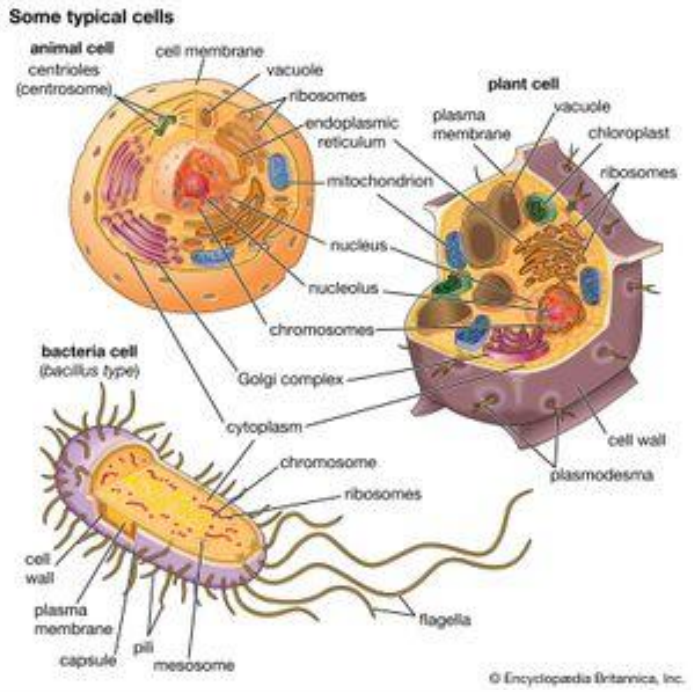
Выравнивание биологических последовательностей

С.А. Спирин
sas@belozersky.msu.ru
МФК "Биоинформатика", 5 марта 2025

Введение

Ещё немного биологии

Место действия: клетка



Бактериальная клетка окружена **фосфолипидной мембраной** (или двумя) и клеточной стенкой.
В мембрану встроены различные белки. Цитоплазма — раствор малых, средних и крупных молекул в воде.

Клетка животного тоже окружена мембраной, а её цитоплазма содержит многочисленные **органеллы**, некоторые из которых имеют собственные мембраны.

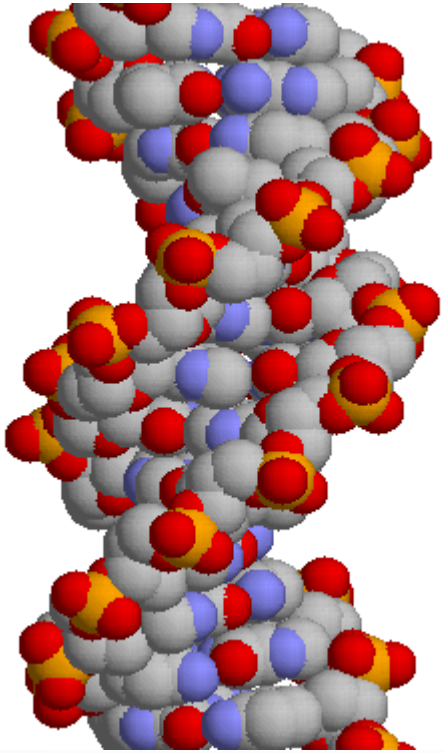
Основные процессы

- Транскрипция
  «Переписывание» ДНК на РНК
- Трансляция
  «Перевод» РНК в белок
- Репликация
  Удвоение ДНК с сохранением информации
- Репарация
  Исправление ошибок в тексте, записанном на ДНК
- Регуляция
  Определение, каким генам экспрессироваться в данный момент

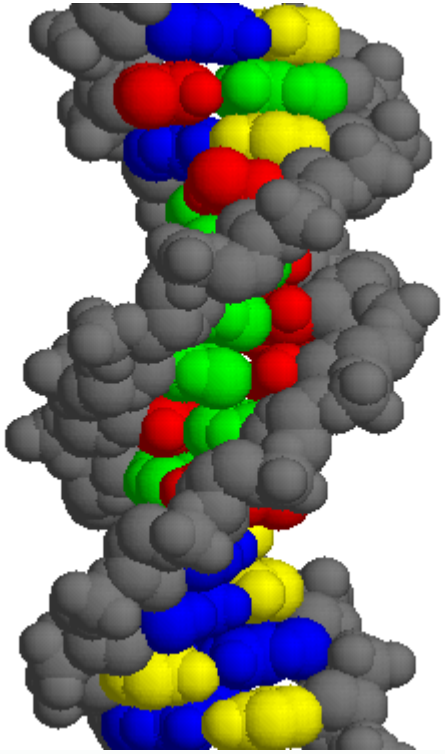
Важные термины

- Ген
  Участок генома, транскрибирующийся в функциональную РНК
- Экспрессия гена
  Производство окончательного продукта (белка или РНК) на основе гена

Ещё раз про ДНК



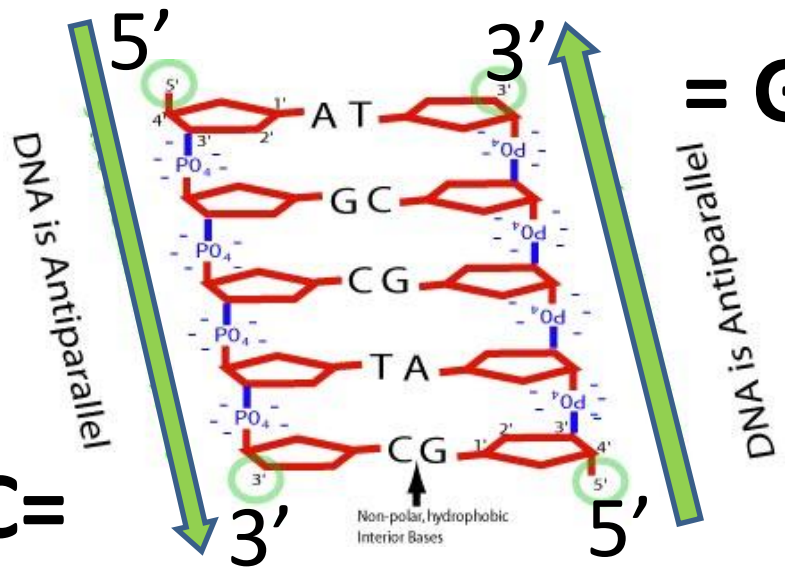
C N O P



A C G T

В живой клетке ДНК — комплекс из двух молекул, каждая из которых — гетерополимер из нуклеотидов четырёх типов: A, C, G, T

ДНК состоит из двух антипараллельных комплементарных цепочек



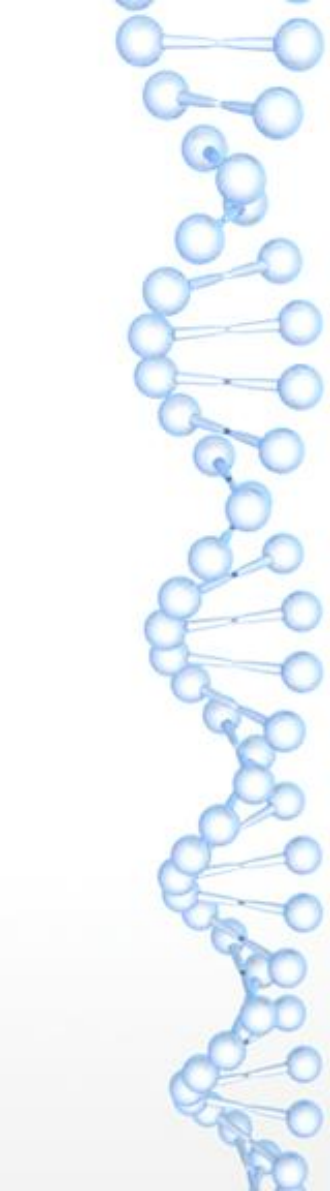
= GAGCT

AGCTC=

Гуанин (G) связан тремя **водородными связями** с цитозином (C)
Аденин (A) связан двумя водородными связями с тимином (T)

Последовательность одной цепи ДНК

1. Состоит из букв **A, T, G, C**
2. Всегда пишется в направлении от 5’-конца к 3’-концу
3. Химическая формула ДНК однозначно определяется последовательностью
4. Последовательность несёт всю наследственную информацию.

