

Парное выравнивание

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

8 апреля 2025

Сергей Александрович Спирин
Юлия Алешина

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

MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFK GHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_MOUSE

MGLSDGEWQLVLNVWGKVEADLAGHGQEV LIGLFK THPETLDKFDKFKNLKSEEDMKGSE
DLKKHGCTVLTALGTILKKKGQHA AEIQPLAQSHATKHKIPVKYLEFISEIIIIEVLKKRH
SGDFGADAQGA MSKALELFRNDIAAKYKELGFQG

>MYG_BOVIN

MGLSDGEWQLVLNAWGKVEADVAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASE
DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH
PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG

Разместим последовательности друг под другом, чтобы было видно сходство

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHHPETLEKFDKFKHLKSEDEMKASE 60
MYG_MOUSE MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIIGLFKTHPETLDKFDKFKNLKSEEDMKGSE 60
MYG_BOVIN MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIIRLFTGHPETLEKFDKFKHLKTEAEMKASE 60
*****: ***** ** . *****:*****:***: * :**. **
MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_MOUSE DLKKHGCTVLTALGTILKKKGQHA AEIQPLAQSHATKHKIPVKYLEFISEIIIIEVLKKRH 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
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGHPEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_DANRE
MADHDLVLKCGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKGELLKAKGDHAALLKPLANTHANIHKVALNRFRLITEVLVKVMAEKAGLDAA
GQALRRVMDAVIDIDIGYYKEIGFAG
```

Как сравнивать, если разная длина?

Сравнение последовательностей

Последовательности миоглобинов человека и рыбы:

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGH PETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_DANRE
MADHDLVLKCGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNRFRLITEVLVKVMAEKAGLDAA
GQGALRRVMDAVIDIDGYYKEIGFAG
```

Как сравнивать, если разная длина?

Ответ: выравнивание!

