Мутации и выравнивание

Понятие выравнивания последовательностей. Выравнивание как отражение эволюции последовательностей. Вес выравнивания. Глобальное и локальное выравнивание последовательностей.

Сергей Александрович Спирин
Юлия Александровна Алешина

7 апреля 2026

План

1. Введение: гомологичные белки
2. Источники гомологичных белков
   - Мутации:
     - Ошибки репликации.
     - Повреждения ДНК и их репарация.
   - Закрепление мутаций

2. Выравнивание:
   - последовательностей потомков относительно предка;
   - двух потомков одного предка.

3. Формализация: вес выравнивания

4. Программы парного выравнивания в EMBOSS

5. Редактор выравниваний Jalview

Последовательности миоглобинов человека, мыши и быка

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>MYG_MOUSE
MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE
DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH
SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG
>MYG_BOVIN
MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE
DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH
PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG
```

Разместим последовательности друг под другом, чтобы было видно сходство

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_MOUSE MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60
MYG_BOVIN MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
          **************.*******: ******** **. *****:******:**:* :**.**
MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_MOUSE DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH 120
MYG_BOVIN DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH 120
          ****** ******* ******:* **:: **:***.**************: **.**: :*
MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_MOUSE SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG 154
MYG_BOVIN PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG 154
           .*******.**.*******:*:*::** ***:*
```

Видно, что большинство букв совпадает, но некоторые различаются.
Это последовательности **гомологичных белков**. Это означает, что данные белки произошли от общего предка.
За время, прошедшее от существования общего предка, некоторые позиции поменялись, но большинство остались прежними.

Сравнение последовательностей

Последовательности миоглобинов человека и рыбы:

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>MYG_DANRE
MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
GQGALRRVMDAVIGDIDGYYKEIGFAG
```

Как сравнивать, если разная длина?

Сравнение последовательностей

Последовательности миоглобинов человека и рыбы:

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
>MYG_DANRE
MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
GQGALRRVMDAVIGDIDGYYKEIGFAG
```

Как сравнивать, если разная длина?
Ответ: выравнивание!

Сравнение последовательностей

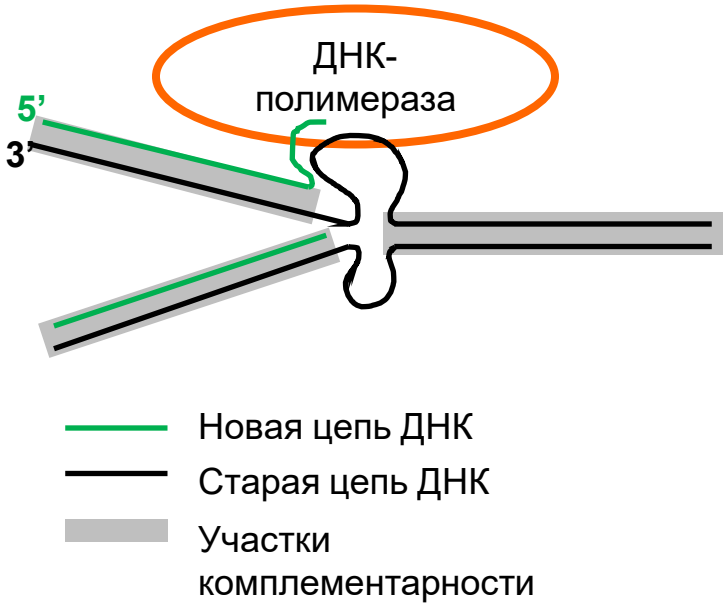
```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_DANRE ----MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQG-DLAGSP 55
              .: :***: ** ****  .:* *** **** :*:**: * **. :..  :: .*

MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_DANRE AVAAHGATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKA 115
           :  ******. ** :** **.* * :****:**. **: :: :.:*:* :::*: .*

MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_DANRE --GLDAAGQGALRRVMDAVIGDIDGYYKEIGFAG 147
            .:.* .***:.:.:: .  *: . ***:** *
```

- Выравнивание – размещение двух (или более) последовательностей друг под другом так, чтобы **сходные участки** этих последовательностей оказались друг под другом.
- Против букв, которым ничего не соответствует в другой последовательности, ставят знак «гэпа» (=пропуска, пробела). Здесь это дефис.

Источники разнообразия геномов

Ошибки репликации

- Генетическая информация (последовательности белков закодированы в ДНК) передается от поколения к поколению за счет процесса **репликации** ДНК, который предшествует клеточному делению
- При копировании ДНК-полимераза бывают **ошибки** - мутации
- ДНК любого ныне живущего организма получилась переписыванием ДНК организма, жившего примерно 3,5 млрд лет тому назад (LUCA).
- Текст ДНК, конечно, изменился до неузнаваемости. Но родство (гомологичность) последовательностей некоторых белков во всех современных организмах устанавливается достаточно надежно.



Источники разнообразия геномов

Повреждения ДНК и их репарация

- Имеется много источников повреждений ДНК:
  - ультрафиолетовое излучение,
  - различные химические вещества, содержащиеся в пище, воздухе, табачном дыме…
  - некоторые ферменты самого организма
- Повреждения ДНК контролируются клеткой и **репарируются**.

Увы, не всегда правильно. Одно из следствий – онкологические заболевания.

Гомологичные последовательности

```
>First
CGTTCCCGGGTCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCG
GTGGAGTAACCATTTGGAGCTAGCCGTCGAAGGTGGG
>Second
CGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGAAGTCTGCAATGCCCAAAGTCG
GTGGGATAACCTTTATAGGAGTCAGCCGCCTAAGGCAGG
```

Выравнивание
(демонстрирует сходство)

```
>First
CGTTCCCGGGTCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCG
GTGGAGTAACCATTTGGAGCTAGCCGTCGAAGGTGGG
>Second
CGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGAAGTCTGCAATGCCCAAAGTCG
GTGGGATAACCTTTATAGGAGTCAGCCGCCTAAGGCAGG
```

88/100 нт идентичны

```
First    1 CGTTCCCGGGTCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAAC  50
           ||||||||||.|||||||||||||||||||||||.|..|||.||.||.
Second   1 CGTTCCCGGGCCTTGTACACACCGCCCGTCACACCATGGAAGTCTGCAAT  50

First   51 ACCCGAAGCCGGTGGAGTAACCAT--TTGGAGCTAGCCGTCGAAGGTGGG  98
           .|||.|||.||||||..|||||.|  |.||||..|||||.|.||||..||
Second  51 GCCCAAAGTCGGTGGGATAACCTTTATAGGAGTCAGCCGCCTAAGGCAGG 100
```

Негомологичные последовательности

```
>First
CGTTCCCGGGTCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCG
GTGGAGTAACCATTTGGAGCTAGCCGTCGAAGGTGGG
>Third
CCTGCCTTAGGCGGCTGACTCCTATAAAGGTTATCCCACCGACTTTGGGCATTGCAGACTT
CCATGGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACG
```

