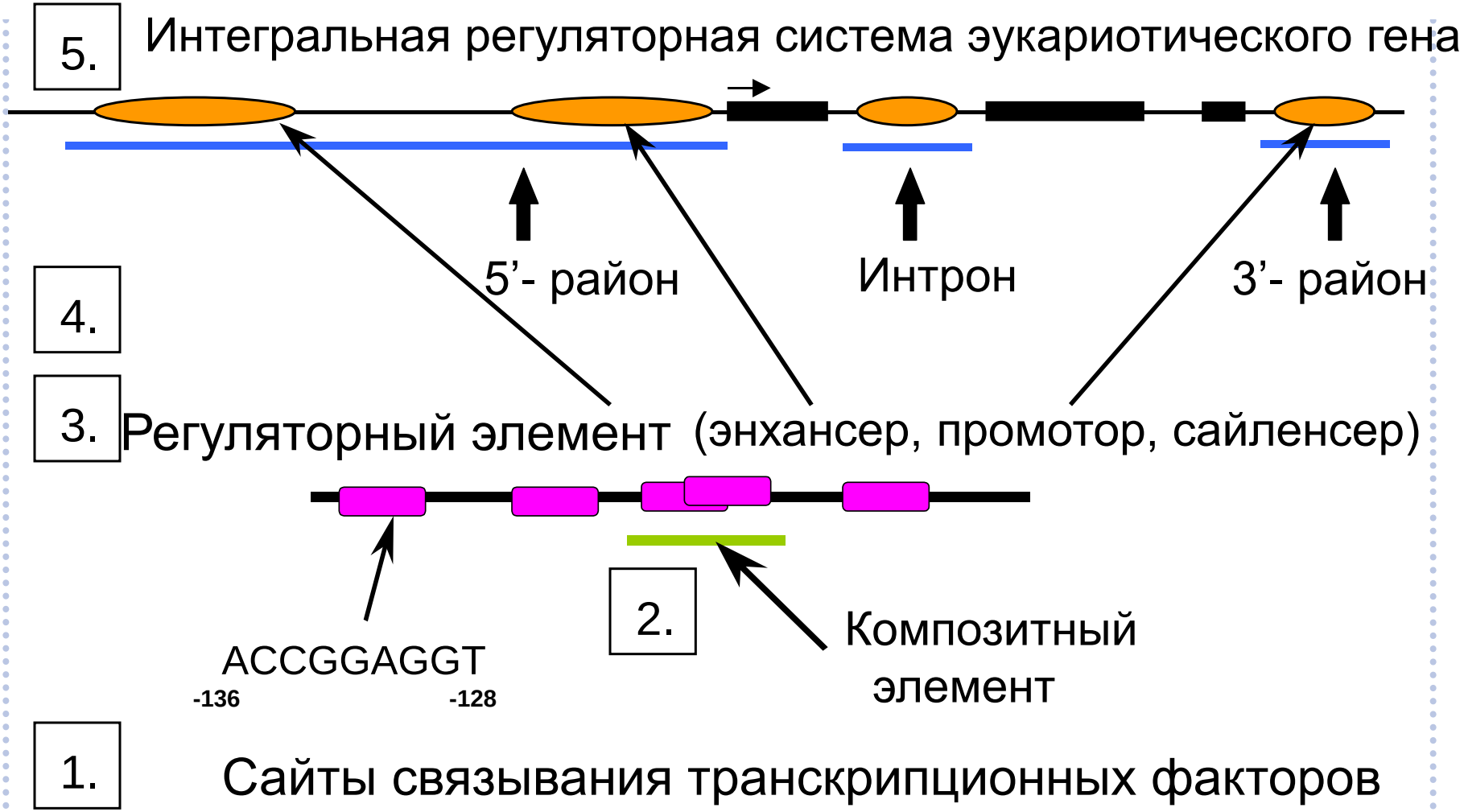
Институт Цитологии и Генетики СО РАН, Новосибирск

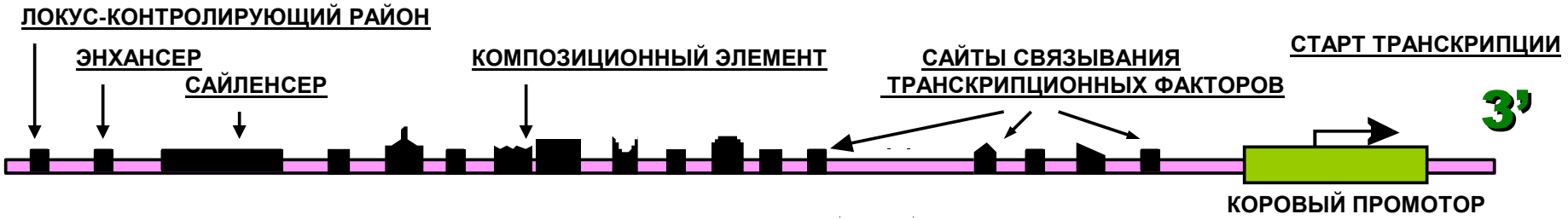
Распознавание и анализ сайтов связывания транскрипционных факторов

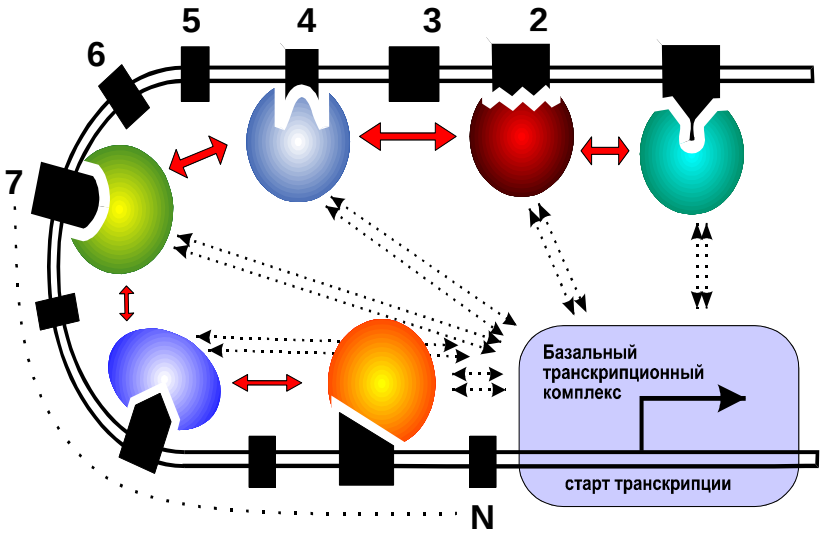
О.В.Вишневский

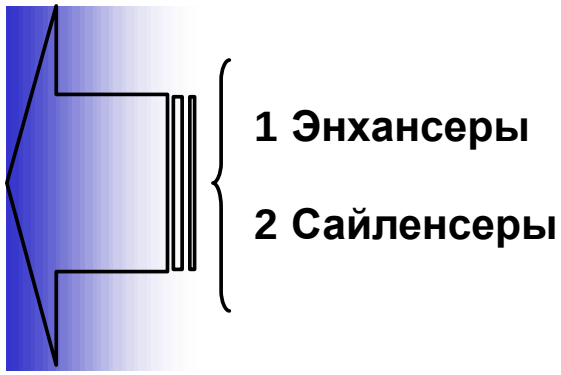
Общая модель транскрипционного регуляторного района эукариотического гена



