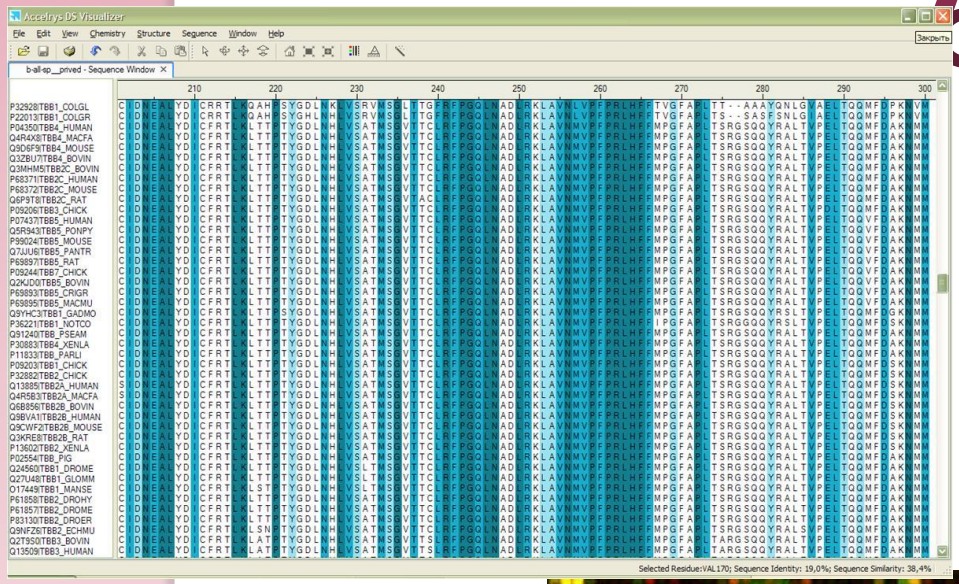
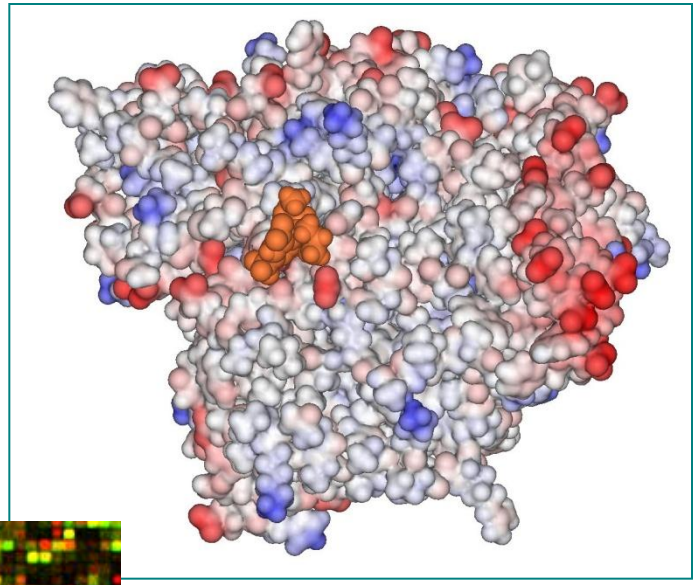
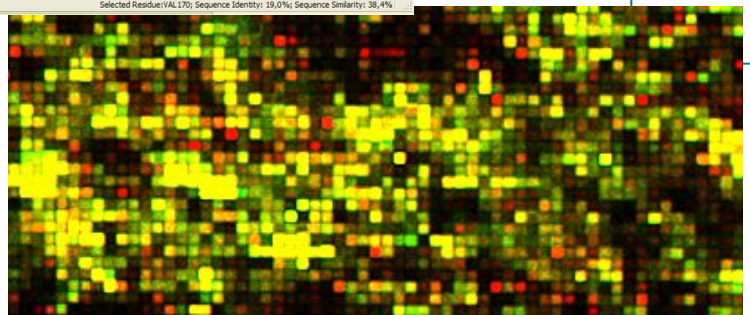
Київський національний університет ім. Тараса Шевченка
Інститут високих технологій

БІОІНФОРМАТИК

к.б.н. Нилорко О.А.







Парное выравнивание

Основний спосіб визначити схожість двох послідовностей - вирівняти їх

```
>EC_Tr  :  MQNRLTI KDI ARLSGVGKSTVSRVLNNEYR
>EC_Fr  :  MKLDEI ARLAGVSRTTASYVI NGKAKQYR
```

- При аналізі первинних структур процедура вирівнювання виявляє сходство між послідовностями (**sequence similarity**), яке може свідчити про гомологію **(homology)**, тобто еволюційну спорідненість макромолекул.

Геп – пропуск в послідовності

```
>EC_Tr  :  MQNRLTIKDIARLSGVGKSTVSRVLNNE---YR
>EC_Fr  :  ----MKLDEIARLAGVSRTTASYVINGKAKQYR
```

Гомологичные последовательности – последовательности, имеющие общее происхождение (общего предка).

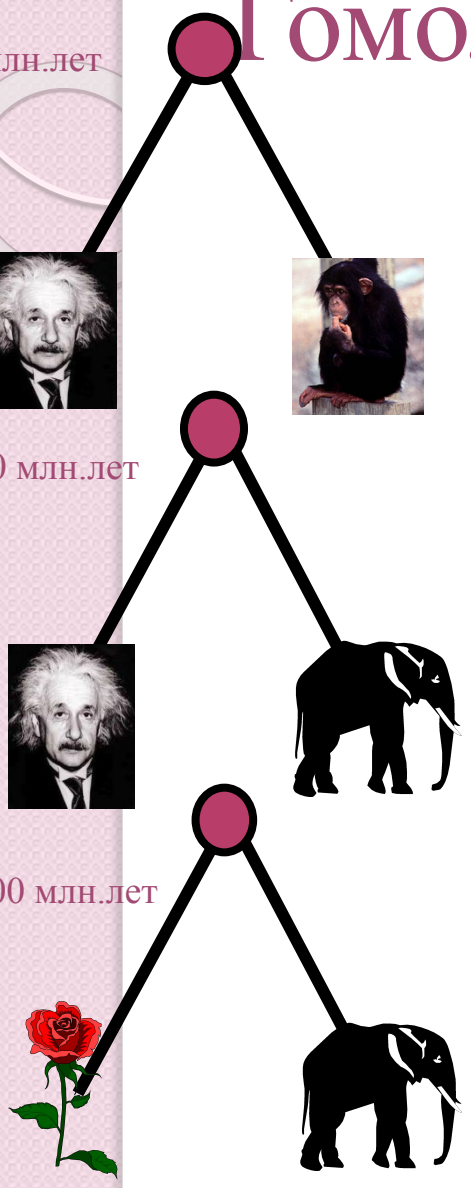
Признаки гомологичности белков

сходная 3D-структура

в той или иной степени похожая аминокислотная последовательность

- **разные другие соображения…**

Гомологи (?)



Усе живе походить від одного загального предка, отже, усі послідовності є «гомологами».