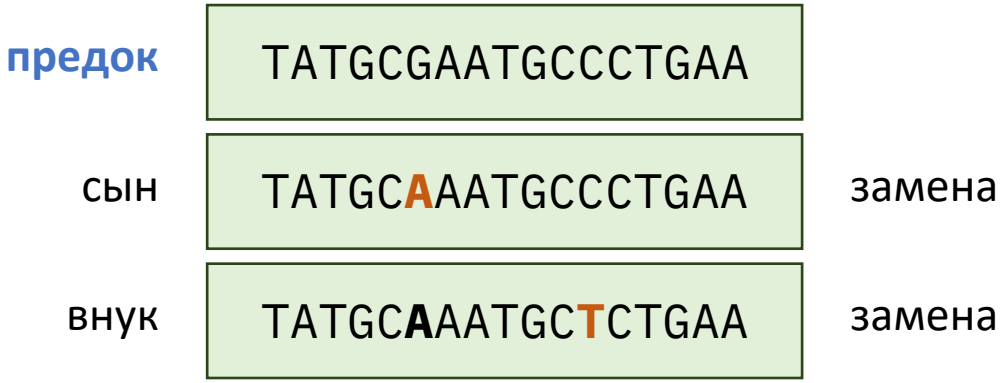
Выравнивание (бессмысленное)

```
>First
CGTTCCCGGGTCTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCG
GTGGAGTAACCATTTGGAGCTAGCCGTCGAAGGTGGG
>Third
CCTGCCTTAGGCGGCTGACTCCTATAAAGGTTATCCCACCGACTTTGGGCATTGCAGACTT
CCATGGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACG
```

28/100 позиций совпадают

```
First   1 -----------------------------------CGTTCCCGGGT  11
                                              ||.....|||.
Third   1 CCTGCCTTAGGCGGCTGACTCCTATAAAGGTTATCCCACCGACTTTGGGC  50

First  12 CTTGTACACACCGCCCGTCACACCACGAGAGTTTGTAACACCCGAAGCCG  61
          .|||.|.||..|....||...||.....|.||.|..||..||||....||
Third  51 ATTGCAGACTTCCATGGTGTGACGGGCGGTGTGTACAAGGCCCGGGAACG 100
```

Выравнивание последовательностей потомков относительно предка



Выравнивание последовательностей потомков относительно предка

11 из 17 позиций (колонок) выравнивания консервативны – не изменились от предка

Выравнивание как отражение эволюции последовательностей

- **Гомологичные последовательности** – последовательности, произошедшие от одного предка путем многократного переписывания генома
- При репликации почти всегда каждый нуклеотид потомка происходит «от» определенного нуклеотида предка.
- В выравнивании гомологичных последовательностей у разных потомков одного и того же предка **гомологичные нуклеотиды (аминокислоты) должны стоять в одной колонке**.

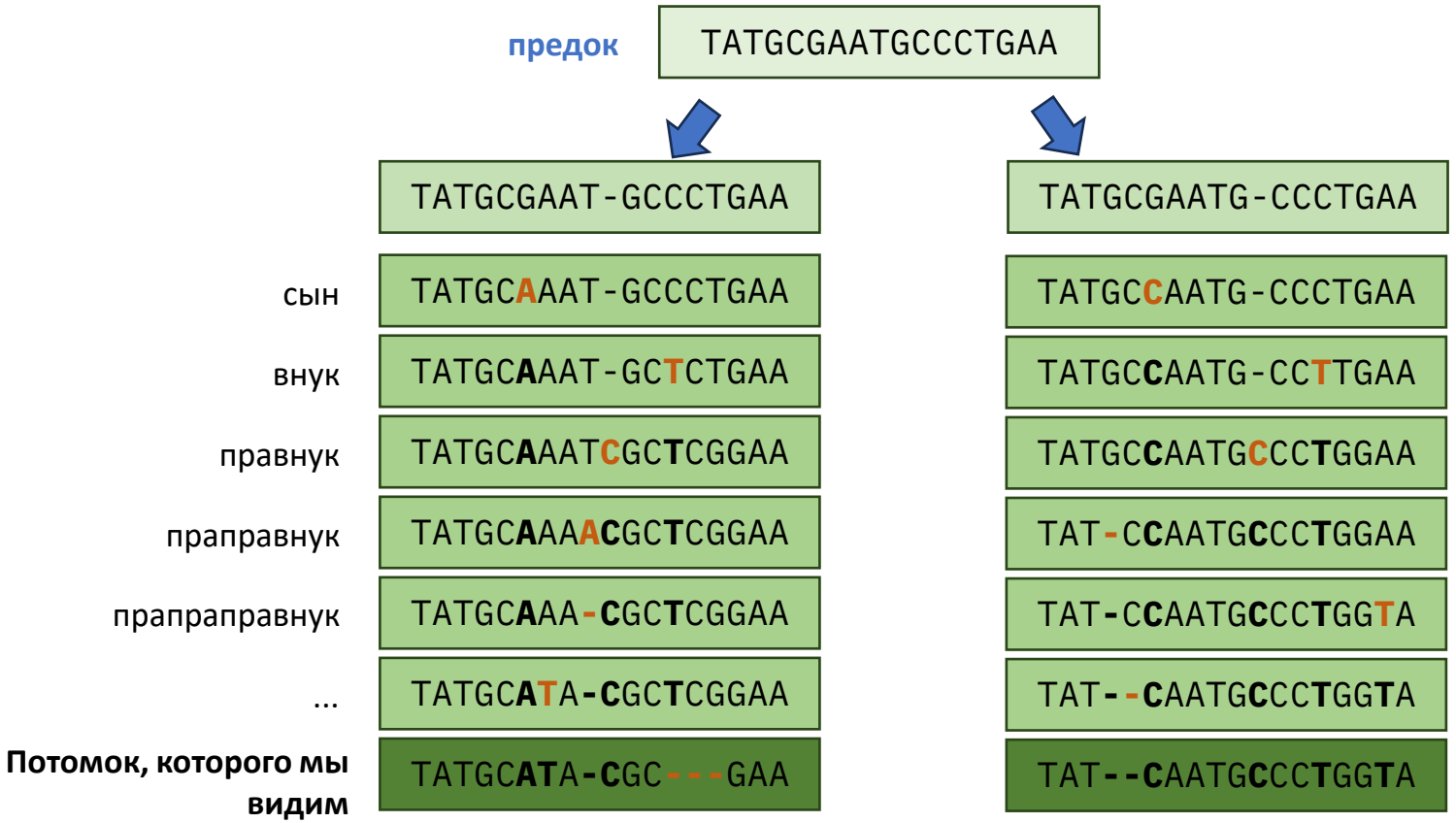
предок	TATGCGAAT-GCCCTGAA
сын	TATGC**A**AAT-GCCCTGAA
внук	TATGC**A**AAT-GC**T**CTGAA
правнук	TATGC**A**AAT**C**GC**T**CGGAA
праправнук	TATGC**A**AA**A**C**G**C**T**CGGAA
прапраправнук	TATGC**A**AA**-**C**G**C**T**CGGAA
…	TATGC**AT**A**-C**GC**T**CGGAA
…	TATGC**AT**A**-C**GC**---**GAA

Выравнивание как отражение эволюции последовательностей

- Такое выравнивание бывает только в экспериментах по изучению эволюции:
  - E.coli (Ленски)
  - Schizophyllum (А.Кондрашов)
  - вирусы и др.