Сравнение последовательностей

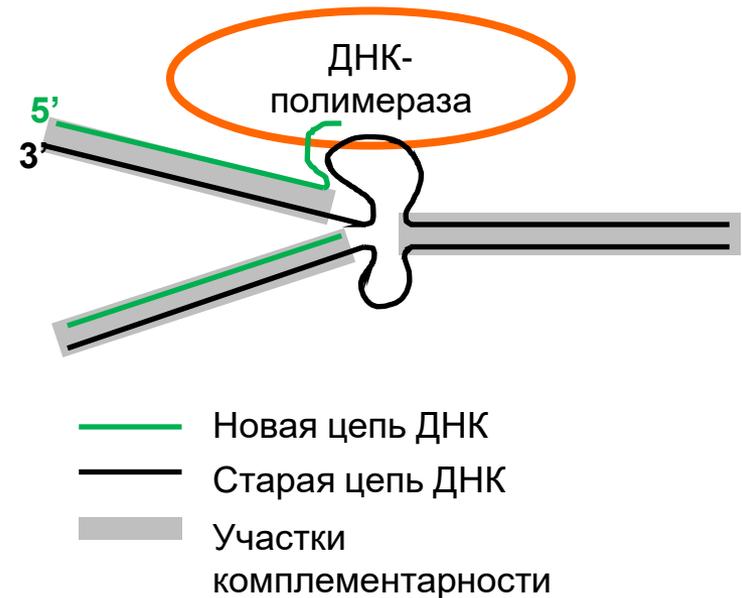
```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGH PETLEKFDKFKHLKSEDEMKASE 60
MYG_DANRE ---MADHDLVLK CWGAVEADYAANGGEV LNRLFKEYPDTLKLFPKFSGISQG-DLAGSP 55
      .: :***: ** ***** .:* *** ***** :*:** : * **. :.. :: .*
MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_DANRE AVAAHGATVLK KLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKA 115
      : *****. ** :** *. * * :*****:**. **: :: :.**: * :***: .*
MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_DANRE --GLDAAGQGALRRVMDAVIDGIDGYYKEIGFAG 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
GTGGGATAACSTTTATAGGAGTCAGCCGCCTAAGGCAGG

Негомولوجичные 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

предок	TATGCGAAT - GCCCTGAA	
сын	TATGCA A AAT - GCCCTGAA	замена
внук	TATGCA AA AAT - GC T CTGAA	замена
правнук	TATGCA AA AAT C GC T CGGAA	вставка 1 п.н.
праправнук	TATGCA AAA A C GC T CGGAA	замена
прапраправнук	TATGCA AAA - C GC T CGGAA	делеция 1 п.н.
...	TATGCA ATA - C GC T CGGAA	замена
...	TATGCA ATA - C GC - - - GAA	делеция 3 п.н.

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	TATGCA ^A AAT - GCCCTGAA
внук	TATGCA ^{AA} AAT - GC ^T CTGAA
правнук	TATGCA ^{AAAT} CGCTCGGAA
праправнук	TATGCA ^{AAA} A ^A CGCTCGGAA
прапраправнук	TATGCA ^{AAA} -CGCTCGGAA
...	TATGC ^{ATA} -CGCTCGGAA
