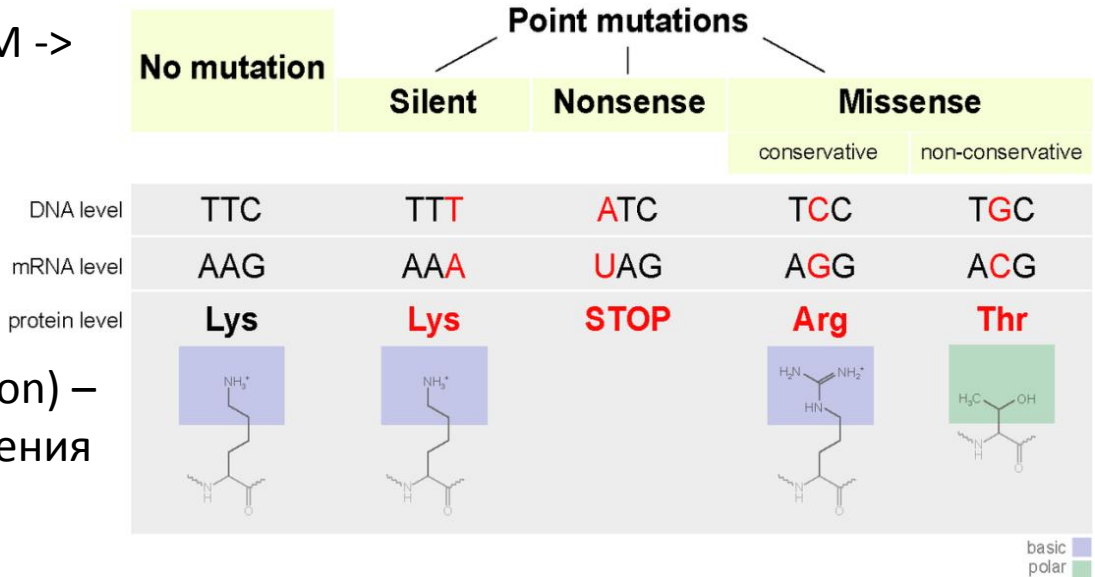
Оценка выравнивания (во избежание ошибок округления используют логарифмирование)



# 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 – 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 получают **матричным умножением**.

Матрица PAM250 предполагает не более 250 мутаций на 100 остатков (M.O.Dayhoff, 1978).

# Матрица вероятностей замен РАМ1

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Λ	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

(р · 10 000)

# Составление оценочных матриц для PAM

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

(упрощенно)

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		
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W		

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

(M.O.Dayhoff, 1978)

# РАМ. Что насчет консервативности?

<b>РАМ</b>	<b>0</b>	<b>1</b>	<b>30</b>	<b>80</b>	<b>110</b>	<b>200</b>	<b>250</b>
~ % идентичности	100	99	75	50	60	25	20

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

- 1) Как оценить консервативность, если матрицы РАМ2...РАМ250 получены чисто теоретически?
- 2) Верны ли числа в этой таблице?

