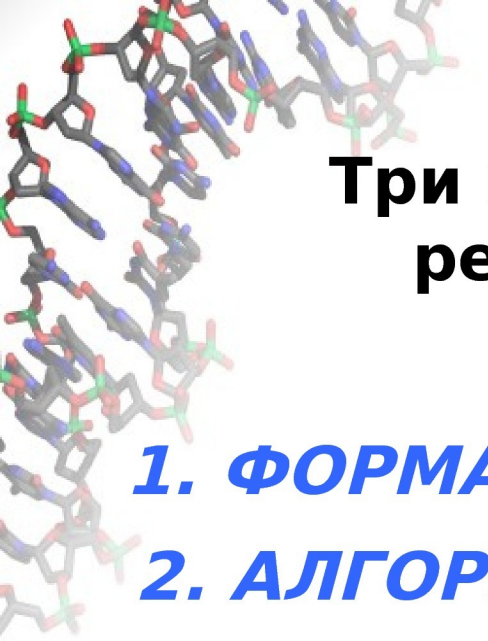


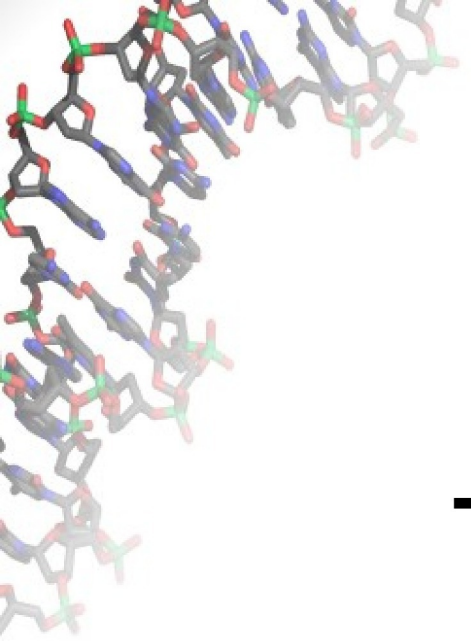
Три вещи, которые нужно уточнить, решая содержательную задачу

1. ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ.

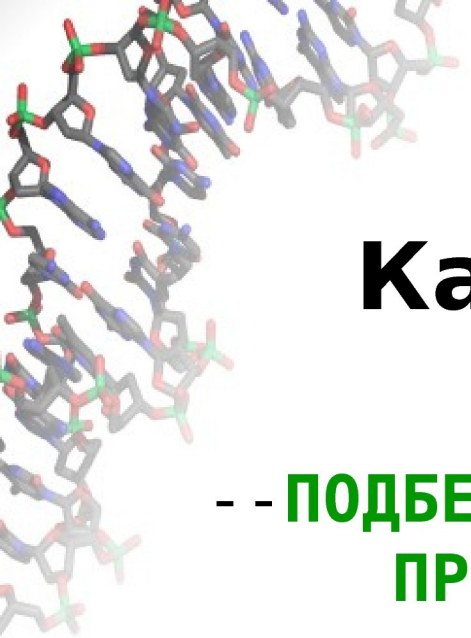
2. АЛГОРИТМ РЕШЕНИЯ

3. ПРОВЕРКА – АДЕКВАТНО ЛИ ТО, ЧТО ПОЛУЧАЕТСЯ (ИСХОДНОЙ СОДЕРЖАТЕЛЬНОЙ ЗАДАЧЕ).

Как правило, уточнение этих вещей происходит взаимосвязано и итеративно



Часть 1.
Точность и достоверность выравниваний.



Варианты выравниваний

Какой вариант выбрать?

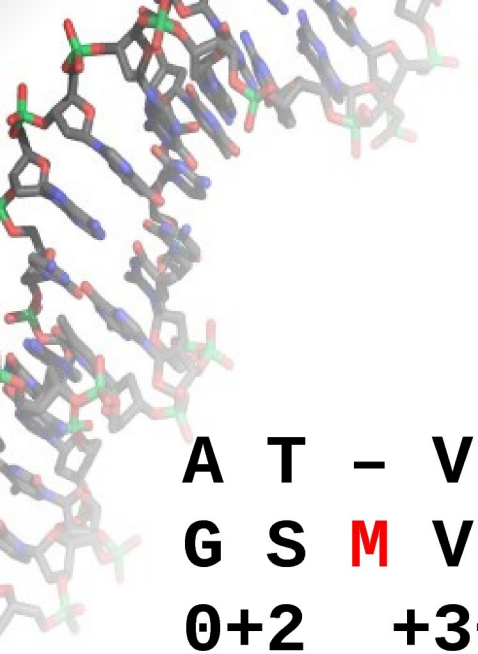
--ПОДБЕРЕЗОВИК
ПРЕДОСИНОВИЧКИ

ПОДБЕРЕЗОВИК--
ПРЕДОСИНОВИЧКИ

ПО-ДБЕРЕЗОВИК--
ПРЕДОСИН-ОВИЧКИ

ПО-ДБЕРЕЗОВИК--
ПРЕД-ОСИНОВИЧКИ

ПО-ДБЕРЕЗОВИ-К-
ПРЕД-ОСИНОВИЧКИ



Вес выравнивания

Пример: A=0; B = 2

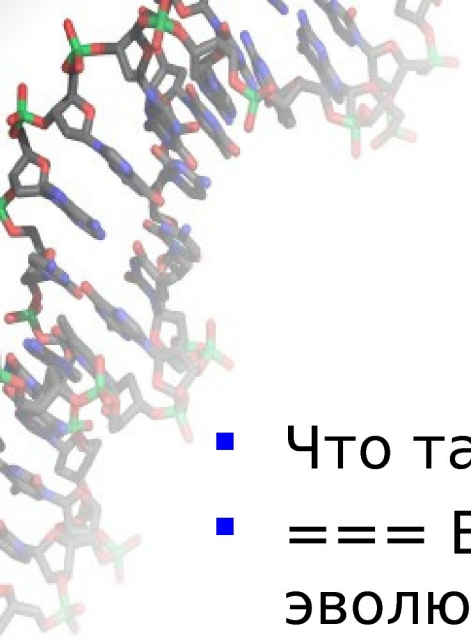
```
A T – V V I – - T G S
G S M V L L E F S G T
0+2  +3+2+3    +2+7+2= 21
    -2       -4       = -6
```

Штраф за удаление фрагиента длины L:

d(L) = A + B*L

$$Score = \Sigma\, m(i,j) - \Sigma\, d(L_k)$$

$$Score = 21 - 6 = 15$$



Правильно ли оптимальное выравнивание?