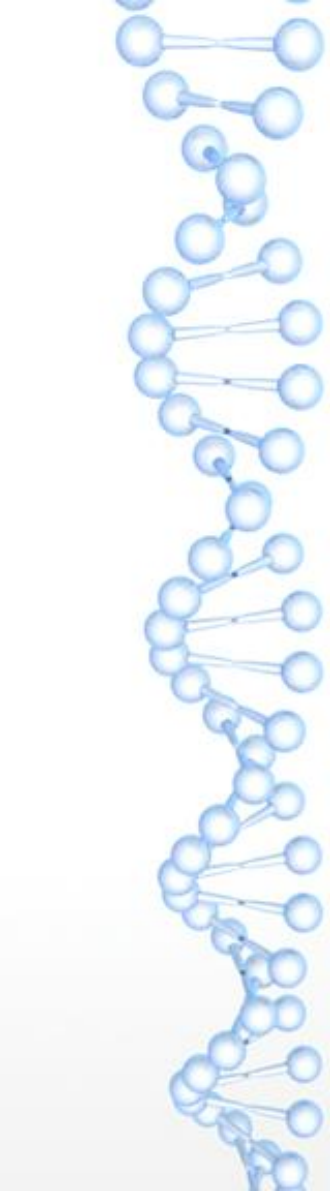


Последовательность двухцепочечной ДНК

— это последовательность одной из цепей.

Какой?

gatcaacactacttgacttcaagacttaccataaagaaaactatagtgtggtattggcaa
aagacaagacaaatagatcaacataacaaaataaagggccatgaaatagacccatatagt
caattgatttttgacaaagaaggattggcaatagaatggggtaaagatagtcttctcaac
aaacggtaccagaatgactgaatacccacatgcaaaaagaaaaagaaatgaacctagaca
cagatcttatacagttcacaaaaatgtaactcaaaatgaatcatagacctaaatataata
ttcaagactataaaaccctaaaatataacataggggaaaatctaaacaatcttgagtttg
ttaatgacttttagatacaataccaaaggcaggatccaggaaagaatcgataagctggg
cttcattaaaattaaaatatttctgctctatgaagccactgtcaagagaaggaaaaggca
agccatagactgggagaaaatatttacaaaagacatacatgataaaggactattatccaa



Последовательность двухцепочечной ДНК

— это последовательность одной из цепей.

Какой?

```
gatcaacactacttgacttcaagacttaccataaagaaaactatagtgtggtattggcaa
aagacaagacaaatagatcaacataacaaaataaagggccatgaaatagacccatatagt
caattgatttttgacaaagaaggattggcaatagaatggggtaaagatagtcttctcaac
aaacggtaccagaatgactgaatacccacatgcaaaaagaaaaagaaatgaacctagaca
cagatcttatacagttcacaaaaatgtaactcaaaatgaatcatagacctaaatataata
ttcaagactataaaaccctaaaatataacataggggaaaatctaaacaatcttgagtttg
ttaatgacttttagatacaataccaaaggcaggatccaggaaagaatcgataagctggg
cttcattaaaattaaaatatttctgctctatgaagccactgtcaagagaaggaaaaggca
agccatagactgggagaaaatatttacaaaagacatacatgataaaggactattatccaa
```

А всё равно, какой!

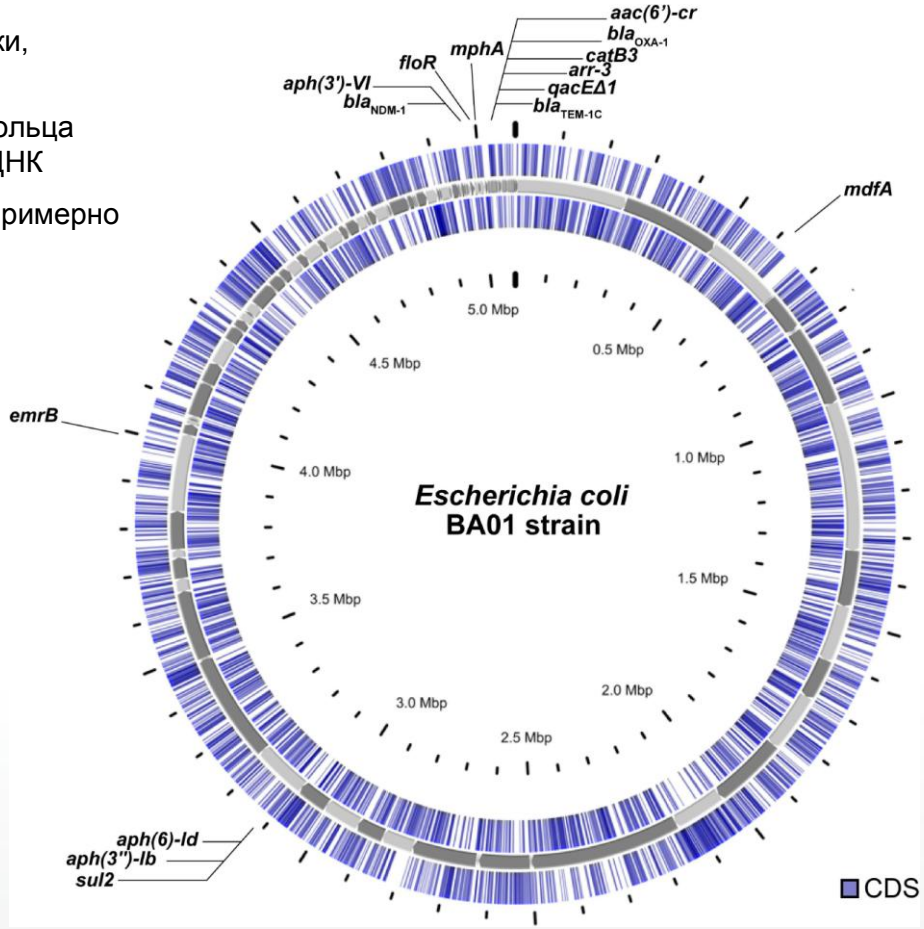
Записывается одна из цепей, вторую всегда можно восстановить по комплементарности

Из чего состоит геном

Синие полоски = участки, кодирующие белки

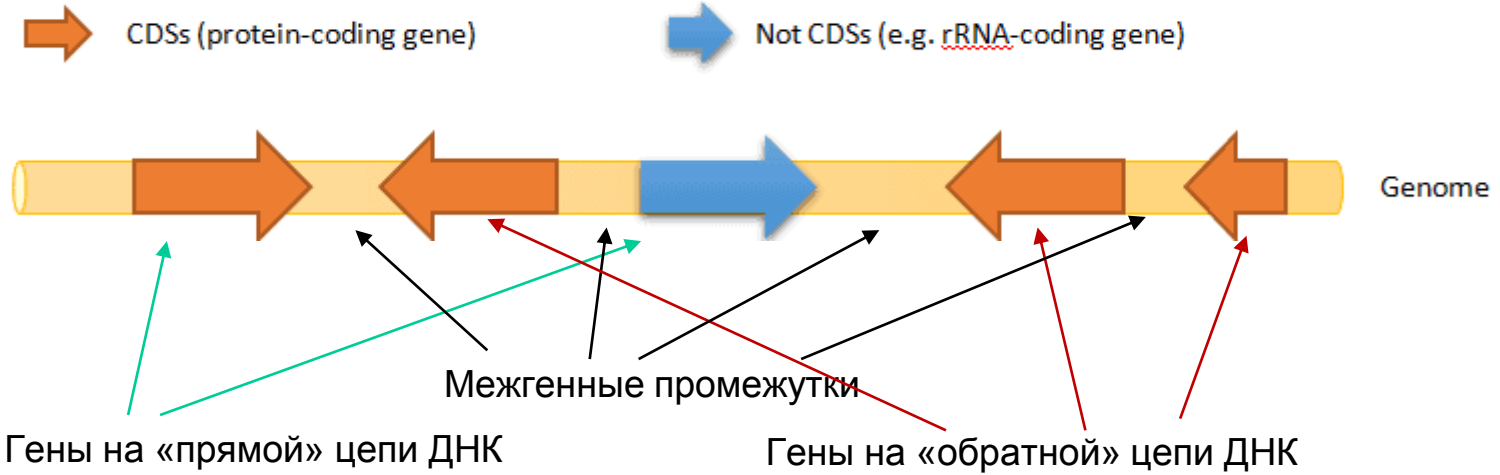
Два концентрических кольца изображают две цепи ДНК

