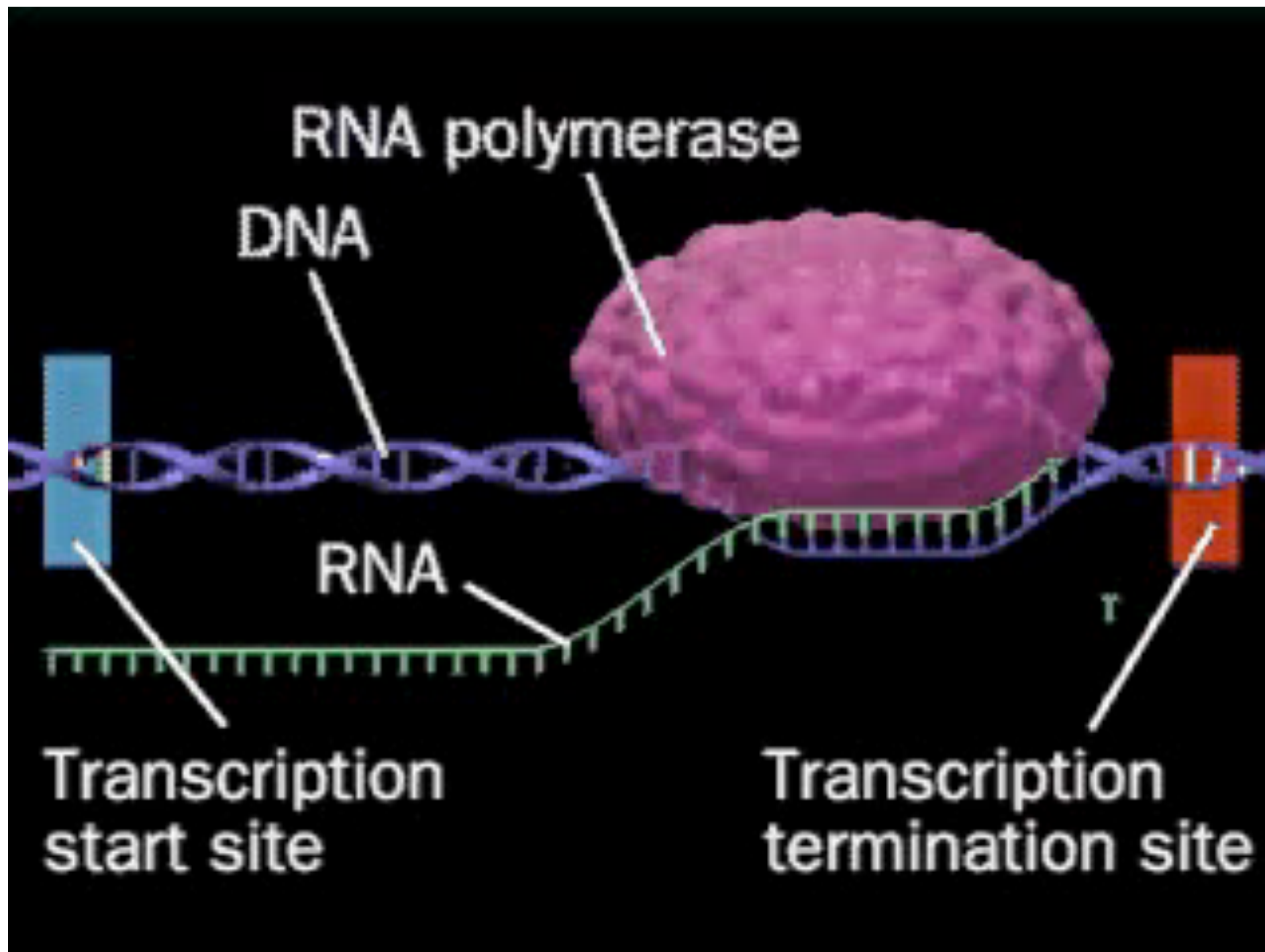
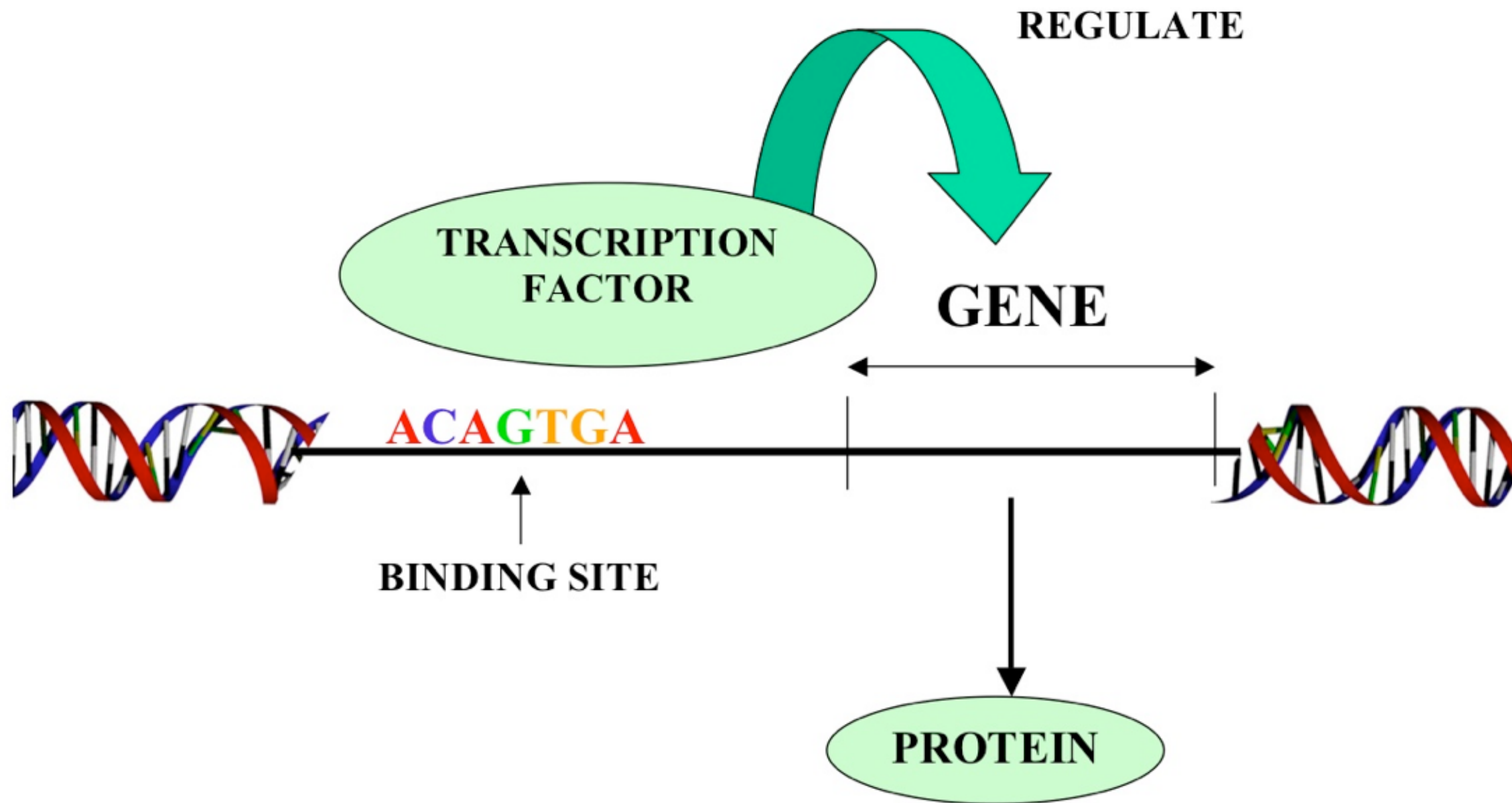


