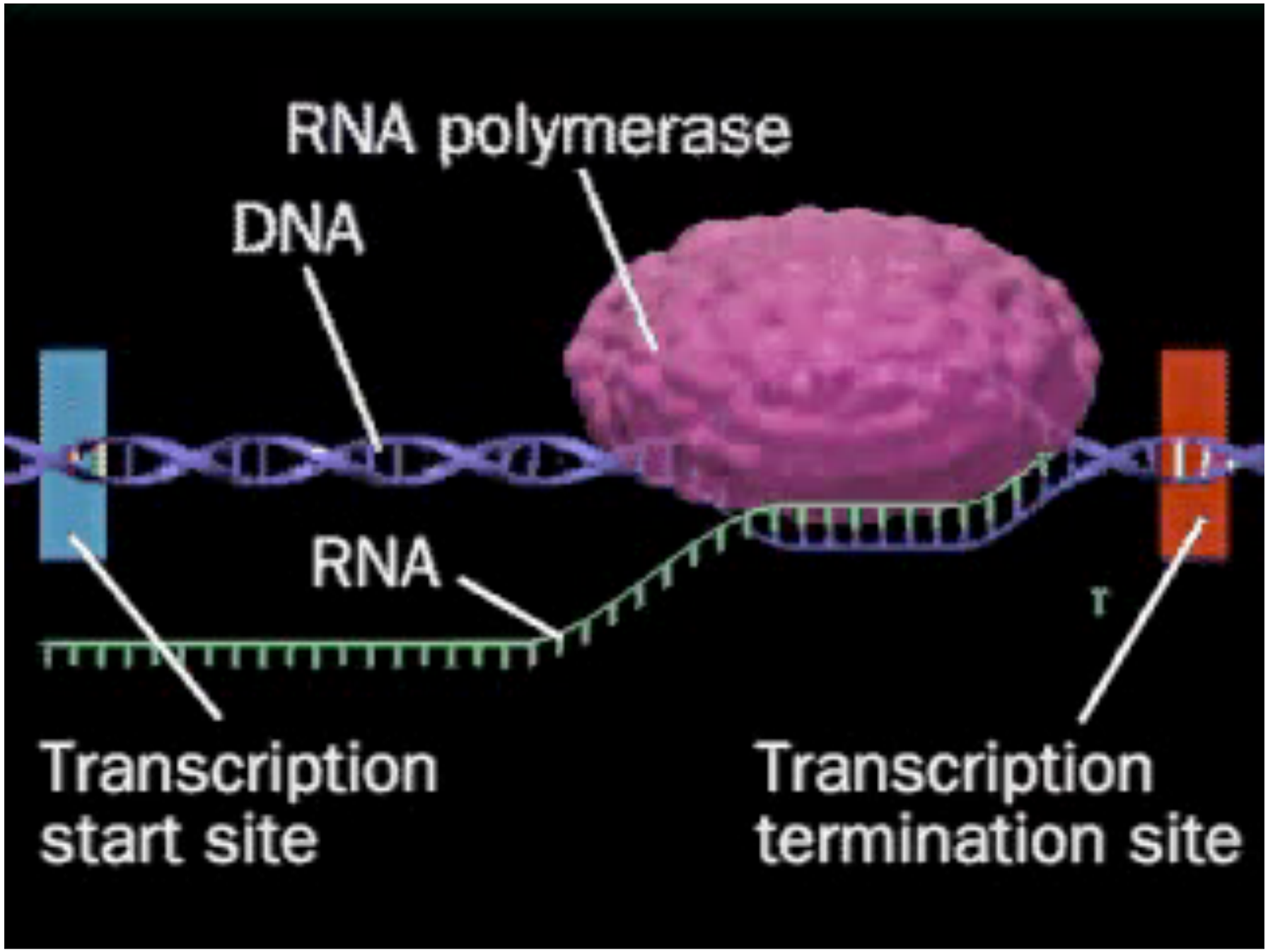
