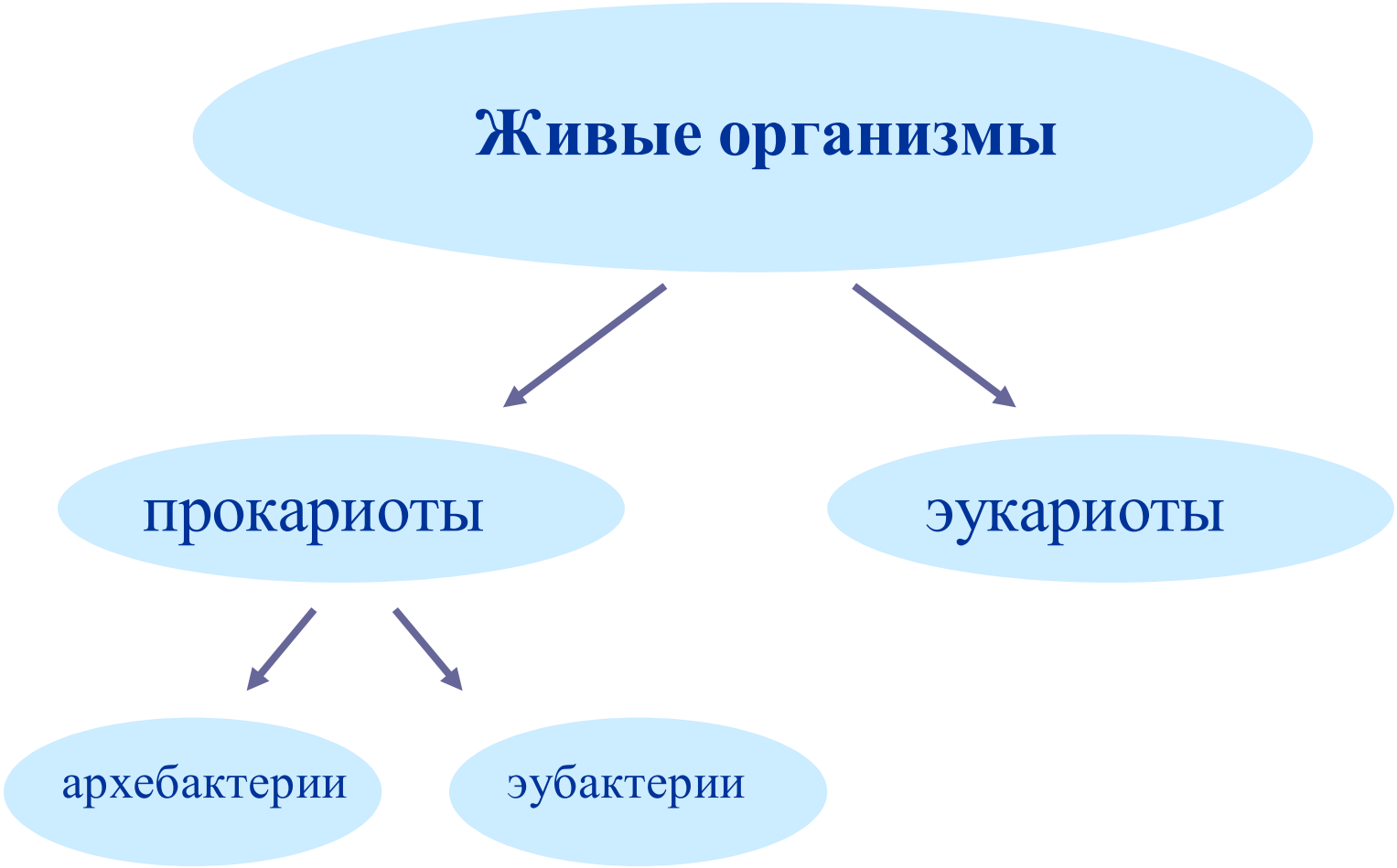
Контекстный анализ и распознавание сайтов связывания транскрипционных факторов

Поздняков М.А.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Некоторые сведения из систематики



Основные отличия эукариот от прокариот

1) наличие ядра

2) наличие других компартментов клетки: вакуоли, комплекс Гольджи, лизосомы, митохондрии и др.

3) Наличие окислительного фосфорилирования (дыхательной цепи).

4) Отличия в клеточной границе

(1) наличие клеточной стенки (у растений),

(2) отсутствие муреина и других составляющих, свойственных бактериям

5) Молекулы ДНК линейные (у прокариот - кольцевые)

6) Отсутствие плазмид как средства обмена информацией

7) Диплоидность ДНК

8) Значительно более богатый арсенал репарации ДНК.

Основные отличия эукариот от прокариот (продолжение)

9) Прерывистая структура генов.

10) Значительно бо'льшие размеры геномов

11) Намного бо'льшая доля некодирующей ДНК

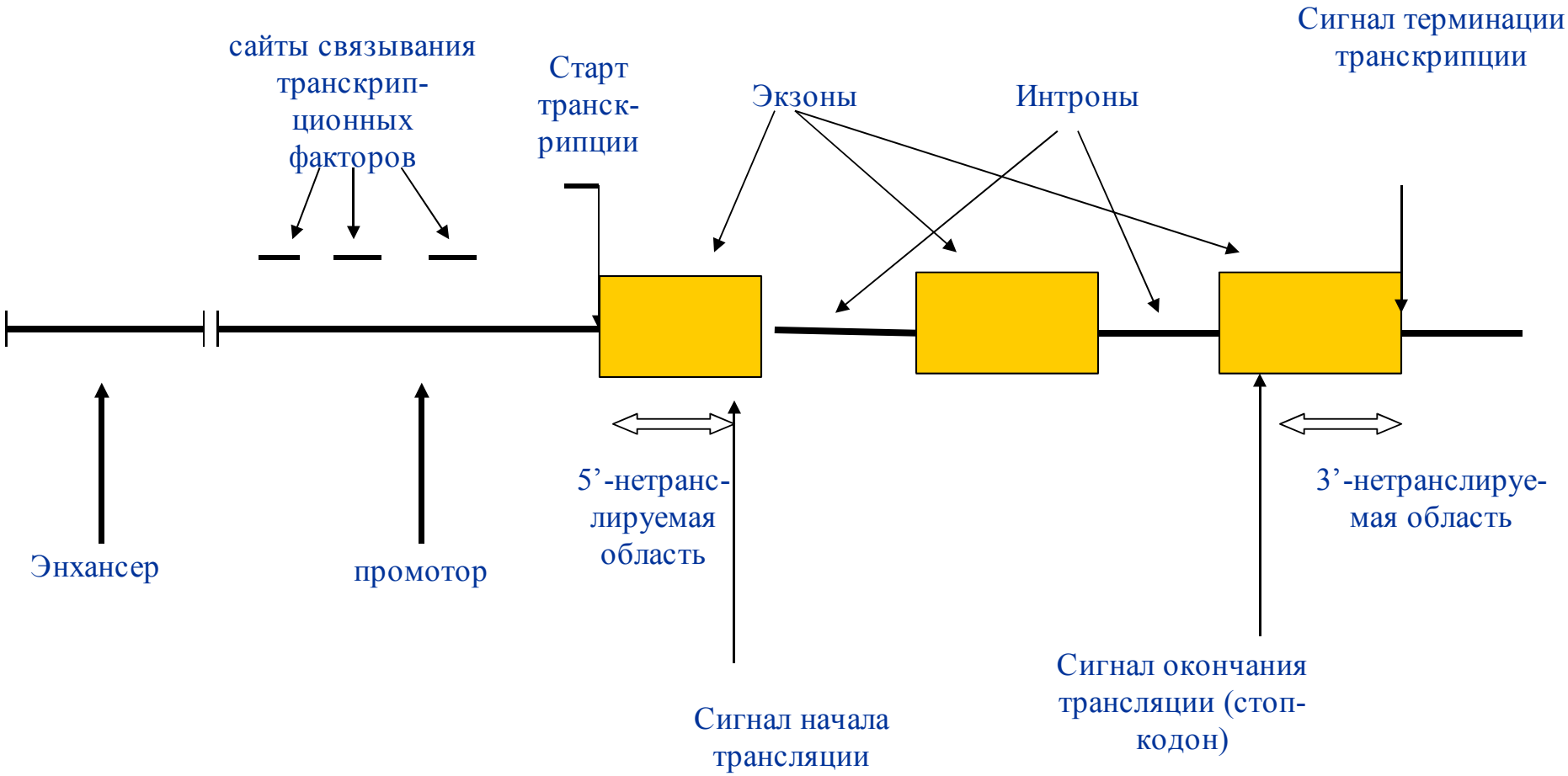
12) Значительно больше повторяющихся последовательностей:
(а) генов - кластеры изофункциональных генов;
(б) повторы в некодирующих областях геномов.

13) Значительно бо'льшая упаковка ДНК

14) Есть многоклеточные организмы. Дифференцировка клеток по функциям.

15) Особенности экспрессии эукариотического гена.
Значительно более сложная регуляция уровня экспрессии генов.

