

Вводная лекция

Алгоритмы в биоинформатике

Антон Елисеев
eliseevantoncoon@gmail.com

О чем курс?

Первый модуль

- От ДНК до белков. Почему важны короткие участки генома?
- Что означает эволюция для ДНК?
- Зачем и как сравнивать две последовательности ДНК?
- ~ многие последовательности ДНК с образцом?
- Филогения: смысл и значение для одного гена.
- Что бывает кроме мутаций, инсерций и делеций?

О чем курс?

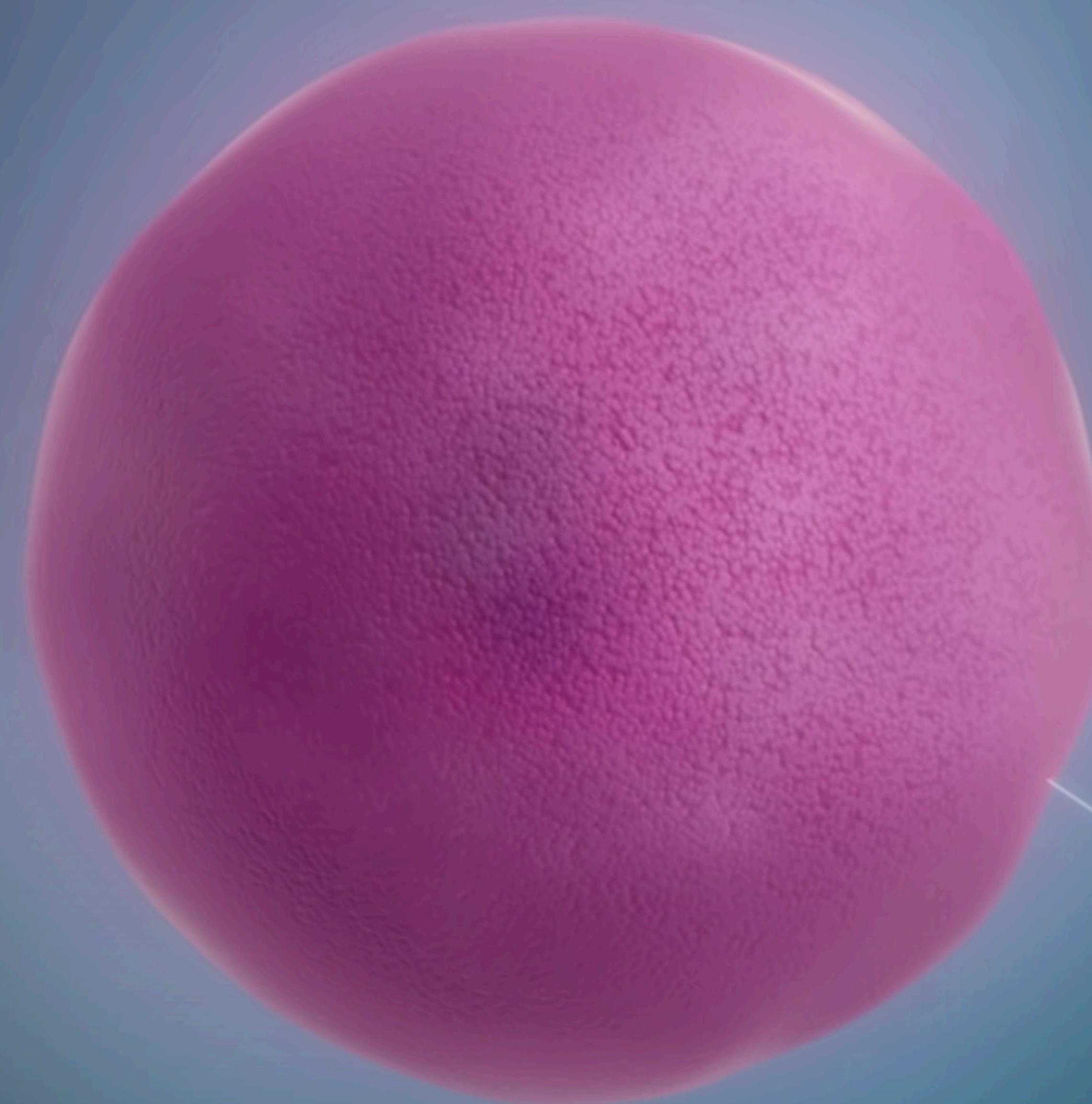
Второй модуль

- Секвенирование! Артефакты и важная информация.
- NGS данные. Как эффективно хранить короткие прочтения?
- Сборка генома из коротких прочтений.
- Сборка многих геномов. Метагеном, гаплотипы и связанные задачи.
- Эволюция и ее параметры. Зачем нужны вероятностные модели?
- Вторичная структура РНК.

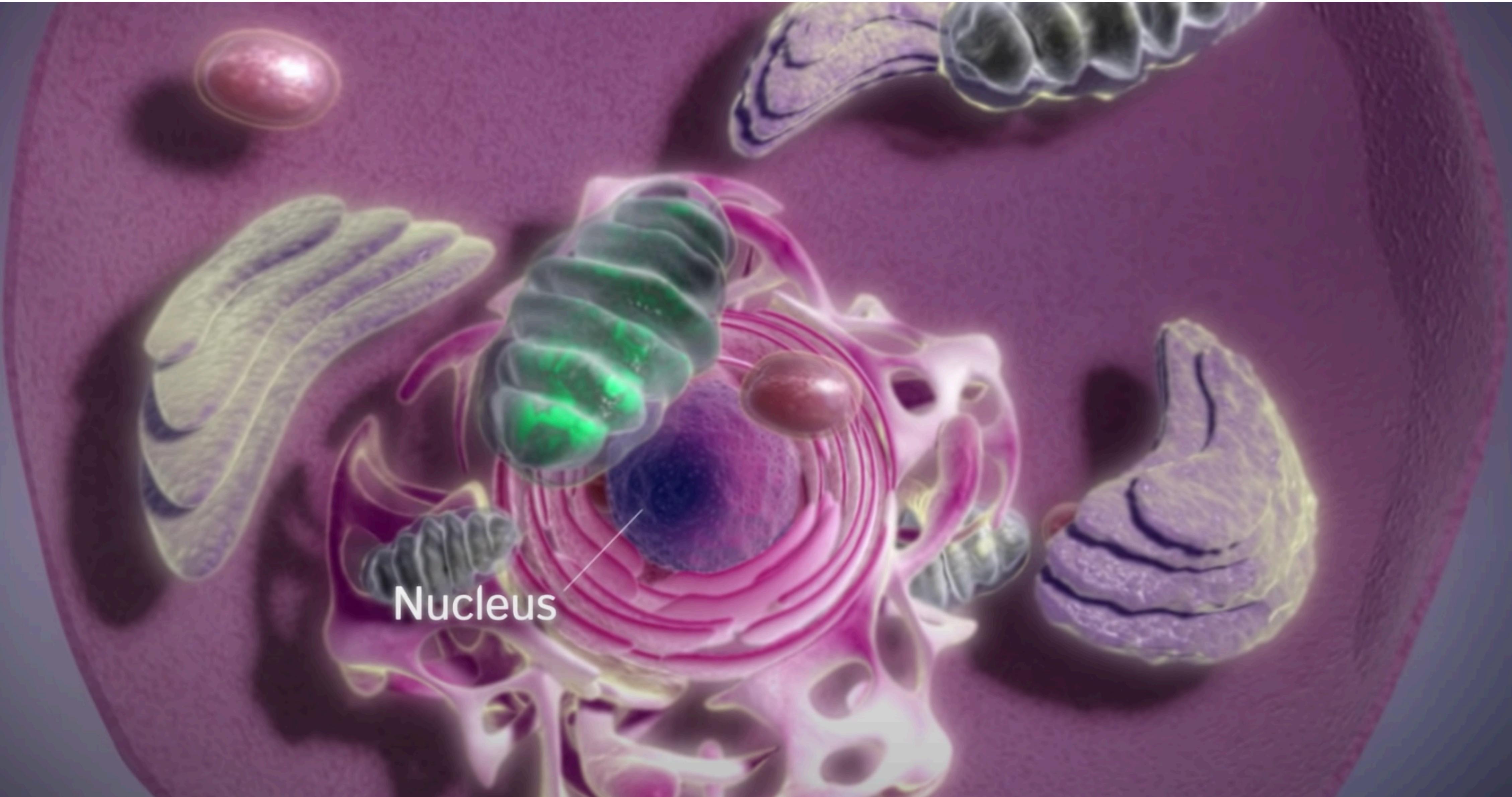
Какие темы не затронем

- ML в биоинформатике.
- Математические модели для анализа динамики популяций.
- Анализ 3D структур. Фолдинг и докинг.
- Анализ экспрессе генов.
- Гуманизация геномов.
- ...

