ВВЕДЕНИЕ В БИОИНФОРМАТИКУ

Лекция №3

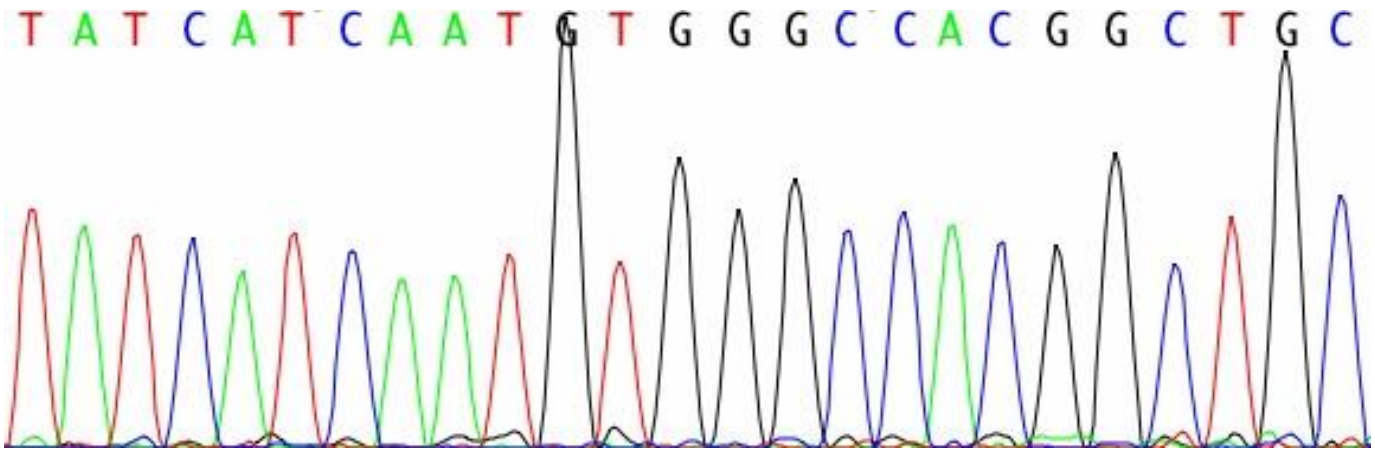
Сравнение последовательностей: матрицы сходства, динамическое программирование, локальное и глобальное выравнивание.
Матрицы PAM, Blosum.

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

Сайт курса http://intbio.org/bioinf2018

Представьте, что...

...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 вариабельные области
- Предположить эволюционные взаимосвязи

Выравнивание последовательностей – основной инструмент биоинформатики

Рассмотрим последовательности gctgaacg и ctataatc. Их можно выровнять:

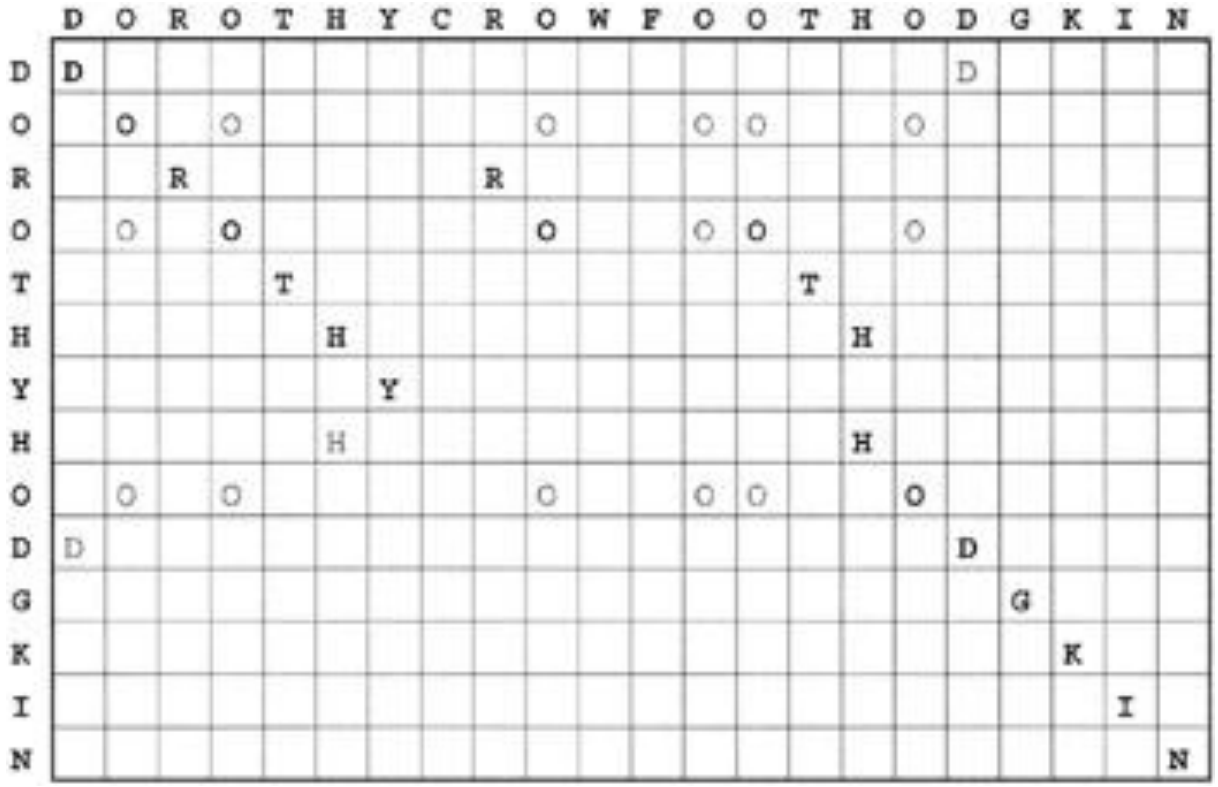
неинформативно	- - - - - - - g c t g a a c g c t a t a a t c - - - - - - - -
без пропусков	g c t g a a c g c t a t a a t c
с пропусками так	g c t g a - a - - c g - - c t - a t a a t c
или так	g c t g - a a - c g - c t a t a a t c -

Точечная матрица сходства (dotplot)

Рассмотрим две последовательности:

- DOROTHYHODGKIN
- DOROTHYCROWFOOTHODGKIN

Матрица сходства для них будет выглядеть так:



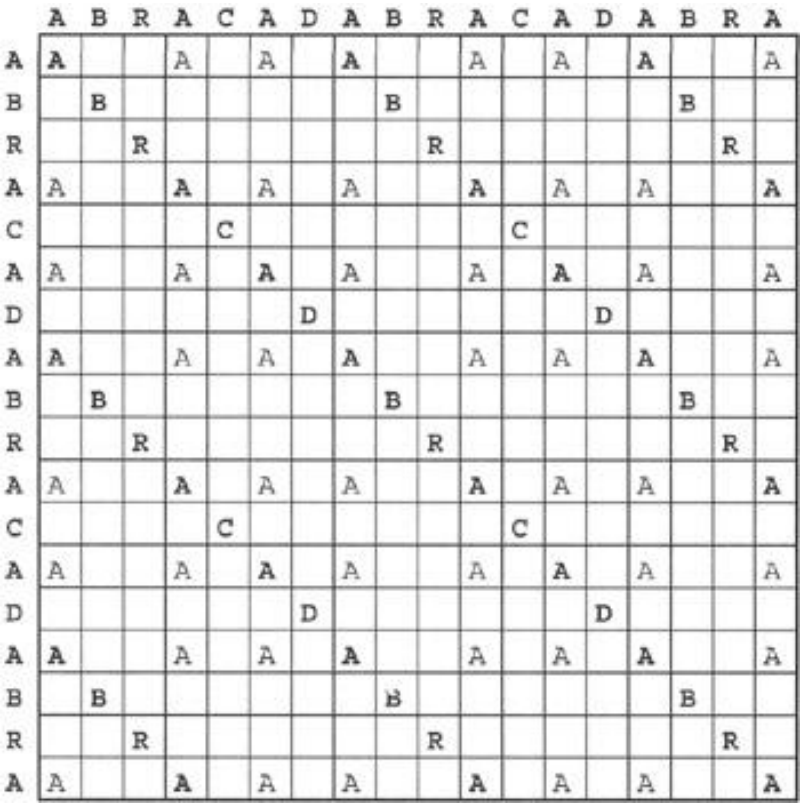
(1910 - 1994)

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

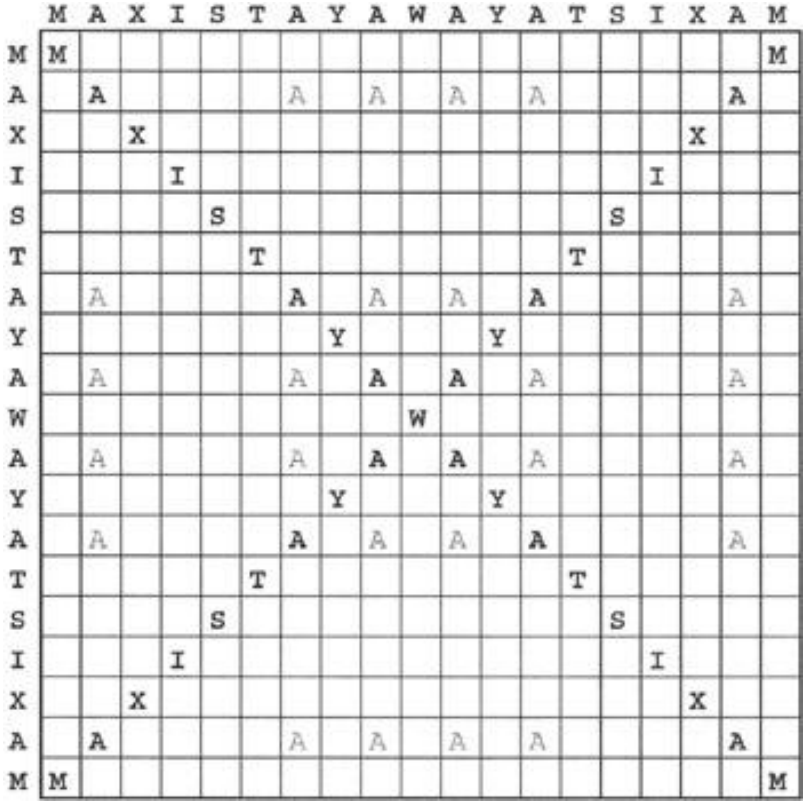
Нобелевский лауреат 1964 года

Точечная матрица сходства (dotplot)

Последовательность с повторами (ABRACADABRACADABRA):



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



Точечная матрица сходства (dotplot)

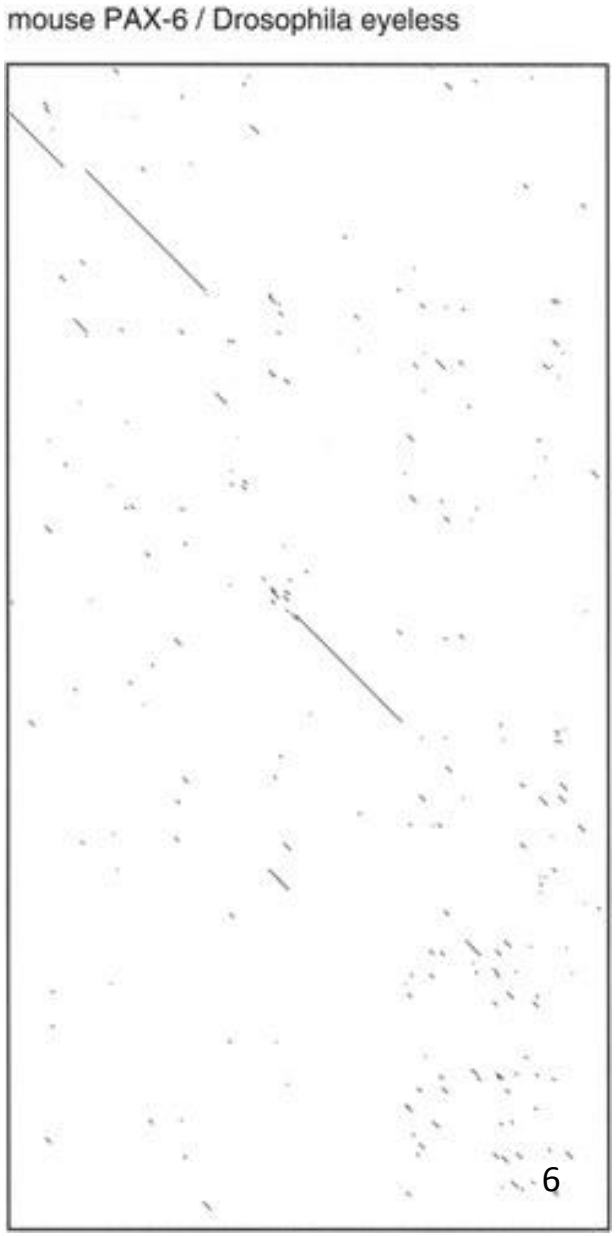
Для реальных последовательностей (белок PAX-6 мыши и белок eyeless плодовой мушки) матрица сходства имеет более сложный вид. Хорошо заметны три протяженных участка сходства.

