

Межфакультетский курс «Биоинформатика»
Факультет биоинженерии и биоинформатики МГУ
весна 2025

Лекция 9, часть 1

Алгоритмы выравнивания биологических последовательностей

С.А.Спирин
15 апреля 2025

Задача выравнивания

>CYB5_CHICK

MVGSSEAGGEAWRGRYYRLEEVQKHNNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAETLITTVQSNSSSWSNWVI
PAIAAIIIVALMYRSYMSE

>CYB5_HUMAN

MAEQSDEAVKYYTLEEIQKHNHNSKSTWLIILHMKVYDLTKFLEEHPGGEEVLREQAGGDAT
ENFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKPPETLITIDSSSSSWWTNWVIPAISA
VAVALMYRLYMAED

Как сопоставить буквы одной последовательности буквам другой?

Хочется не возиться с каждой парой последовательностей, а поручить это дело компьютеру...

Задача выравнивания

Биологическая задача: сопоставить буквы одной последовательности буквам другой, чтобы соответствующие буквы имели общее происхождение (построить правильное выравнивание)

Формализация: каждому возможному выравниванию сопоставить число – его качество (обычно называется “вес”, по-английски Score), так, чтобы правильное выравнивание по возможности отличалось от неправильных бóльшим весом.

Алгоритмическая задача: придумать алгоритм, находящий выравнивание с самым бóльшим весом за разумное время.

Змей-горыныч биоинформатики



Формализация 1: вес глобального выравнивания

CYB5_CHICK	1	MVGSSEAGGEAWRGRYYRLEEVDQKHNNSQSTWIIVHHRIYDITKFLDEHP	50
		.:: . . .: . .: .: .: .: .: .: .: .:	
CYB5_HUMAN	1	---MAEQSDEA--VKYYTLEEIQKHNHNSKSTWLILHHKVYDLTKFLEEHP	45
CYB5_CHICK	51	GGEEVLREQAGGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKP	100
		: :	
CYB5_HUMAN	46	GGEEVLREQAGGDATENFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKP	95
CYB5_CHICK	101	AETLITTVQSNSSSSWSNWNWVIPAIAAIIIVALMYRSYMSE-	138
		. : . : . : : : : : : : . :	
CYB5_HUMAN	96	PETLITPIDSSSSWWTNWNWVIPAISAVAVALMYRLYMAED	134

Вес выравнивания = сумма весов **позиций** выравнивания

Вес позиции, если в ней сопоставлены две буквы, равен соответствующему элементу **матрицы замен**.

За каждый символ несоответствия («гэп», в данном случае это минус) из веса вычитается некоторое положительное число («штраф за гэп»)

Матрица замен аминокислот

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

Формализация 1: вес глобального выравнивания

CYB5_CHICK	1	MVGSSEAGGEAWRGRYYRLEEVOKHNNSQSTWIIVHHRIYDITKFLDEHP	50
		.: .:. . .: . . . : . : . : . : . : . : . : . : . : . : . : . : . : .	
CYB5_HUMAN	1	---MAEQSDEA--VKYYTLEEIQKHNHNSKSTWLILHVKVYDLTKFLEEHP	45
CYB5_CHICK	51	GGEEVLREQAGGDATEFEDVGHSTDARALSETFIIGELHPDDRPKLQKP	100
		: . : . : . : . : . : . : . : . : . : . : . : . : . : . : .	
CYB5_HUMAN	46	GGEEVLREQAGGDATEFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKP	95
CYB5_CHICK	101	AETLITTVQSNSSWSNWNWVIPAIAAIIVALMYRSYMSE-	138
		. : . : . : . : . : . : . : . : . : . : . : . : . : . : . : .	
CYB5_HUMAN	96	PETLITPIDSSSSWWTNWNWVIPAISAVAVALMYRLYMAED	134

Вес выравнивания = сумма весов **позиций** выравнивания

Вес позиции, если в ней сопоставлены две буквы, равен соответствующему элементу **матрицы замен** (например, для позиции 4 он равен -1, потому что в ней сопоставлены M и S).