...	TATGC ^{ATA} -CGC - - -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	TATGCA A AAT - GCCCTGAA
внук	TATGCA AA AAT - GC T CTGAA
правнук	TATGCA AAA AAT CG CTCGGAA
праправнук	TATGCA AAAA AAT CG CTCGGAA
прапраправнук	TATGCA AAAAA AAT CG CTCGGAA
...	TATGCA ATA - CG CTCGGAA
...	TATGCA ATA - CGC - - - GAA

предок

TATGCGAATGCCCTGAA



TATGCGAAT - GCCCTGAA

TATGCGAATG - CCCTGAA

сын

TATGC**A**AAT - GCCCTGAA

TATGCC**C**AATG - CCCTGAA

внук

TATGC**A**AAT - G**C**TCTGAA

TATGCC**C**AATG - C**C**TTGAA

правнук

TATGC**A**AAT**C**GCTCGGAA

TATGCC**C**AATG**C**CCCTGGAA

праправнук

TATGC**A**AA**A**CGCTCGGAA

TAT - **C**CAATG**C**CCCTGGAA

прапраправнук

TATGC**A**AA - **C**GCTCGGAA

TAT - **C**CAATG**C**CCCTGG**T**A

...

TATGC**A**T**A** - **C**GCTCGGAA

TAT - - **C**AATG**C**CCCTGG**T**A

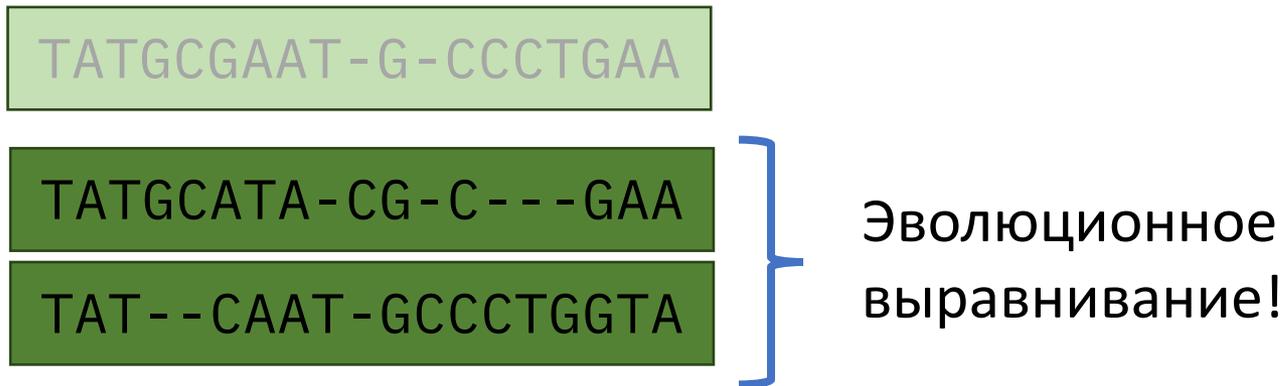
Потомок, которого мы видим

TATGC**A**T**A** - **C**GC - - - GAA

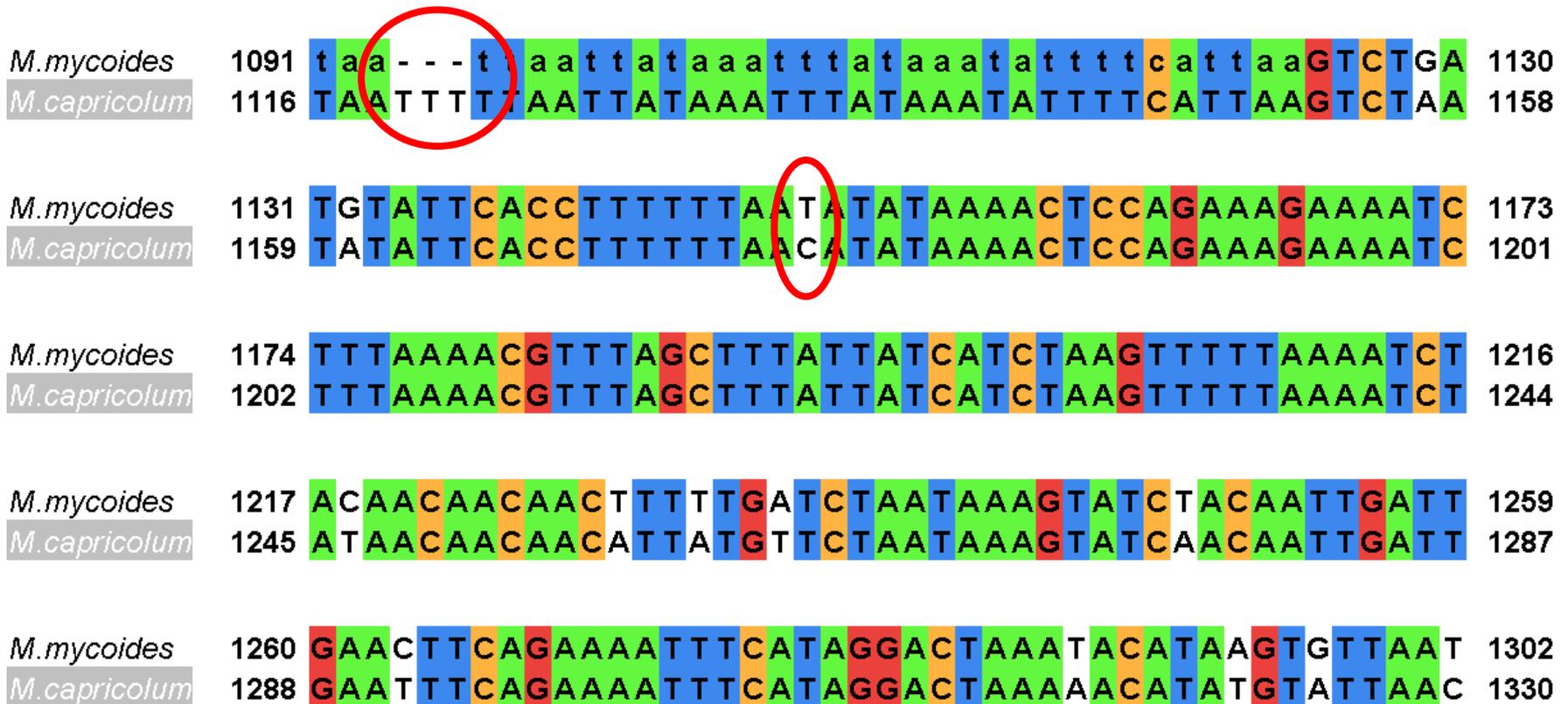
TAT - - **C**AATG**C**CCCTGG**T**A

Выравнивание отражает эволюцию

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

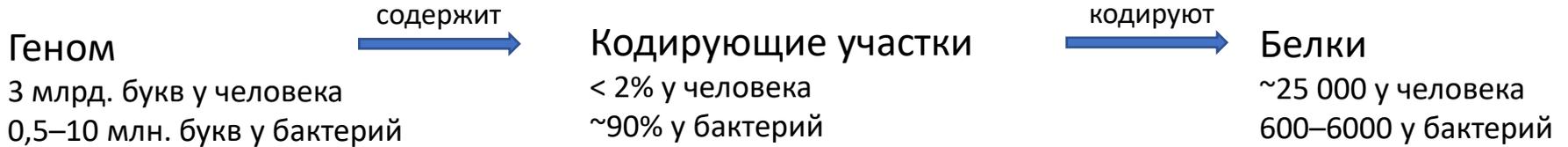


Выравнивание геномов двух потомков общего предка микоплазм: *M. capricolum* и *M. mycoides* (маленький фрагмент)



В среднем 92% совпадающих букв на гомологичных участках 20

ДНК и белки



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