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

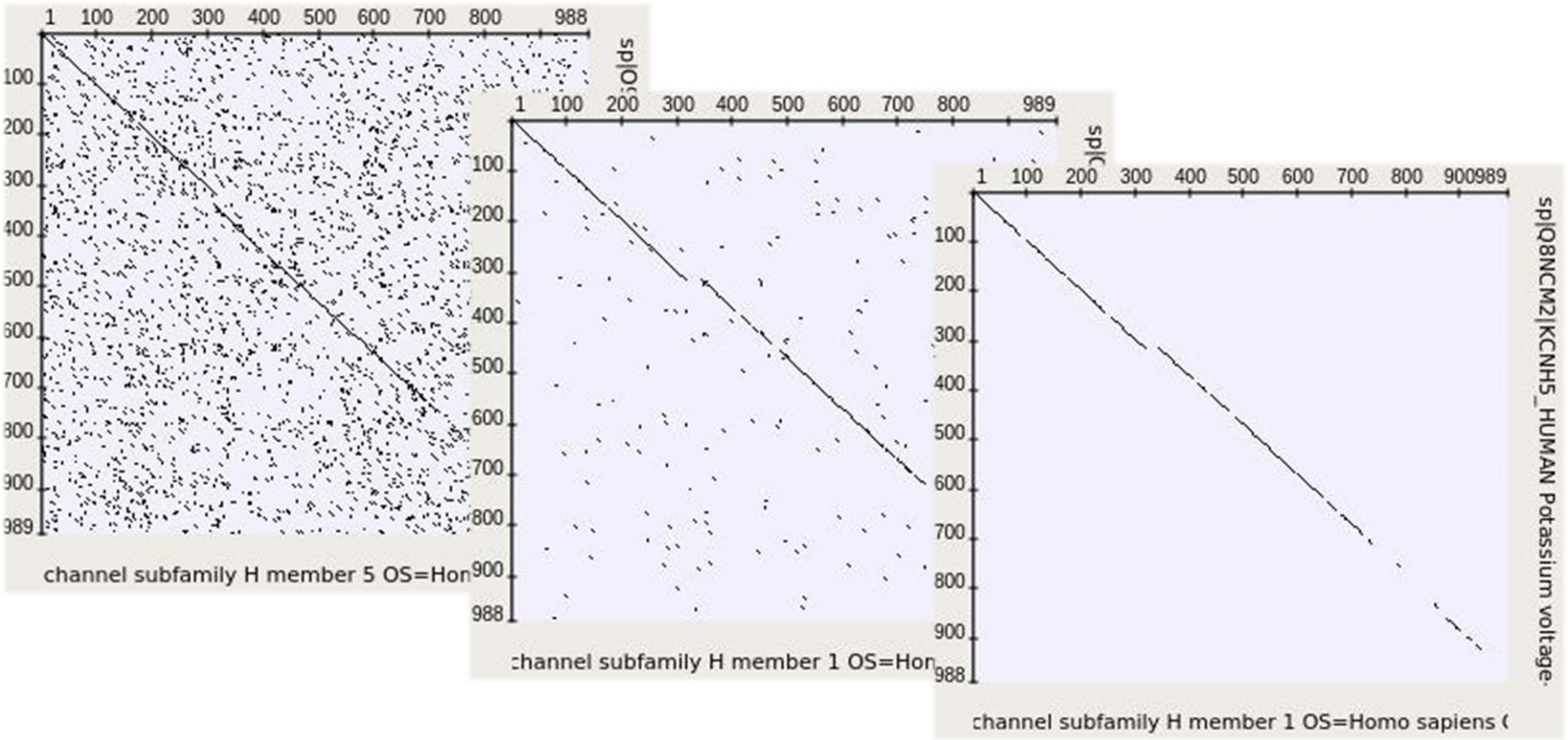
Интерактивное построение точечных матриц:

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

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



Точечная матрица сходства (dotplot)



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

Free bioinformatics software:

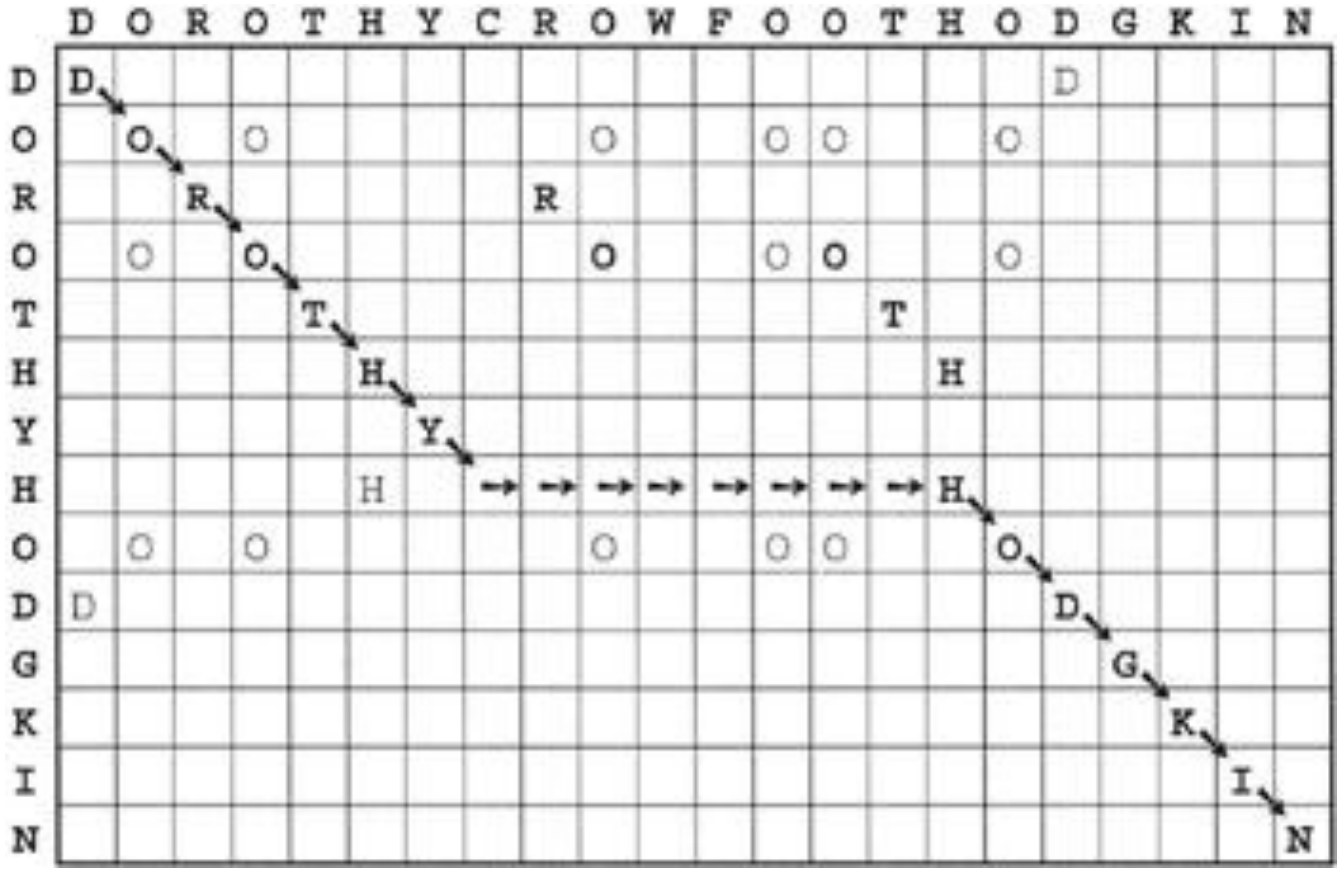
http://ugene.net

Unipro UGENE

Точечные матрицы и выравнивание последовательностей

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

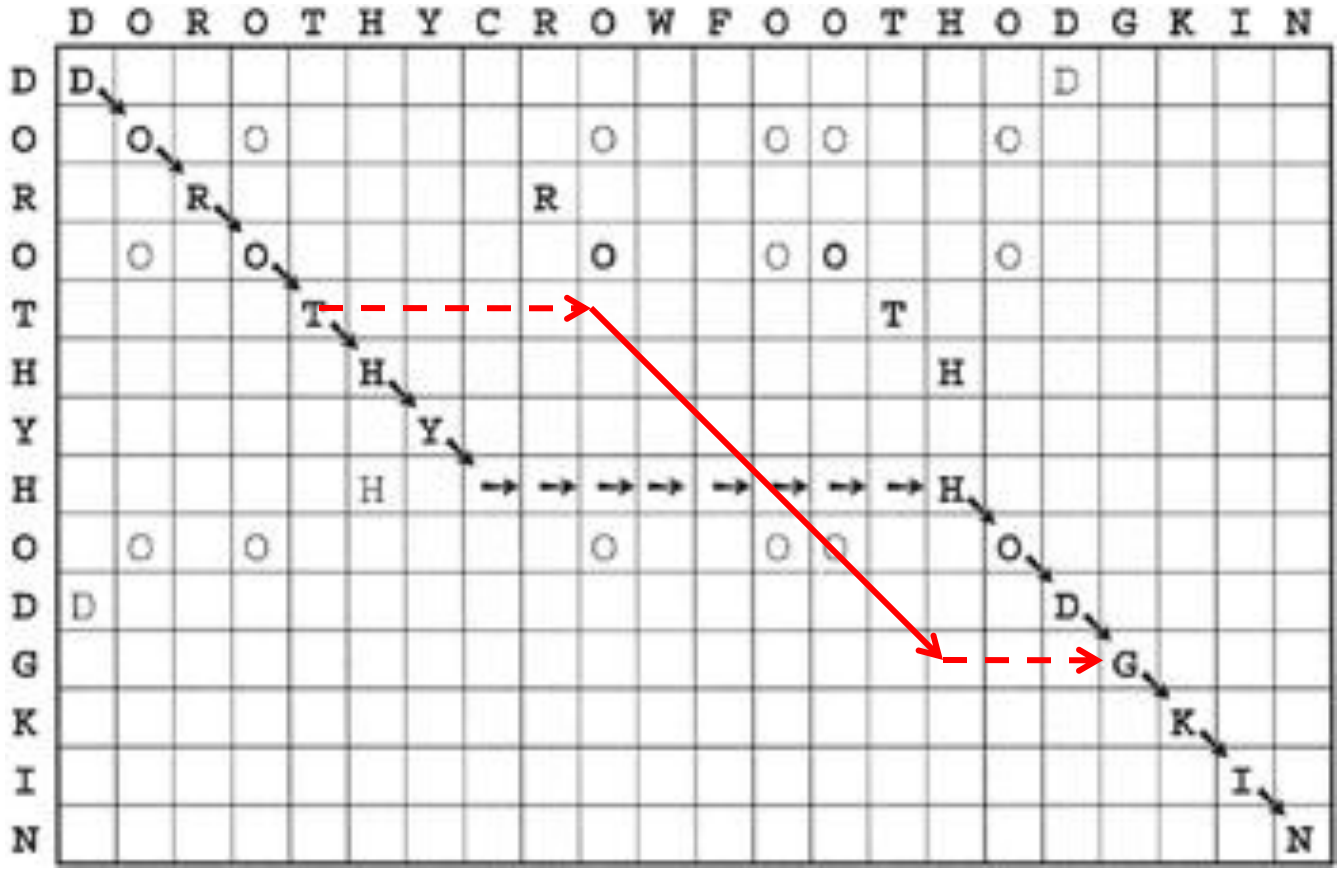
```
DOROTHY--------HODGKIN
DOROTHYCROWFOOTHODGKIN
```



Точечные матрицы и выравнивание последовательностей

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

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



Мера сходства последовательностей

Как выбрать оптимальное выравнивание из всех возможных? Нужна мера сходства.

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

(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).

(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 -	=	

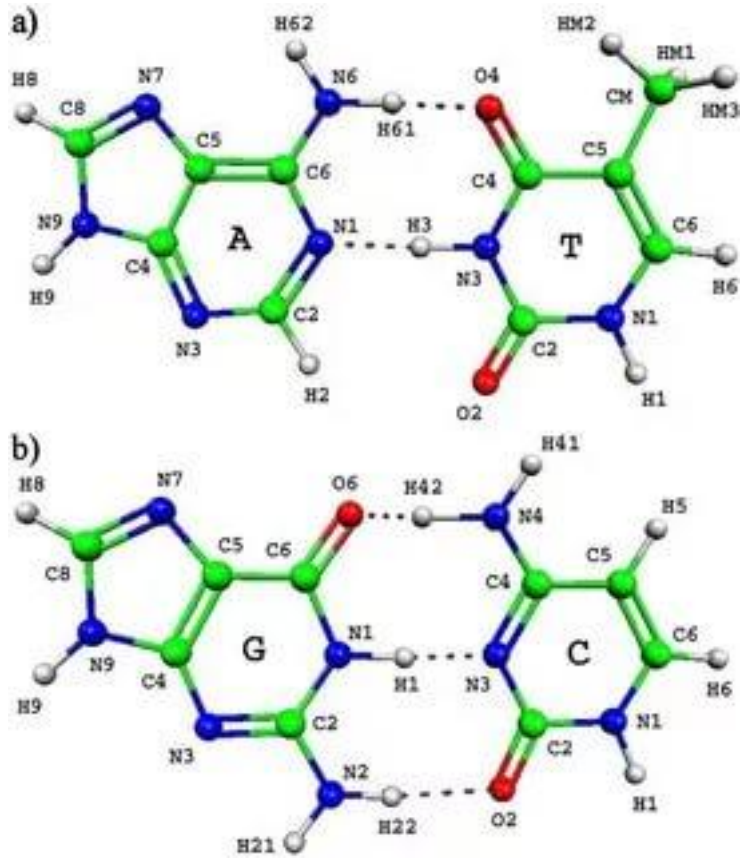
Мера сходства последовательностей

Выравнивание		Расстояние по Левенштейну
- - - - - - - g c t g a a c g 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:

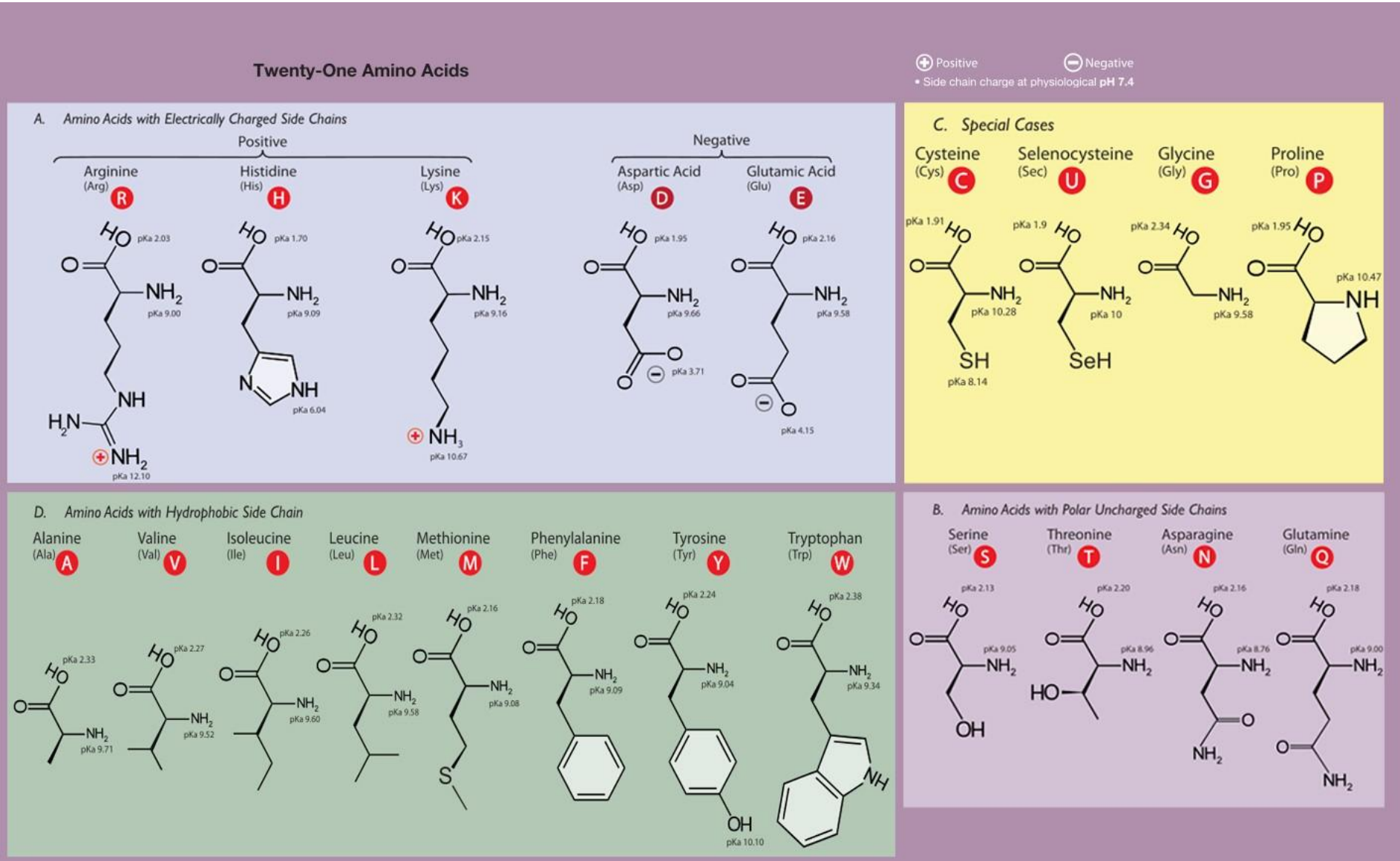
- Некоторые замены происходят вероятнее других (особ. аминокислотные);
- Совместная делеция нескольких остатков более вероятна, чем независимая;
- ...



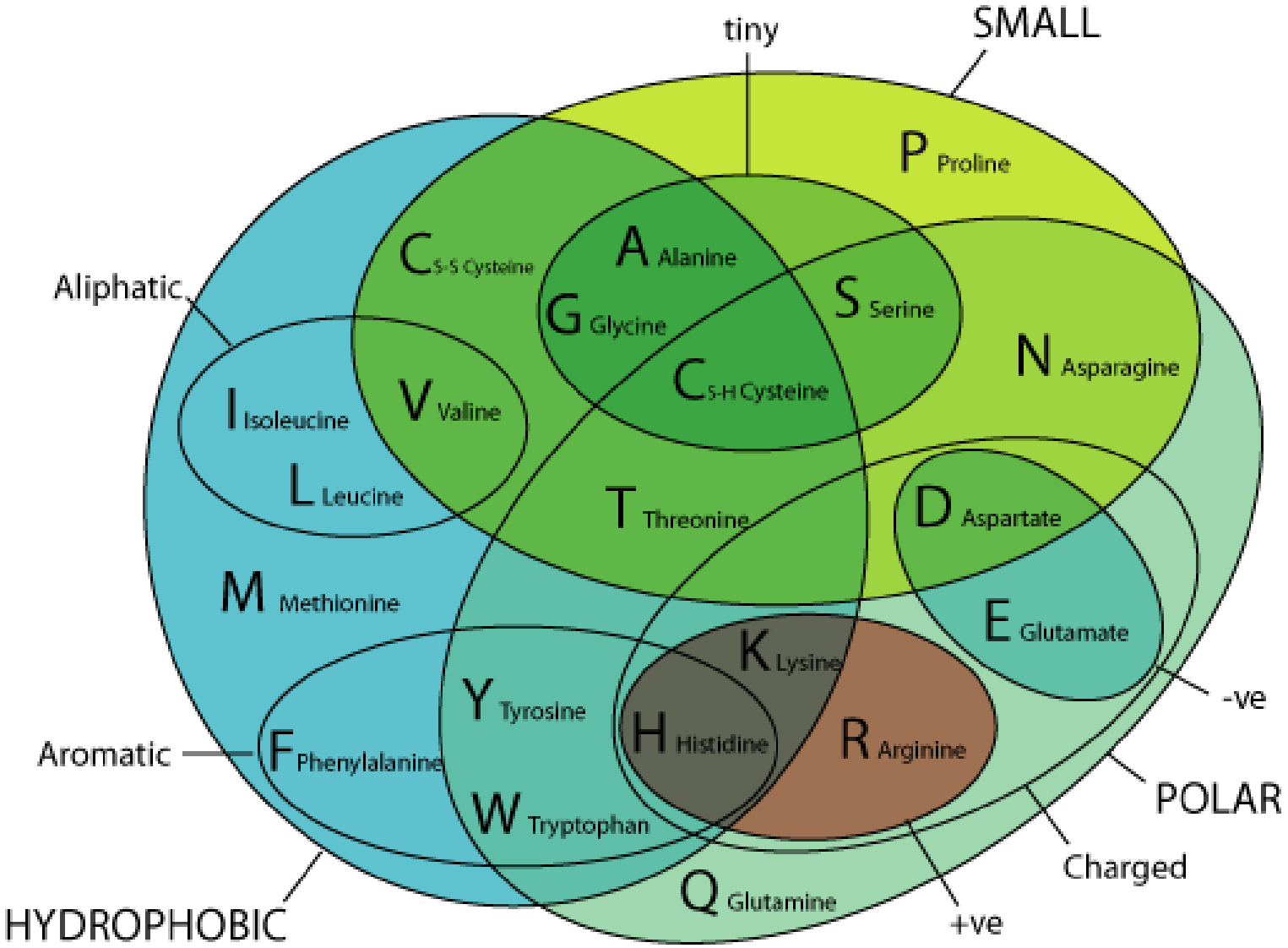
Одна из возможных матриц замен, учитывающая, что транзиции встречаются чаще трансверсий:

	a	g	t	c
a	20	0	−5	−5
g	0	20	−5	−5
t	−5	−5	20	0
c	−5	−5	0	20

Структура аминокислот



Свойства аминокислот



Составление матриц замен

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

2) «Эмпирический» – собираем статистические данные о встречаемости аминокислотных замен в известных белках