- Что такое «правильное выравнивание»?
- === Выравнивание, отражающее эволюционный процесс (не обязательно биологический).

Две одинаковые буквы скорее имеют общего предка, чем две разные буквы

Две буквы «одинаковой гласности» скорее имеют общего предка, чем две буквы «разные гласности»

Г)	Д)
П-ОДБЕРЕЗОВИК--	ПО-ДБЕРЕЗОВИ-К-
ПРЕД-ОСИНОВИЧКИ	ПРЕД-ОСИНОВИЧКИ

При любом штрафе за удаление символа

Д) лучше, чем Г)

(удаление – посимвольное)

Две одинаковые буквы скорее имеют общего предка, чем две разные буквы

Две буквы «одинаковой гласности» скорее имеют общего предка, чем две буквы «разные гласности»

Г)

П-ОДБЕРЕЗОВИК--

ПРЕД-ОСИНОВИЧКИ

Д)

ПО-ДБЕРЕЗОВИ-К-

ПРЕД-ОСИНОВИЧКИ

НО:

Д) – «неправильно» («эволюционно»)

ПОД -БЕРЕЗ-ОВИ-К

ПРЕД-ОСИН –ОВИ-К

ПРЕД-ОСИН –ОВИ-Ч-ЕК

ПРЕД-ОСИН –ОВИ-Ч- К-И

Эталон: эволюционное выравнивание

ASVVLDFTGT

ASVVLDFTGT
ATVVI—-TGS

AS-VVLDFTGT
GSMVLLEFSGT

AS-VVLDFTGT
AT-VVI—-TGS

AS-VVLDFTGT
GSMVLLEFSGT

AS-VVLDFTGT
AT-VVI—-TGS
GSMVLLEFSGT

Эталон: эволюционное выравнивание

ASVVLDFTGT

ASVVLDFTGT
ATVVI—-TGS

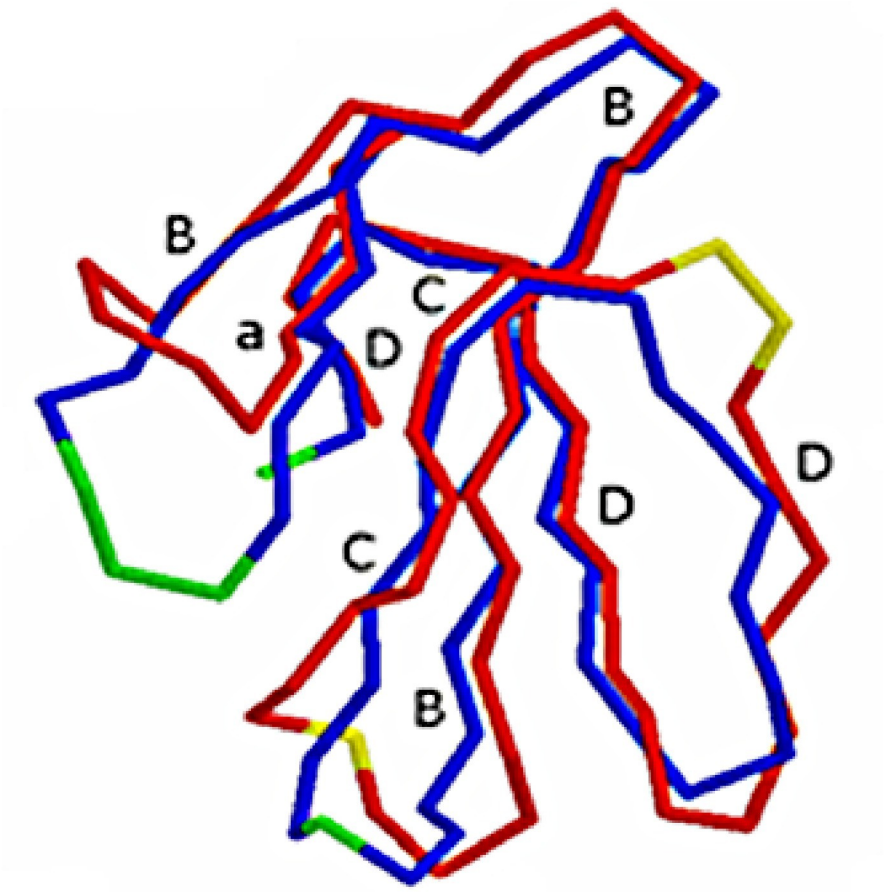
AS-VVLDFTGT
GSMVLLEFSGT

AS-VVLDFTGT
AT-VVI—-TGS

AS-VVLDFTGT
GSMVLLEFSGT

AT-VVI—-TGS
GSMVLLEFSGT

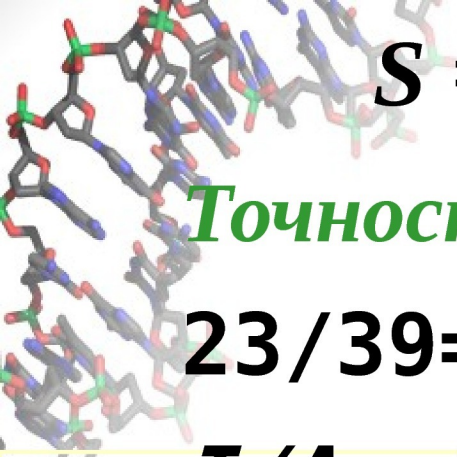
Приближение: структурные выравнивания



Структурное и алгоритмическое выравнивания

```
Str)                                    40
сопоставлений
  lkCnqli...PPFWKTCPKGKNLCYKmtmraapmvPVKRGCidv
   riCfnhqssqPQTTKTCSPGESSCYHkqwsdfrgtIIERGCg..
     *       ******************        ******
     1              16                   6


AlgSW)
     1              16                    6
     *       ******************        ******
lk..C...nqliPPFWKTCPKGKNLCYK...mtmraapmvPVKRGCidv
..riCfnhqssqPQTTKTCSPGESSCYHkqwsdfrgt...IIERGC..g
                                  35 сопоставлений
```