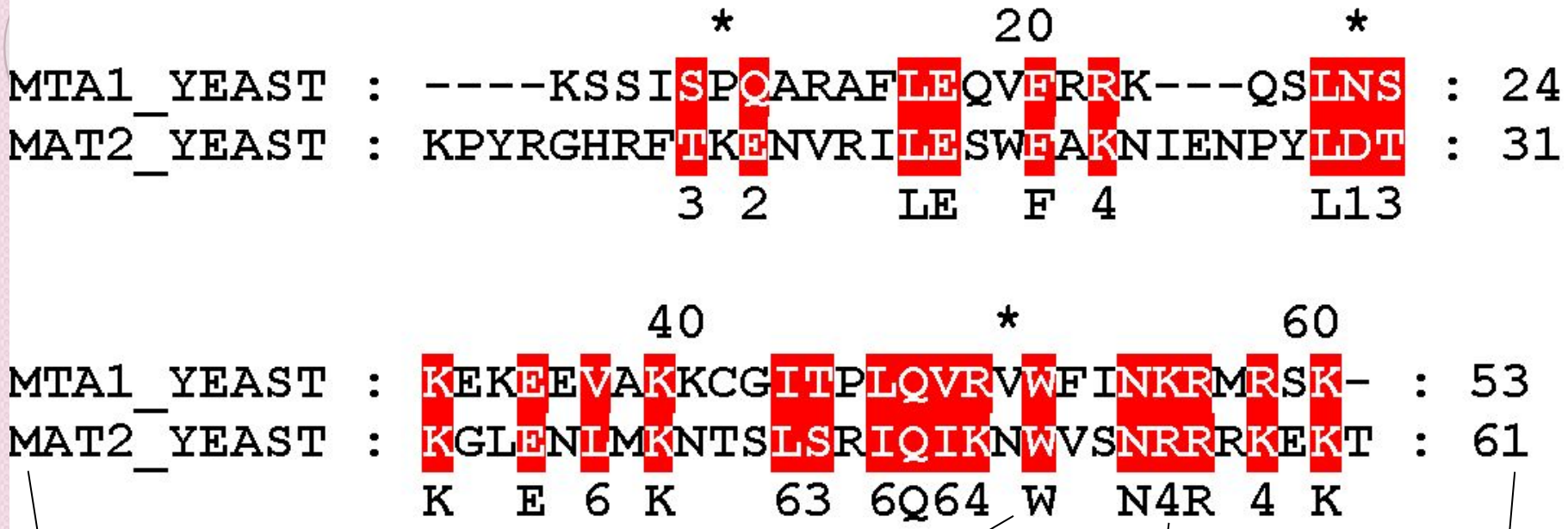
Насправді гомологи – тільки ті послідовності, подібність яких можна підтвердити існуючими методами з певною чутливістю:

Білок у двох різних організмах виконує подібну функцію й це можна підтвердити експериментально.

Определение

Выравнивание (alignment) – сравнение двух (парный) или нескольких (множественный) последовательностей. Поиск серий идентичных символов в последовательностях

Что изображено?



Номер столбца выравнивания

Название последовательности

Консервативный остаток

Функционально консервативная позиция

Номер последнего в строке остатка ИЗ ЭТОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

«Идеальное» выравнивание – запись последовательностей одна под другой так, чтобы гомологичные фрагменты оказались друг под другом.

```
      домовой
скупидом
    водомерка ?
```

Гэп – пропуск в последовательности

```
лесовоз      ?--лесо---воз
ледоход       лед---оход---
```

1	підберезовик підосиновик-	
2	підберезовик -підосиновик	
3	підберезовик підосин-овик	
4	підберезовик під-осиновик	
5	підберез----овик під-----осиновик	

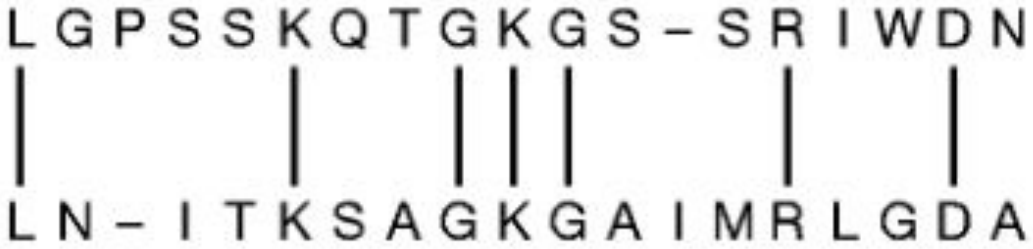
Какие задачи решает парное выравнивание?

- Нуклеотиды
  - Изучение эволюционных связей
  - Поиск генов, доменов, сигналов …
- Белки
  - Изучение эволюционных связей
  - Классификация белковых семейств по функции или структуре
  - Идентификация общих доменов по функции или структуре.

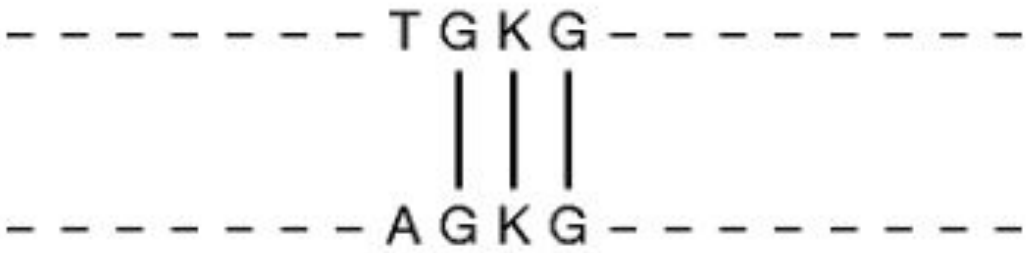
Парное выравнивание – методы сравнения

- Глобальное выравнивание – находит лучшее решение для целых последовательностей.

- Локальное выравнивание – находит похожие районы в двух последовательностях.



Global alignment



Local alignment


Біологічна задача
Формалізація
Алгоритм
Алгоритм
Параметри
Параметри
Тестування
Визначення області застосуємості

Пример: сравнение последовательностей

- ***Тестирование:*** алгоритм должен распознавать последовательности, для которых известно, что они биологически (структурно и/или функционально) сходны

Формалізація задачі

- через визначення **редакційної відстані**
- через визначення **ваги вирівнювання**.

Редакционное расстояние

- ***Элементарное преобразование последовательности:*** замена буквы или удаление буквы или вставка буквы.
- ***Редакционное расстояние:*** минимальное количество элементарных преобразований, переводящих одну последовательность в другую.
- Формализация задачи сравнения последовательностей: найти редакционное расстояние и набор преобразований, его реализующий

Вага вирівнювання (alignment **score**)

Якість співставлення двох послідовностей може бути описана за допомогою певного чисельного критерія. Цей критерій дістав назву ва вирівнювання (англ. alignment score) і представляє собою суму позитивних і штрафних балів (числових коефіцієнтів), які нараховуються в залежності від того, які символи (залишки) опиняються в тій самій позиції вирівнювання. В загальному вигляді ва вирівнювання може бути обрахована як:

+ Кількість балів за ідентичні залишки
+ Кількість балів за подібні залишки
+ Кількість балів за неспівпадаючі залишки
– Кількість балів за відкриття проміжка
– Кількість балів за продовження проміжка

Сумарна вага вирівнювання

Вычисление наилучшего выравнивания путем прохождения по Dot matrix для двух белков по 300 аминокислот требует 10^88 сравнений

Парное выравнивание

Человеческий гемоглобин (HH):

VLSPADKTNVKAAWGKVGAHAGYEG

Миоглобин кашалота (SWM):

VLSEGEWQLVLHVWAKVEADVAGHG

Парное выравнивание - идентичность

```
(HH)   VLSPADKTNVKAAWGKVGAHAGYEG
       |||       |    | || |      |
(SWM)  VLSEGEWQLVLHVWAKVEADVAGHG
```

Процент идентичности: 36.000

Парное выравнивание - сходство

```
(HH)  VLSPADKTNVKAAWGKVGAHAGYEG
      |||  .   |   | ||  |     |
(SWM) VLSEGEWQLVLHVWAKVEADVAGHG
```

Процент похожести: 40.000 (| и .)

