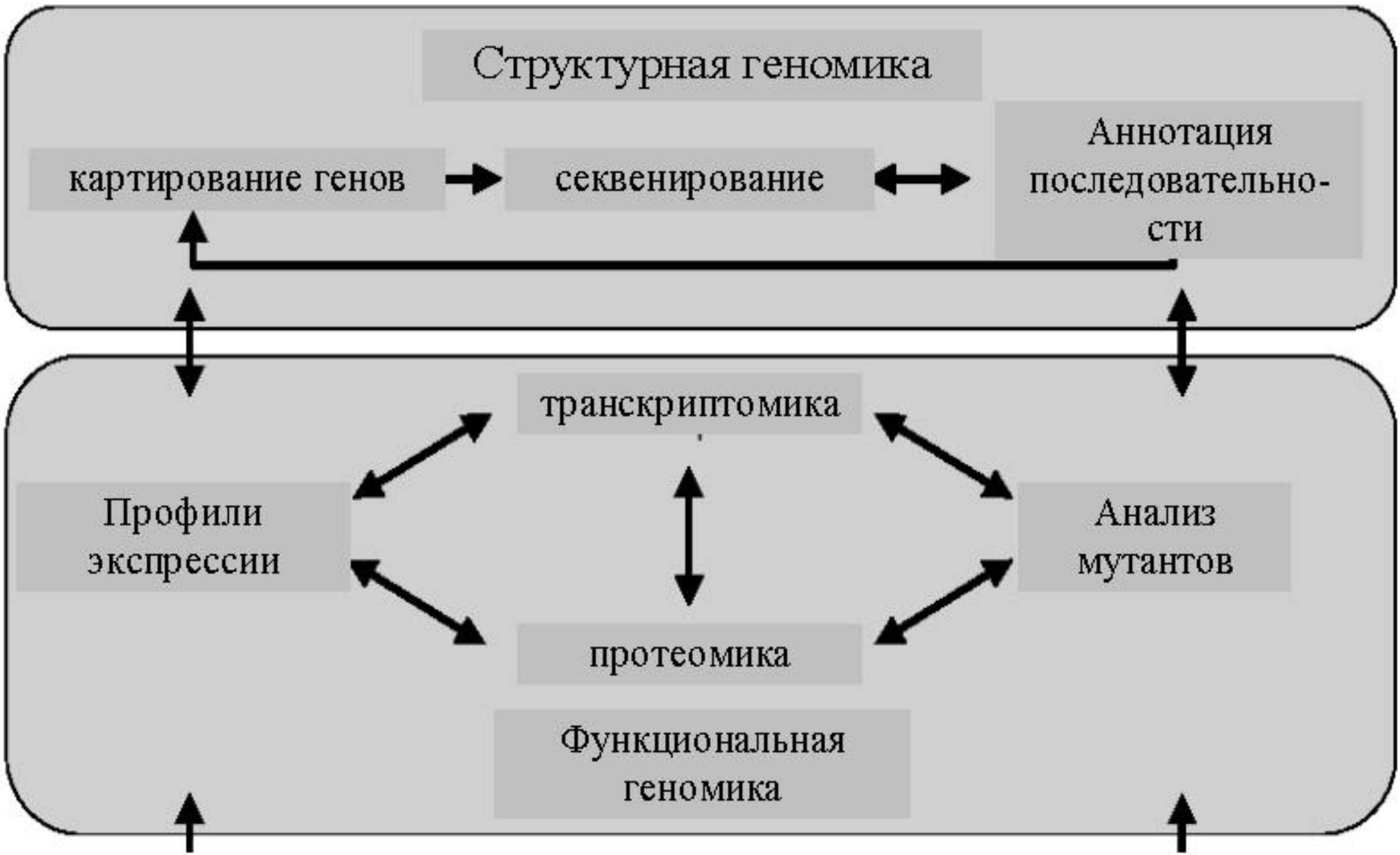
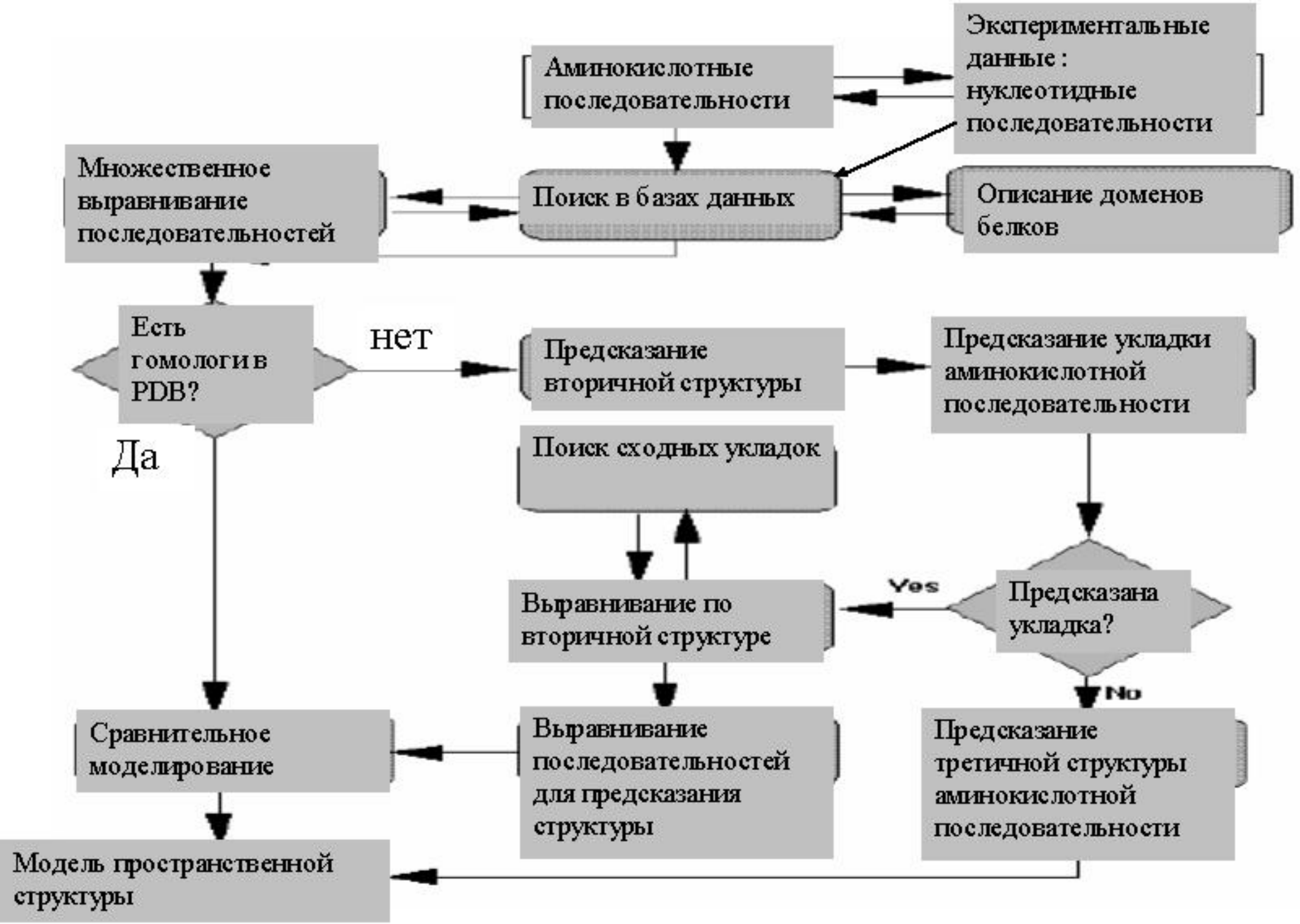
Уровни изучения генома



Регуляторные последовательности ДНК

Для изучаемой последовательности ДНК определяют:

- Сайты рестрикции для экспериментов по клонированию
- Сайты связывания транскрипционных факторов для понимания регуляции транскрипции гена
- Открытые рамки считывания и аминокислотные последовательности, соответствующие им
- Для протяженных последовательностей определяют потенциальную экзон-интронную структуру для поиска генов
- Вторичную структуру РНК – продуктов транскрипции генов
- Частоты использования кодонов в открытых рамках считывания


Аминокислотные последовательности
Экспериментальные данные : нуклеотидные последовательности
Множественное выравнивание последовательностей
Поиск в базах данных
Описание доменов белков
Есть гомологи в PDB?
нет
Да
Предсказание вторичной структуры
Предсказание укладки аминокислотной последовательности
Поиск сходных укладок
Выравнивание по вторичной структуре
Yes
Предсказана укладка?
No
Сравнительное моделирование
Выравнивание последовательностей для предсказания структуры
Предсказание третичной структуры аминокислотной последовательности
Модель пространственной структуры

Укладка ДНК и регуляция транскрипции

Хромосомы эукариот упакованы в хроматин, первым уровнем упаковки является нуклеосомный уровень

Нуклеосомы содержат гистоны, гистоны подавляют транскрипцию, поскольку конкурируют с регуляторными белками за связывание с ДНК

Установлено с помощью экспериментов по чувствительности к ДНК-азе I

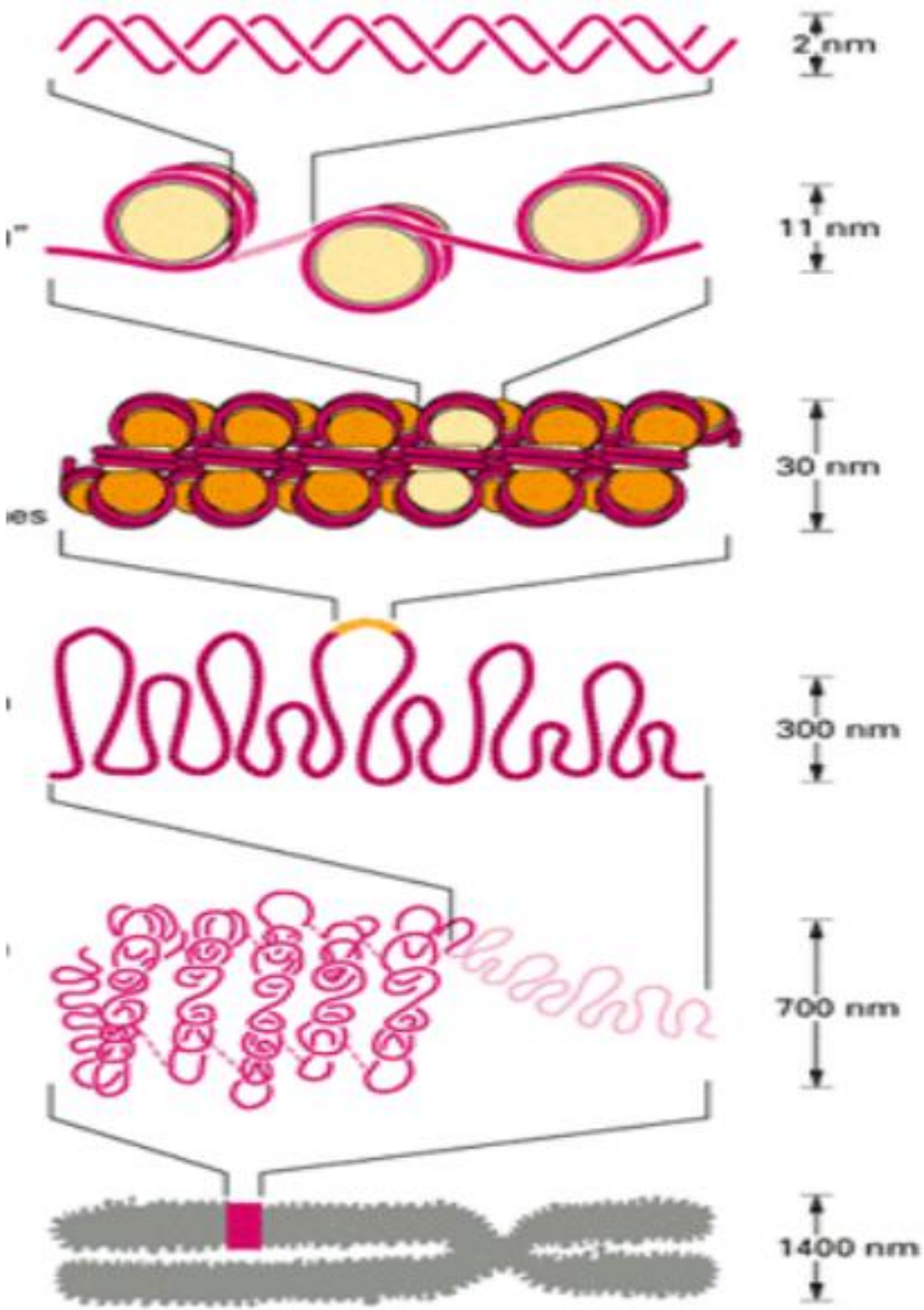
- ДНК-аза I быстро расщепляет транскрипционно активную ДНК
- Гистоны защищают ДНК от ДНК-азы I и быстрого расщепления не происходит
- Эксперименты по добавлению гистонов и регуляторных белков транскрипции показали, что гистоны конкурентно связываются с промоторами и подавляют транскрипцию

Решение «проблемы гистонов»:
Транскрипционно активные гены имеют менее развитую структуру хроматина, чем неактивные гены