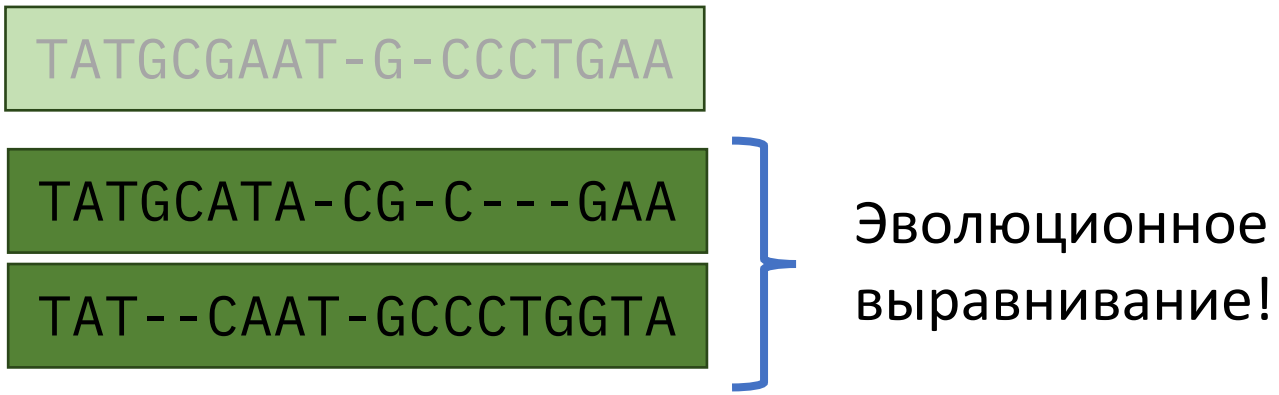
Schizophyllum commune

предок	TATGCGAAT-GCCCTGAA
сын	TATGC**A**AAT-GCCCTGAA
внук	TATGC**A**AAT-GC**T**CTGAA
правнук	TATGC**A**AAT**C**GC**T**CGGAA
праправнук	TATGC**A**AA**A**C**G**C**T**CGGAA
прапраправнук	TATGC**A**AA**-C**GC**T**CGGAA
…	TATGC**AT**A**-C**GC**T**CGGAA
…	TATGC**AT**A**-C**GC**---**GAA


предок
TATGCGAATGCCCTGAA
TATGCGAAT-GCCCTGAA
TATGCGAATG-CCCTGAA
сын
TATGCAAAT-GCCCTGAA
TATGCCAATG-CCCTGAA
внук
TATGCAAAT-GCTCTGAA
TATGCCAATG-CCTTGAA
правнук
TATGCAAATCGCTCGGAA
TATGCCAATGCCCTGGAA
праправнук
TATGCAAAACGCTCGGAA
TAT-CCAATGCCCTGGAA
прапраправнук
TATGCAAA-CGCTCGGAA
TAT-CCAATGCCCTGGTA
...
TATGCATA-CGCTCGGAA
TAT--CAATGCCCTGGTA
Потомок, которого мы видим
TATGCATA-CGC---GAA
TAT--CAATGCCCTGGTA

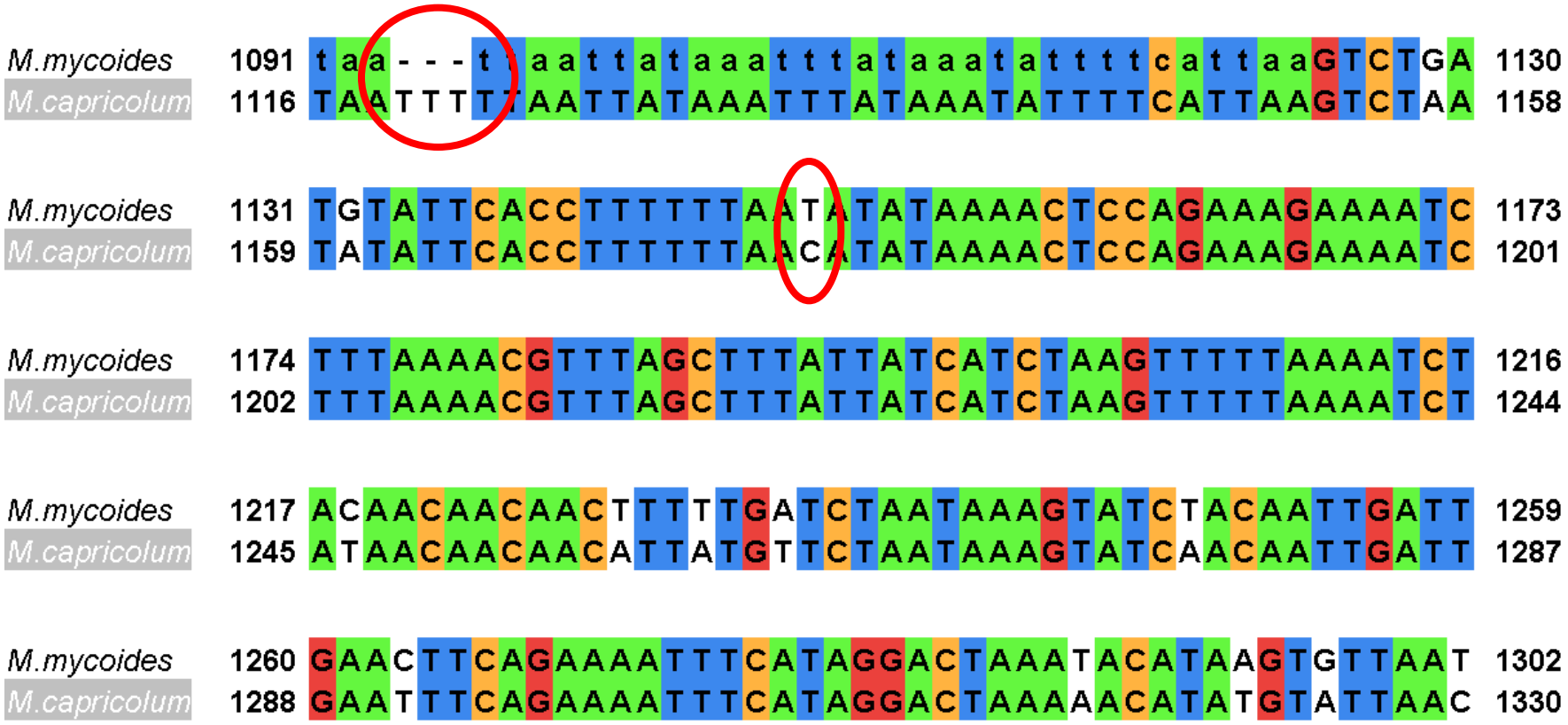
Выравнивание отражает эволюцию

- Как правило нам известны последовательности только **современных** организмов
- Мы решаем обратную задачу – построить эволюционное выравнивание, зная последовательности потомков



Выравнивание геномов двух потомков общего предка микоплазм: *M.capricolum* и *M.mycoides*

(маленький фрагмент)



В среднем 92% совпадающих букв на **гомологичных** участках

ДНК и белки

Геном
3 млрд. букв у человека
0,5–10 млн. букв у бактерий

содержит →

Кодирующие участки
< 2% у человека
~90% у бактерий

кодируют →

Белки
~25 000 у человека
600–6000 у бактерий

Генетический код

	T(U)	C	A	G
T(U)	TTT Phe TTC Phe TTA Leu TTG Leu	TCT Ser TCC Ser TCA Ser TCG Ser	TAT Tyr TAC Tyr TAA Stop TAG Stop	TGT Cys TGC Cys TGA Stop TGG Trp
C	CTT Leu CTC Leu CTA Leu CTG Leu	CCT Pro CCC Pro CCA Pro CCG Pro	CAT His CAC His CAA Gln CAG Gln	CGT Arg CGC Arg CGA Arg CGG Arg
A	ATT Ile ATC Ile ATA Ile ATG Met	ACT Thr ACC Thr ACA Thr ACG Thr	AAT Asn AAC Asn AAA Lys AAG Lys	AGT Ser AGC Ser AGA Arg AGG Arg
G	GTT Val GTC Val GTA Val GTG Val	GCT Ala GCC Ala GCA Ala GCG Ala	GAT Asp GAC Asp GAA Glu GAG Glu	GGT Gly GGC Gly GGA Gly GGG Gly

Аминокислоты

A Ala Alanine Аланин
R Arg Arginine Аргинин
N Asn Asparagine Аспарагин
D Asp Aspartic Acid Аспарагиновая кислота
C Cys Cysteine Цистеин
Q Gln Glutamine Глютамин
E Glu Glutamic Acid Глутаминовая кислота
G Gly Glycine Глицин
H His Histidine Гистидин
I Ile Isoleucine Изолейцин
L Leu Leucine Лейцин
K Lys Lysine Лизин
M Met Methionine Метионин
F Phe Phenylalanine Фенилаланин
P Pro Proline Пролин
S Ser Serine Серин
T Thr Threonine Треонин
W Trp Thryptophan Триптофан
Y Tyr Tyrosine Тирозин
V Val Valine Валин
"Stop" в таблице кода означает стоп-кодон – сигнал окончания трансляции.