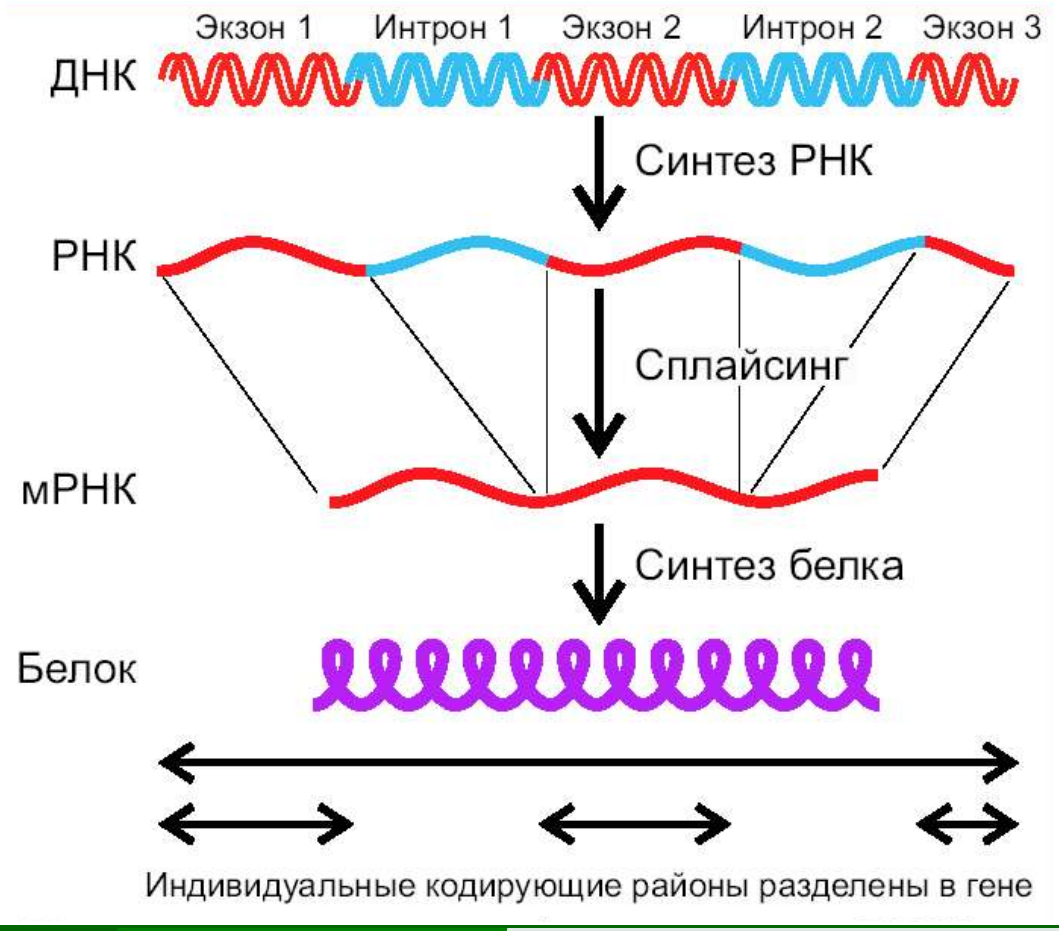
Структура эукариотического гена



Основные этапы экспрессии эукариотического гена

Физико-химическое содержание: ***вырезание фрагментов из пре-мРНК.***

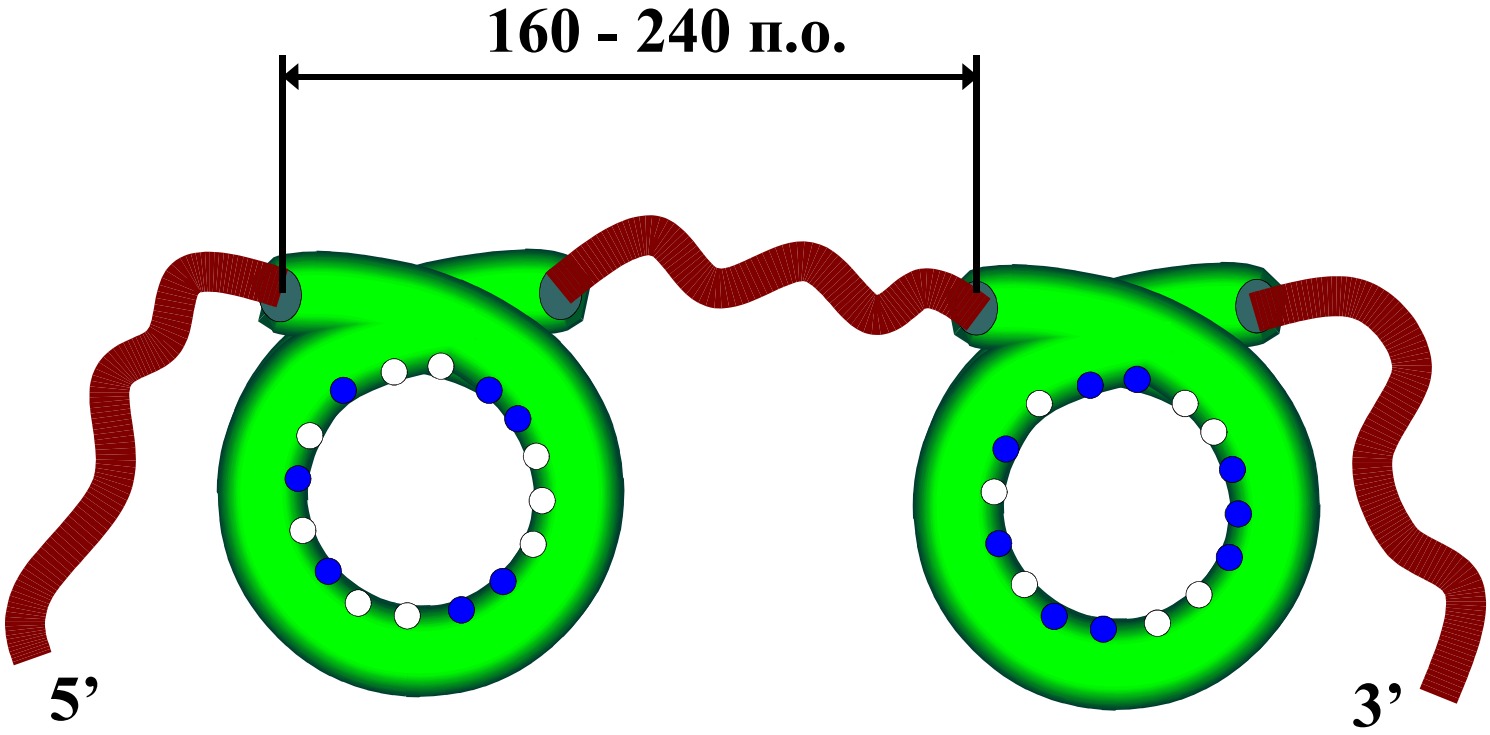
Информационное содержание: ***удаление фрагментов, не несущих генетической информации из генетического РНК-текста***



Размеры геномов некоторых организмов

Организм	Примерный размер гаплоидного генома
E.coli	3×10^{7}
Миксомицеты	7×10^{7}
Трипаносома	8×10^{7}
Нематода	8×10^{7}
Резушник Талля	7×10^{7}
Шелкопряд	5×10^{8}
Плодовая мушка	1.7×10^{8}
Морской еж	8×10^{8}
Шпорцевая лягушка	3×10^{9}
Протей	5×10^{10}
Курица	1.2×10^{9}
Мышь	3×10^{9}
Корова	3.1×10^{9}
Человек	2.9×10^{9}
Кукуруза	5×10^{9}
Лук	1.5×10^{10}

Первый уровень укладки хроматина: нуклеосома



Нуклеосома состоит из октамера – восьми гистоновых белков, и ДНК, делающей вокруг этого октамера полтора оборота (примерно 150 п.о.) Размер свободного участка между соседними нуклеосомами составляет около 50 п.о.

Укладка хроматина

Рентгеноструктурный анализ структуры нуклеосомы

Luger K. et al., Nature 1997, 389, 251-260

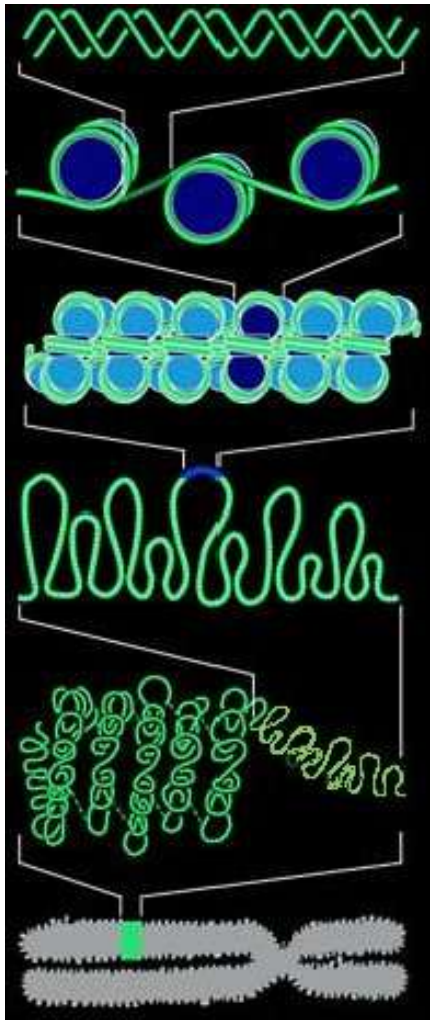
Двойная спираль ДНК

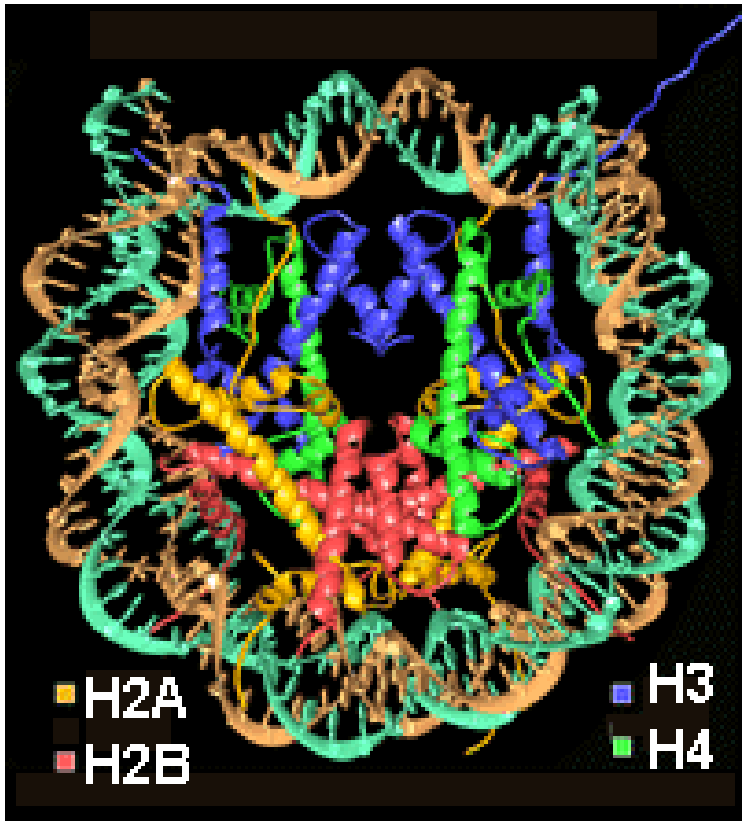
Хроматиновая нить “бусины на нити”

30 нм хроматиновая нить

Высшие порядки укладки хроматина

Хромосома





Нуклеосомная ДНК имеет форму соленоида из ~1.8 витков и длину ~ 146 пар оснований.

Механизмы регуляция уровня экспрессии гена на всех этапах экспрессии гена

Упаковка ДНК	1) Гены в сильноупакованных участках генома не экспрессируются или экспрессируются слабо. 2) Играют роль также химические модификации ДНК (матилирование), 3) вероятность посадки нуклеосомы, 4) химические модификации белков нуклеосомы, др.
транскрипция	Сложный аппарат регуляции транскрипции
трансляция	1) Время жизни РНК 2) Использование кодонов, которым соответствуют часто- или редко-встречающиеся tРНК в цитоплазме клетки, др.