- Гистоны ацетилируются и фосфорилируются, изменяя способность связываться с ДНК

- Белки, связывающиеся с энхансерами, блокируют гистоны и

Укладка ДНК



Направления геномики

Анализ экспрессии генов

Изменения экспрессии во времени

Определение регуляторных районов

Функциональная классификация и рассмотрение ортологов

Сравнения геномов

Изучения семейств ортологов

Выявление и анализ часто встречающихся мотивов

Аннотация геномов

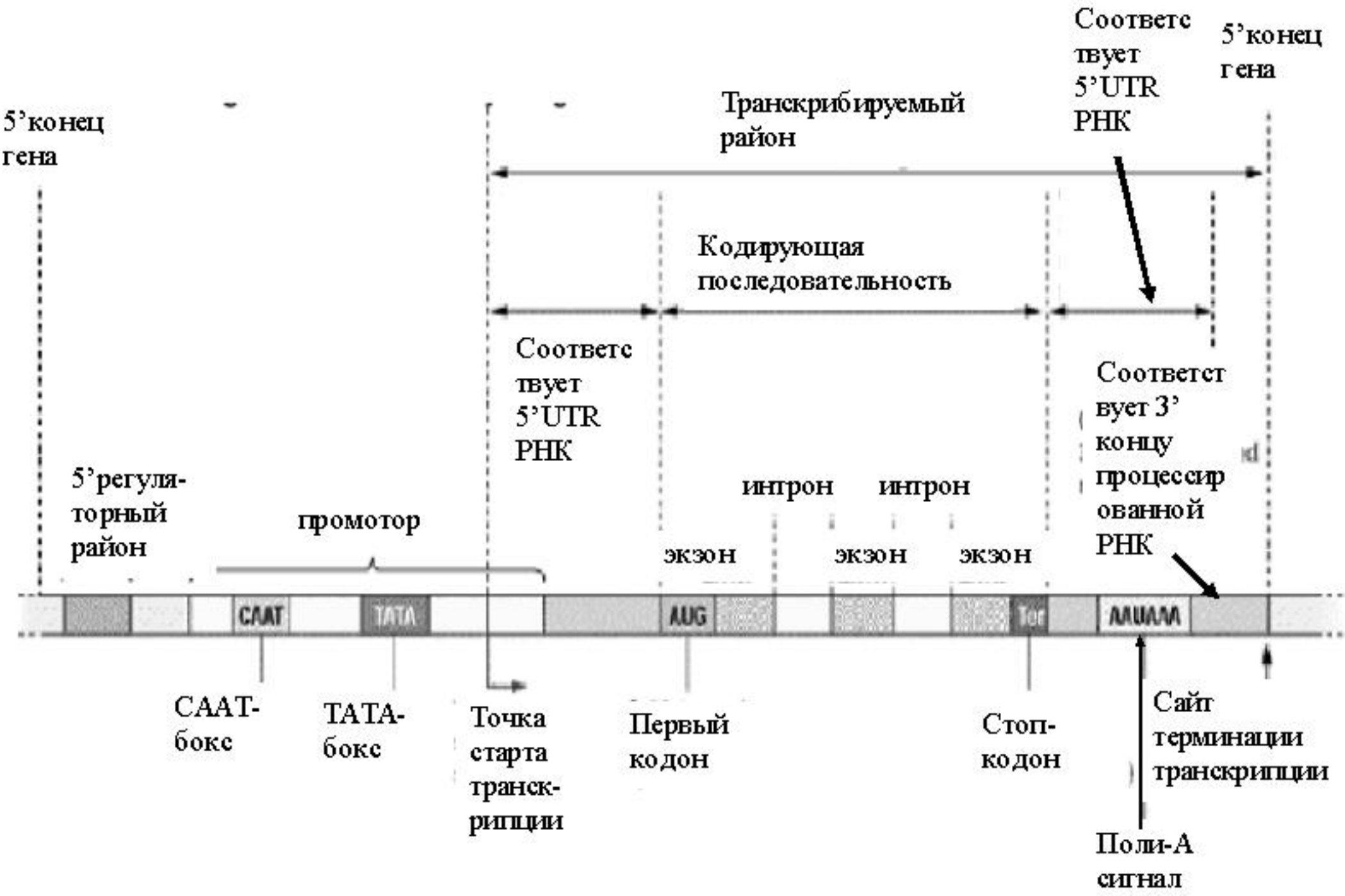
Построение деревьев сходств геномов

Структурная геномика

Предсказание генов и других элементов геномов

Направления биоинформатики

Схема структуры эукариотического гена

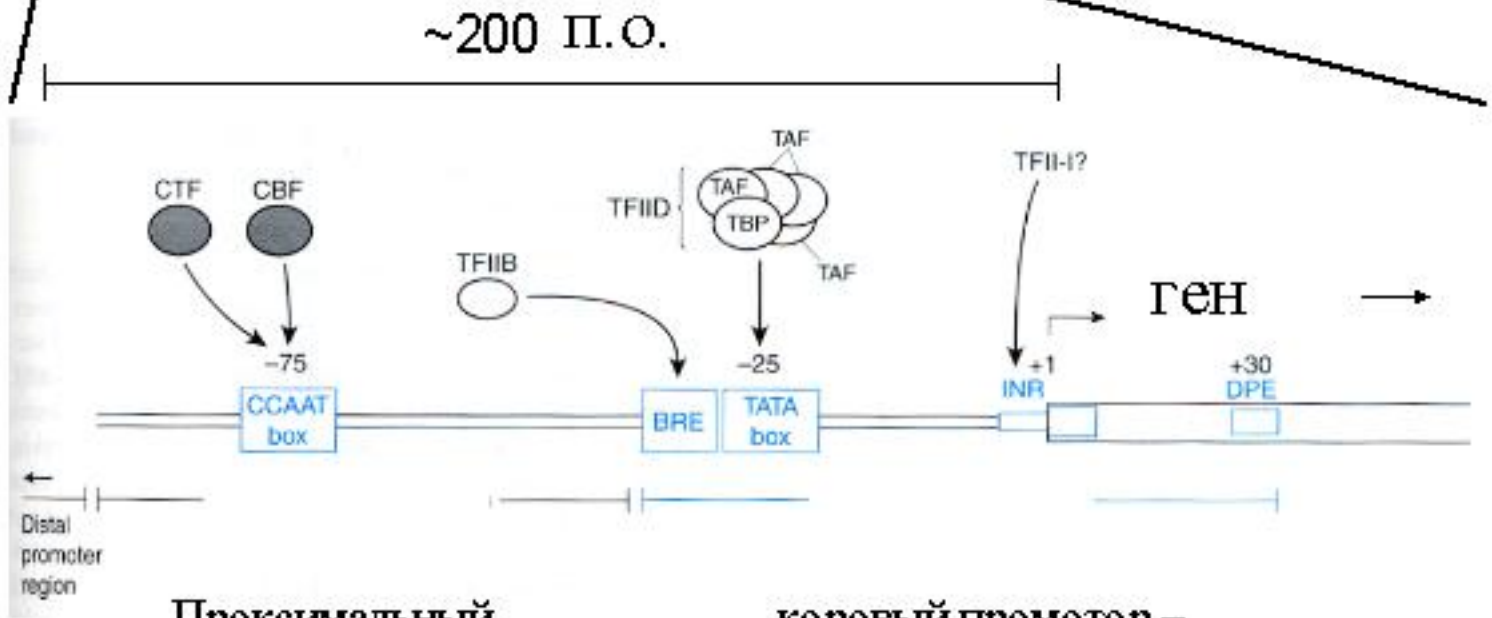


Регуляция гена

- Гены регулируются на уровне транскрипции регуляторными белками, которые связываются с сайтами ДНК, расположенными выше, ниже и в пределах последовательности гена
  - Промоторы
  - Энхансеры
  - Сайленсеры
  - Локус-контролирующие районы
- Гены регулируются:
  - Метилированием ДНК
  - Пост-транскрипционно мРНК-белковыми взаимодействиями
  - Пост-трансляционно белок-белковыми взаимодействиями

Промоторы

- РНК-полимераза связывается с ДНК промотора и базальным транскрипционным комплексом, и начинается транскрипция
- Регуляция транскрипции в большой степени связана с регуляторными элементами промотора



Для транскрипции гена высших эукариот базальные транскрипционные факторы должны связаться с ДНК промотора

Проксимальный промотор –200 –50 п.о.

коровый промотор – 200 –50 п.о.

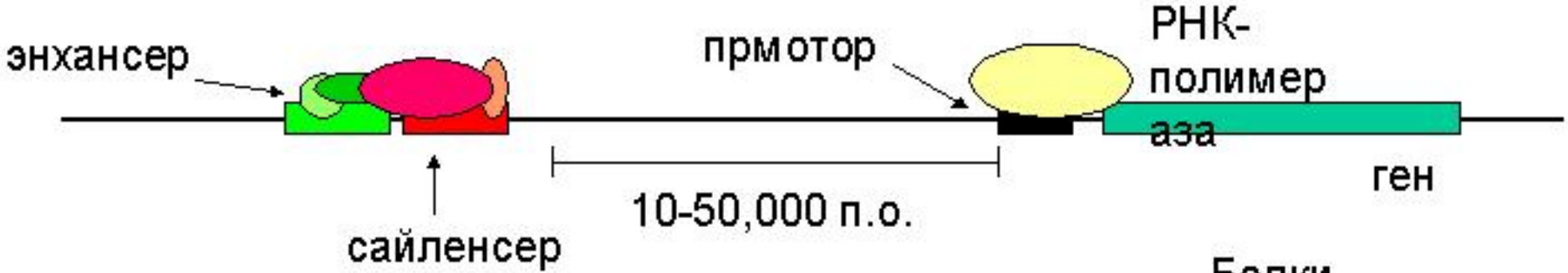
Транскрипция гена регулируется промоторами

Промоторы:

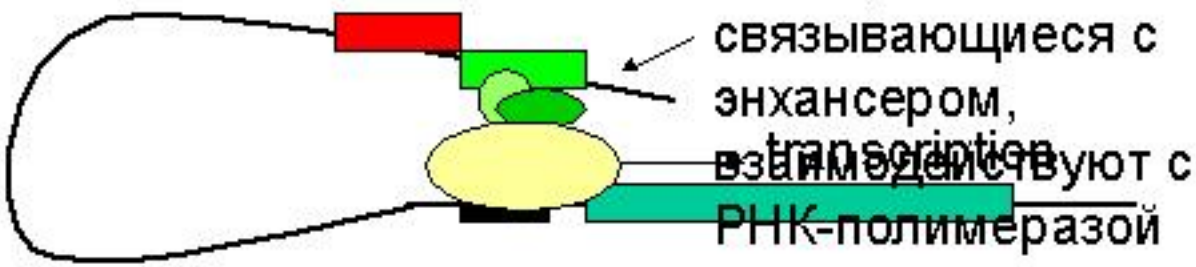
- **Расположены выше точки старта транскрипции**
- **Содержат регуляторные элементы, определяющие (1) где находится точка транскрипции (например, ТАТА-бокс), (2) будет ли транскрипция (сайты связывания транскрипционных факторов)**
- **С сайтами, расположенными в пределах промоторов, связываются транскрипционные факторы.**
- **Для одного гена может быть один или несколько промоторов**
- **Промоторы могут содержать позитивные и негативные регуляторные элементы**

Энхансеры и сайленсеры

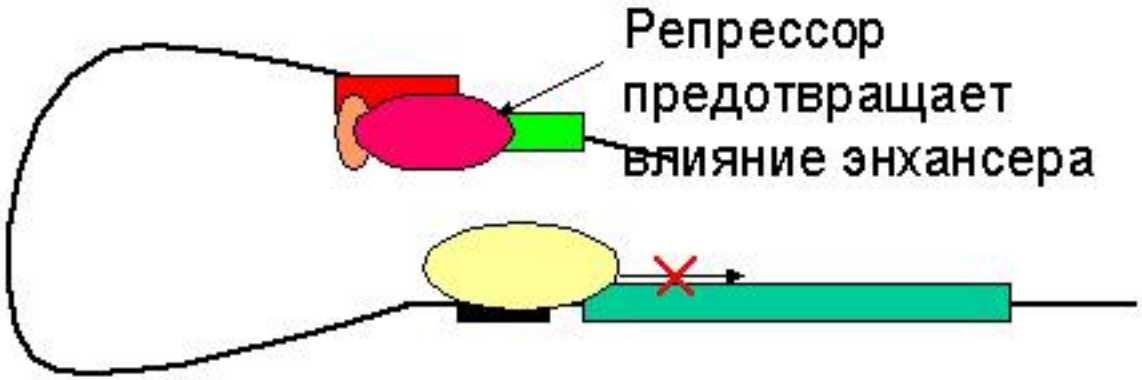
Районы, расположенные более удаленно выше промотора содержат сайты связывания регуляторных белков, которые влияют позитивно или негативно на транскрипцию гена



Энхансеры усиливают транскрипцию



Репрессоры предотвращают действие энхансеров



Энхансеры

Находятся выше и ниже точки старта транскрипции.

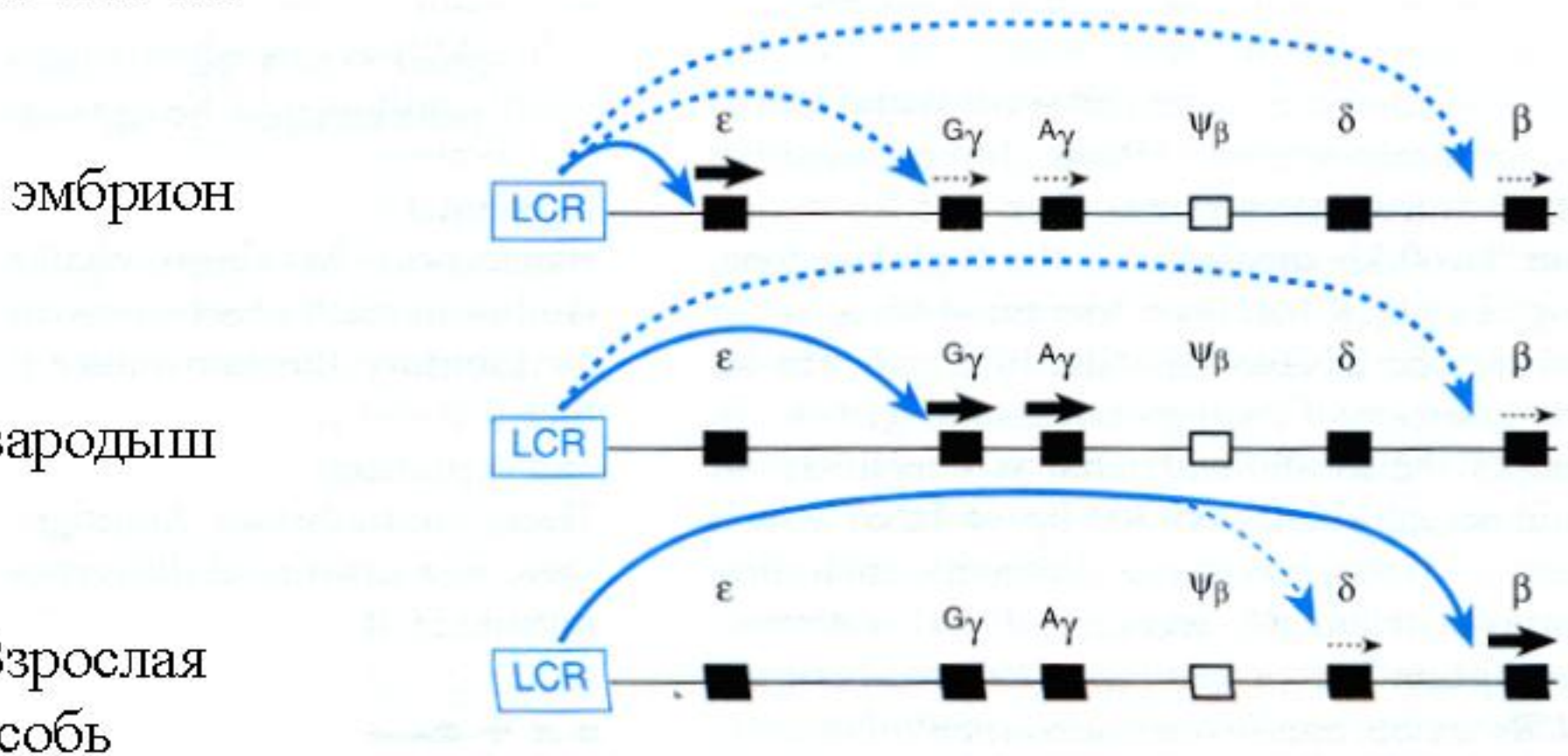
- Регуляторные белки связываются со специфическими последовательностями энхансера; связывание определяется последовательностью ДНК.
- Могут формироваться петли ДНК, которые позволяют транскрипционным факторам связываться элементами энхансера.
- Взаимодействия регуляторных белков определяют, будет транскрипция активирована или подавлена.

Промоторы и энхансеры

- Некоторые регуляторные белки являются общими для прмоторов и энхансеров, другие – специфичными.
- Каждый промотор и энхансер содержит специфичный набор сайтов связывания белков, определяющих уровень экспрессии.
- Уровень экспрессии регулируется взаимодействием позитивных и негативных регуляторных белков.
- Число комбинаций связывающихся с промоторами и энхансерами белков чрезвычайно велико.

Локус-контролирующие районы

"Локус-контролирующий район" регулирует множество генов, находящихся на удаленном расстоянии. Механизм неизвестен.



Характеристики, которые различаются у генов и другими районами ДНК

Характеристики гена:

i. динуклеотидные, кодонные и дикодонные характеристики

ii. Регуляторные районы

iii. Точка старта транскрипции, сайты сплайсинга, сайты терминации транскрипции, другие сайты

iv. Родственные гены: отношение числа неконсервативных к числу консервативных замен < 1, мало делеций, инсерций, в основном длина кодирующей части кратна трем.

Характеристики районов ДНК, не содержащих гены:

i. Много повторов

ii. В течение эволюции: отношение числа неконсервативных к числу консервативных замен = 1, много инсерций, делеции, сдвигов рамки считывания.