Мутации

gatcaacactacttgacttcaa**g**acttaccataaagaaaac

gatcaacactacttgacttcaa**a**acttaccataaagaaaac

точечная замена

---

gatcaacactacttgacttcaa**ga**cttaccataaagaaaac

gatcaacactacttgacttcaacttaccataaagaaaac

делеция

---

gatcaacactacttgacttcaagacttaccataaagaaaac

gatcaacactacttgacttcaaga**ta**cttaccataaagaaaac

инсерция
(вставка)

Классификация мутаций в кодирующих последовательностях ДНК

❖ **Синонимичная** (амк остаток не меняется)

❖ **Миссенс (missense)**: меняет амк остаток

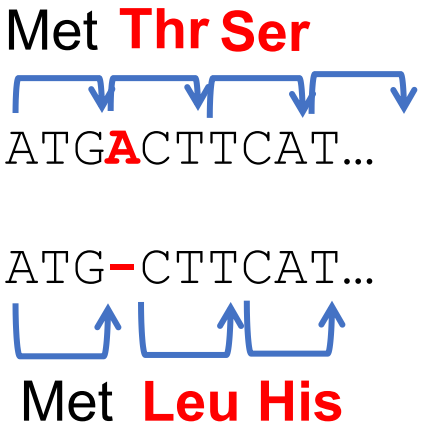
- ✓ на близкий по свойствам
- ✓ с отличными свойствами

❖ **Нонсенс (nonsense):** заменяет кодон остатка на стоп-кодон

Классификация мутаций в кодирующих последовательностях ДНК

- Сдвиг рамки (**frameshift**): вставка или делеция размера, не кратного трём

Результатом являются совсем другие аминокислоты после мутации и, как правило, стоп-кодон сравнительно недалеко (в среднем через 21 триплет после мутации)



Судьба мутации

Бактерия разделилась, и у одного из потомков произошла мутация. (ошибка репликации, или повреждение ДНК и ошибка репарации).

Что будет с потомством мутанта? Увидим ли мы эту мутацию, если отсеквенируем 1 000 000 бактерий этого штамма через 10 лет?

Потомство бактерии

В благоприятных условиях бактерия может делиться каждый час.

Сколько будет бактерий через 24 часа? А через год????

Потомство бактерии

В благоприятных условиях бактерия может делиться каждый час.

Сколько будет бактерий через 24 часа? А через год????

Ответ: примерно столько же, сколько сейчас.

Потомство бактерии

В благоприятных условиях бактерия может делиться каждый час.

Сколько будет бактерий через 24 часа? А через год????

Ответ: примерно столько же, сколько сейчас.

Численность подавляющего большинства популяций **постоянна** (по крайней мере на отрезках времени порядка лет) – погибает примерно столько же, сколько рождается.
Современная популяция человека – исключение!

Если члены популяции генетически идентичны, то вероятность оставить потомство для всех **одинакова** (точнее, зависит от только от внешних факторов).

Следствие: математическое ожидание числа потомков одной бактерии через достаточно большой промежуток времени равно 1.

Судьба нейтральной мутации

Предположим, что мутация **нейтральна** = никак не влияет на матожидание числа потомков (таких мутаций довольно много).

Мутация произошла и передаётся потомкам мутанта. Значит, в популяции появился новый **полиморфизм**. У данного варианта кода есть **частота** (сначала очень маленькая).

Судьба нейтральной мутации

Предположим, что мутация **нейтральна** = никак не влияет на матожидание числа потомков (таких мутаций довольно много).

Мутация произошла и передаётся потомкам мутанта. Значит, в популяции появился новый **полиморфизм**. У данного варианта генома есть **частота** (сначала очень маленькая).

Что произойдёт с частотой через пару суток?

Судьба нейтральной мутации

Предположим, что мутация **нейтральна** = никак не влияет на матожидание числа потомков (таких мутаций довольно много).

Мутация произошла и передаётся потомкам мутанта. Значит, в популяции появился новый **полиморфизм**. У данного варианта генома есть **частота** (сначала очень маленькая).

Что произойдёт с частотой через пару суток?

Ответ: частота либо немного возрастёт, либо немного упадёт. То и другое примерно равновероятно.

Случайное блуждание

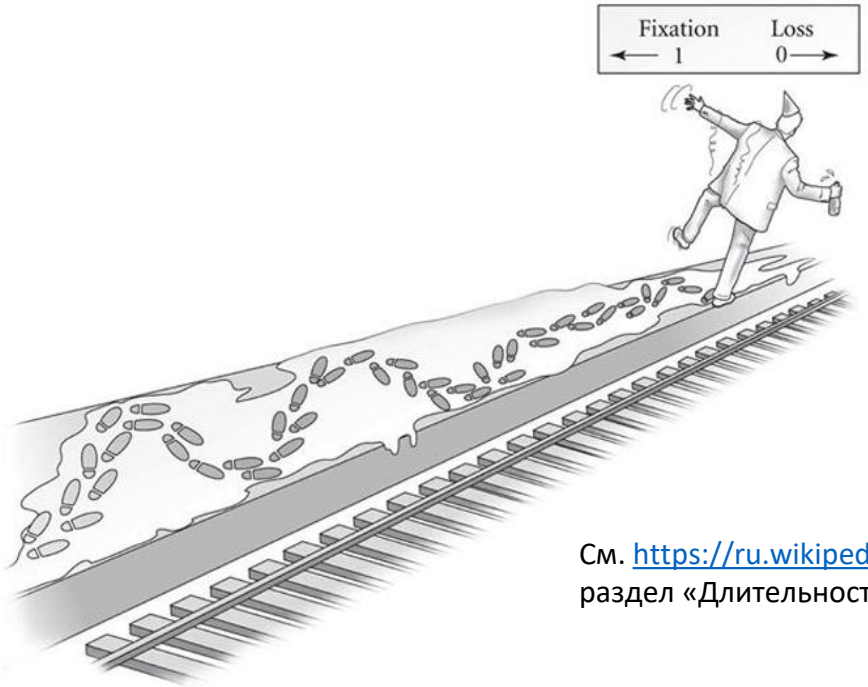
Частота любого нейтрального полиморфизма постоянно колеблется случайным образом (это называется «генетический дрейф»). Математическая модель такого процесса называется «случайное блуждание».

На тротуаре стоит пьяный и каждые 10 сек. делает шаг либо направо, либо налево, случайно выбирая направление. Как далеко он уйдёт за время T?

Ответ: в среднем на расстояние, пропорциональное корню квадратному из T.

Случайное блуждание с поглощением

По длинной дамбе идёт пьяный и с каждым шагом отклоняется либо на полметра вправо, либо на полметра влево. Как скоро он свалится с дамбы?
Ответ: *скоро…*



См. https://ru.wikipedia.org/wiki/Задача_о_разорении_игрока, раздел «Длительность случайного блуждания»

Когда частота генетического варианта достигает 100% или 0%, процесс её изменения прекращается.