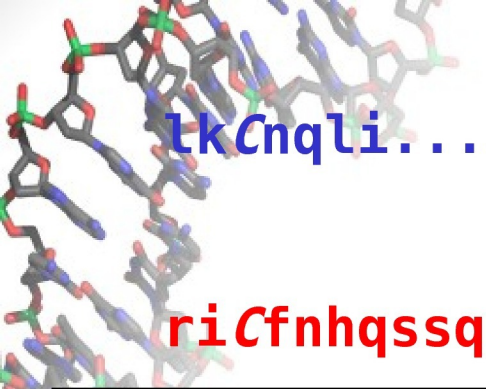
S = 39 I = 23 A= 34

Точность *Acc = I/S=*

23/39=0.59 *Достоверность* *Conf =*

I/A= 23/34=0.68

```
Str)
   1234567   8901234567890123456789012345678 9
   lkCnqli...PPFWKTCPKGKNLCYKmtmraapmvPVKRGCidv
   riCfnhqssqPQTTKTCSPGESSCYHkqwsdfrgtIIERGCg..
     *       ******************          ******
     1              16                      6
AlgSW)
     1              16                      6
     *       ******************          ******
lk..C...nqliPPFWKTCPKGKNLCYK...mtmraapmvPVKRGCidv
..riCfnhqssqPQTTKTCSPGESSCYHkqwsdfrgt...IIERGC..g
     1   2345678901234567890   234567   890123  4
```

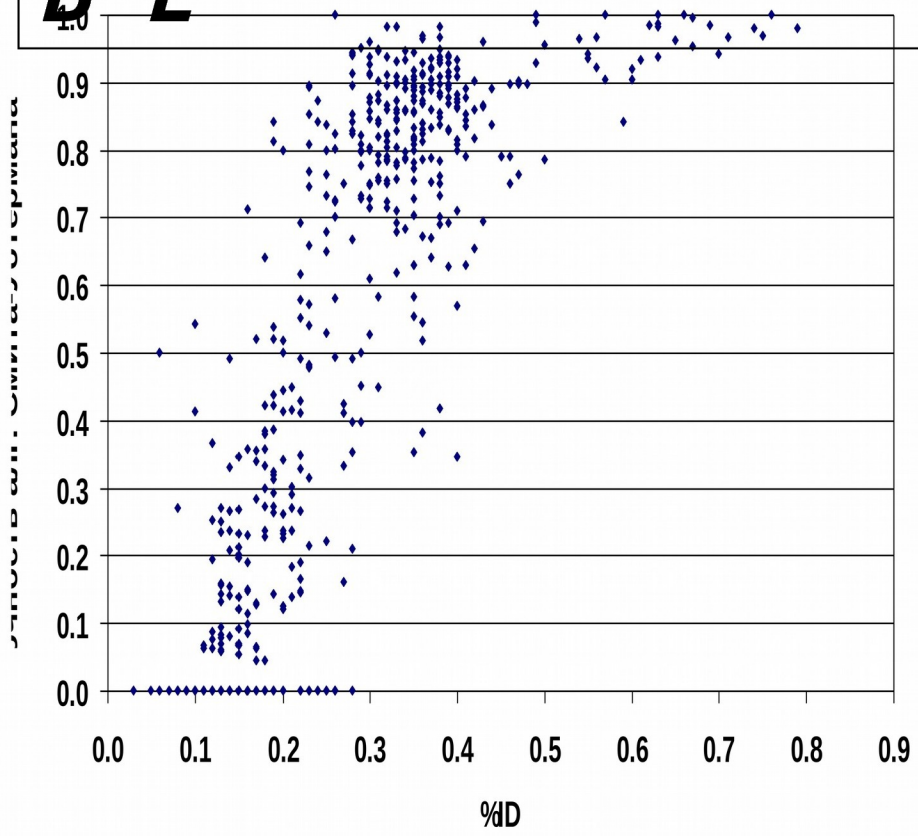


lk*C*nqli...*P*PF*W**KTC*PK*G*KNL*CY*Kmtmraapmvpv*K**RGC*i
dv

ri*C*fnhqssq*P*QTT*KTC*SP*G*ESS*CY*Hkqwsdfrgtiie*RGC*g
..

%ID=
=11/39 =
= ***0.28***
Acc =***0.59***

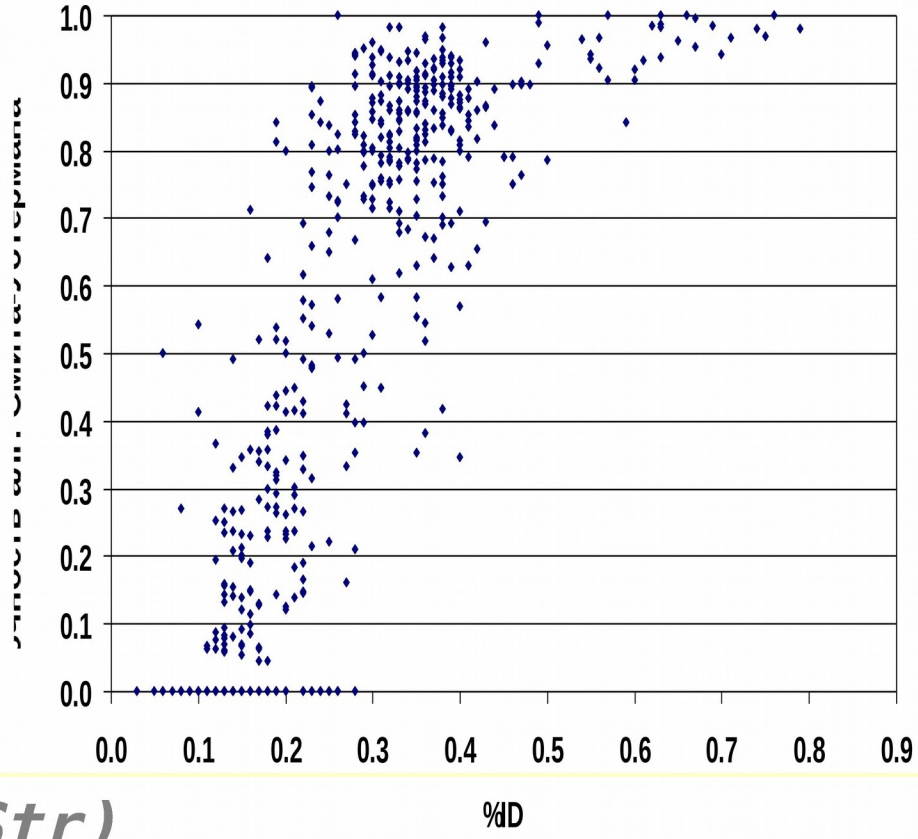
Алгоритм Смита-Уотермана (SW). d(L) = A + B*L