Различия между прокариотами и эукариотами:

- **Экспрессия прокариотических генов регулируется опероном, содержащим сайты связывания регуляторных белков.**

- **Экспрессия эукариотического гена также регулируется с участием регуляторных районов, содержащих сайты связывания регуляторных белков. Гены не объединены в цистроны.**

- **Экспресссия эукариотического гена носит более сложный характер – транскрипция проходит в ядре, трансляция в цитоплазме.**

- **Рассматривают два типа регуляции эукариотического гена:**

  - **Кратковременную регуляцию – гены «включаются» и «выключаются» в ответ на воздействия окружающей среды и потребности клетки.**

  - **Долговременную регуляцию – регуляцию экспрессии генов в течение онтогенеза, дифференциация клеток.**

Различные уровни регуляции эукариотического гена

Экспрессия эукариотического гена регулируется следующими уровнями:

3. транскрипция

5. Процессинг РНК

7. Транспорт мРНК

9. Трансляция мРНК

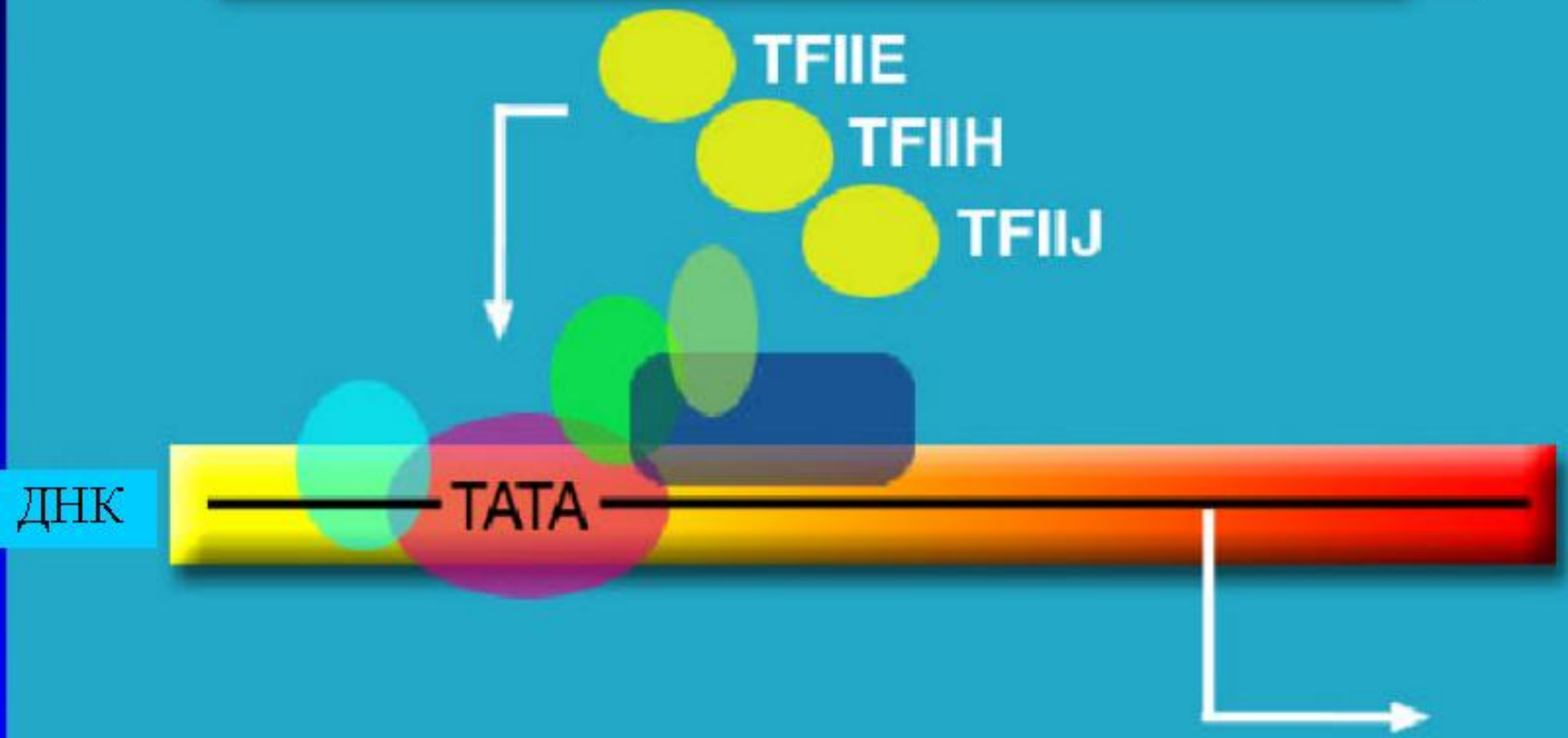
11. Деградация мРНК

13. Деградация белка

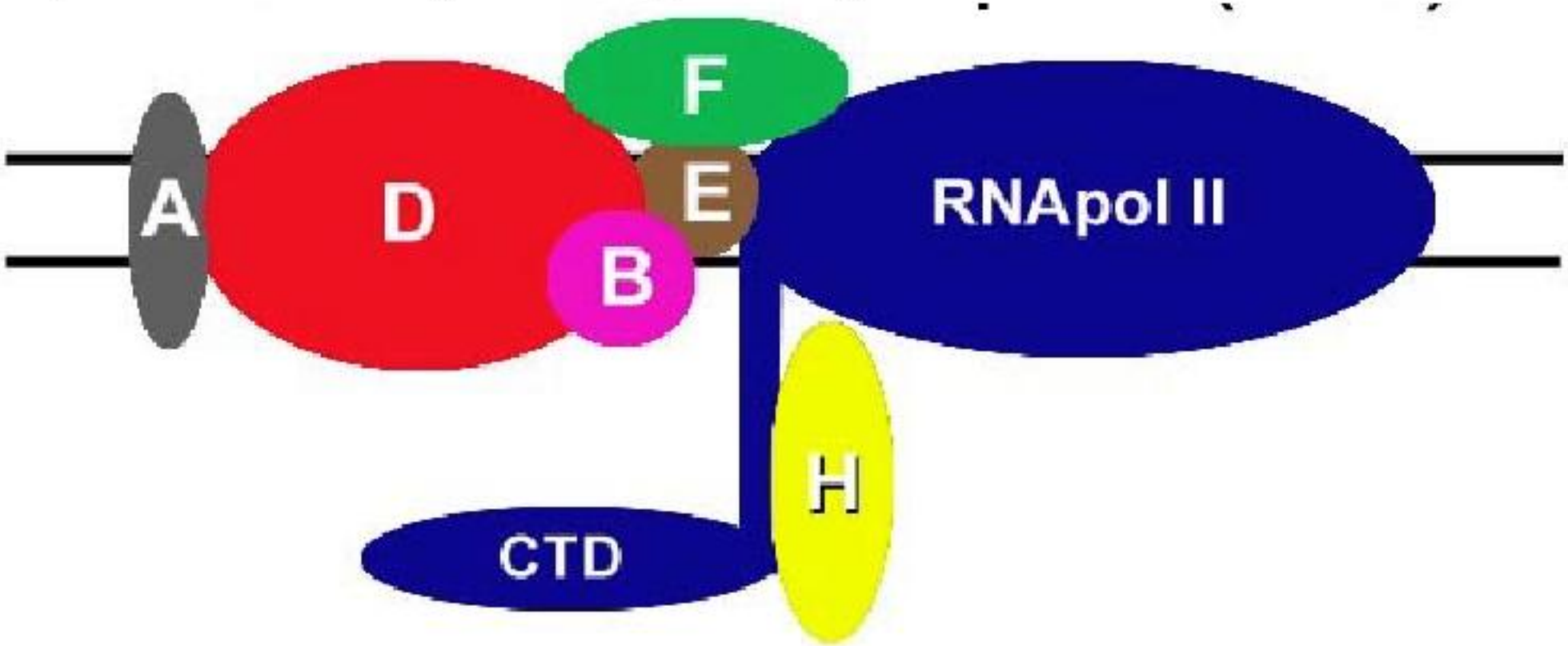
Экспрессия прокариотического гена регулируется:

транскрипцией

Образование преинициаторного транскрипционного комплекса



преинициаторный транскрипционный комплекс



RNApol: РНК-полимераза

CTD : Карбоксильный терминальный домен

TFIID $_{11}$**:** TBP + 10 TPB-связанных факторов

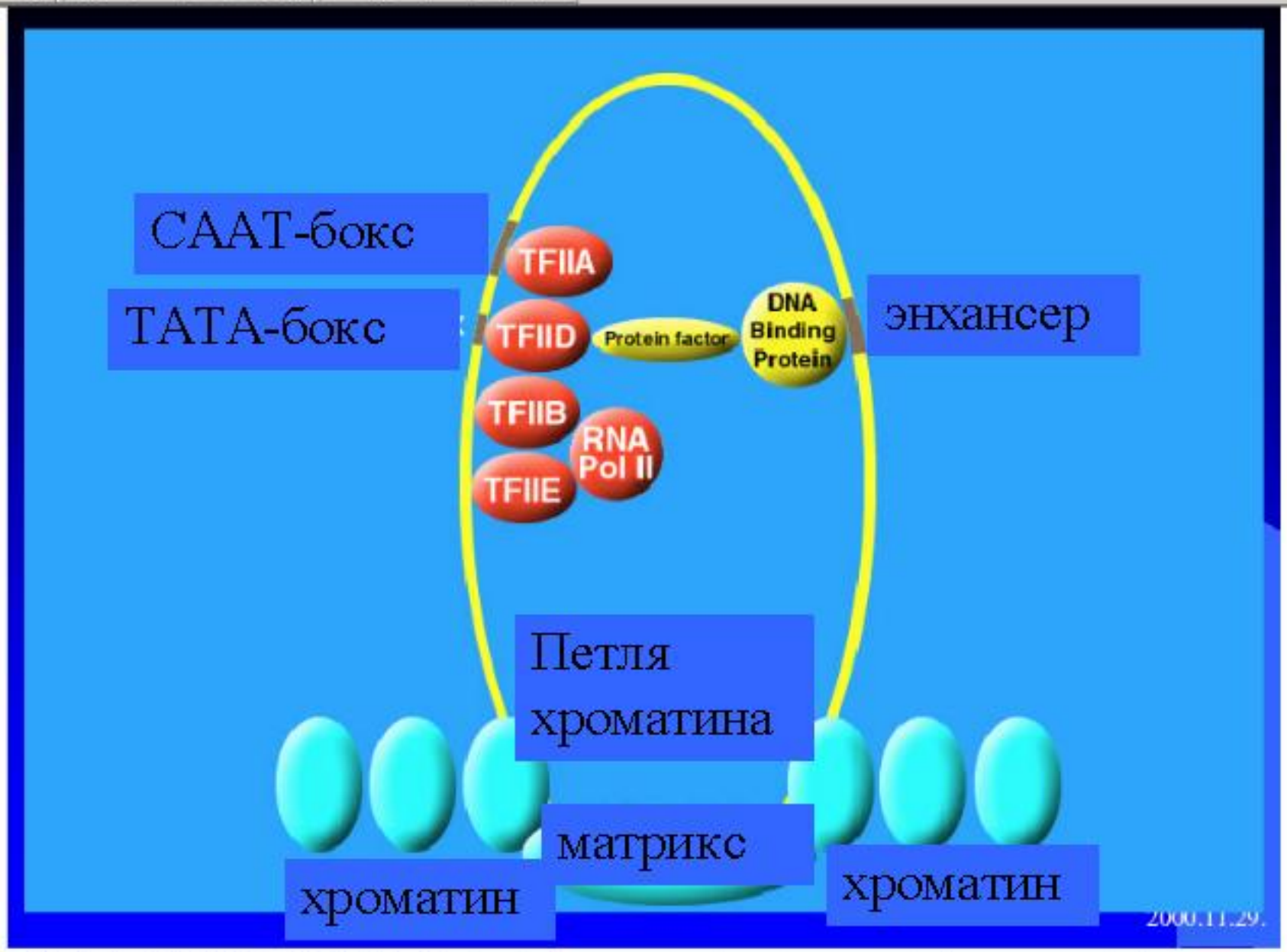
TFIIB $_{1}$**:** Одиночный пептид связывается после TBP

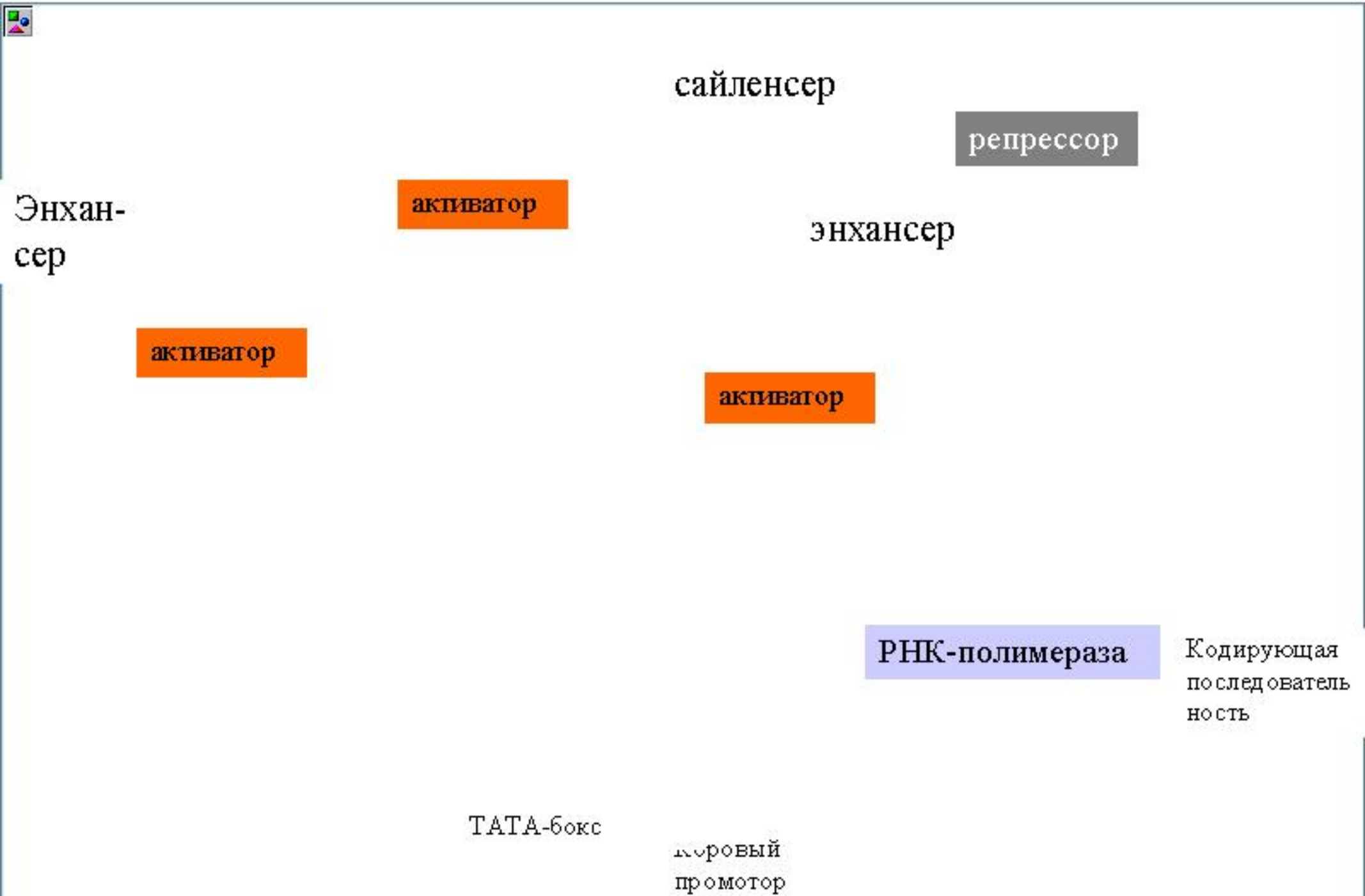
TFIIF $_{4}$**:** Стабилизирует комплекс TBP-TFIIB-pol-III, необходимый для связывания IIE

TFIIE $_{4}$**:** Помогает связыванию TFIIH

TFIIH $_{9}$**:** Имеет каталитическую активность (геликазную и киназную)

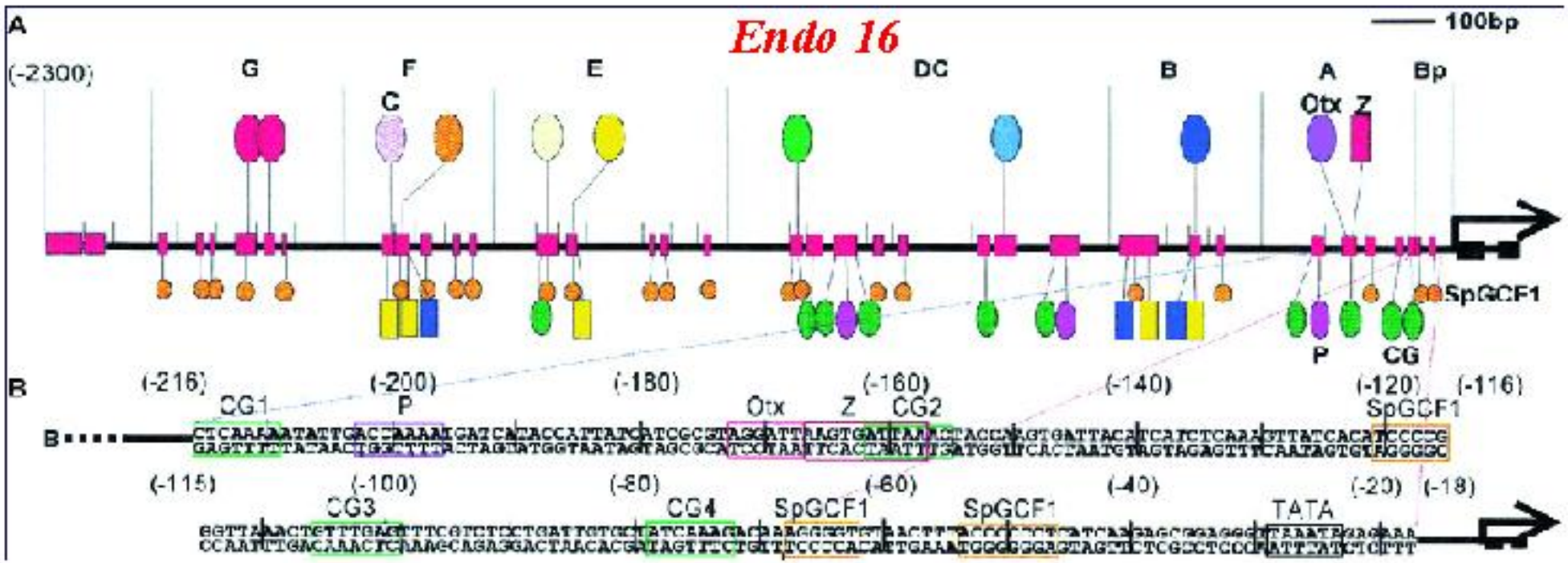
* Число субъединиц


CAAT-бокс
TATA-бокс
TFIIA
TFIID
Protein factor
DNA Binding Protein
энхансер
TFIIB
RNA Pol II
TFIIE
Петля хроматина
матрикс
хроматин
хроматин
2000.11.29.