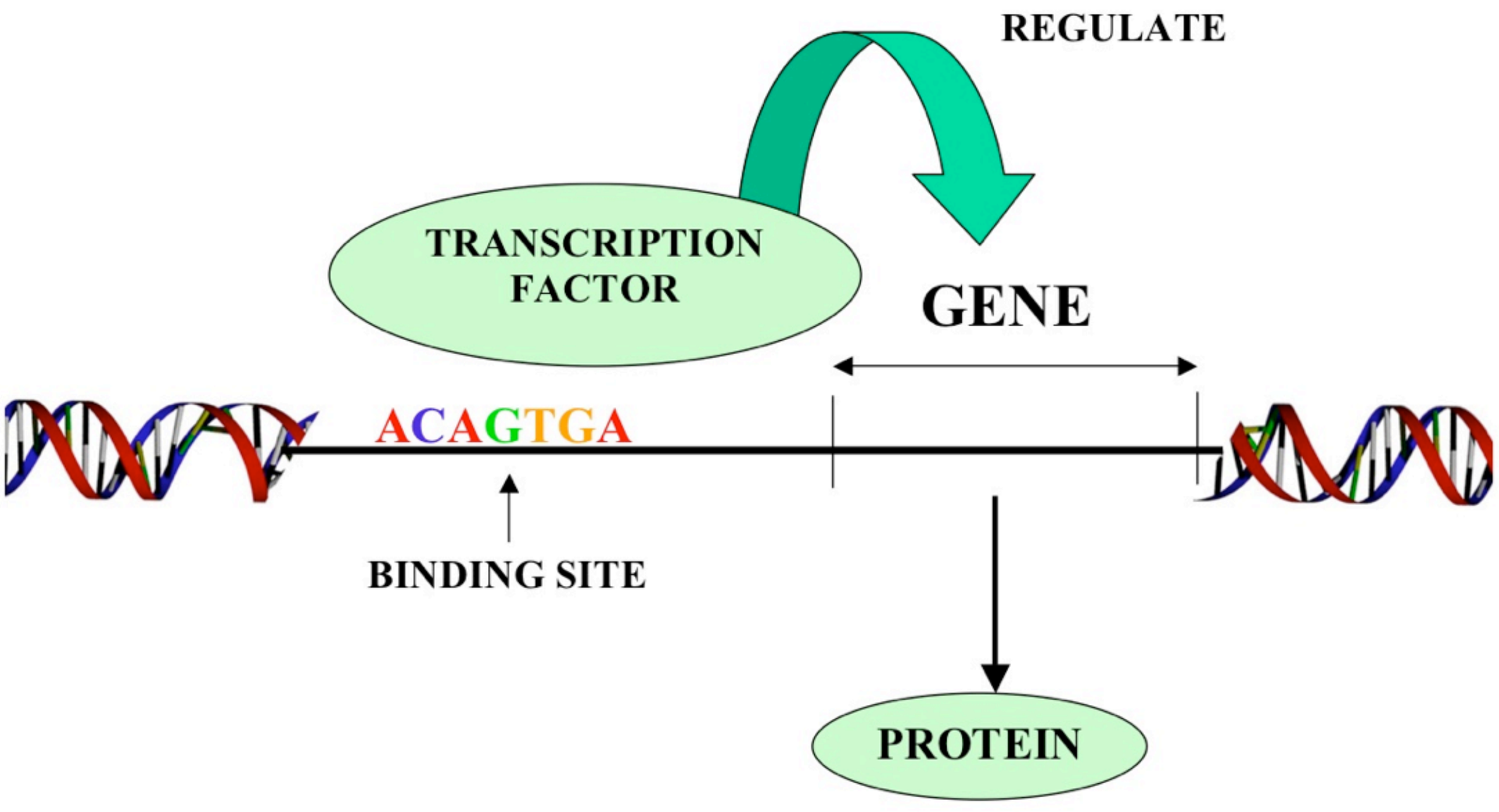
Поиск мотивов ДНК