При ID< 0.3
SW-выравнивание
НЕ ПОХОЖЕ
на структурное

%ID	SW точность (acc)
< 0,1	0,037
0,1-0,3	0,306
0,3-0,4	0,818
>0,4	0,893

Причины плохого качества выравниваний SW



Острова **– безделеционные фрагменты выравниваний.**

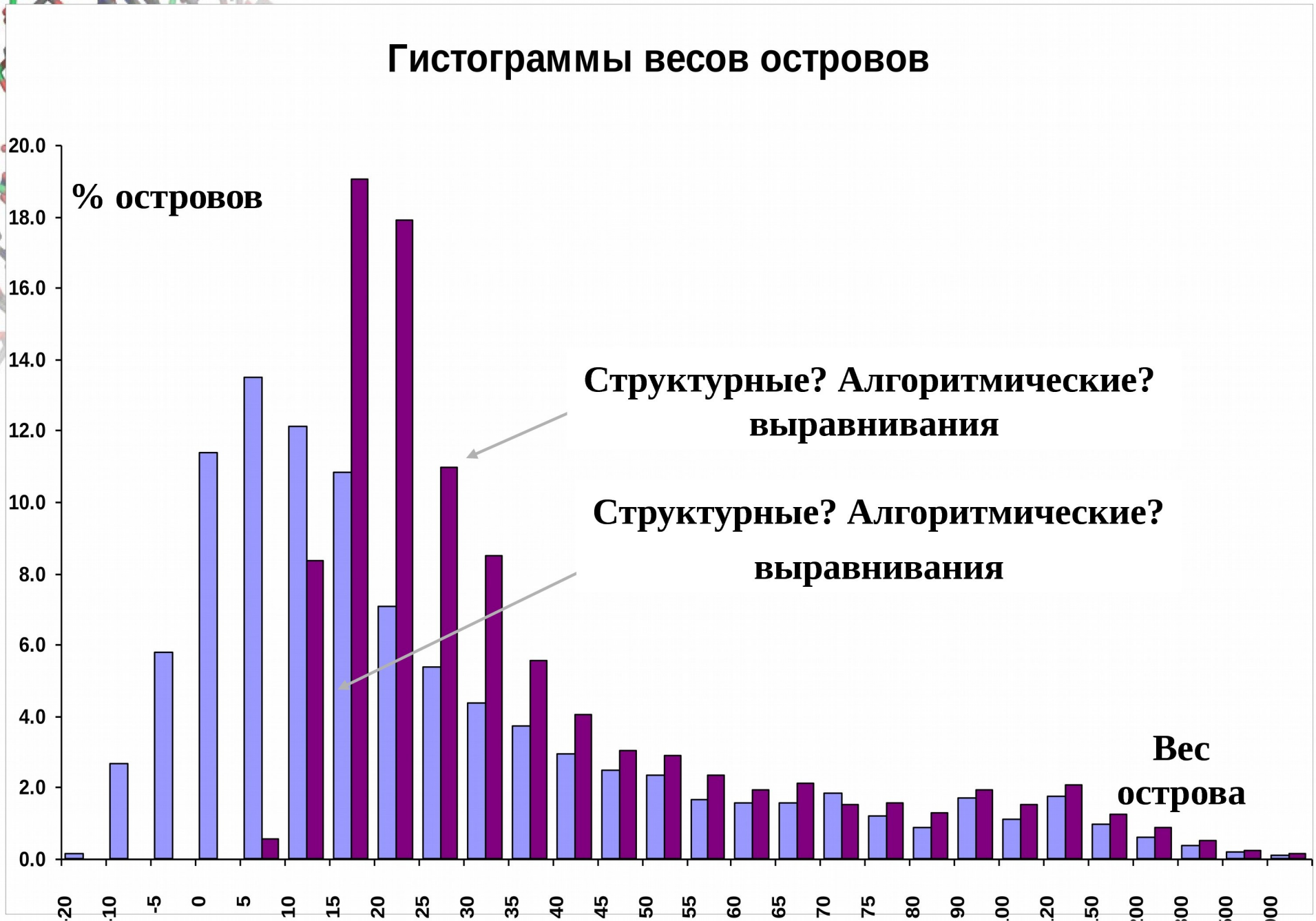
Вес острова **– сумма весов сопоставлений**

Str)

```
lkCnqli...PPFWKTCPKGKNLCYKmtmraapmvPVKRGCidv
riCfnhqssqPQTTKTCSPGESSCYHkqwsdfrgtIIERGCg..
^^^^^^^^    ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
```