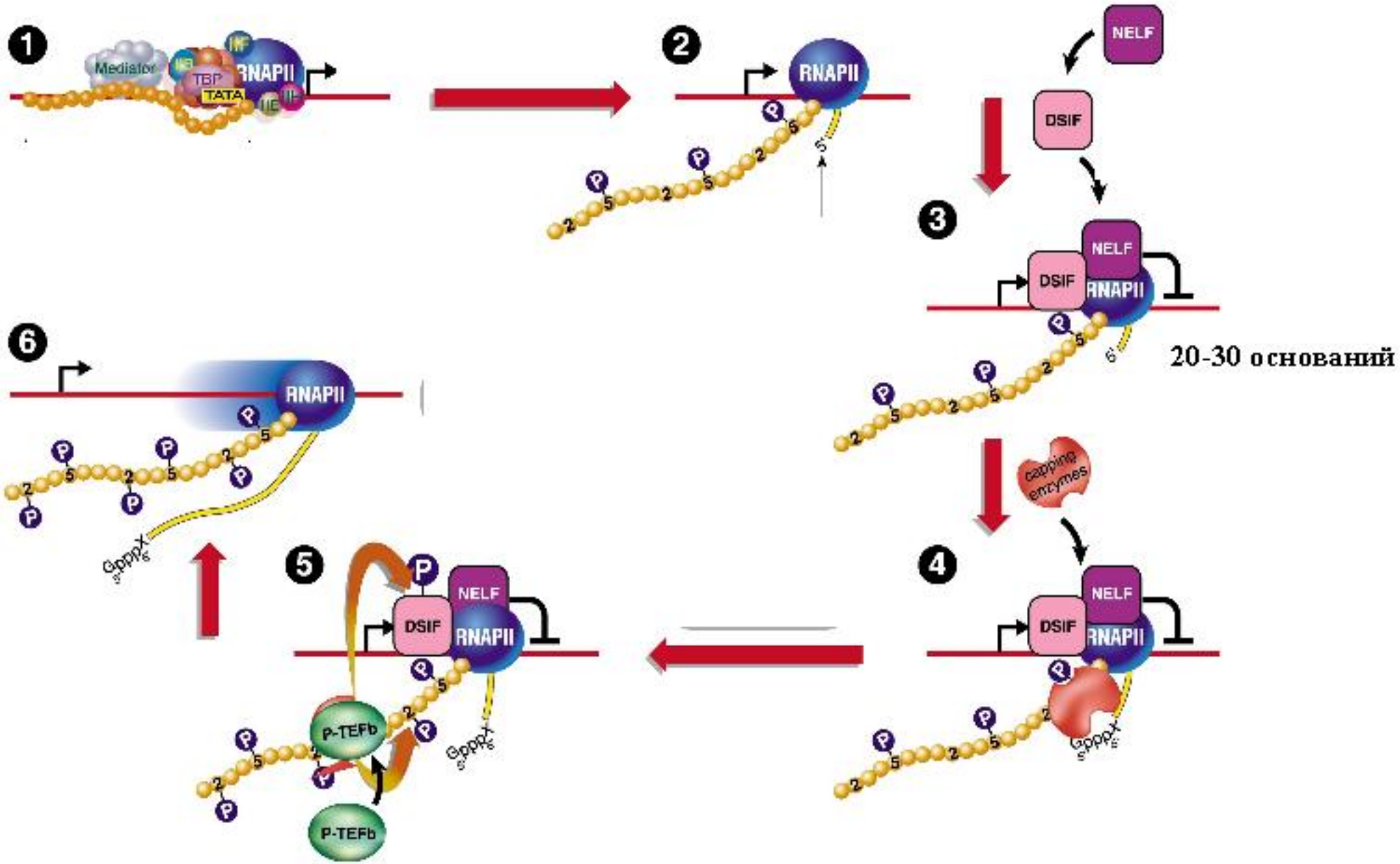

сайленсер
репрессор
Энхан-
сер
активатор
энхансер
активатор
активатор
РНК-полимераза
Кодирующая
последователь
ность
ТАТА-бокс
ровый
промотор

Транскрипционные факторы могут взаимодействовать друг с другом

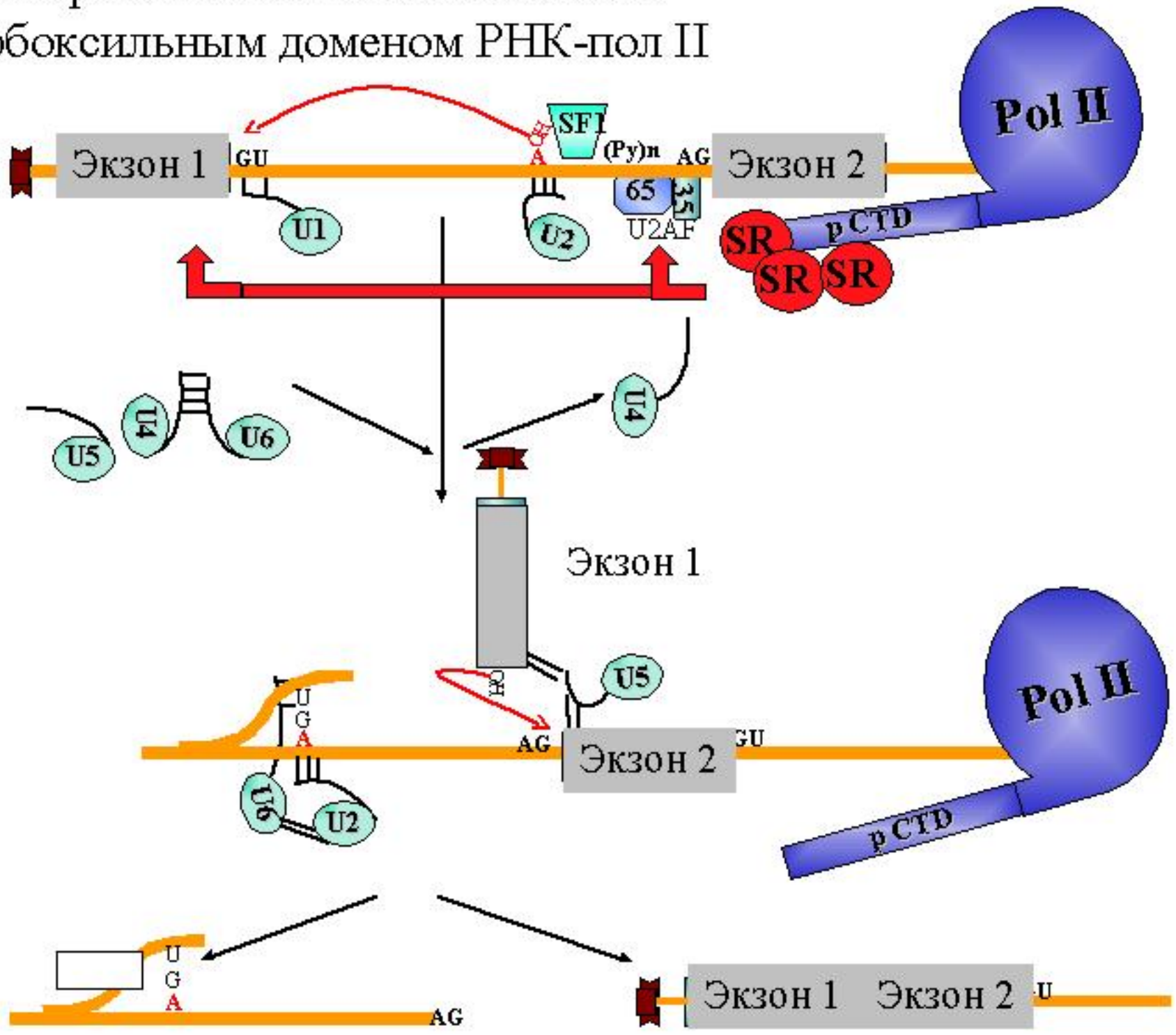
- Регуляторные элементы эукариотических генов часто собраны в «модули»
- Транскрипционные факторы часто действуют синергично (кооперативно)



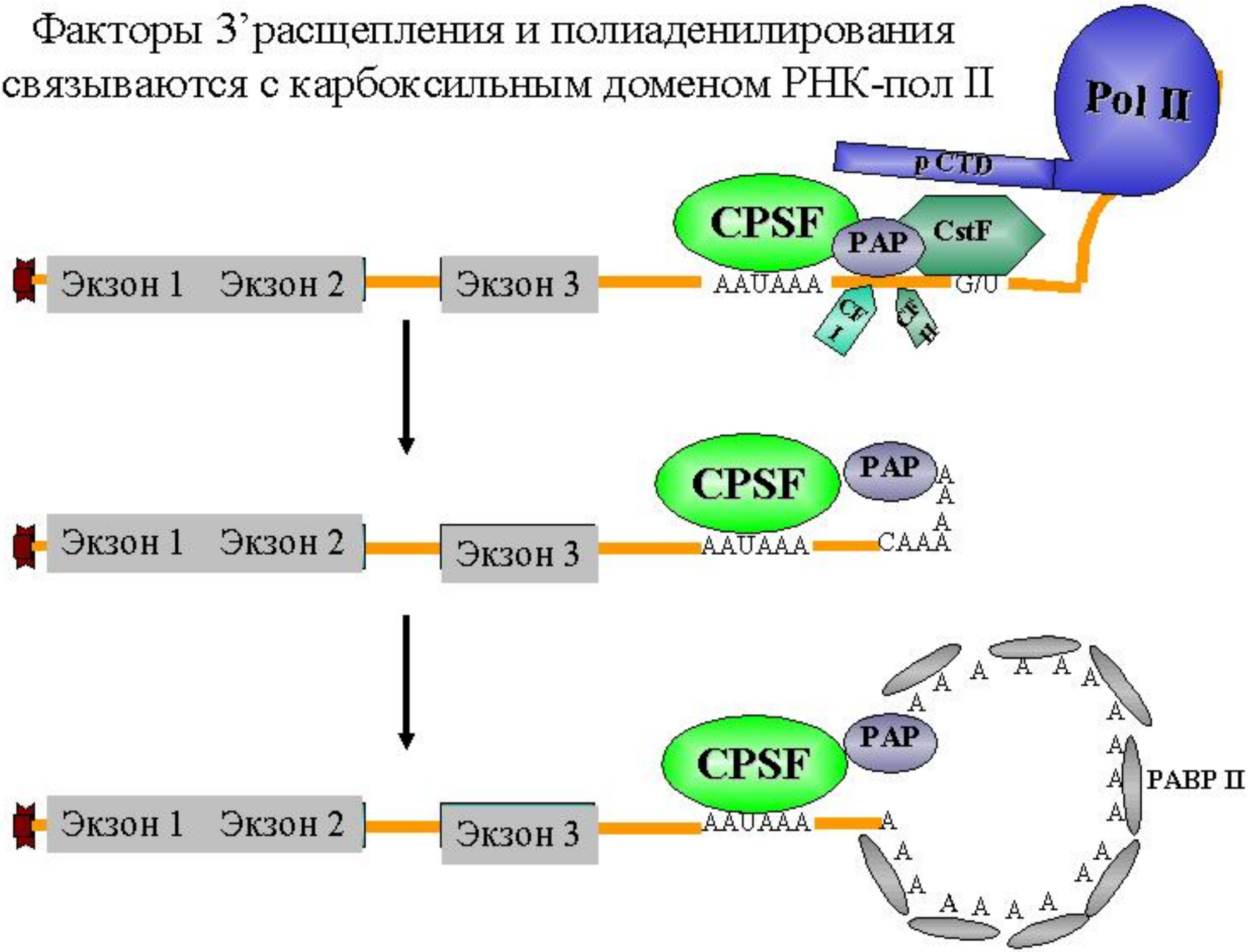
Кепирование 5’ пре-мРНК и элонгация транскрипции



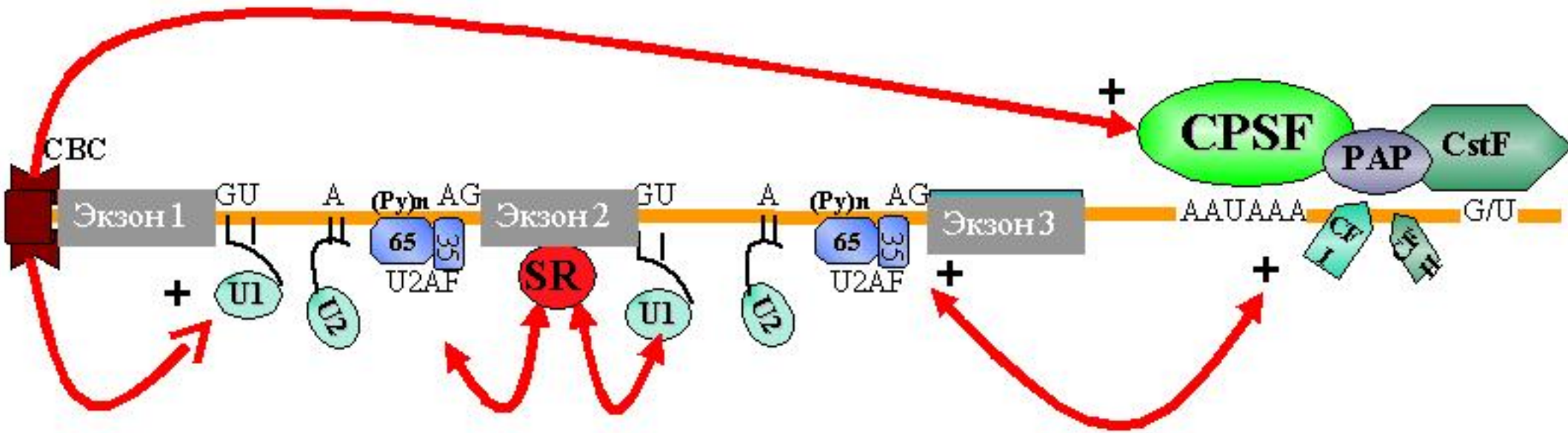
Факторы сплайсинга связываются
с карбоксильным доменом РНК-пол II