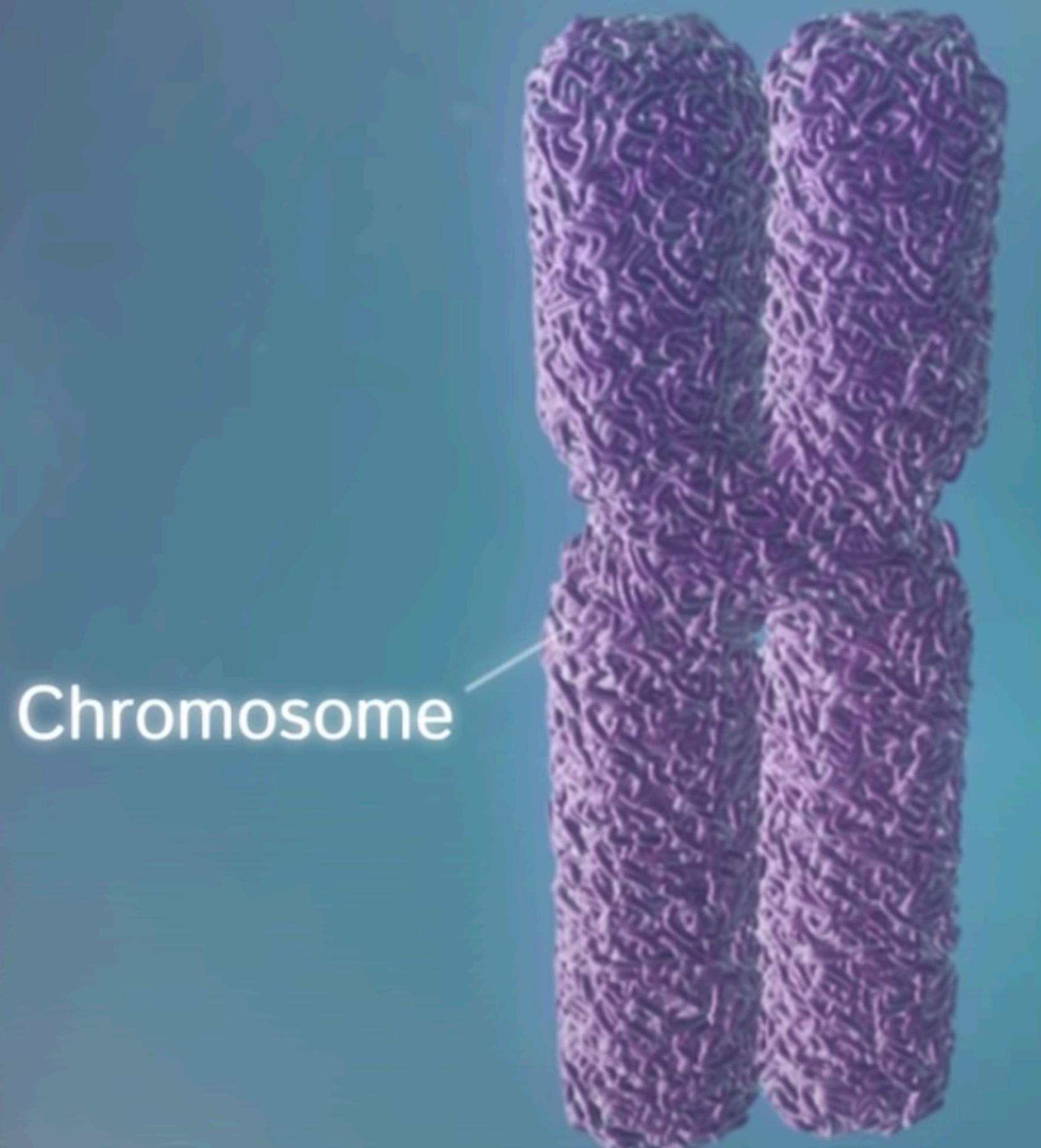
От ДНК до белка



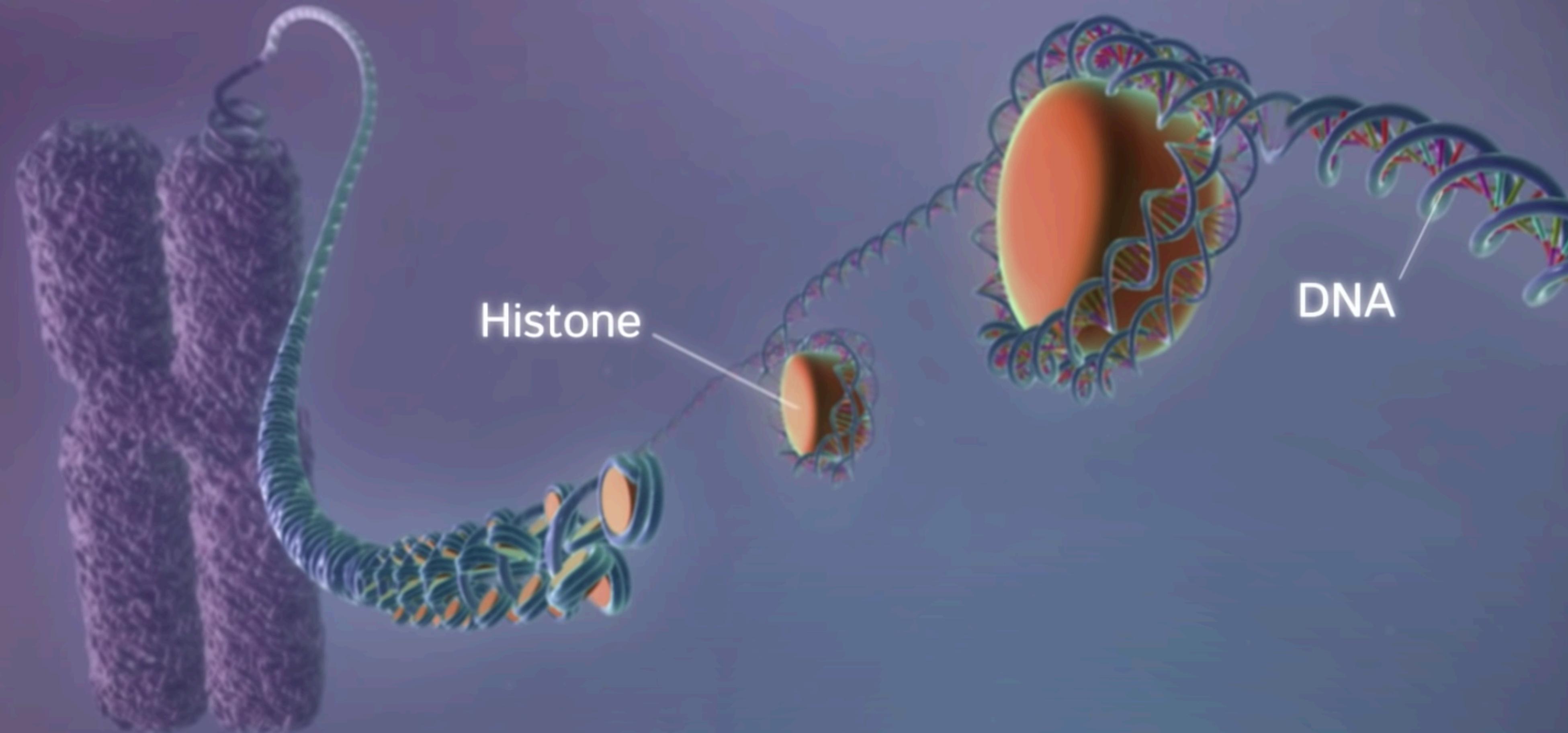
От ДНК до белка



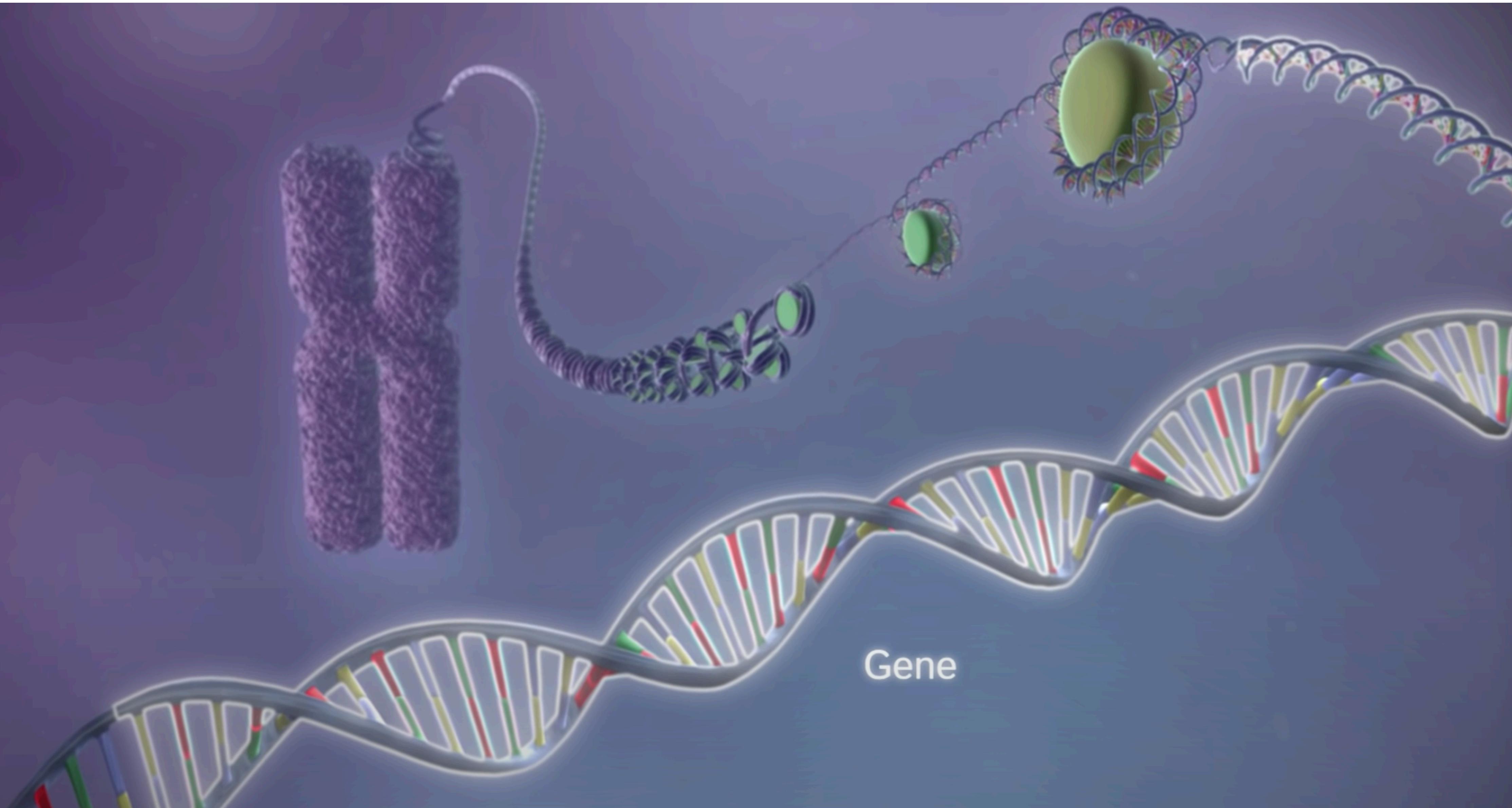
От ДНК до белка



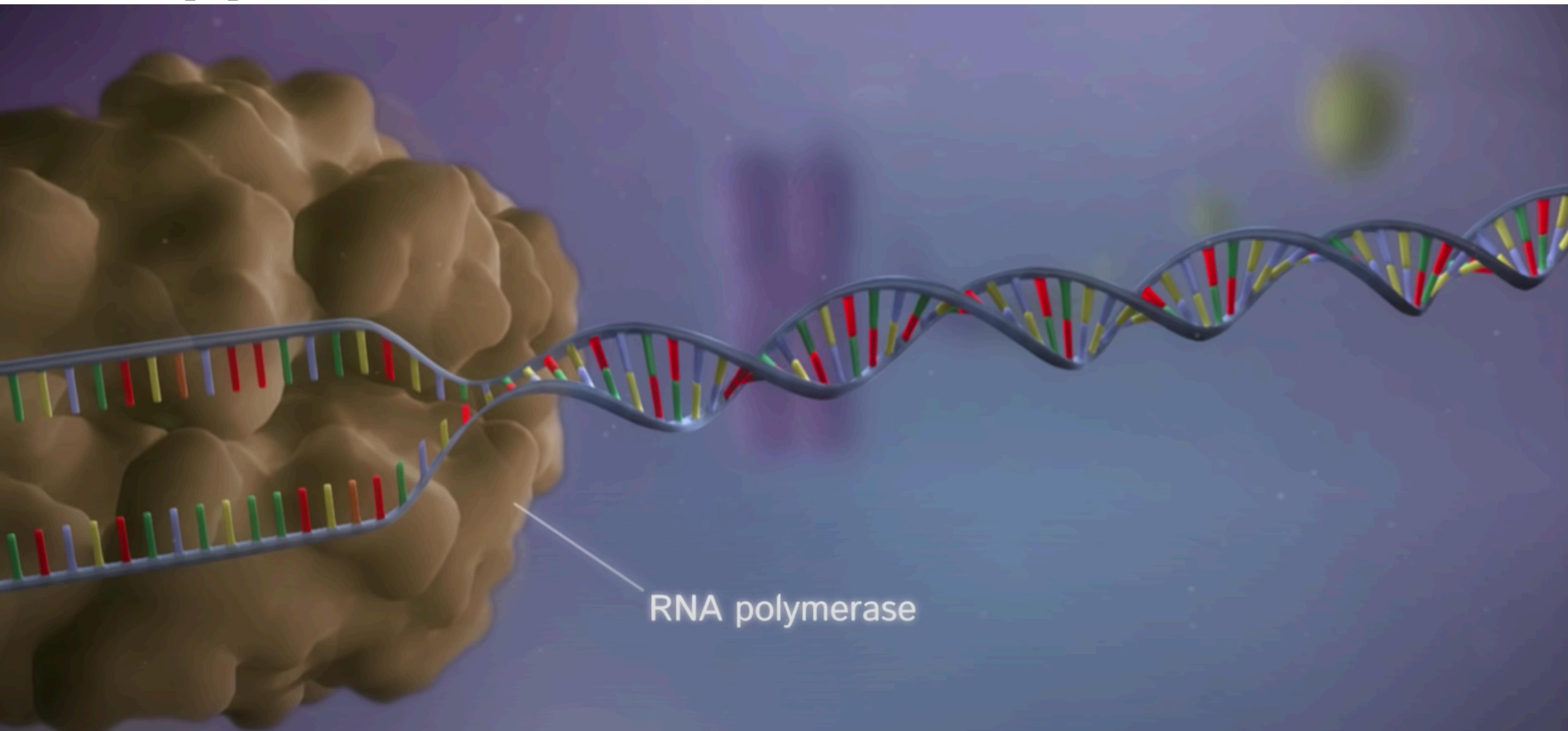