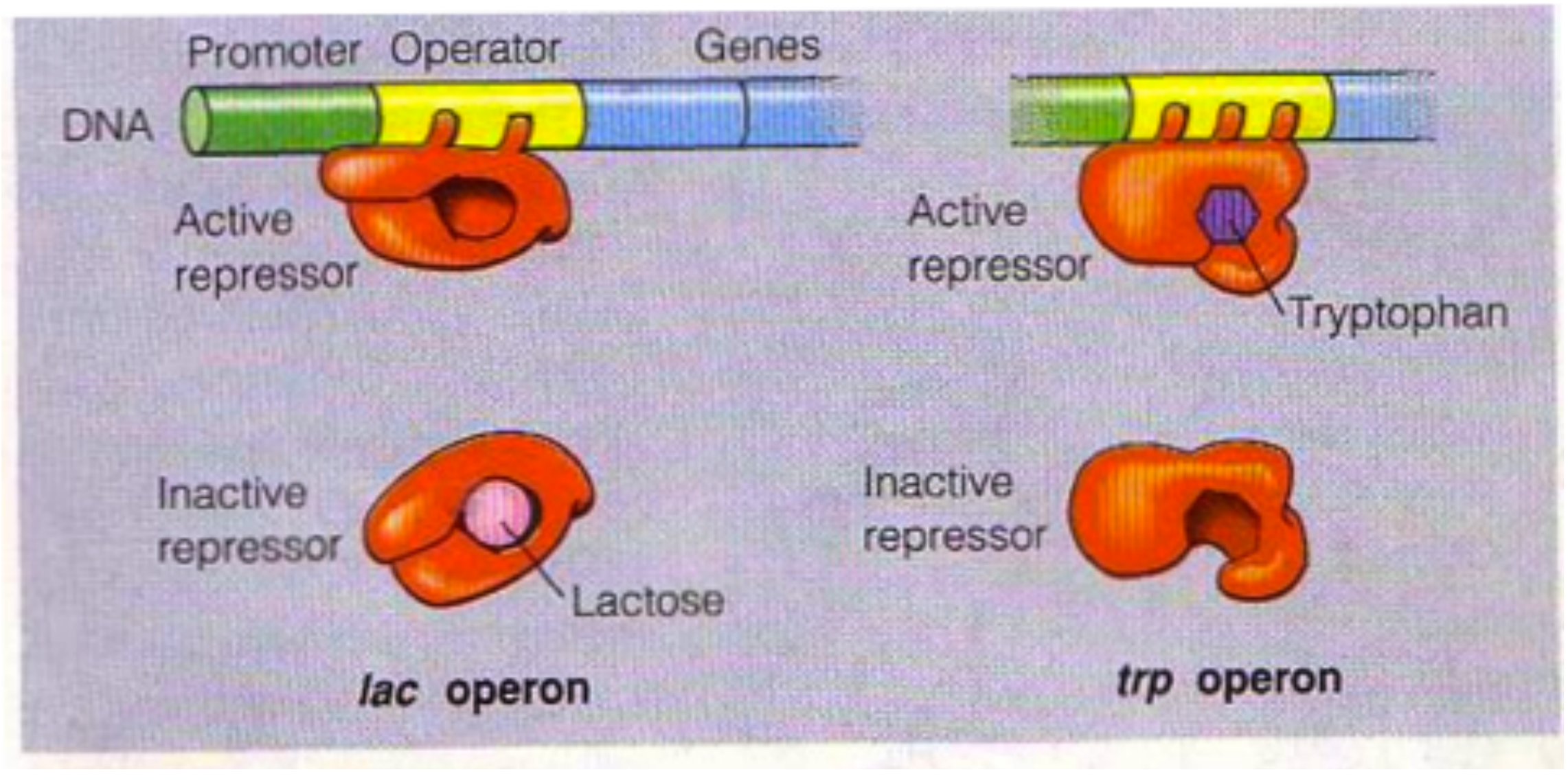
Транскрипция



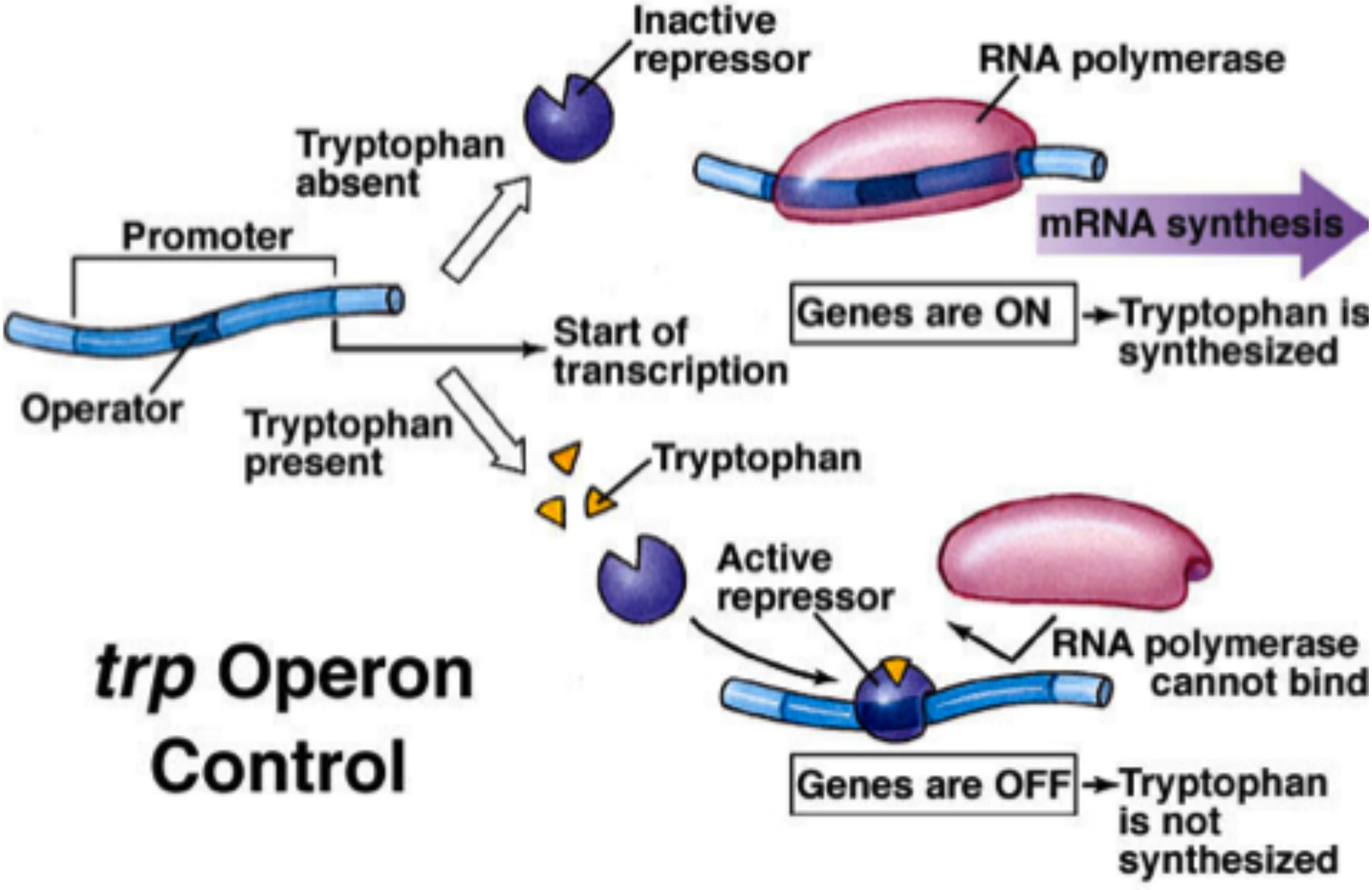