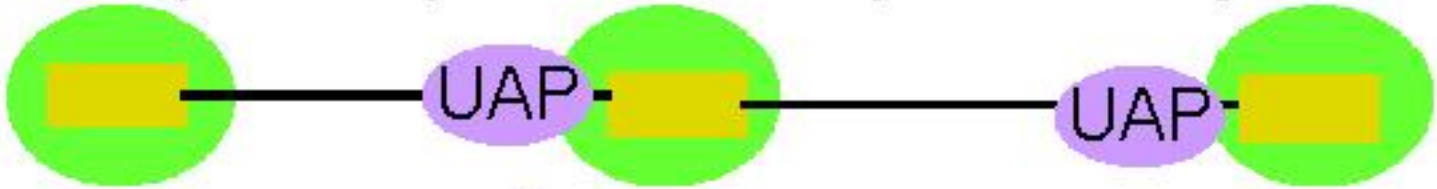
Факторы 3’расщепления и полиаденилирования связываются с карбоксильным доменом РНК-пол II



Сложное взаимодействие аппарата транскрипции и факторов процессинга мРНК

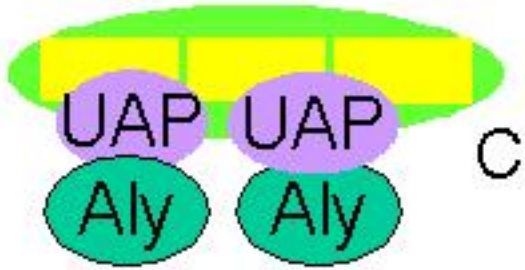


Сплайсинг нужен для эффективного транспорта мРНК в цитоплазму

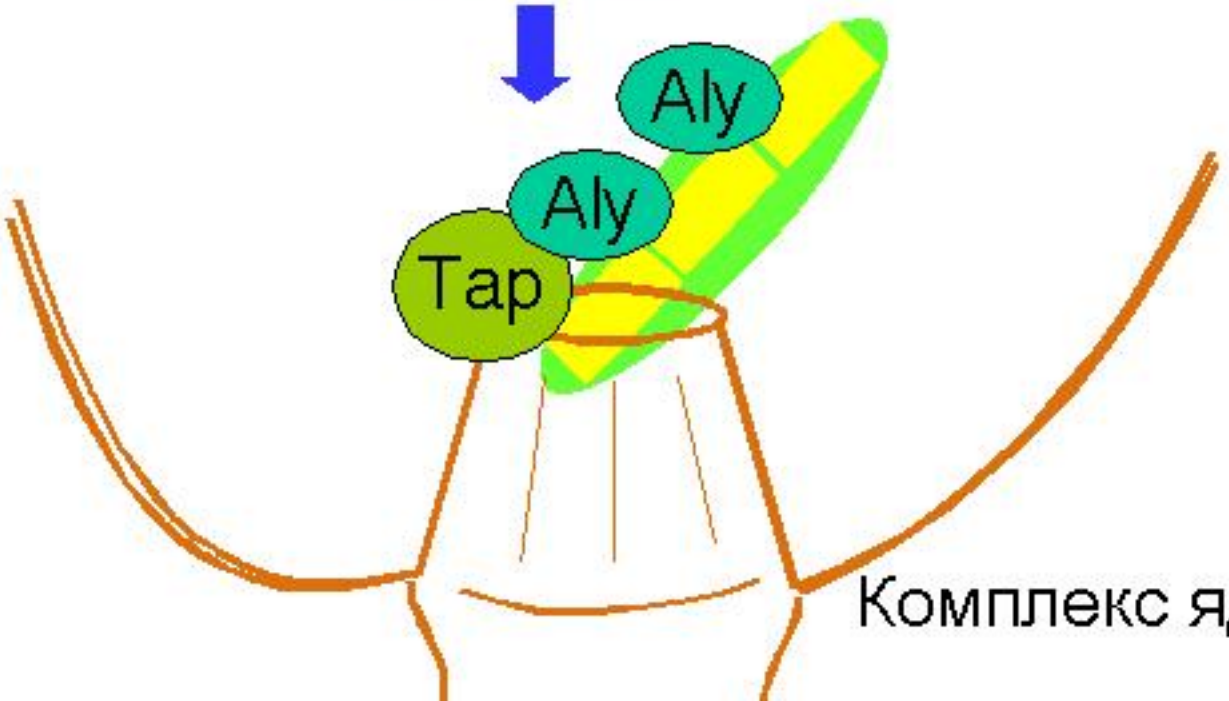


сплайсосома

UAP56 – фактор сплайсинга; он присоединяет Aly к мРНК



Сплайсированная мРНП

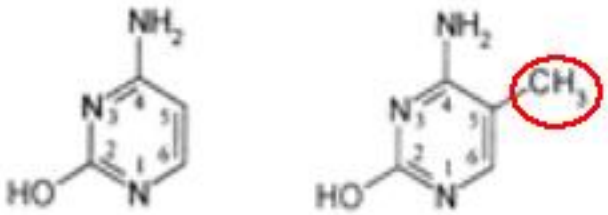


Комплекс ядерной поры

Процессы, связанные с экспрессией эукариотического гена

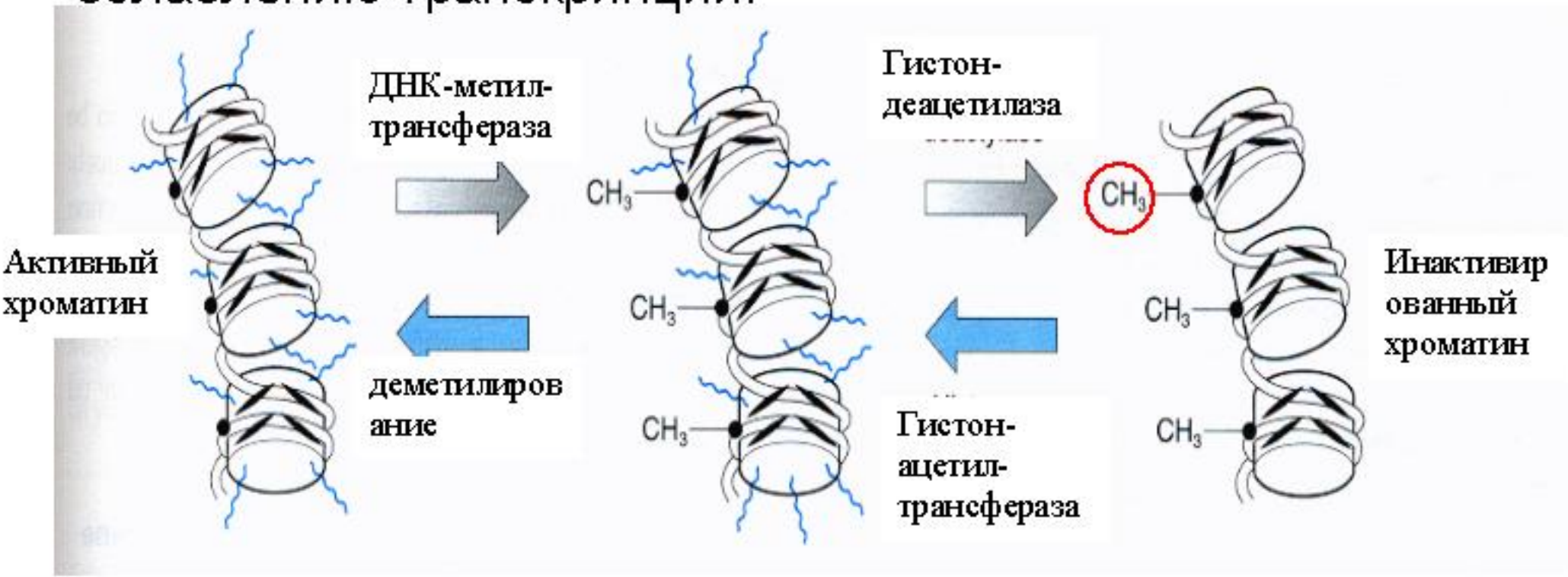
Метилирование ДНК

- Нуклеотидное основание “С” может быть метилировано:

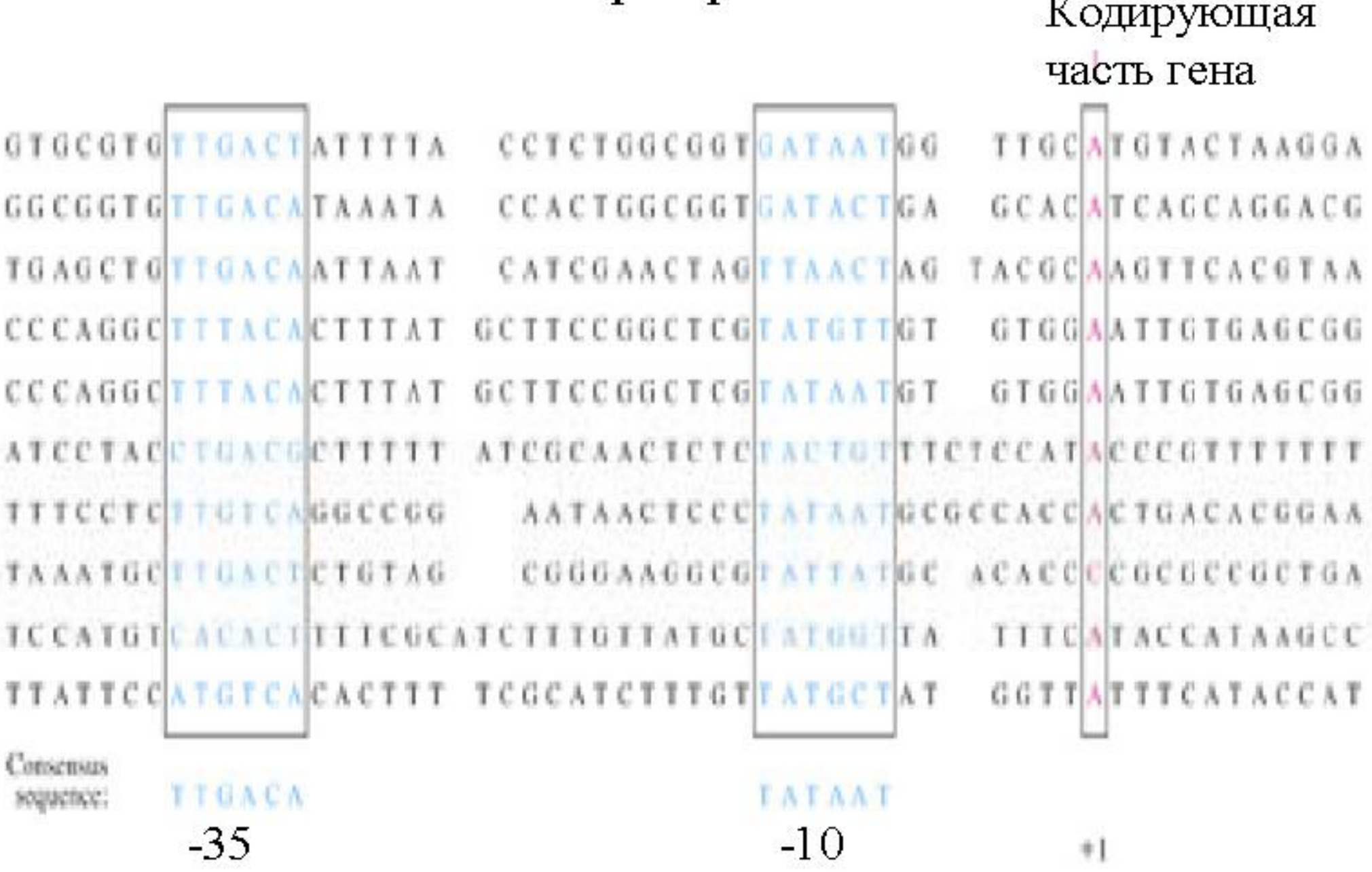


цитозин 5-метил-цитозин

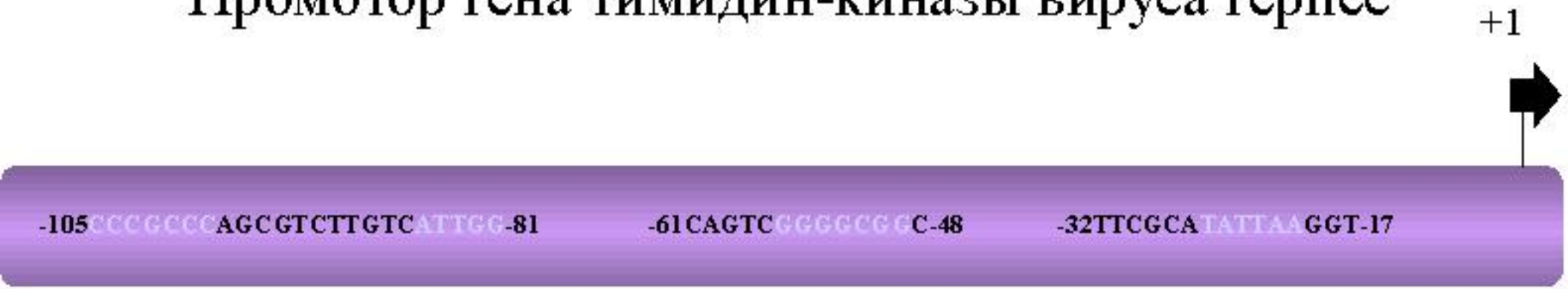
- Это приводит к более плотной упаковке ДНК и ослаблению транскрипции:



Выравнивание 10 сильных прмоторов генов бактерий и бактериофагов



Промотор гена тимидин-киназы вируса герпес



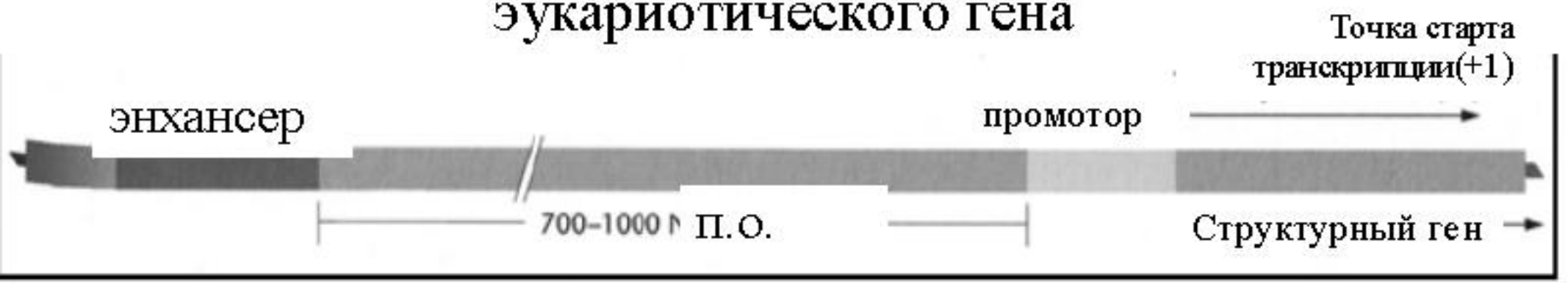
GC богатый элемент

CCAAT – бокс

GC богатый элемент

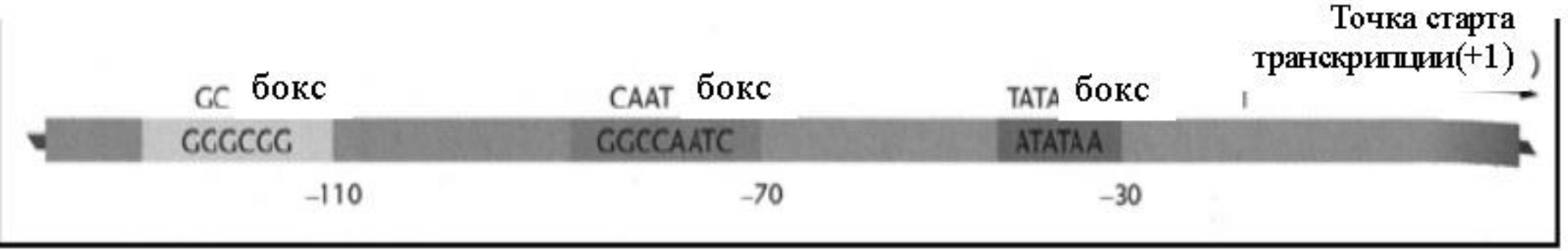
TATA бокс

Схема расположения регуляторных районов эукариотического гена



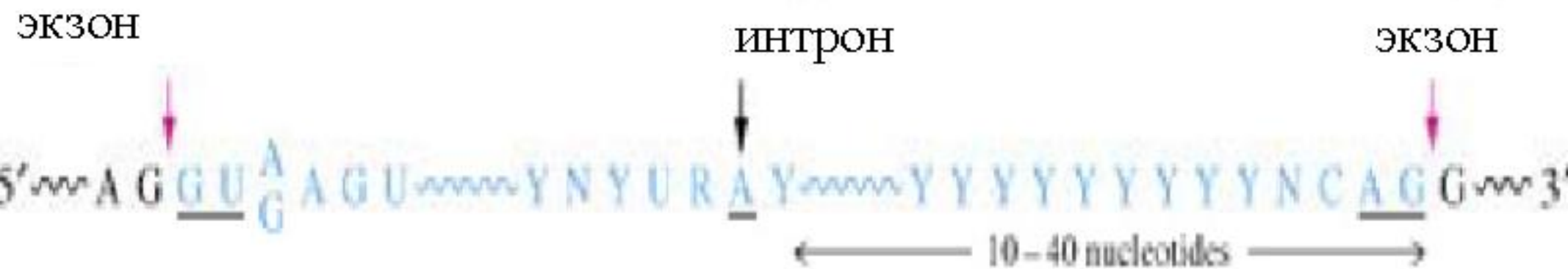
•Промоторный район содержит регуляторные элементы, обычно расположенные на расстоянии около 100. п.о. от точки старта транскрипции, включающие: ТАТА-бокс (-25 –35 п.о.), СААТ-бокс (-70 –80 п.о.) и GC-бокс (примерно –110 п.о.)