Схематичное расположение генов на ДНК

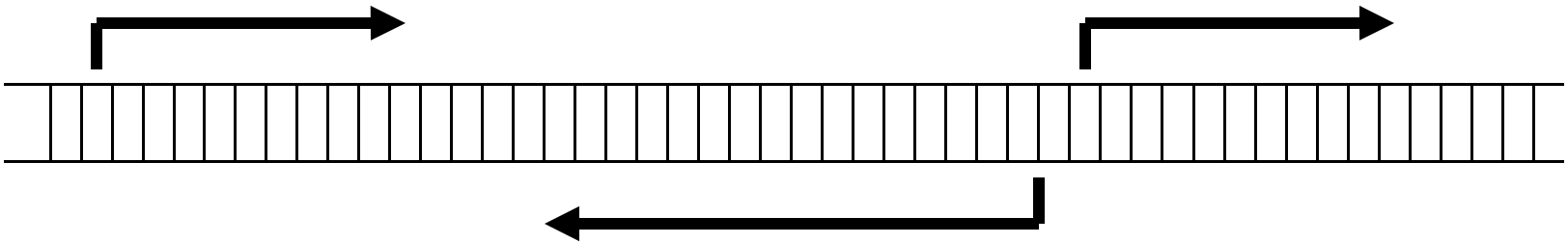


Схема строения оперона прокариот

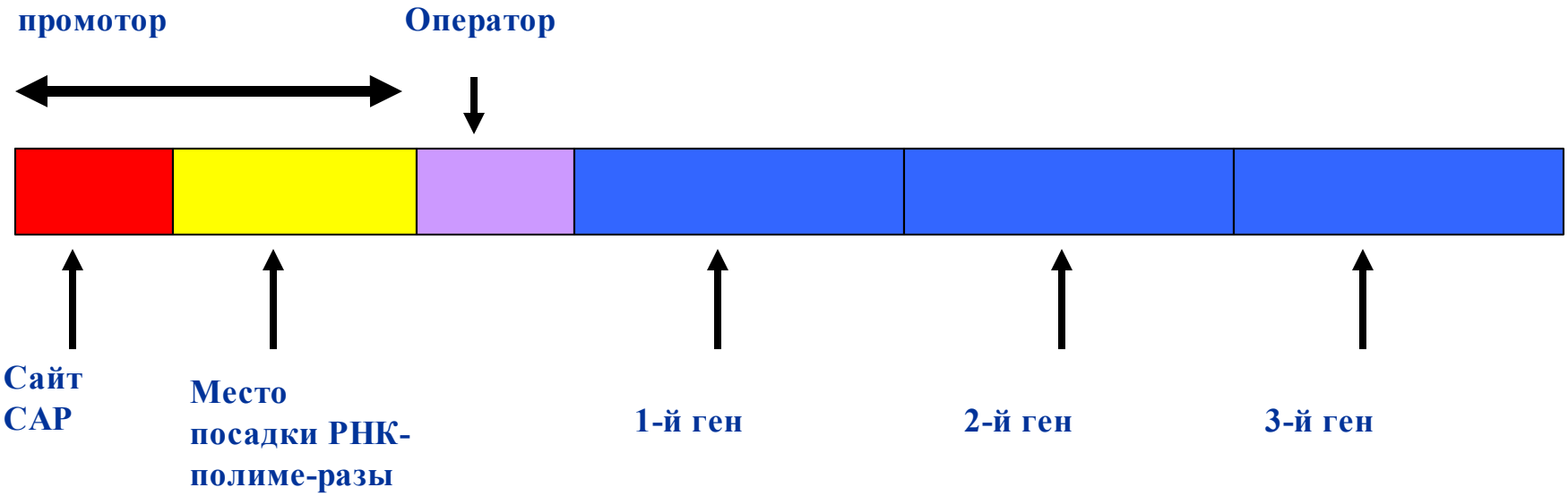
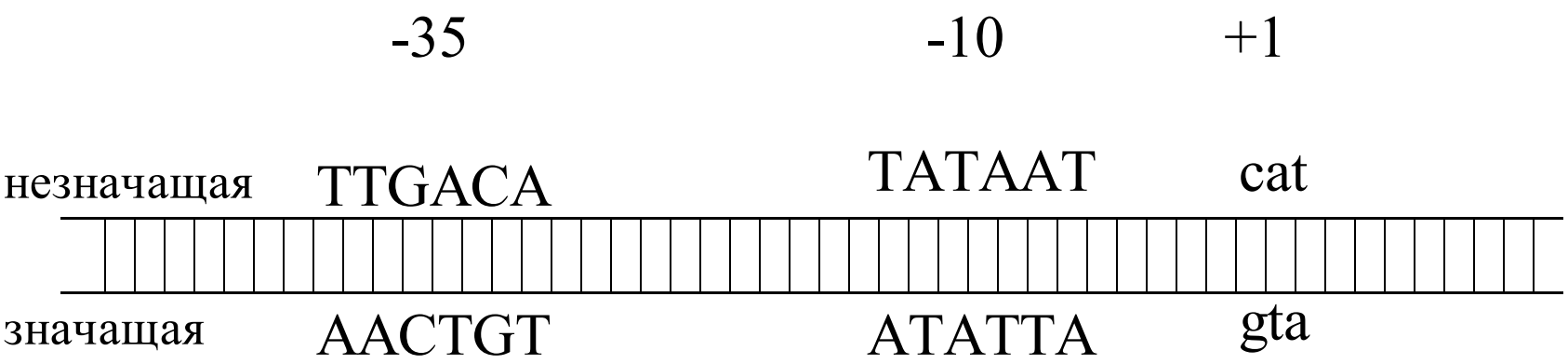


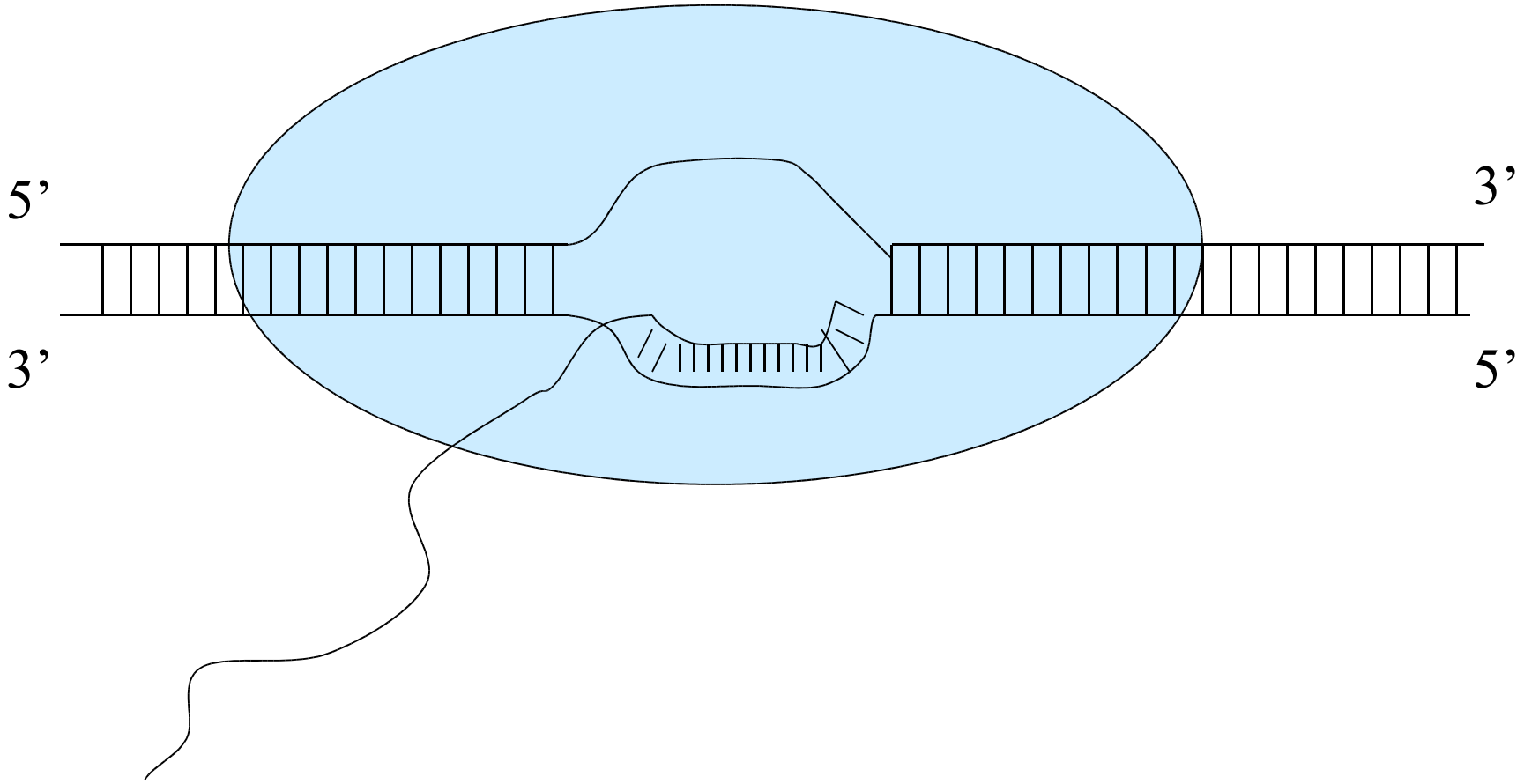
Схема строения промотора прокариот



Стадии транскрипции

1) связывание РНК-полимеразы с ДНК
2) инициация цепи РНК
3) рост (элонгация) цепи РНК
4) терминация цепи РНК

РНК-полимераза, транскрибирующая ДНК



Особенности транскрипции у эукариот

1) Транскрипцию осуществляет 3 разные РНК-полимеразы.

2) РНК-полимераза эукариот не может самостоятельно инициировать транскрипцию. Для этого нужно большое число белков

3) Многие регуляторные белки у эукариот могут влиять на скорость транскрипции

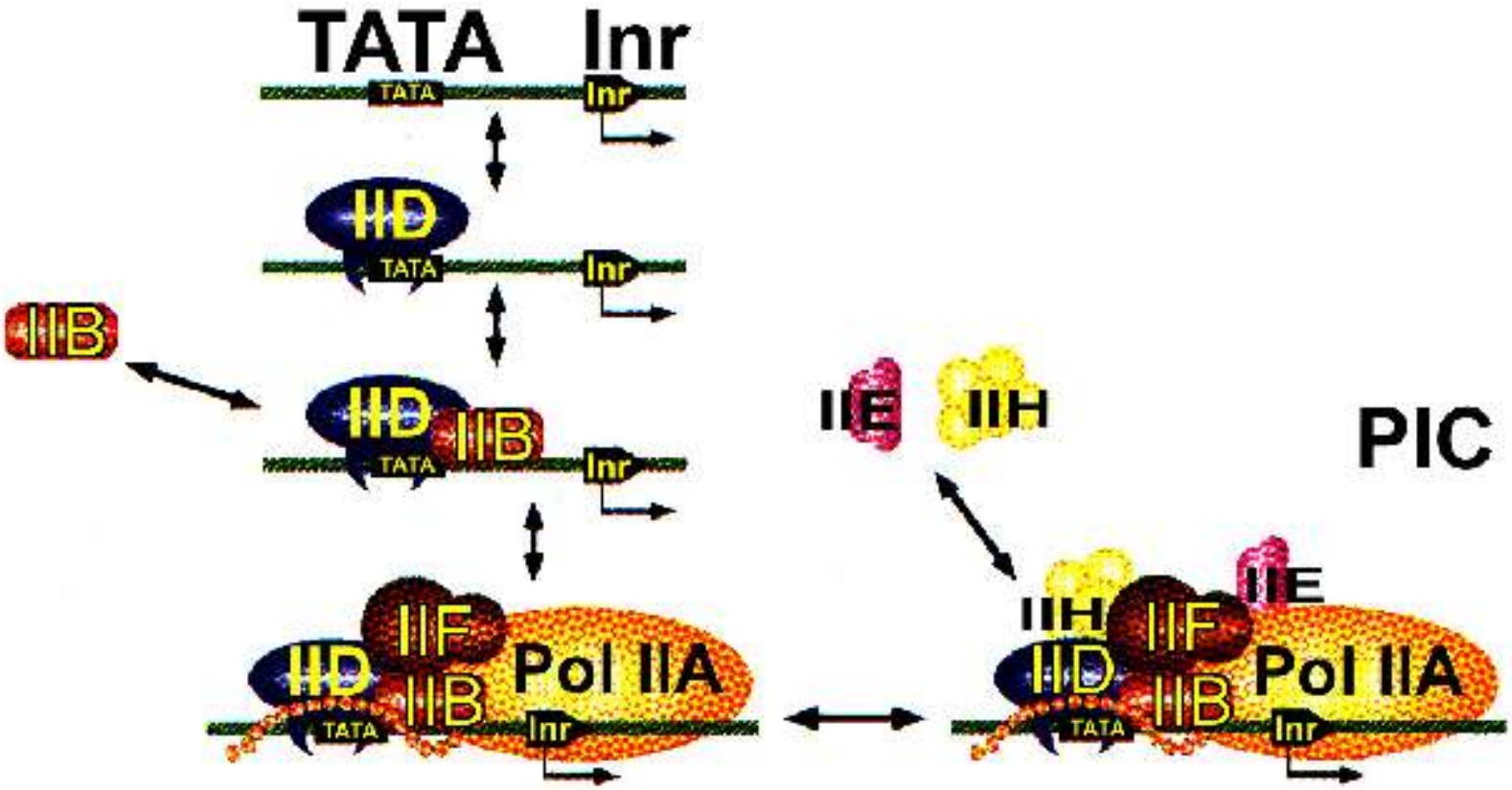
РНК-полимеразы эукариот

1. РНК-полимераза I : гены рибосомных РНК (5,8S рРНК, 18S рРНК и 28S рРНК).

2. РНК-полимераза II : гены, кодирующие белки, а также малых ядерных РНК (за исключением гена U6).

3. РНК-полимераза III : гены тРНК, 5S рРНК, 7SL РНК и U6 РНК.