РЕГУЛЯЦИЯ ТРАНСКРИПЦИИ У ЭУКАРИОТ





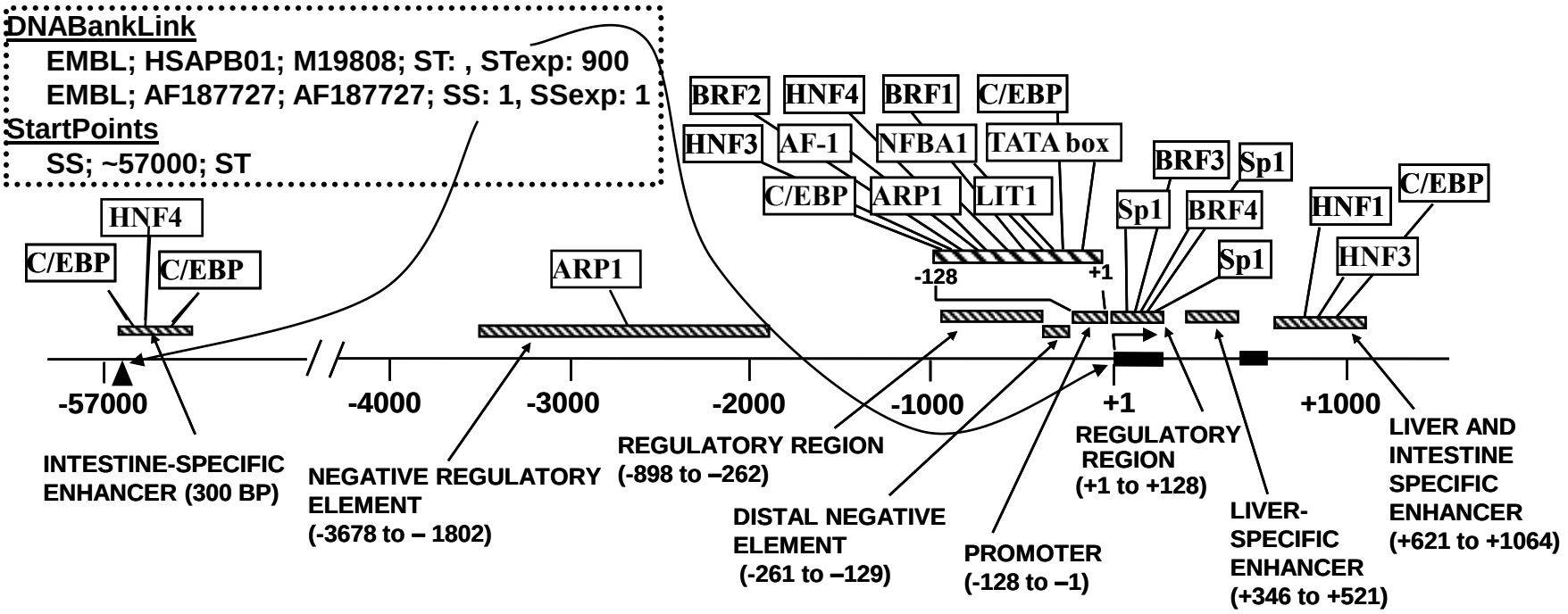


Емкость кода регуляции транскрипции

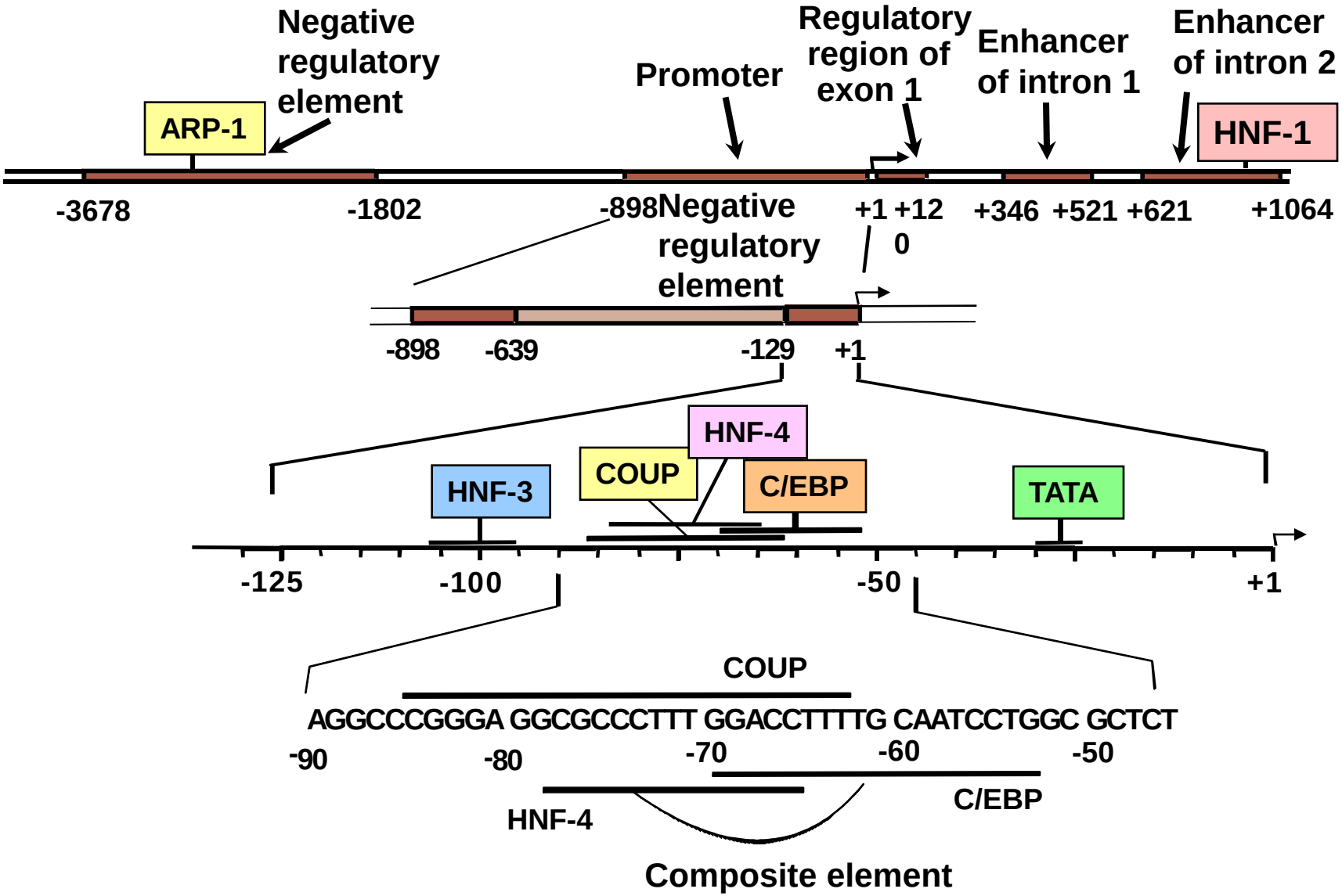
$$W = \sum_{n=1}^{N} C_N^n \cdot 2^{C_n^2}$$

(Для 20 сайтов $W=10^{80}$)

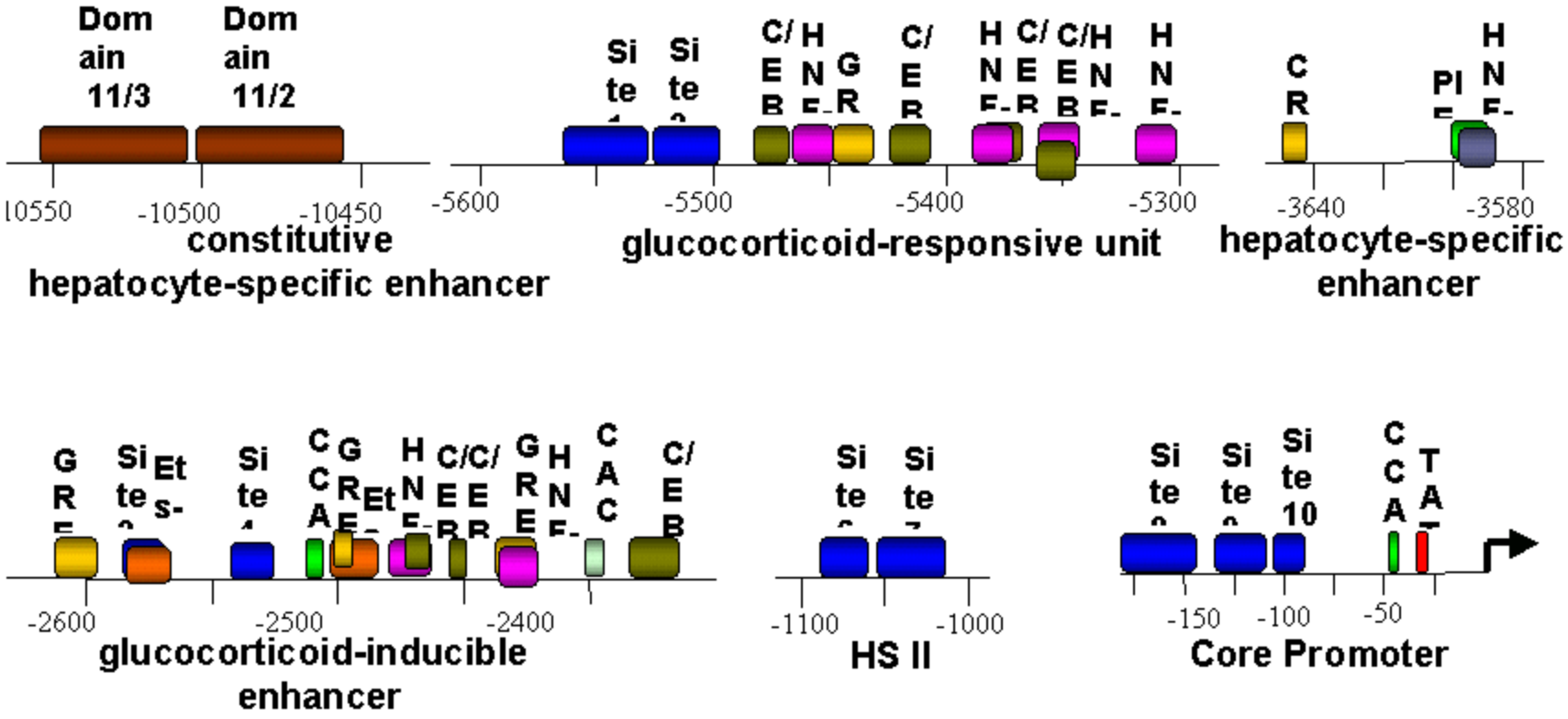
РЕГУЛЯТОРНЫЕ РАЙОНЫ , КОНТРОЛИРУЮЩИЕ ТРАНСКРИПЦИЮ ГЕНА АПОЛИПОПРОТЕИНА В ЧЕЛОВЕКА



Организация регуляторных районов транскрипции гена аполипопротеина В человека

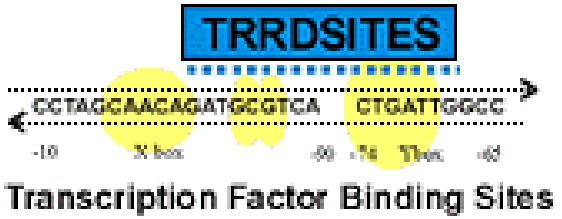


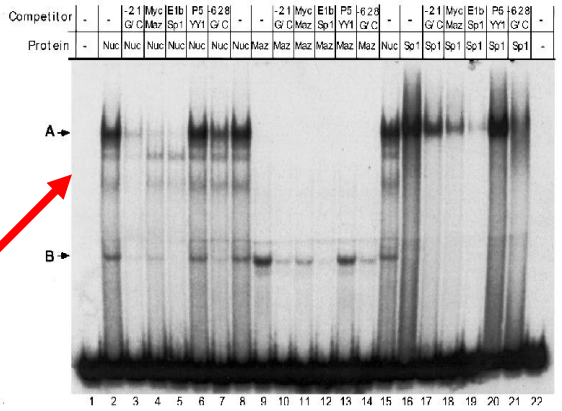
Регуляторный район гена тирозин аминотрансферазы крысы



TRRDSITES

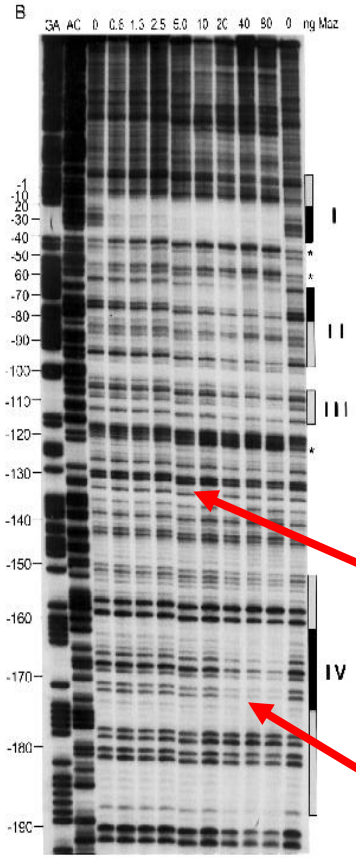
ОПИСАНИЕ САЙТА СВЯЗЫВАНИЯ ТРАНСКРИПЦИОННОГО ФАКТОРА



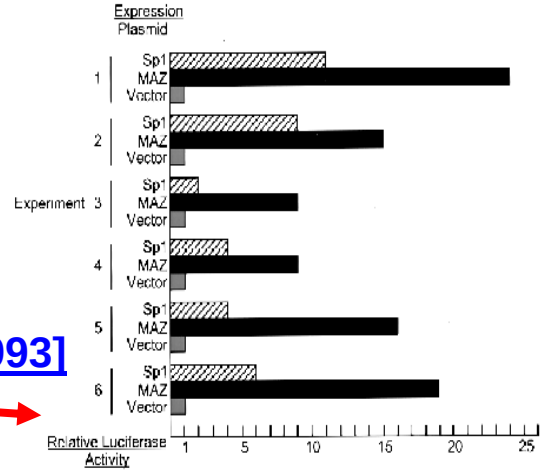


GEL-MOBILITY SHIFT ASSAY

AN S1160
ID Gene: Hs:APOB
AP REGULATORY UNIT: P00670
NM HNF-4 bs; HNF-4 binding site
NY AF-1 binding site
NY BA1 binding site
NS R01612
TF HNF-4; hepatic nuclear factor 4
AT increase
SQ cccgggaggCGCCCTTTGGACCTtttg
PQ -88 to -62
PF -82 to -62
BF EMBL: M15053 : 68
AG 1.1.5, 3.5 [Metzger S. et al., 1993]
AG rat liver cells: 3.6 [Metzger S. et al., 1993]
AG human HepG2 cells: 6.2 [Metzger S. et al., 1993]
AG HeLa cells: 6.2, 6.6 [Metzger S. et al., 1993]
AG 1.1.5, 3.3, 3.5, 4.2 [Ladias J.A. et al., 1992]
//