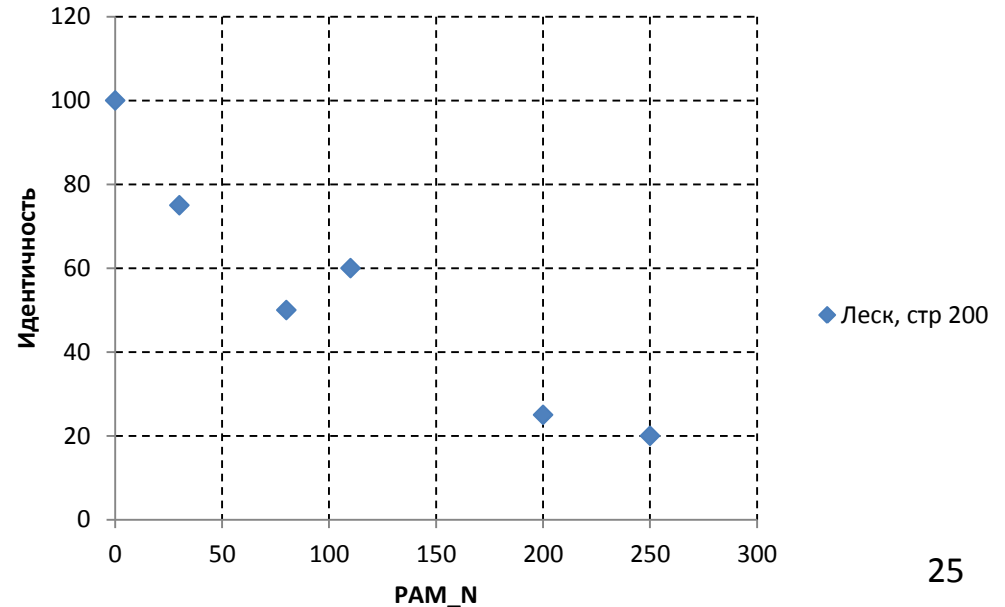


# РАМ. Что насчет консервативности?

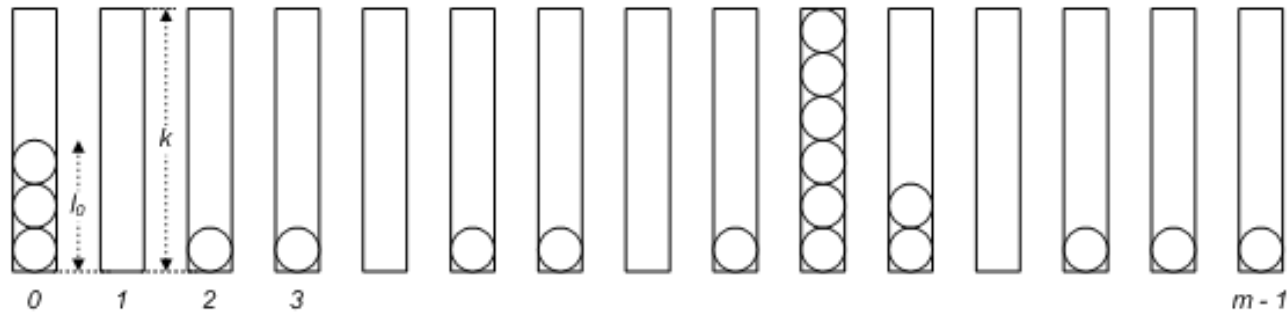
РАМ	0	1	30	80	110	200	250
~ % идентичности	100	99	75	50	60	25	20

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

- 1) Как оценить консервативность, если матрицы РАМ2...РАМ250 получены чисто теоретически?
- 2) Верны ли числа в этой таблице?



# РАМ. 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

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

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

# РАМ. Что насчет консервативности?

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

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

$N$  – длина последовательности

$K$  – общее число мутаций

$M$  – число позиций с мутациями

$E(M)$  – математическое ожидание  $M$

<b>РАМ</b>	<b>0</b>	<b>1</b>	<b>30</b>	<b>80</b>	<b>110</b>	<b>200</b>	<b>250</b>
~ % идентичности	100	99	74	45	33	13	8
	100	99	75	50	60	25	20