Схема сборки и функционирования базального транскрипционного комплекса

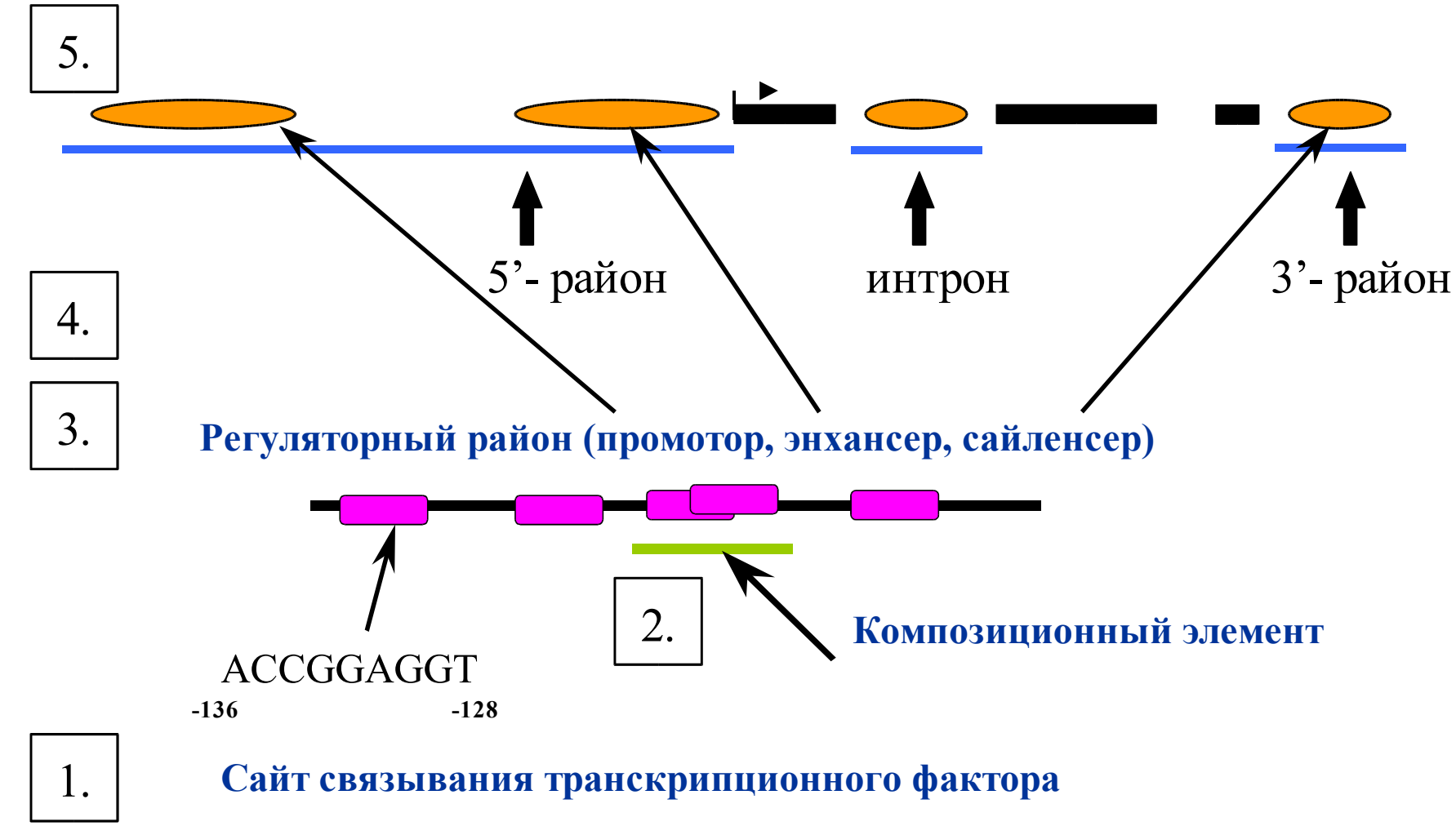


(Nikolov and Burley, 1997)

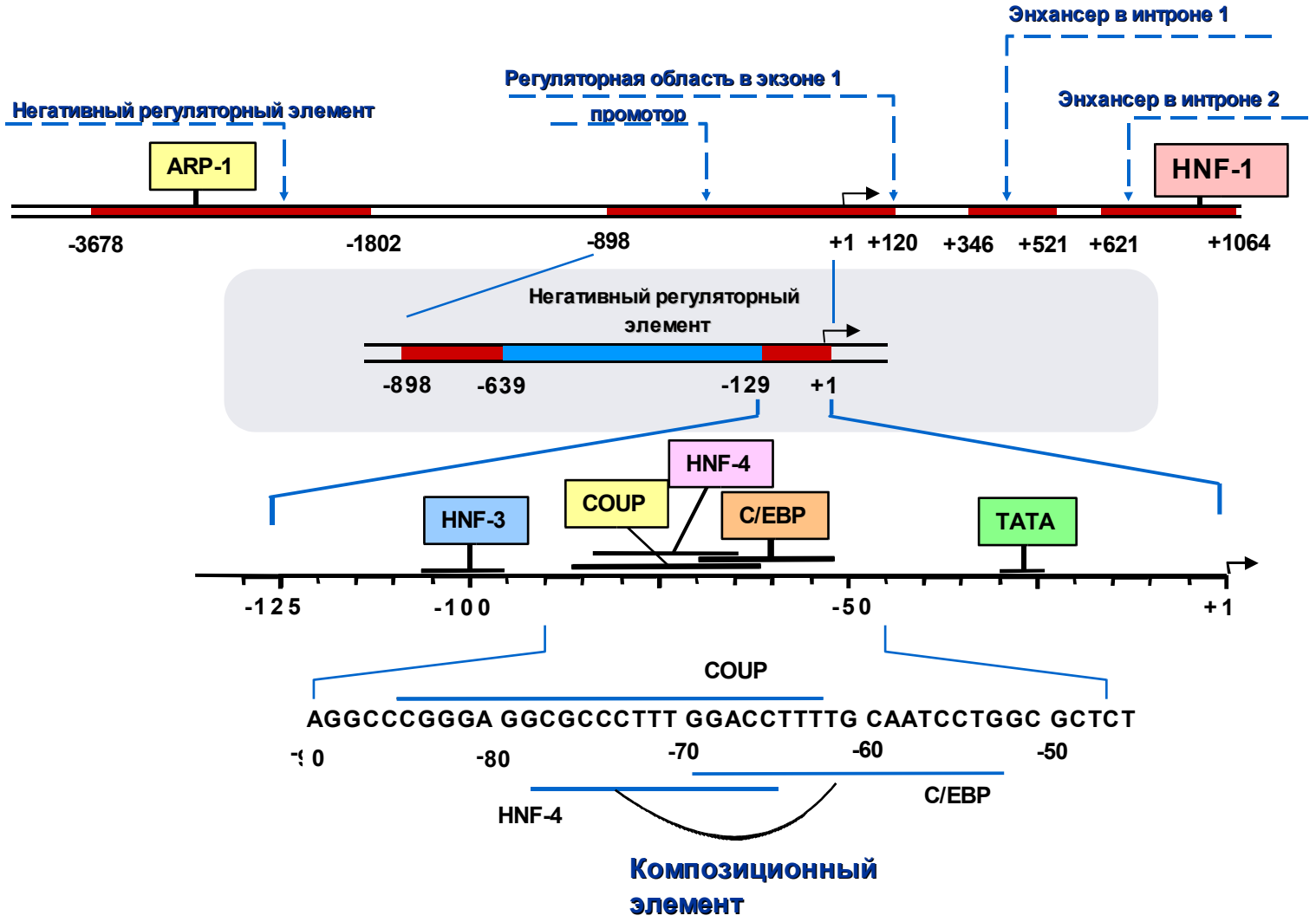
Основные элементы промоторов эукариот

элемент	консенсус	локализация
TATA-бокс	TATAAAAA STWTAWADRSSSSSS	-35
CAAT-бокс	GGCCAATCT	-212 .. –57
GC-бокс	GGCCGG	-164 .. +1
Inr-элемент	$YYA_{(+1)}NWYY$	Содержит старт
DPE-элемент (drosophila)	RGWCGTG	+30

Схема района регуляции транскрипции эукариотического гена

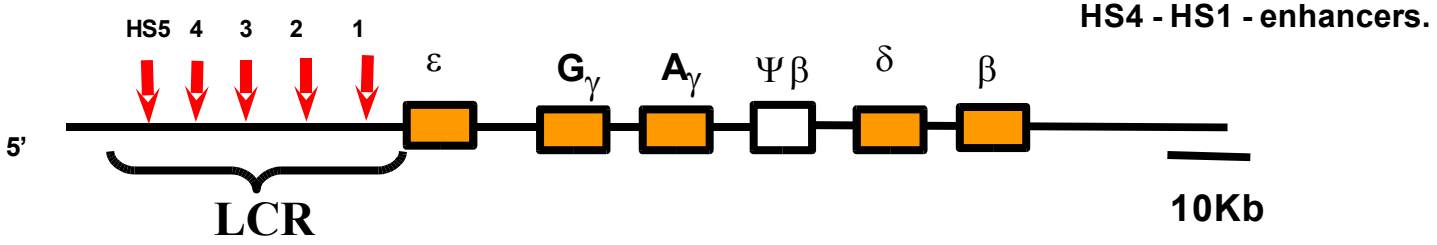


База данных TRRD: организация регуляторных районов гена человека, кодирующего белок аполипопротеин В



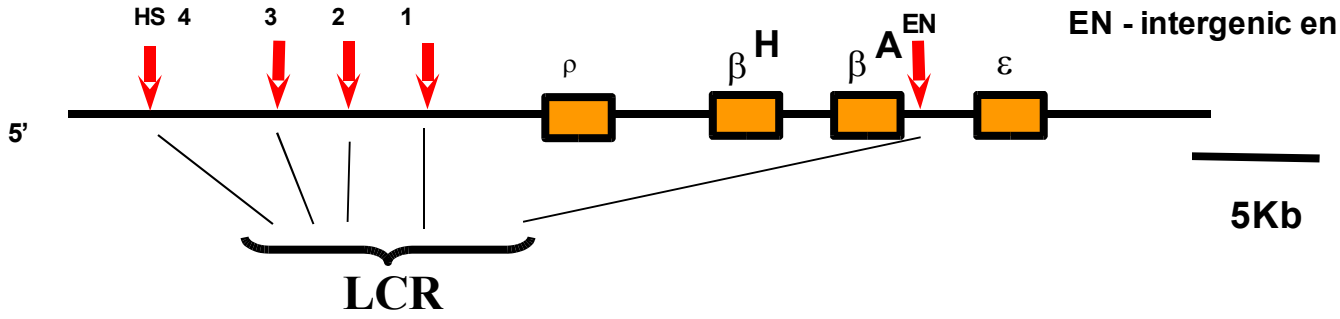
Локус контролирующие районы обеспечивают координированную ткане- и стадиеспецифичную экспрессию генов

HUMAN β-GLOBIN GENE CLUSTER



HS5 - insulator;
HS4 - HS1 - enhancers.

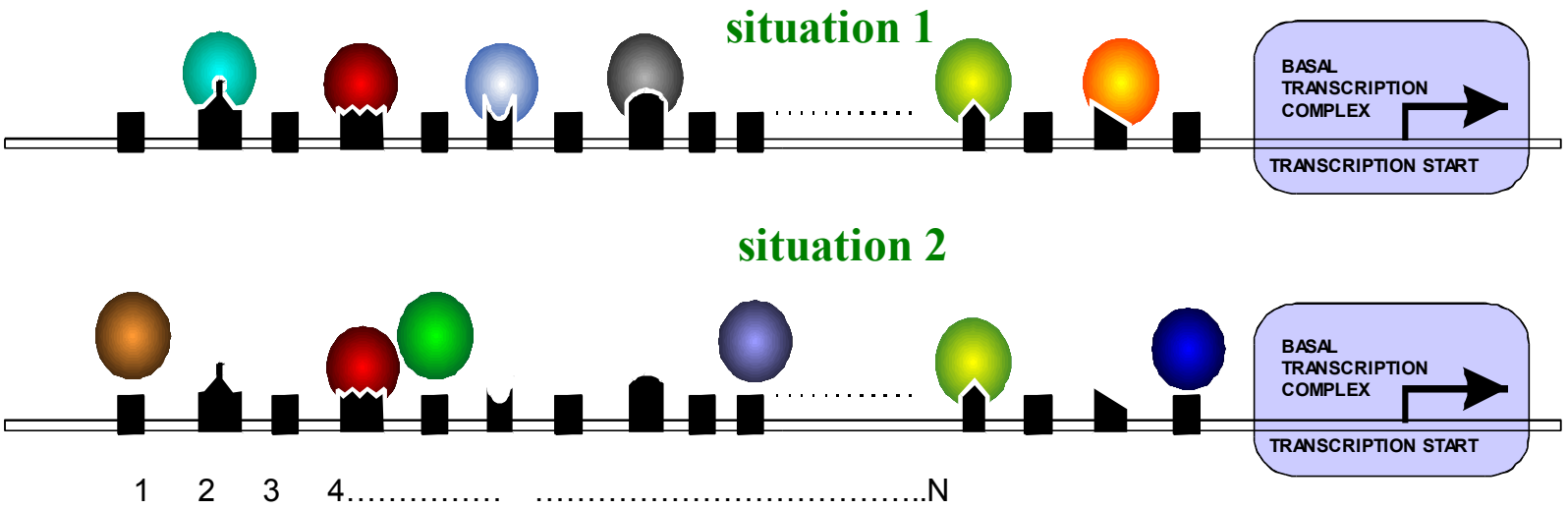
CHICKEN β-GLOBIN GENE CLUSTER



HS4 - insulator;
HS3, HS2 - enhancers
EN - intergenic enhancer

Codes of transcription regulation Estimation of informational capacity of transcription regulation code

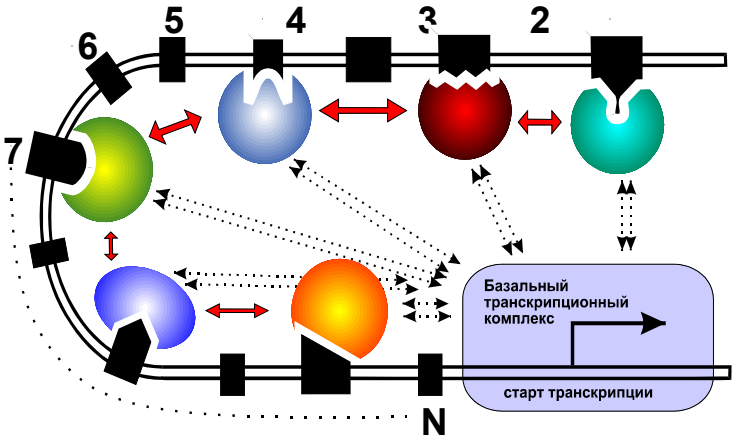
In a multi-cell organism, one and the same gene has different expression patterns in various conditions (depending upon the cell cycle stage; the type of a cell, tissue, organ, stage of development, action of inductors, environmental conditions, etc.)