DNASE I FOOTPRINTING



TRANSIENT EXPRESSION ANALYSIS

Методы анализа и распознавания сайтов связывания основанные на выравнивании экспериментально полученных участков посадки белковых факторов

ttagcctg**ATGCTAC**ccaattgcgatt
gccggg**GCTACGGT**gcggtacggccga
tacggtccgatg**ATGCTACGGT**ggatt

ttagcctg**ATGCTAC**ccaattgcgatt
gccggg**GCTACGGT**gcggtacggccga
tacggtccgatg**ATGCTACGGT**ggatt

Сайты рестрикции, на которых впервые применялся подобный подход были идеальным объектом, для анализа и распознавания, поскольку, зачастую, представляют собой очень простую последовательность ДНК, например GAATTC для EcoRI белка.

Причем все участки немодифицированной ДНК имеющие точно такие последовательности нуклеотидов будут распознаваться соответствующими рестриктазами и разрезаться.

Мутация хотя бы в одной позиции приводит к снижению эффективности рестрикции на порядки.

До сих пор сайты рестрикции остаются наиболее легко распознаваемыми сигналами в ДНК.

МЕТОД КОНСЕНСУСА

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

TATAAT

Консенсус **TATAAT**
распознает 2 сайта из 6
реальных и предскажет
1 ложный сайт на 4000 п.н.

Выборка из 6 –10 районов промоторов E.coli
(Pribnov,1975).

Расширенный IUPAC алфавит

1	A	A	adenine
2	T	T	timidine
3	G	G	guanine
4	C	C	citosine
5	R	G/A	purine
6	Y	T/C	pirimidine
7	M	A/C	amino
8	K	T/G	keto
9	W	A/T	weak
10	S	G/C	strong
11	B	not A	
12	V	not T	
13	H	not G	
14	D	not C	
15	N	any	

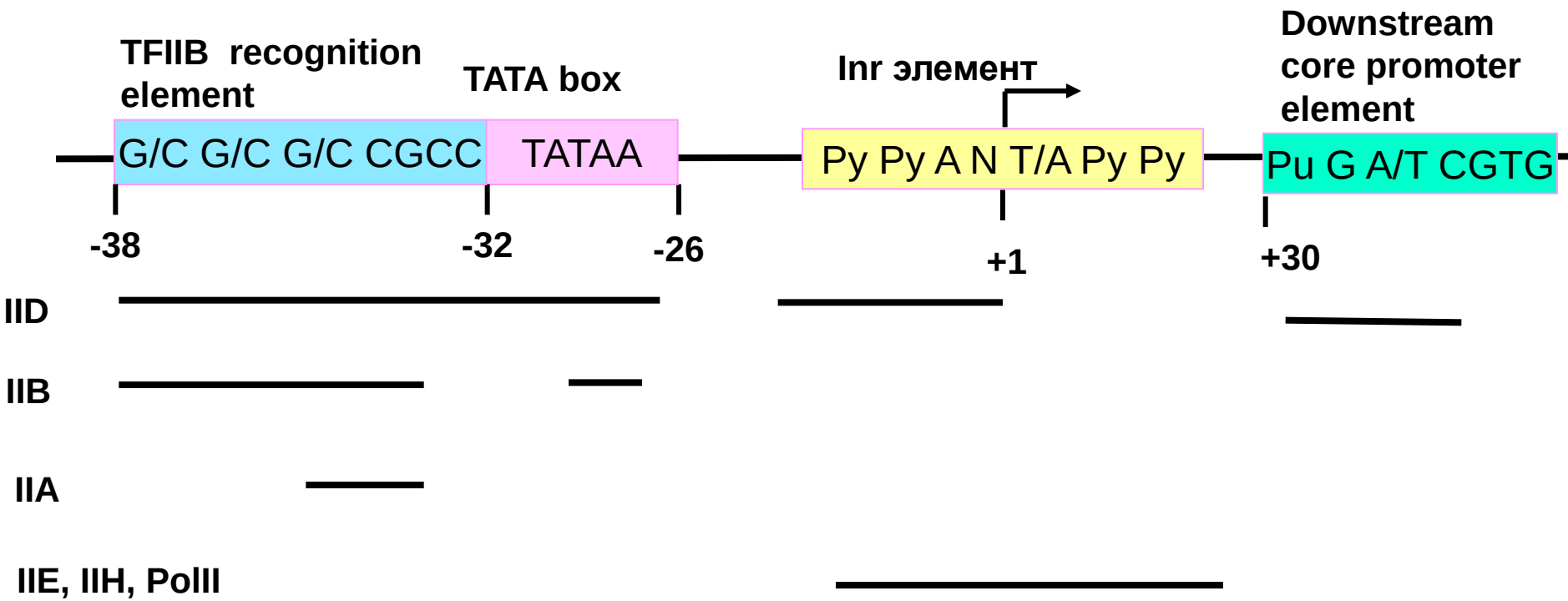
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

TATRHT

Консенсус **TATRHT**
распознает 4 реальных сайта
из 6 и предскажет
1 ложный сайт на 770 п.н.

Выборка из 6 –10 районов промоторов E.coli
(Pribnov,1975).

ТИПИЧНЫЕ ЭЛЕМЕНТЫ КОРОВОГО ПРОМОТОРА Pol II ТРАНСКРИБИРУЕМЫХ ГЕНОВ.



дрозофила

Типы коровых промоторов:
TATA+ Inr+;
TATA+ Inr-;
TATA- Inr+;
TATA- Inr-;

«Py» – пиримидин «C» либо «T»
«Pu» пурин «A» либо «G»

ЧАСТОТНАЯ МАТРИЦА

Каждый $n_{i,b}$ элемент матрицы содержит информацию о представленности b-й буквы в i-й позиции сайта

	T	A	T	R	H	T
A	0	6	0	3	4	0
T	5	0	5	0	1	6
G	1	0	0	3	0	0
C	0	0	1	0	1	0

МАТРИЦА ОТНОСИТЕЛЬНЫХ ЧАСТОТ

Каждый $f_{i,b}$ элемент матрицы равен $f_{i,b} = n_{i,b} / N$

	T	A	T	R	H	T
A	0	1	0	1/2	2/3	0
T	5/6	0	5/6	0	1/6	1
G	1/6	0	0	1/2	0	0
C	0	0	1/6	0	1/6	0

ПОЗИЦИОННАЯ ВЕСОВАЯ МАТРИЦА

$$w_{i,b} = \ln\left(\frac{f_{i,b}}{p_b} + \frac{s}{100}\right) + c_i$$

f_{ib} – относительная частота b-й буквы в i-й позиции сайта,
p_b – средняя частота присутствия b-й буквы в геноме,
а c_i и s – константы, подбираемые эмпирически.

Пример позиционной весовой матрицы для –10 района промоторов E.coli.

	T	A	T	R	H	T
A	-0.69	1.34	-0.69	0.77	1.00	-0.69
T	1.18	-0.69	1.18	-0.69	0.05	1.34
G	0.28	-0.69	-0.69	1.09	-0.69	-0.69
C	-0.69	-0.69	0.28	-0.69	0.28	-0.69

Оценка веса W (score) нуклеотидной последовательности на основе позиционной весовой матрицы

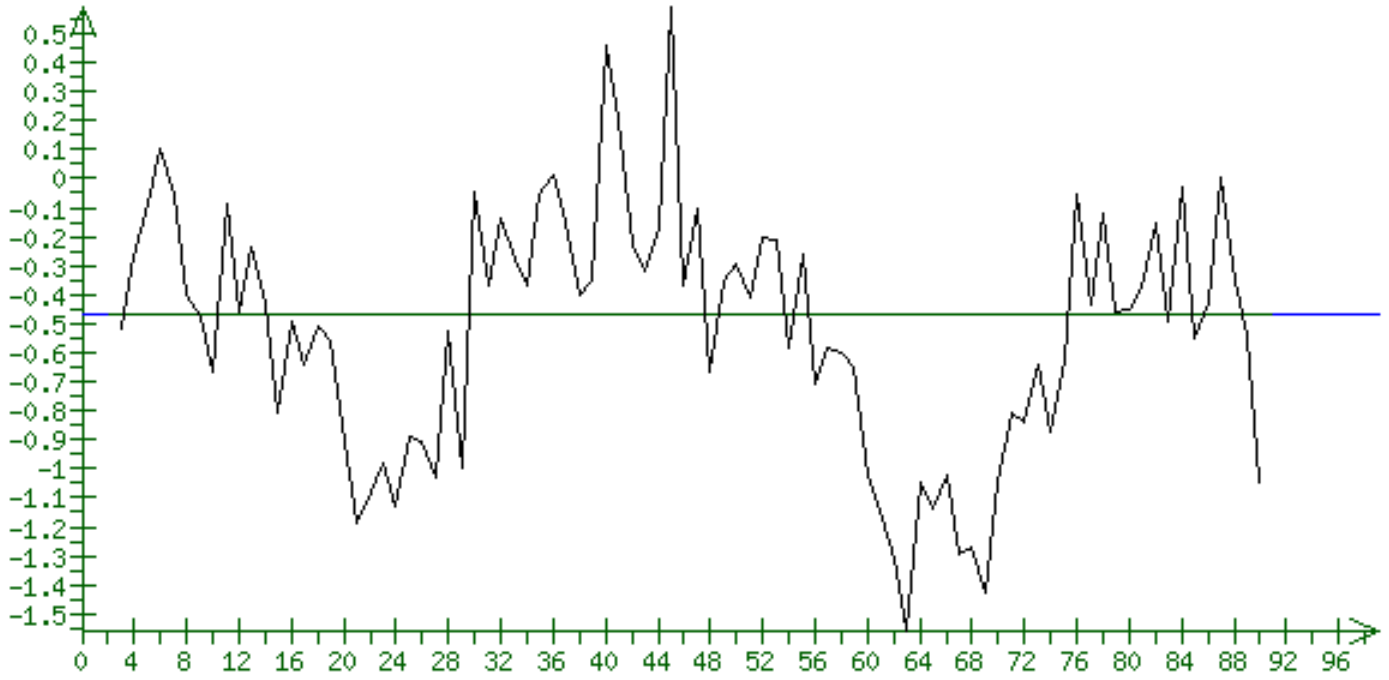
$$W = \sum_{i=1}^{L} w_{i,b}$$

Пример оценки весов нуклеотидной последовательности на основе весовой матрицы –10 района промоторов E.coli.