От ДНК до белка



От ДНК до белка

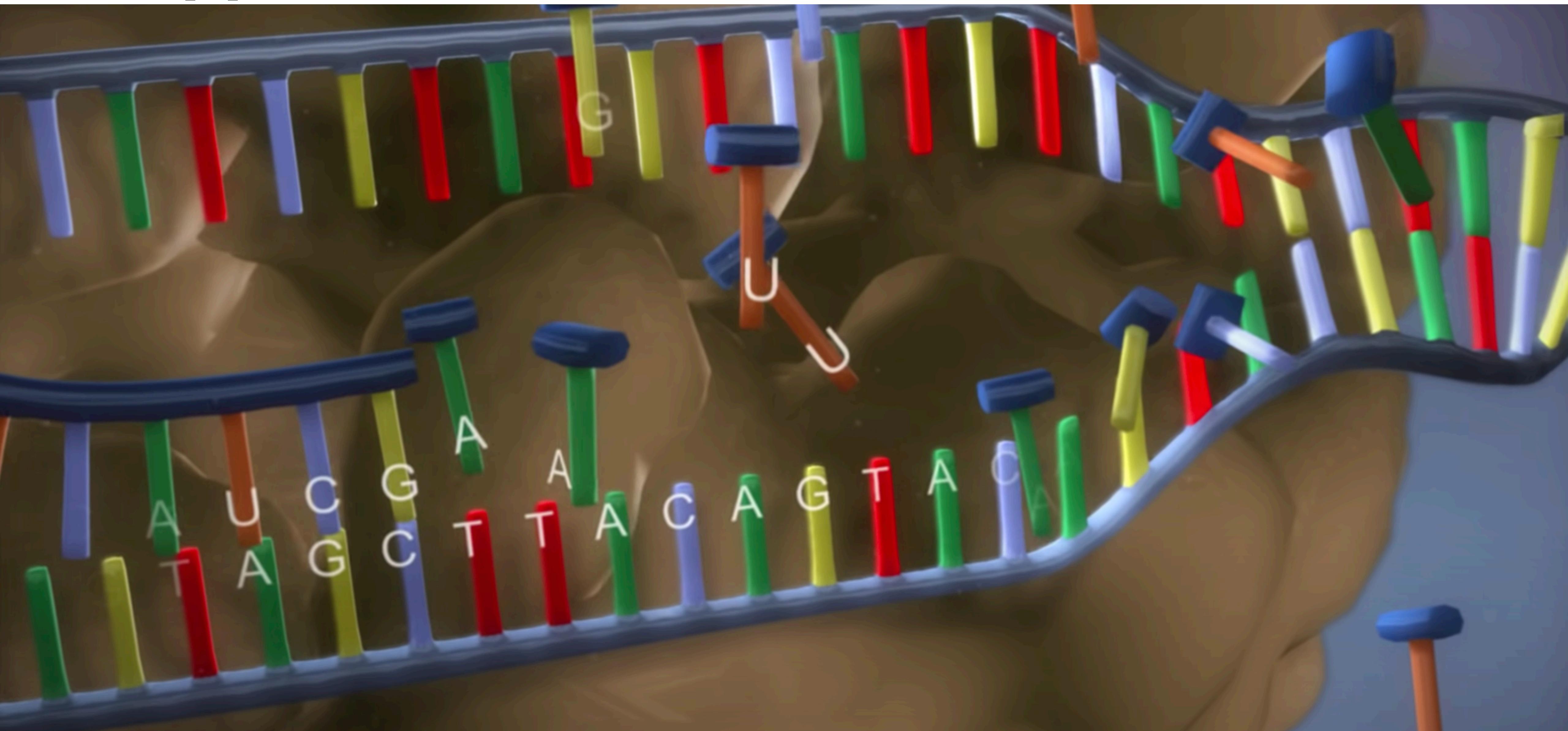


От ДНК до белка

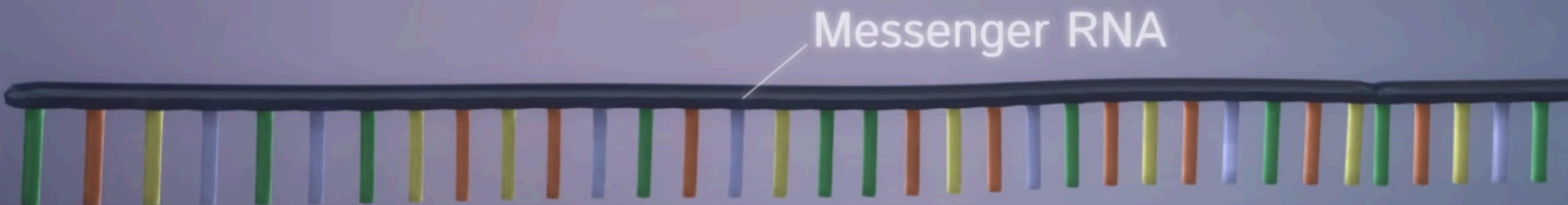


RNA polymerase

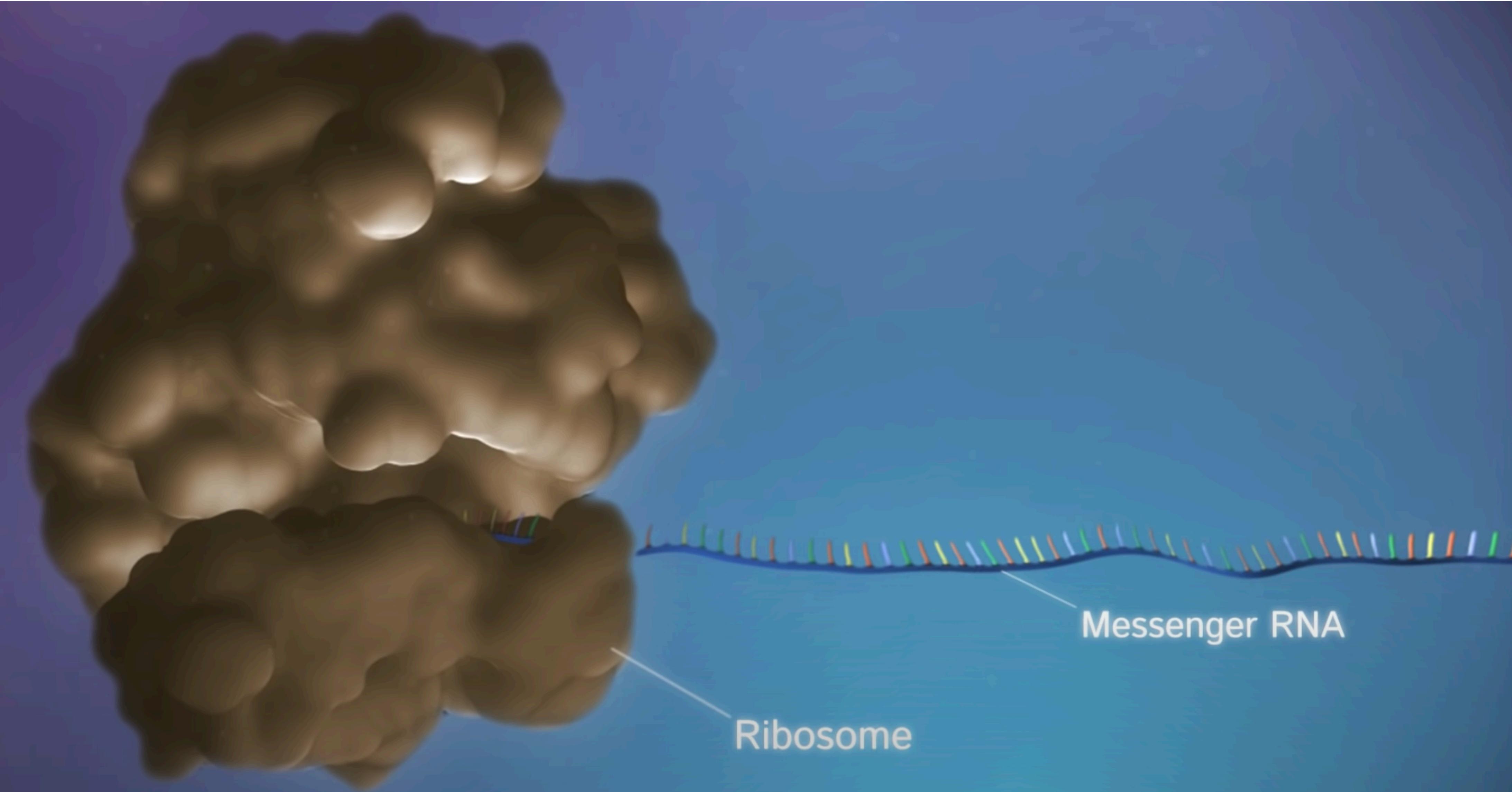