Процент идентичности: 36.000 (только |)

Парное выравнивание – вставка промежутков (gaps)

```
(HH)   VLSPADKTNVKAAWGKVGAH-AGYEG
        |||   .     |    | || | ||   |
(SWM)  VLSEGEWQLVLHVWAKVEADVAGH-G
```

- Gap Weight: 4
- Gaps: 2
- Процент похожести: 54.167
- Процент идентичности: 45.833

Парное выравнивание – вставка промежутков

```
AKWTNLK----WAKV-ADVAGH-G
|| ||  |    | ||  | |||  |
AK-TNVKAKLPWGKVGAHVAGEYG
```

- вставка\удаление промежутка
- продление промежутка

Парное выравнивание - Scoring

```
(HH)   VLSPADKTNVKAAWGKVGAH-AGYEG
       |||   .    |    | ||      ||   |
(SWM)  VLSEGEWQLVLHVWAKVEADVAGH-G
```

Final score:

(V,V) + (L,L) + (S,S) + (D,E) + …

- (penalty for gap insertion)*(number of gaps)
- (penalty for gap extension)*(extension length)

Парное выравнивание

- Алгоритмы парного выравнивания пробуют все возможные варианты выравнивания.
- Результат – выравнивание с наивысшей оценкой.
- Различные системы оценки дают разные лучшие выравнивания!!!

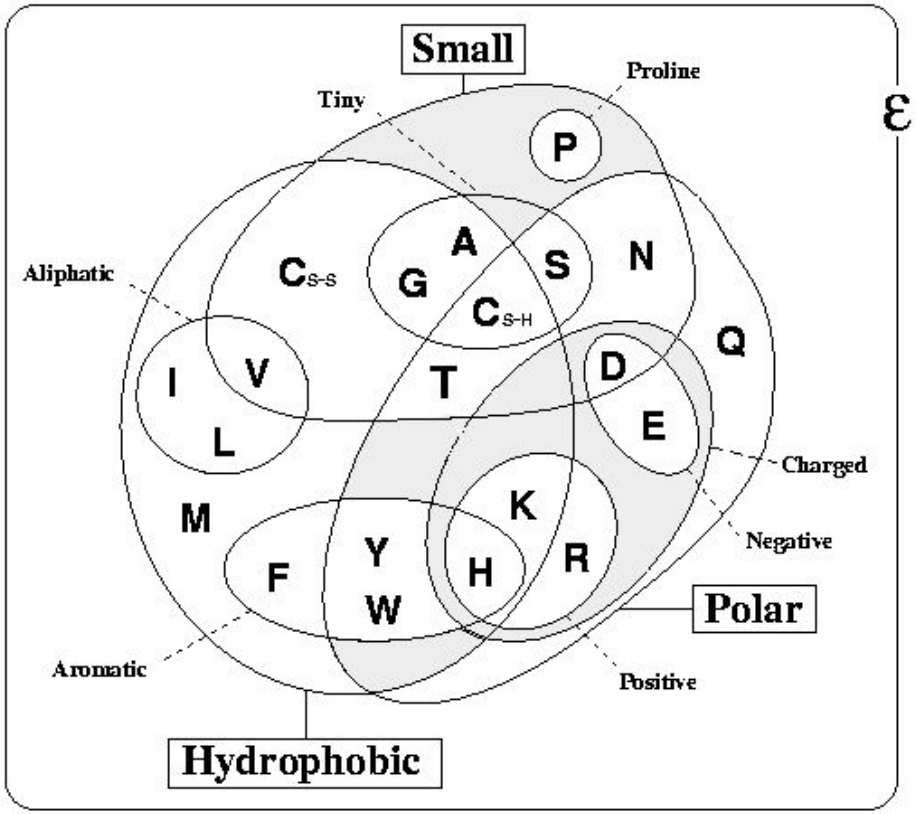
Система оценки - белки

- **Идентичность**: подсчитывается количество совпадений и делится на длину выравниваемого региона
- **Similarity**: Менее формализованная величина

Amino Acid	**Category**
Asp (D) Glu(E) Asn (N) Gln (Q)	Кислоты\амиды
His (H) Lys (K) Arg (R)	Основания
Phe (F) Tyr (Y) Trp (W)	Ароматические
Ala (A) Cys (C) Gly (G) Pro (P) Ser (S) Thr (T)	Гидрофильные
Ile (I) Leu (L) Met (M) Val (V)	Гидрофобные

Система оценки - белки

Сходство: Положительная оценка для выравниваемых аминокислот из одной и той же группы.



Парное выравнивание

Весовые матрицы (матрицы для оценки) – PAM, BLOSUM, Gonnet

Системы оценки выравнивания различны для белков и для ДНК\РНК

Margaret Oakley Dayhoff

1972 год

Сформулировала первую вероятностную модель эволюции белков

Матрицы сравнения белков

Семейство матриц, которые отражают вероятность замены одной аминокислоты на другую во время эволюции.

PAM = Point Accepted Mutation

Набор матриц, которые используются для выравнивания аминокислотных последовательностей белков

Substitution Matrix

Матрица 20X20: в узлах – вероятности замены одной аминокислоты на другую

Еволюція терміна APM/PAM

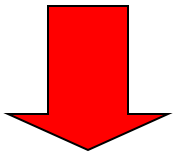
зафіксовані (прийняті) точкові мутації (accepted point mutation), тобто амінокислотні заміни, що закріпилися в процесі еволюції відповідних білкових послідовностей. Абревіатура APM згодом була трансформована у PAM, яка в деяких випадках розшифровується дослідникам як percent accepted mutation (процент зафіксованих мутацій) У одиницях PAM виміряють еволюційну відстань між амінокислотними послідовностями (кількість PAM на 100 амінокислотних залишків), а кількість PAM за певний проміжок часу (зазвичай на 100 млн. років) є показником швидкості еволюційних змін, що відбуваються в білковому ланцюзі.

PAM матрица

- PAM единицы отображают эволюционную дистанцию.
- 1 PAM единица – вероятность 1 точечной мутации на 100 аминокислот.
- Умножение PAM 1 на себя даёт более высокие матрицы, применимые для сравнения белков, удалённых эволюционно.

PAM матрица

- PAM матрица базируется на последовательностях с 85% идентичности.