Генов на обеих цепях примерно поровну



https://rotel.pressbooks.pub/genetics/chapter/prokaryotic-genome-structure/

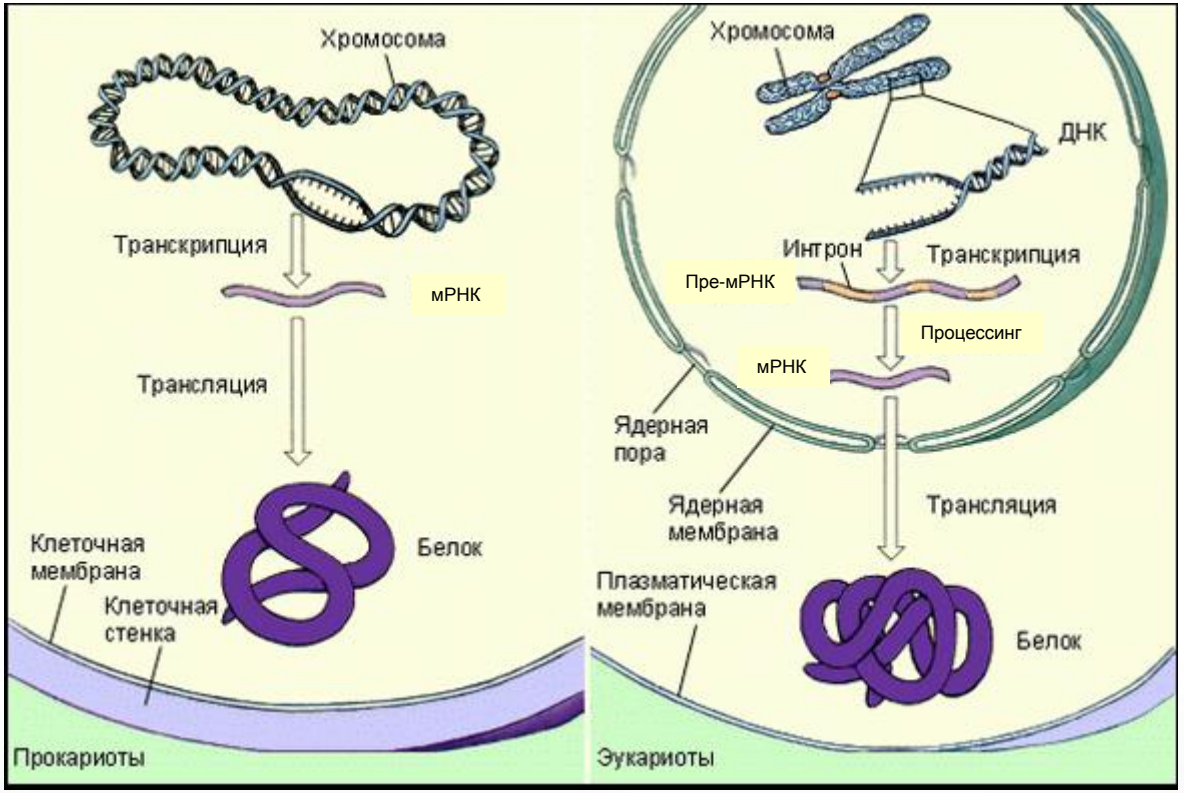
Из чего состоит геном (маленький участок)



Выбор, какую цепь считать прямой, произволен!

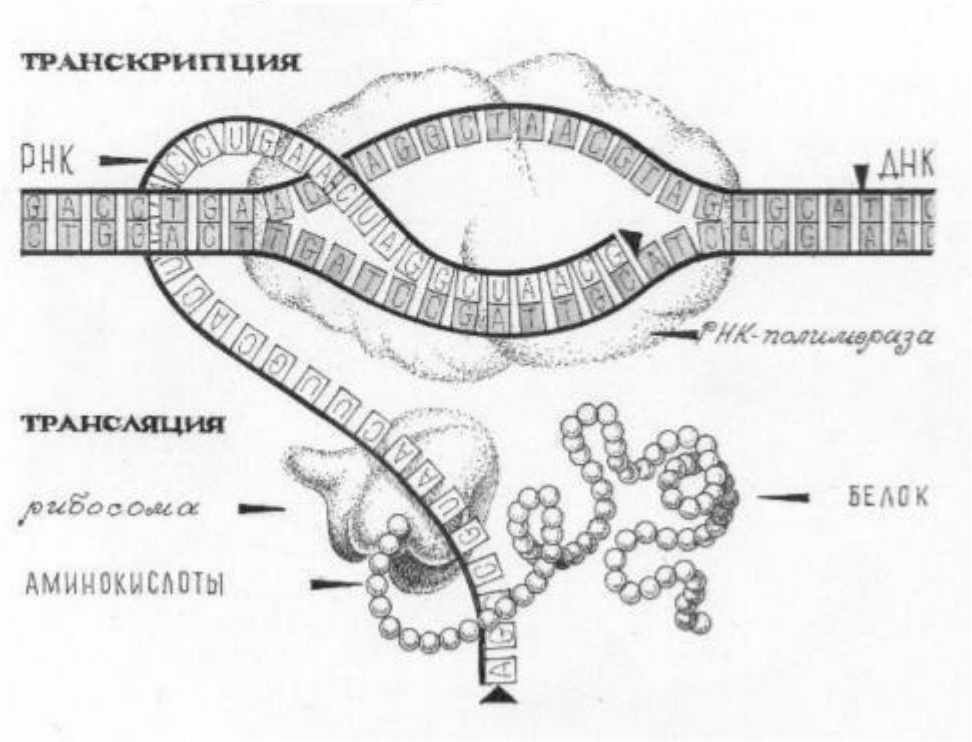
https://help.ezbiocloud.net/intergenic-region/

Поток информации ДНК→белок



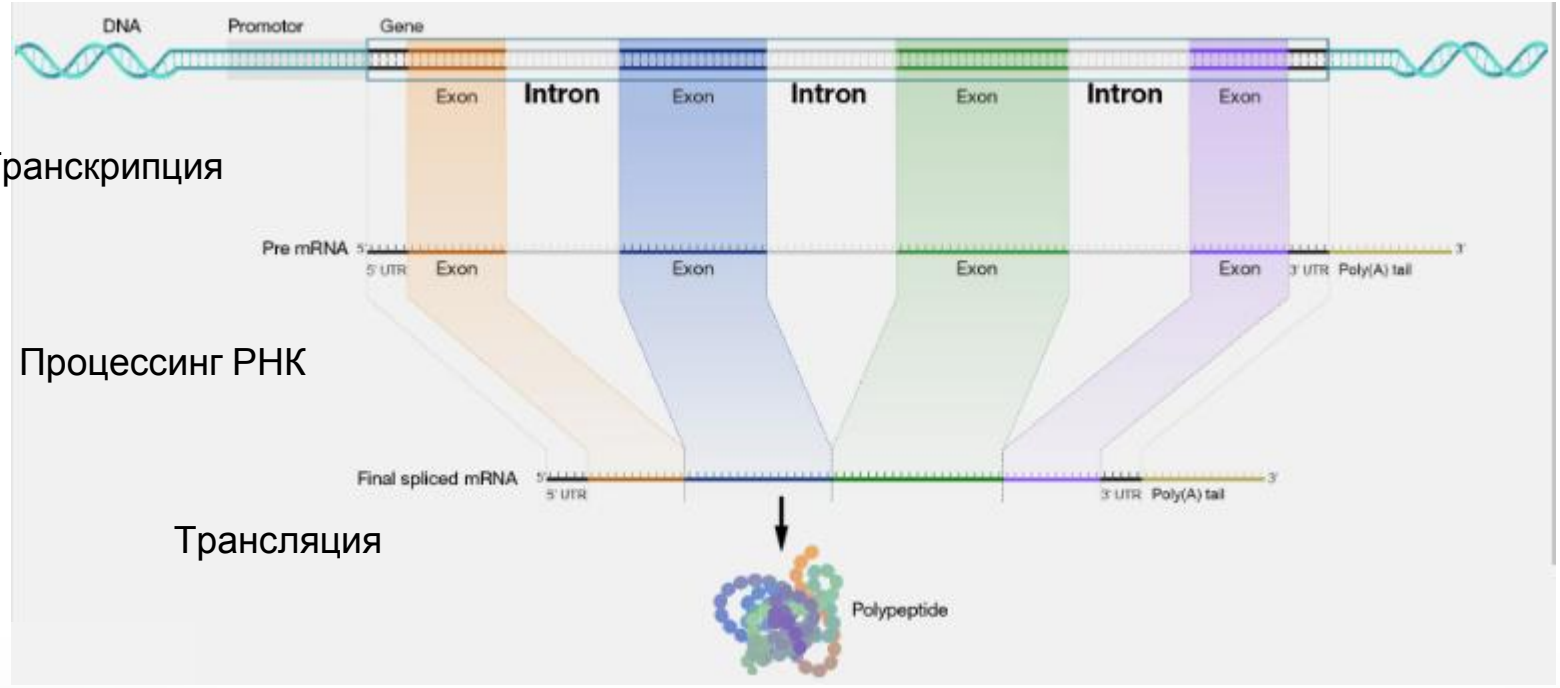
https://www.s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/bgf/nilmb_bgf/Лекция%203.pdf