•Другие регуляторные районы энхансеры. Их расположение нестрого фиксировано относительно точки старта транскрипции. Они могут быть расположены выше, ниже и в пределах гена



Сайты сплайсинга

Консенсусные последовательности сайтов спайсинга позвоночных



Консенсус 5’сайта сплайсинга

Консенсус точки ветвления

Консенсус 3’сайта сплайсинга

Функциональные сайты

•Функциональный сайт – наименьшая функциональная единица генома (также называемая функциональный мотив, сигнал)

•Комбинация функциональных сайтов составляет функциональный район

•Функциональной единицей белок-кодирующей последовательности является кодон.

•Различные функциональные сайты могут перекрываться и взаимодействовать

•Функциональные сайты могут присутствовать в кодирующих частях гена

•часто функциональные сайты являются сайтами связывания регуляторных белков

Функциональные сайты

Анализ и распознавание функциональных сайтов основано на поиске контекстных характеристик, общих для всех сайтов, выполняющих специфическую функцию

Функциональные сайты, как правило, характеризуются вариабельностью контекстных характеристик, что затрудняет создание точных методов распознавания.

Требования к выборке сайтов

- Для адекватного анализа выборки сайтов требуется гомогенность выборки (то есть сходство анализируемых структур сайтов)

- Гетерогенность выборки приведет к тому, что такие методы как консенсус и весовая матрица будут построены с усреднением различных характеристик и проявят низкую точность

Классификация сайтов сплайсинга

Одним из предложенных методов классификации сайтов сплайсинга является метод, строящий набор консенсусов. Он основан на следующих допущениях:

- Высокая частота опрпдпленных нуклеотидных оснований в определенных позициях сайта отражает функциональную важность этого нуклеотидного основания в данной позиции.

- Нуклеотидные основания в различных позициях сайтов могут быть взаимнозависимы, формируя таким образом структуру, которую могут распознавать какие-либо факторы

Классификация сайтов сплайсинга

Предложенный метод классификации сайтов сплайсинга, строящий набор консенсусов, характеризуется недостаточной точностью распознавания сайтов сплайсинга. Для распознавания генов более важно не пропустить сайты сплайсинга, повышенная частота ложно-положительных предсказаний не играет негативной роли

Поли-А сайты

- Поли-А сайт, расположенный ниже экзон-богатого района является хорошим свидетельством в пользу того, что нуклеотидная последовательность содержит ген, который завершает кодирующую часть найденным поли-А сайтом

- Предсказание поли-А сайтов в настоящее время характеризуется низкой точностью и дает много ложно-положительных и ложно-отрицательных предсказаний

Поли-А сайты

- Наиболее известный вариант поли-А сайта имеет вид AATAAA, расположен на 15-20 п.о. выше точки расщепления РНК и присоединения поли-А хвоста
- Около 90% РНК содержат точную копию этой последовательности. Другой часто встречающийся вариант ATTAAA
- Анализ соседних оснований показал, что некоторые другие основания могут быть важны для распознавания поли-А сайта.
- Выявлен дополнительный сигнал YGTGTTYY на расстоянии 20-30 п.о. ниже сайта расщепления РНК

Районы прикрепления к ядерному матриксу

Районы прикрепления к ядерному матриксу связаны с сайтами, которые регулируют экспрессию генов. Пример программы поиска этих районов:

www.ncgr.org/MarFinder

Островки CpG

- На 5’конце около половины генов млекопитающих расположен островок CpG
- Предполагают, что на 5’конце всех генов домашнего хозяйства млекопитающих расположен островок CpG
- Наличие CpG островков рядом с генами других позвоночных не столь постоянно

Островки CpG и структура генома

•Первичная структура генома очень гетерогенна. Наиболее известным примером такой гетерогенности являются островки CpG.

•Островки CpG составляют специфическую фракцию генома и, в отличие от большей части ДНК, они неметилированы и характеризуются повышенным содержанием CpG

•Островки CpG характеризуются значительно повышенным содержанием C:G по сравнению с остальной ДНК. Островки CpG могут служить маркерами генов

•В гаплоидном геномме человека присутствует около 45 000 островков CpG

Сайты связывания транскрипционных факторов

•Белок-кодирующие гены эукариот транскрибируются РНК-полимеразой II. Транскрипция инициируется на промоторном районе с участием комплекса различных факторов.

•Элементарные регуляторные сигналы короткие (5-30 п.о.) и очень вариабельны.

Часто встречающимися сигналами промоторов генов, транскрибируемых РНК-полимеразой II, являются ТАТА-бокс, сайт кепирования, СААТ-бокс, GC-бокс. ТАТА-бокс присутствует в большинстве промоторных районов белок-кодирующих генов.

Методы распознавания функциональных сайтов

- консенсусные последовательности
  - Весовые матрицы
- решающие деревья
  - Скрытые марковские модели (Hidden Markov Models, HMMs)
  - Нейронные сети
  - другие

Пример консенсусной последовательности

- В самом простом случае в каждую позицию консенсуса ставится наиболее часто встречающееся нуклеотидгное основание в 4- или 15-буквенном коде

Консенсус может приводить к потере информации и к ложным предсказаниям

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATCTT

TATAAT

consensus sequence

TATRNT

consensus (IUPAC)

MELON
MANGO
HONEY
SWEET
COOKY

MONEY

Пример (позиционной) весовой матрицы

- Computed by measuring the frequency of every element of every position of the site (weight)

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

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

- Вес потенциального сайта определяется как сумма весов матрицы, типу основания в каждой позиции потенциального сайта
- Недостатки:
  - Требуется пороговое значение веса
  - Предполагает независимость встреч оснований в позициях

Пример частотной матрицы ТАТА-бокса

Позиции

Posi-	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11
(a)															
A	16	4	90	1	91	69	92	57	40	14	21	21	21	17	20
C	37	12	0	2	0	0	1	1	11	35	38	33	30	28	26
G	39	5	1	1	1	0	5	11	40	39	33	33	33	36	36
T	8	79	9	96	8	31	2	31	9	12	8	13	16	19	18
(b)	G	T	A	T	A	A	A	A	G	G	C	G	G	G	G
	C		T		T	T		T	A	C	G	C	C	C	C

Пример частотной матрицы сайта, содержащего инициаторный кодон

5′ Нетранслируемый район | Кодирующая последовательность

Позиции	-6	-5	-4	-3	-2	-1	+1	+2	+3
	G	C	C	A G	C	C	A	T	G
A	18	19	24	68	23	15	100	0	0
C	21	40	58	2	55	53	0	0	0
G	47	23	12	30	16	23	0	0	100
T	13	18	6	0	7	9	0	100	0

Пример частотной матрицы сайта, содержащего инициаторный кодон

Позиции	-3	-2	-1		+1	+2	+3	+4	+5	+6
A	28	59	8	/	0	0	54	74	5	16
C	40	14	5	/	0	0	2	8	6	18
G	17	13	81	/	100	0	42	11	85	21
T	14	14	6	/	0	100	2	8	4	45

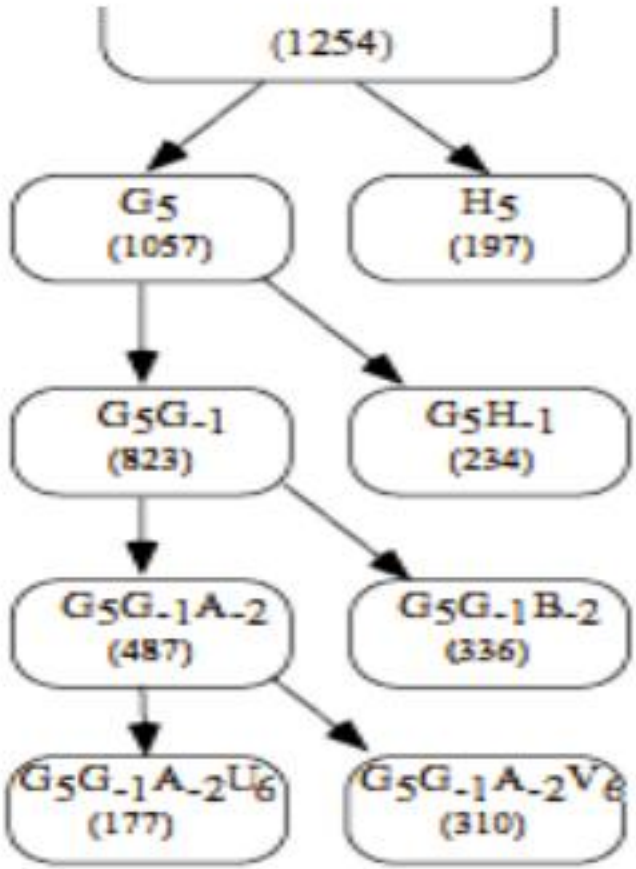
A	G	/	G	T	R		
G	/	G	T	N	A	G	
	/	G	T	R	A	G	
	/	G	T	R	N	G	T
G	/	G	T	A			

Акцепторные сайты сплайсинга

Позиции	интрон												экзон			
	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	/	+1
	T	T	T	T	T	T	T	T	T	T	N	C	A	G	/	G
				C	C	C	C	C	C	C	C	C				
A	10	8	6	6	9	9	8	9	6	6	23	2	100	0		28
C	31	36	34	34	37	38	44	41	44	40	28	79	0	0		14
G	14	14	12	8	9	10	9	8	6	6	26	1	0	100		47
T	44	43	48	52	45	44	40	41	45	48	23	18	0	0		11

Пример решающего дерева

Все донорные сайты сплайсинга



A%	C%	G%	U%	Pos	A%	C%	G%	U%
33	36	19	13	-3	35	44	16	6
56	15	15	15	-2	85	4	7	5
9	4	78	9	-1	2	1	97	0
44	3	51	3	+3	81	3	15	2
75	4	13	9	+4	51	28	9	12
14	18	19	49	+6	22	20	30	28
34	37	18	11	-3	29	31	21	18
59	10	15	16	-2	43	30	17	11
40	4	53	3	+3	56	0	43	0
70	4	16	10	+4	93	2	3	3
17	21	21	42	+6	5	10	10	76
37	42	18	3	-3	29	30	18	23
39	5	51	5	+3	42	1	56	1
62	5	22	11	+4	80	4	8	8
19	20	25	36	+6	14	21	16	49
32	40	23	5	-3	39	43	15	2
27	4	59	10	+3	46	6	46	3
51	5	25	19	+4	69	5	20	7

Все сайты

позиция

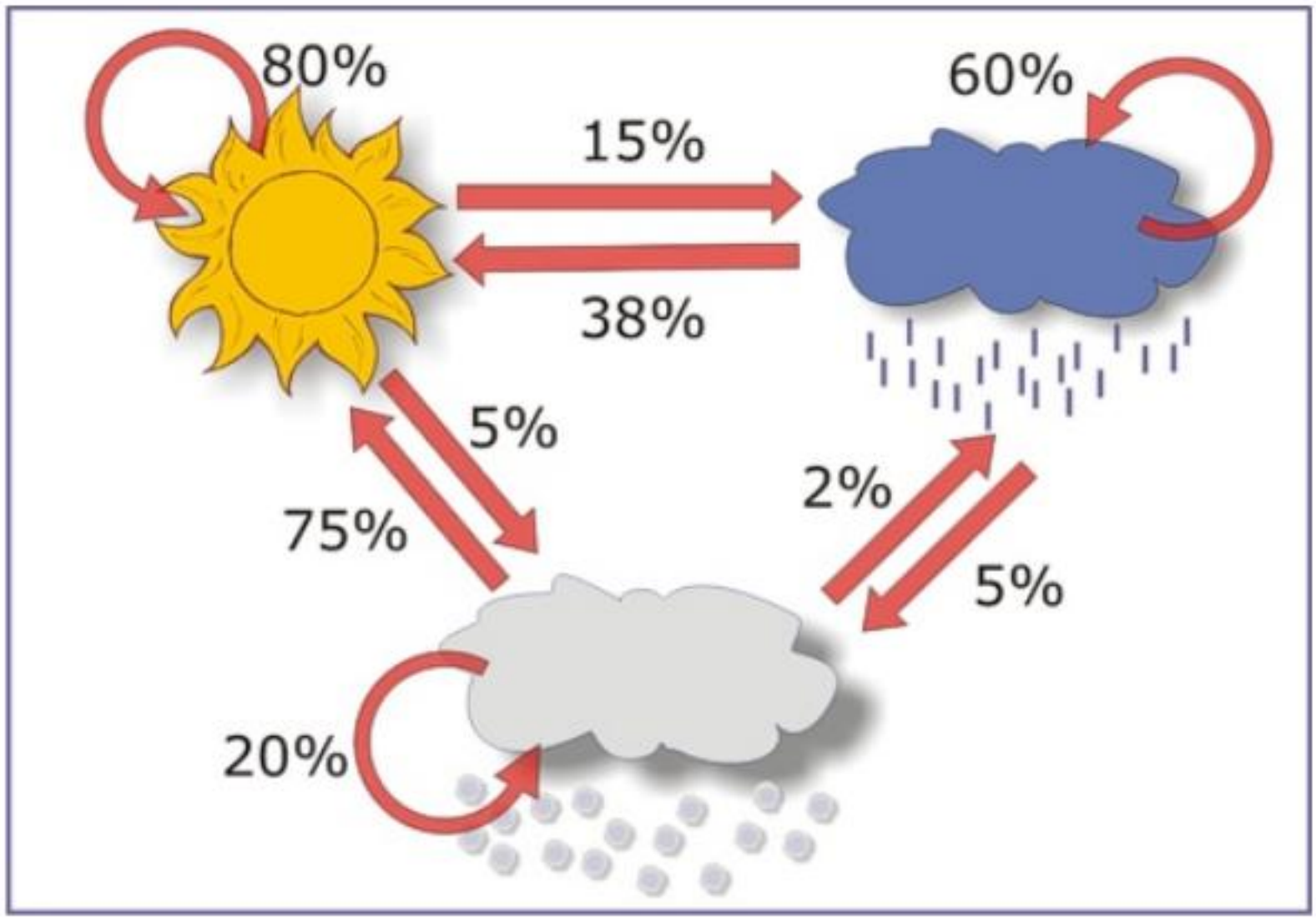
Base	-3	-2	-1	+1	+2	+3	+4	+5	+6
A%	**33**	**60**	8	0	0	**49**	**71**	6	15
C%	**37**	13	4	0	0	3	7	5	19
G%	18	14	**81**	**100**	0	**45**	12	**84**	20
U%	12	13	7	0	**100**	3	9	5	**46**

мяРНК 3' G U C C A U U C A 5'

Распознавание сайтов сплайсинга в кДНК

- Предсказание сайтов сплайсинга в лишенных интронов кДНК важно для некоторых экспериментальных работ
- Сайты сплайсинга в кДНК состоят из экзонных частей донорного и акцепторного сайтов сплайсинга
- Поэтому консенсусной последовательностью для таких сайтов является MAG/G

Марковские модели



Составляющие марковской модели

- Набор состояний:

$\{S_1, S_2, \ldots, S_N\}$

- Вероятности переходов между состояниями (матрица транзиций)

$A_{ij} = P(q_{t+1} = S_i \mid q_t = S_j)$

- Распределение начального состояния

$\pi_i = P(q_1 = S_i)$

Методы распознавания сайтов сплайсинга

- **Весовые матрицы** Staden, 1984, Shapiro and Senapathy, 1987, Senapathy et al., 1990; Zhang and Marr, 1993),
- **перцептрон** Nakata et al., 1985), quantification method (Iida, 1987; Iida, 1988),
- **learning technique** (Kudo et al., 1987); Quinqueton and Moreau, 1985),
- **Нейронные сети** (Brunak et al., 1991; Brunak et al., 1990; Lapedes et al., 1990),
- **K-tuple статистика** Bougueleret et al., 1988; Solovyev, 1993; kel et al., 1993),
- **Дискриминантная** (Gelfand, 1989; Penotti, 1991).