От ДНК до белка



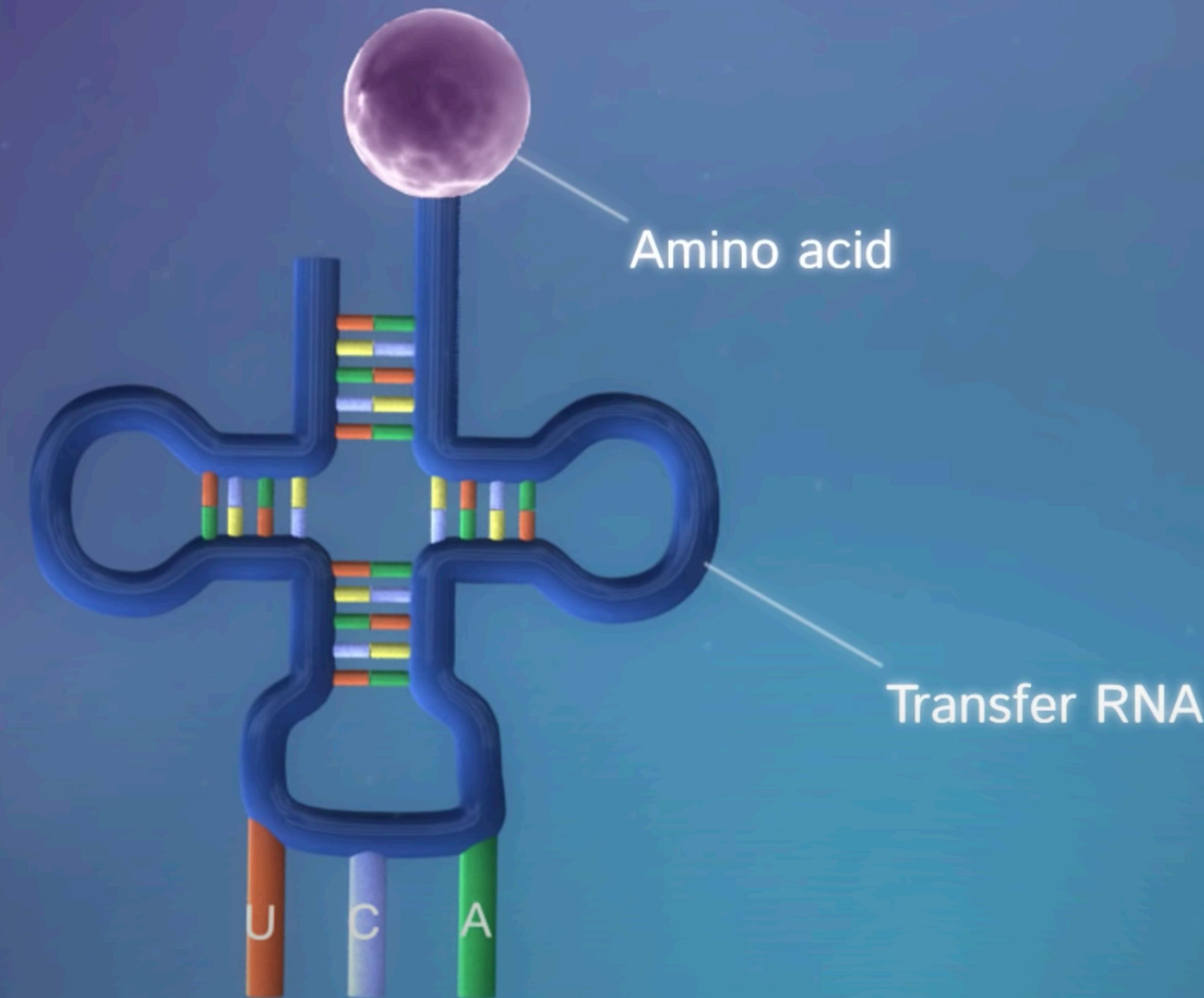
От ДНК до белка



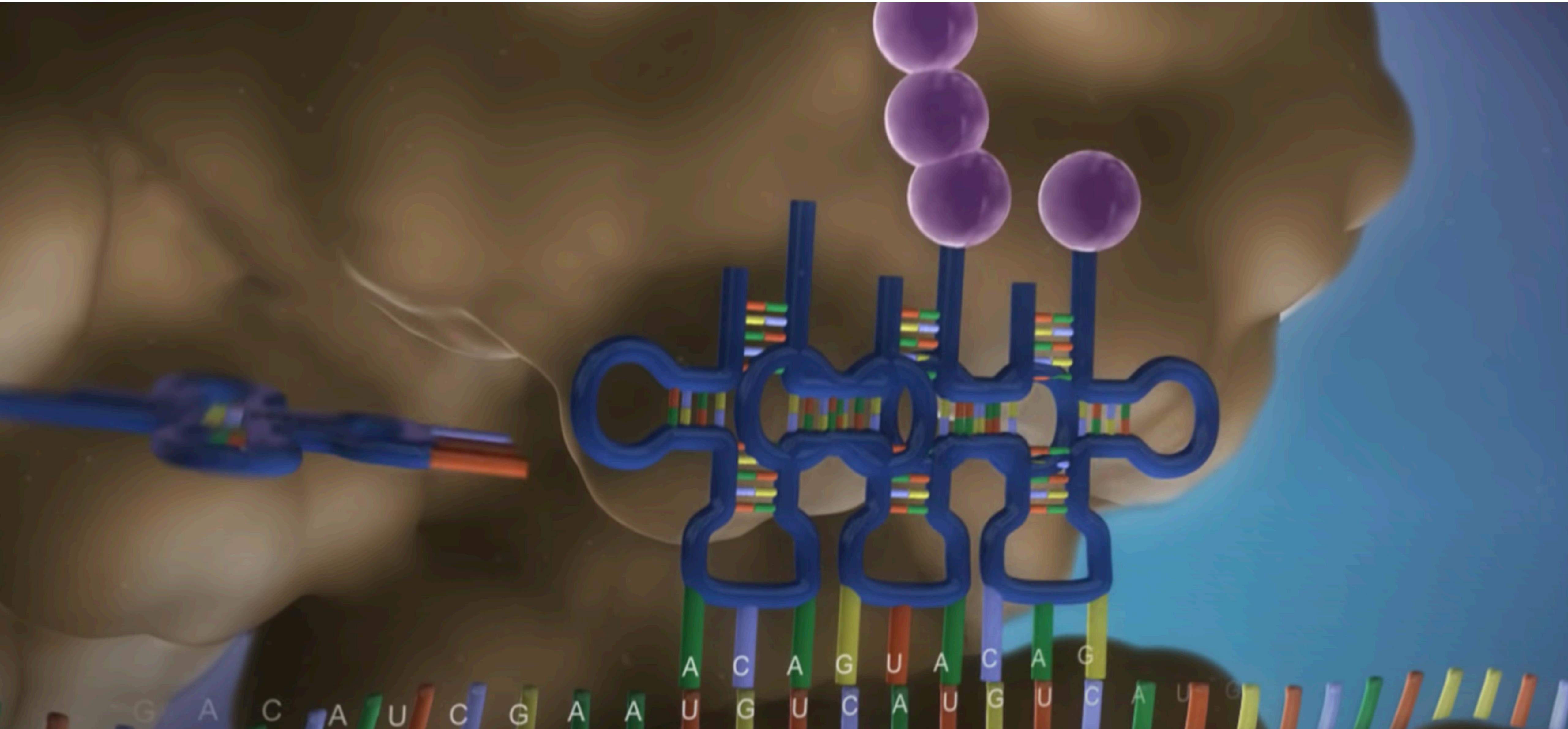
От ДНК до белка



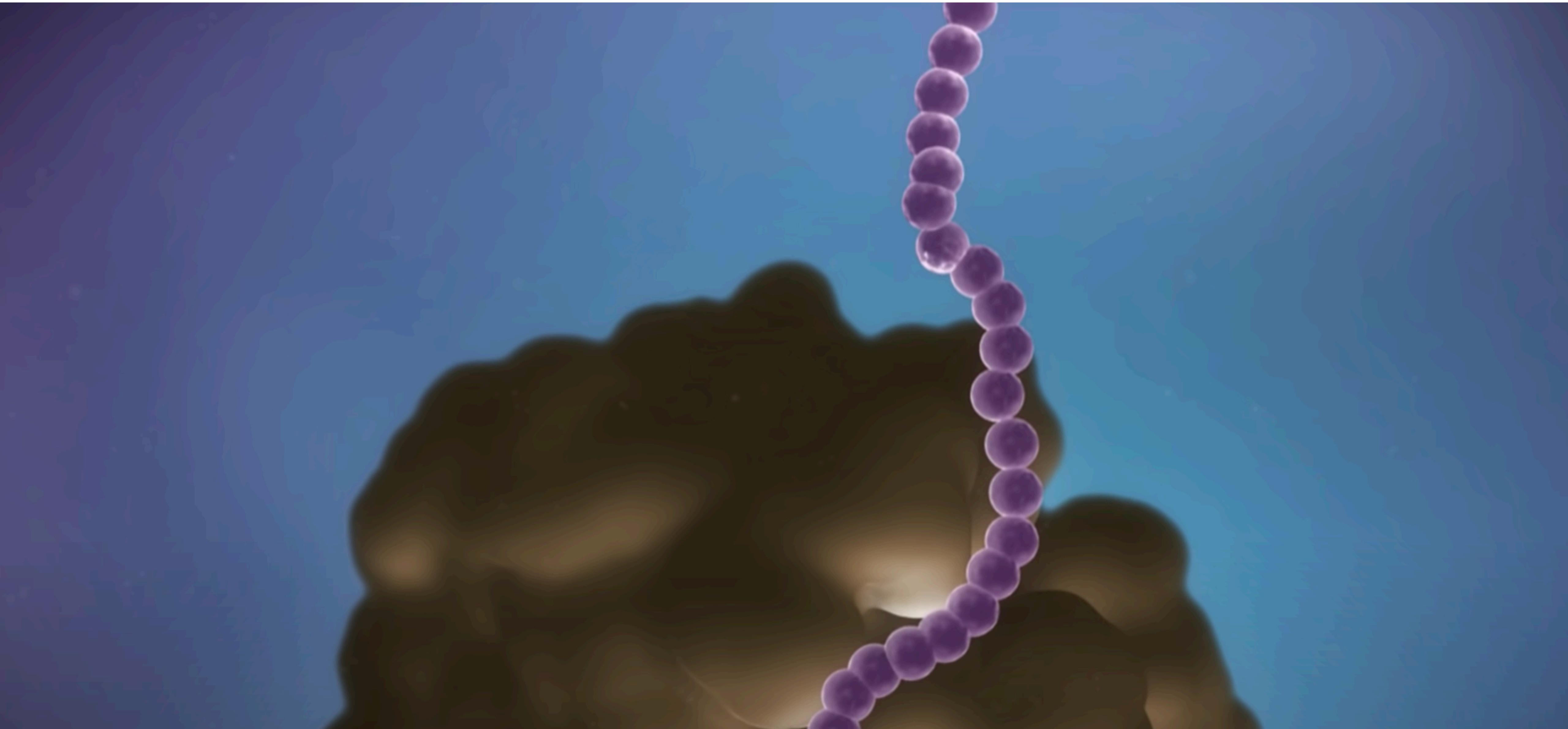
От ДНК до белка