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

Первой начала использование компьютеров для сравнения последовательностей

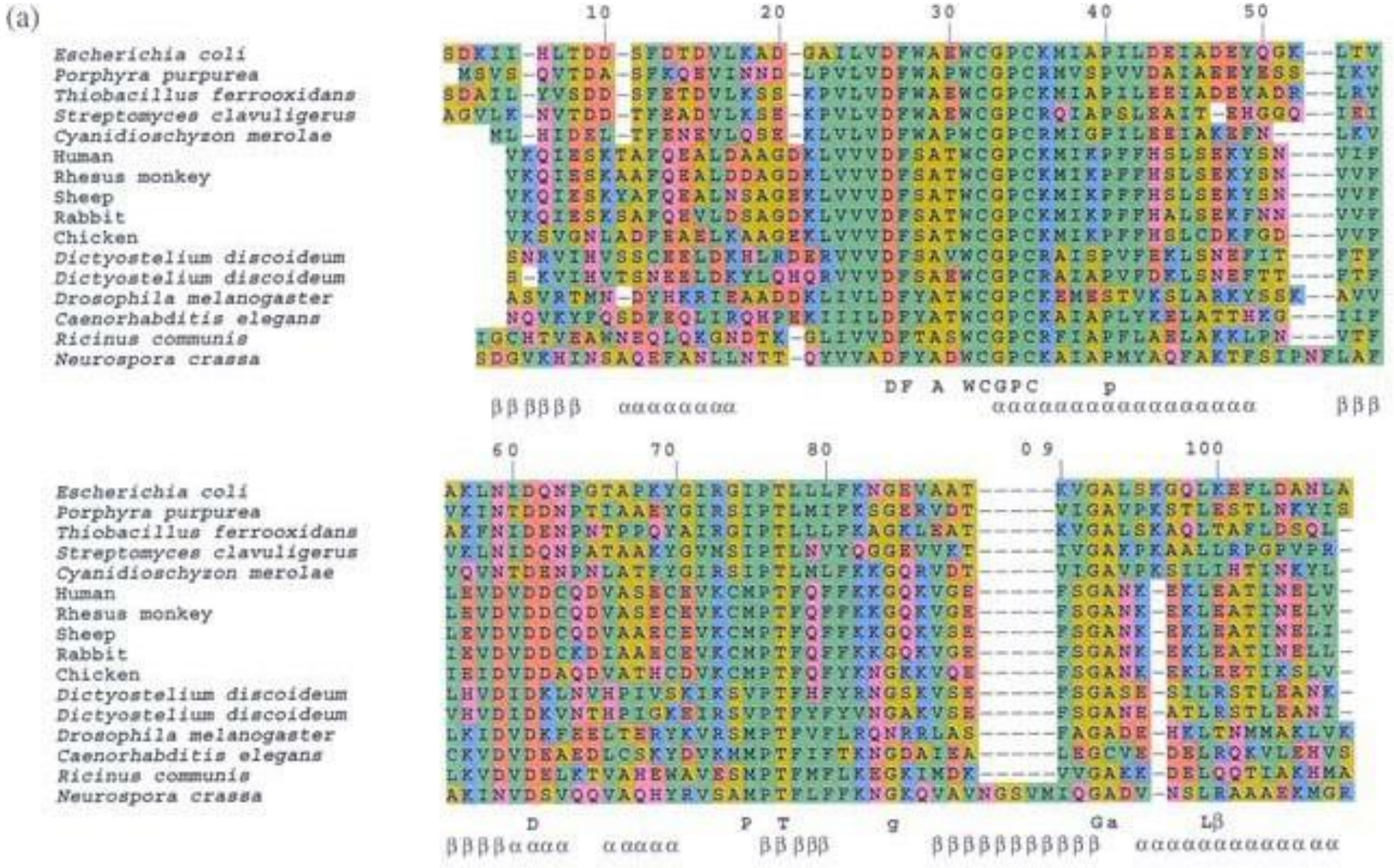
Предложила однобуквенный код для аминокислот

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

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

Маргарет Дэйхоф
(1925 – 1983)

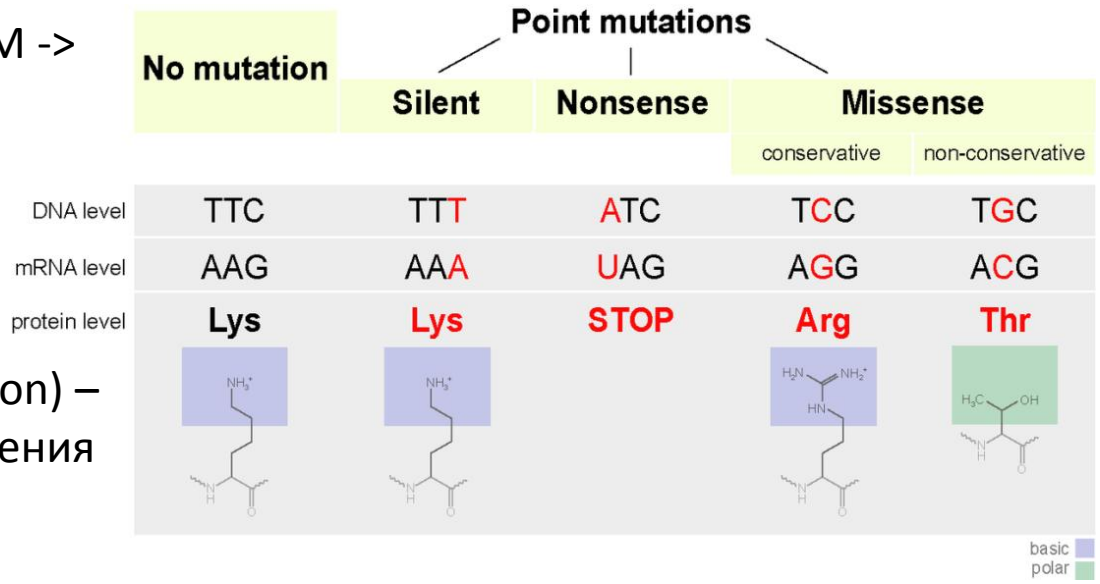
Пример множественного выравнивания



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

Составление матриц замен. 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

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?

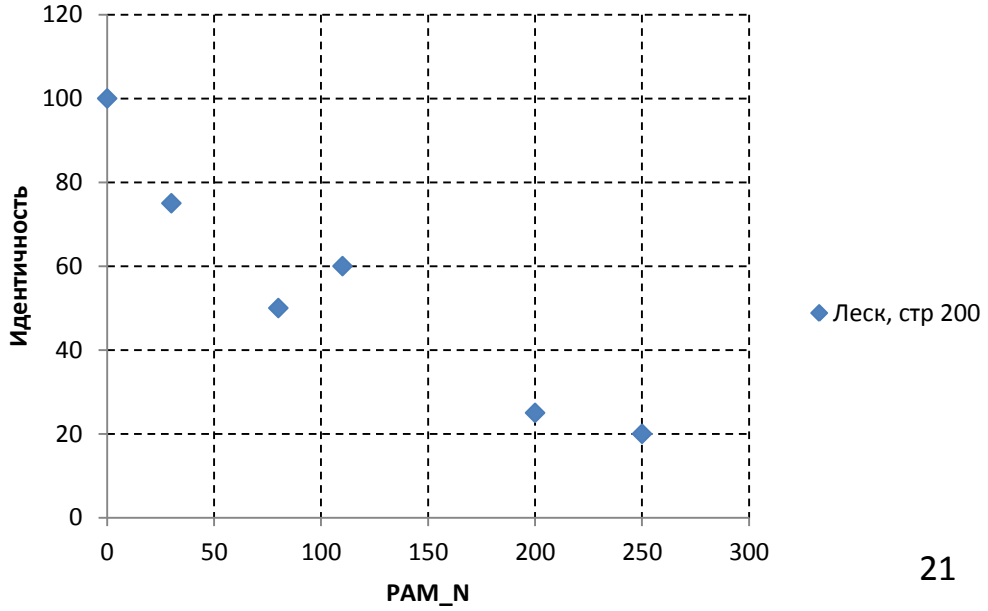
PAM. Что насчет консервативности?