A	-0.69	1.34	-0.69	0.77	1.00	-0.69
T	1.18	-0.69	1.18	-0.69	0.05	1.34
G	0.28	-0.69	-0.69	1.09	-0.69	-0.69
C	-0.69	-0.69	0.28	-0.69	0.28	-0.69

$$W(GCTATG)=0.28-0.69+1.18+0.77+0.05-0.69=0.9$$

Распознавание сайта связывания транскрипционного фактора Sp1 в протяженной последовательности с помощью позиционной весовой матрицы.



Оценка информационного содержания позиций построенных весовых матриц

$$I_i = 2 + \sum_{b=A,T,G,C} f_{b.i} \log_2 f_{b.i}$$

$$I_i = \sum_{b=A,T,G,C} f_{b.i} \log_2 \frac{f_{b.i}}{p_b}$$

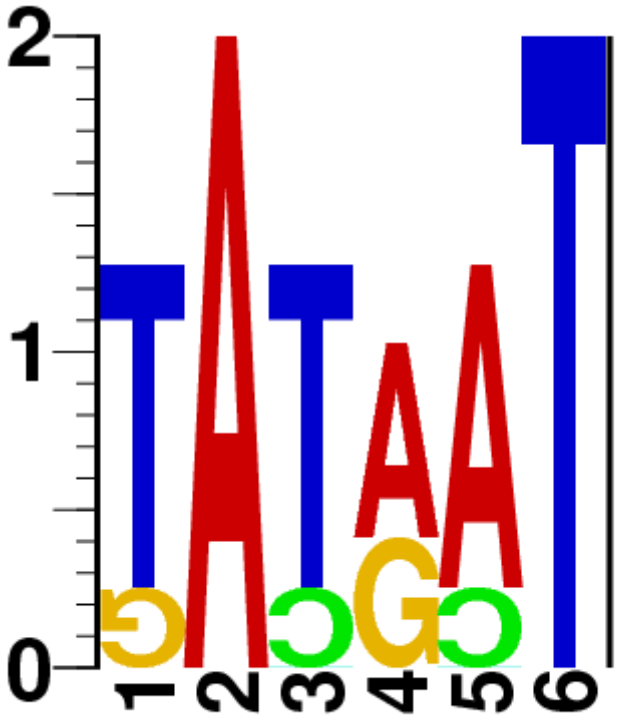
Оценка полного информационного содержания на примере информационной матрицы –10 района промоторов E.coli.

$$I = \sum_{i=1}^{L} I_i$$

A	0	1.34	0	0.38	0.66	0
T	0.98	0	0.98	0	0.00	1.34
G	0.04	0	0	0.54	0	0
C	0	0	0.04	0	0.04	0
I_i	1.03	1.34	1.03	0.93	0.72	1.34

I=1.03+1.34+1.03+0.93+0.72+1.34=6.39

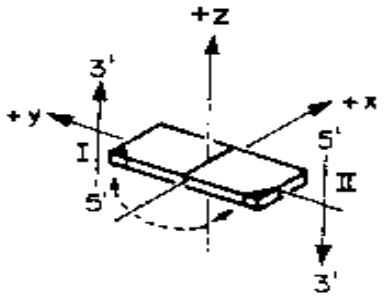
Logo- представление весовой матрицы



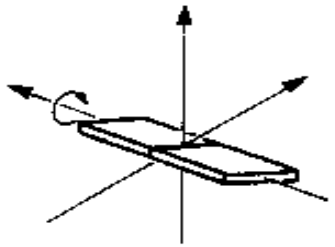
Logo- представление частотной матрицы для выборки –10 районов промоторов E.coli.

http://www.cbs.dtu.dk/gorodkin/appl/slogo.html

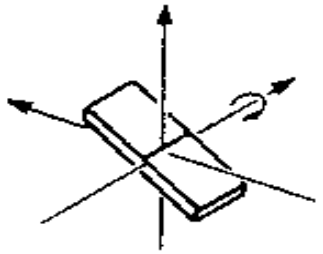
Локальные конформационные свойства двойной цепи ДНК.



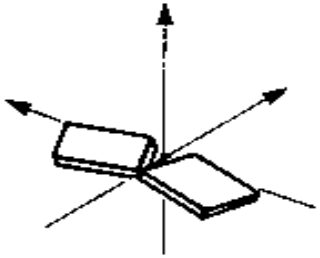
Coordinate frame



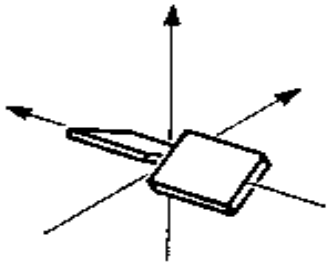
Tip (θ)



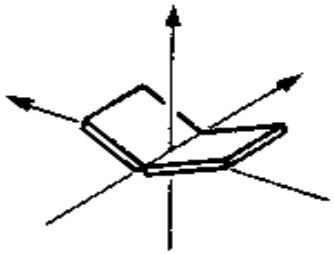
Inclination(η)



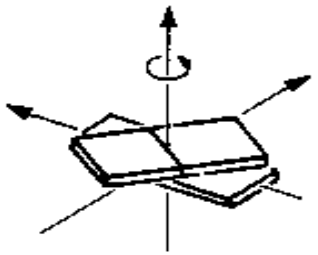
Opening (σ)



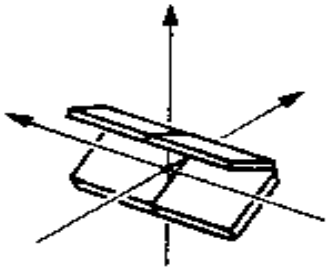
Propeller twist(ω)



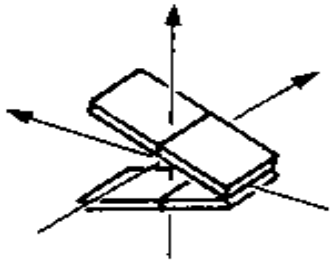
Buckle(κ)



Twist(Ω)



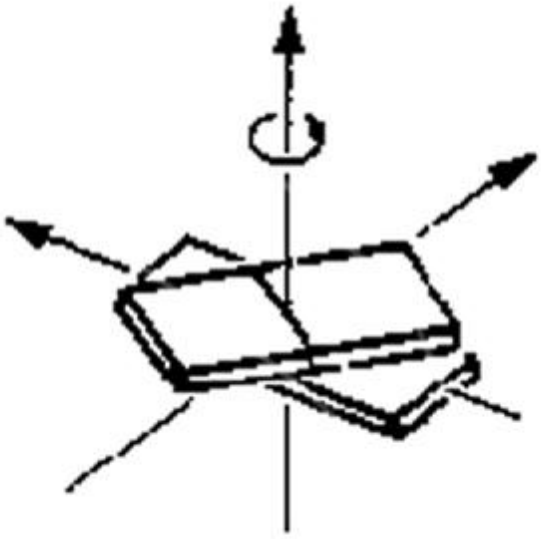
Roll(ρ)



Tilt(τ)

Описание зависимости угла TWIST от динуклеотидного контекста

```
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
```