У близких белков функции не должны сильно различаться

Protein	PAMs per 100 million years
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome *c*	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

Относительная мутабельность аминокислот

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Нормализованные частоты аминокислот

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

PAM 1 – матрица вероятностей

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	98.67	0.02	0.09	0.10	0.03	0.08	0.17	0.21	0.02	0.06	0.04	0.02	0.06	0.02	0.22	0.35	0.32	0.00	0.02	0.18
R	0.01	99.13	0.01	0.00	0.01	0.10	0.00	0.00	0.10	0.03	0.01	0.19	0.04	0.01	0.04	0.06	0.01	0.08	0.00	0.01
N	0.04	0.01	98.22	0.36	0.00	0.04	0.06	0.06	0.21	0.03	0.01	0.13	0.00	0.01	0.02	0.20	0.09	0.01	0.04	0.01
D	0.06	0.00	0.42	98.59	0.00	0.06	0.53	0.06	0.04	0.01	0.00	0.03	0.00	0.00	0.01	0.05	0.03	0.00	0.00	0.01
C	0.01	0.01	0.00	0.00	99.73	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.05	0.01	0.00	0.03	0.02
Q	0.03	0.09	0.04	0.05	0.00	98.76	0.27	0.01	0.23	0.01	0.03	0.06	0.04	0.00	0.06	0.02	0.02	0.00	0.00	0.01
E	0.10	0.00	0.07	0.56	0.00	0.35	98.65	0.04	0.02	0.03	0.01	0.04	0.01	0.00	0.03	0.04	0.02	0.00	0.01	0.02
G	0.21	0.01	0.12	0.11	0.01	0.03	0.07	99.35	0.01	0.00	0.01	0.02	0.01	0.01	0.03	0.21	0.03	0.00	0.00	0.05
H	0.01	0.08	0.18	0.03	0.01	0.20	0.01	0.00	99.12	0.00	0.01	0.01	0.00	0.02	0.03	0.01	0.01	0.01	0.04	0.01
I	0.02	0.02	0.03	0.01	0.02	0.01	0.02	0.00	0.00	98.72	0.09	0.02	0.21	0.07	0.00	0.01	0.07	0.00	0.01	0.33
L	0.03	0.01	0.03	0.00	0.00	0.06	0.01	0.01	0.04	0.22	99.47	0.02	0.45	0.13	0.03	0.01	0.03	0.04	0.02	0.15
K	0.02	0.37	0.25	0.06	0.00	0.12	0.07	0.02	0.02	0.04	0.01	99.26	0.20	0.00	0.03	0.08	0.11	0.00	0.01	0.01
M	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.08	0.04	98.74	0.01	0.00	0.01	0.02	0.00	0.00	0.04
F	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.01	0.02	0.08	0.06	0.00	0.04	99.46	0.00	0.02	0.01	0.03	0.28	0.00
P	0.13	0.05	0.02	0.01	0.01	0.08	0.03	0.02	0.05	0.01	0.02	0.02	0.01	0.01	99.26	0.12	0.04	0.00	0.00	0.02
S	0.28	0.11	0.34	0.07	0.11	0.04	0.06	0.16	0.02	0.02	0.01	0.07	0.04	0.03	0.17	98.40	0.38	0.05	0.02	0.02
T	0.22	0.02	0.13	0.04	0.01	0.03	0.02	0.02	0.01	0.11	0.02	0.08	0.06	0.01	0.05	0.32	98.71	0.00	0.02	0.09
W	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	99.76	0.01	0.00
Y	0.01	0.00	0.03	0.00	0.03	0.00	0.01	0.00	0.04	0.01	0.01	0.00	0.00	0.21	0.00	0.01	0.01	0.02	99.45	0.01
V	0.13	0.02	0.01	0.01	0.03	0.02	0.02	0.03	0.03	0.57	0.11	0.01	0.17	0.01	0.03	0.02	0.10	0.00	0.02	99.01

PAM 1 – нормализованная матрица вероятностей

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

[top row shows original amino acid; left column shows replacement amino acid]

PAM 250

ORIGINAL AMINO ACID

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

PAM матрицы

Observed % difference	Evolutionary distance (PAM)
1	1
10	11
20	23
30	38
40	56
50	80
60	120
70	159
80	250

- Значення елементів вагової PAM-матриці розраховується за формулою
- $S(i,j) = 10 \log_{10}(M_{ij}/p_j)$
- де S – вага співставлення амінокислоти *i* та амінокислоти *j*, M_{ij} – імовірність заміни *i* на *j* (з відповідної матриці імовірностей), p_j – нормалізована частота зустрічаємості амінокислоти *j* (імовірність зустріти амінокислоту *j* при випадковому вирівнюванні).

PAM 250 – **весовая** матрица

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	−2	6																		
N	0	0	2																	
D	0	−1	2	4																
C	−2	−4	−4	−5	12															
Q	0	1	1	2	−5	4														
E	0	−1	1	3	−5	2	4													
G	1	−3	0	1	−3	−1	0	5												
H	−1	2	2	1	−3	3	1	−2	6											
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5										
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	−2	6									
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5								
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6							
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9						
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6					
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2				
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3			
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17		
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4

BLOSUM Matrices

- **Blo**cks **Su**bstitution **M**atrices.

Матрицы PAM обладают ограниченными возможностями, так как их «рейтинги замен» были получены из выравниваний последовательностей с как минимум 85% идентичности.

- Henikoff and Henikoff (1992) разработали набор матриц, базирующийся на большем количестве данных (dataset of alignments).
  BLOSUM учитывает значительно больше замен, чем PAM, даже для редких пар.

BLOSUM

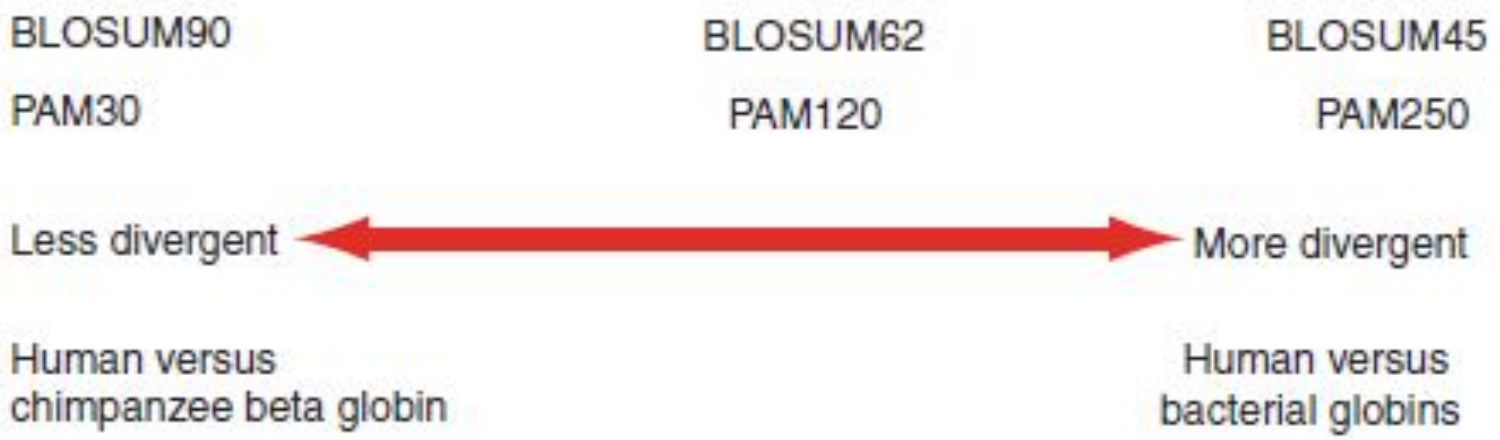
- Блоки – короткие стабильные образы «шаблоны» по 3-60 аа длиной.