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

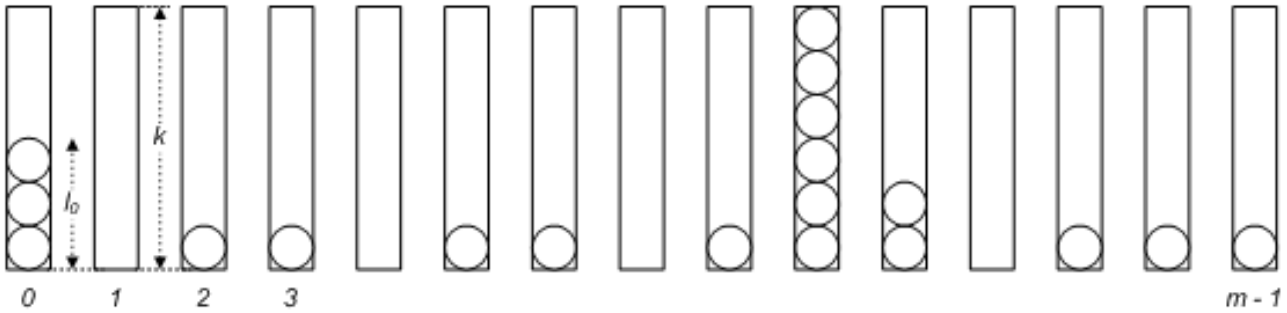
(А. Леск. Введение в биоинформатику (2009), стр. 200)

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

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

N – длина последовательности
K – общее число мутаций
M – число позиций с мутациями
E(M) – математическое ожидание M

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

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

N – длина последовательности
K – общее число мутаций
M – число позиций с мутациями
E(M) – математическое ожидание M

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

BLOSUM62 (S. Henikoff and J.G. Henikoff, 1992)

	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

Контроль посещаемости сегодня

Заполнить форму по адресу

http://intbio.org/3

возможность закрывается сразу после перерыва

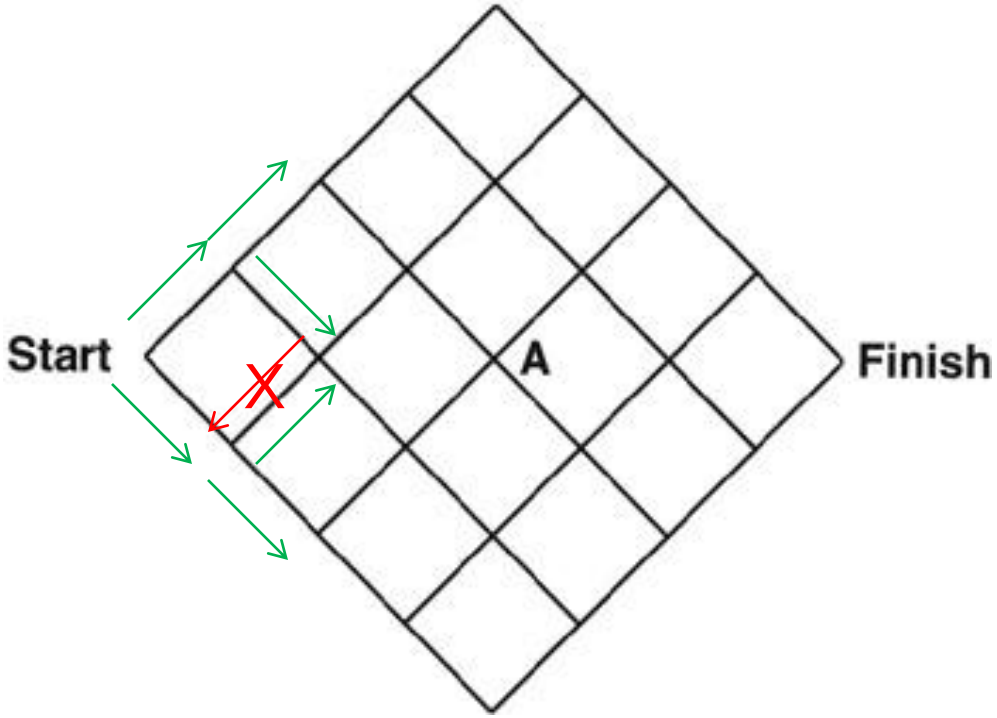
либо

записаться в список на перерыве

Расчет выравнивания двух последовательностей

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

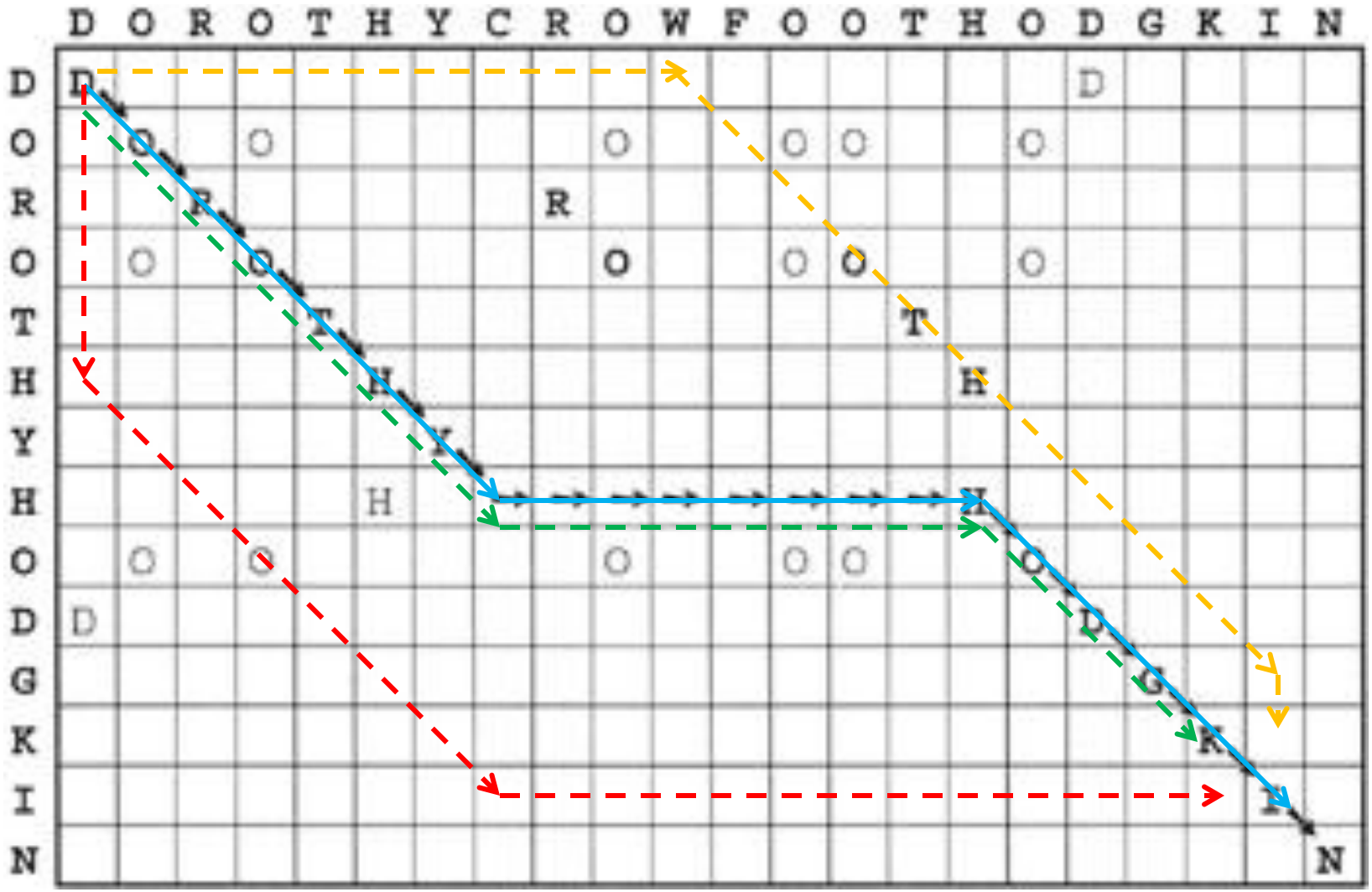
Сколько путей надо оценить, чтоб выбрать лучший?