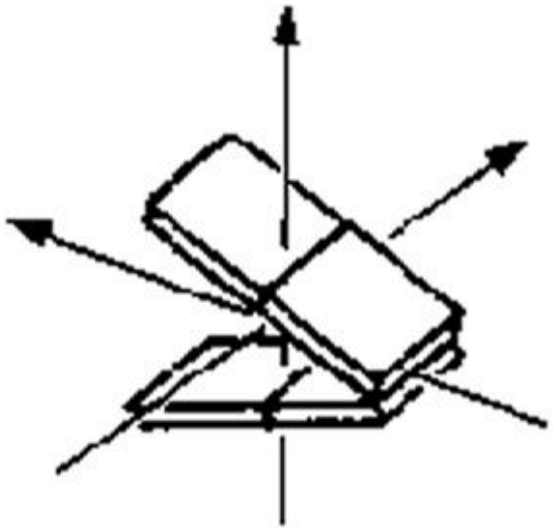


Twist (Ω)

```
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

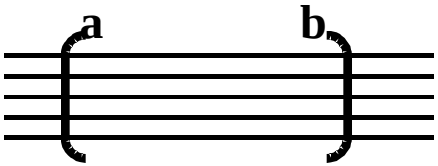
```
MI P0000016
MN Conformational
MD DNA/protein-complex
ML dinucleotide step
PN Tilt
PM Averaged for X-rays
PV TiltCompl
PU Degree
```



Tilt (τ)

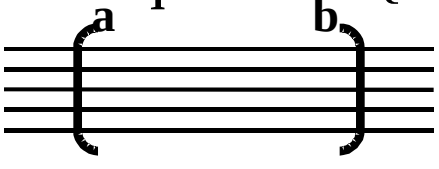
```
DINUCLEOTIDE
   AA 1.9  *
   AT 0.0
   AG 1.3
   AC 0.3
   TA 0.0
   TT 1.9  *
   TG 0.3
   TC 1.7
   GA 1.7
   GT -0.1 **
   GG 1.0
   GC 0.0
   CA 0.3
   CT 1.3
   CG 0.0
   CC 1.0
//
```

Sites {S}

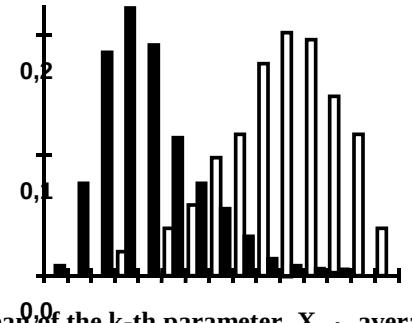


$$X_{k,a,b}(S) = \frac{\sum_{i=a}^{b-5} X_k(S_i, S_{i+1})}{b-a-4};$$

Random Sequences {R}



$$X_{k,a,b}(R) = \frac{\sum_{i=a}^{b-5} X_k(R_i, R_{i+1})}{b-a-4};$$



LET Xk,a,b(S) and Xk,a,b(R) be the mean of the k-th parameter, X_k, averaged for a region [a, b] of the site S and random sequence R, respectively .

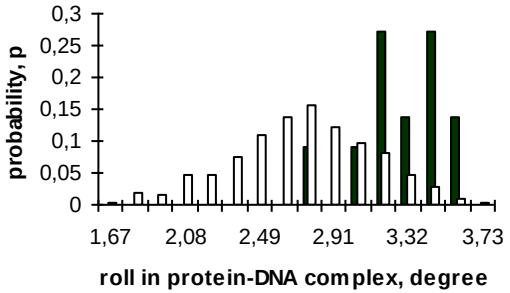
The difference between distributions $X_{k,a,b}\{S\}$ and $X_{k,a,b}\{R\}$ is tested for significance using the following criteria:

(1) the difference between the means of $X_{k,a,b}\{S\}$ and $X_{k,a,b}\{R\}$;
(2) the difference between the variances of $X_{k,a,b}\{S\}$ and $X_{k,a,b}\{R\}$;
(3) the difference between the densities of $X_{k,a,b}\{S\}$ and $X_{k,a,b}\{R\}$;
(4) the difference between the ranges of $X_{k,a,b}\{S\}$ and $X_{k,a,b}\{R\}$.
(5) $X_{k,a,b}\{S\}$ distribution is Gaussian,
(6) $X_{k,a,b}\{R\}$ distribution is Gaussian.

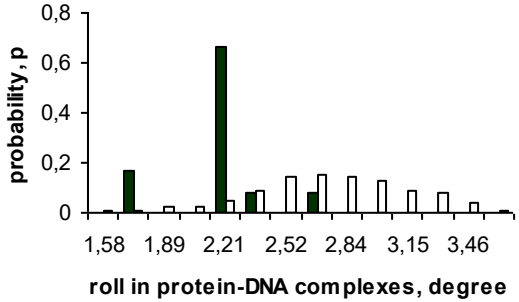
Выбор лучших физико-химических и конформационных параметров сайтов методом B-DNA-VIDEO

Fishburn's Theory of Utility for Decision Making

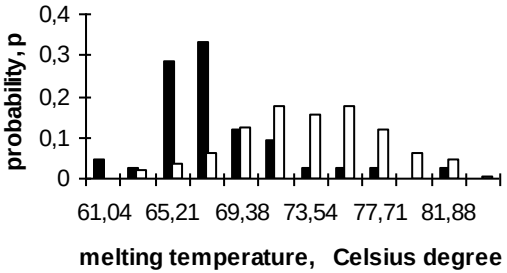
AP-1 (region [-5;17])



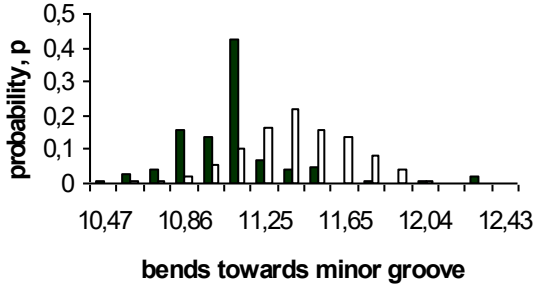
EN (region [-10;2])



HNF3 (region [-21;4])



ER (region [-6; 17])



Сравнение распределений значимых конформационных и физико-химических параметров для сайтов связывания транскрипционных факторов (черные колонки) и случайных последовательностей (белые колонки)

Система ACTIVITY

КОНТЕКСТНО-ЗАВИСИМЫЕ ХАРАКТЕРИСТИКИ:

Облигатные:

1) Универсальные для сайтов одного типа
2) Определяют базальный уровень активности сайтов (F_0)

Факультативные:

1) Специфичны для каждой последовательности
2) Модулируют активность сайта относительно базального уровня

$$F(S) = F_0(S) + \sum_{k=1}^{K} F_k \times X_k(S)$$

F(S) - Общее значение активности сайта с нуклеотидной последовательностью S

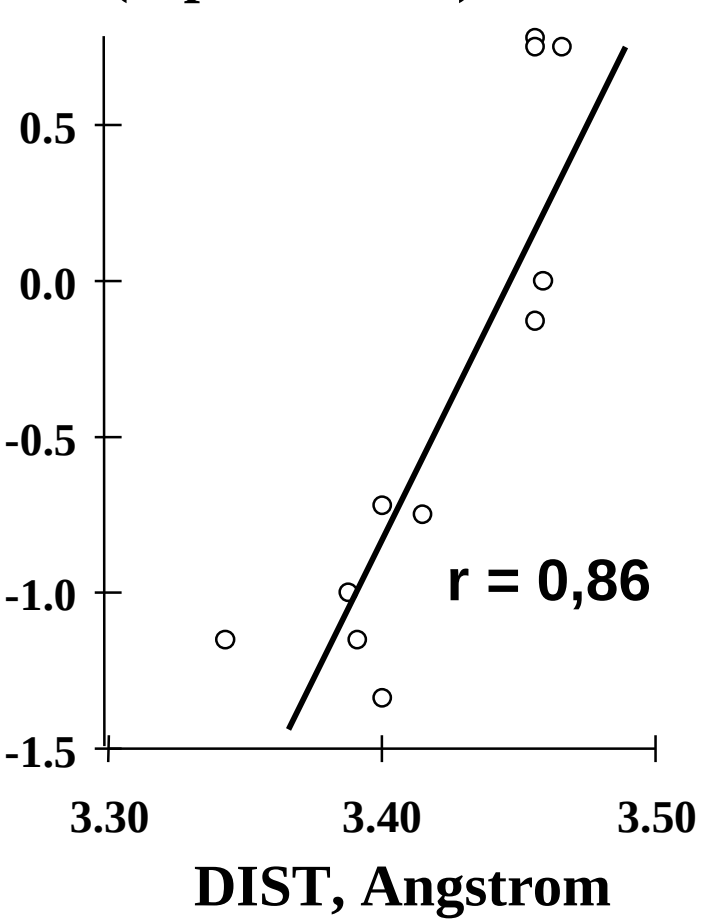
F_0(S) - Базальная активность сайтов данного типа

F_K - Вклад факультативного свойства X_k в активность сайта

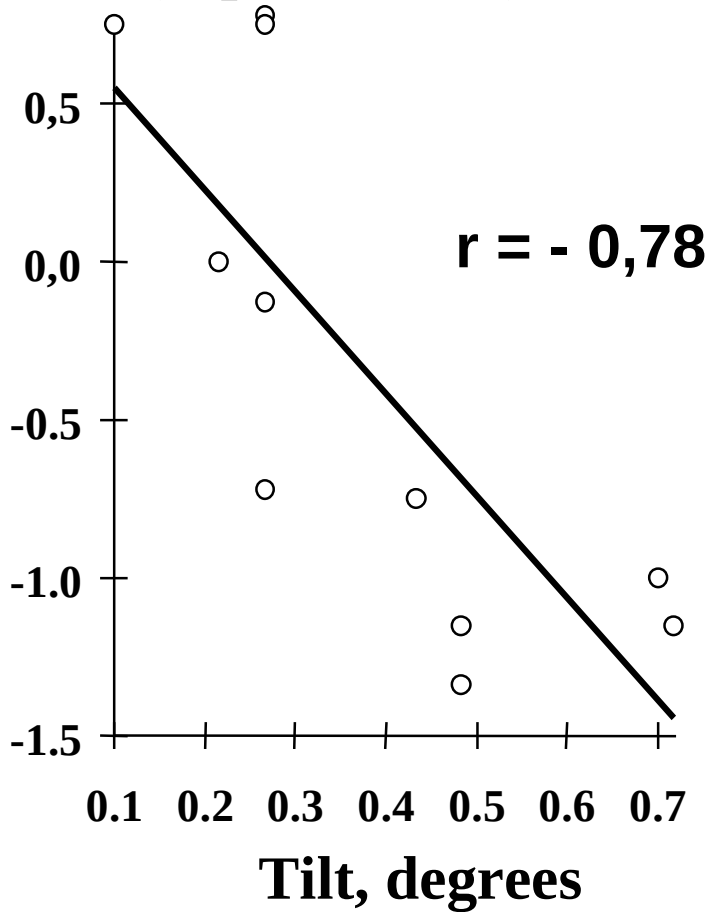
X_K(S) - Значение факультативного свойства X_k для последовательности S

Линейно аддитивная модель

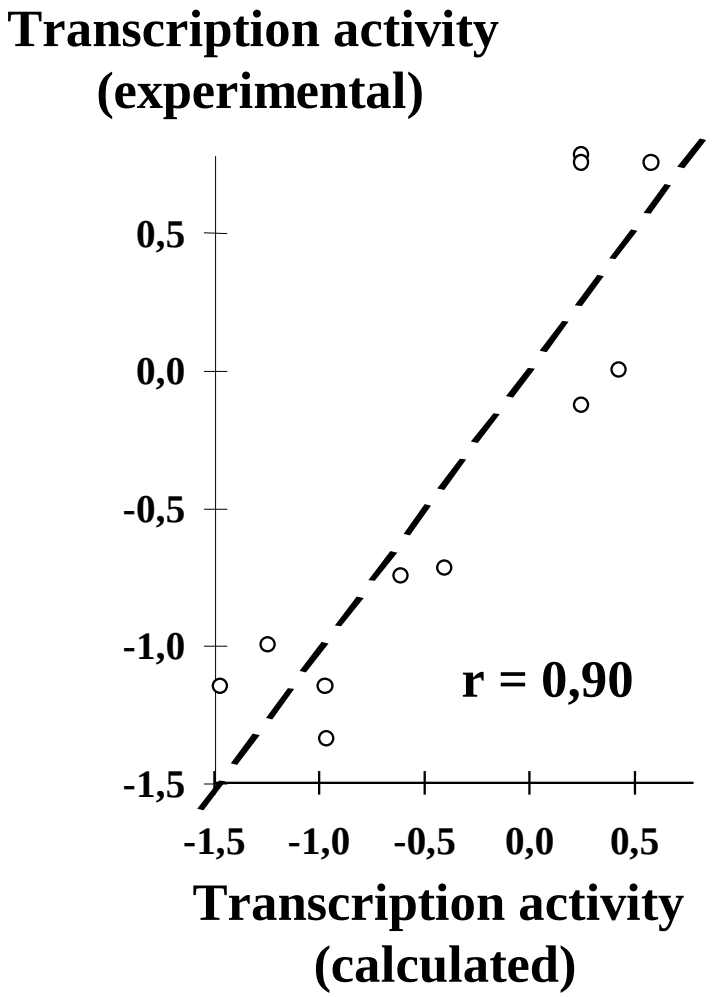
a) **Transcription activity (experimental)**



b) **Transcription activity (experimental)**



Транскрипция гена αA-crystalline зависит и от ширины большой бороздки ДНК и от угла Tilt



Транскрипционная активность гена αA-crystallin мыши: сравнение предсказанных и экспериментально определенных значений

$$F = -39 - 0.1 * Pnucl + 12 * DIST - Tilt$$

Методы распознавания сайтов связывания транскрипционных факторов в невыровненных нуклеотидных последовательностях.

Не смотря на значительный прогресс в разработке новых экспериментальных подходов анализа регуляторных районов и накопление большого количества данных о локализации сайтов связывания транскрипционных факторов, компьютерные методы анализа и распознавания регуляторных элементов остаются одним из эффективных инструментов исследователей.

Такие методы позволяют выявлять характерные контекстные сигналы в выборках регуляторных районов объединенных общей функцией без использования экспериментальных данных о локализации в них известных сайтов связывания. Подобный анализ может дать информацию о присутствии нового сайта или о конформационных характеристиках анализируемого района. В то же время решение этой задачи требует применения более сложных методов анализа и способов оценки значимости выявленных сигналов.

Методы k-плетов.

Для обнаружения коротких и высоко консервативных мотивов широко используются методы, основанные на обнаружении слов заданной длины достоверно перепредставленных в обучающей выборке по сравнению с ожидаемым по случайным причинам.

Метод WORDUP, предложенный Pesole et al. заключается в поиске относительно коротких слов в 4-х буквенном алфавите, представленных в большей части последовательностей обучающей выборки и наращивания их длины до тех пор, пока значимость удлиненного слова, определенная по критерию χ^2 не станет ниже значимости более короткого входящего в него слова.

Дана последовательность $S= x_1x_2…x_n$; x_i=A,T,G,C
содержащая n-w+1 олигонуклеотидов длины w, определенных как
$s_j= x_j,x_{j+1},…,x_{j+w-1}$; j=1, n-w+1