За каждый символ несоответствия («гэп», в данном случае это минус) из веса вычитается некоторое положительное число («штраф за гэп»)

Проблема 1: адекватность формализации

- Оптимальное выравнивание = максимальное по весу
- Эволюционно правильное выравнивание: в каждой колонке стоят гомологичные (имеющие общее происхождение) буквы.

Это не всегда одно и то же!

Проблема 2: реализация

Нужен алгоритм, находящий оптимальное выравнивание.

Решение 1: перепробуем все варианты выравнивания и для каждого посчитаем вес, потом выберем лучшее.

Можно посчитать, что для двух последовательностей длины 100 имеется около 10^{59} различных вариантов их выравнивания.

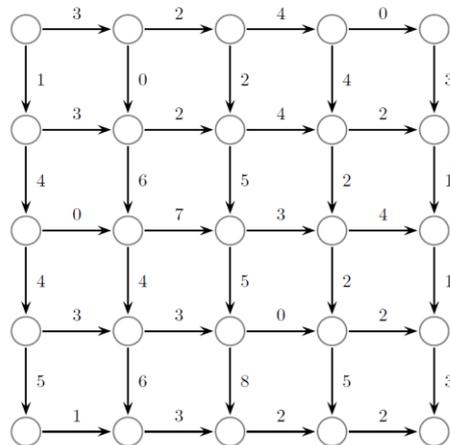
Ни один компьютер не переберёт все варианты даже за тысячу лет...

Проблема 2: реализация

Нужен алгоритм, находящий оптимальное выравнивание.

Решение 2: алгоритм динамического программирования – от части к целому.

Примером алгоритма динамического программирования может служить поиск оптимального пути по городу с прямоугольной планировкой — **“Manhattan Tourist problem”** (см., например, <https://slideplayer.com/slide/8282966/>)



Проблема 2: реализация

Нужен алгоритм, находящий оптимальное выравнивание.

Решение 2: алгоритм динамического программирования – от части к целому.

Идея: посчитаем оптимальное выравнивание сначала для начальных отрезков – «префиксов» обеих последовательностей.

Формула для веса выравнивания подобрана так, что если мы знаем оптимальное выравнивание для:

1. первых k букв (k -префикса) первой последовательности и m -префикса второй
2. $(k+1)$ -префикса первой и m -префикса второй
3. k -префикса первой и $(m+1)$ -префикса второй

то мы можем найти оптимальное выравнивание $(k+1)$ -префикса первой и $(m+1)$ -префикса второй, перебрав всего три варианта!

Идея динамического программирования

Пусть известны:

1. Оптимальное выравнивание первых k букв первой и m букв второй

```
Seq1  MVGSSEAGGEAWRGRYYRLEEV           $k$   
      |...||  .:|.|.||||:  
Seq1  ---MAEQSDEA--VKYYTLEEI           $m$ 
```

2. Оптимальное выравнивание первых k букв первой и $m + 1$ букв второй

```
Seq1  MVGSSEAGGEAWRGRYYRLEEV-           $k$   
      |...||  .:|.|.||||:  
Seq1  ---MAEQSDEA--VKYYTLEEIQ           $m+1$ 
```

3. Оптимальное выравнивание первых $k + 1$ букв первой и m букв второй

```
Seq1  MVGSSEAGGEAWRGRYYRLEEVQ           $k+1$   
      |...||  .:|.|.||||:  
Seq1  ---MAEQSDEA--VKYYTLEEI-           $m$ 
```

Идея динамического программирования

Теперь, чтобы найти оптимальное выравнивание первых $k+1$ букв первой и $m+1$ букв второй, надо выбрать всего между тремя вариантами:

Seq1 MVGSSEAGGEAWRGRYYRLEEVQ $k+1$

|...|| .:|.|.::|

Seq1 ---MAEQSDEA--VKYYTLEEIQ $m+1$

Seq1 MVGSSEAGGEAWRGRYYRLEEV-Q $k+1$

|...|| .:|.|.::|

Seq1 ---MAEQSDEA--VKYYTLEEIQ- $m+1$

Seq1 MVGSSEAGGEAWRGRYYRLEEVQ- $k+1$

|...|| .:|.|.::|

Seq1 ---MAEQSDEA--VKYYTLEEI-Q $m+1$

Алгоритм Нидлмана – Вунша (Needleman&Wunsch)