Транскрипция и трансляция в прокариотах



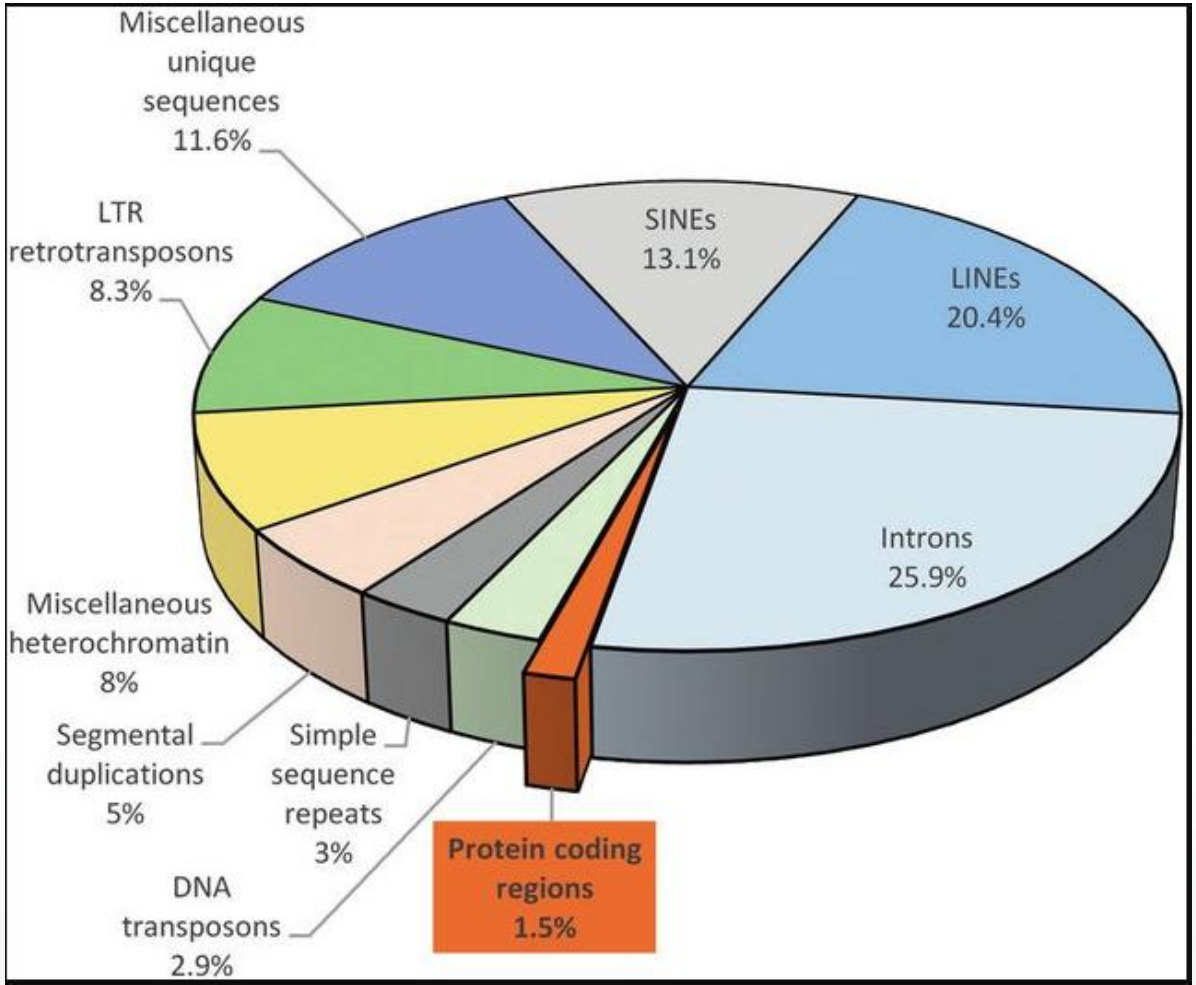
https://www.slideserve.com/merry/4764583

Структура гена у эукариот

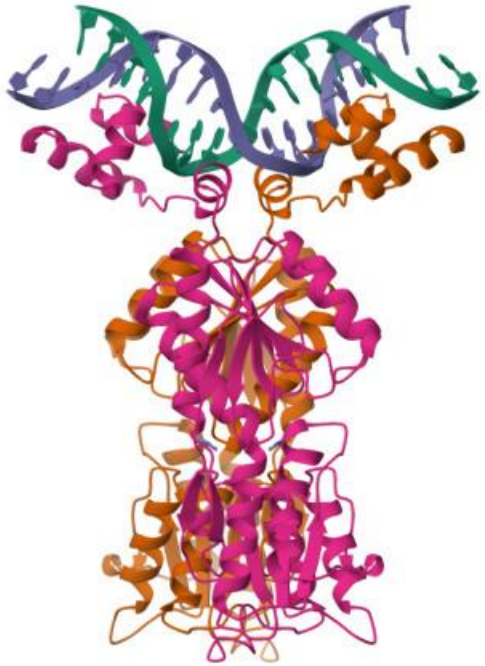


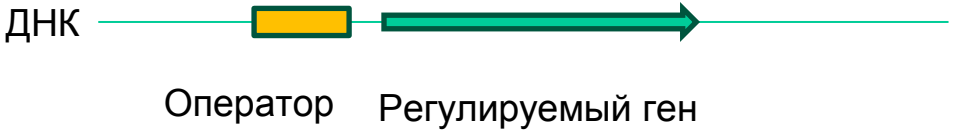
https://www.differencebetween.com/difference-between-utr-and-intron/

Состав генома человека



Пример регуляции: пуриновый репрессор





Пурины — важные для клетки вещества, в частности из них синтезируются нуклеотиды аденин и гуанин, а также АТФ.

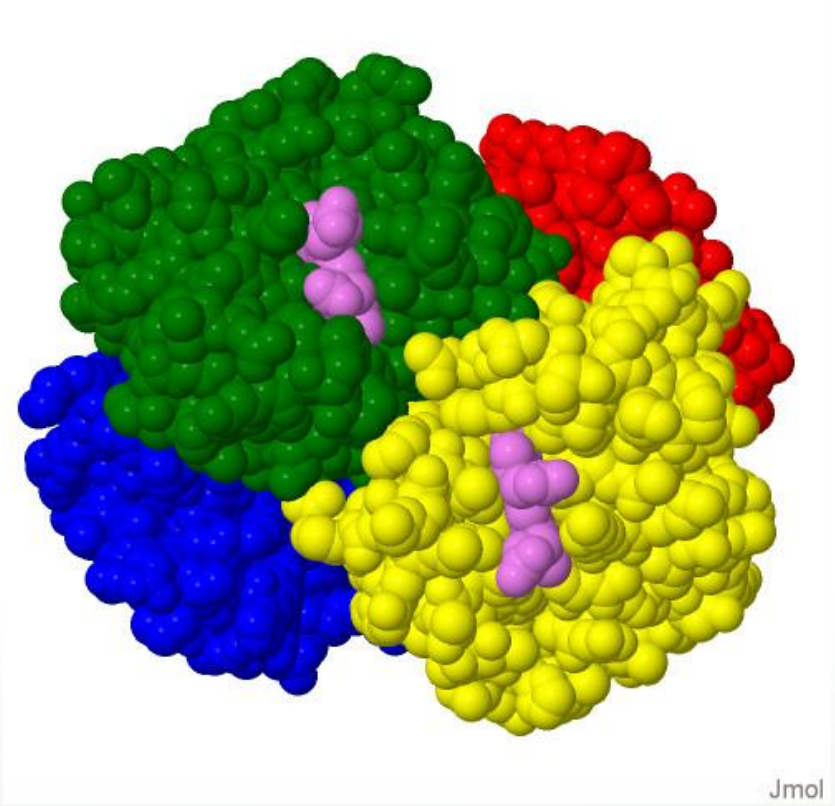
При избытке пуринов генам их синтеза и импорта лучше не экспрессироваться, чтобы зря не тратить ресурсы.

Пуриновый репрессор — белок, который, связываясь с пурином, приобретает способность связываться с операторами генов синтеза и импорта пуринов. Связавшись с оператором, репрессор блокирует экспрессию гена.