Поиск мотивов ДНК

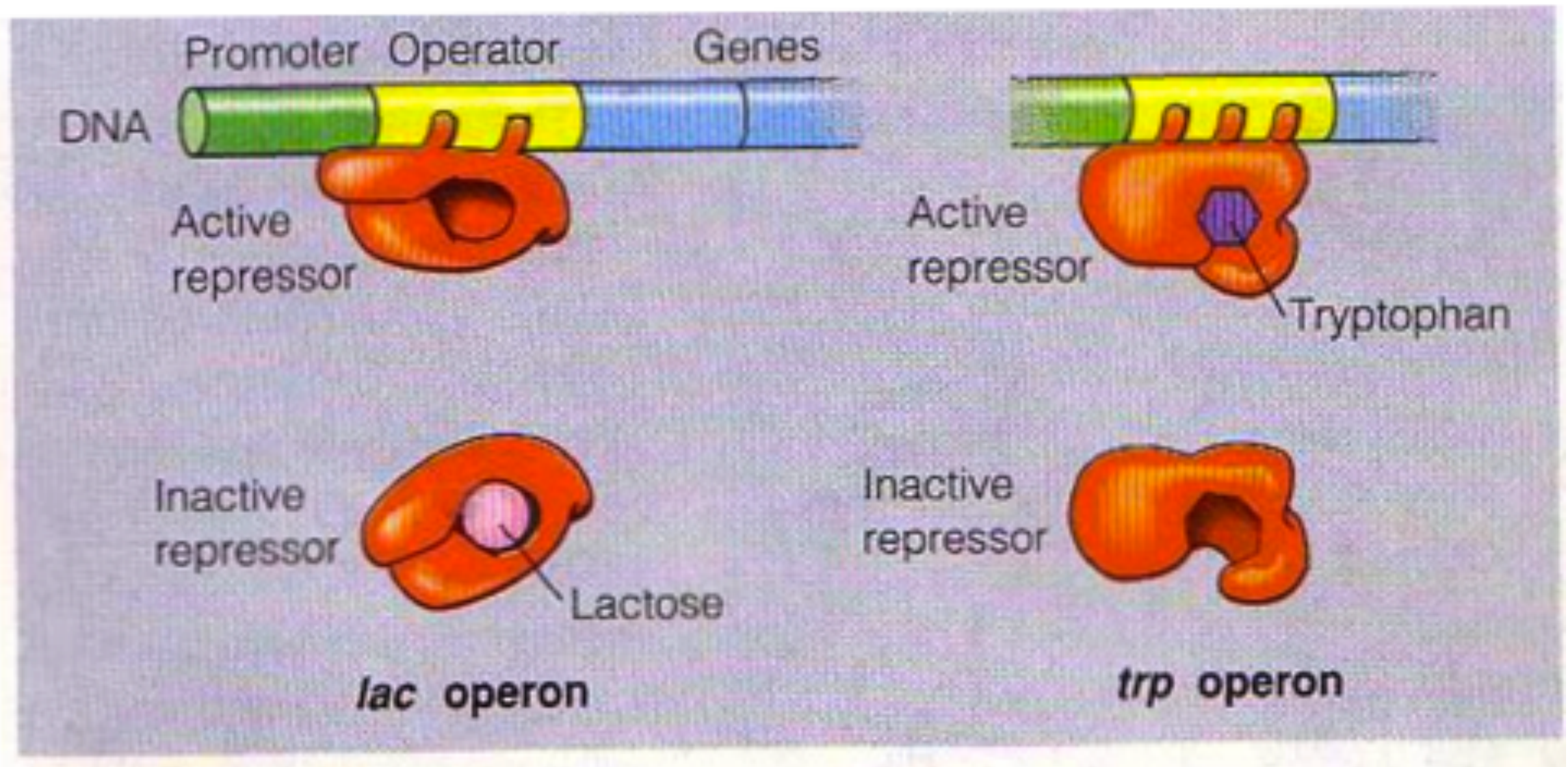
Транскрипция



Транскрипционная регуляция

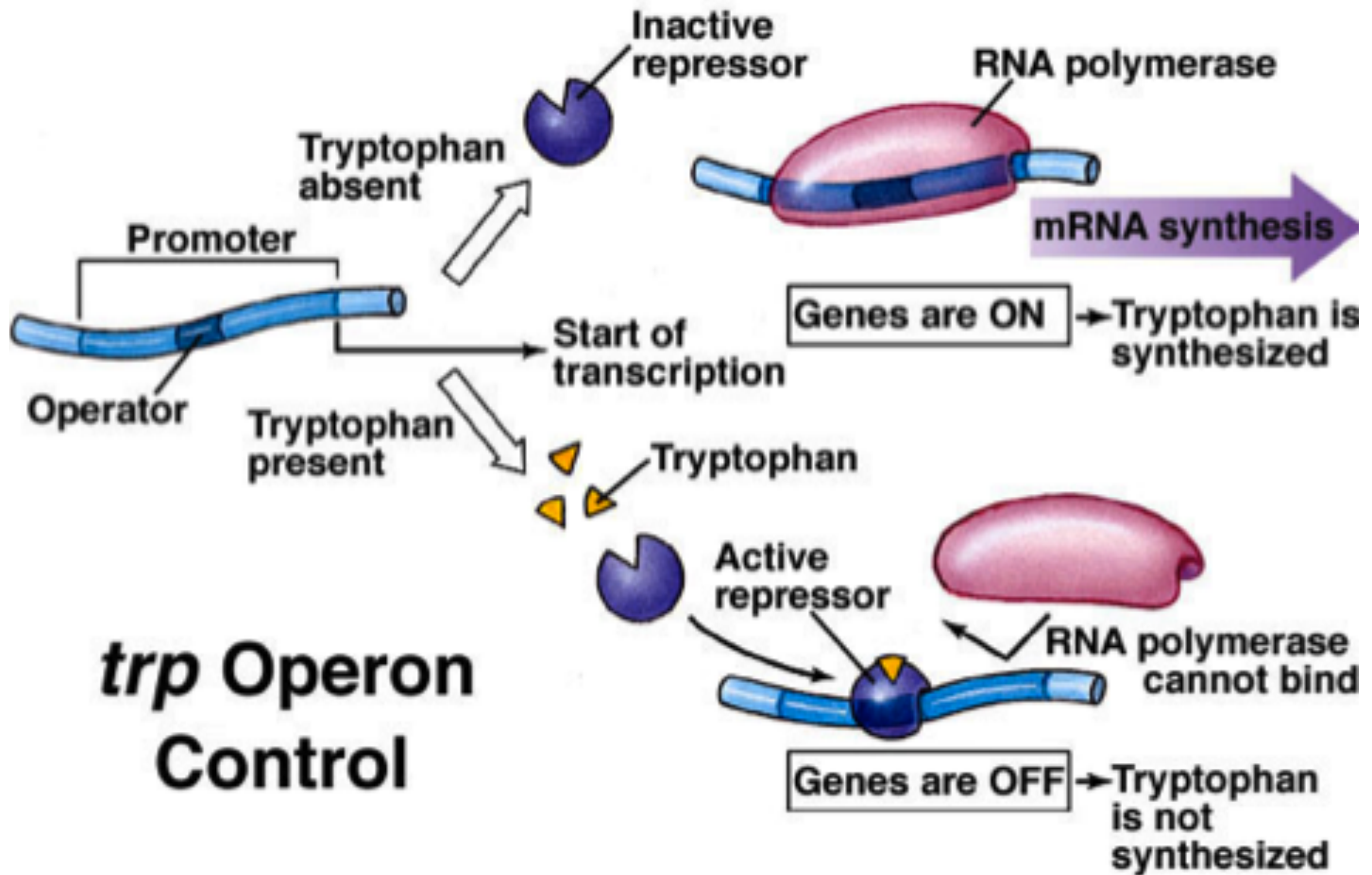