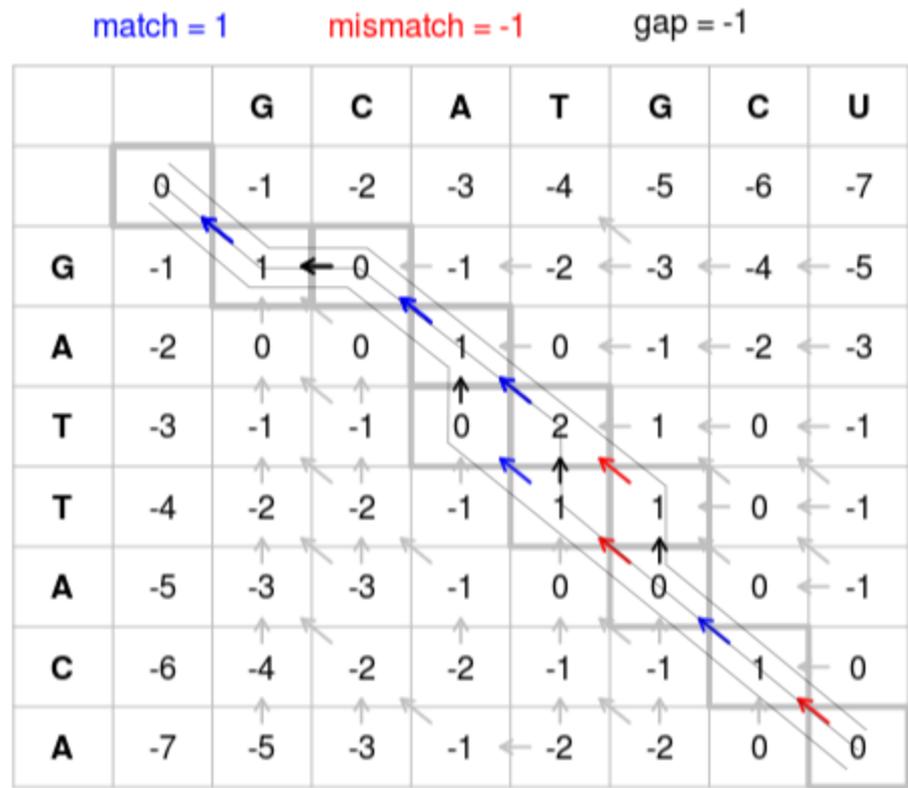


Рис. из Википедии

Матрица $(n_1+1) \times (n_2+1)$ заполняется слева направо и сверху вниз весами лучших выравниваний двух **префиксов** исходных последовательностей и стрелками. Стрелка показывает последний шаг **лучшего** пути в данную клетку. После заполнения матрицы выравнивание восстанавливается движением по стрелкам, начиная с правого нижнего угла.

Алгоритм Нидлмана – Вунша (Needleman&Wunsch)