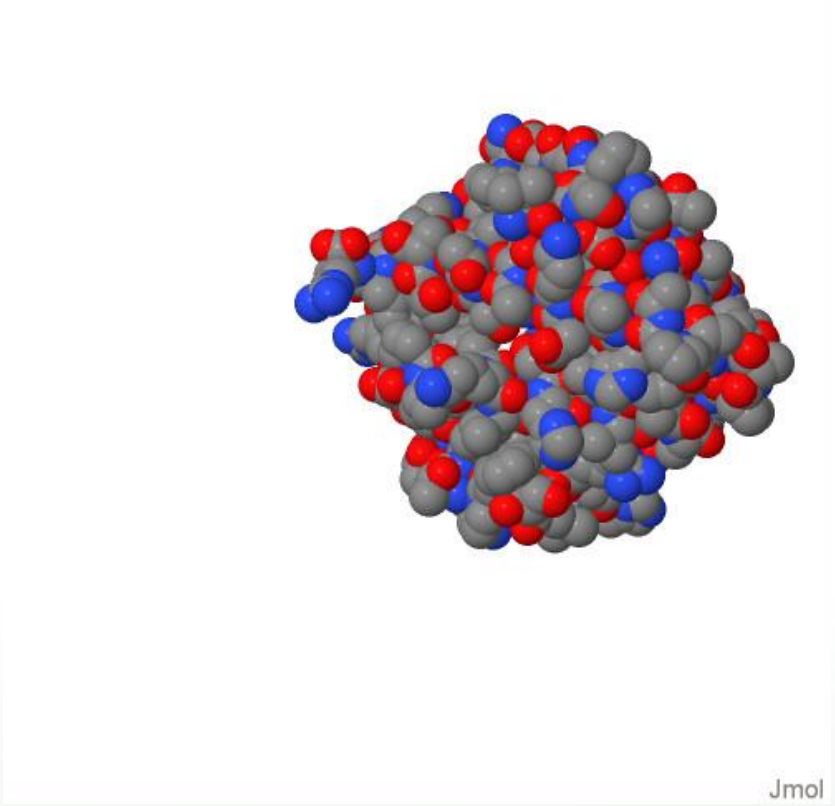
Этот пример – из бактерии кишечной палочки (*E.coli*).
У эукариот всё много сложнее

Молекула гемоглобина

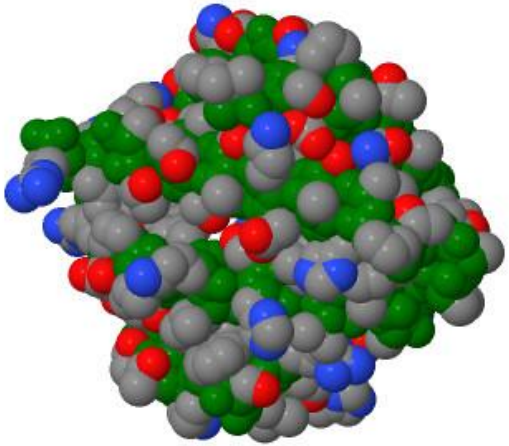
– это, строго говоря , комплекс из нескольких молекул



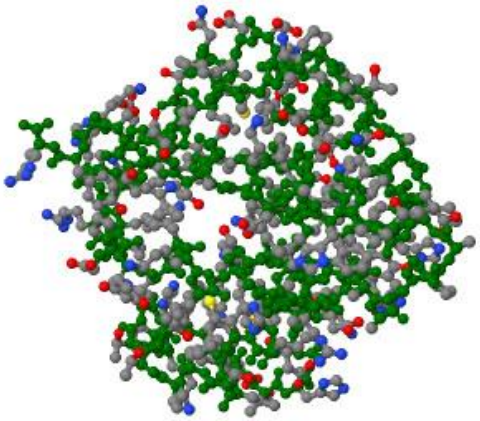
Альфа-цепь гемоглобина



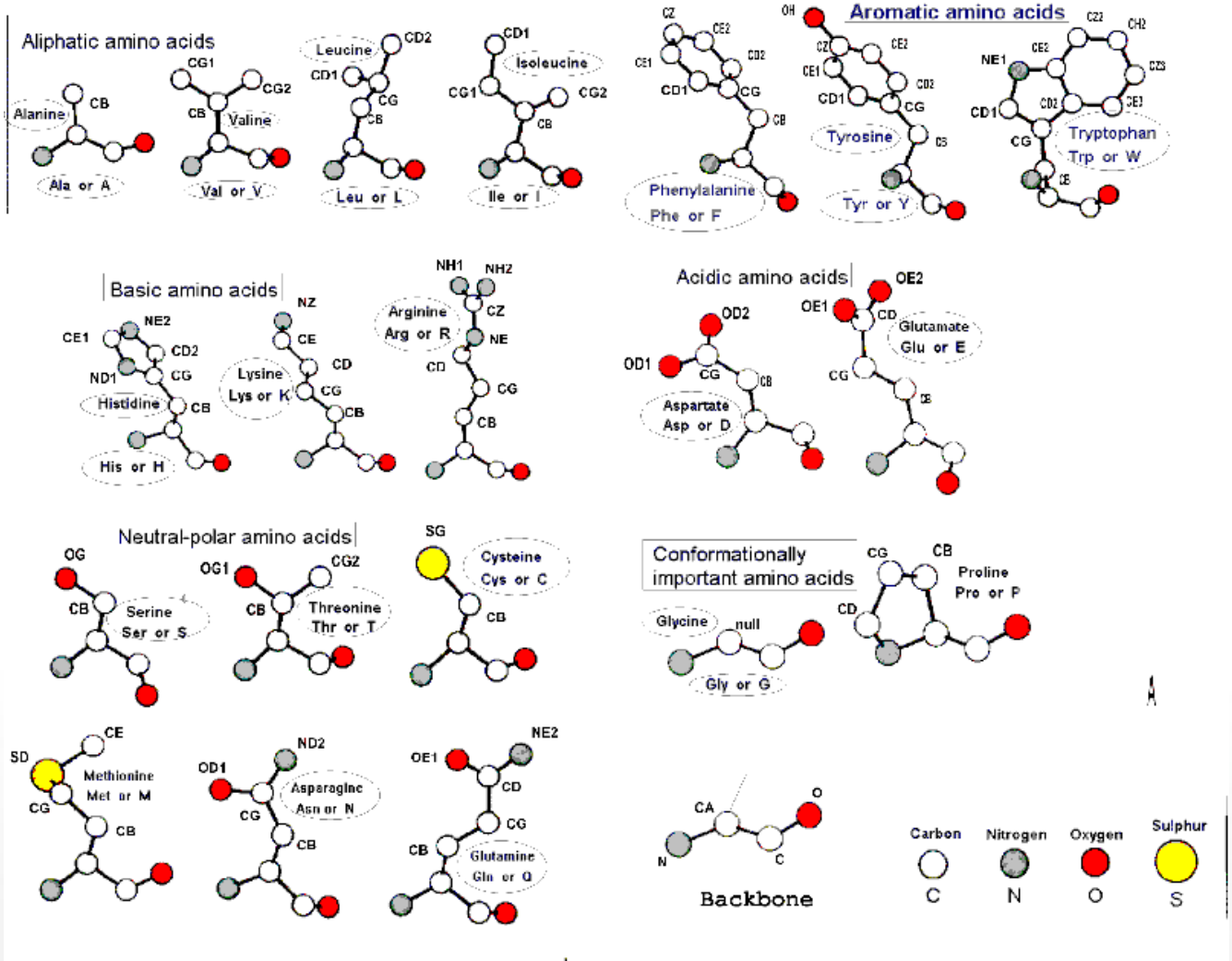
Альфа-цепь гемоглобина



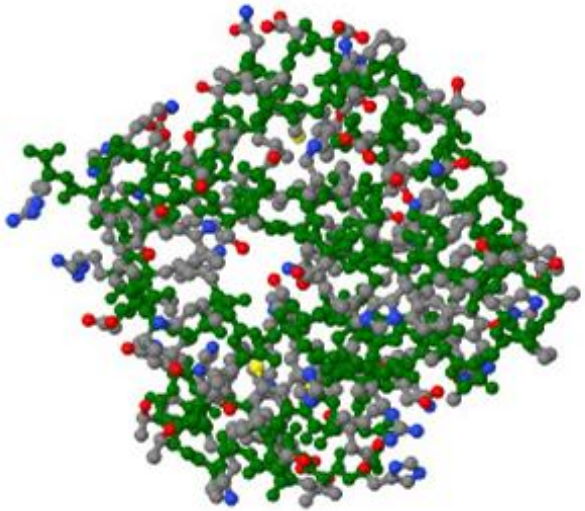
Альфа-цепь гемоглобина



Аминокислотные остатки



Альфа-цепь гемоглобина



Третичная структура (3D-структура)

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

Первичная структура (последовательность)

Последовательности альфа-цепей гемоглобина

```
HBA_HUMAN       MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
HBA_MOUSE       MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
HBA_ELEMA       MVLSDKDKTNVKATWSKVGDHASDYVAEALERMFFSFPTTKTYFPHFDLSHGSGQVKGHG

HBA_HUMAN       KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
HBA_MOUSE       KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
HBA_ELEMA       KKVGEALTQAVGHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLSSHQPTEFTP

HBA_HUMAN       AVHASLDKFLASVSTVLTSKYR
HBA_MOUSE       AVHASLDKFLASVSTVLTSKYR
HBA_ELEMA       EVHASLDKFLSNVSTVLTSKYR
```

HBA_HUMAN — белок человека, HBA_MOUSE — мыши, HBA_ELEMA — слона
Названия последовательностей – из общедоступного банка Uniprot
https://www.uniprot.org/

Последовательности альфа-цепей гемоглобина

```
HBA_HUMAN       MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
HBA_MOUSE       MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
HBA_ELEMA       MVLSDKDKTNVKATWSKVGDHASDYVAEALERMFFSFPTTKTYFPHFDLSHGSGQVKGHG

HBA_HUMAN       KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
HBA_MOUSE       KKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTP
HBA_ELEMA       KKVGEALTQAVGHLDDLPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLSSHQPTEFTP

HBA_HUMAN       AVHASLDKFLASVSTVLTSKYR
HBA_MOUSE       AVHASLDKFLASVSTVLTSKYR
HBA_ELEMA       EVHASLDKFLSNVSTVLTSKYR
```

Последовательности похожи, но немного различаются

Почему белки разные, но похожие?