# РАМ. Что насчет консервативности?

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

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

$N$  – длина последовательности

$K$  – общее число мутаций

$M$  – число позиций с мутациями

$E(M)$  – математическое ожидание  $M$

РАМ	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

	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	

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**

# 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 ☹

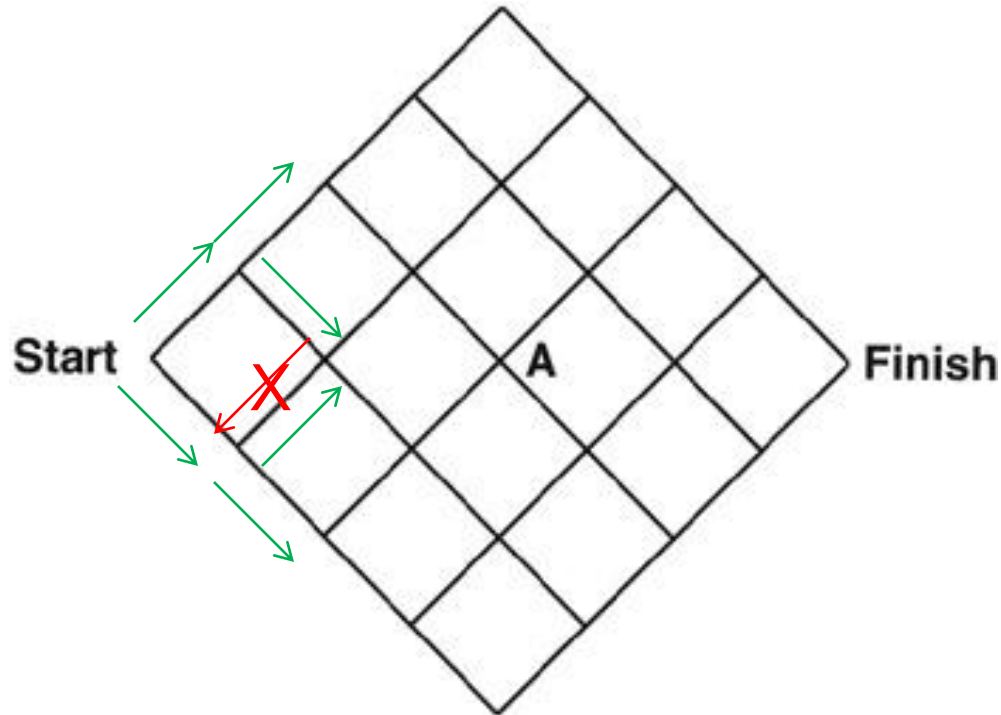
┌  
| - аналоги PAM по данным на 1992 год  
└

┌  
|  
| 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?