Остров1 Остров 2

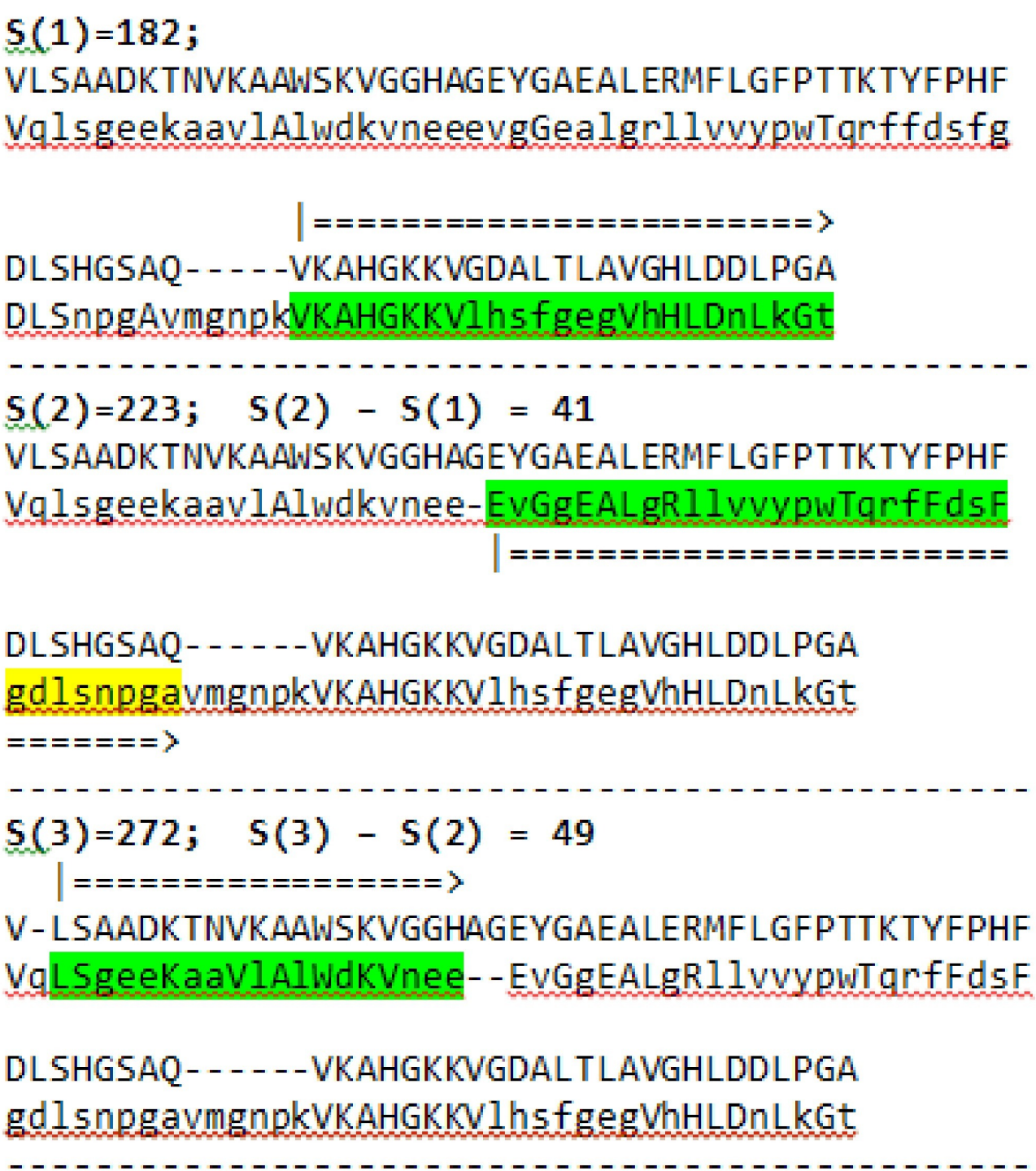
Причины плохого качества выравниваний SW



Пример.
α и β цепи гемоглобина лошади

```
          10        20          30        40         50          60       70        80
1 2345678901234567890  12345678901234567890123456 56789012     67890124567890123456789012
V-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLSHGSAQ-----VKAHGKKVGDALTLAVGHLDDLPGA
VqLSgeeKaaVlAlWdKVnee--EvGgEALgRllvvypwTqrfFdsFgDLSnpgAvmgnpkVKAHGKKVlhsfgegVhHLDnLkGt
1 23   4  5 6 7 89     0 1 234 5       6   7  8 901   2      34567890       1 234 5 6
                       10                          20                 30
```