gatcaacactacttgacttcaag**g**acttaccataaagaaaac



gatcaacactacttgacttcaaa**a**acttaccataaagaaaac

точечная замена

gatcaacactacttgacttcaag**g**acttaccataaagaaaac



gatcaacactacttgacttcaaa**actt**accataaagaaaac

делеция

gatcaacactacttgacttcaag**g**acttaccataaagaaaac



gatcaacactacttgacttcaag**g**acttaccataaagaaaac

инсерция
(вставка)

Классификация мутаций в кодирующих последовательностях ДНК

... ААТССГТСААГТСТА...
... Asn Pro Ser Ser Leu ...

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

... ААТССГТС**G**АГТСТА...
... Asn Pro Ser Ser Leu ...

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

- ✓ на близкий по свойствам
- ✓ с отличными свойствами

... ААТССГ**A**СААГТСТА...
... Asn Pro **Thr** Ser Leu ...

... ААТССГТСААГ**A**СТА...
... Asn Pro Ser **Arg** Leu ...

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

... ААТССГТ**G**ААГТСТА...
... Asn Pro **Stop** Ser Leu ...

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)

Met **Thr Ser**

ATG**A**CTTCAT...

ATG-CTTCAT...

Met **Leu His**

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».

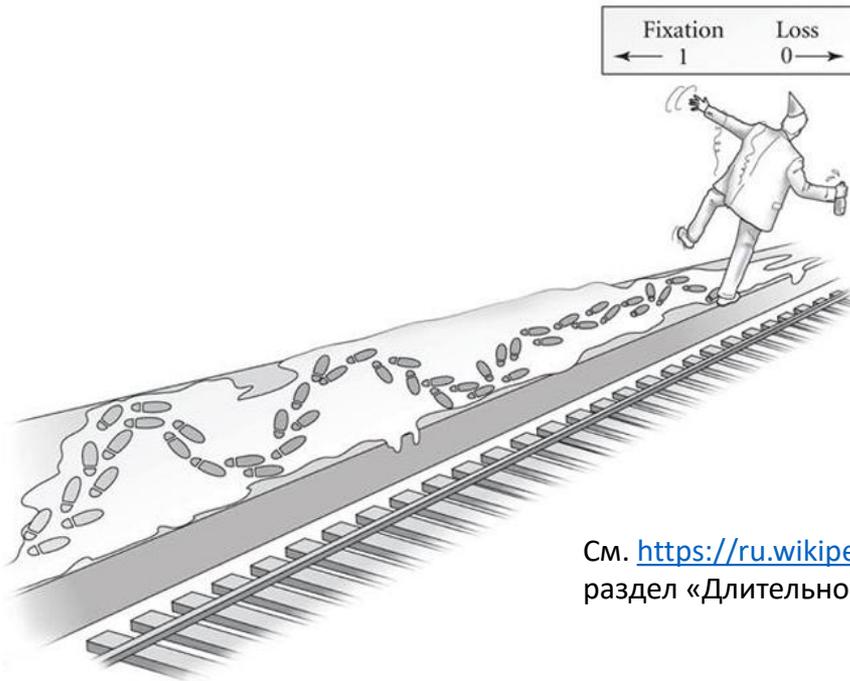
На тротуаре стоит пьяный и каждые 10 сек. делает шаг либо направо, либо налево, случайно выбирая направление. Как далеко он уйдёт за время T ?

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