Гены и белки

Геном
$3 \cdot 10^9$ букв у человека
Несколько млн. букв у бактерий

содержит →

Кодирующие участки
Около 1,5% генома у человека
Около 95% у бактерий

кодируют →

Белки
Около 25 000 у человека
Несколько тысяч у бактерий

Генетический код

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

Репликация ДНК

При делении клетки происходит **репликация ДНК**
Двойная спираль расплетается, и к каждой из двух цепей пристраивается новая комплементарная цепь.
Получаются **две** копии исходной двойной спирали.

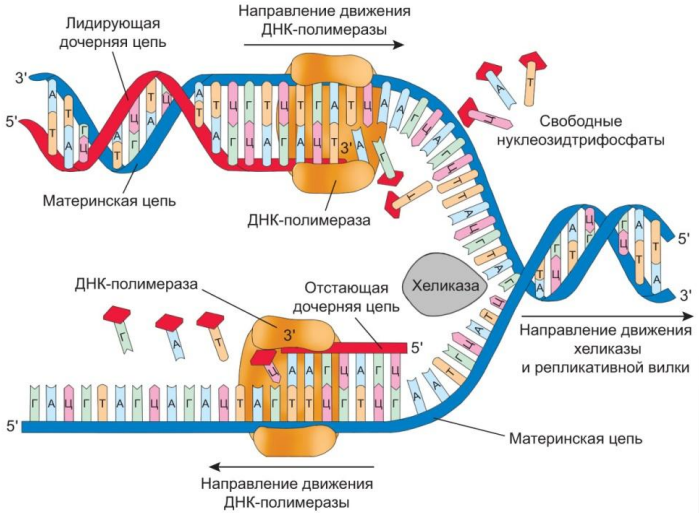


Рис. 16.4. Схема процесса репликации ДНК

http://profil.adu.by/pluginfile.php/4323/mod_book/chapter/11948/16.4.jpg

Передача информации между поколениями

Благодаря репликации ДНК двух клеток, получившихся при делении, идентична ДНК исходной клетки

… или не совсем идентична:

1. Бывают ошибки репликации (вставляется не тот нуклеотид)
2. Между делениями ДНК может испортится, в том числе какая-то пара нуклеотидов может замениться (из-за ошибок репарации. Репарация — это «починка» неминуемых повреждений)

Мутации

gatcaacactacttgacttcaa**g**acttaccataaagaaaac

gatcaacactacttgacttcaa**a**acttaccataaagaaaac

точечная замена

---

gatcaacactacttgacttcaa**ga**cttaccataaagaaaac

gatcaacactacttgacttcaacttaccataaagaaaac

делеция

---

gatcaacactacttgacttcaagacttaccataaagaaaac

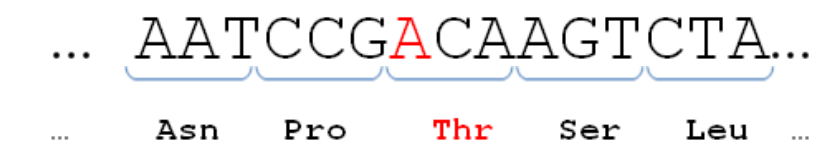
gatcaacactacttgacttcaaga**ta**cttaccataaagaaaac

инсерция
(вставка)

Точечные замены в гене

1) “молчащая”(синонимическая)мутация

2) замена остатка на близкий по свойствам



3) замена остатка на остаток с иными свойствами

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».

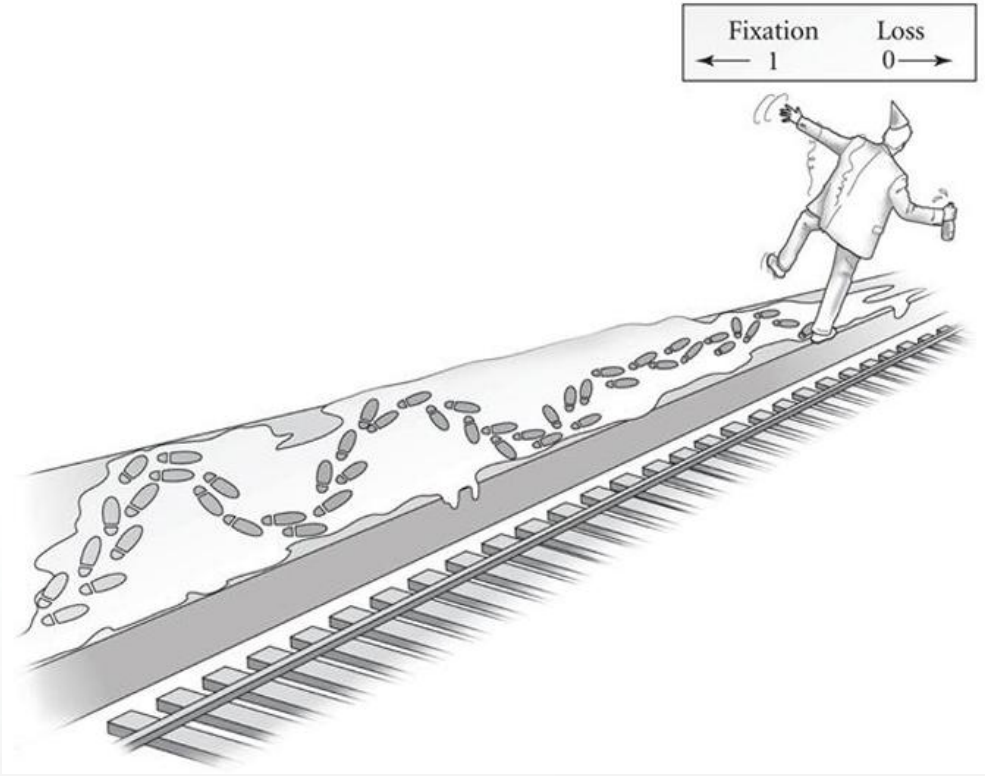
На тротуаре стоит пьяный и каждые 10 сек. делает шаг либо направо, либо налево, случайно выбирая направление. Как далеко он уйдёт за время T?

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

Случайное блуждание с поглощением

По длинной дамбе идёт пьяный и с каждым шагом отклоняется либо немного вправо, либо немного влево. Как скоро он свалится с дамбы?

Ответ: *скоро…*



Случайное блуждание с поглощением

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!

В результате генетического дрейфа происходит закрепление новых нейтральных (или почти нейтральных) вариантов.

Накопление нейтральных мутаций

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.
Во втором случае говорят, что мутация **закрепилась.**

Дупликация генов

Кроме точечных мутаций, бывают крупные перестройки генома

В частности, случаются дупликации генов

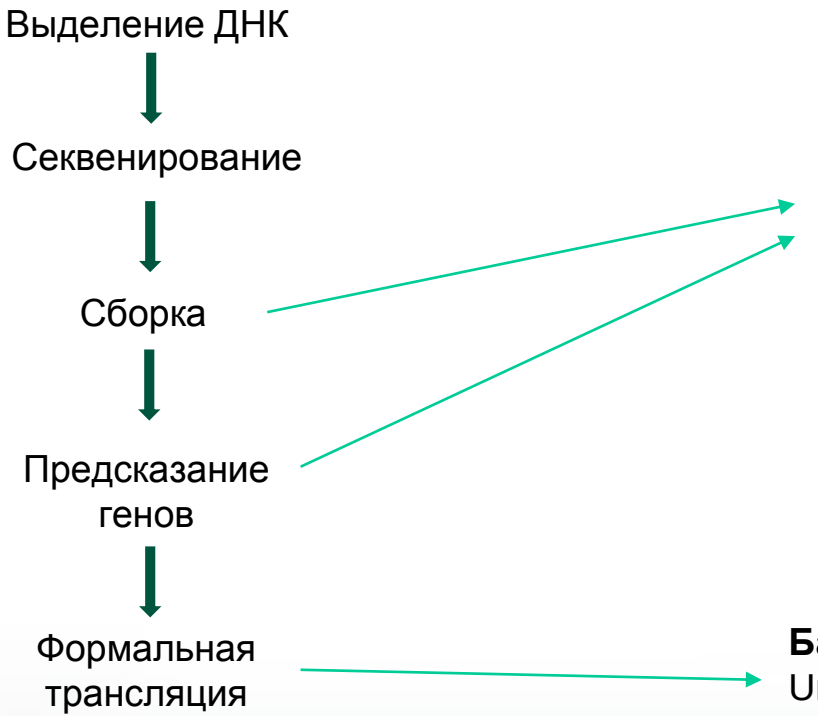
В большинстве случаев одна из копий накапливает мутации (нейтральные, поскольку есть вторая копия!) и превращается в нефункциональный **псевдоген**