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

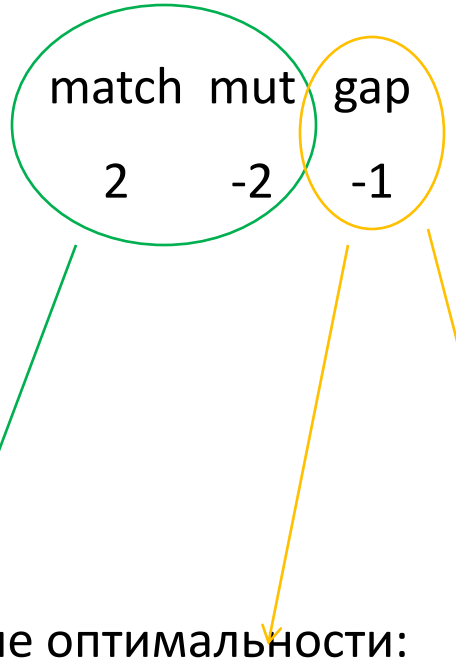
Поиск глобального выравнивания

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:

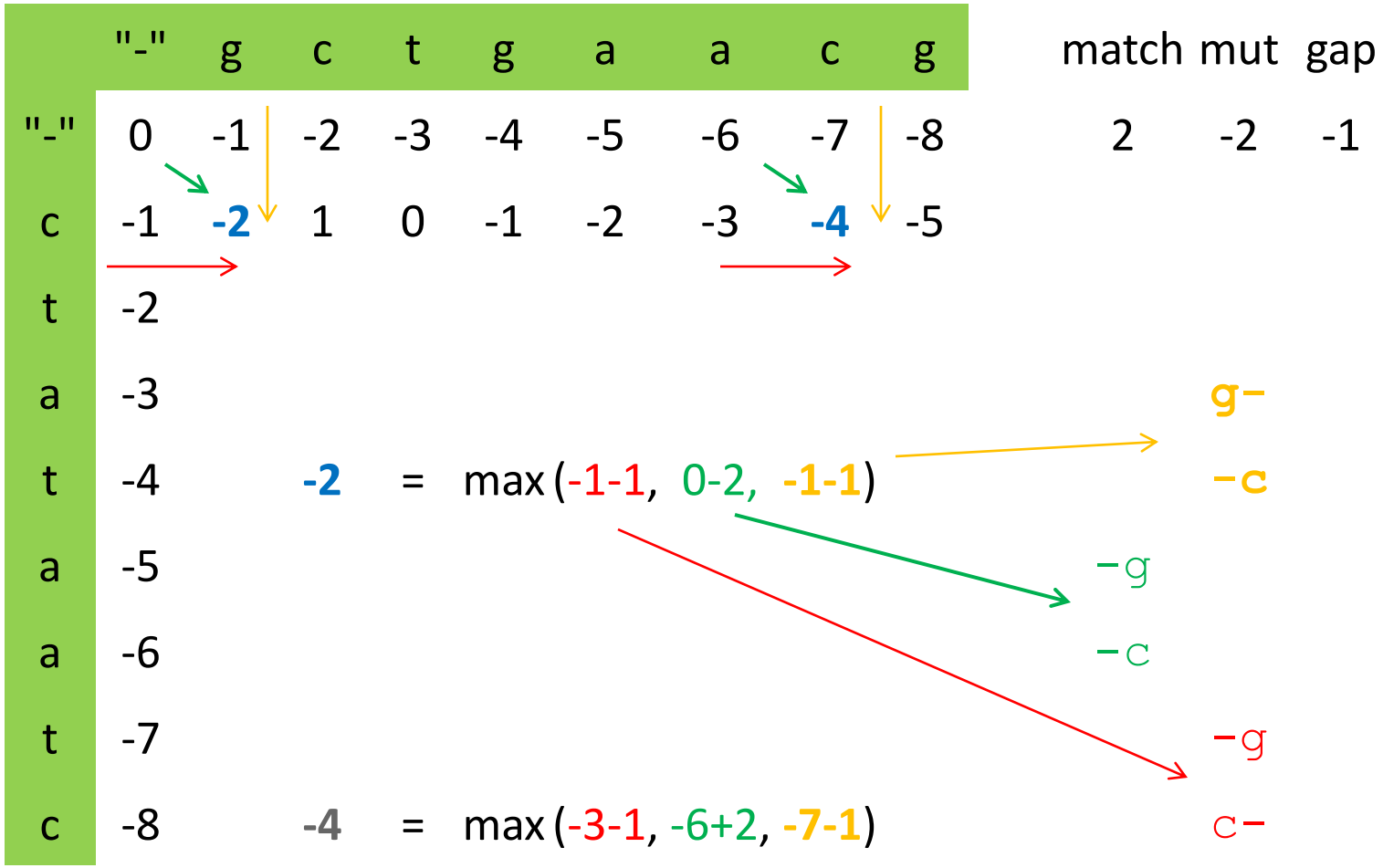
$$F_{i0} = d \cdot i$$

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

Рекурсия, основанная на принципе оптимальности:

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

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша



Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

	"-"	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

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

	"-"	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

Расчет выравнивания двух последовательностей. Алгоритм Нидлмана-Вунша

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

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

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

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

Другие приложения:

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

Б) Компьютерное стереозрение