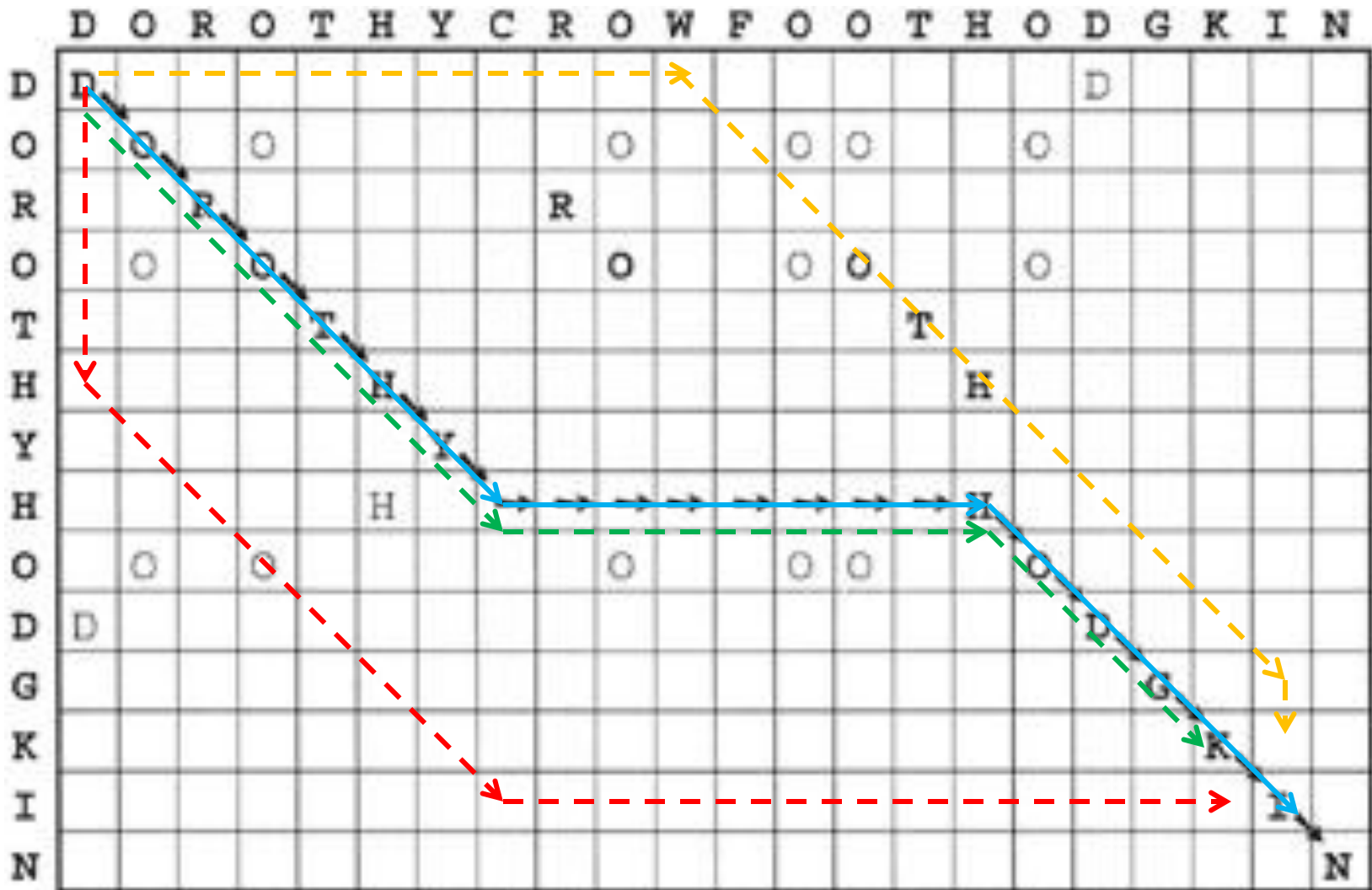
Сколько путей надо оценить, чтоб выбрать лучший?



**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								

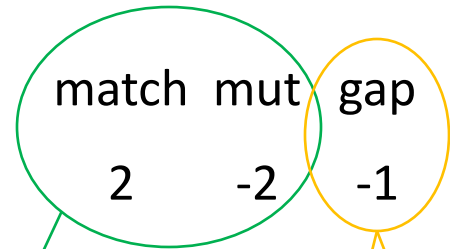
Базис:

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

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

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

	"-"	g	c	t	g	a	a	c	g	match	mut	gap
"-"	0	-1	-2	-3	-4	-5	-6	-7	-8	2	-2	-1
c	-1	<b>-2</b>	1	0	-1	-2	-3	<b>-4</b>	-5			
t	-2											
a	-3											
t	-4		<b>-2</b>									
a	-5											
a	-6											
t	-7											
c	-8											