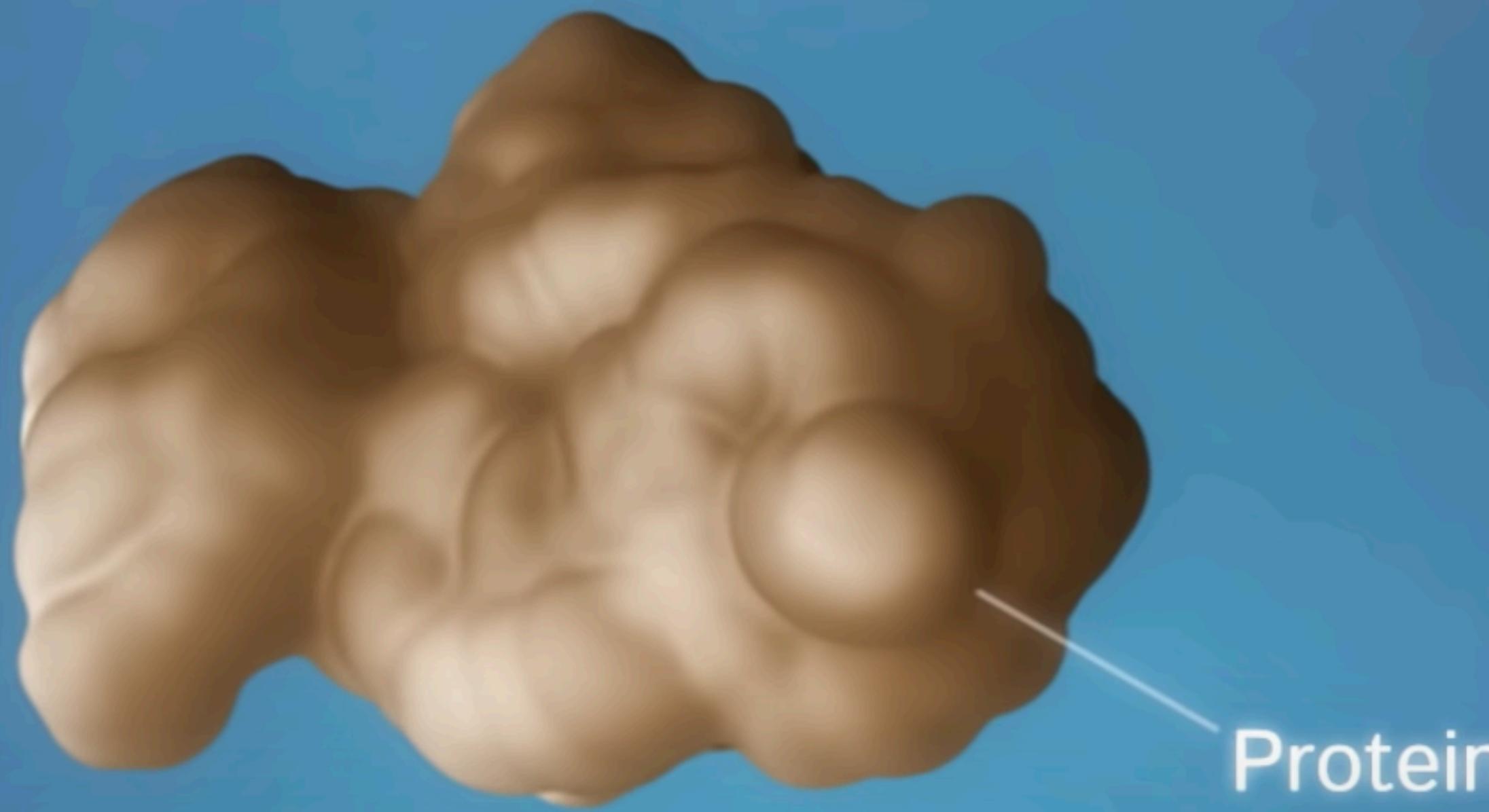
От ДНК до белка



От ДНК до белка



От ДНК до белка



Protein

Эволюция и ДНК

“Ничто в биологии не имеет смысла, кроме как в свете эволюции”
(Добржанский)