энергия

Программы распознавания сайтов сайтов связывания транскрипционных факторов

- В настоящее время программы распознавания сайтов связывания транскрипционных факторов интегрированы с базами данных о сайтах связывания транскрипционных факторов, консенсусах, весовых матрицах
- SIGNAL SCAN (Prestridge, 1991),
- MATRIX SEARCH (Chen et al., 1995),
- ConsInspector (Frech et al., 1993) and MatInspector (Quandt et al., 1995).
- PromoterView
- (http://www.itba.mi.cnr.it/tradat)
- (http://transfac.gbf-braunschweig.de)
- (http://www.gsf.de).

Пример ДНК-белкового взаимодействия

Структуры белков

- **Первичная структура** – последовательность аминокислот
- **Вторичная структура** – локальная пространственная организация белка; короткие участки укладки – альфа-спирали и бета-складки
- **Третичная структура** – пространственная структура белка – как отдельные альфа-спирали и бета-складки соотносятся друг с другом в пространстве (собираются в домены)
- **Четвертичная структура** – белок может состоять более чем из одной полипептидной цепи

последовательности белков

Суперсемейство

Семейство

домен

Мотив

Сайт

Аминокислотный остаток

Сложность ССТФ как объекта изучения

1) Консервативные и вариабельные участки в последовательности сайта

2) Число консервативных участков в сайте

3) Направление, в котором работает сайт.

4) Границы и локализация консервативных участков.

5) Оптимальная длина флангов.

Определение локализации кора. Пример – выборка сайтов связывания SF-1

atgtcaaggccgtgac	atgt**CAAGGCCG**tgac
aggctcaaggtcatca	aggct**CAAGGTCA**tca
tcaaggagaaggtcag	tcaagga**GAAGGTCA**g
aaagtagaggtcagga	aaagt**AGAGGTCA**gga
gaggcaaggccactgg	gagg**CAAGGCCA**ctgg
taccaaggtcagaaat	tac**CAAGGTCA**gaaat
gagttcaaggtaataa	gagtt**CAAGGTAA**taa
tttcgaggtcatggcc	ttt**CGAGGTCA**tggcca
gacttcaaggtcccaa	gactt**CAAGGTCC**caa
cccgccaaggcccatg	ggggg**CAAGCCCC**atg

Подбор направлений последовательностей в выборке. Пример – выборка сайтов связывания SF-1

caggccaaggtcaaac

gtagttcaaggcaata

ctcctaaggtcatcct

aaattaccttgacac

caggc**CAAGGT**caaac

gtagtt**CAAGGC**aata

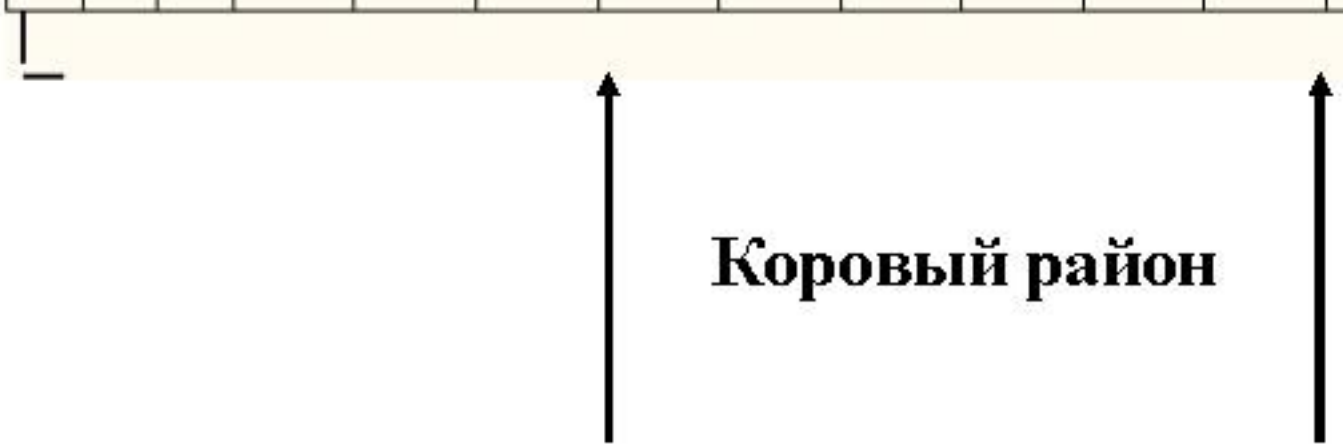
ctcc**TAAGGC**catcct

gtggt**CAAGGT**aattt

ggcac**CAAGGC**tagag

Поиск консервативных участков в последовательностях выборки. Матрица частот оснований для выборки сайтов SF-1

A		9	16	7	6	2	**35**	**39**	0	0	3	4	**26**	7	11	7	9	3
C		8	10	7	13	**33**	2	1	0	0	10	**28**	7	12	7	5	8	1
T		8	3	7	19	0	0	0	0	0	**26**	8	4	10	7	11	3	3
G		9	9	19	2	5	4	1	**41**	**41**	2	1	4	12	14	15	3	0
		a	**a**	**g**	**t**	**C**	**A**	**A**	**G**	**G**	**T**	**c**	**a**	c	g	g	a	a



Оценка $\chi 2$ для колонки весовой матрицы

Wa, Wc, Wt, Wg – абсолютные частоты оснований в данной позиции

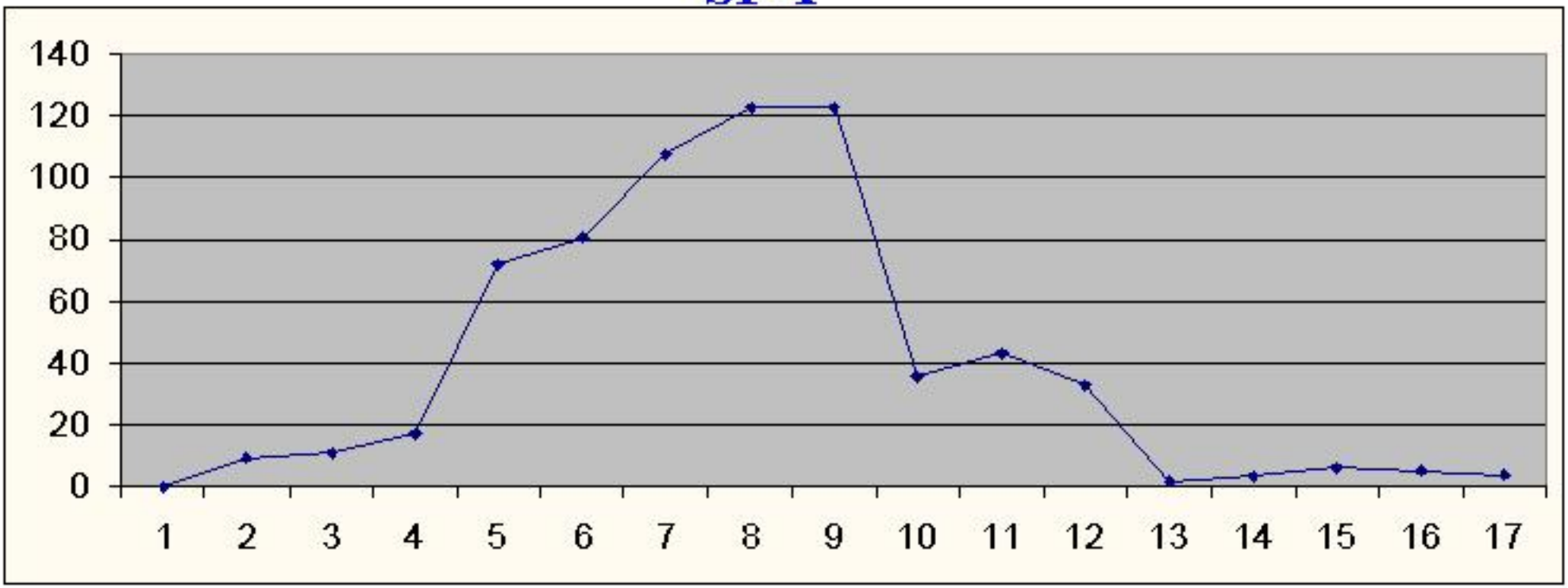
pa, pc, pt, pg – относительные частоты нуклеотидных оснований в геноме

$$N = N_a + N_c + N_t + N_g$$

$$E_a = N \times p_a$$

$$\chi^2 = \frac{(N_a - E)^2}{E_a} + \frac{(N_c - E)^2}{E_c} + \frac{(N_t - E)^2}{E_t} + \frac{(N_g - E)^2}{E_g}$$

Подбор консервативных участков в последовательностях выборки. Оценки величины хи-квадрат по матрице, построенной для выборки сайтов SF-1



	a	g	t	C	A	A	C	C	T	C	A	c	c	c	a	a

Резюме

При построении частотной матрицы для сайтов, нужно определить:

1) 1) направление каждого сайта

2) 2) число и локализацию консервативных участков в каждом сайте, выравнивание сайтов

3) 3) оптимальные фланги, окружающие консервативные районы

4) 4) Пороговое значение веса последовательности для поиска сайтов

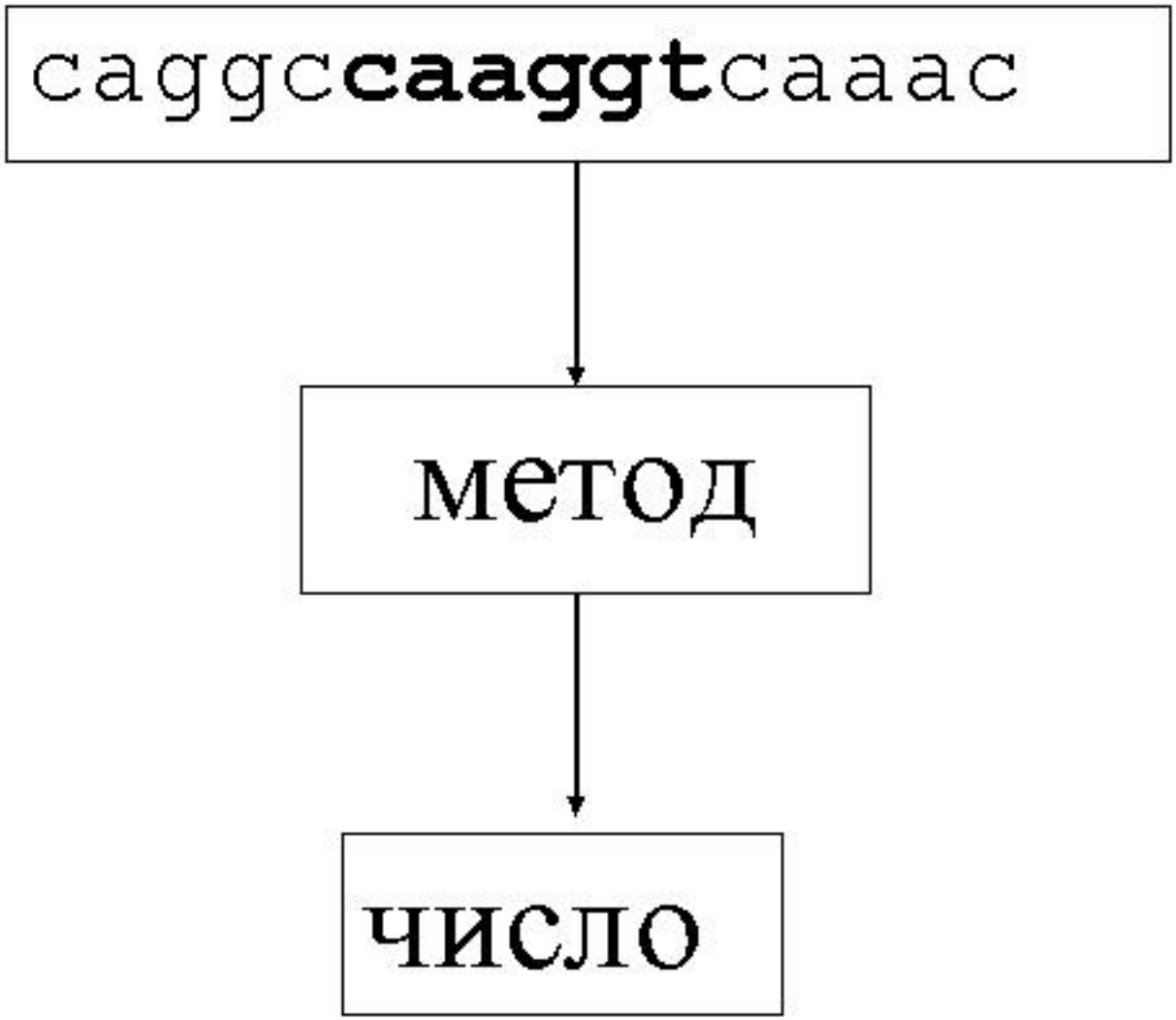
Задачи, возникающие при построении метода распознавания

1) Процедура «взвешивания» последовательности (приписания последовательности числа).

2) Получение порога для разделения позитивной выборки от негативной.

3) Оценка точности метода на контроле

Построение метода распознавания сайтов



Оценка веса последовательности с помощью весовой матрицы

A	w11	w12	w13	w14	w15	w16	w17	w18
C	w21	w22	w23	w24	w25	w26	w27	w28
G	w31	w32	w33	w34	w35	w36	w37	w38
T	w41	w42	w43	w44	w45	w46	w47	w48

Весовая Матрица

C	***A***	***A***	***G***	***G***	***C***	***C***	***G***

Тестовая последовательность

Вес тестовой последовательности равен:

$$W(X) = \sum_{j=1,..L} W(a_j, j)$$

где aj – нуклеотидное основание тестовой последовательности в позиции j.

Способы построения весовой матрицы

N_{ij} Матрица абсолютных частот, i=1,2,3,4 – номера нуклеотидных оснований, j=1,2,3,…L, где L – длина матрицы

$F_{ij} = \frac{N_{ij}}{N}$ Матрица относительных частот. N – число сайтов в выборке

$W_{ij} = \log \frac{F_{ij}}{P_i}$ Весовая матрица, которая максимально разделяет выборку сайтов от выборки случайных последовательностей, имеющих частоты оснований Pi i=1,2,3,4 (номера нуклеотидных оснований)

$W_{ij} = F_{ij} \times \log \frac{F_{ij}}{P_i}$ Информационная матрица

$W_{ij} = \log \frac{F_{ij}}{F_{\max j}}$ Матрица дискриминантной энергии Берга-вон Хиппеля. Максимальная энергия сайта равна нулю. Сайты с отклонением от консенсусной последовательности имеют энергию меньше нуля.