Факторы транскрипции: связывание с ДНК

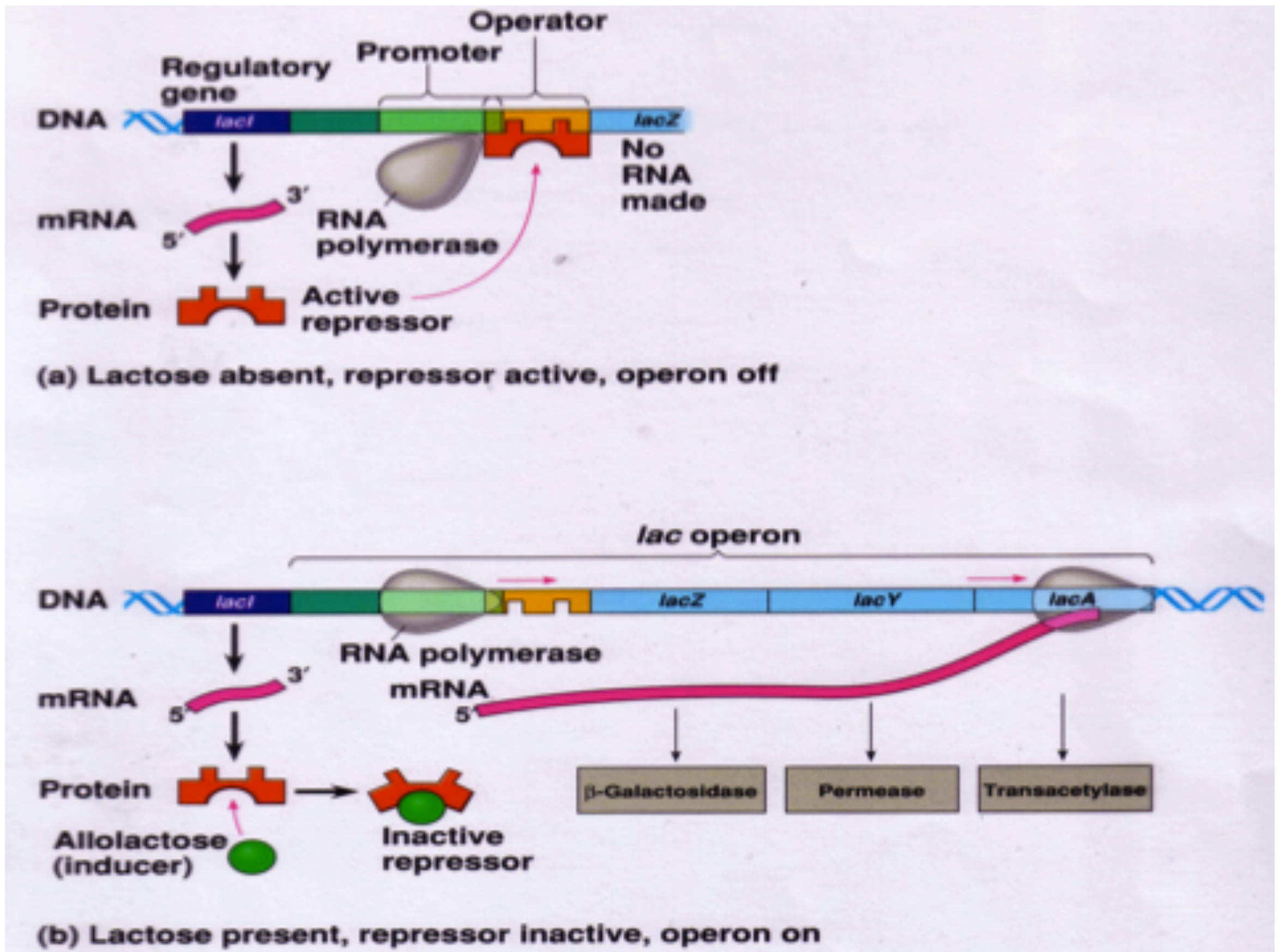


Транскрипционная