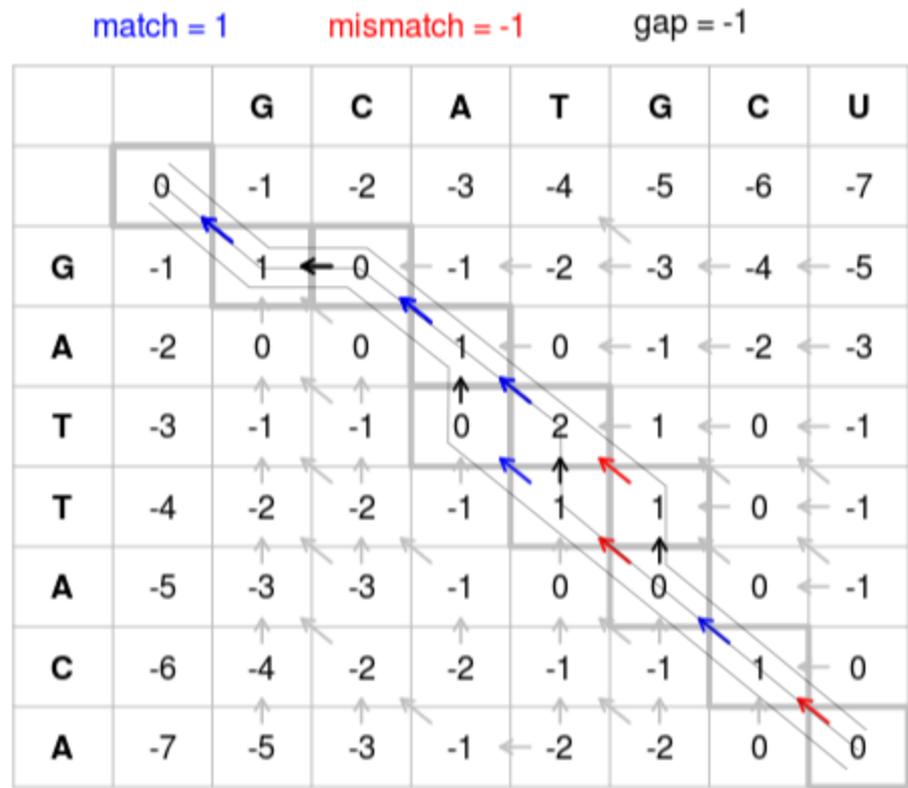


Рис. из Википедии

Время работы пропорционально произведению длин последовательностей (что намного меньше числа различных выравниваний, равного числу сочетаний из n_1+n_2 по n_1)

Формализация 1а: аффинные штрафы за гэпы

В реальности одна большая делеция более вероятна, чем две малых той же суммарной длины. Поэтому хочется вычитать из веса выравнивания штраф за делецию/вставку не в виде суммы по всем гэпам, а более умным способом, зависящим от длины инделя.

Но: не удаётся создать аналог алгоритма Нидлмана – Вунша для произвольной функции зависимости величины штрафа от длины инделя 😞

Компромисс: штраф имеет вид $G + E \cdot L$, где L – длина инделя, G и E – положительные числа, причём $G > E$.

Для такой (аффинной) зависимости штрафа от длины существует аналог алгоритма Нидлмана – Вунша, работающий всего в три раза медленнее исходного.

Формализация 2: вес локального выравнивания

Часто бывает, что нужно выровнять короткую и длинную последовательность, причём в длинной есть небольшой кусок, сходный с короткой. Алгоритм Нидлмана – Вунша (точнее, соответствующая формализация) в этом случае работает плохо из-за большого груза штрафов за концевые гэпы.

Вес локального выравнивания вычисляется только по позициям, заключённым между первым и последним сопоставлением букв (концевые гэпы игнорируются).

Разработан алгоритм Смита – Уотермана (Smith&Waterman), похожий на алгоритм Нидлмана – Вунша, но выдающий оптимальное по такому весу (так называемое **оптимальное локальное**) выравнивание.

Локальное выравнивание

Задачу локального выравнивания можно сформулировать так:

1. выбрать фрагмент первой последовательности и
2. фрагмент второй последовательности и
3. выравнивание этих фрагментов
так чтобы вес был максимальным (среди всех возможных в пунктах 1, 2, 3 выборов).