Транскрипционная регуляция



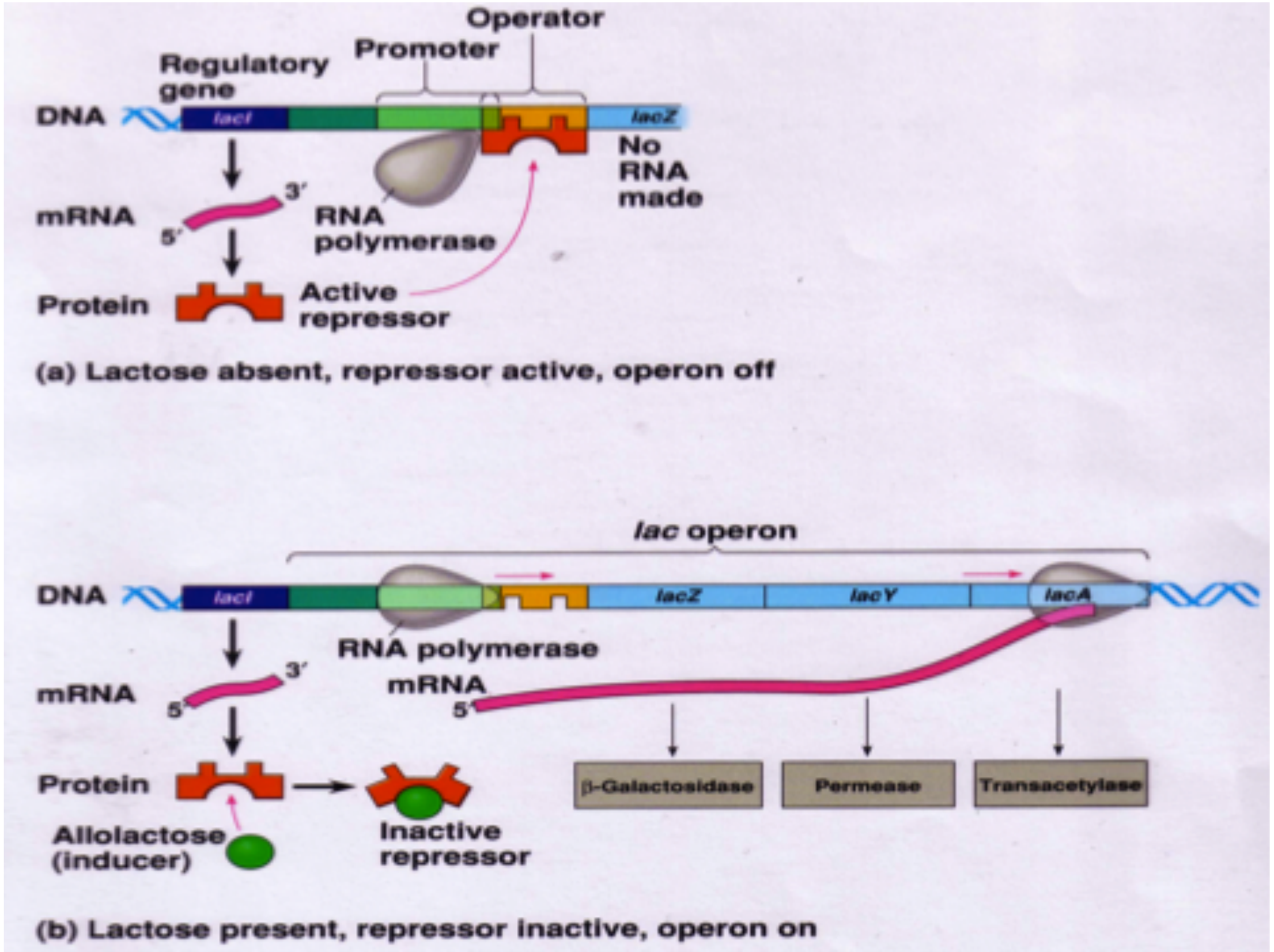
Факторы транскрипции: связывание с ДНК