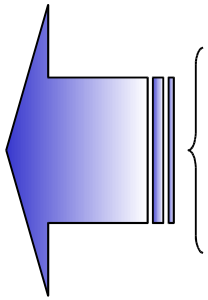


Capacity W of transcription regulation code (underestimates)
In the context of the absence of the protein-protein interactions between transcription factors $W = 2^N$; where $N=20$ $W = 10^6$;

Код регуляции транскрипции

Ген-специфический **транскрипционный комплекс**





1 Энхансеры

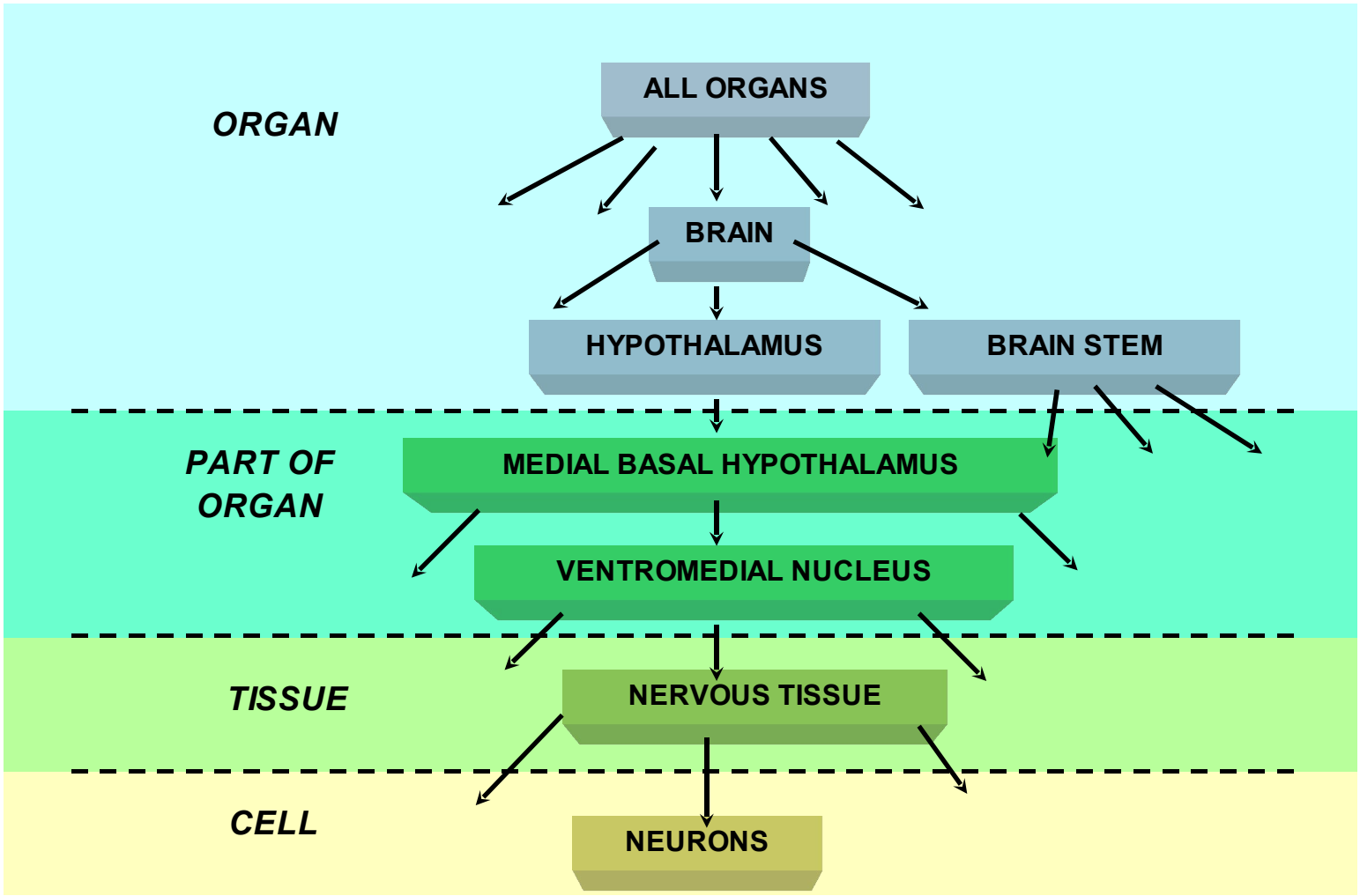
2 Сайленсеры

Емкость кода регуляции транскрипции:

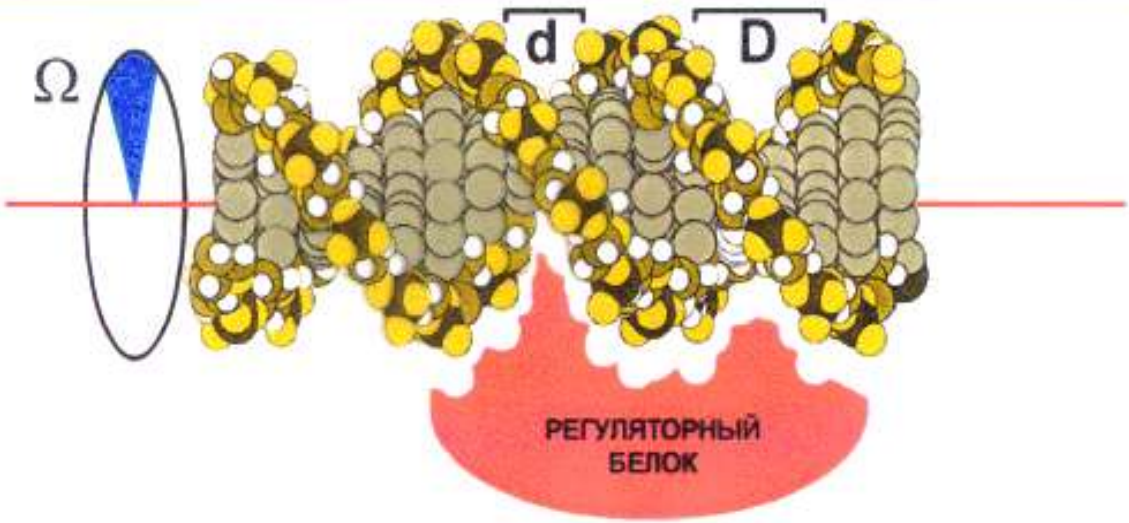
$$W = \sum_{n=1}^{N} C_N^n \cdot 2^{C_n^2}$$ **- число возможных вариантов комплекса**

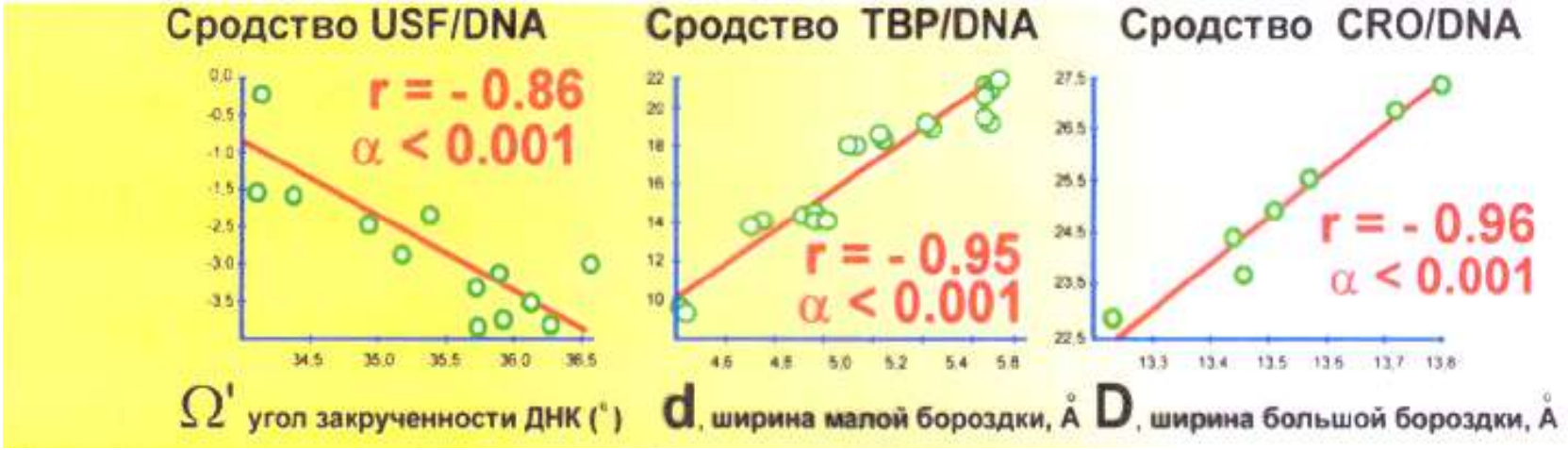
(при N= 20 сайтам, W=10^{30})

Hierarchical organization of controlled vocabularies of morphological terms in the trrd database