За исторически короткое время любой нейтральный вариант либо исчезает из популяции, либо закрепляется в ней!

Закрепление мутаций как результат генетического дрейфа

Вероятность закрепиться для новой нейтральной мутации очень мала, но не 0.

Организмов в популяции много, мутаций в них происходит тоже много (примерно 10^{-8} на п.н. на поколение – каждая сотая новорождённая бактерия несёт новую мутацию). Значительная доля мутаций нейтральна.

Итог: геномы независимых популяций начинают различаться, чем дальше, тем больше – в них независимо накапливаются нейтральные мутации.

А если мутация не нейтральна?

Каждому варианту генома можно сопоставить его «приспособленность» f = матожидание числа потомков организма с таким геномом (через какой-то фиксированный промежуток времени).

В подавляющем большинстве случаев новая мутация порождает либо нейтральный вариант
($f = 1$) либо вредный ($f < 1$).

Вредный вариант тоже начинает «блуждать», но вероятность «шага вверх» оказывается меньше вероятности «шага вниз». Это очень сильно уменьшает вероятность закрепления – тем сильнее, чем меньше f, и тем сильнее, чем больше популяция.

Явление невозможности закрепления вредной мутации называется **стабилизирующий отбор** или же **отрицательный отбор**.

Положительный отбор

Если вдруг $f > 1$, то вероятность закрепления мутации вырастает во много раз. Процесс закрепления полезных мутаций называется **положительным отбором.**

Собственно, полезных мутаций так мало именно потому, что большинство возможных полезных мутаций уже закрепились.

Обычно полезные мутации начинают появляться в заметном количестве только при изменении условий жизни организмов – например при появлении нового источника пищи или новой опасности или попадании части популяции в другой климат…

Эволюция белков

Мутации возникают случайно.

Конкретная мутация может быть:

- летальной;
- вредной;
- слабовредной;
- нейтральной;
- полезной.

Мутация порождает **полиморфизм данного белка в популяции.**
Доля каждого варианта подвержена случайным изменением (модель: «случайное блуждание с поглощением»).
За исторически короткое время один из вариантов (старый или новый) исчезает. В первом случае говорят, что мутация **закрепилась.**

Как правило, пространственная структура белка почти не меняется при эволюции его последовательности. В первом приближении верно утверждение: **гомологичные белки имеют почти одинаковые 3D-структуры.**

Множественное выравнивание белковых последовательностей

Мы видим только закрепившиеся мутации!

Гомология – общность происхождения

- При репликации почти всегда каждый нуклеотид потомка происходит от определенного нуклеотида предка.
- В выравнивании гомологичных последовательностей у разных потомков одного и того же предка гомологичные нуклеотиды должны стоять в одной колонке.
- Как правило, нам известны геномы только современных организмов, и потому у нас нет способа проверить, какие нуклеотиды гомологичны.
- Гомологичность последовательностей часто можно установить анализом их выравнивания.
- Проблема построения выравнивания обсуждается ниже.

Выравнивание последовательностей касается всех студентов МГУ!

Положение об обеспечении самостоятельности выполнения письменных работ в МГУ имени М.В.Ломоносова на основе системы «Антиплагиат»

Самостоятельное выполнение письменных работ обучающимися в МГУ имени М.В.Ломоносова (далее – МГУ) является необходимым условием эффективности этих работ как элементов учебного процесса, развития у обучающихся навыков научной работы.

К обучающимся в Университете относятся студенты, аспиранты, докторанты, слушатели и соискатели (ст.ст. 123-128 Устава МГУ).

www.msu.ru/projects/antiplagiat/antiplagiat.doc

Для данных двух последовательностей существует много разных выравниваний

```
TGGAGTAACCAT-
TGGGATAACCTTG
```

```
TGGAGTAACCAT-----------
-----------TGGGATAACCTTG
```

```
-TGGAGTAACCAT
TGGGATAACCTTG
```

```
TGGA--GTAACCAT--
TGGGATAA---CCTTG
```

Всего для двух последовательностей одинаковой длины *n* имеется около C_{2n}^{n} разных выравниваний

~ $9*10^{58}$ для последовательностей длины 100

Биоинформатическая задача: выбрать среди множества выравниваний правильное

Алгоритм выравнивания решает математическую задачу, а не биологическую

Математическая задача разбивается на две:

- Любому выравниванию сопоставить число – его **вес**
- Для данных последовательностей построить выравнивание с наибольшим весом

Три понимания «правильного» выравнивания

1. Оптимальное выравнивание: наилучшее по весу

Его ищут программы. *Существует для любого набора последовательностей!*

2. Эволюционное выравнивание: запись, отражающая ход эволюции

Не поддается достоверной реконструкции в большинстве реальных случаев; может отличаться от оптимального выравнивания.

Алгоритм вычисления веса стараются выбрать так, чтобы можно было ожидать, что эволюционное выравнивание будет среди нескольких оптимальных.

Не существует для *негомологичных последовательностей!*

3. Функциональное выравнивание: сопоставление функционально идентичных частей белков или нуклеиновых кислот

Объясняет сохранение в эволюции одних частей белка и варьирование других. Поскольку функция и 3D-структура белка очень тесно связаны, функционально выровненные аминокислотные остатки должны иметь примерно одинаковое расположение в пространстве

Вес парного выравнивания

Простейший вариант

За каждую колонку с совпадающими буквами прибавляем число *A*
За каждую колонку с разными буквами вычитаем число *B*
За каждую «чёрточку» (гэп) вычитаем число *C*

Вес парного выравнивания

Простейший вариант

За каждую колонку с совпадающими буквами прибавляем число *A*
За каждую колонку с разными буквами вычитаем число *B*
За каждую «чёрточку» (гэп) вычитаем число *C*

```
First  TGGAGTAACCAT--TAGGAGCTAGCCG
       |||..||||||  ||||||..|||||
Second TGGGATAACCTTGATAGGAGTCAGCCG
```

Здесь 20 совпадений, 5 несовпадений, два гэпа, значит вес 20*A* – 5*B* – 2*C*
Например, при *A* =5, *B* = 4, *C* = 6 вес равен 68.
Проверьте, что ни у какого выравнивания этих последовательностей вес не будет бóльшим. Это выравнивание является оптимальным при данных параметрах *A, B, C*.

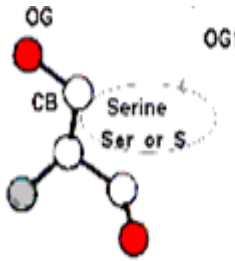
Вес парного выравнивания – белки

Какое выравнивание имеет бóльшие шансы оказаться правильным?

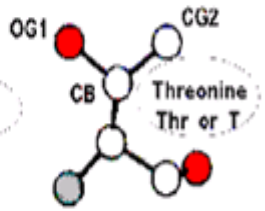
```
AF**T**GAHAYL
AY**S**---AYM
```

```
AFTGA**H**AYL
AY---**S**AYM
```

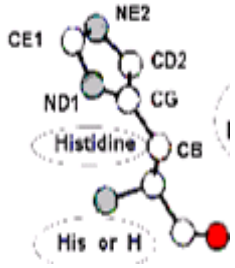
Вес парного выравнивания – белки



Серин S



Треонин T



Гистидин H

Мутация серина в треонин закрепляется с гораздо большей вероятностью по сравнению с мутацией серина в гистидин.