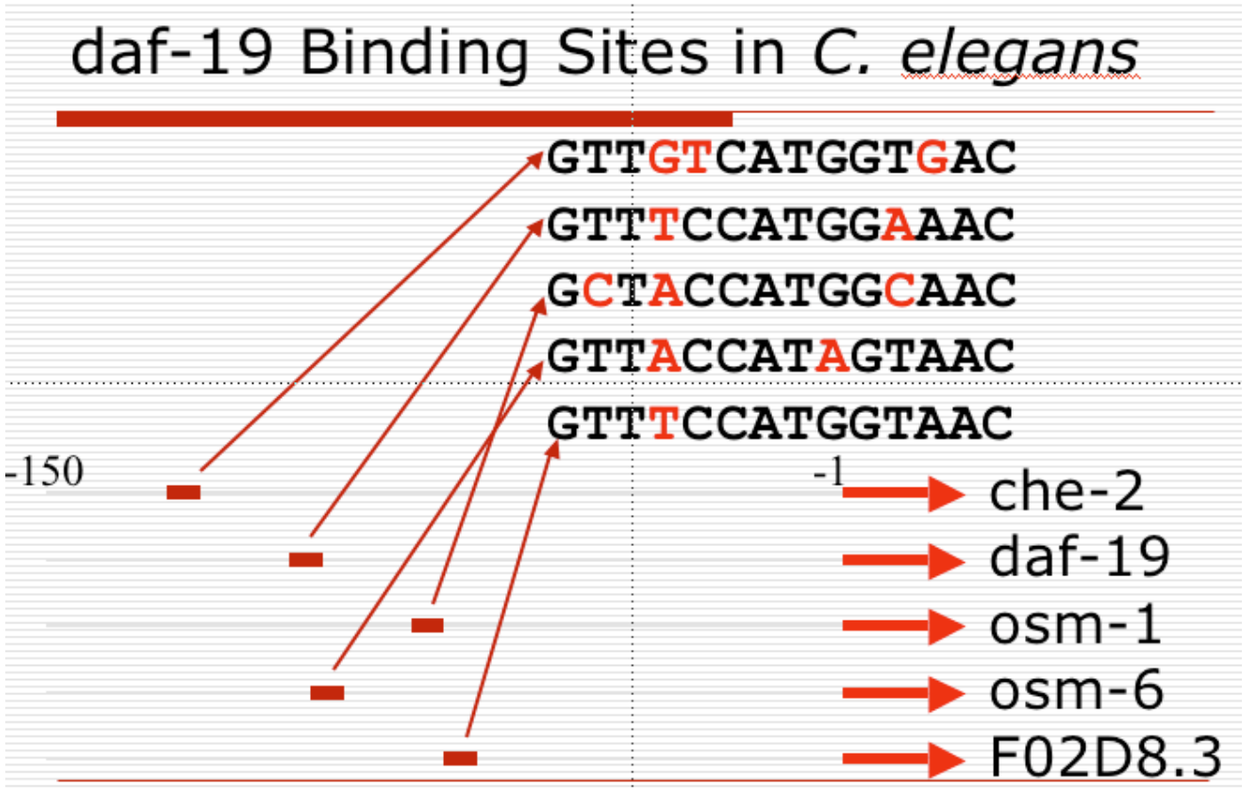
Транскрипционная регуляция: trp оперон - синтез триптофана