Но изредка вторая копия приобретает новую функцию! Так возникают **паралогичные гены**

Примеры паралогов у человека: девять гемоглобинов, красный, зелёный и синий опсины, многие ферменты, чуть различающиеся специфичностью…

У человека подавляющее большинство генов имеет паралоги

Банки последовательностей

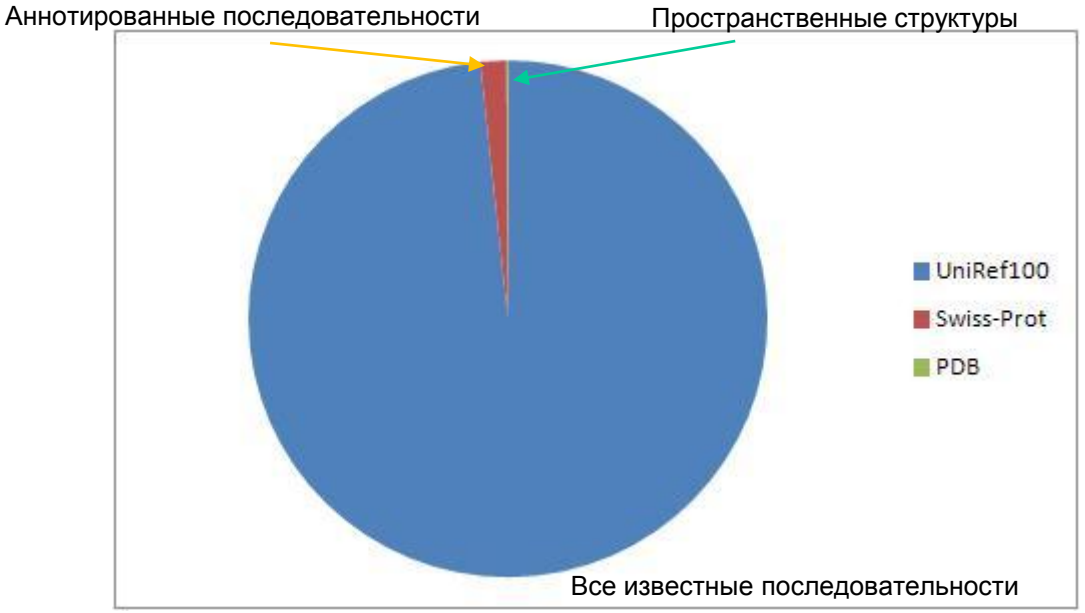


Банки нуклеотидных последовательностей:
GenBank, ENA, DDBJ, RefSeq genomic

Банки аминокислотных последовательностей:
Uniprot и RefSeq protein

Банк 3D структур PDB пополняется независимо!

Число белков в разных банках



Последовательностей определено во много раз больше, чем структур
Большинство последовательностей не аннотированы

Основная часть

Наконец-то про выравнивания

Последовательности миоглобинов человека, мыши и быка

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_MOUSE
MGLSDGEWQLVLNVWGKVEADLAGHGQEVLIGLFKTHPETLDKFDKFKNLKSEEDMKGSE
DLKKHGCTVLTALGTILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEIIIEVLKKRH
SGDFGADAQGAMSKALELFRNDIAAKYKELGFQG

>MYG_BOVIN
MGLSDGEWQLVLNAWGKVEADVAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE
DLKKHGNTVLTALGGILKKKGHHEAEVKHLAESHANKHKIPVKYLEFISDAIIHVLHAKH
PSDFGADAQAAMSKALELFRNDMAAQYKVLGFHG
```

Напишем последовательности друг под другом, чтобы было лучше 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, что означает, что эти белки произошли от общего предка. За время, прошедшее от существования общего предка, некоторые буквы менялись, но большинство остались неизменными.

Последовательности миоглобинов человека и рыбы

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_DANRE
MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
GQGALRRVMDAVIGDIDGYYKEIGFAG
```

Разная длина, как сравнивать?

Последовательности миоглобинов человека и рыбы

```
>MYG_HUMAN
MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE
DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH
PGDFGADAQGAMNKALELFRKDMASNYKELGFQG

>MYG_DANRE
MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGDLAGSPAVAAH
GATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKAGLDAA
GQGALRRVMDAVIGDIDGYYKEIGFAG
```

Разная длина, как сравнивать?
Ответ: **выравнивание**

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_DANRE ----MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGD-LAGSP 55
              .: :***: ** ****  .:* *** **** :*:**: * **. :.. * : .*

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

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

Последовательности миоглобинов человека и рыбы

Гэпы показывают, что для данного аминокислотного остатка **нет** гомологичного в другом белке

Биологическая причина — инсерции и делеции, закрепившиеся в эволюции.

Отличить инсерцию от делеции мы не можем.

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_DANRE ----MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGISQGD-LAGSP 55
               .: :***: ** ****  .:* *** **** :*:**: * **. :.. * : .*

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

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

Выравнивание белков, гомологичных не по всей длине

```
ANTP_DROME MTMSTNNCESMTSYFTNSYMGADMHHGHYPGNGVTDLDAQQMHHYSQN----ANHQGNMP 56
HXA1_HUMAN ----------------------------------MDNARMNSFLEYPILSSGDSGTCS 24
                                              :*  :*: : :     :...*.
ANTP_DROME YPRFPPYDRMPYYN-----------------GQGMDQQQQHQVYSRPDSPSSQVGGVMPQ 99
HXA1_HUMAN ARAYPSDHRITTFQSCAVSANSCGGDDRFLVGRGVQIGSPHHHHHH-----HHHHPQPAT 79
              :*  .*:  ::                  *:*::  . *: : :      :
ANTP_DROME AQTNGQLGVPQQQQQQQQQPSQNQQQQQAQQAPQQLQQQLPQVTQQVTHPQQQQQQPVVY 159
HXA1_HUMAN YQTSGNLGVSYSHSS--CGPSYGSQNF----------------------------SAPY 108
            **.*:***  .:..    ** ..*:                               . *
ANTP_DROME ASCKLQAAVGGLGMVPEGGSPPLVDQMSGHHMNAQMTLPHHMGHPQAQLGYTD--VGVPD 217
HXA1_HUMAN SPYALNQEAD-----VSGGYPQCAPAVYSGNLSSPMVQHH-----HHHQGYAGGAVGSPQ 158
           :   *:  ..      .** *  .  : . ::.: *.  *     : : **:.  ** *:
ANTP_DROME VTE--VHQNHHNMGMYQQQSGVPPVGAPPQGMMHQGQGPPQMHQGHPGQHTPPSQNPNSQ 275
HXA1_HUMAN YIHHSYGQEHQSLALATYNNSLSP-----------------LHASHQE----ACRSP-AS 196
             .    *:*:.:.:   :..: *                  :* .*       .:.* :.
ANTP_DROME SSGMPSPLYPWMRSQFGK-----------CQERKRGRQTYTRYQTLELEKEFHFNRYLTR 324
HXA1_HUMAN ETSSPAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTR 256
           .:. *:  : **: : .              : :  * .:*  *  *********:****
ANTP_DROME RRRIEIAHALCLTERQIKIWFQNRRMKWKKENKTKGEPGSG-------GEGDEITP---- 373
HXA1_HUMAN ARRVEIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSS 316
            **:*** :* *.* *:********** **.:*    * *         :.:* :
ANTP_DROME -PNSPQ------------ 378
HXA1_HUMAN SPCVPSPGSSTSDTLTTSH 335
               *  *.
```

Локальное выравнивание белков

```
ANTP_DROME   QFGKCQERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTER 340
HXA1_HUMAN   EYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTRARRVEIAASLQLNET 275
             ..*   .    *  .*  *  *********.**** **.*** .* * *

ANTP_DROME   QIKIWFQNRRMKWKKENK 358
HXA1_HUMAN   QVKIWFQNRRMKQKKREK 289
             *.********** **  *
```

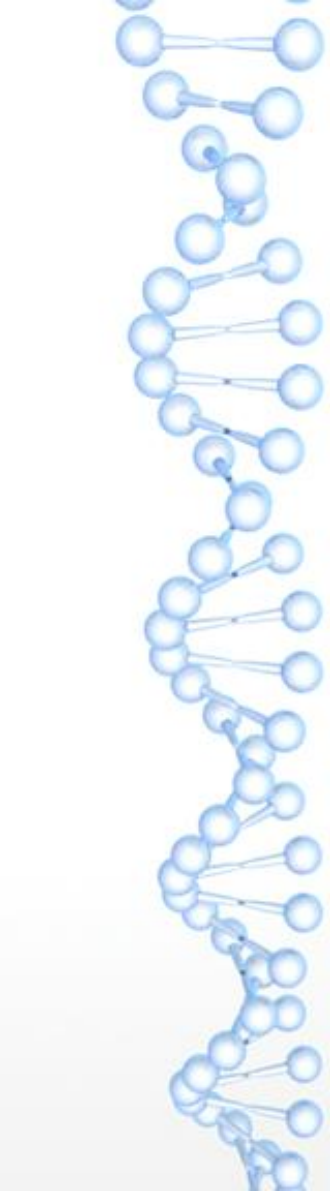
Программа **глобального** выравнивания только расставляет гэпы = выравнивает последовательности по всей длине

Программа **локального** выравнивания:

1) выбирает в каждой последовательности по участку;
2) выравнивает между собой эти участки.