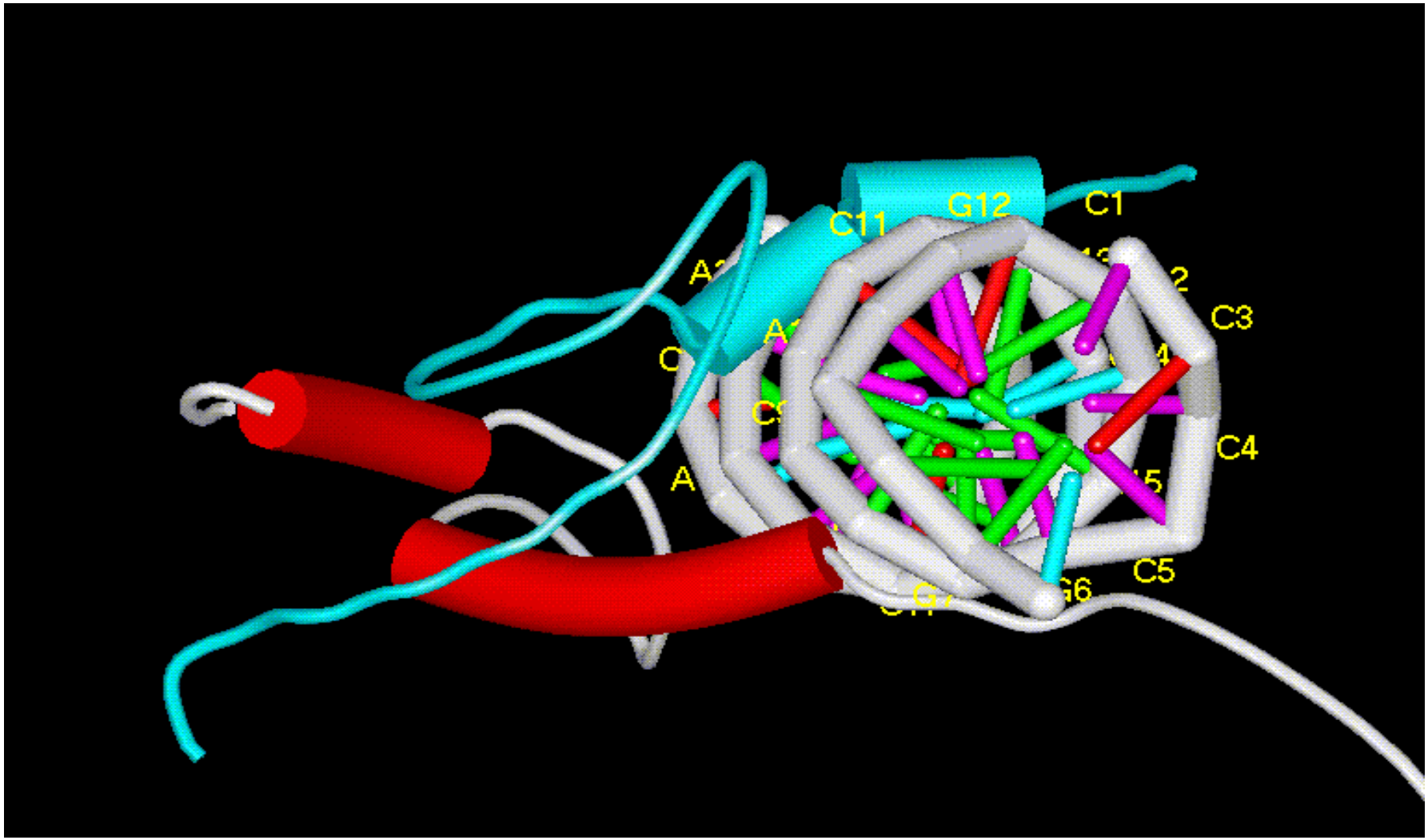


Сродство регуляторных белков к сайтам их связывания определяется конформационными свойствами днк



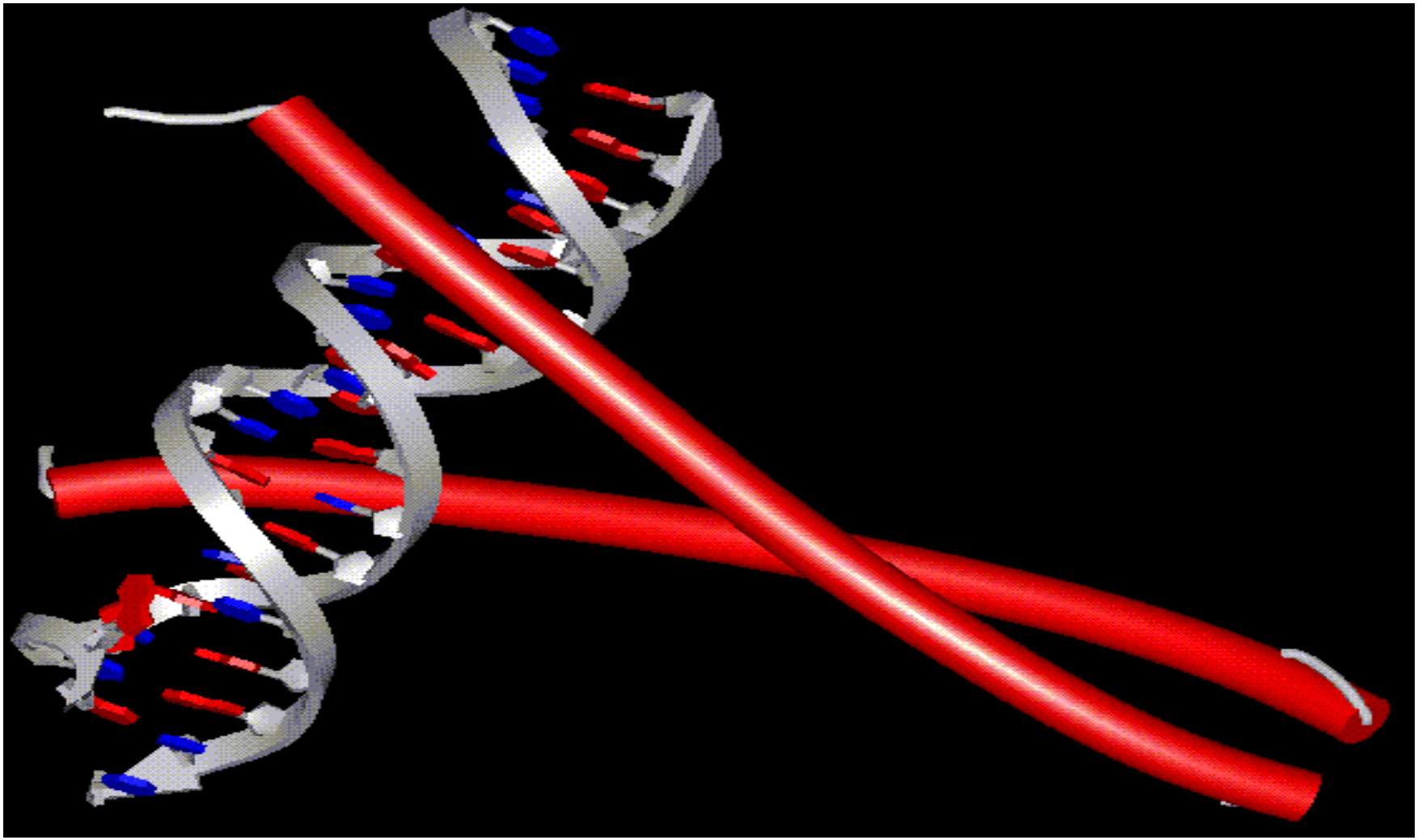


Трехмерная структура днк-белкового комплекса



http://wwwmgs.bionet.nsc.ru/mgs/programs/acts2/images/USF_X.html

Трехмерная структура ДНК-белкового комплекса (bZIP)



Пространственное строение двойной спирали ДНК

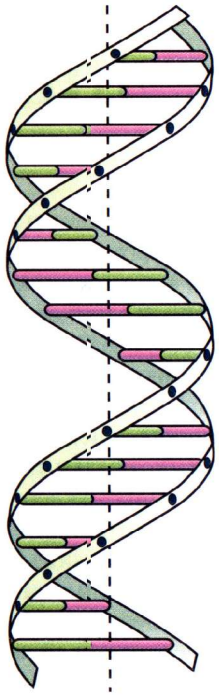
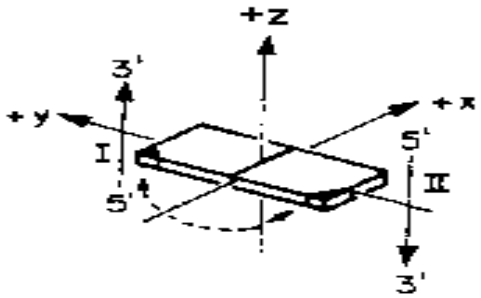
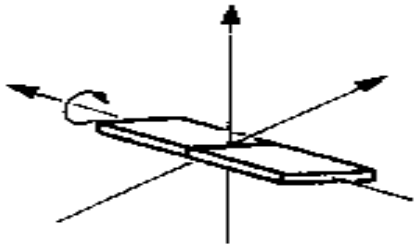


Рис. 3.3. Модель двойной спирали ДНК. Поперечные перекладины – комплементарные пары оснований, «боковины» – сахарофосфатный остов.

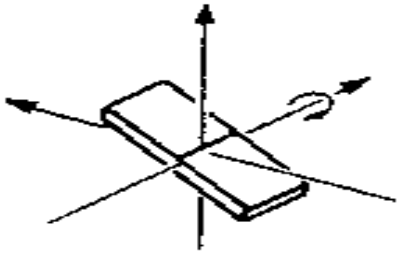
Локальные конформационные параметры двойной спирали ДНК



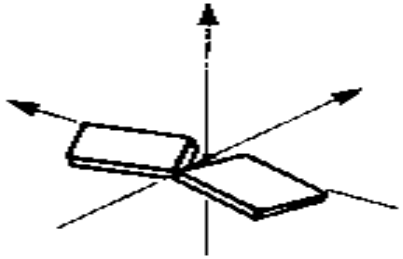
Coordinate frame



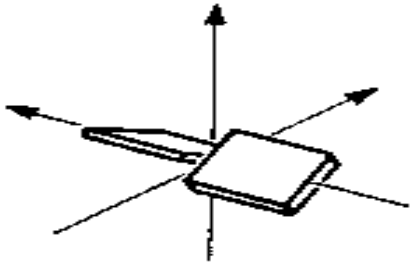
Tip (θ)



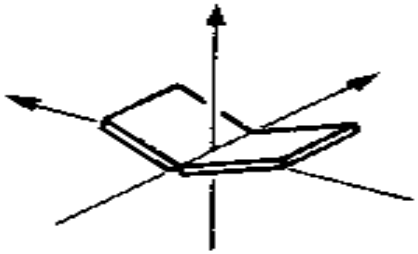
Inclination(η)



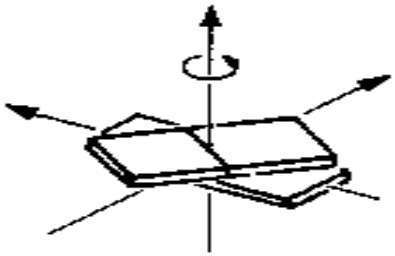
Opening (σ)



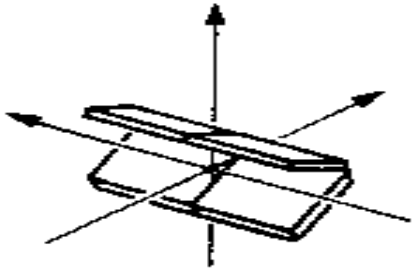
Propeller twist(ω)



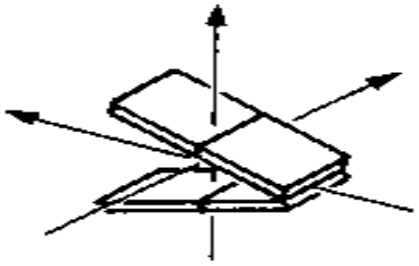
κ)Buckle(



Twist(Ω)



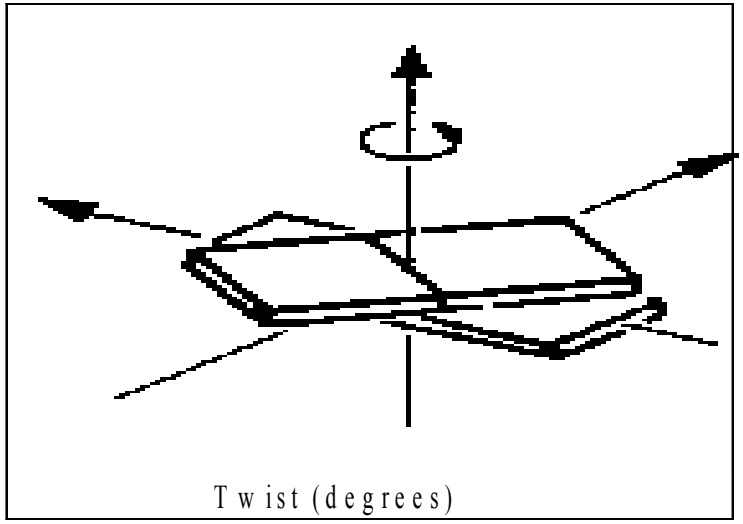
Roll(ρ)



Tilt(τ)