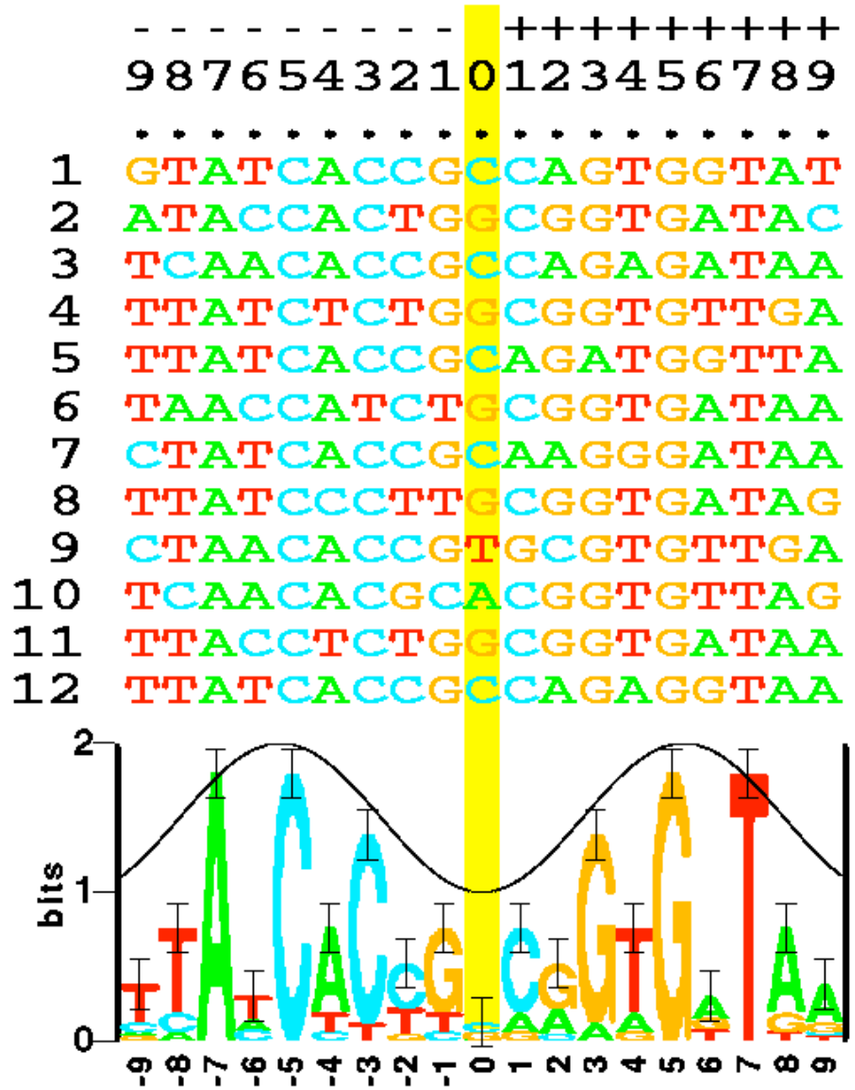


Транскрипционная регуляция: lac оперон - утилизация лактозы



Регуляторные мотивы ДНК





Дискретные подходы к поиску мотивов ДНК.

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

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=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 [Waterman1985]):

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

Искомый мотив W=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 [Galas1985]):

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=argmin(d(W',S))*

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

Алгоритм CONSENSUS [Stormo1999]:

Цикл №1:
Для каждого слова W из S *(слово фиксированной длины L)*
Для каждого слова W’ из S *(в одном выравнивании по одному слову из одной последовательности)*
Построить выравнивание W и W’

Для следующего шага оставляем C_1 наилучших выравниваний, $A_1, \ldots, A_{C1}$

ACGGTTG , CGAACTT , GGGCTCT …
ACGCCTG , AGAACTA , GGGGTGT …

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

Цикл t:

Для каждого выравнивания A_j из цикла t-1

Для каждого слова W из S (из неиспользованных последовательностей)

Построить выравнивание W и A_j

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

ACGGTTG , CGAACTT , GGGCTCT …

ACGCCTG , AGAACTA , GGGGTGT …

… … …

ACGGCTC , AGATCTT , GGCGTCT …

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

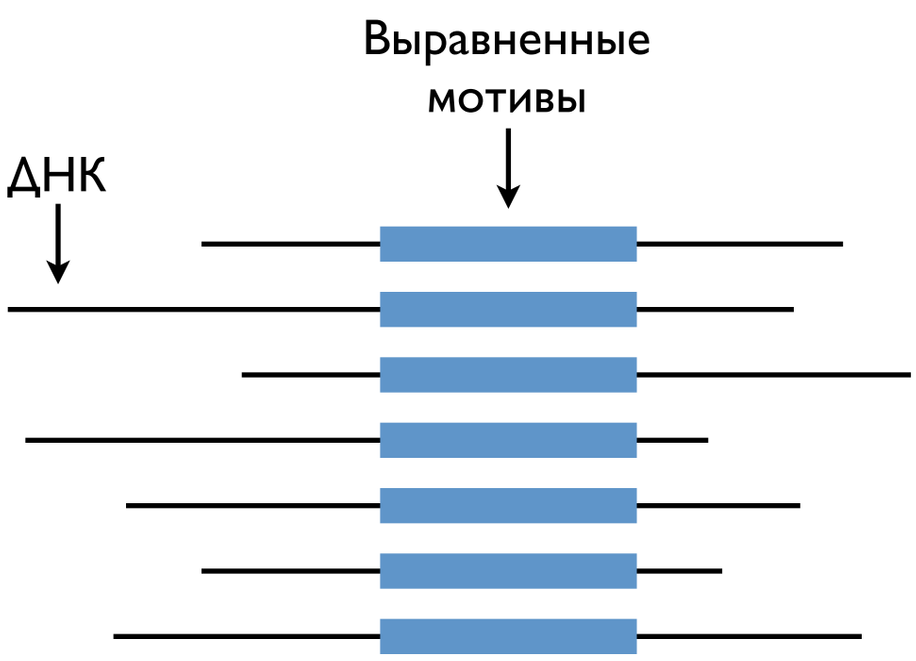
Вычислительная сложность: $O(N^2) + O(N\,C_1) + O(N\,C_2) + \ldots + O(N\,C_n) = O(N^2 + NC_{total})$