Условия окружающей
среды

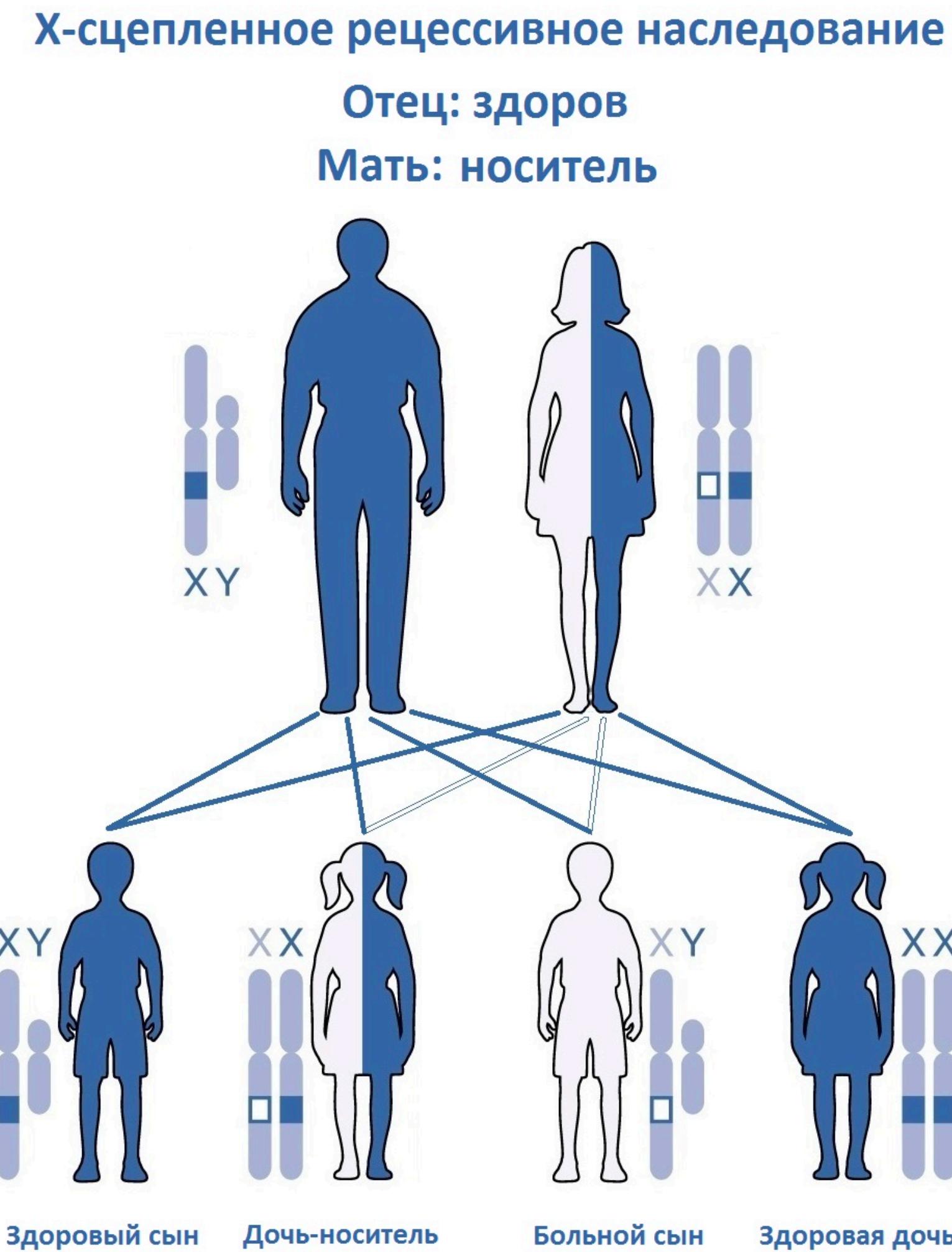
+

Мутации, вставки и
удаления в ДНК

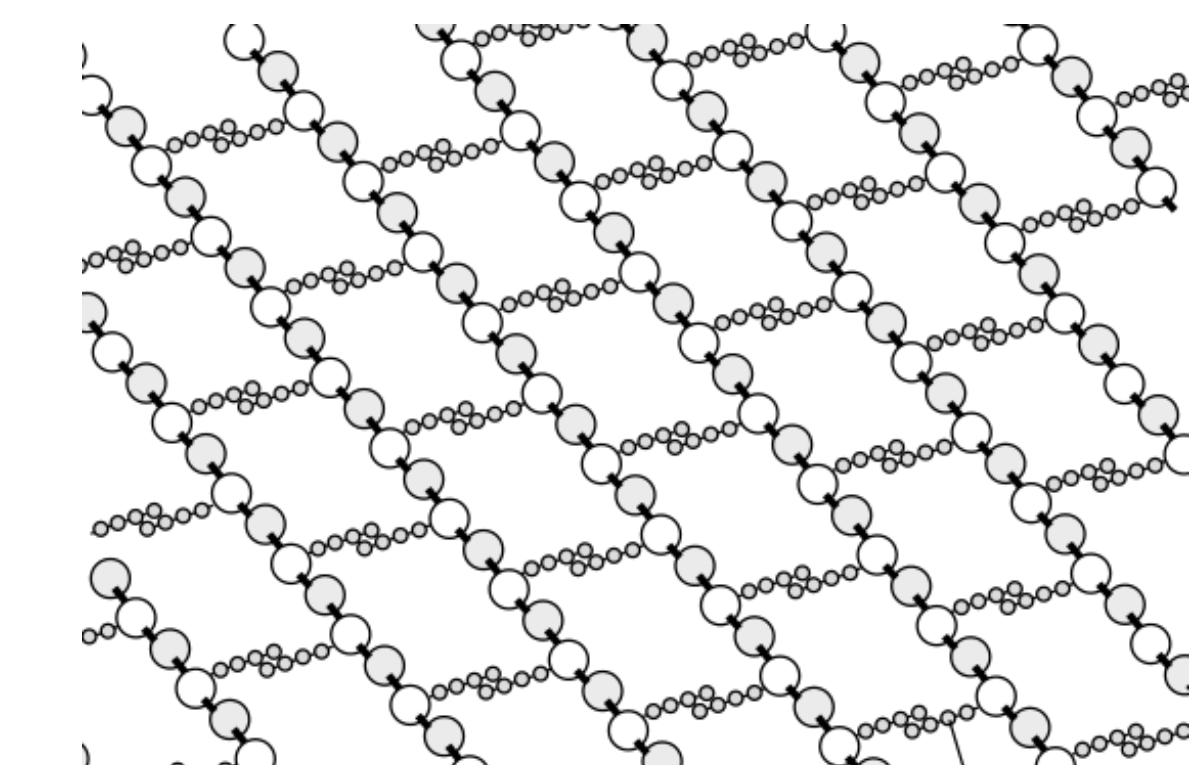
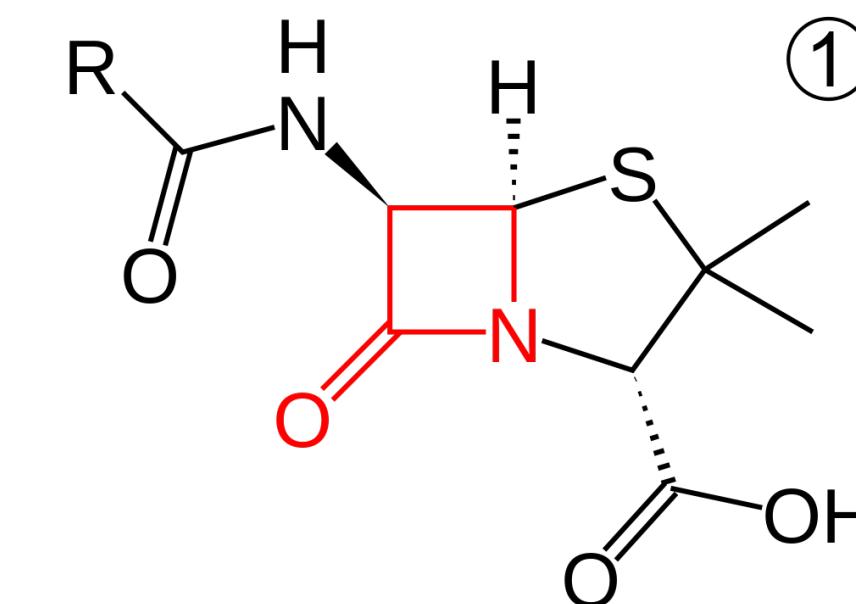


Эволюция и ДНК. Примеры.