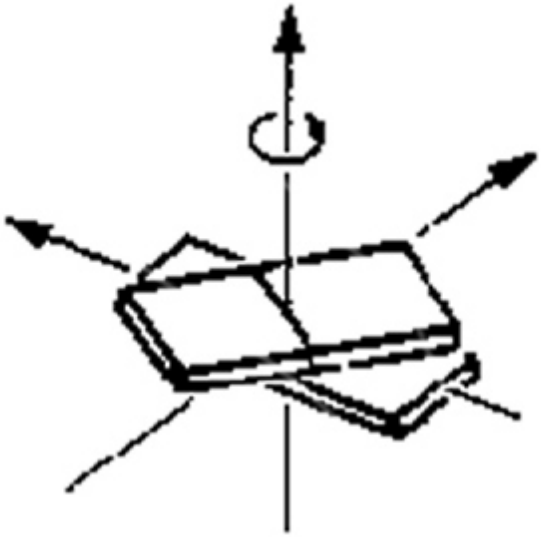
Описание угла twist в компьютерной системе B-DNA-VIDEO

```
MT  PROPERTY COMPILATION "ACTIVITY"
//
MN  Conformational
MD  B-DNA
ML  dinucleotide step
//
RN  [1]
RA  Suzuki M, Yagi N, Finch JT
RT  Role of base-backbone and base-base interactions in
RT  alternatingRT DNA conformations.
RJ  FEBS Lett (1996) 379:148-152
//
PN  Helical twist
//
PU  degrees
AA  35.6**
AT  29.3
AG  31.9
AC  31.1
TA  39.5
TT  35.6
TG  36.0
TC  35.9
GA  35.9
GT  31.1**
GG  33.3
GC  34.6
CA  35.9
CT  31.9
CG  34.9
CC  33.3
```



Описание зависимости значений угла TWIST от динуклеотидного контекста в базе знаний

```
MI P0000001
MN Conformational
MD B-DNA
ML dinucleotide step
PN Twist
PM Calculated by Sklenar,
PM and averaged by
Ponomarenko
PV TwistCalc
PU Degree
```

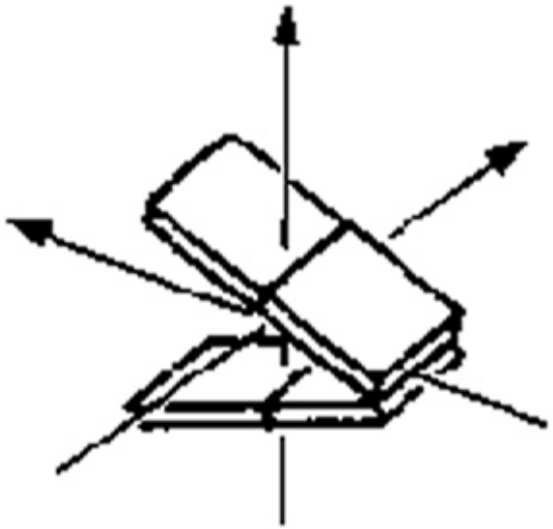


Twist (Ω)

```
DINUCLEOTIDE
   AA 38.90
   AT 33.81
   AG 32.15
   AC 31.12 **
   TA 33.28
   TT 38.90
   TG 41.41 *
   TC 41.31
   GA 41.31
   GT 31.12 **
   GG 34.96
   GC 38.50
   CA 41.41 *
   CT 32.15
   CG 32.91
   CC 34.96
//
```

Описание зависимости угла TILT от динуклеотидного контекста в базе знаний

```
MI P0000016
MN Conformational
MD DNA/protein-complex
ML dinucleotide step
PN Tilt
PM Averaged for X-rays
PV TiltCompl
PU Degree
```



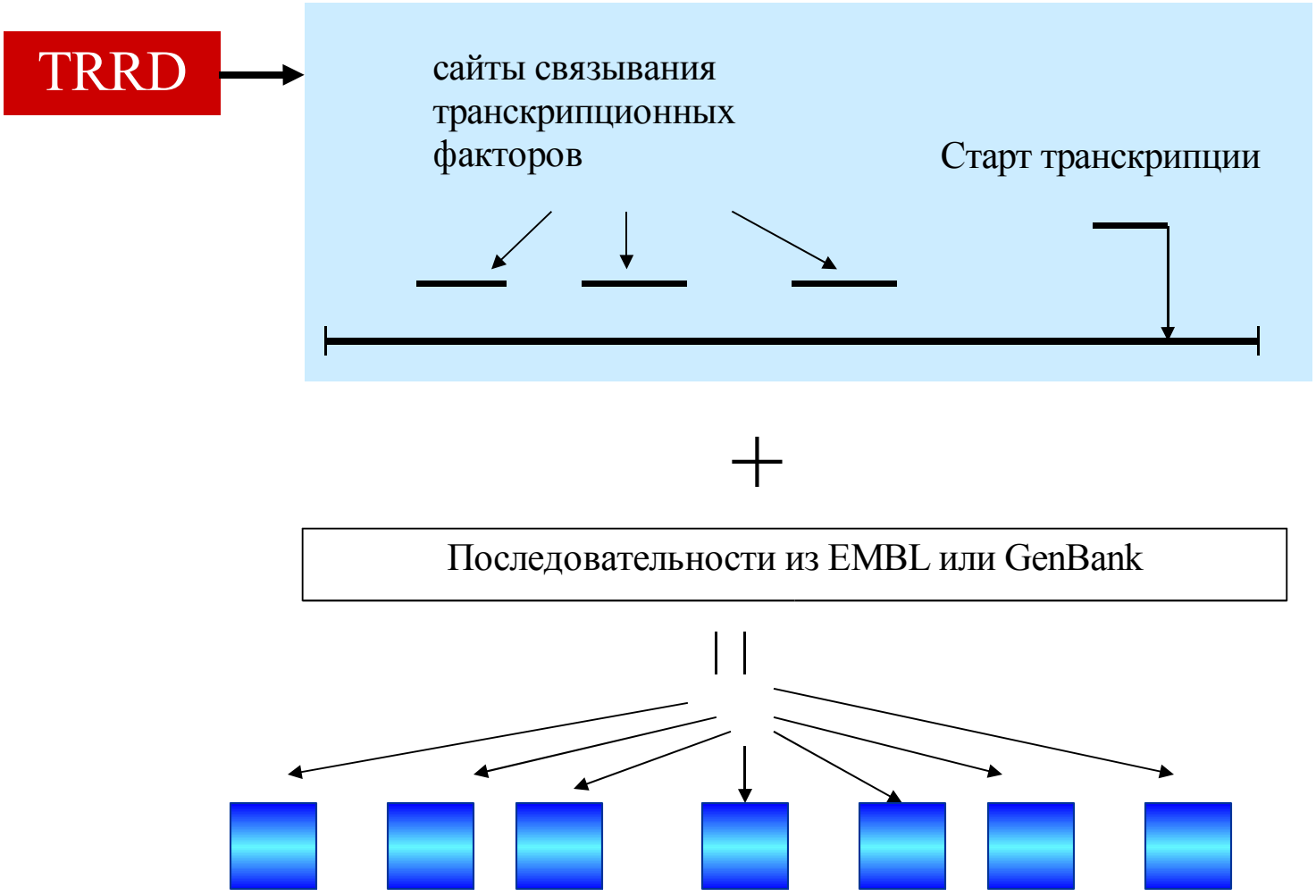
Tilt (τ)

```
DINUCLEOTIDE
   AA 1.9   *
   AT 0.0
   AG 1.3
   AC 0.3
   TA 0.0
   TT 1.9   *
   TG 0.3
   TC 1.7
   GA 1.7
   GT -0.1 **
   GG 1.0
   GC 0.0
   CA 0.3
   CT 1.3
   CG 0.0
   CC 1.0
//
```

Пример записи TRRD

```
GeneID
     Hs:IFNB
     Links:(TRRD Viewer, Transcription factors, Gene expression regulation,
   GeneAC
     A00274
   Annotators
     Ananko E.
   Species
     human, Homo sapiens
   GeneName Full
     interferon-beta
   DNABankLink
     EMBL; HSIFD4; V00534; J00218; K00616; M11029; ST:285
     SWISS-PROT; INB HUMAN; P01574;
     GN_GENE; Hs:IFN-beta;
   EPD Class
     6.1.5.8.
   KeyWords
     virus-induced, T-cell activation,(Medline,GenBank)
     regulation of cell growth and differentiation,(Medline,GenBank)
   Chromosome
     9p22-p21
   RegRegion
     5'region
   REGULATORY UNIT: P00097
   RegUnit
     IRE, interferon regulatory element; ST; -110 to -36; S2294, S1382, S1383,
   SitePosition
     Site:( S2294) -122 to -93; NRE II; negative regulatory domain II
     Site:( S1382) -116 to -88; HMG I binding site (2);
   //
```

Схема работы программы TRRD-Pars экстракции последовательностей сайтов и регуляторных районов из TRRD и EMBL/GenBank



Построение консенсуса для выборки NF-IL6

TGTAGTAAG

TTATGGAAT

TTCCGCAAT

TGGTGAAAT

TTACGCAAG

TGTAGTAAG

TTCGGAAAT

TTGGGCAAG

TGATGGAAG

консенсус TKNNGNAAK

Расширенный 15-буквенный алфавит IUPAC

CODE	NUCLEOTIDE SET	INTERPRETATION
A	A	ADENINE
T	T	THYMIDINE
G	G	GUANINE
C	C	CYTOSINE
W	A, T	WEAK H-BOND
R	A, G	PURINE
M	A, C	AMINO GROUP
K	T, G	KETO GROUP
Y	T, C	PYRIMIDINE
S	G, C	STRONG H-BOND
B	T, G, C	NON-ADENINE
V	A, G, C	NON-THYMIDINE
H	A, T, C	NON-GYANINE
D	A, T, G	NON-CYTOSINE
N	A, T, G, C	ANY NUCLEOTIDE

Консенсусные последовательности сайтов связывания некоторых транскрипционных факторов

фактор	консенсус
AP 1	T G A V T C A
ARNT	A C G T G
CJUN	T G W C N Y H
CKROX	G G G G G S A G G S G
CMYB	H B H V M T D N C M S Y H D B Y A K
COUP	R G K T C A $(N)_{2-6}$ M R G D T C
E2F	T T T S S C G S S V D D
EGR1	N G H G G G G G Y G G S V V K G
ER	M G G T C A T G A B C
GATA	W G A T A R
HNF1	G T T A A T N W T T V W Y
HNF3	G T T T G H Y T
HNF4	G D B C A R A G K K C A
HSE	K W M B V K W M
JUN B	T K A Y T C A
JUN D	G A S T C A
PIT 1	W W W A W W C A T
TTF 1	T C A C R R K

Шаги построения весовой матрицы для сайтов связывания фактора SF1

Выборка сайтов связывания фактора SF1	atgtcaaggccgtgac aggctcaaggtcatca tcaaggagaaggtcag aaagtagaggtcagga gaggcaaggccactgg taccaaggtcagaaat gagttcaaggtaataa tttcgaggtcatggcc
Выравнива-ние	atgt**CAAGGCCG**tgac aggct**CAAGGTCA**tca tcaagga**GAAGGTCA**g aaagt**AGAGGTCA**gga gagg**CAAGGCCA**ctgg tac**CAAGGTCA**gaaat gagtt**CAAGGTAA**taa ttt**CGAGGTCA**tggcca

Шаги построения весовой матрицы для сайтов связывания фактора SF1

C	A	A	G	G	C	C	G	Район кора
C	A	A	G	G	T	C	A	
G	A	A	G	G	T	C	A	
A	G	A	G	G	T	C	A	
C	A	A	G	G	C	C	A	
C	A	A	G	G	T	C	A	
C	A	A	G	G	T	A	A	
C	G	A	G	G	T	C	A	
C	A	A	G	G	T	C	C	
C	A	A	G	G	C	C	C	

1	8	10	0	0	0	1	7	Матрица
8	0	0	0	0	3	9	2	абсолютных
1	2	0	10	10	0	0	1	частот
0	0	0	0	0	7	0	0	

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 P_i 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.