$$-2 = \max(-1-1, 0-2, -1-1)$$

$$-4 = \max(-3-1, -6+2, -7-1)$$

$$-g-$$

$$-c$$

$$-g$$

$$-c$$

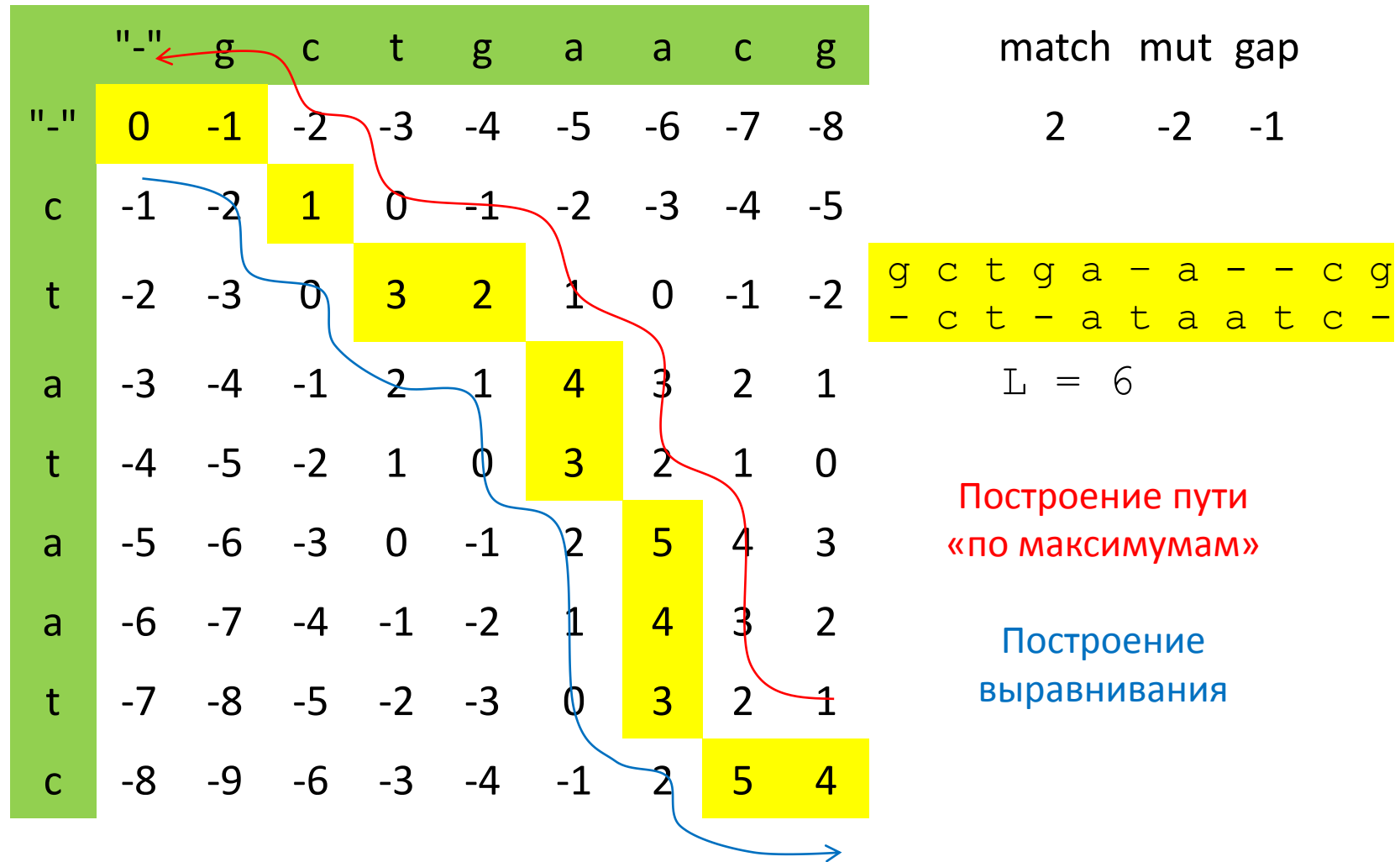
$$-g$$

$$c-$$

# Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

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

# Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша



# Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

	"-"	g	c	t	g	a	a	c	g		match	mut	gap
"-"	0	-1	-2	-3	-4	-5	-6	-7	-8		0	-1	-1
c	-1	-1	-1	-2	-3	-4	-5	-6	-7				
t	-2	-2	-2	-1	-2	-3	-4	-5	-6	g c t - g a a c g			
a	-3	-3	-3	-2	-2	-2	-3	-4	-5	- c t a t a a t c	L = 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				