Рассмотрим N нуклеотидных последовательностей S_1, S_2, …,S_N, длиной L_i нуклеотидов (i=1,…,N).
Если олигонуклеотиды распределяются согласно распределению Пуассона, то вероятность $\pi_i(s_k)$ встретить олигонуклеотид k, длиной w нуклеотидов (k=1, …,4^w), в последовательности S_i хотя бы один раз:

$$\pi_i(s_k) = 1 - e^{-q_i(s_k)*(L_i - w+1)}$$, где

$$q_i(s_k) = \frac{f(u_{1,2})*f(u_{2,3})*\dots*f(u_{w-1,w})}{f(u_2)*\dots*f(u_{w-1})}$$

– ожидаемая вероятность олигонуклеотида s_k в последовательности i исходя из марковской модели первого порядка. $f(u_{x,y})$ и $f(u_z)$– частоты ди и мононуклеотидов в каждой последовательности.

Наблюдаемое число последовательностей, содержащих олигонуклеотид s_k из N равно:

$$P(s_k) = \sum_{i=1}^{N} p_i(s_k)$$

где $p_i(s_k) = 1$, если олигонуклеотид s_k обнаружен в i последовательности, и $p_i(s_k) = 0$, если олигонуклеотид s_k не обнаружен в i последовательности.

Ожидаемое число последовательностей, содержащих олигонуклеотид s_k из N равно:

$$\prod(s_k) = \sum_{i=1}^{N} \pi_i(s_k)$$

Статистическая значимость наблюдения слова s_k может быть вычислена по критерию χ^2.

$$\chi_k^2 = [P(s_k) - \prod(s_k)]^2 / \prod(s_k)$$

Олигонуклеотиды, чья значимость превышает пороговое значение могут быть отсортированы на основании этого критерия.

Пусть $W=\{s_1,\ldots, s_z\}$ – набор значимых олигонуклеотидов длины w, отсортированных согласно их значениям χ^2. Генерируются все олигонуклеотиды длины w+1, содержащие слова s_i и s_j длины w из множества W и рассчитывается их значение χ^2. Полученный олигонуклеотид длины w+1, s^{w+1} будет считаться статистически значимым, если

$$\chi^2(s^{w+1})>\max[\chi^2(s_i), \chi^2(s_j)],$$

В этом случае олигонуклеотиды s_i и s_j длины w заменяются на s^{w+1} длины w+1.

Данная процедура повторяется для слов длины w+2 и так далее до тех пор, пока выполняется данное условие.

Метод реализаций.

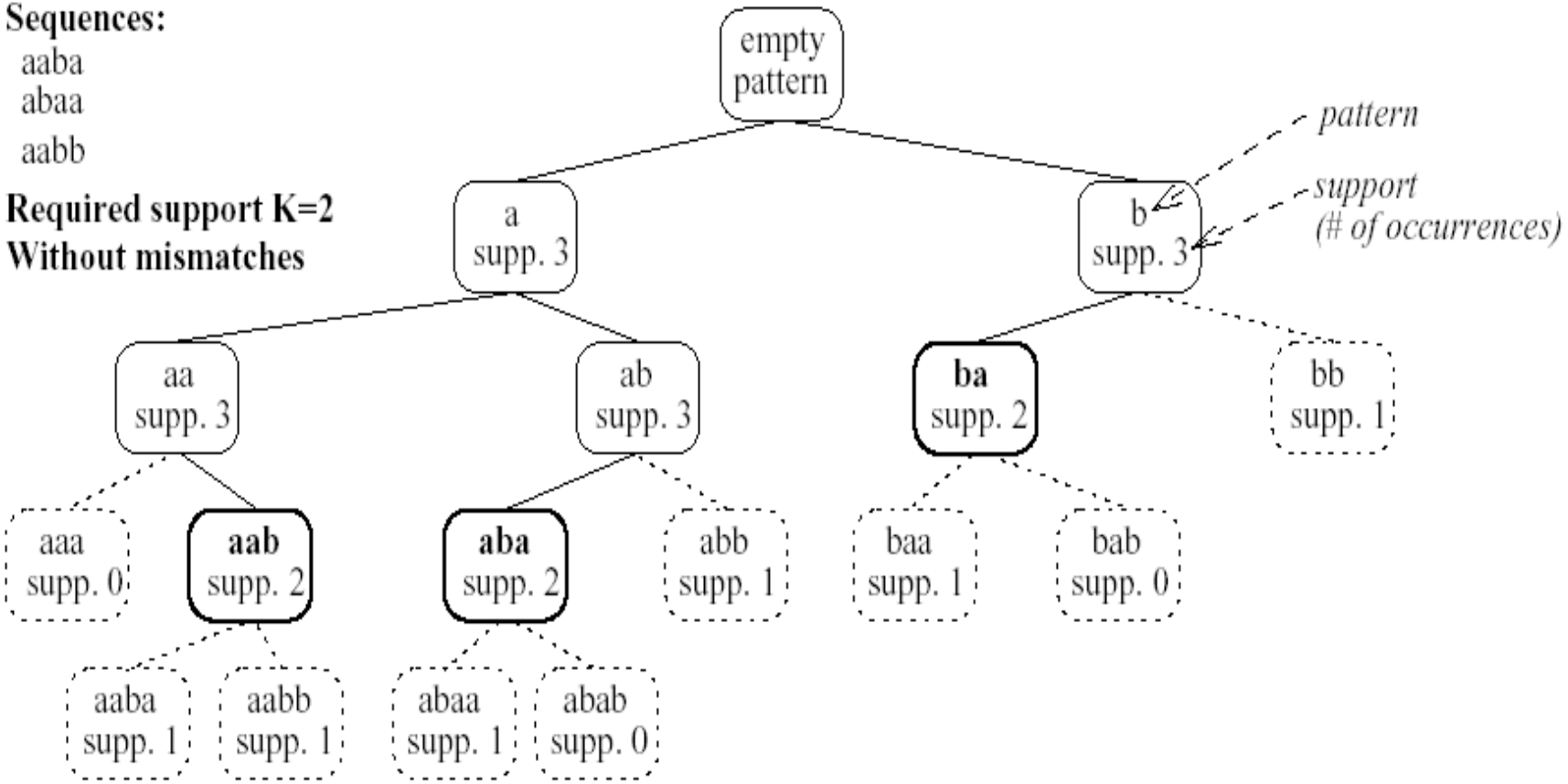
Метод выявления сайтов связывания предложенный Кондрахиным основан на представлении сайтов связывания как набора конкретных реализаций $R = \{R_0, R_1, ..., R_{k-1}\}$. Каждая из реализаций представляет собой слово длины τ в 15-символьном коде IUPAC-IUB. Такой подход позволяет избежать усреднения нуклеотидного контекста в описании функционального сайта, как это имеет место при построении одиночного консенсуса. Построение набора реализаций определятся двумя параметрами: 1) длиной олигонуклеотидного слова τ ; 2) максимально допустимым различием (расстояние Хэмминга) $t^{(miss)}$ между этими олигонуклеотидными словами.

На первом шаге в анализируемом наборе $U_0=\{u_1,\ldots, u_m\}$ последовательностей экспериментально определенного сайта путем полного перебора выявляется так называемая главная реализация R_0, представляющая из себя олигонуклеотидное слово длины τ с наибольшей частотой встречаемости в выборке . После этого все последовательности, содержащие это слово, удаляются из выборки U_0 образуя множество U_1. На следующем этапе в получившейся выборке последовательностей путем полного перебора производится поиск следующей реализации, то есть такого слова R_1, которое обладает наименьшим расстоянием от слова R_0 и наибольшей представленностью в выборке. Этот процесс итерационно продолжается до тех пор, когда множество U_r окажется пустым, либо в нем присутствуют только слова, отличающиеся от R_0 больше чем на $t^{(miss)}$.

Для автоматического выявления наиболее эффективных значений параметров производится запуск описанной процедуры для всех возможных заданных значений пар τ и $t^{(miss)}$. При этом максимизируется функционал f(τ, $t^{(miss)}$) соответствующий проценту покрываемых последовательностей, то есть последовательностей содержащих хотя бы одну из реализаций R при текущих значениях параметров.

Метод распознавания на основе набора реализаций был реализован в программе FUNSITE-SIG-REAL. Считается, что фрагмент ДНК длины может рассматриваться как функциональный сайт определенного типа, если он совпадает с одной из реализаций этого сайта.

Методы суффиксных деревьев.



Дерево суффиксов для выборки, состоящей из трех последовательностей.

Алгоритм WINNOWER.

$n = 4,\ d = 1,\ L = 3$

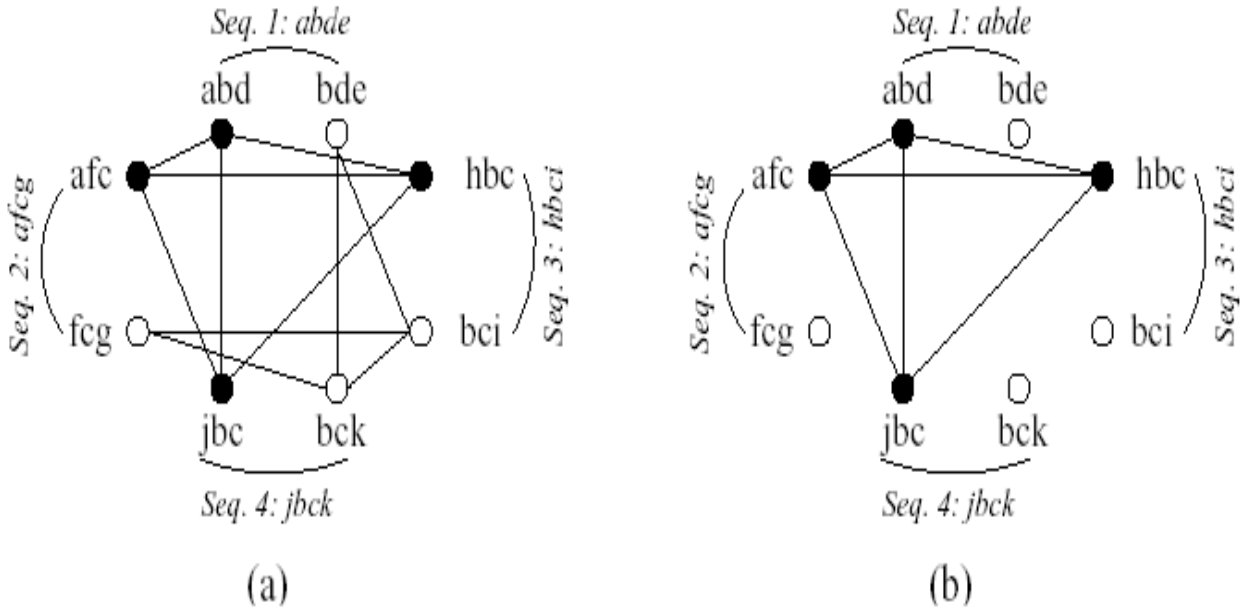
Sequences:

abde

afcg

hbci

jbck



Демонстрация работы алгоритма WINNOWER. (a) Граф подслов длины L=3 построенный для выборки четырех последовательностей. (b) Клика размера 4, полученная в результате удаления всех вершин, которые не могут входить в клики заданного размера.

Методы локального множественного выравнивания регуляторных последовательностей.

Для нахождения более длинных и более вырожденных сигналов используют методы локального множественного выравнивания. В этом случае задачей алгоритма является нахождение оптимального локального выравнивания максимизирующего целевую функцию.

Программа CONSENSUS

A