Гемофилия А

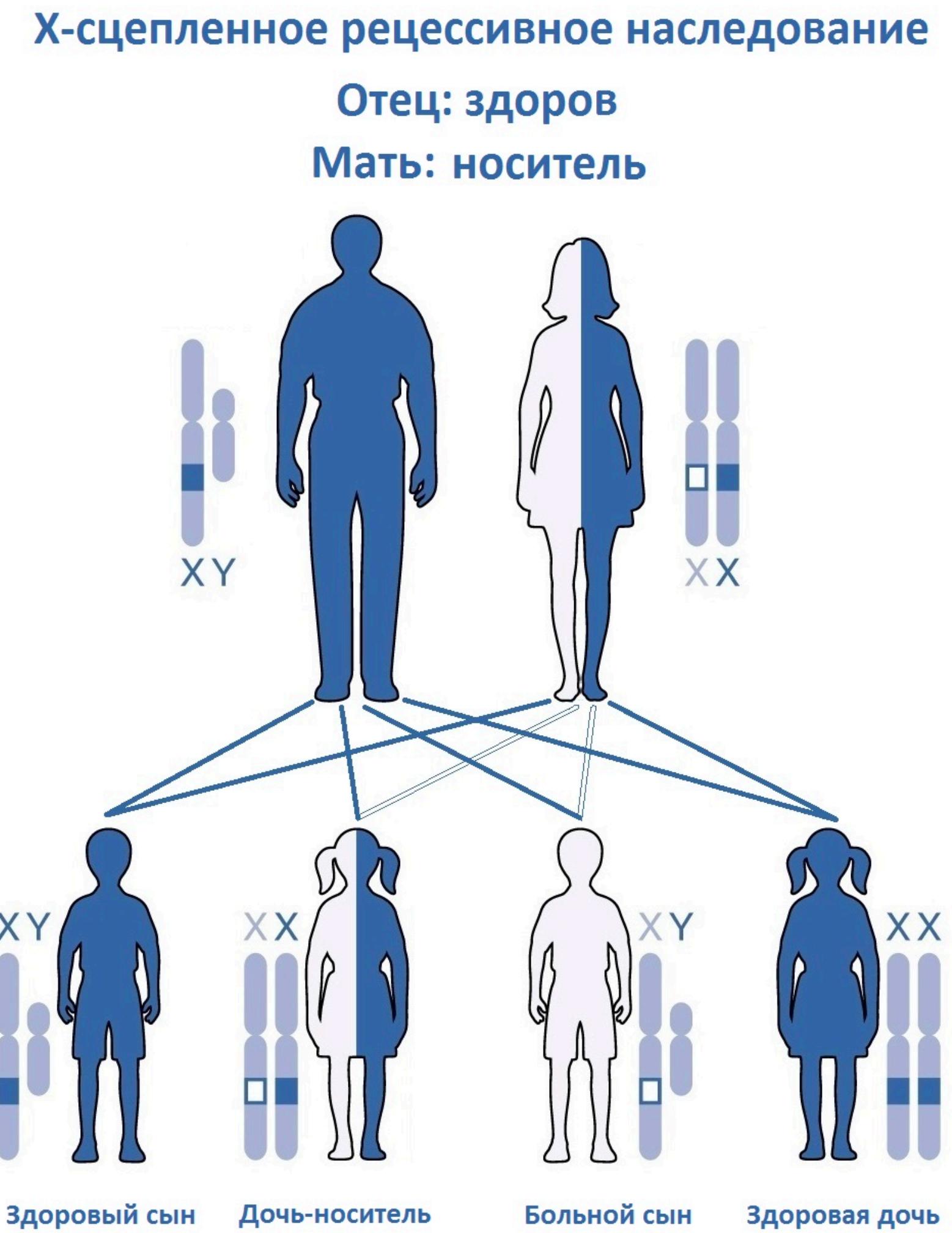


Устойчивость к антибиотикам

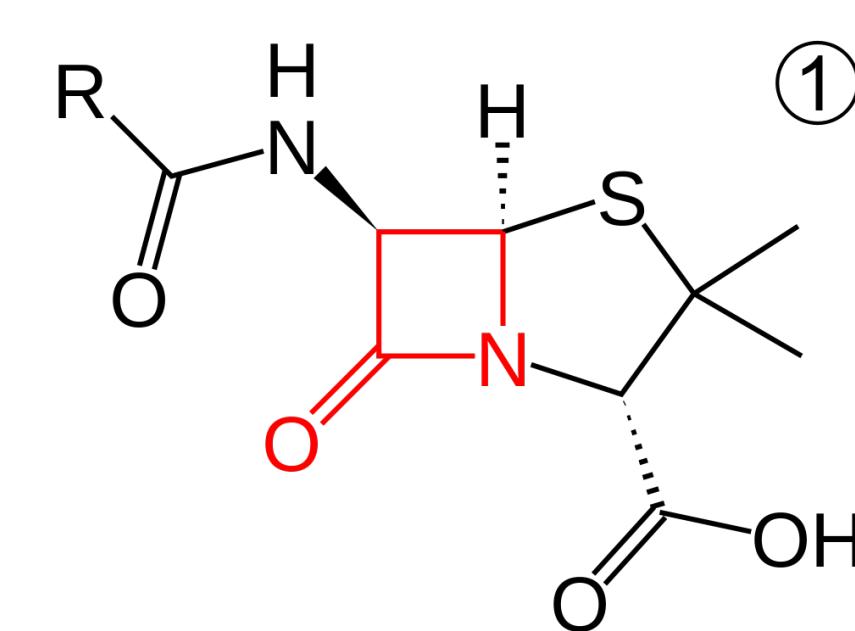


Эволюция и ДНК. Примеры.