регуляция: trp оперон

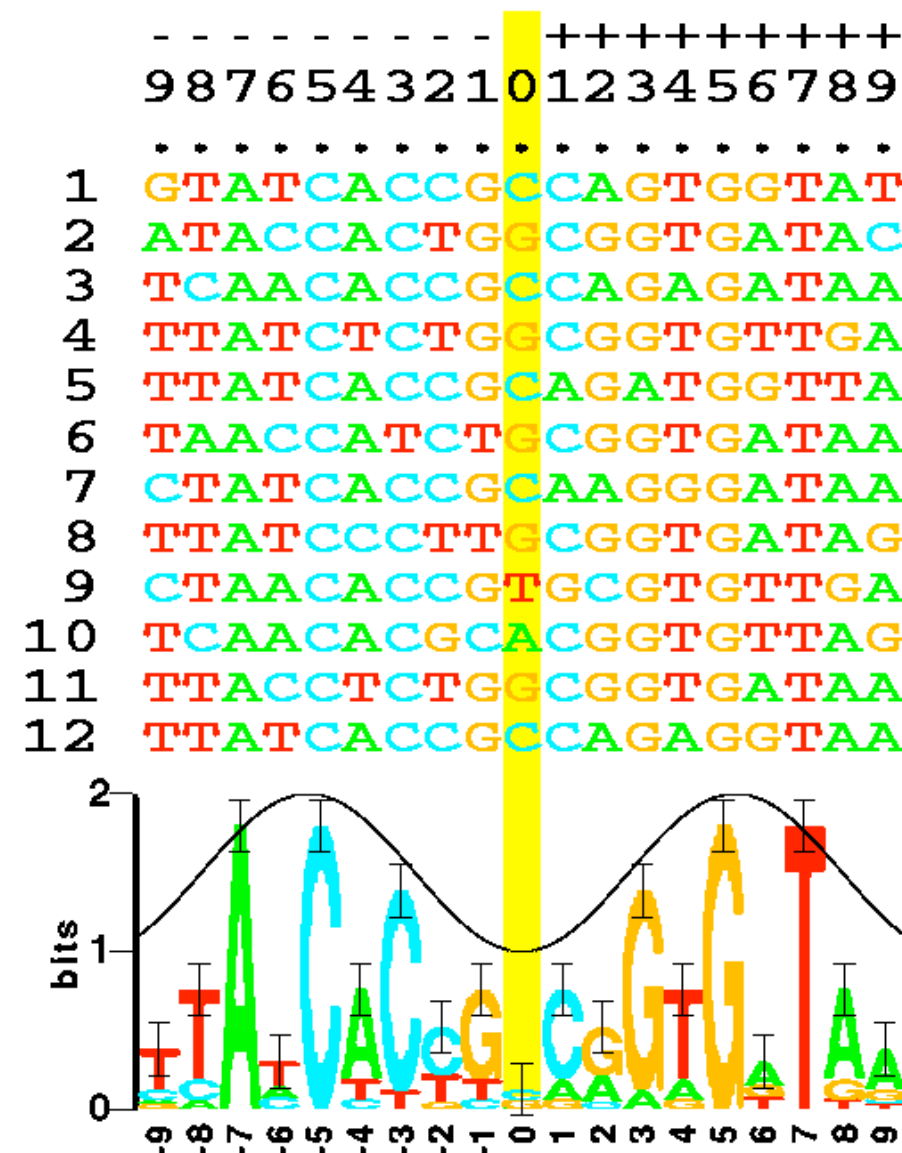
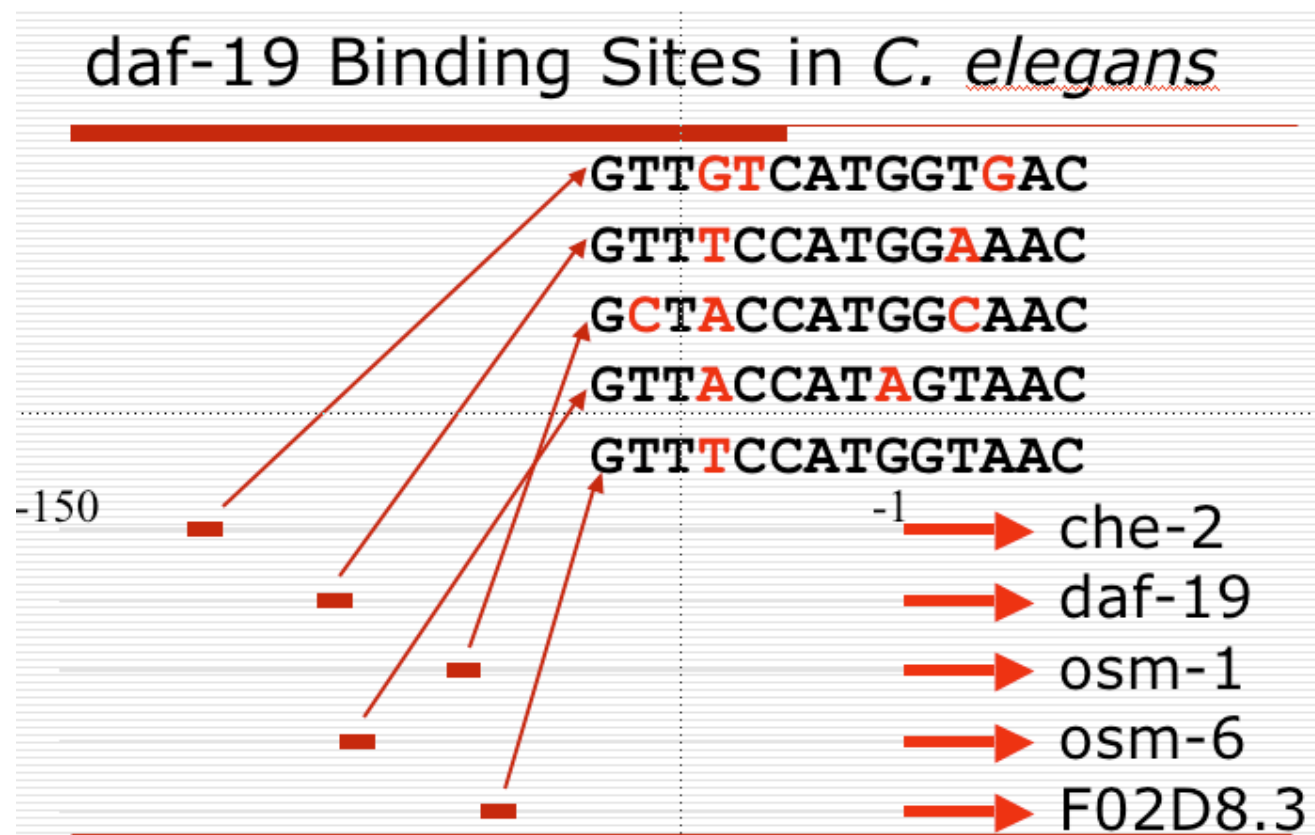
- синтез триптофана