Построение метода распознавания сайтов

Негативная выборка :

tcccagtcgat

acagtcgtagc

gggtcgtcgaa

ggtacgaacga

acagtgctgca

Позитивная выборка:

taccaaggtca

agacaaggtca

ggacaaggtca

ggccaaggtca

agacaaggtca

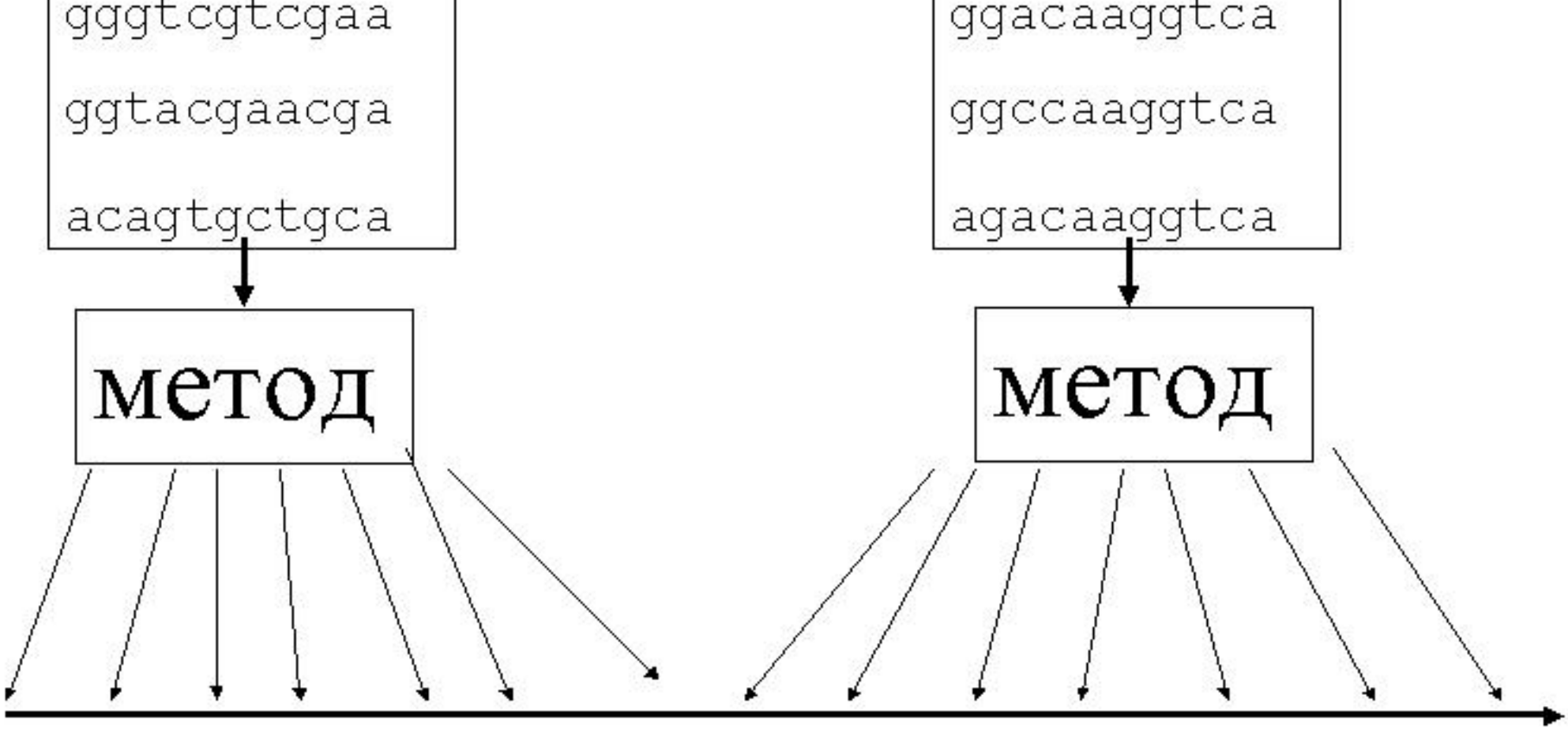
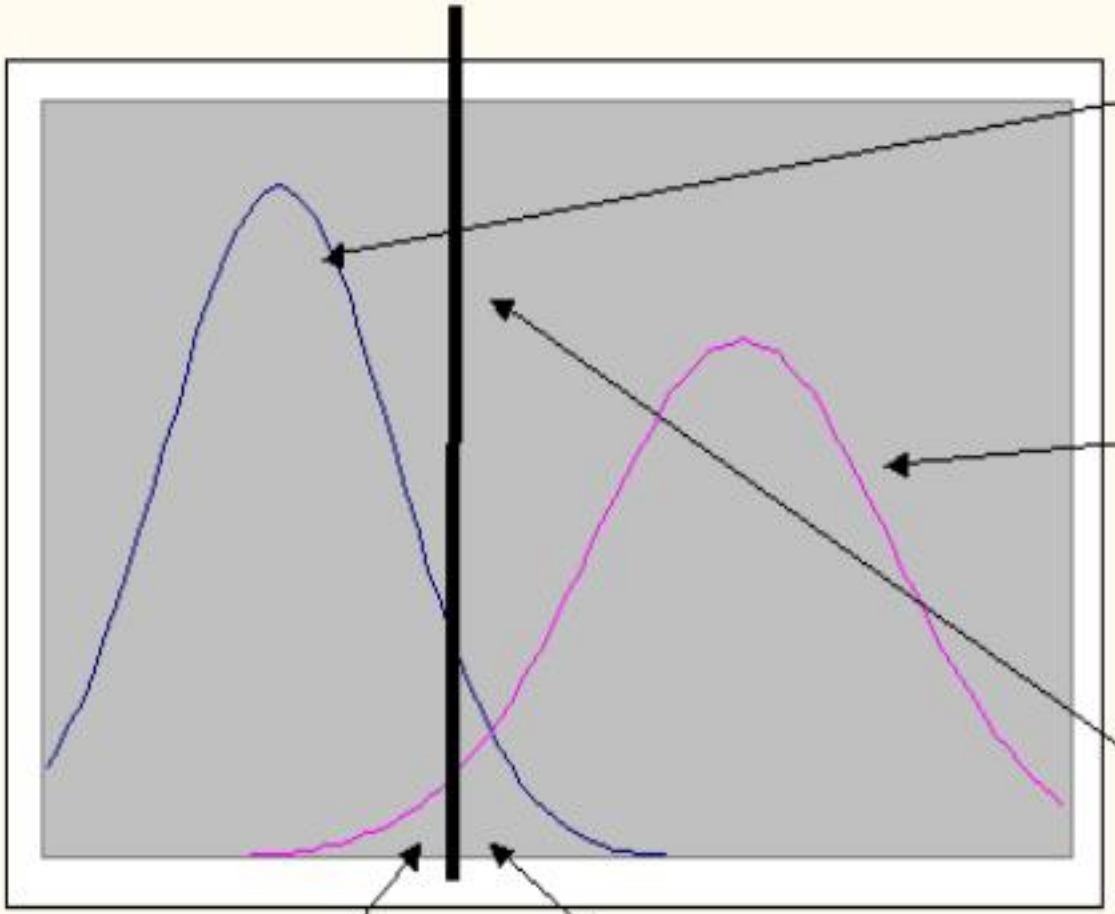


Схема разделения распределения весов сайтов от распределения весов случайных последовательностей



Распределение весов W_{Rnd} для случайных последовательностей

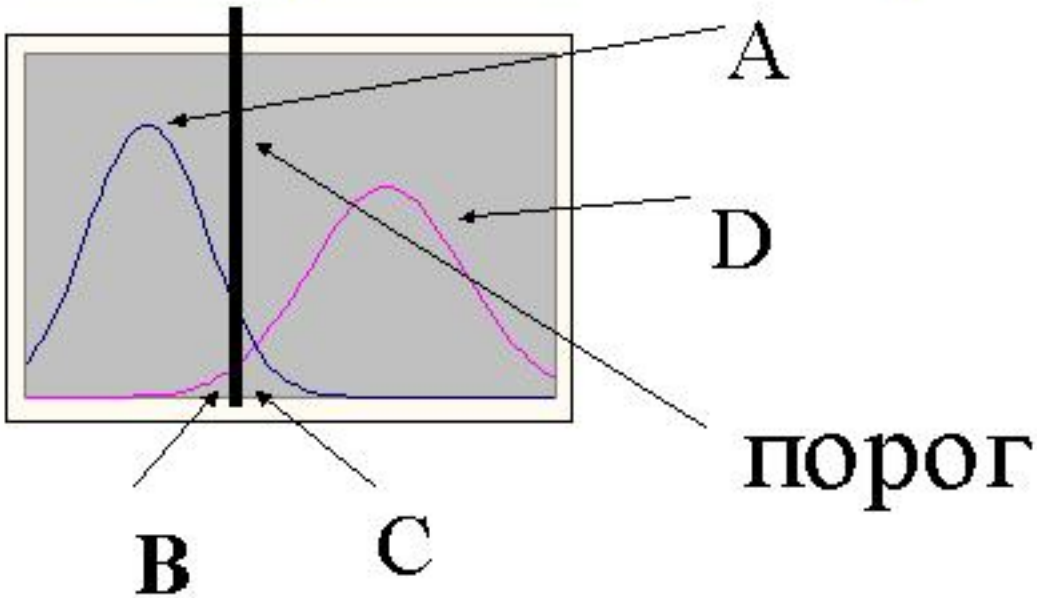
Распределение весов W_{site} реальных сайтов

Значение веса W, выбранное для разделения распределений W_{site} и W_{no}.

Площадь S_1, оценка ошибки первого рода

Площадь S_2, оценка ошибки второго рода

Оценка разделения распределения весов сайтов от распределения весов случайных последовательностей



A – число негативных последовательностей, вес которых меньше порога

B – число сайтов, вес которых меньше порога

C – число негативных последовательностей, вес которых выше порога

D – число сайтов, вес которых выше порога

$$cc = \frac{A \times D - B \times C}{\sqrt{(A+C)(A+B)(C+D)(B+D)}}$$

Некоторые способы оценки точности распознавания

$E_1 = \frac{N^-_{site}}{N_{site}}$ Оценка ошибки первого рода (доля сайтов, которые не были распознаны)

N^-_{site} - число распознанных сайтов из контрольной выборки

N_{site} - размер контрольной выборки сайтов

$E_2 = \frac{N^+_{no}}{N_{no}}$ Оценка ошибки второго рода (доля негатиных последовательностей, которые были распознаны как сайты)

N^+_{no} - число негативных последовательностей, распознанных как сайты

N_{no} - размер контрольной выборки негативных последовательностей

Некоторые способы оценки точности распознавания

		Предсказание	
		сайты	не-сайты
Реальность	сайты	TP	FN
	не-сайты	FP	TN

TP – число верно предсказанных сайтов,

TN – число верно предсказанных негативных последовательностей (не-сайтов),

FN – число неверно предсказанных сайтов,

FP – число неверно предсказанных не-сайтов.

Некоторые способы оценки точности распознавания

В этих обозначениях оценка ошибки первого рода:

$$E_1 = \frac{FN}{FN + TP}$$

Оценка ошибки второго рода:

$$E_2 = \frac{FP}{TN + FP}$$

Некоторые способы оценки точности распознавания

Чувствительность – доля правильно предсказанных сайтов:

$$Sn = \frac{TP}{TP + FN}$$

Специфичность (доля правильно отвергнутых негативных последовательностей):

$$Sp = \frac{TP}{TP + FP}$$

Другое определение специфичности (это вероятность того, что предсказанный методом сайт действительно является сайтом):

$$Sp = \frac{TP}{TP + FP}$$

Коэффициент корреляции (мера связи между предсказанными и реальными объектами, одновременно учитывающая все элементы таблицы сопряженности)

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(FN + TN)}}$$

Точность предсказания

•Функциональная аннотация последовательностей в основном рассматривает два набора данных:

•Набор последовательностей, несущих функцию

•Набор последовательностей, не выполняющих данную функцию

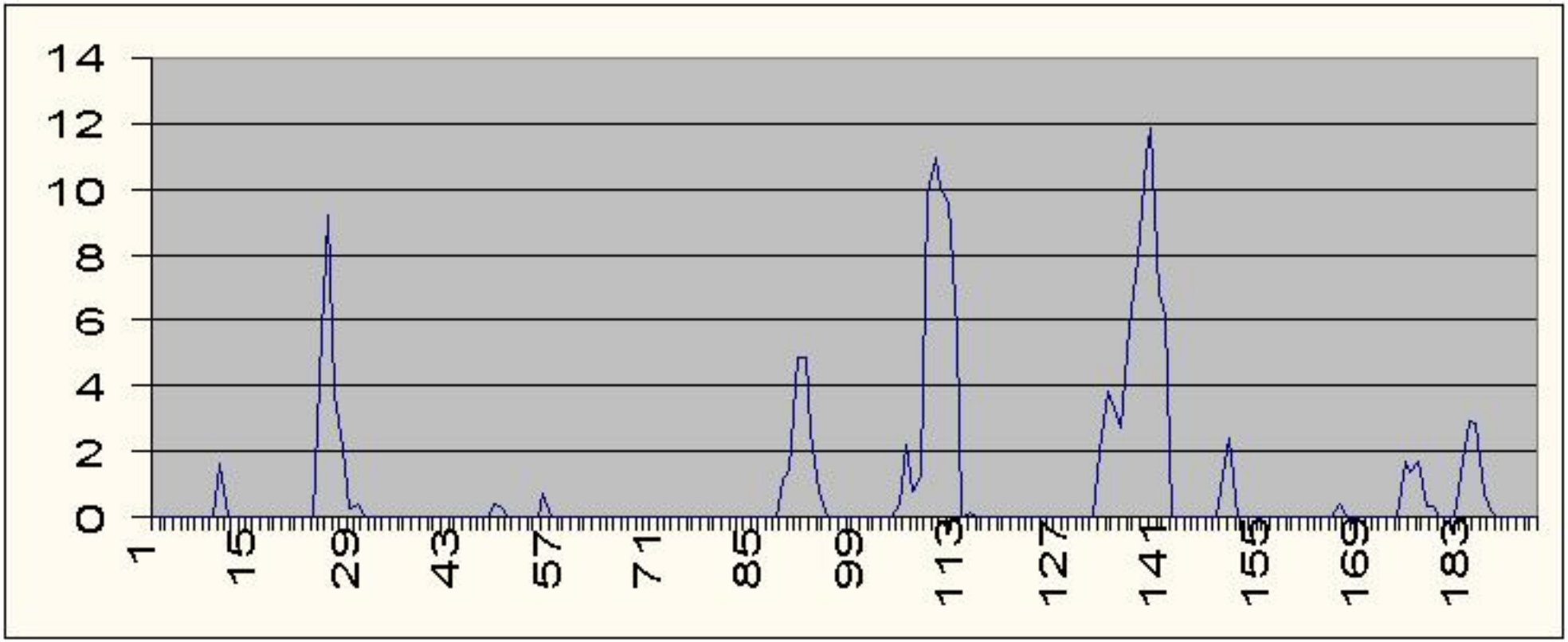
•Каждый набор последовательностей нужно разбить на обучающую и контрольную выборку

•Контрольные выборки следует использовать только для оценки точности

Проблемы, возникающие при оценке точности предсказания

- **Базы данных о нуклеотидных последовательностях сильно вырожденны (гомологи, повторы последовательностей)**
- **Нужно следить за тем, чтобы последовательности в обучающей и контрольной выборке были негомологичны**
- **Исключение гомологов сокращает размер выборки**
- **Разработан ряд критериев оценки точности, из которых следует выбрать подходящий для решения конкретной биологической задачи.**

Поиск сайтов связывания фактора SF-1 вдоль последовательности регуляторного района P00121 из TRRD



Name	Position	Strand	Site
SF-1	25	+	gcctagcgcag
SF-1	109	+	gaccagaggag
SF-1	138	+	tgtcaaggccg

Следующий шаг после анализа последовательности

Более полное понимание функции гена

Протеомика, анализ экспрессии, структурная геномика, белок-белковые взаимодействия

Изучаемые гены, расположение в геноме

Выявление регуляторных раионов генов, повторов, псевдогенов и др.