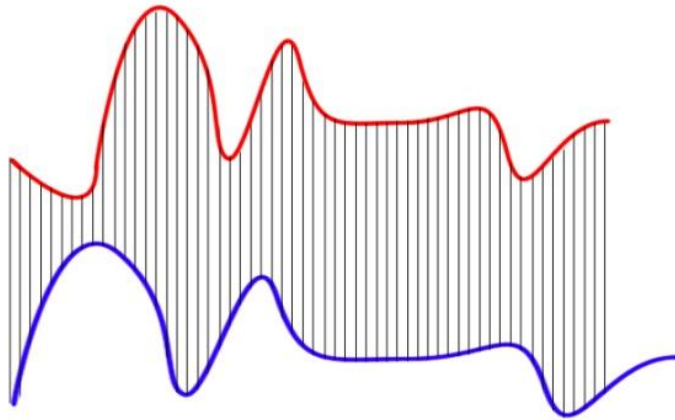
# Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

Варианты штрафов за вставку: постоянный, линейный, аффинный, логарифмический...

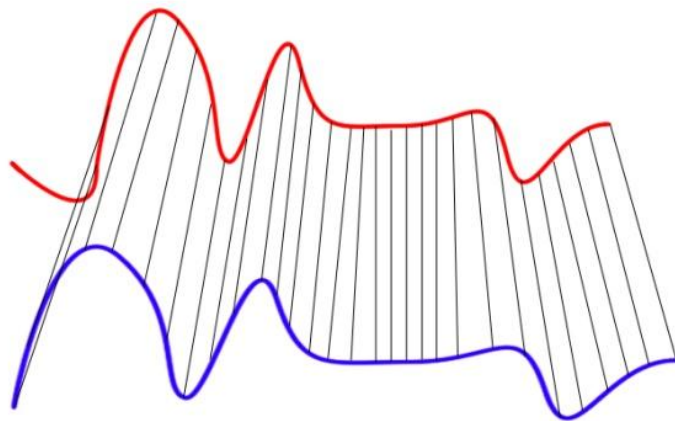
Построение глобального выравнивания **не годится** для поиска по БД (сотни тысяч последовательностей), т.к. требует  $(m \times n)$  времени и памяти для хранения матриц.

# Другие приложения динамического программирования

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



Euclidean Matching

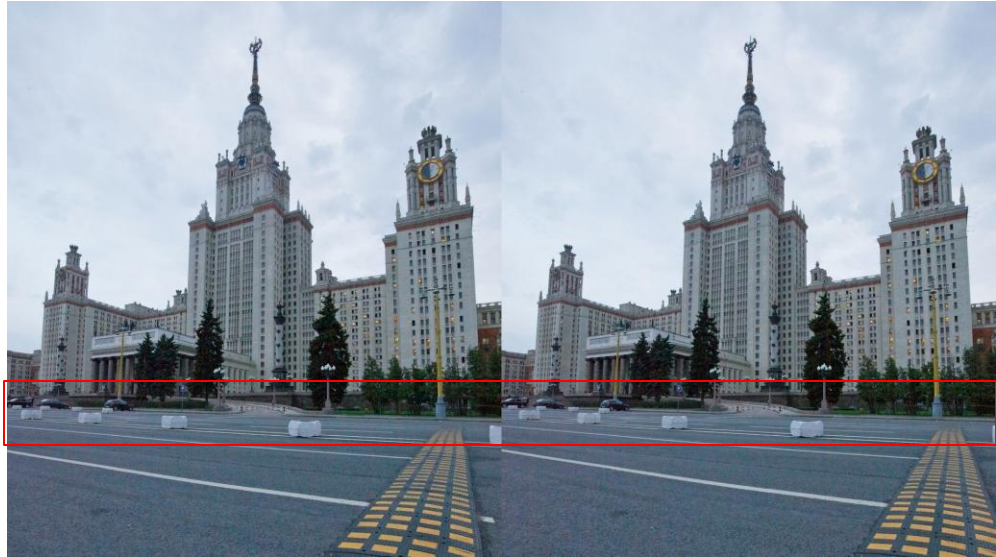


Dynamic Time Warping Matching



# Другие приложения динамического программирования

Компьютерное  
стереозрение



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