```
11 DRREIRHIWDDVWSSSF TDRRVAIVRAVFDDL FKHYP TSKALFERVK IDEPESGEF 66
 8 DRHEVLDNWKGIWSAEFTGRRVAIGQAIFQELFALDPNAKGVFGRVNVDPKSEADW 63
 8 DRREVQALWRSIWSAEDTGRRTLIGRLLFEELFEIDGATKGLFKRVNVDDTHSPEE 63
 7 QRKVKVQQAQVYSGES--RTDFAIDVFNNFFRTNPD-RSLFN RVNGDNVYSPEF 59
 9 QRLKVKQWAKAYGVGHE--RVELGIALWKSMFAQDNDARDLFRVHGEDVHSPAF 62
 8 EGLKVKSEWGRAYGSGHD--REAFSQA IWRATFAQVPESRSLFKRVHGDDTSHPAF 61
 6 QRFKVKHQWAEAFGTSHH--RLDFGLKLWNSIFRDAPEIRGLFKRVDGDNAYSAEF 59
 7 QRLKVKRQWAEAYGSGND--REEFGHF IWTHVFKDAPSARDLFRVVRGDN IHTPAF 60

67 KSHLVRVANGLDLLINLLDDTLVLQSHLGH LADQHIQRKGV TKEYFRGIGEAFA 120
64 KAHVIRVINGLDLAVNLLLEDPKALQEELKHLARQHRERSGVKAVYFDEMEKALL 117
64 FAHVLRVVNGLDTLIGVLGDSDTLNSLIDHLAEQHKARAGFKTVYFKEFGKALN 117
60 KAHMVRVFAFGFDILISVLDDKPVLDQALAHYAAFHKQF-GTIP--FKAFGQTMF 110
63 EAHMARVFNGLDRV ISSLTDEPVLNAQLEHLRQQHIKL-GITGHMFNLMRTGLA 115
62 IAHAEVRLGGLDIAISTLDQPATLKEELDHLQVQHEGR-KIPDNYFDAFKTAIL 114
60 EAHAEVRLGGLDMTISLLDDQAAFDAQLAHLKSQHAER-NIKADYYGVFVNELL 112
61 RAHATRVLGGLDMCIAALLDDEGV LNTQLAHLASQHSSR-GVSAAQYDVVEHSVM 113
```

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).

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

$\sim 9 \cdot 10^{58}$ 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 $20A - 5B - 2C$

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 .

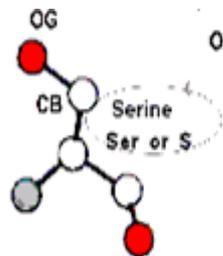
Вес парного выравнивания – белки

Какое выравнивание имеет бóльшие шансы оказаться правильным?

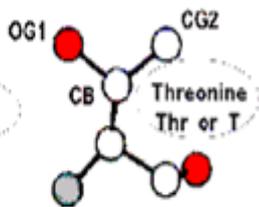
AFTGAHAYL
AYS---AYM

AFTGAHAYL
AY---SAYM

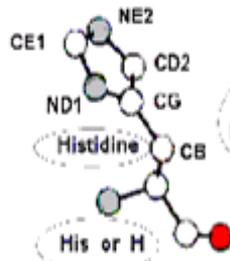
Вес парного выравнивания – белки



Серин S



Треонин T



Гистидин H

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

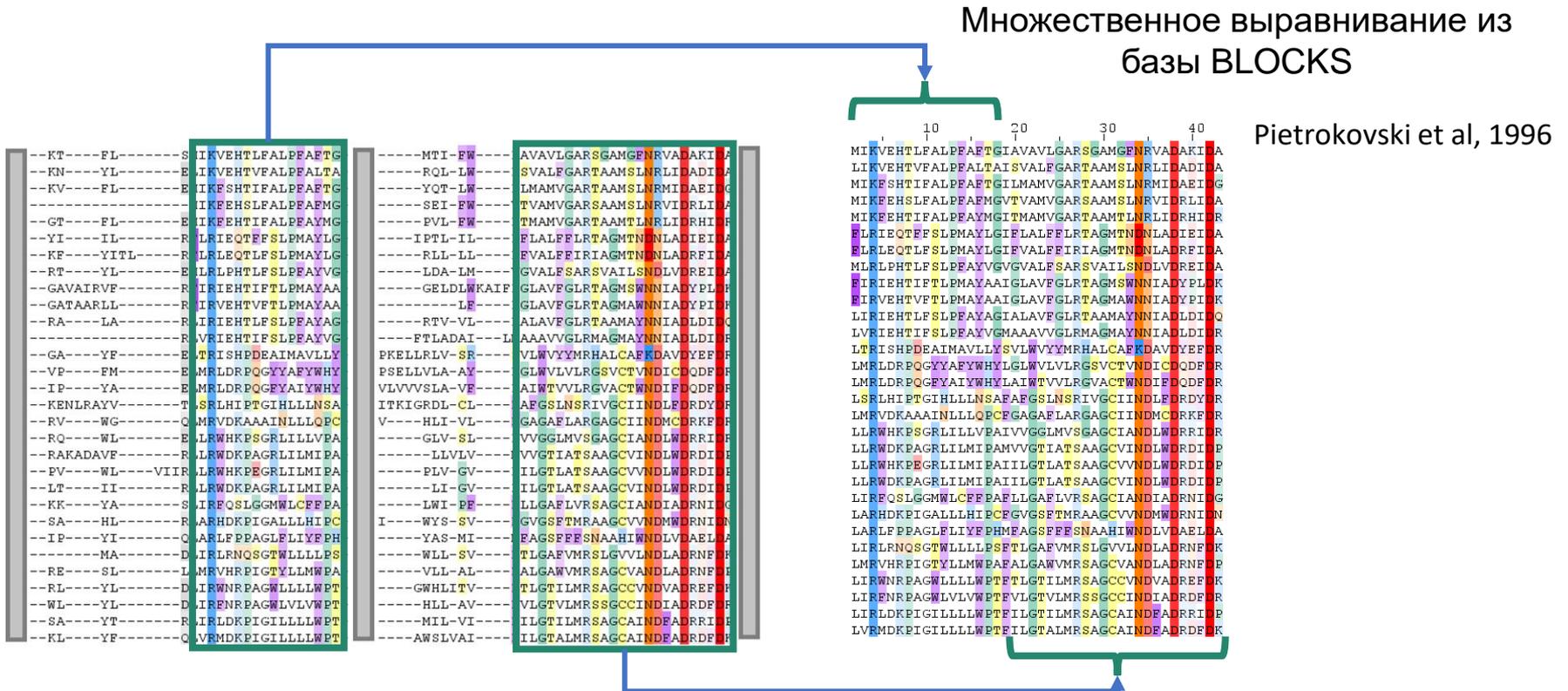
Поэтому если в одной колонке выравнивания оказались буквы S и T, это скорее аргумент за данное выравнивание, чем против него. Значит, за такую колонку лучше увеличивать вес, чем уменьшать.