- Белки могут быть поделены на семейства по наличию тех или иных блоков (семейство X содержит блоки a,b,c,d).
  Blosum использует ~500 семейств и ~2000 блоков.

- Различные матрицы Blosum выведены из блоков с различной степенью идентичности: blosum62 получена из выравнивания последовательностей с по меньшей мере 62% идентичности

BLOSUM62

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


BLOSUM90
BLOSUM62
BLOSUM45
PAM30
PAM120
PAM250
Less divergent
More divergent
Human versus chimpanzee beta globin
Human versus bacterial globins

Observed Differences in 100 Residues	Evolutionary Distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

Параметры по умолчанию

- Параметры для открытия\продления промежутков индивидуальны для каждой матрицы
- PAM30: open=9, extension=1
- PAM250: open=14, extension=2

Параметры по умолчанию

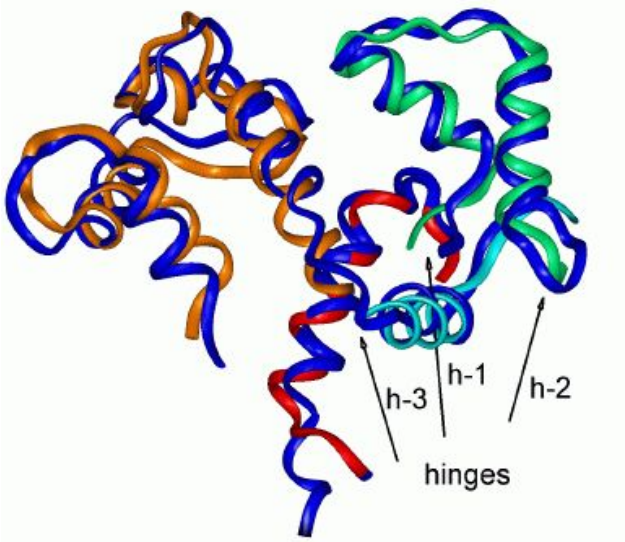
Выравнивания будут сильно отличаться при использовании различных параметров для промежутков.

Для каждой матрицы параметры по умолчанию генерируют оптимальное выравнивание.

Матрицы были тестированы с разными параметрами до тех пор, пока не был получено «правильное выравнивание».

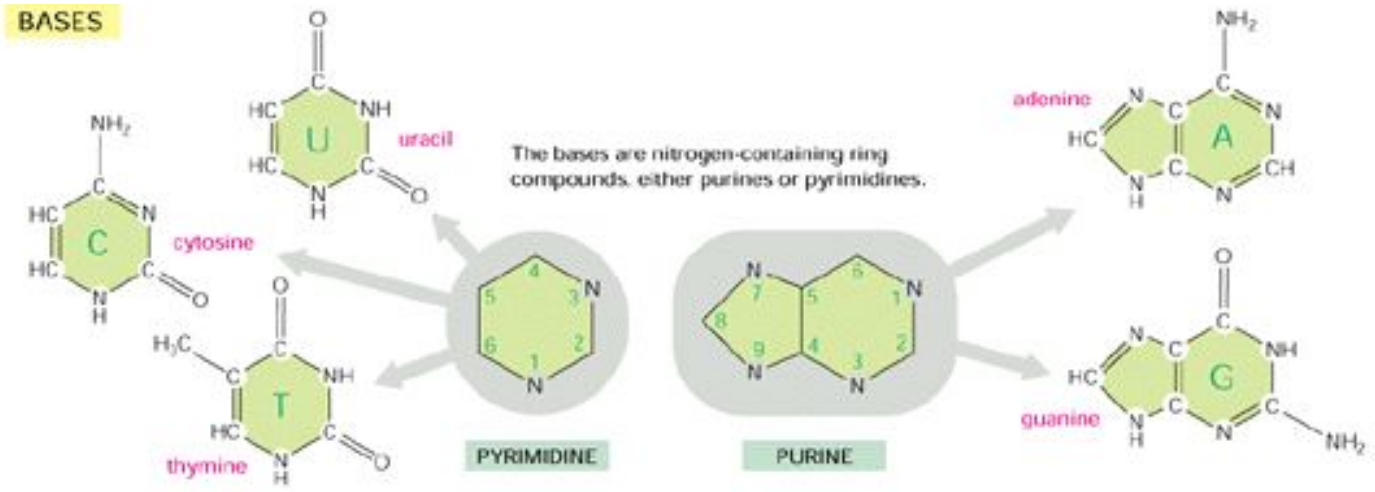
Параметры по умолчанию

Мы можем использовать выравнвание последовательностей, базирующееся на структурном выравнивании. В этом случае структурное выравнивание является «правильным» для наших целей



Матрицы оценки DNA

- Похожесть нуклеотидов DNA определить невозможно.
- Основания делятся на 2 группы: пурины (A,G) и пиримидины (C,T)



Матрицы оценки DNA

Мутации делятся на переходы (transitions) и превращения (transversions).

Транзиции – пурин на пурин, пиримидин на пиримидин.

Трансверсии – пурин на пиримидин или пиримидин на пурин.

Матрицы оценки DNA

- De-facto транзиции происходят чаще.

Матрицы оценки DNA

Унифицированная матрица подстановок нуклеотидов:

T	C	G	A	From To
			2	A
		2	-6	G
	2	-6	-6	C
2	-6	-6	-6	T

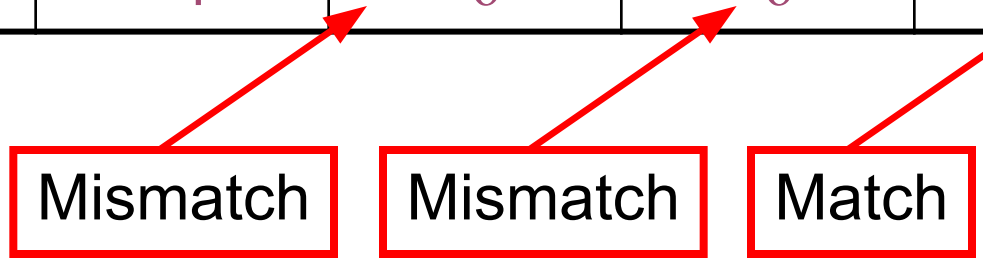
Mismatch

Match

Матрицы оценки DNA

Неунифицированная матрица подстановок нуклеотидов:

T	C	G	A	From To
			2	A
		2	-4	G
	2	-6	-6	C
2	-4	-6	-6	T



Глобальное выравнивание

- Алгоритм Needleman and Wunsch (1970)
- Находит выравнивание двух полных последовательностей:

```
ADLGAVFALCDRYFQ
||||      ||||  |
ADLGRTQN-CDRYYQ
```

Локальное выравнивание

- Алгоритм Smith and Waterman (1981).
- Выполняет оптимальное выравнивание наиболее идентичного\похожего сегмента двух последовательностей.

```
вересень
береста
нерест

    марево
катамаран
корчмар

    гумореска
      море
голодомор
```

```
ADLG      CDRYFQ
||||      |||| |
ADLG      CDRYYQ
```

Выравнивание последовательностей методами динамического программирования

Динамічне програмування – спосіб вирішення складних задач шляхом їх розбиття на більш прості підзадачі. Він може бути застосований для так званих задач з оптимальною підструктурою, що виглядають як набір задач, які перекриваються між собою, і складність яких трішки менше вихідної (загальної) задачі. Термін «оптимальність підструктури» в динамічному програмуванні означає, що оптимальне рішення під задач меншого розміру може бути використано для розв’язання вихідної задачі.