где $C_{total} = \sum_i C_i$

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

~ CAG

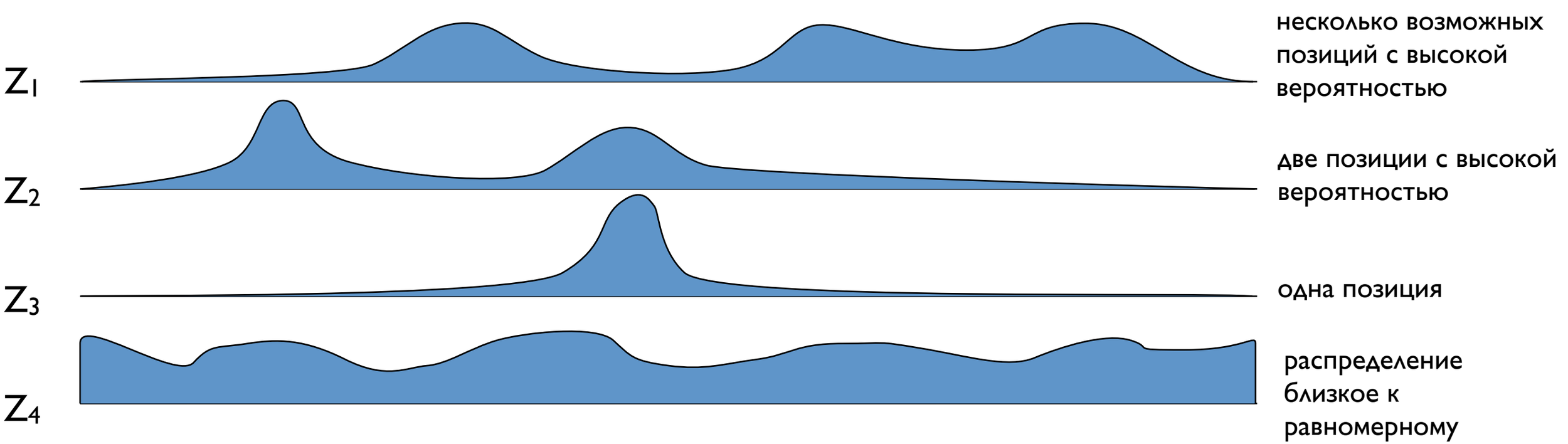
Максимизация ожидания: вероятности позиций сайтов

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

$Z =$

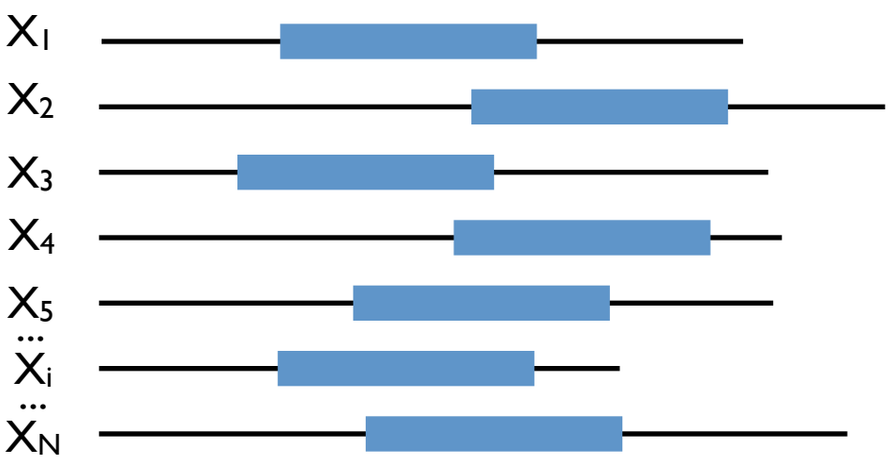
	1	2	3	4
Последовательность 1	0.1	0.1	0.2	0.6
Последовательность 2	0.4	0.2	0.1	0.3
Последовательность 3	0.3	0.1	0.5	0.1
Последовательность 4	0.1	0.5	0.1	0.3