Транскрипционная регуляция: *lac* оперон - утилизация лактозы



Регуляторные мотивы ДНК



Дискретные подходы к поиску мотивов ДНК.

Дан набор последовательностей $S=\{x^1, x^2, \dots, x^n\}$.

Для слова W (мотива) длины L и набора последовательностей S определена функция оценки (scoring function) $d(W, S)$, например:

$$d(W, S) = \sum_i d(W, x^i)$$

где $d(W, x^i)$ - минимальное расстояние Хэмминга для всех слов длины L из последовательности x^i .

Дискретные подходы к поиску мотивов ДНК: pattern-driven approach

Основная идея: полный перебор слов длины L
(Pattern-driven algorithm [Pevzner2000]):

Для всевозможных слов W длины L , начиная с $AA..A$ по $TT..T$ (4^L слов):
Вычислить $d(W, S)$

Искомый мотив $W^* = \operatorname{argmin}(d(W, S))$

Вычислительная сложность: $O(L N 4^L)$, где $N = \sum_i |x^i|$

Преимущества: метод находит наилучший мотив (глобальный экстремум для заданной функции оценки)

Недостатки: время выполнения

Дискретные подходы к поиску мотивов ДНК: sample-driven approach

Основная идея: поиск мотивов на основе слов, представленных в последовательностях (sample-driven algorithm [Waterman 1985]):

Для всевозможных слов W длины L встречающихся в S :
Вычислить $d(W, S)$

Искомый мотив $W^* = \operatorname{argmin}(d(W, S))$

Вычислительная сложность: $O(L N^2)$

Преимущества: время выполнения

Недостатки: метод может не находить лучшие мотивы

Дискретные подходы к поиску мотивов ДНК: extended sample-driven approach

Основная идея: поиск мотивов на основе слов, представленных в последовательностях и близких к ним слов (extended sample-driven algorithm [Galas 1985]):

Определение: α -окрестностью слова W называются набор всевозможных слов W' , таких что $d(W, W') \leq \alpha$

Для всевозможных слов W' из α -окрестностей слов W , встречающихся в S :
Вычислить $d(W', S)$

Искомый мотив $W^* = \operatorname{argmin}(d(W', S))$

Дискретные подходы к поиску мотивов ДНК: CONSENSUS - пример “жадного” алгоритма поиска

Алгоритм CONSENSUS [Stormo 1999]:

Цикл №1:

Для каждого слова W из S (слово фиксированной длины L)

Для каждого слова W' из S

Построить выравнивание W и W'

Для следующего шага оставляем C_1 наилучших выравниваний, $A_1, \dots, A_{C_1}$

ACGGTTG , CGAACTT , GGGCTCT ...
ACGCCTG , AGAACTA , GGGGTGT ...