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}

Protein1	F	L	R	T	A	G	M	T	N	D	N	L	A	D	I	E	I	D	A
Protein2	F	I	R	I	A	G	M	T	N	D	N	L	A	D	R	F	I	D	A
...	S	A	R	S	V	A	I	L	S	N	D	L	V	D	R	E	I	D	A
	G	L	R	T	A	G	M	S	W	N	N	I	A	D	Y	P	L	D	K
	G	L	R	T	A	G	M	A	W	N	N	I	A	D	Y	P	I	D	K
	G	L	R	T	A	A	M	A	Y	N	N	I	A	D	L	D	I	D	Q
	G	L	R	M	A	G	M	A	Y	N	N	I	A	D	L	D	I	D	R
	Y	M	R	H	A	L	C	A	F	K	D	A	V	D	Y	E	F	D	R
	V	L	R	G	S	V	C	T	V	N	D	I	C	D	Q	D	F	D	R
	V	L	R	G	V	A	C	T	W	N	D	I	F	D	Q	D	F	D	R

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

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 <td>T <td>A <td>G <td>P <td>D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td>	T <td>A <td>G <td>P <td>D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td>	A <td>G <td>P <td>D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td></td></td></td>	G <td>P <td>D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td></td></td>	P <td>D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td></td>	D <td>E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td></td>	E <td>Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td></td>	Q <td>N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td></td>	N <td>H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td></td>	H <td>R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td></td>	R <td>K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td></td>	K <td>M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td></td>	M <td>I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td></td>	I <td>L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td></td>	L <td>V <td>W <td>Y <td>F <td></td> </td></td></td></td>	V <td>W <td>Y <td>F <td></td> </td></td></td>	W <td>Y <td>F <td></td> </td></td>	Y <td>F <td></td> </td>	F <td></td>	

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

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.ii
# * 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  1
```