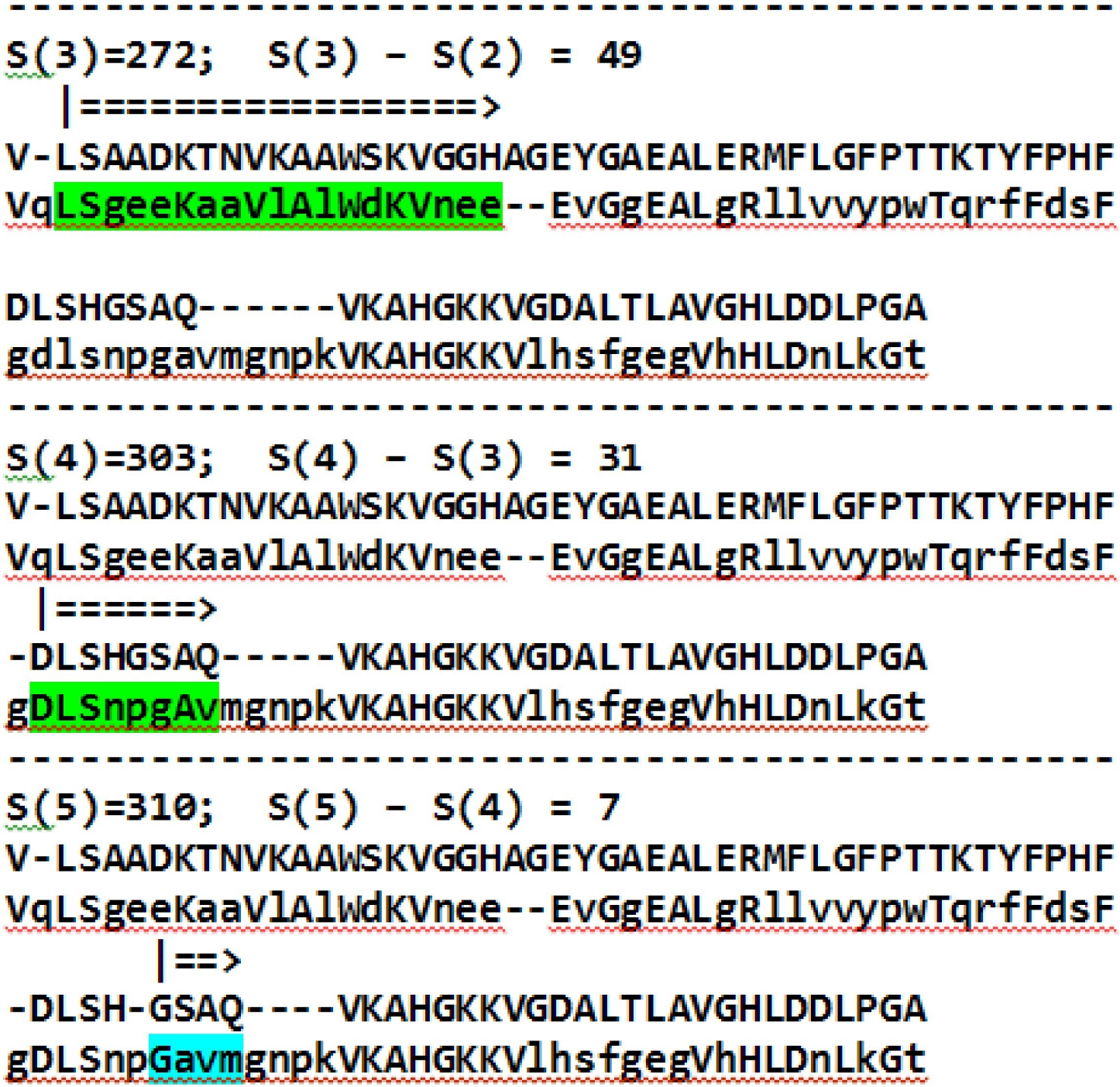
%ID = 36/81 ≈ 0.44

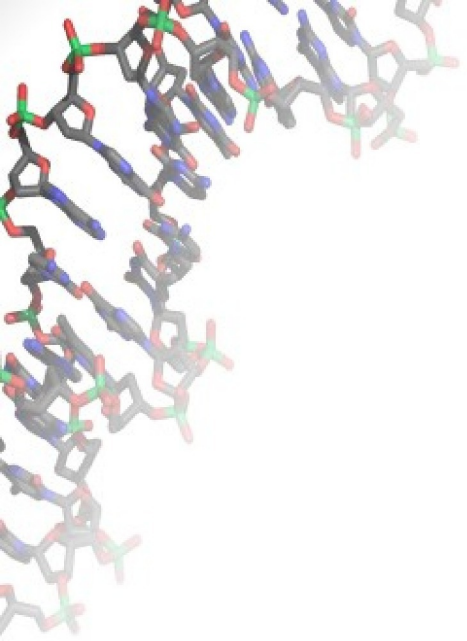


Пример.
α и β цепи гемоглобина лошади

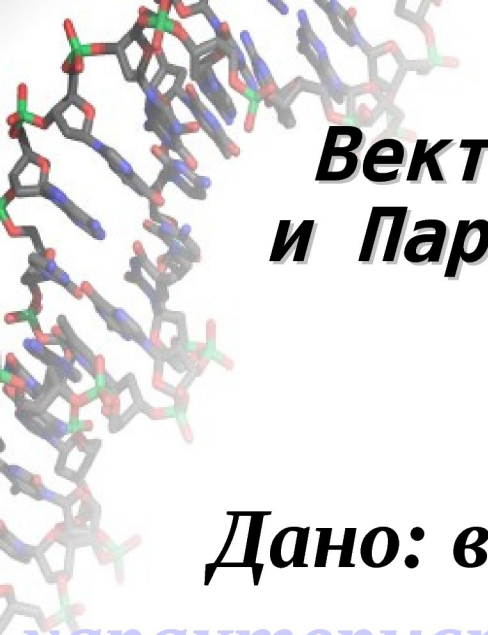
Оптимальные выравнивания, содержащие 1, 2 и 3 удаленных фрагмента

S – сумма весов сопоставлений


S(3)=272; S(3) - S(2) = 49
V-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
VqLSgeeKaaVlAlWdKVnee--EvGgEALgRllvvypwTqrfFdsF
DLSHGSAQ------VKAHGKKVGDALTLAVGHLDDLPGA
gdlsnpgavmgnpkVKAHGKKVlhsfgegVhHLDnLkGt
S(4)=303; S(4) - S(3) = 31
V-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
VqLSgeeKaaVlAlWdKVnee--EvGgEALgRllvvypwTqrfFdsF
-DLSHGSAQ-----VKAHGKKVGDALTLAVGHLDDLPGA
gDLSnpgAvmgnpkVKAHGKKVlhsfgegVhHLDnLkGt
S(5)=310; S(5) - S(4) = 7
V-LSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF
VqLSgeeKaaVlAlWdKVnee--EvGgEALgRllvvypwTqrfFdsF
-DLSH-GSAQ----VKAHGKKVGDALTLAVGHLDDLPGA
gDLSnpGavmgnpkVKAHGKKVlhsfgegVhHLDnLkGt



КАК УЛУЧШИТЬ КАЧЕСТВО ВЫРАВНИВАНИЙ?

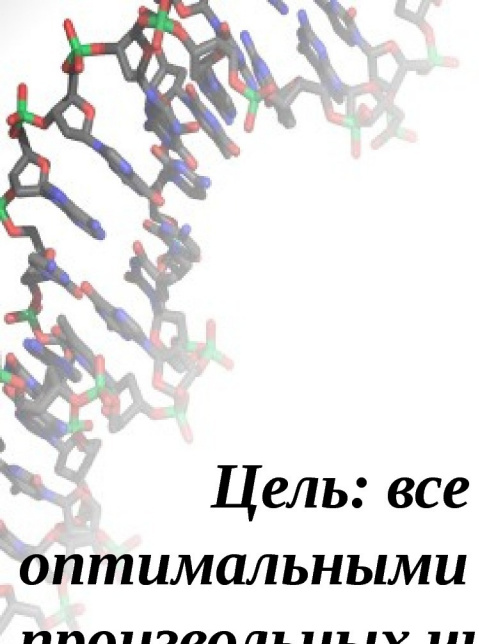


Векторные веса выравниваний и Парето-оптимальные выравнивания

$$W(G) = c \bullet p - f \bullet r - d \bullet s$$

Дано: *вес выравнивания, зависящий от характеристик выравнивания и числовых параметров*

Цель: *все выравнивания, которые могут быть оптимальными при заданном виде весовой функции и каких-либо значениях числовых параметрах.*



$$W(G) = c \cdot p - f \cdot r - d \cdot s$$

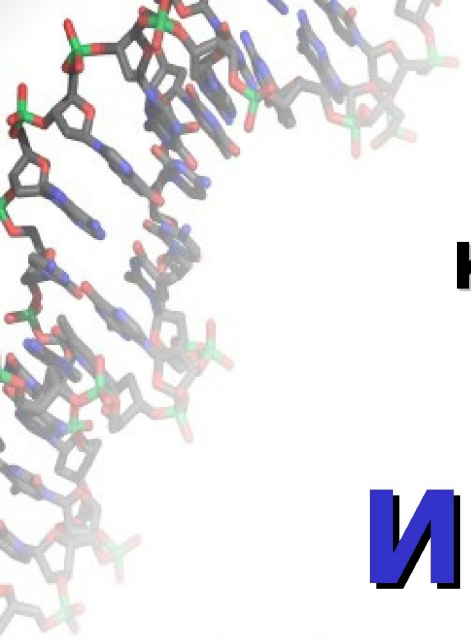
Цель: все выравнивания, которые могут быть оптимальными при заданном виде весовой функции и произвольных числовых параметрах.