Дискретные подходы к поиску мотивов ДНК: CONSENSUS - пример “жадного” алгоритма поиска

Цикл t :

Для каждого слова W из S

Для каждого выравнивания A_j из цикла $t-1$

Построить выравнивание W и A_j

Для следующего шага оставляем C_t наилучших выравниваний, $A_1, \dots, A_{C_t}$

ACGGTTG , CGAACTT , GGGCTCT ...
ACGCCTG , AGAACTA , GGGGTGT ...
...
ACGGCTC , AGATCTT , GGCGTCT ...

$C_1, \dots, C_n$ - константы задаваемые в качестве параметров алгоритма

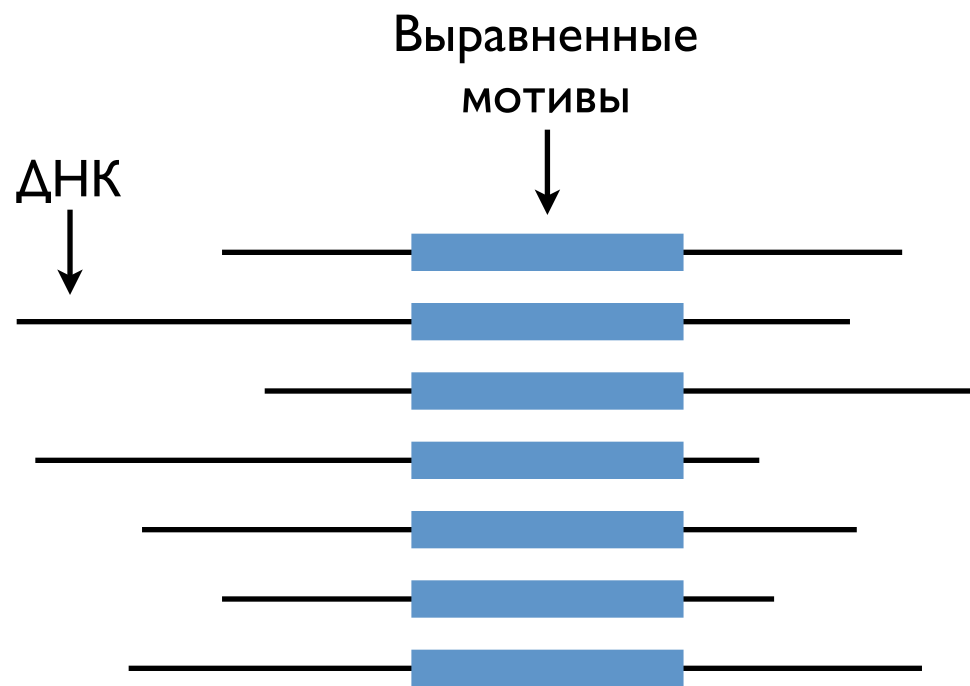
Вычислительная сложность: $O(N^2) + O(N C_1) + O(N C_2) + \dots + O(N C_n) = O(N^2 + N C_{\text{total}})$

где $C_{\text{total}} = \sum_i C_i$

Алгоритм максимизации ожидания для поиска мотивов ДНК

- В основе алгоритма максимизации ожидания лежит итеративная сходящаяся процедура оценки вероятностной модели мотива и позиций сайтов на ДНК.

Имея выравнивание мотивов, можно построить вероятностную модель



Вероятностная модель мотива -
позиционная весовая матрица

	1	2	3	4	5	6	7
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3



Имея вероятностную модель, можно найти вероятные позиции сайтов на ДНК

Алгоритм максимизации ожидания (Expectation Maximization): общая схема

Дано: длина сайта W , набор последовательностей ДНК.

Установим начальные значения для параметров мотива.

В цикле:

1. Используя вероятностную модель мотива оценим позиции сайтов на ДНК (E-шаг).

2. Используя найденные позиции сайтов обновим параметры мотива (M-шаг).

Остановим процесс по достижению сходимости.

Результат: мотив, позиции сайтов на ДНК.

Максимизация ожидания: вероятностные представления мотива и фона

- Вероятностная модель мотива P длины W - матрица вероятностей p_{ck} размерности $4 \times W$, где p_{ck} - вероятность наблюдения остатка c в позиции мотива k .

$P =$

	1	2	3
A	0.1	0.5	0.2
C	0.4	0.2	0.1
G	0.3	0.1	0.6
T	0.2	0.2	0.1

$\sim \text{CAG}$

- Вероятностная модель P_0 фонового распределения остатков.

$P_0 =$

	0
A	0.26
C	0.24
G	0.23
T	0.27

$\sim$ равномерное распределение

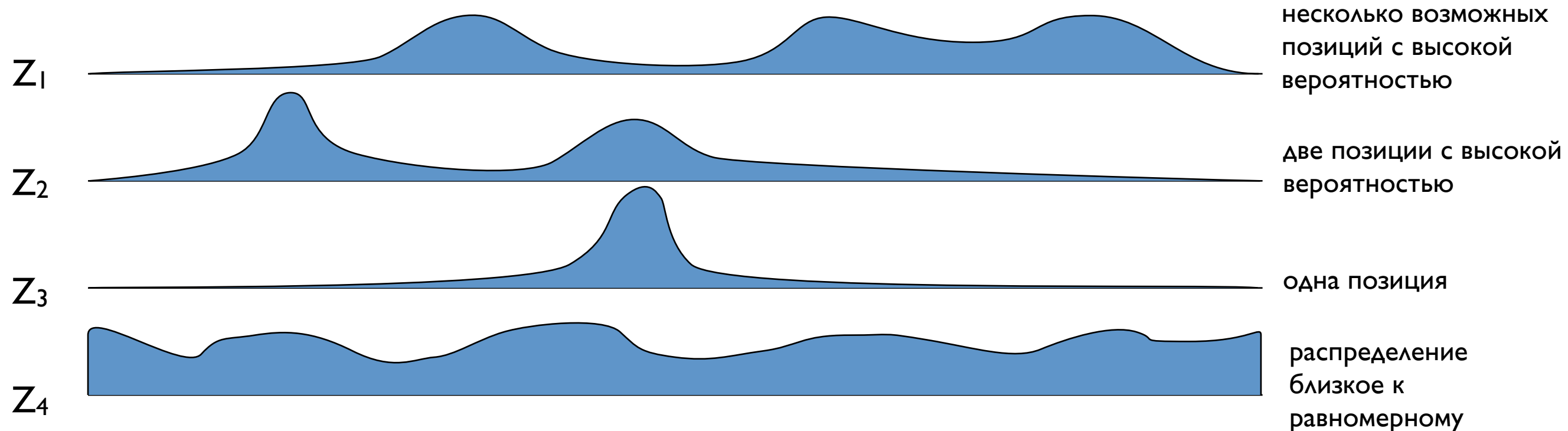
Максимизация ожидания: вероятности позиций сайтов

- Элементы Z_{ij} матрицы Z представляют собой вероятности нахождения сайта в позиции j последовательности i .