```
A C T G A A T
A G C G T C C
C T T G C C G
```

B

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

I = 12.0

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

I = 12.0

C

```
  A C T G A A
  A G C G T C
A 2 0 0 0 1 1
C 0 1 1 0 0 1
G 0 1 0 2 0 0
T 0 0 1 0 1 0
```

I = 8.0

```
  A C T G A A
  G C G T C C
A 1 0 0 0 1 1
C 0 2 0 0 1 1
G 1 0 1 1 0 0
T 0 0 1 1 0 0
```

I = 7.0

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

I = 6.0

```
  C T G A A T
  G C G T C C
A 0 0 0 1 1 0
C 1 1 0 0 1 1
G 1 0 2 0 0 0
T 0 1 0 1 0 1
```

I = 7.0

D

```
  A C T G A A
  A G C G T C
  C T T G C C
A 2 0 0 0 1 1
C 0 1 1 0 0 1
G 0 1 0 2 0 0
T 0 0 1 0 1 0
```

I = 6.1

```
  A C T G A A
  A G C G T C
  T T G C C G
A 1 0 0 0 1 1
C 0 2 0 0 1 1
G 1 0 1 1 0 0
T 0 0 1 1 0 0
```

I = 3.8

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

I = 5.8

```
  C T G A A T
  G C G T C C
  T T G C C G
A 0 0 0 1 1 0
C 1 1 0 0 1 1
G 1 0 2 0 0 0
T 0 1 0 1 0 1
```

I = 5.4

Метод поиска максимального правдоподобия (EM) метод.

Метод EM является стандартным и широко используемым подходом для решения задач поиска максимального правдоподобия.

Он представляет собой два итерационных шага:

1. expectation (E) –шаг оценки вероятностных параметров модели и

2. maximization (M)- шаг максимизации параметров модели,

которые повторяются до тех пор, пока не будет выполнено условие конвергенции (сойденности алгоритма).

Lawrnce and Reilly реализовали и применили EM алгоритм для анализа сайтов связывания белка CRP.

На первом шаге его работы происходит начальная оценка и установка частот нуклеотидов входящих в сайт p_{bx} и не входящих в него p_b^0.

Затем производится вероятностное взвешивание каждой позиции возможного начала сайта связывания на основе формулы Байеса.

Согласно этой формуле вероятность встретить сайт связывания B_{jy} начинающийся в y-й позиции j-й последовательности может быть рассчитана как

$$P(B_{jy} \mid p, S) = \frac{P(S \mid B_{jy}, p) * P^0(B_{jy})}{\sum_x P(S \mid B_{jx}, p) * P^0(B_{jy})} = \frac{P(S \mid B_{jy}, p)}{\sum_x P(S \mid B_{jx}, p)}$$

где $p=\{p_{bx}, p_b^0\}$, S- рассматриваемая последовательность, исходная (prior) вероятность $p^0(B_{jy})$ постоянна и равна $1/(L-k+1)$, а $P(S|B_{jy},p)$ равна произведению вероятностей оснований входящих в сайт, начинающийся в y позиции, и оснований не входящих него на основе модели $p=\{p_{bx}, p_b^0\}$.

Шаг E заканчивается расчетом ожидаемого числа n_{bx} каждого основания b в каждой позиции x сайта связывания путем суммирования вероятностных весов позиций потенциальных сайтов, содержащих данное основание b, и числа n_b^0 каждого основания b в не - сайтах.

Шаг M заключается в пересчете p_{bx} и p_b^0 исходя из полученных n_{bx} и n_b^0.

Эти процедуры последовательно продолжаются до тех пор, пока алгоритм не сойдется, и параметры p_{bx} и p_b^0 не перестанут изменяться.

Gibbs sampler

Алгоритм стартует со случайного выбора одной стартовой позиции в каждой последовательности выборки. Работа программы заключается в итерационном повторении двух операций.

A. Шаг построения модели. На этом шаге происходит генерирование модели **p** = $\{p_{b,x}, p^0_b\}$, для первой выбранной последовательности. Как и в случае EM метода модель **p** оценивается на основе позиционных частот оснований, входящих в последовательности предполагаемых сайтов. При этом сайт входящий в выбранную последовательность при расчете модели **p** не учитывается.

B. Шаг вероятностного выбора (sampling). На этом шаге происходит выбор новой позиции сайта для первой последовательности путем вероятностного взвешивания всех возможных позиций. При этом вес каждой позиции оценивается как $A_x=Q_x/P_x$, где Q_x – вероятность генерирования сегмента x на основе модели $p_{b,x}$, а P_x – на основе p^0_b. А вероятность случайного выбора сегмента x рассчитывается, как $\frac{A_x}{\sum_i A_i}$.

MOTIF sampler

В начале все последовательности выборки объединяются в одну последовательность.

Затем производится выравнивание случайно разбросанных сайтов. При этом сайты не должны перекрываться или попадать на границы последовательностей. Оставшиеся районы считаются не-сайтами.

После этого программа берет первую позицию объединенной последовательности и проверяет, не перекрывается ли сайт, начинающийся в этой позиции с границами последовательностей или с ранее определенными сайтами других типов.

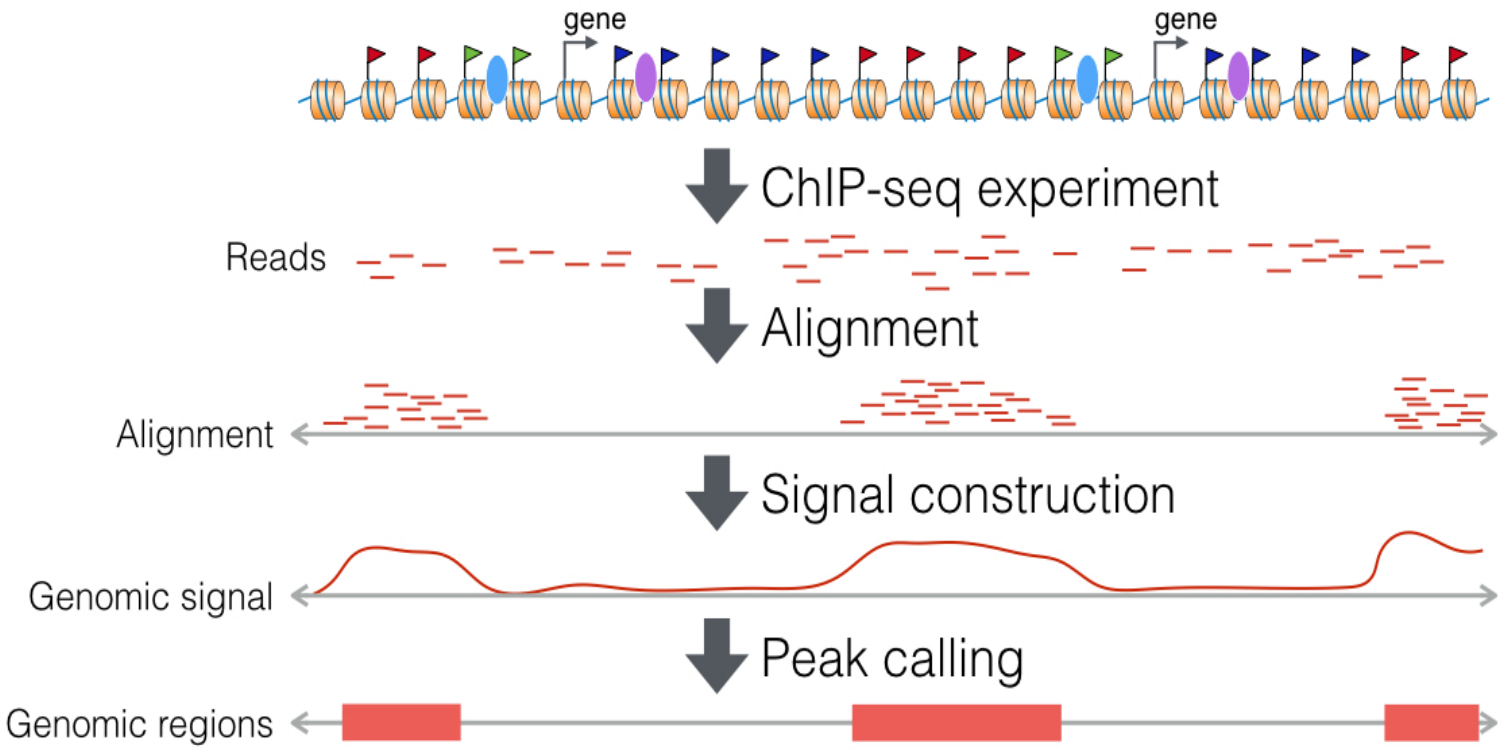
Затем из выравненного набора текущих анализируемых сайтов удаляются сайты, с которыми он перекрывается и производится пересчет вероятностей $p=\{p_{bx}, p_b^0\}$.

Следующим шагом программы является вероятностное помещение анализируемой позиции либо в набор сайтов, либо не сайтов согласно отношению

$p(site)* p_{bx}/(p(non\text{-}site)*p_b^0)$,

где p(site) и p(non-site) –вероятности принадлежности какого либо участка к сайтам или не сайтам, определяемые пользователем.
Затем программа берет следующую позицию и т.д. до последней позиции объединенной последовательности.
Процесс снова повторяется с первой позиции и т.д. до тех пор, пока алгоритм не сойдется.
Данный подход позволяет последовательно находить сайты разных типов и оптимизировать их длину.

Chromatin immunoprecipitation sequencing (ChIP-seq)



DREME

http://meme-suite.org/tools/dreme