Пример построения матрицы относительных частот для выборки выровненных последовательностей сайтов связывания фактора ACP2

	g	c	a	c	a	a	c	c	c	a	g
	g	c	c	c	t	a	a	c	a	a	g
	g	g	t	a	g	a	g	c	a	a	g
	g	c	a	c	a	a	a	c	c	a	g
	c	c	c	a	g	c	c	c	c	a	g
матрица относительных частот:											
A	0	0	0.4	0.4	0.4	0.8	0.4	0	0.4	1	0
T	0	0	0.2	0	0.2	0	0	0	0	0	0
C	0.2	0.8	0.4	0.6	0	0.2	0.4	1	0.6	0	0
G	0.8	0.2	0	0	0.4	0	0.2	0	0	0	1

Весовая матрица для сайтов связывания фактора ACP2

-0.2	-0.2	0.9	0.9	0.9	1.4	0.9	-0.2	0.9	1.6	-0.2
1.4	-0.2	0.5	-0.2	0.5	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
0.5	1.4	0.9	1.2	-0.2	0.5	0.9	1.6	1.2	-0.2	-0.2
-0.2	0.5	-0.2	-0.2	0.9	-0.2	0.5	-0.2	-0.2	-0.2	1.6

Информационная матрица для сайтов связывания фактора ACP2

-0.02	**-0.02**	**0.3**	**0.3**	**0.3**	**0.8**	**0.3**	**-0.02**	**0.3**	**1.1**	**-0.02**
0.8	**-0.02**	**0.1**	**-0.02**	0.8	**-0.02**	**-0.02**	**-0.02**	**-0.02**	**-0.02**	**-0.02**
0.1	0.8	**0.3**	**0.5**	**-0.02**	0.1	**0.3**	**1.1**	**0.5**	**-0.02**	**-0.02**
-0.02	0.1	**-0.02**	**-0.02**	**0.3**	**-0.02**	0.1	**-0.02**	**-0.02**	**-0.02**	**1.1**

Поиск потенциальных сайтов с помощью весовой матрицы

Весовая матрица

Изучаемая нуклеотидная последовательность

Участок нуклеотидной последовательности проверяется на наличие сайта

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

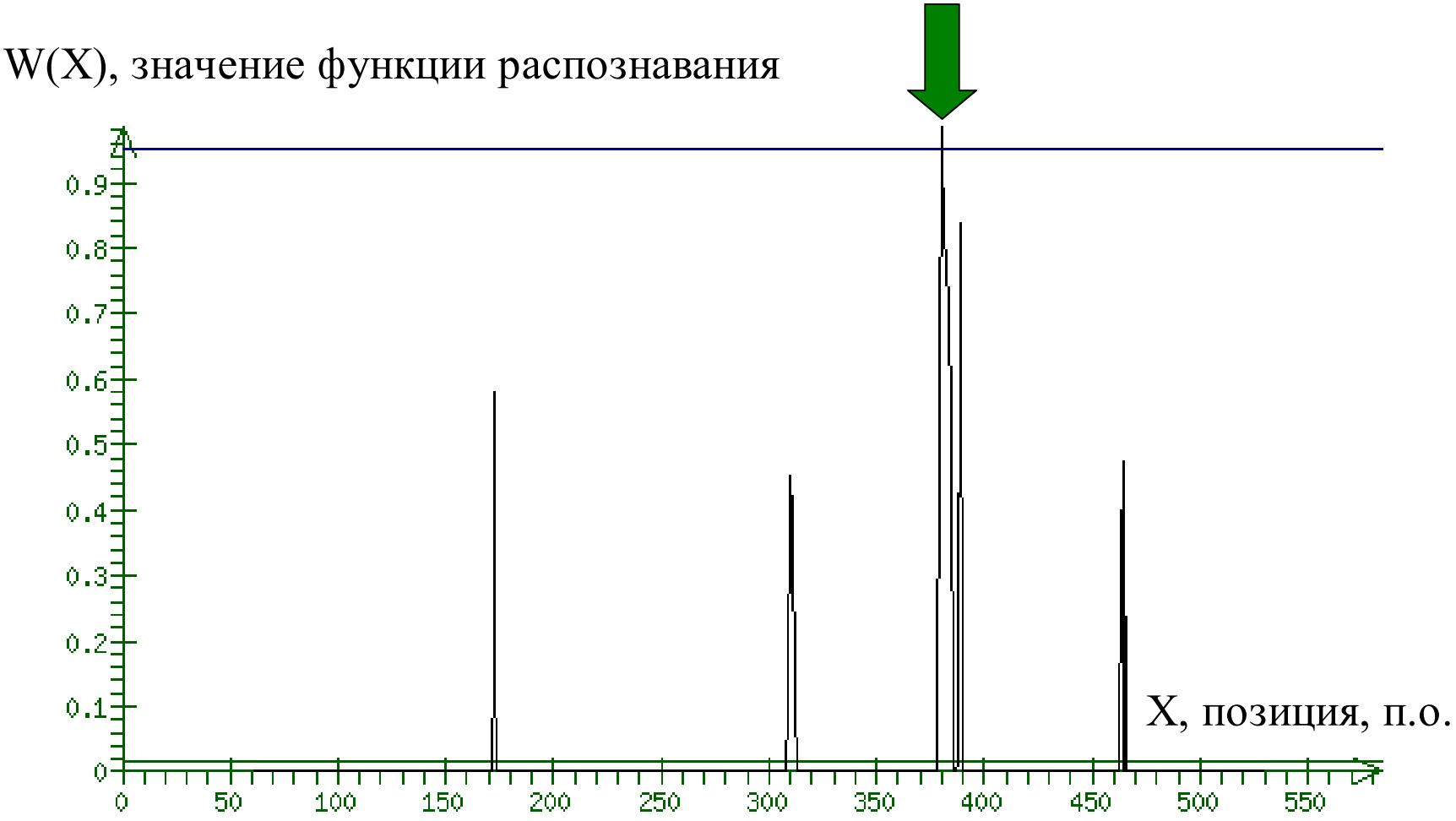
C	A	A	G	G	C	C	G

Вес тестовой последовательности равен:

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

Профиль функции распознавания сайтов NFκB для гена IRF2 человека (EMBL AC D14082, L24442; TRRD S1378)

Сайт расположен в районе (327..446)



Построение метода распознавания сайтов

Выборка негативных последовательностей

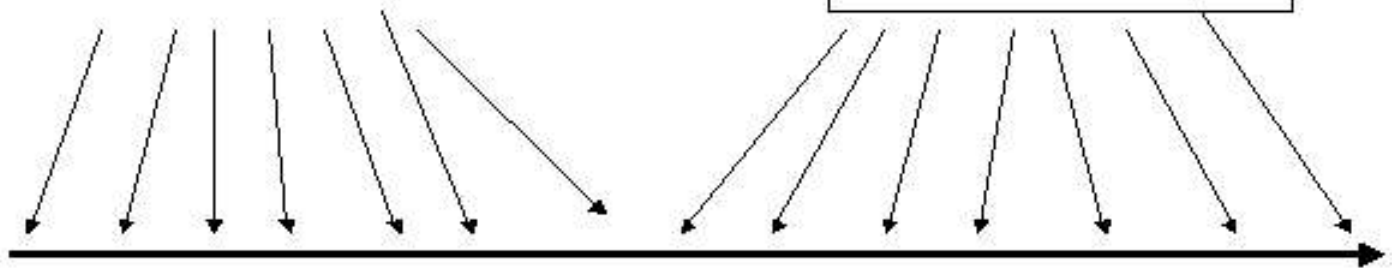
tcccagtcgat
acagtcgtagc
gggtcgtcgaa
ggtacgaacga
acagtgctgca

Весовая матрица

Выборка сайтов

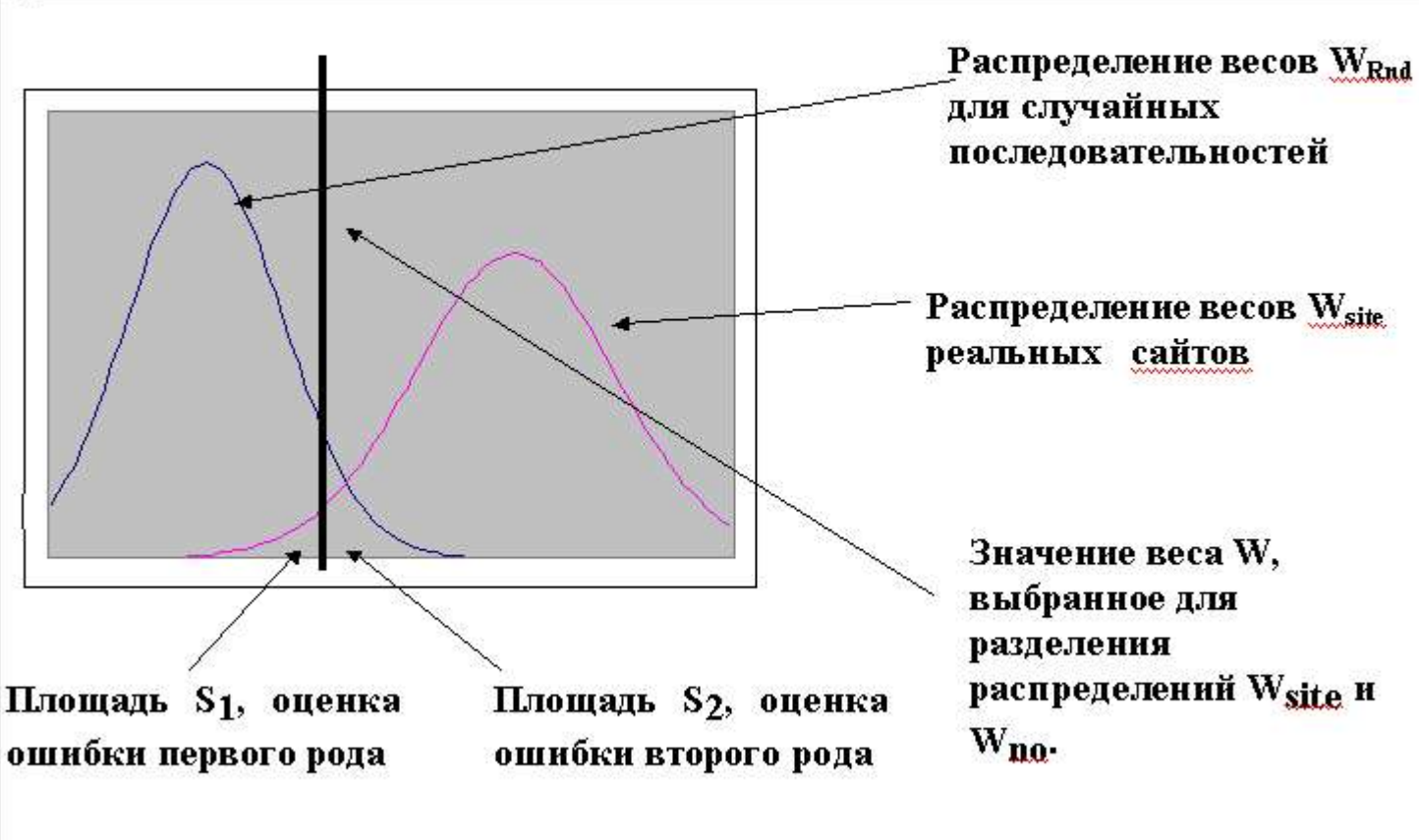
taccaaggtca
agacaaggtca
ggacaaggtca
ggccaaggtca
agacaaggtca

Весовая матрица



Числовая прямая

Схема распределений весов сайтов и негативных последовательностей



Некоторые способы оценки точности распознавания

1) Оценка ошибки первого рода (доля сайтов, которые не были распознаны)

$$E_1 = \frac{N^-_{site}}{N_{site}}$$ **где**

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

N_{site} **- размер контрольной выборки сайтов**

Некоторые способы оценки точности распознавания

2) Оценка ошибки второго рода (доля негатиных последовательностей, которые были распознаны как сайты)

$$E_2 = \frac{N_{no}^{+}}{N_{no}}$$ **где**

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),

FP – число неверно предсказанных сайтов,

FN – 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:

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

Другие оценки точности распознавания

Чувствительность – доля правильно предсказанных сайтов

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

Специфичность (доля правильно отвергнутых негативных последовательностей)

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

Другие оценки точности распознавания

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

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

Коэффициент корреляции (мера связи между предсказанием и реальностью, одновременно учитывающая все элементы таблицы сопряженности)

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

Зависимость оценки ошибки второго рода от ошибки первого рода для сайтов связывания фактора GATA