$Z =$

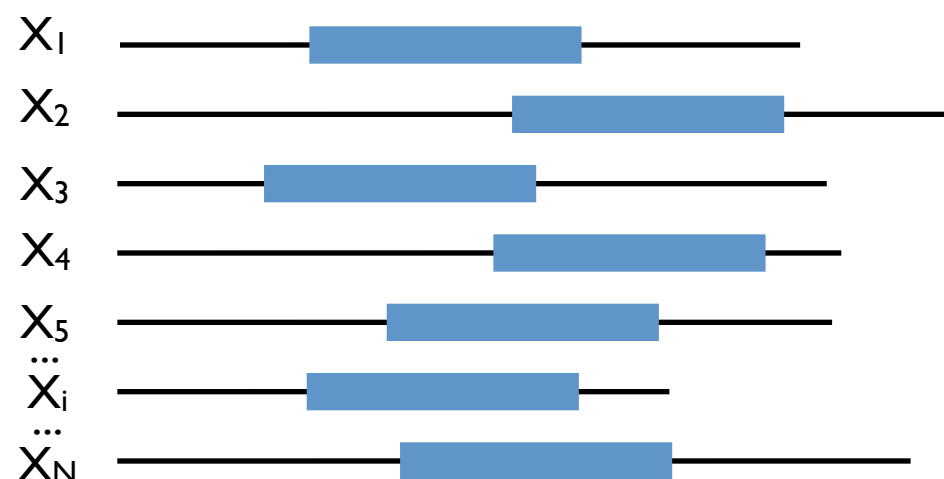
	1	2	3	4
Последовательность 1	0.1	0.1	0.2	0.6
Последовательность 2	0.4	0.2	0.1	0.3
Последовательность 3	0.3	0.1	0.5	0.1
Последовательность 4	0.1	0.5	0.1	0.3