DREME (Discriminative Regular Expression Motif Elucidation) это программа из Интернет- доступного пакета MEME, предназначенная для выявления коротких (до 8 п.н.) вырожденных мотивов, записанных 15-ти буквенном коде и соответствующих коровому району сайтов связывания. DREME создавался для быстрого анализа больших данных ChIP-seq экспериментов.

На вход программы подается две выборки последовательностей. Первая – выборка последовательностей, соответствующая районам пиков ChIP-seq. В качестве второй выборки может выступать либо контрастная выборка последовательностей другого ChIP-seq эксперимента, либо выборка из сгенерированных последовательностей, полученная путем перемешивания последовательностей первой выборки.

DREME

1. Оценка значимости всех возможных совершенных (невырожденных) олигонуклеотидов длины от 3 до 8 п.н. с помощью точного теста Фишера.
2. Ранжирование олигонуклеотидов по значимости. 100 наиболее значимых поступают на второй этап.
3. Для каждого из 100 слов производится попытка генерализации. После чего производится оценка его полезности.
4. Оценка значимости всех полученных генерализаций и их сортировка. 100 наиболее значимых мотивов поступают на дальнейшую генерализацию (шаг 3). Этот процесс продолжается до тех пор, пока в ходе генерализации продолжают появляться новые значимые мотивы.
5. Выделяется наиболее значимый мотив. Все олигонуклеотиды в анализируемой выборке, входящие в область значений этого мотива маскируется и процесс нахождения следующего по значимости мотива запускается заново.

На выходе программы получается набор значимых мотивов (консенсусов), записанных в 15-ти буквенном коде.

CisFinder

http://lgsun.grc.nia.nih.gov/CisFinder

CisFinder это еще один популярный Веб-ресурс для выявления значимых мотивов, соответствующих сайтам связывания транскрипционных факторов в выборках ChIP-seq экспериментов.

На вход программы подается анализируемая выборка последовательностей и контрольная выборка. В качестве контрольной выборки программа CisFinder может сгенерировать случайную выборку на основе марковости частот нуклеотидов третьего порядка.

Метод CisFinder основан на выявлении перепредставленных в анализируемой выборке коротких слов со вставками и без вставок, и их кластеризации.

Результатом работы программы является позиционная частотная матрица произвольной длины и ее logo-представление.

A. Nucleotide substitution matrix for word 'ATGCAAAT'

$[W_{pi}] =$

Position	i = 1 (A)	i = 2 (C)	i = 3 (G)	i = 4 (T)
p = 1	**A**TGCAAAT	**C**TGCAAAT	**G**TGCAAAT	**T**TGCAAAT
p = 2	A**A**GCAAAT	A**C**GCAAAT	A**G**GCAAAT	A**T**GCAAAT
p = 3	AT**A**CAAAT	AT**C**CAAAT	AT**G**CAAAT	AT**T**CAAAT
p = 4	ATG**A**AAAT	ATG**C**AAAT	ATG**G**AAAT	ATG**T**AAAT
p = 5	ATGC**A**AAT	ATGC**C**AAT	ATGC**G**AAT	ATGC**T**AAT
p = 6	ATGCA**A**AT	ATGCA**C**AT	ATGCA**G**AT	ATGCA**T**AT
p = 7	ATGCAA**A**T	ATGCAA**C**T	ATGCAA**G**T	ATGCAA**T**T
p = 8	ATGCAAA**A**	ATGCAAA**C**	ATGCAAA**G**	ATGCAAA**T**

B. Frequency substitution matrices for the test and control sequences

$[T_{pi}] =$

Position	i = 1	i = 2	i = 3	i = 4
p = 1	200	46	43	120
p = 2	42	52	44	200
p = 3	45	40	200	37
p = 4	38	200	57	55
p = 5	200	48	43	145
p = 6	200	52	48	100
p = 7	200	59	42	30
p = 8	80	47	65	200

$[C_{pi}] =$

Position	i = 1	i = 2	i = 3	i = 4
p = 1	50	33	48	43
p = 2	57	58	43	50
p = 3	46	52	50	53
p = 4	42	50	51	46
p = 5	50	52	44	52
p = 6	50	38	43	47
p = 7	50	56	41	31
p = 8	48	37	46	50

C. Subtraction of matrices

Position	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$p = 1$	150	13	-5	77
$p = 2$	-15	-6	1	150
$p = 3$	-1	-12	150	-16
$p = 4$	-4	150	6	9
$p = 5$	150	-4	-1	93
$p = 6$	150	14	5	53
$p = 7$	150	3	1	-1
$p = 8$	32	10	19	150

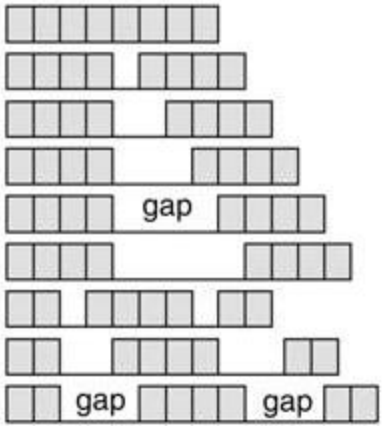
D. Negative values are replaced by zero

Position	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$p = 1$	150	13	0	77
$p = 2$	0	0	0	150
$p = 3$	0	0	150	0
$p = 4$	0	150	6	9
$p = 5$	150	0	0	93
$p = 6$	150	14	5	53
$p = 7$	150	3	1	0
$p = 8$	32	10	19	150

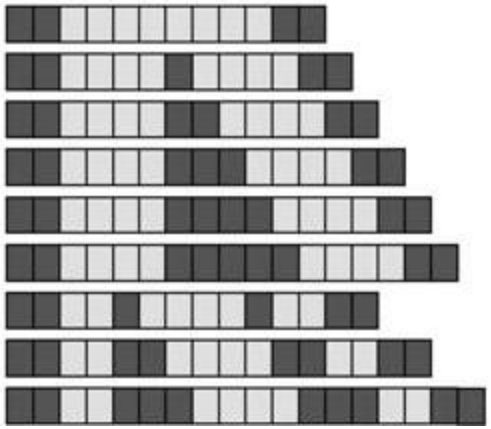
E. Normalized position frequency matrix (PFM)

Position	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$p = 1$	63	5	0	32
$p = 2$	0	0	0	100
$p = 3$	0	0	100	0
$p = 4$	0	91	4	5
$p = 5$	62	0	0	38
$p = 6$	68	6	2	24
$p = 7$	97	2	1	0
$p = 8$	15	5	9	71

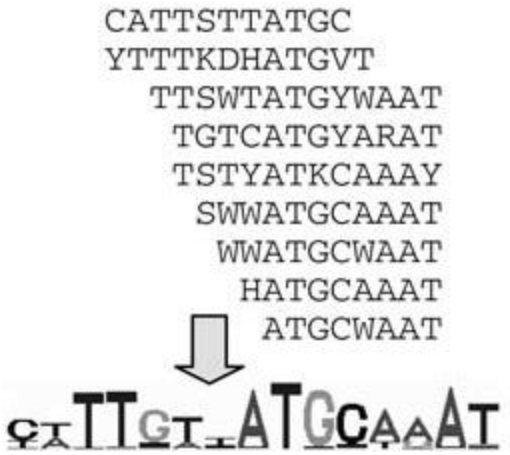
F. Patterns of 8-mer words with and without gaps



G. Filling the gaps and extending PFMs



H. Clustering and combining PFMs



Оценка качества работы программ предсказания сайтов связывания.

TP- true positive, количество правильного обнаружения сайтов в тех последовательностях, в которых эти сайты реально представлены;

TN- true negative, количество не обнаружения сайтов в последовательностях реально не содержащих данный сайт;

FP- false positive, количество обнаружения сайтов в последовательностях реально не содержащих данный сайт;

FN- false negative, количество не обнаружения сайтов в последовательностях реально содержащих данный сайт;

$$E_1 = \frac{FP}{TN + FP}$$

Ошибка первого рода (перепредсказания)

$$E_2 = \frac{FN}{TP + FN}$$

Ошибка второго рода(недопредсказания)

$$E_a = \frac{E_1 + E_2}{2}$$

Средняя ошибка

$$Sn = 1 - E_2 = \frac{TP}{TP + FN}$$

Sn – чувствительность, отражает долю правильно предсказанных сайтов среди всех реальных сайтов

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

Sp – специфичность, отражает долю правильно предсказанных сайтов среди всех предсказаний сайтов

$$C(D,M) = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

Спасибо за внимание!