Идея:

- *использовать векторные веса:*

$$V(G) = \langle p,\ r,\ s \rangle$$

- *искать Парето-оптимальные выравнивания*



КАК УЛУЧШИТЬ КАЧЕСТВО ВЫРАВНИВАНИЙ?

ИСПОЛЬЗОВАТЬ СПЕЦИФИКУ ЗАДАЧИ!

8. Векторные веса выравниваний и Парето-оптимальные выравнивания

Характеристики:

NMatch(G), *-NDel(G),*

WeightMatch(G) *- NGap(G)*

Алгоритм на основе обобщенных статистических сумм.

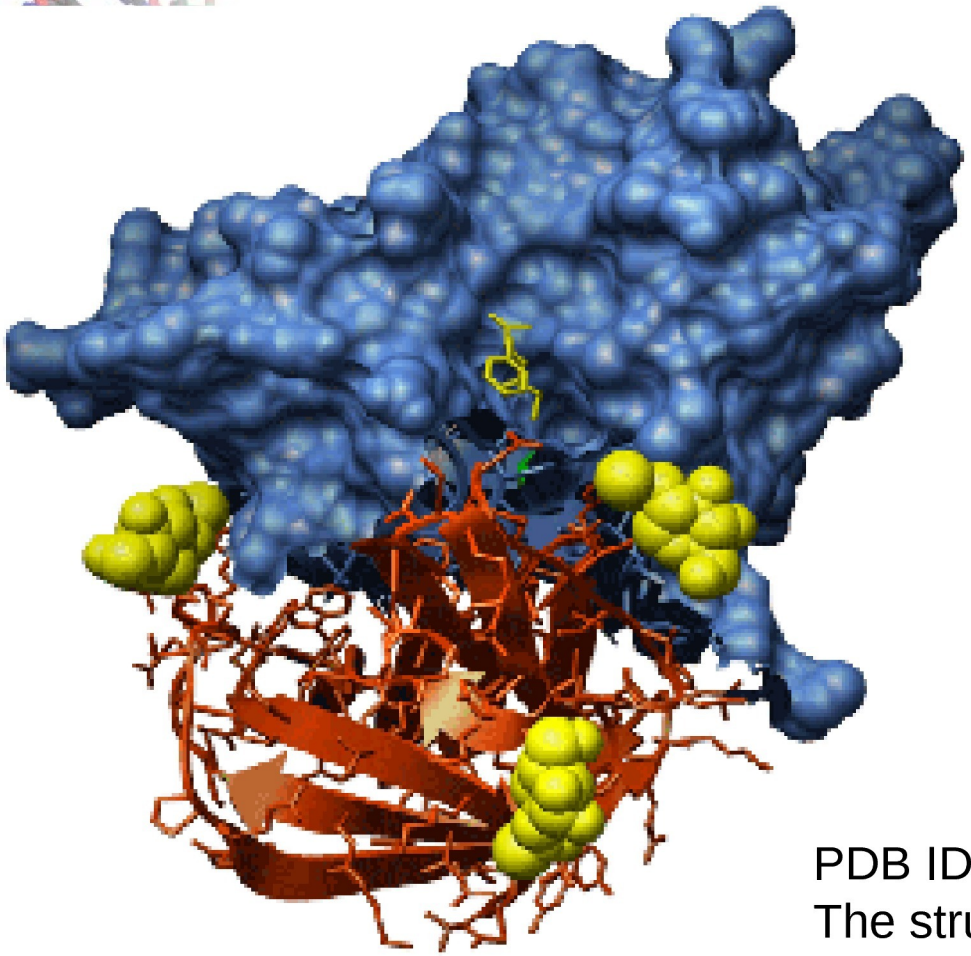
Время для 2-компонентных весов:

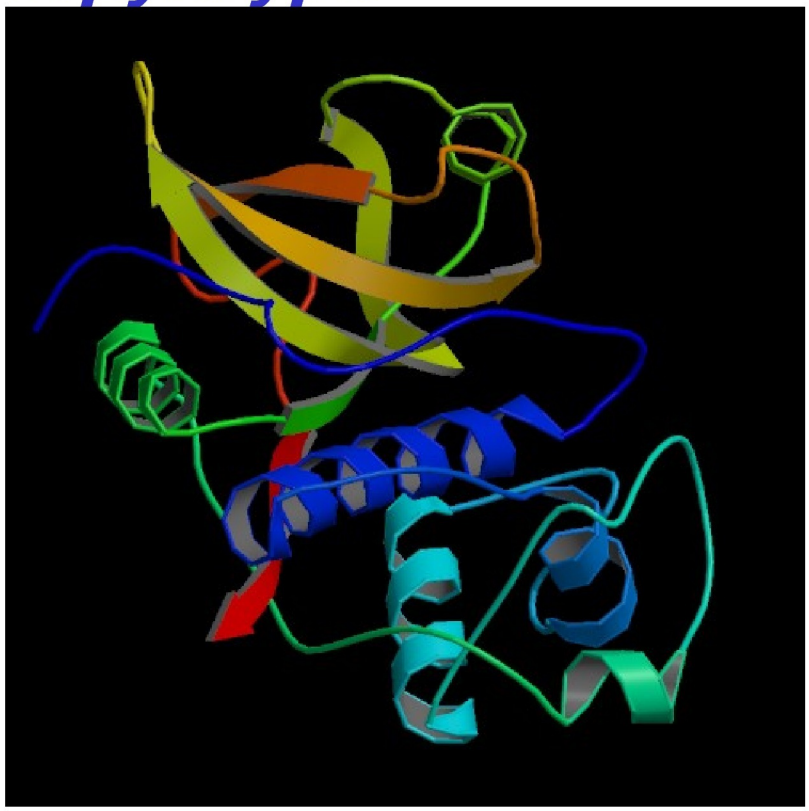
$$\sim \min(L_1, L_2) L_1 L_2$$

Белки: 1 нить; L ~ 10^2 – 10^3

аминокислоты (20)