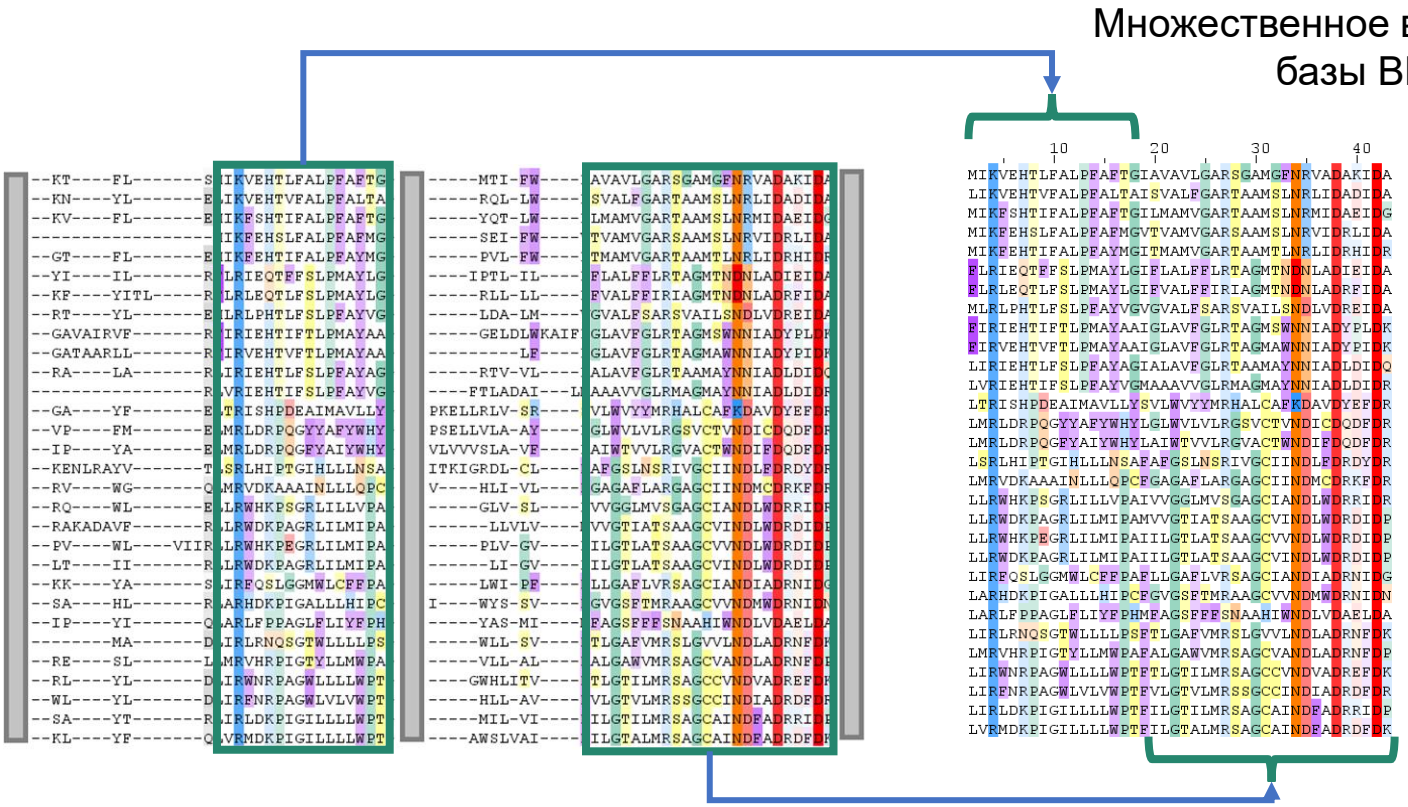
Поэтому если в одной колонке выравнивания оказались буквы S и T, это скорее аргумент за данное выравнивание, чем против него. Значит, за такую колонку лучше увеличивать вес, чем уменьшать.

Матрица весов аминокислотных замен BLOSUM62

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

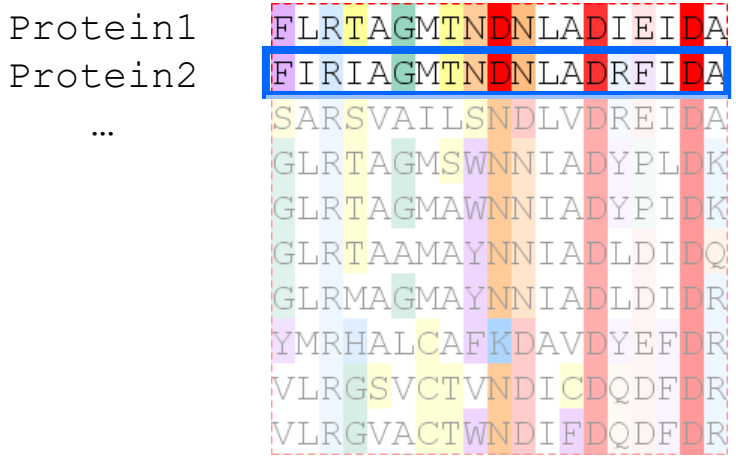
Из работы (Henikoff&Henikoff, **1992**, *PNAS*)

Матрицы замен серии BLOSUM



Матрицы замен серии BLOSUM

- Подсчитываем частоты сопоставлений пар аминокислот - q_{ij}
- Из частоты встречаемости каждой аминокислоты высчитываем ожидаемую вероятность сопоставления каждой пары - e_{ij}



$$s_{ij} = 2 \cdot \log_2 \frac{q_{ij}}{e_{ij}}$$

Наблюдаемая вероятность для пары ***i, j***

Ожидаемая вероятность для пары ***i, j***

Матрицы замен серии BLOSUM

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

$$\frac{q_{ij}}{e_{ij}} = 1$$

Частота встречаемости пары не отличается от ожидаемой случайно

$$\frac{q_{ij}}{e_{ij}} > 1$$

Пара встречается чаще, чем ожидалось бы случайно

$$\frac{q_{ij}}{e_{ij}} < 1$$

Пара встречается реже, чем ожидалось бы случайно

Матрицы замен серии BLOSUM

BLOSUM N – последовательности, использованные для построения матрицы, имеют менее N% сходства

- BLOSUM80
- BLOSUM62
- BLOSUM50
- BLOSUM45

Матрицы замен серии BLOSUM

BLOSUM N – последовательности, использованные для построения матрицы, имеют менее N% сходства

- BLOSUM80
- BLOSUM62
- BLOSUM50
- BLOSUM45

близкородственность

PHAT – матрица для трансмембранных белков

Матрица BLOSUM62

```
#  Matrix made by matblas from blosum62.iij
#  * column uses minimum score
#  BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
#  Blocks Database = /data/blocks_5.0/blocks.dat
#  Cluster Percentage: >= 62
#  Entropy =   0.6979, Expected =  -0.5209
   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  1 -1 -4
Z -1  0  0  1 -3  3  4 -2  0 -3 -3  1 -1 -3 -1  0 -1 -3 -2 -2  1  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
```

B - Asx, Z - Glx, X – unknown, * – stop

Вес парного выравнивания – белки

Посчитаем веса выравниваний, используя матрицу BLOSUM62

```
AFTGAHAYL          AFTGAHAYL
AYS---AYM          AY---SAYM
```

Обозначим значения матрицы на пересечении строки A и столбца B через M(A,B)

Тогда вес левого выравнивания равен:
M(A,A) + M(F,Y) + **M(T,S)** + M(A,A) +M(Y,Y) + M(L,M) – штраф за гэпы,
а правого:
M(A,A) + M(F,Y) + **M(H,S)** + M(A,A) +M(Y,Y) + M(L,M) – штраф за гэпы.

Штрафы за гэпы одинаковы, значит веса различаются слагаемым M(T,S) слева против M(H,S) справа. Но M(T,S) = 1, а M(H,S) = –1, поэтому вес левого выравнивания больше на 2.