Выравнивание последовательностей методами динамического программирования

У загальному випадку задача, що має оптимальну підструктуру, можна розв'язати за допомогою стратегії «трьох кроків»:

1 розбиття задачі на підзадачі меншого розміру;

2 знаходження оптимального розв'язання задач рекурсивно, застосовуючи такий самий трьохкроковий алгоритм;

3 використання отриманого рішення під задач для конструювання рішення вихідної задачі.

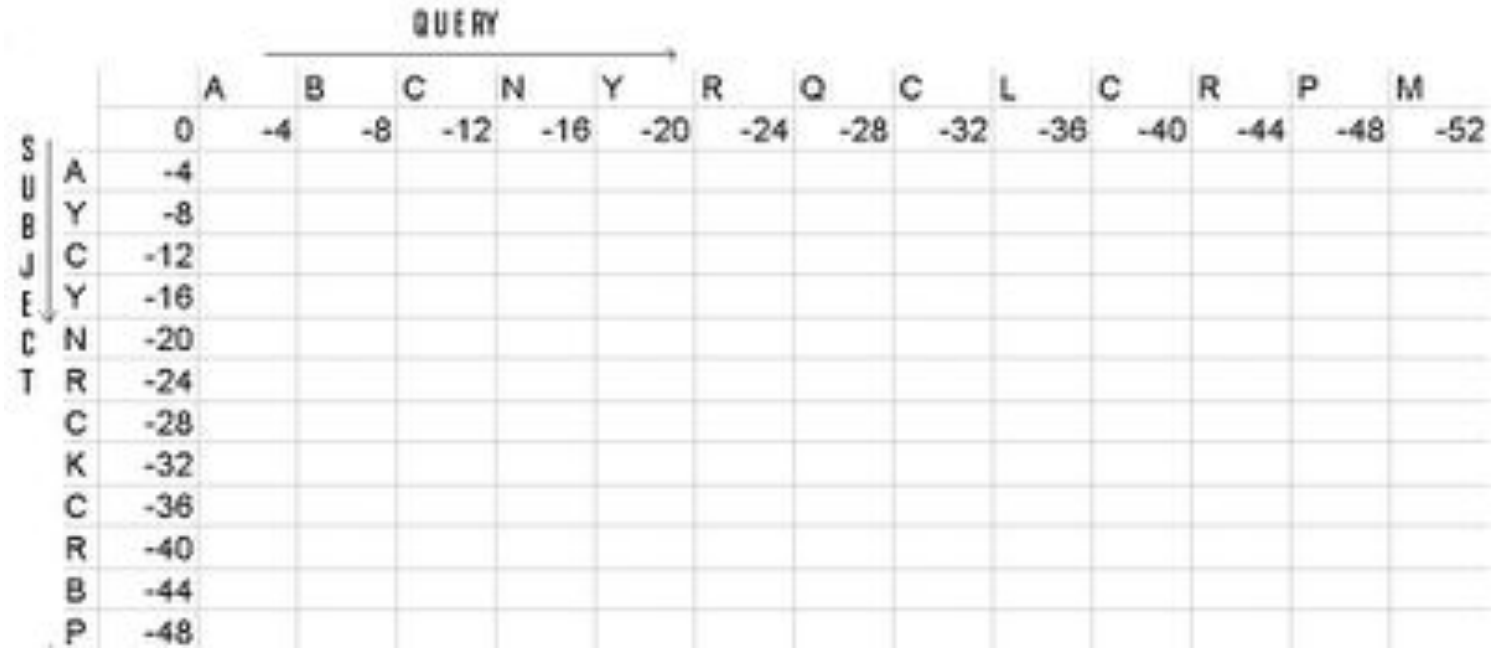
Під задачі розв'язуються поділом їх на підзадачі другого порядку (меншого розміру). Процес продовжується до тих пір, доки ми не прийдемо до тривіального випадку задачі, що розв'язується за константний час (відповідь можна знайти одразу).

Алгоритм Ніделмана-Вунша

- 1. Побудова ініціюючої матриці
- 2. Заповнення матриці
- 3. Пошук шляху вирівнювання

Алгоритм Ніделмана-Вунша

- І. Побудова ініціюючої матриці



Дано: 2 последовательности x[1…n] и y[1…m]

При сопоставлении x[1...i] и y[1…j] есть 3 варианта:

Совпадение x[1…i-1] и y[1…j-1]: x[i]=y[j]

x[1…i-1] i

y[1…j-1] j

Совпадение x[1…i] и y[1…j-1] и совпадение пропуска в x и y[j]

x[1… i] -

y[1…j-1] j

Совпадение x[1…i-1] и y[1…j] и совпадение x[i] и пропуска в y

x[1…i-1] i

y[1… j] -

Scoring matrix s(a,b), $s(-, x) = s(x, -) = -d$

F_{ij} – лучшая score-функция выравнивания x[1…i] and y[1…j]

for 1 <= i <= n, 1 <= j <= m

$$F_{ij} = \max \begin{cases} F_{i-1,j-1} + s(x_i, y_j) \\ F_{i-1,j} - d \\ F_{i,j-1} - d \end{cases}$$

Needleman-Wunsch 1970

Алгоритм Ніделмана-Вунша

- Заповнення матриці

		A	B	C	N	Y	R	Q	C	L	C	R	P	M
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	-48	-52
A	-4	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35	-39	-43
Y	-8	1	2	-2	-6	-2	-6	-10	-14	-18	-22	-26	-29	-33
C	-12	-3	-2	7	3	-1	-5	-9	-5	-9	-13	-17	-21	-25
Y	-16	-7	-6	3	4	8	4	0	-4	-8	-12	-16	-20	-24
N	-20	-11	-10	-1	8	4	5	1	-3	-7	-11	-15	-19	-23
R	-24	-15	-14	-5	4	5	9	5	1	-3	-7	-6	-10	-14
C	-28	-19	-18	-9	0	1	5	6	10	6	2	-2	-6	-10
K	-32	-23	-22	-13	-4	-3	1	2	6	7	3	-1	-5	-9
C	-36	-27	-26	-17	-8	-7	-3	-2	7	3	12	8	4	0
R	-40	-31	-30	-21	-12	-11	-2	-6	3	4	8	17	13	9
B	-44	-35	-26	-25	-16	-7	-6	-5	-1	0	4	13	14	10
P	-48	-39	-30	-29	-20	-19	-10	-9	-5	-4	0	9	18	14

Neddleman & Wunsch

1970 год

Алгоритм:

1) Начинает с конца последовательностей и продвигается, за каждый цикл сравнивая по одной букве
2) Генерирует все возможные варианты (сходство, различие, делеция, инсерция)
3) Определяют очки:

Например, сходство = +1, различие = 0, gap = -0.5

		A	B	C	N	Y	R	Q	C	L	C	R	P	M
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	-48	-52
A	-4	D	L	L	L	L	L	L	L	L	L	L	L	L
Y	-8	T	D	D	L	L	L	L	L	L	L	L	L	L
C	-12	T	DT	D	L	L	DL	DL	D	L	D	L	L	L
Y	-16	T	DL	T	D	D	L	L	L	DL	DL	DL	DL	DL
N	-20	T	DL	T	D	TL	D	DL	DL	DL	DL	DL	DL	DL
R	-24	T	DL	T	T	D	D	L	L	L	L	D	L	L
C	-28	T	DT	DT	T	DT	T	D	D	L	DL	L	L	L
K	-32	T	DT	T	T	DT	T	DT	T	D	DL	DL	DL	DL
C	-36	T	DT	DT	T	DT	T	T	D	DTL	D	L	L	L
R	-40	T	DT	T	T	DT	D	DTL	T	D	T	D	L	L
B	-44	T	D	T	T	DT	T	D	T	DT	T	T	D	DL
P	-48	T	T	DT	T	DT	T	DT	D	DT	T	T	D	L

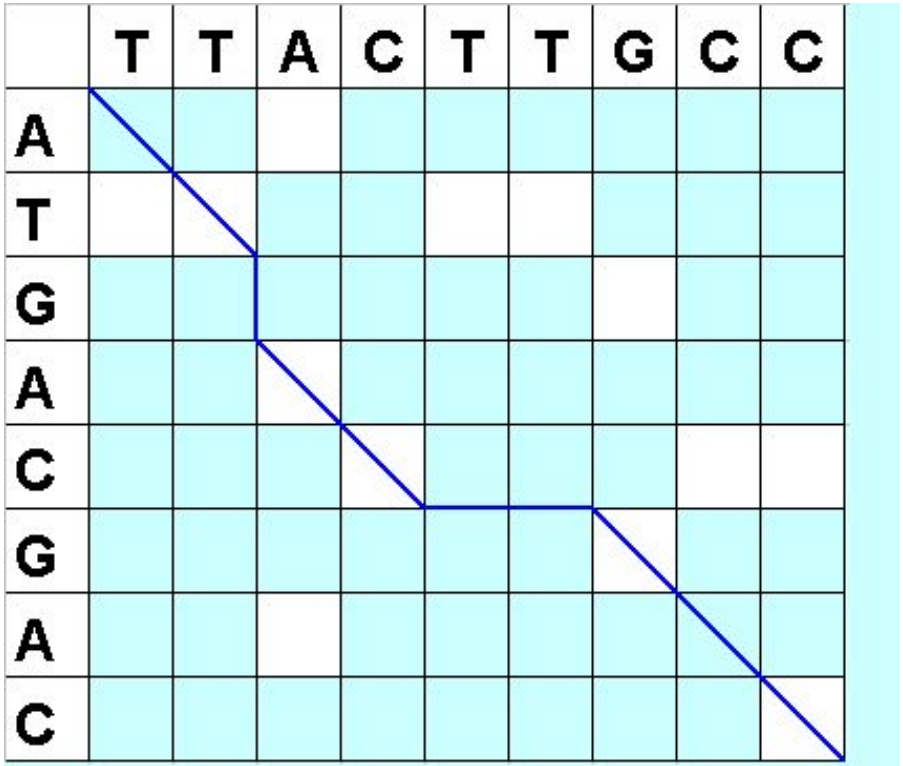
		A	B	C	N	Y	R	Q	C	L	C	R	P	M
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36	-40	-44	-48	-52
A	-4	5	1	-3	-7	-11	-15	-19	-23	-27	-31	-35	-39	-43
Y	-8	1	2	-2	-6	-2	-6	-10	-14	-18	-22	-26	-29	-33
C	-12	-3	-2	7	3	-1	-5	-9	-5	-9	-13	-17	-21	-25
Y	-16	-7	-6	3	4	8	4	0	-4	-8	-12	-16	-20	-24
N	-20	-11	-10	-1	8	4	5	1	-3	-7	-11	-15	-19	-23
R	-24	-15	-14	-5	4	5	9	5	1	-3	-7	-6	-10	-14
C	-28	-19	-18	-9	0	1	5	6	10	6	2	-2	-6	-10
K	-32	-23	-22	-13	-4	-3	1	2	6	7	3	-1	-5	-9
C	-36	-27	-26	-17	-8	-7	-3	-2	7	3	12	8	4	0
R	-40	-31	-30	-21	-12	-11	-2	-6	3	4	8	17	13	9
B	-44	-35	-26	-25	-16	-7	-6	-5	-1	0	4	13	14	10
P	-48	-39	-30	-29	-20	-19	-10	-9	-5	-4	0	9	18	14

Маршрут выравнивания

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar;48(3):443-453.

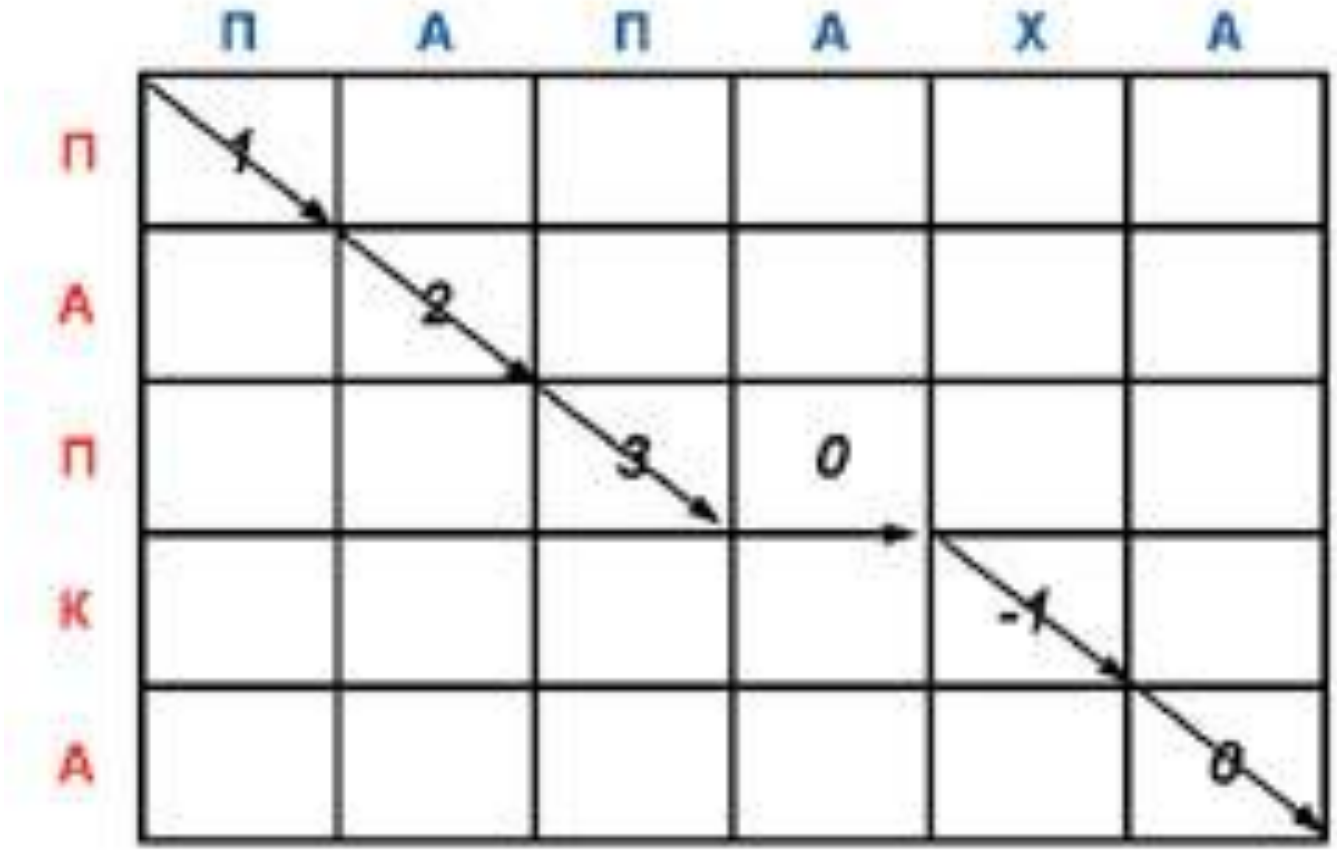
Матрица «0 / 1»