- Примеры плотности распределения Z



Максимизация ожидания: итеративная процедура

Вероятные положения сайтов в позициях последовательностей: Z_{ij}



M-шаг

E-шаг

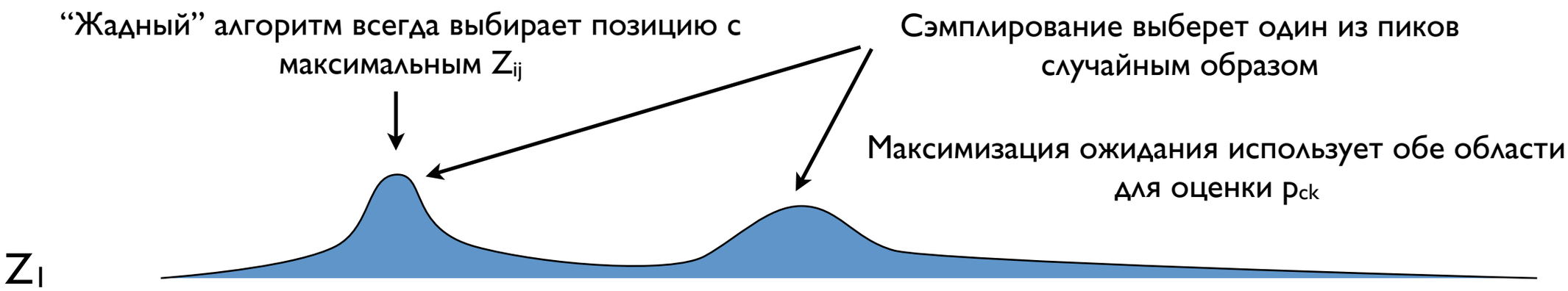
Вероятности остатков в позициях мотива: p_{ck}

	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

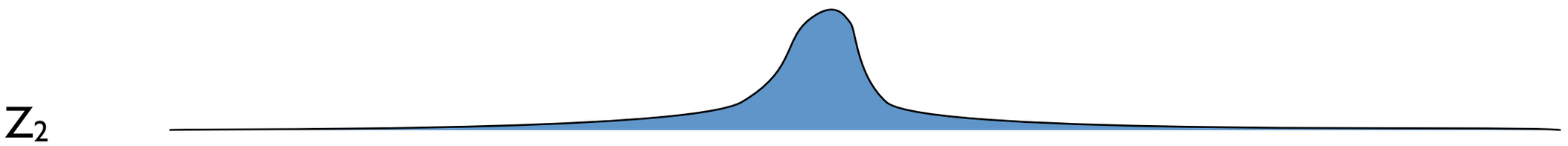
- Вычисление матрицы Z_{ij} по известным p_{ck}:
  - в каждой позиции последовательности вычислим вероятность существования сайта путем перемножения вероятностей, соответствующих наблюдаемым отстаткам в позициях.

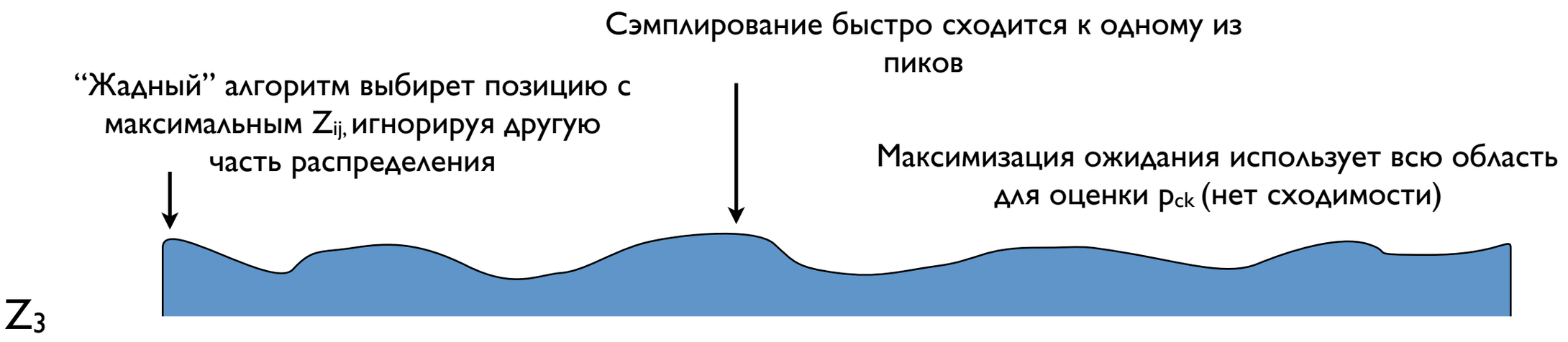
- Обновление параметров модели мотива p_{ck} на основе вычисленной матрицы Z_{ij}:
  - Максимизация ожидания: частоты остатков взвешиваются с весами *Zij* и вычисляются по всем позициям последовательностей.
  - “Жадный” подход: для каждой последовательности *i* выбирается одна позиция *j* - позиция с максимальной вероятностью *Zij* вдоль последовательности.
  - Сэмплирование: выбирается одна позиция для каждой последовательности согласно распределению *Zij*.

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



Все три подхода приведут к одинаковому результату





Благодарности

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