Вес парного выравнивания: аффинные штрафы за гэпы

```
First  TGGAGTAACCAT--TTGGAGCTAGCCG
       |||..|||||.|  |.||||..|||||
Second TGGGATAACCTTTATAGGAGTCAGCCG
```

Выравнивание 1

```
First  TGGAGTAACCAT-TT-GGAGCTAGCCG
       |||..|||||.| .| ||||..|||||
Second TGGGATAACCTTTATAGGAGTCAGCCG
```

Выравнивание 2

Вес парного выравнивания: аффинные штрафы за гэпы

```
First  TGGAGTAACCAT--TTGGAGCTAGCCG
       |||..|||||.|  |.||||..|||||
Second TGGGATAACCTTTATAGGAGTCAGCCG
```

Выравнивание 1

```
First  TGGAGTAACCAT-TT-GGAGCTAGCCG
       |||..|||||.| .| ||||..|||||
Second TGGGATAACCTTTATAGGAGTCAGCCG
```

Выравнивание 2

Выравнивание 1 биологически более вероятно, чем выравнивание 2
(потому что одна делеция в две буквы случается чаще, чем две делеции в одну букву)

Чтобы выравнивание 1 имело больший вес, чем выравнивание 2, штрафы за гэпы делают зависимым от числа подряд идущих гэпов.

Стандартный способ: за первый гэп вычитается «штраф за открытие», за каждый последующий — меньший «штраф за удлинение»

Терминология: гэпы и индели

Один знак "–", означающий отсутствие в данной последовательности **одной** буквы, гомологичной другим буквам данного столбца, мы будем называть **«гэп»**

Совокупность подряд идущих гэпов мы будем называть «**индель**», от **ин**серция/**дел**еция.

```
First  TGGAGTAACCAT--TTGGAG-CTAGCCG
       |||..|||||.|  |.||||.|.|||||
Second TGGGATAACCTTTATAGGAGTCCAGCCG
```

Тут три гэпа и два инделя

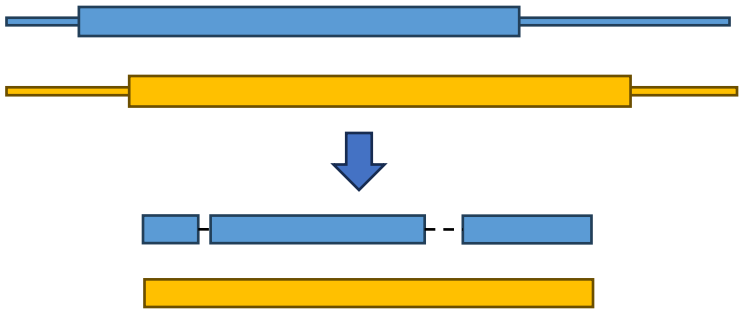
К сожалению, терминология не вполне устоялась. В литературе и описаниях программ вы можете встретить употребление термина «гэп» для обозначения инделя.

Глобальное выравнивание
(алгоритм Нидлмана-Вунша)

- Последовательности должны быть гомологичны по всей длине
- Задача - найти выравнивание с наибольшим весом
- Последовательности выравниваются целиком

Локальное выравнивание
(алгоритм Смита-Ватермана)

- Последовательности могут содержать гомологичные и негомологичные участки
- Задача – найти участки локального сходства и их выравнивание с наибольшим весом



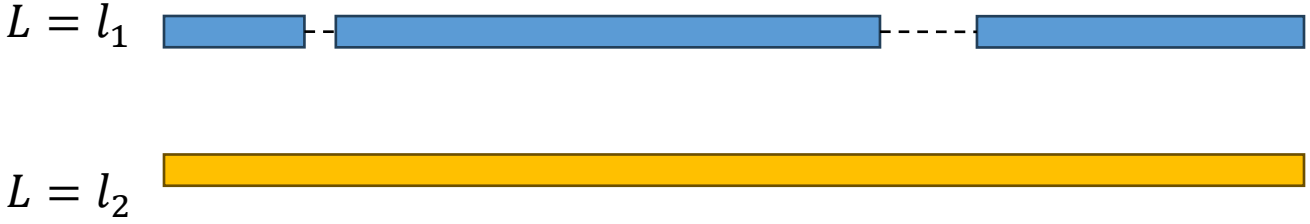
Интерпретация: вне выровненных участков гомология не обнаружена

Покрытие локального выравнивания

Seq_1, $L = L_1$

Seq_2, $L = L_2$

Нашли локальные участки сходства и выровняли:



Покрытие Seq_1: l_1/L_1

Покрытие Seq_2: l_2/L_2

Программы

Парное глобальное выравнивание:

в EMBOSS: **needle**, **stretcher**

Парное локальное выравнивание:

в EMBOSS: **water**, **matcher**

прочие: пакет BLAST (**blastn** для НК, **blastp** для белков)

Множественное выравнивание:

в EMBOSS: **emma**, **edialign**

прочие: **Muscle**, **MAFFT**, **ClustalO**, **Pride**, …

Редакторы выравниваний: **Jalview**, **GeneDoc**, …

Для самостоятельного обдумывания

1. (формальный вопрос). Если не штрафовать гэпы до первого сопоставления и после последнего, то глобальное выравнивание сведётся к локальному. Иначе говоря, задачу локального выравнивания можно сформулировать так: найти выравнивание с наибольшим весом, но при этом вес считаем хитрее: штрафуем только те гэпы, которые оказались между сопоставлениями букв.
2. (содержательный вопрос). Почему локальное выравнивание довольно часто имеет больший биологический смысл по сравнению с глобальным?

Форматы хранения выравниваний

Fasta-формат

```
>CHICK
MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQAGGDATENFEDVG
HSTDARALSETFIIGELH-PDDRPKLQK--PAETLITTVQSNSSSWSN---WVIP-AIAAIIVALMYRSYMS
E-
>HUMAN
---MAEQSDEAVK--YYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVG
HSTDAREMSKTFIIGELH-PDDRPKLNK--PPETLITTIDSSSSWWTN---WVIP-AISAVAVALMYRLYMA
ED
>CUSRE
------MGGSKV----YSLAEVSEHSQPNDCWLVIGGKVYDVTKFLDDHPGGADVLLSSTAKDATDDFEDIG
HSSSARAMMDEMCVGDID-SSTIPTKTSYTPPKQPLYNQDKTPQFIIKLLQFLVPLIILGVAVGIRFYKKQS
SD
```

Aln-формат (он же Clustal)

```
CHICK   MVGSSEAGGEAWRGRYYRLEEVQKHNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
HUMAN   ---MAEQSDEAVK--YYTLEEIQKHNHSKSTWLILHHKVYDLTKFLEEHPGGEEVLREQA
CUSRE   ------MGGSKV----YSLAEVSEHSQPNDCWLVIGGKVYDVTKFLDDHPGGADVLLSST

CHICK   GGDATENFEDVGHSTDARALSETFIIGELH-PDDRPKLQK--PAETLITTVQSNSSSWSN
HUMAN   GGDATENFEDVGHSTDAREMSKTFIIGELH-PDDRPKLNK--PPETLITTIDSSSSWWTN
CUSRE   AKDATDDFEDIGHSSSARAMMDEMCVGDID-SSTIPTKTSYTPPKQPLYNQDKTPQFIIK

CHICK   ---WVIP-AIAAIIVALMYRSYMSE-
HUMAN   ---WVIP-AISAVAVALMYRLYMAED
CUSRE   LLQFLVPLIILGVAVGIRFYKKQSSD
```

Изменение формата в EMBOSS

Из fasta (или любого другого) в clustal:

```
seqret alignment.fasta clustal::alignment.aln
```

Из любого в fasta:

```
seqret alignment.aln fasta::alignment.fasta
```

Важно: формат задаётся префиксом вида «формат::», а не расширением имени файла!