вторичная структура





PDB ID: **2act** E.N. Baker, E.J. Dodson (1980): The structure of actinidin at 1.7 Ångstroms

9. Выбор биологически корректного выравнивания

Эволюционные события:

- замены символов;

- удаления/вставки фрагментов

Вес:

<WeightMatch(G), - NGap(G)>

2. Улучшение качества выравниваний – учет вторичной структуры.

$$W(i,j) = \max \begin{cases} W(i-1,j-1) + M(a_i, b_j) + SBON \times Q(i,j) \\ W(i-1,j) - GOP - GEP \\ WA(i-1,j) - GEP \\ W(i,j-1) - GOP - GEP \\ WB(i,j-1) - GEP \\ 0 \end{cases}$$

Q(i, j) – мера сходства вторичной структуры в i-й и j-й позициях

SBON – вес учета вторичной структуры

Выбор биологически корректного выравнивания.

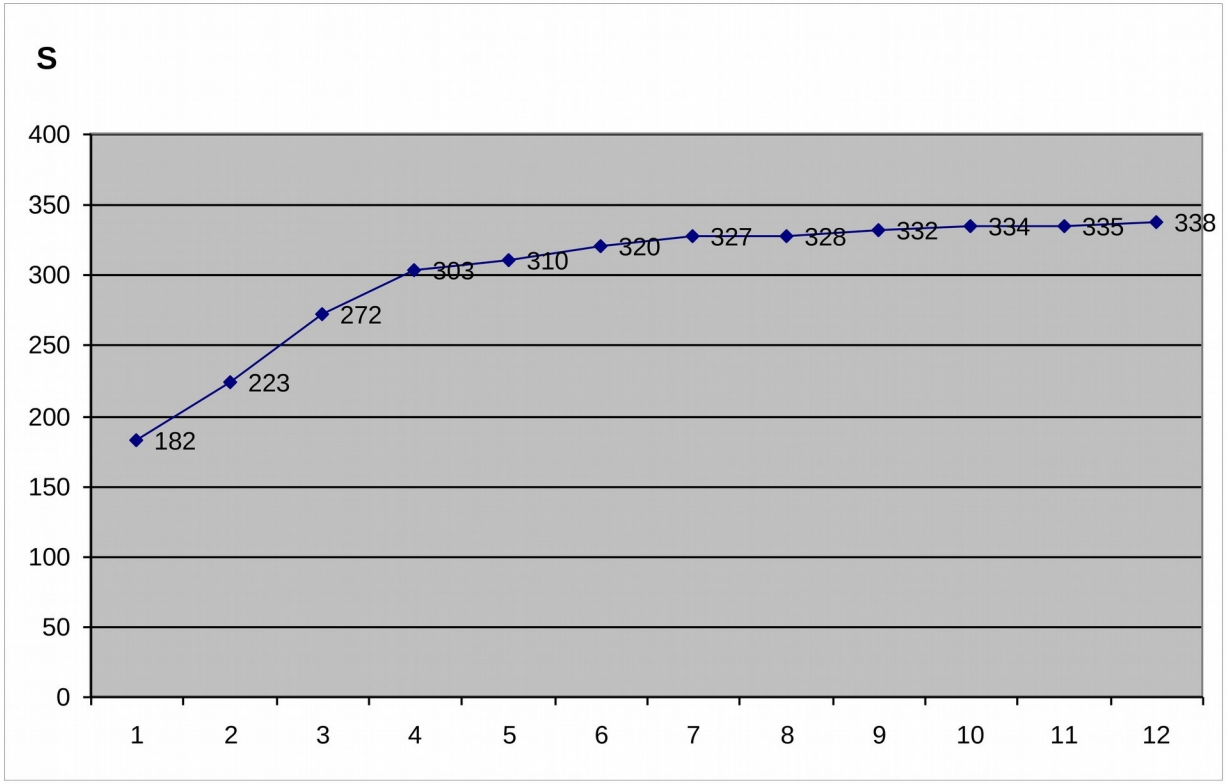
Ось X:

NGap, g

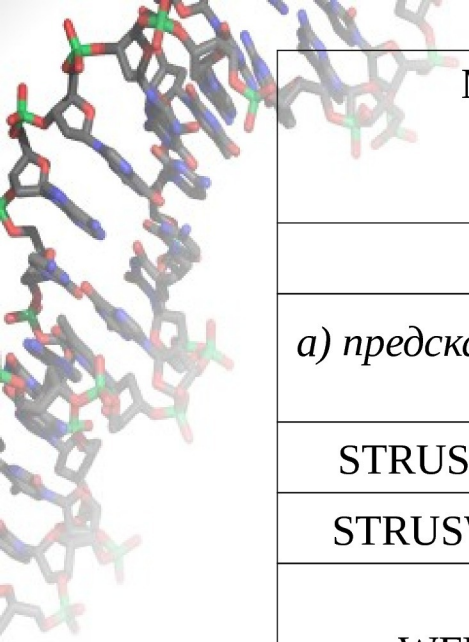
Ось Y:

WeightMatch, S(g)

$$Dcr \leq \log(L)/\log(1/(1-p)) - 1$$



$$D(g) = S(g) - S(g-1)$$



Метод	bonus	GOP	GEP	Acc	Conf	Acc, ID < 30%	Conf, ID < 30%
SW	-	7	1	0.53	0.59	0.353	0.429
а) предсказание вторичной структуры по последовательности							
STRUSWER_SIN_S	2	10	1	0.58	**0.62**	0.428	**0.482**
STRUSWER_SIN_%	7	8	2	**0.6**	0.62	**0.461**	0.477
WFMFL_SIN	матрица	13	1	0.4	0.49	0.263	0.346
б) предсказание вторичной структуры с привлечением данных о гомологичных белках							
STRUSWER_PSI_S	8	9	1	0.66	0.68	0.546	0.573
STRUSWER_PSI_%	17	6	2	**0.68**	**0.7**	**0.579**	**0.589**
WFMFL_PSI	матрица	16	1	0.63	0.67	0.503	0.56
в) экспериментально известная структура							
STRUSWER_EXP	8	10	1	**0.68**	**0.7**	**0.577**	0.601
WFMFL_EXP	матрица	15	1	0.64	0.7	0.527	**0.602**