Когда какое выравнивание нужно

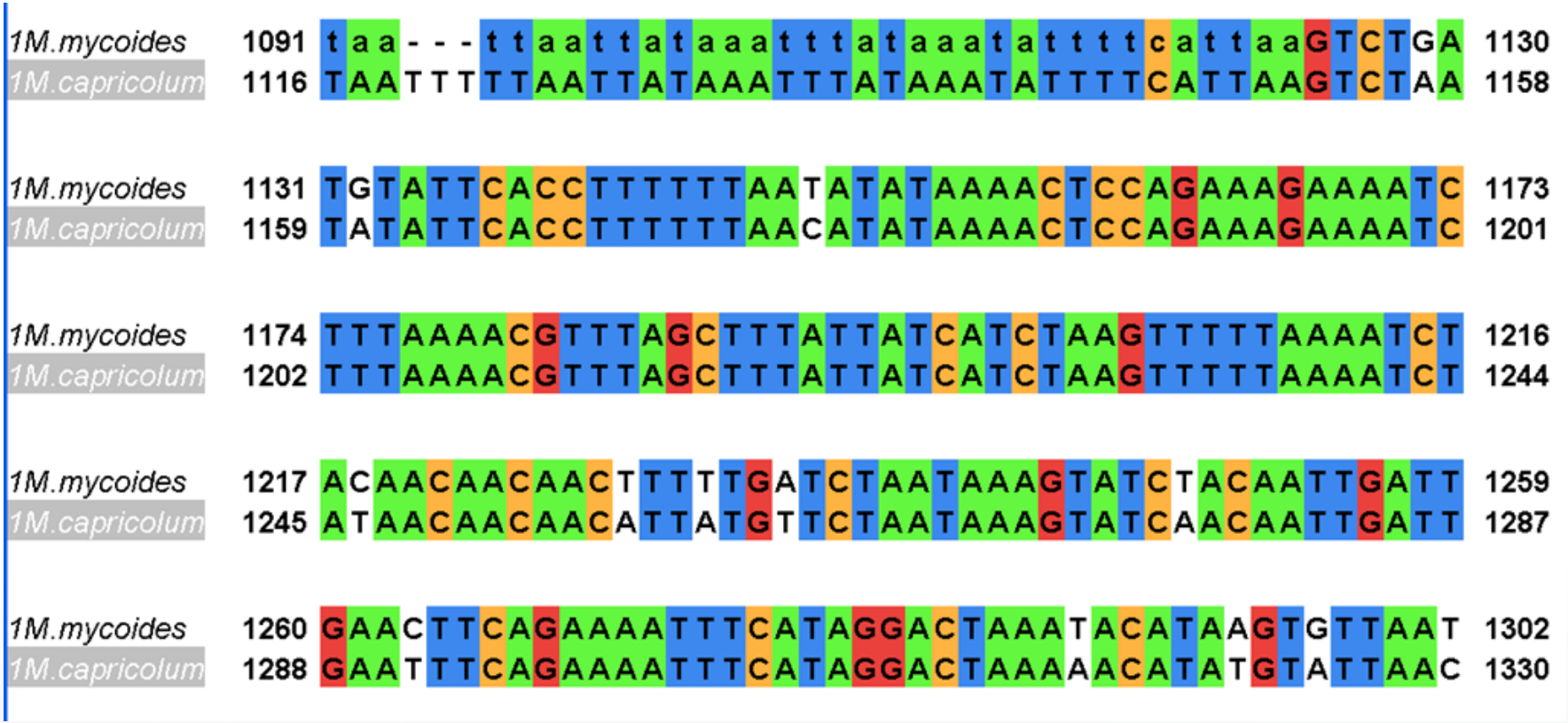
- Если мы уверены, что две последовательности гомологичны по всей длине, то глобальное
- Если две последовательности содержат (относительно небольшие) гомологичные участки, то локальное
- Если мы ничего заранее не знаем, то предпочтительно тоже локальное (на глобальном можно не увидеть хороший участок сходства — признак гомологии)



Гомология и выравнивание

- Гомология – происхождение от общего предка
- Выравнивание последовательностей должно отражать эволюцию
- Выравнивание имеет биологический смысл только для гомологичных участков геномов или белков

Участок выравнивания двух геномов



Алгоритмы выравнивания

- Парное глобальное выравнивание (global alignment of two sequences) — алгоритм Нидлмана – Вунша (Needleman & Wunsch)
- Парное локальное выравнивание (local alignment of two sequences) — алгоритм Смита – Уотермена (Smith & Waterman)
- Множественное выравнивание (multiple alignment), когда последовательностей больше двух — великое множество алгоритмов, самые известные: ClustalW, Muscle, MAFFT, T-Coffee.

Принцип работы алгоритмов парного выравнивания

- Имеется процедура вычисления **веса** выравнивания (alignment score)
- Алгоритм Нидлмана – Вунша находит оптимальную расстановку гэпов: такую, чтобы вес получившегося выравнивания был максимальным
- Алгоритм Смита – Уотермена находит оптимальные (то есть дающие максимальный вес):
  - ✔ пару участков в последовательностях;
  - ✔ для этой пары — расстановку гэпов.
- Оба алгоритма используют **динамическое программирование**

Как вычисляется вес выравнивания

Цель: вес должен быть тем больше, чем больше выравнивание похоже на правильное (отражающее эволюцию).
В настоящее время устоялся следующий подход:

- Для последовательностей ДНК за каждую пару совпадающих букв к весу выравнивания **прибавляется** некоторое число (например, 1), а за каждую пару разных букв **вычитается** другое число (например, 2)
- Для последовательностей белков есть понятие веса замены букв: чем реже буквы заменяются друг на друга в эволюции, тем меньше этот вес

  Всего нужно задать 210 чисел, некоторые из них положительные, другие отрицательные. Они составляют **матрицу весов аминокислотных замен**.
- За гэпы вычитается **штраф** (gap penalty)

Матрица весов аминокислотных замен BLOSUM 62

Треугольная (симметричная) матрица

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

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

Вес выравнивания

L Y P W M R S

T F D W M K V

Буква первой посл-ти	Буква второй посл-ти	Значение из матрицы
L	T	–1
Y	F	3
P	D	–1
W	W	11
M	M	5
R	K	2
S	V	–2

Итого: –1+3–1+11+5+2–2 = **17**

Замечания

- Программы выравнивания действуют формально и выдадут выравнивание, даже если на вход им подать негомологичные белки. Смысла такое выравнивание иметь не будет
- Даже если белки гомологичны, в выравнивании, выданном программой, не обязательно всюду будут сопоставлены гомологичные аминокислотные остатки или нуклеотиды

  Алгоритмы выравнивания сделаны так, чтобы максимизировать долю правильных сопоставлений. Но невозможно придумать такой алгоритм, который бы выдавал биологически правильный ответ всегда. Выравнивание, выданное программой — это всего лишь реконструкция причины (эволюции) по последствиям (т. е. современным последовательностям).

Зачем выравнивать?

- Чтобы предсказать функцию белка на основании установленной гомологии с экспериментально изученными белками
- Чтобы выяснить, какие аминокислотные остатки консервативны (следовательно, важны для функции белка)
- Чтобы оценить время расхождения последовательностей, а если последовательностей много — реконструировать их эволюционную историю

Вопросы и ответы

Что такое гомология?
Ответ: общность происхождения
(*НЕПРАВИЛЬНО говорить «последовательности гомологичны на 56%. Последовательности либо гомологичны, либо нет*)

Как определить, гомологичны ли два белка?
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 укладку полипептидной цепи в пространстве.

Какой процент идентичности служит надёжным признаком гомологии?
Ответ: для белков обычно более 20–25% на достаточно длинном участке
(а точнее будет в лекции про BLAST)