Если вы напишете `seqret one.fasta clustal::two.fasta`, то в файле two.fasta окажется выравнивание в формате clustal, а не fasta (так делать не надо, чтобы не запутать себя)

Список форматов: http://emboss.open-bio.org/html/use/ch05s04.html

Программы needle, water, stretcher, matcher по умолчанию используют собственные (не переформатируемые) форматы. Заставить их выдать выравнивание в одном из стандартных форматов можно опцией –aformat <название формата>, например –aformat fasta

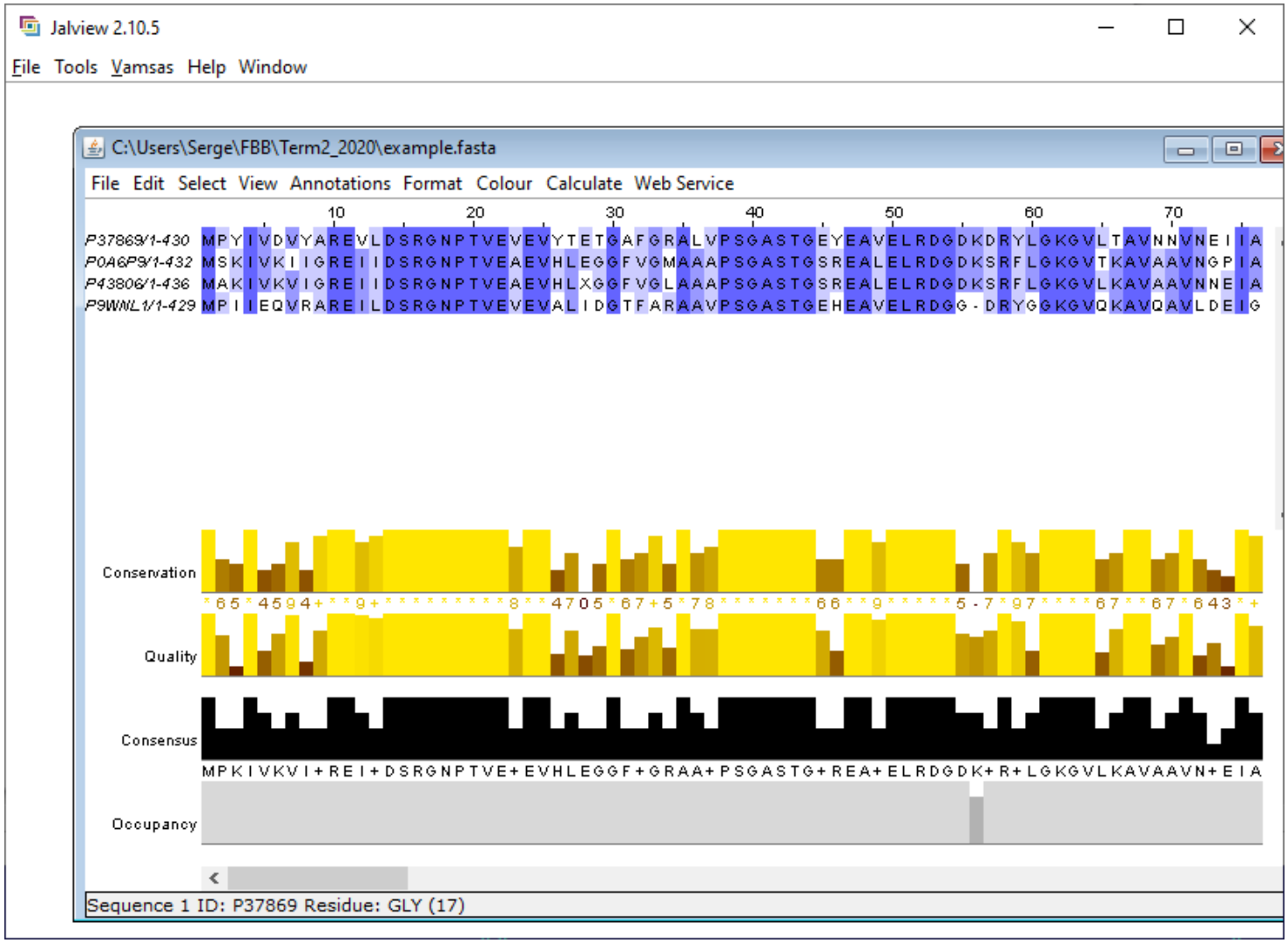
Изменение формата через BioPython

Если установлен BioPython:

```
from Bio import AlignIO
inh = open("input_file.fasta", "r")
outh = open("output_file.aln", "w")
alignment = AlignIO.parse(inh, "fasta")
AlignIO.write(alignment, outh "clustal")
inh.close()
outh.close()
```

Список форматов см. https://biopython.org/wiki/AlignIO
Он включает форматы stockholm и phylip-relaxed, которых нет в EMBOSS

Jalview


Jalview 2.10.5
File Tools Vamsas Help Window
C:\Users\Serge\FBB\Term2_2020\example.fasta
File Edit Select View Annotations Format Colour Calculate Web Service
10 20 30 40 50 60 70
P37869/1-430 MPYIVDVYAREVLDSRGNPTVEVEVYTETGAFGRALVPSGASTGEYEAVELRDGDKDRYLGKGVLTAVNNVNEIIA
P0A6P9/1-432 MSKIVKIIGREIIDSRGNPTVEAEVHLEGGFVGMAAAPSGASTGSREALELRDGDKSRFLGKGVTKAVAAVNGPIA
P43806/1-436 MAKIVKVIGREIIDSRGNPTVEAEVHLXGGFVGLAAAPSGASTGSREALELRDGDKSRFLGKGVLKAVAAVNNEIA
P9WNL1/1-429 MPIIEQVRAREILDSRGNPTVEVEVALIDGTFARAAVPSGASTGEHEAVELRDGG-DRYGGKGVQKAVQAVLDEIG
Conservation
Quality
Consensus
MPKIVKVI+REI+DSRGNPTVE+EVHLEGGF+GRAA+PSGASTG+REA+ELRDGDK+R+LGKGVLKAVAAVN+EIA
Occupancy
Sequence 1 ID: P37869 Residue: GLY (17)

Словарик

Alignment	Выравнивание
Gap	Гэп
Indel	Индель
Gap penalty	Штраф за гэпы
Gap opening penalty	Штраф за индель (за первый гэп в нём)
Gap extension penalty	Штраф за удлинение инделя (за каждый последующий гэп)
Score	Вес выравнивания
Scoring matrix	Матрица замен аминокислот

Вопросы и ответы

Что такое гомология?

Ответ: общность происхождения

(*НЕПРАВИЛЬНО говорить «последовательности гомологичны на 56%». Последовательности либо гомологичны, либо нет*)

Как определить, гомологичны ли белки?

Ответ: в большинстве случаев единственный способ — выровнять их последовательности и посмотреть на процент совпадающих букв. Если он достаточно велик, то белки, вероятно, гомологичны. Если нет, то всякое может быть.

Если для обоих белков известны пространственные структуры, то есть гораздо более чувствительный способ: сравнить ход полипептидной цепи в пространстве.

Какой процент идентичности служит надёжным признаком гомологии?

Ответ: для белков обычно более 20–25% на достаточно длинном участке

(более точный ответ будет дан в следующих лекциях)