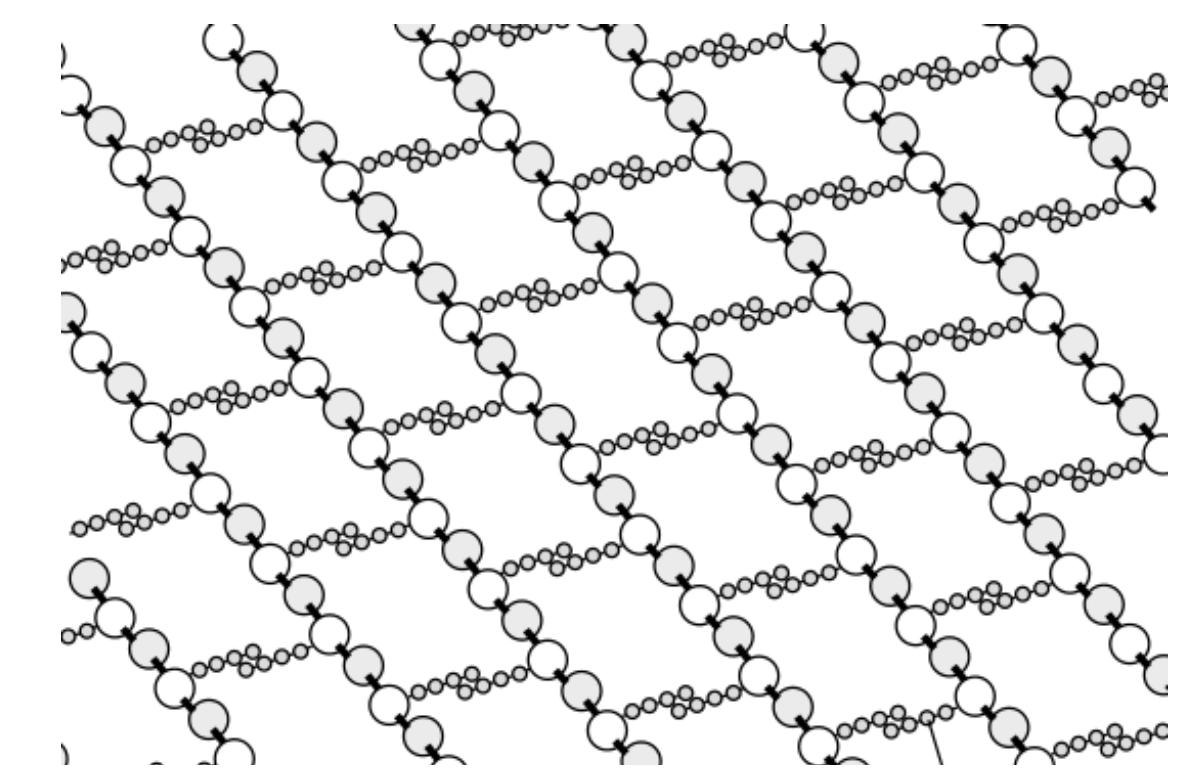
Гемофилия А



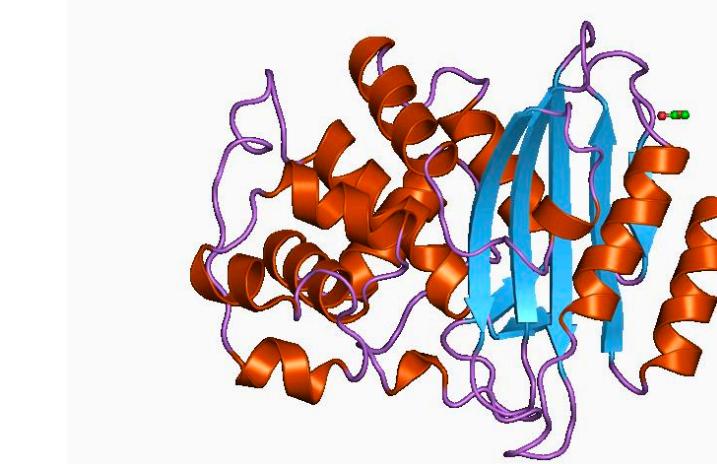
Устойчивость к антибиотикам



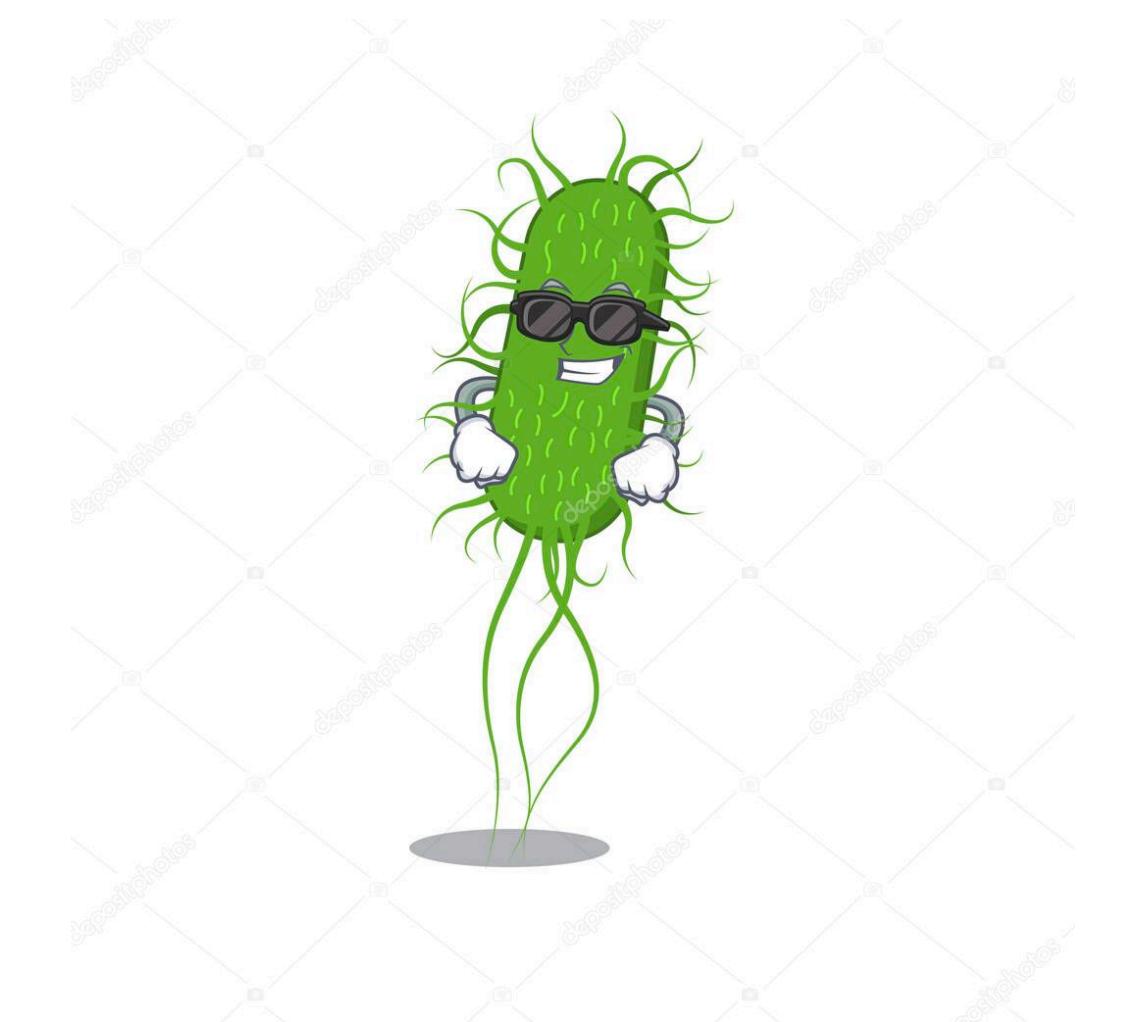
Пеницилин



Пептидогликан

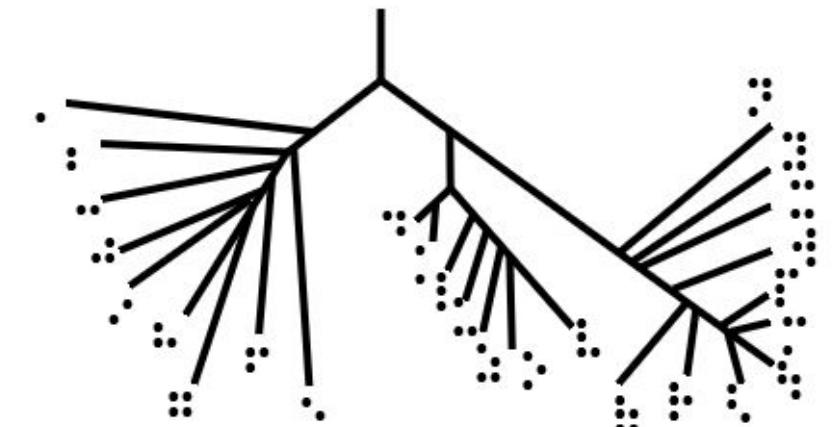
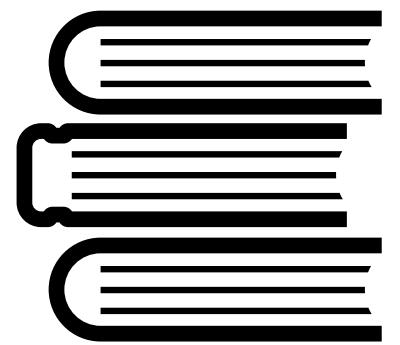


β-лактамазы



Свойства ДНК

- ДНК определяет синтез полипептидной цепи из аминокислот.
Однаковый ген -> “одинаковый” протеин.
- Хранится внутри ядра клетки.
- Может изменяться и сохранять изменения



Сравнение участков ДНК

Если бы происходили только мутации, то как сравнивать 2 последовательности?

GATTACA

CATTTAGA

Сравнение участков ДНК

Если бы происходили только мутации, то как сравнивать 2 последовательности?

GATTACA

CATTA~~G~~A

Сравнение участков ДНК.

Расстояние Хэмминга

Если бы происходили только мутации, то как сравнивать 2 последовательности?

Посчитаем расстояние Хэмминга

```
a, b - входные строки  
distance = 0  
for i = 1 to a.length  
    if a[i] != b[i]:  
        distance += 1
```

Сравнение участков ДНК

Расстояние Хэмминга

Расстояние Хэмминга, какая сложность?

По времени:

По памяти:

Сравнение участков ДНК

Расстояние Хэмминга

Расстояние Хэмминга, какая сложность?

По времени:

$O(n)$

По памяти:

Сравнение участков ДНК

Расстояние Хэмминга

Расстояние Хэмминга, какая сложность?

По времени:

$O(n)$

По памяти:

$O(1)$

Сравнение участков ДНК

Если бы происходили только **мутации, делеции и инсерции** то как тогда сравнивать 2 последовательности?

GATTACA

AAGAGTAC

Сравнение участков ДНК

Расстояние редактирования

Если бы происходили только **мутации, делеции и инсерции** то как тогда сравнивать 2 последовательности?

__GATTACA
AAGAGTAC_

Просто посчитаем минимальное количество таких операций
(Расстояние редактирования)

Сравнение участков ДНК

Расстояние редактирования

Перебор?

Сравнение участков ДНК

Расстояние редактирования

Динамика!

Сравнение участков ДНК

Расстояние редактирования

1. Сколько существует подзадач для данной задачи?

Сравнение участков ДНК

Расстояние редактирования

1. Сколько существует подзадач для данной задачи?
 $n \times m$ подзадач выравниваний $a[1,i]$ с $b[1,j]$

Сравнение участков ДНК

Расстояние редактирования

1. Сколько существует подзадач для данной задачи?
 $n \times m$ подзадач выравниваний $a[1,i]$ с $b[1,j]$
2. Какая из подзадач является решением задачи?

Сравнение участков ДНК

Расстояние редактирования

1. Сколько существует подзадач для данной задачи?
 $n \times m$ подзадач выравниваний $a[1,i]$ с $b[1,j]$
2. Какая из подзадач является решением задачи?
 $a[1,n]$ с $b[1,m]$

Сравнение участков ДНК

Расстояние редактирования

1. Сколько существует подзадач для данной задачи?
 $n \times m$ подзадач выравниваний $a[1,i]$ с $b[1,j]$
2. Какая из подзадач является решением задачи?
 $a[1,n]$ с $b[1,m]$
3. Можно ли решить подзадачу, пользуясь решением более мелких подзадач?
Да, давайте разберемся, как.

Сравнение участков ДНК

Расстояние редактирования

Допустим мы знаем расстояние для пар

1. $(a[1,i-1], b[1,j-1])$
2. $(a[1,i-1], b[1,j])$
3. $(a[1,i], b[1,j-1])$

Как найти расстояние между $(a[1,i], b[1,j])$?

Сравнение участков ДНК

Расстояние редактирования

Допустим мы знаем расстояние для пар

1. $(a[1,i-1], b[1,j-1])$
2. $(a[1,i-1], b[1,j])$
3. $(a[1,i], b[1,j-1])$

Как найти расстояние между $(a[1,i], b[1,j])$?

Просто выбрать из какого состояния было бы оптимальнее
прийти в данное!

1. Добавим по 1 символу
2. Вставим символ
3. Удалим символ

Сравнение участков ДНК

Расстояние редактирования

	G	A	T	T	A	C	A
0	1						
A							
A							
G							
A							
G							
T							
A							
C							

Сравнение участков ДНК

Расстояние редактирования

	G	A	T	T	A	C	A
0	1	2	3	4	5	6	7
A	1						
A	2						
G	3						
A	4						
G	5						
T	6						
A	7						
C	8						

Сравнение участков ДНК

Расстояние редактирования

	G	A	T	T	A	C	A
0	1	2	3	4	5	6	7
A	1	1					
A	2						
G	3						
A	4						
G	5						
T	6						
A	7						
C	8						

Сравнение участков ДНК

Расстояние редактирования

	G	A	T	T	A	C	A
0	1	2	3	4	5	6	7
A	1	1	1	2	3	4	5
A	2	2	1	2	3	3	4
G	3	2	2	2	3	4	4
A	4	3	2	3	3	3	4
G	5	4	3	3	4	4	4
T	6	5	4	3	3	4	5
A	7	6	5	4	4	3	4
C	8	7	6	5	5	4	3

Сравнение участков ДНК

Расстояние редактирования

Замена
Удаление
Вставка

	G	A	T	T	A	C	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
A	2	2	1	2	3	3	4
G	3	2	2	2	3	4	4
A	4	3	2	3	3	3	4
G	5	4	3	3	4	4	4
T	6	5	4	3	3	4	5
A	7	6	5	4	4	3	4
C	8	7	6	5	5	4	3

CA
C

Сравнение участков ДНК

Расстояние редактирования

Замена
Удаление
Вставка

		G	A	T	T	A	C	A
	0	1	2	3	4	5	6	7
A	1	1	1	2	3	4	5	6
A	2	2	1	2	3	3	4	5
G	3	2	2	2	3	4	4	5
A	4	3	2	3	3	3	4	4
G	5	4	3	3	4	4	4	5
T	6	5	4	3	3	4	5	5
A	7	6	5	4	4	3	4	5
C	8	7	6	5	5	4	3	4

— GATTACA
AA **GAGTAC** —

Сравнение участков ДНК

Расстояние редактирования

```
def levensteinIn(a, b):
    D[0][0] = 0
    for j = 1 to n:
        D[0][j] = D[0][j-1] + 1
    for i = 1 to m:
        D[i][0] = D[i-1][0] + 1
    for i = 1 to m:
        for j = 1 to n:
            if a[i] != b[j]:
                D[i][j] = min(D[i-1][j] + 1, D[i][j-1] + 1, D[i-1][j-1] + 1)
            else:
                D[i][j] = D[i-1][j-1]
    return D[m][n]
```

Сравнение участков ДНК

Расстояние редактирования

Какая сложность?

По времени:

По памяти:

Сравнение участков ДНК

Расстояние редактирования

Какая сложность?

По времени:

$O(n \times m)$

По памяти:

Сравнение участков ДНК

Расстояние редактирования

Какая сложность?

По времени:

$O(n \times m)$

По памяти:

$O(n \times m)$

Сравнение участков ДНК

Расстояние редактирования

Какая сложность?

По времени:

$O(n \times m)$

По памяти:

$O(n \times m)$

По памяти можно оптимальнее (алгоритм Хиршберга)

Резюмируем

- ДНК кодирует белки!
- Благодаря свойству локальности ДНК, имеет смысл сравнивать участков генома.
- Сравнивать участки генома можно достаточно эффективно.
- Можно считать расстояние между строками и делать выводы о свойствах организмов.

Формальности

Оценка:

дз **S** баллов

Экзамен **E**

дополнительно можно будет набрать **R** баллов решая задачи с Rosalind.

За каждый пункт можно будет получить от 0 до 100 баллов а итоговая оценка будет формироваться так:

$$M = 1/100 * (6 \mathbf{S} + 3 \mathbf{E} + 2 \mathbf{R})$$

Задания

Дз будет появляться в репозитории на GitHub

Теоретические и исследовательские задания в PDF

Практические можно на любом ЯП, но лучше на питоне