- Примеры плотности распределения Z



Максимизация ожидания: итеративная процедура

Вероятные положения сайтов в позициях последовательностей: Z_{ij}



М-шаг



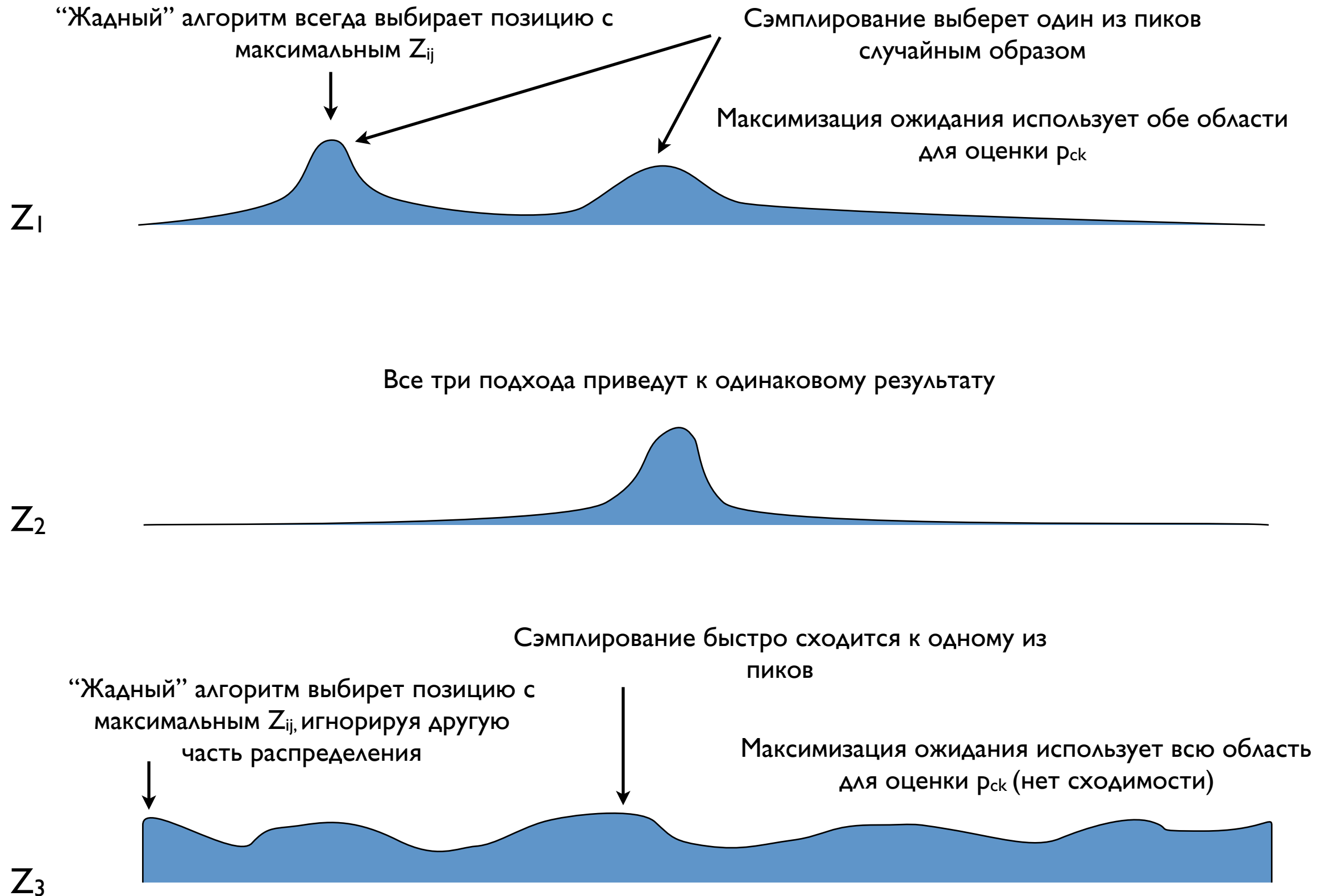
Е-шаг

Вероятности остатков в позициях мотива: p_{ck}

	1	2	3	4	5	6	7
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3

- Вычисление матрицы Z_{ij} по известным p_{ck} :
 - в каждой позиции последовательности вычислим вероятность существования сайта путем перемножения вероятностей, соответствующих наблюдаемым остаткам в позициях.
- Обновление параметров модели мотива p_{ck} на основе вычисленной матрицы Z_{ij} :
 - Максимизация ожидания: частоты остатков взвешиваются с весами Z_{ij} и вычисляются по всем позициям последовательностей.
 - “Жадный” подход: для каждой последовательности i выбирается одна позиция j - позиция с максимальной вероятностью Z_{ij} вдоль последовательности.
 - Сэмплирование: выбирается одна позиция для каждой последовательности согласно распределению Z_{ij} .

Максимизация ожидания, “жадный” алгоритм и сэмплирование по Гиббсу на примере различных случаев распределения Z_{ij}



Е-шаг: вычисление Z_{ij} на основе вероятностной модели мотива

- Вычисление вероятности последовательности X_i , в случае известной позиции мотива j .

$$\Pr(X_i \mid Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k, 0}}_{\text{до сайта}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}_{\text{сайт}} \underbrace{\prod_{k=j+W}^L p_{c_k, 0}}_{\text{после сайта}}$$

- Пример:

$X_i = \text{G C } \boxed{\text{T G T}} \text{ A G}$

$P =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$$\begin{aligned} \Pr(X_i \mid Z_{i3}=1, p) &= p_{G,0} \times p_{C,0} \times p_{T,1} \times p_{G,2} \times p_{T,3} \times p_{A,0} \times p_{G,0} = \\ &= 0.25 \times 0.25 \times \boxed{0.2 \times 0.1 \times 0.1} \times 0.25 \times 0.25 \end{aligned}$$

Е-шаг: вычисление Z_{ij} на основе вероятностной модели мотива

- Используя теорему Байеса найдем вероятности положений сайтов:

$$\Pr(Z_{ij} \mid X_i, p) = \frac{\Pr(X_i \mid Z_{ij} = 1, p) P(Z_{ij} = 1)}{\Pr(X_i)}$$

- Таким образом, на шаге t вероятность $Z_{ij}^{(t)}$ вычисляется, используя известные $p^{(t)}$, следующим образом:

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, p) \cancel{P(Z_{ij} = 1)}}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ij} = 1, p) \cancel{P(Z_{ij} = 1)}}$$

априорные вероятности сокращаются при предположении равновероятности нахождения сайта в любой из позиции

Е-шаг: вычисление Z_{ij} - пример

$X_i =$ G C T G T A G

$P =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.3 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

...

далее нормализовать на $\sum_{j=1}^{L-W+1} Z_{ij}$

М-шаг: пересчет вероятностной модели мотива на основе Z_{ij}

$$p_{c,k}^{(t+1)} = \frac{n_{c,k} + d_{c,k}}{\sum_b (n_{b,k} + d_{b,k})}$$

← псевдоотчеты

$$n_{c,k} = \begin{cases} \sum_i \sum_{\{j | X_{i,j+k-1}=c\}} Z_{ij} & k > 0 \\ n_c - \sum_{j=1}^W n_{c,j} & k = 0 \end{cases}$$

МОТИВ

фон

M-шаг: вычисление параметров мотива на основе Z_{ij} - пример

$$X_1 = A \boxed{C \ A \ G} C \ A$$
$$Z_1 = 0.1 \ 0.7 \ 0.1 \ 0.1$$

$$X_2 = A \ G \ G \boxed{C \ A \ G}$$
$$Z_1 = 0.4 \ 0.1 \ 0.1 \ 0.4$$

$$X_3 = T \boxed{C \ A \ G} T \ C$$
$$Z_1 = 0.2 \ 0.6 \ 0.1 \ 0.1$$

$$n_{A,1} = 0.1 + 0.1 + 0.4 + 0.1 = 0.7$$
$$n_{C,1} = 0.7 + 0.4 + 0.6 = 1.7$$
$$n_{G,1} = 0.1 + 0.1 + 0.1 + 0.1 = 0.4$$
$$n_{T,1} = 0.2 = 0.2$$
$$\text{Total: } T = 0.7 + 1.7 + 0.4 + 0.2 = 3.0$$

$$p_{A,1} = (0.7 + 1) / (T + 4) = 1.7 / 7 = 0.24$$
$$p_{C,1} = (1.7 + 1) / (T + 4) = 2.7 / 7 = 0.39$$
$$p_{G,1} = (0.4 + 1) / (T + 4) = 1.4 / 7 = 0.2$$
$$p_{T,1} = (0.2 + 1) / (T + 4) = 1.2 / 7 = 0.17$$

$P_{c,k} =$

	1	2	3
A	0.24	0.39	0.21
C	0.39	0.21	0.18
G	0.2	0.24	0.44
T	0.17	0.16	0.16

Благодарности

- При подготовке слайдов использовались материалы лекций:
 - Михаила Гельфанда (ИППИ)
 - Андрея Миронова (МГУ)
 - Serafim Batzoglou (Stanford)
 - Manolis Kellis (MIT)
 - Pavel Pevzner (UCSD)