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

Вес парного выравнивания – белки

Посчитаем веса выравниваний, используя матрицу BLOSUM62

AFTGАНAYL
AYS---AYM

AFTGАНAYL
AY---SAYM

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

Тогда вес левого выравнивания равен:

$M(A,A) + M(F,Y) + \mathbf{M(T,S)} + M(A,A) + M(Y,Y) + M(L,M)$ – штраф за гэпы,

а правого:

$M(A,A) + M(F,Y) + \mathbf{M(H,S)} + M(A,A) + M(Y,Y) + M(L,M)$ – 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-ТТ-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 инсерция/делеция.

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

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.

Парное выравнивание: локальное и глобальное

На самом деле это две формализации одной и той же задачи: даны последовательности двух белков, найти гомологичные аминокислотные остатки.

Задача глобального выравнивания: найти выравнивание с наибольшим весом. При вычислении весов учитываются: матрица замен (BLOSUM62) и аффинные штрафы за гэпы.

Задача локального выравнивания: найти

- участок в первой последовательности;
- участок во второй последовательности;
- выравнивание выбранных участков;

так, чтобы вес выравнивания был наибольшим.

Разница в том, что теперь выбираем не только как выравнивать, но и что.

Интерпретация: вне выровненных участков гомология не обнаружена

Алгоритмы и программы оптимального парного выравнивания

- Оптимальное **глобальное** выравнивание: алгоритм **Нидлмана – Вунша**
Needleman & Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*. **48** (3): 443–53
В оригинальной работе предлагалось оценивать выравнивание с линейными штрафами за гэпы (одинаковый штраф за каждый гэп) и описывался алгоритм нахождения оптимального выравнивания с таким весом. Позднее был предложен вес с аффинными штрафами и алгоритм модифицирован для этой ситуации, ещё позднее введены матрицы замен.
Программы в EMBOSS: **needle** и **stretcher**
- Оптимальное **локальное** выравнивание: алгоритм **Смита – Уотермена** (=Смита – Ватермана)
Smith & Waterman (1981). "Identification of Common Molecular Subsequences". *Journal of Molecular Biology*. **147** (1): 195–197
Программа в EMBOSS: **water**
- Близкий алгоритм Уотермена – Эггерта выдаёт не одно оптимальное, а заданное число лучших выравниваний (Waterman & Eggert (1987). *Journal of Molecular Biology*. **197** (4): 723-728)
Программа в EMBOSS: **matcher**

Покрытие локального выравнивания



Seq₁, $L = L_1$



Seq₂, $L = L_2$

Нашли локальные участки сходства и выровняли:



$L = l_1$

Покрытие Seq₁: l_1/L_1

$L = l_2$

Покрытие Seq₂: l_2/L_2

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: штрафует 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
MVGSEAGGEAWRGRYYRLEEVQKHNNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQAGGDATENFEDVG
HSTDARALSETFIIGELH-PDDRPKLQK--PAETLITTVQSNSSSSWSN---WVIP-AIAAIIIVALMYRSYMS
E-
>HUMAN
---MAEQSDEAVK--YYTLEEIQKHNSKSTWLIILHHKVYDLTKFLEEHPGGEEVLREQAGGDATENFEDVG
HSTDAREMSKTFIIGELH-PDDRPKLNK--PPETLITTIDSSSSWWTN---WVIP-AISAVAVALMYRLYMA
ED
>CUSRE
-----MGGSKV----YSLAEVSEHSQPNDWCWLVIGGKVYDVTKFLDDHPGGADVLLSSTAKDATDDFEDIG
HSSSARAMMDEMCGDID-SSTIPTKTSYTPPKQPLYNQDKTPQFI IKLLQFLVPLIILGVAVGIRFYKKQS
SD
```

Aln-формат (он же Clustal)

```
CHICK   MVGSEAGGEAWRGRYYRLEEVQKHNNNSQSTWIIVHHRIYDITKFLDEHPGGEEVLREQA
HUMAN   ---MAEQSDEAVK--YYTLEEIQKHNSKSTWLIILHHKVYDLTKFLEEHPGGEEVLREQA
CUSRE   -----MGGSKV----YSLAEVSEHSQPNDWCWLVIGGKVYDVTKFLDDHPGGADVLLSST

CHICK   GGDATENFEDVGHSTDARALSETFIIGELH-PDDRPKLQK--PAETLITTVQSNSSSSWSN
HUMAN   GGDATENFEDVGHSTDAREMSKTFIIGELH-PDDRPKLNK--PPETLITTIDSSSSWWTN
CUSRE   AKDATDDFEDIGHSSSARAMMDEMCGDID-SSTIPTKTSYTPPKQPLYNQDKTPQFI IK

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

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, ...**

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

Словарик

Alignment

Gap

Indel

Gap penalty

Gap opening penalty

Gap extension penalty

Score

Scoring matrix

Выравнивание

Гэп

Индель

Штраф за гэпы

Штраф за индель (за первый гэп в нём)

Штраф за удлинение инделя (за каждый последующий гэп)

Вес выравнивания

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)