Пример локального выравнивания

```
>HXB7_HUMAN P09629 Homeobox protein Hox-B7
MSSLYYANTLFSKYPASSSVFATGAFPEQTSCAFASNPQRPGYGAGSGASFAASMQLYF
GGGGMAGQSAAGVYAAGYGLEPSSFNMHCAFPEQNLSGVC PGDSAKAAGAKEQRDSDLAA
ESNFRIYPWMRSSGTDRKRGRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAHTLCLTERQ
IKIWFQNRRMKWKKENKTAGPGTTGQDRAEAEAEAEAE
```

```
>ANTP_DROME P02833 Homeotic protein antennapedia
MTMSTNNCESMTSYFTNSYMGADMHHGHYPGNGVTDLDAQQMHHYSQNANHQGNMPYPRF
PPYDRMPYYNGQGMDDQQQHQQVYSRPSDSSQVGGVMPQAQTNGQLGVPQQQQQQQQQPS
QNQQQQQAQQAPQQQLQQQLPQVTQQVTHPQQQQQQPVVYASCKLQAAVGGGLGMVPEGGSP
PLVDQMSGHHMNAQMTLPHHMHPQAQLGYTDVGVDPVTEVHQNHNMGMYYQQSGVPPV
GAPPQGMHQGGQPPQMHQGHQHTPPSQNPNSQSSGMPSPLYPWMRSQFGKQCERKRG
RQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTG
EPGSGGEGDEITPPNSPQ
```

HXB7_HUMAN	126	IYPWMRSS-G--TDRKRGRQTYTRYQTLELEKEFHYNRYLTRRRRIEIAH	172
		: . .: :	
ANTP_DROME	283	LYPWMRSQFGKQCERKGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAH	332
HXB7_HUMAN	173	TLCLTERQIKIWFQNRRMKWKKENKTAG-PGTTGQ	206
		. . : :	
ANTP_DROME	333	ALCLTERQIKIWFQNRRMKWKKENKTGEPGSGGE	367

Алгоритм даёт ответ всегда, даже если осмысленного ответа не существует!

Оптимальное выравнивание двух неродственных последовательностей:

- глобальное:

CYB5_CHICK	1	MVGSSEAGGEAWRGRYYRLEEVE-----QKHNNNSQSTWIIIVHHRIYDI	42
		::: :.....:	
MTAL2_YEAST	1	-----MNKIPIKDLLNPQITDEFKSSILDINKKLFISI	32
CYB5_CHICK	43	TKFLDEHPGG----EEV-LREQAG--GDATENFEDVGHSTDARALSETFI	85
	:. .. :.. .. .:. .:.	
MTAL2_YEAST	33	CCNLPKLPESVTTEEEVELRDILGFLSRANKN-----RKIS----	68
CYB5_CHICK	86	IGELHPDDRPKLQKPAETLITTV-----QSNSSWSNWWIPAIAAIIV	128
		:.. .:.:.. . : : :.....:..	
MTAL2_YEAST	69	-----DEEKKLLQTTSQLTTTITVLLKEMRSIENDRSNYQLTQKNKSAD	112
CYB5_CHICK	129	ALMYRSYMSE-----	138
		. :.....:	
MTAL2_YEAST	113	GLVFNVVTQDMINKSTKPYRGHRFTKENVRILESWFAKNIENPYLDTKGL	162
CYB5_CHICK	138	-----	138
MTAL2_YEAST	163	ENLMKNTSLSRIQIKNWVSNRRRKEKTITIAPELADLLSGEPLAKKKE	210

- локальное:

CYB5_CHICK	92	DDRPKLQKPAETLITTV	108
		:.. .:.:.. . :	
MTAL2_YEAST	69	DEEKKLLQTTSQLTTTI	85

А ещё бывает множественное выравнивание

CLUSTAL W (1.83) multiple sequence alignment

```
CYB5_CHICK      MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
CYB5_HUMAN      ---MAEQSDEAV--KYYTLEEIQKHNHNSKSTWLI LHHKVYDLTKFLEEHPGGEEVLREQA
CYB5_HORSE      ---MAEQSDKAV--KYYTLEEIKKHNHNSKSTWLI LHHKVYDLTKFLEDHPGGEEVLREQA
CYB5_MUSDO      -----MSSEDV--KYFTRAEVAKNNTKDKNWFIIHNNVYDVTAFLNEHPGGEEVLIEQA
CYB5_DROME      -----MSSEET--KTFTRAEVAKHNTNKDTWLLIHNNIYDVTAFLNEHPGGEEVLIEQA
```

```
CYB5_CHICK      GGDATENFEDVGHSTDARALSETFIIGELHPDDRPKLQKPAE-TLITTVQSNSSSSWSNWV
CYB5_HUMAN      GGDATENFEDVGHSTDAREMSKTFIIGELHPDDRPKLNKPPE-TLITTIIDSSSSWWTNWV
CYB5_HORSE      GGDATENFEDIGHSTDARELSKTFIIGELHPDDRSKIAPVE-TLITTVDSNSSWWTNWV
CYB5_MUSDO      GKDATEHFEDVGHSSDAREMMKQYKVGELVAEERSNVPEKSEPTWNTQKTEESSMKS WL
CYB5_DROME      GKDATENFEDVGHNSDARDMMKKYKIGELVESERTSVAQKSEPTWSTEQQTEESSVKS WL
```

```
CYB5_CHICK      IPAIAAIIIVALMYRSYMSE---
CYB5_HUMAN      IPAISAVAVALMYRLYMAED--
CYB5_HORSE      IPAISAVVVALMYRIYTAED--
CYB5_MUSDO      MPFVLGLVATLIYKFFFGTKSQ
CYB5_DROME      VPLVLCLVATLIFYKFFFGGAKQ
```

Визуализация множественного выравнивания

H:\Talks\WFK-2014\cyb5_ali.fasta

File Edit Select View Format Colour Calculate Web Service

```

      10      20      30      40      50      60      70      80      90
CYB5_CHICK/1-138  M V G S S E A G G E A W R G R Y Y R L E E V Q K H N N S Q S T W I I V H H R I Y D I T K F L D E H P G G E E V L R E Q A G G D A T E N F E D V G H S T D A R A L S E T F I I G E L H P D D R
CYB5_HUMAN/1-134  - - - M A E Q S D E A V - - K Y Y T L E E I Q K H N H S K S T W L I L H H K V Y D L T K F L E E H P G G E E V L R E Q A G G D A T E N F E D V G H S T D A R E M S K T F I I G E L H P D D R
CYB5_HORSE/1-134  - - - M A E Q S D K A V - - K Y Y T L E E I K K H N H S K S T W L I L H H K V Y D L T K F L E D H P G G E E V L R E Q A G G D A T E N F E D I G H S T D A R E L S K T F I I G E L H P D D R
CYB5_MUSDOV/1-134  - - - - - M S S E D V - - K Y F T R A E V A K N N T K D K N W F I I H N N V Y D V T A F L N E H P G G E E V L I E Q A G K D A T E H F E D V G H S S D A R E M M K Q Y K V G E L V A E E R
CYB5_DROME/1-134  - - - - - M S S E E T - - K T F T R A E V A K H N T N K D T W L L I H N N I Y D V T A F L N E H P G G E E V L I E Q A G K D A T E N F E D V G H S N D A R D M M K K Y K I G E L V E S E R
  
```

Conservation



.....685745..979754*94*5*5554+*8+9*569**9*5**79*****5*****5*****5**9**9**49676969**4369*

Quality



Consensus



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



Sequence 1 ID: CYB5_CHICK Residue: ARG (13)

Алгоритмы множественного выравнивания

Все алгоритмы **эвристические**. Это значит, что они не решают никакую точно сформулированную задачу (в отличие от алгоритма Нидлмана – Вунша)

- Muscle
- MAFFT
- Probcons
- Pride
- ClustalO
- DiAlign
- ...

Задачи биоинформатики последовательностей

- Выравнивание:
 - парное глобальное
 - парное локальное
 - множественное
- Поиск в банках данных (BLAST)
- Предсказание генов
- Реконструкция филогении
- Поиск функциональных мотивов в белках, РНК, ДНК
- Сборка геномов (по результатам секвенирования)
- Предсказание 3D структуры по последовательности
- ...

Задачи биоинформатики 3D структур

- Визуализация
- Автоматическая разметка структур (выделение доменов, элементов вторичной структуры, полостей и т.д.)
- Выравнивание структур
- Предсказание связывания с лигандами
- ...