Identity, %



T	T	-	A	C	T	T	G	C	C
A	T	G	A	C	-	-	G	A	C
0	1	-1	1	1	-1	-1	1	0	-1

Траектория, соответствующая оптимальному выравниванию

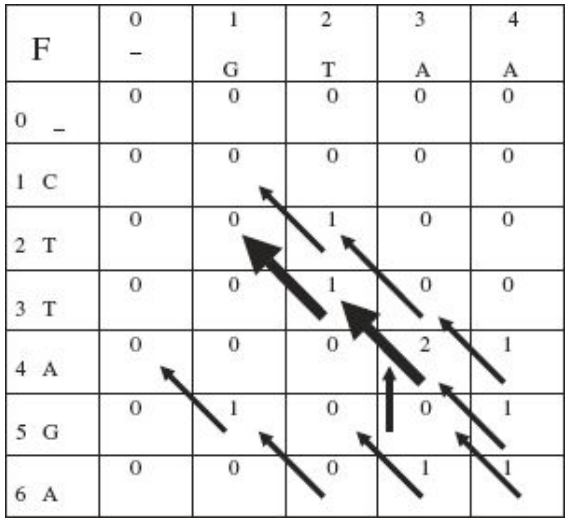


Алгоритм Сміта-Ватермана

Важно:

Выравнивание может не только окончиться, но и начаться в любом месте матрицы.

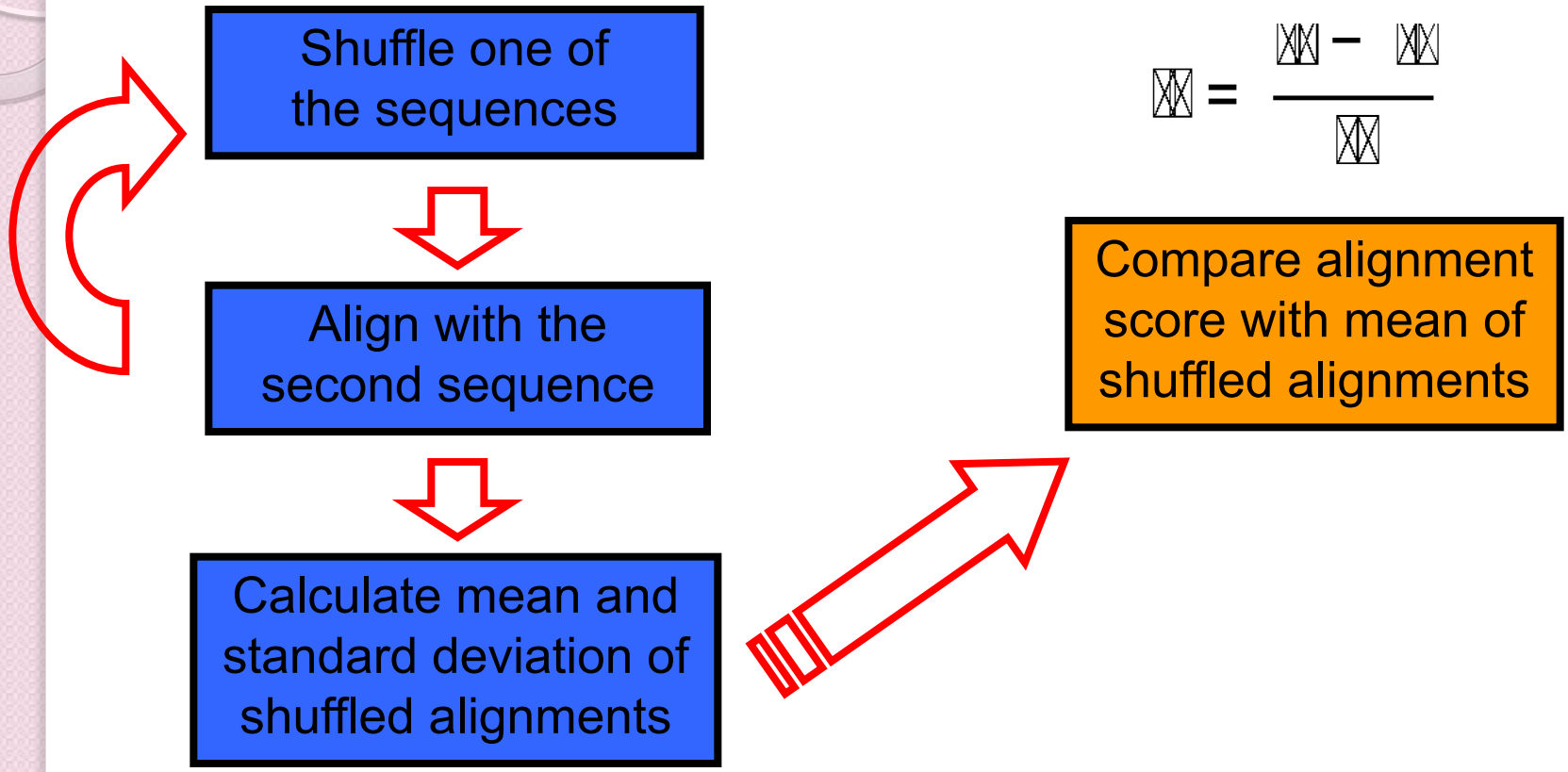
Таким образом, вместо того, чтобы выбирать стартовую точку $F(n,m)$ в правом нижнем углу, выбирают элементы с максимальным скорингом в матрице.



Оценка

- Как можно оценить достоверность выравнивания?
- Какое выравнивание лучше ?

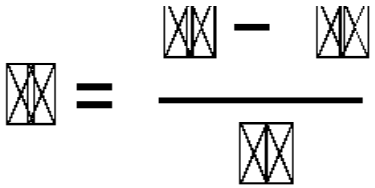
Оценка неслучайности выравнивания



Данные с тем же набором, но с разным порядком:

1. Перемешивание одной последовательности.
2. Повтор выравнивания и его оценка.
3. Повторение 1) и 2) много раз.
4. Посчёт среднего и оценки выравнивания перемешанной последовательности.

Оценка неслучайности выравнивания



x – вага вирівнювання двох вихідних послідовностей

μ – усереднена вага отриманих у результаті вирівнювання перемішаних послідовностей

σ – стандартне відхилення, обраховане для перемішаних послідовностей

Дякую за увагу
Благодарю за внимание
Thank you for your attention