Биоинформатика

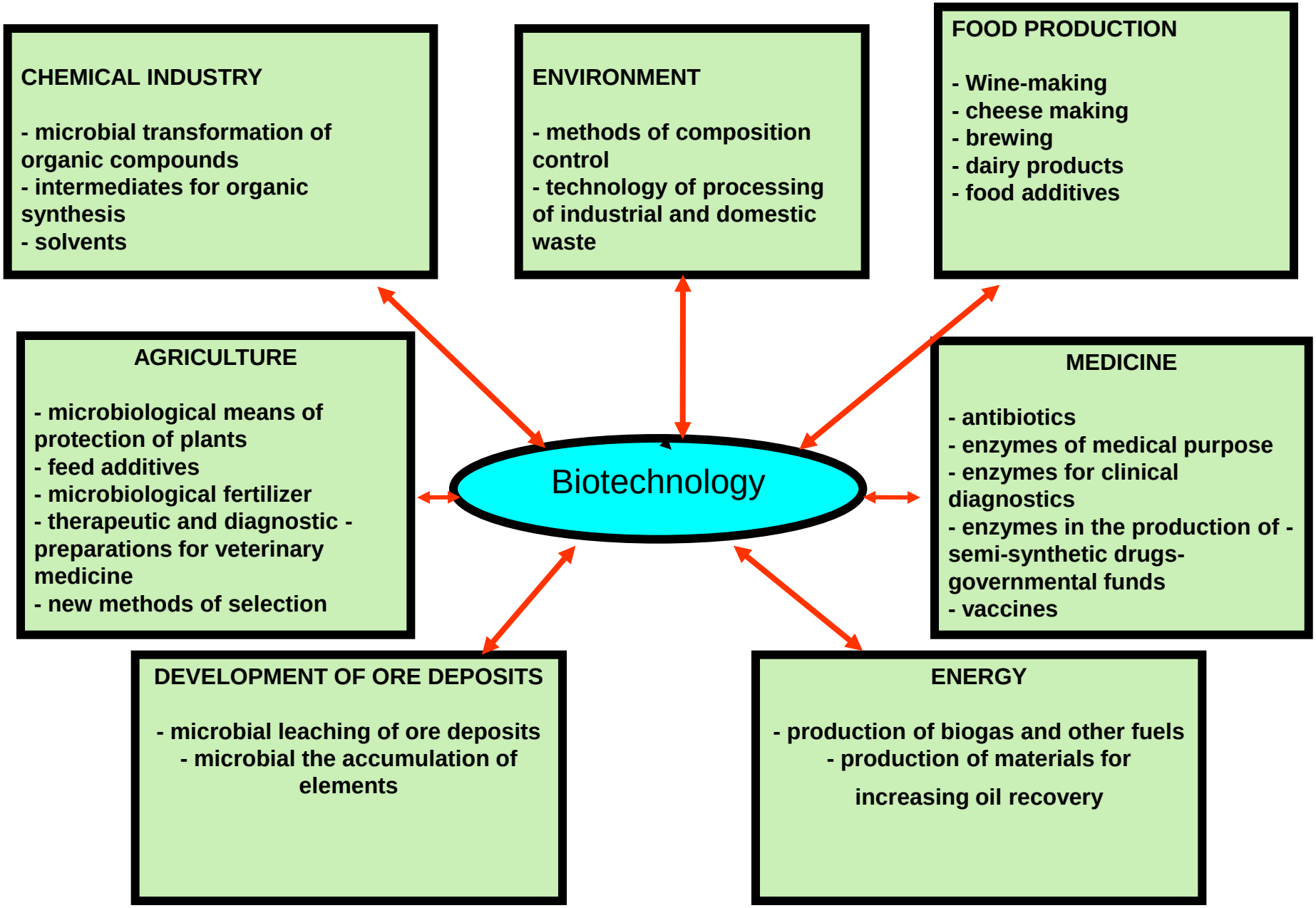
Коротков Евгений Вадимович
Институт Биоинженерии, ФИЦ Биотехнологии РАН

bioinf@yandex.ru

1. Аннотация бактериальных генов.

1. Одиночные и парные точки разладки

1. Поиск мутаций типа сдвиг рамки считывания


CHEMICAL INDUSTRY
- microbial transformation of organic compounds
- intermediates for organic synthesis
- solvents
ENVIRONMENT
- methods of composition control
- technology of processing of industrial and domestic waste
FOOD PRODUCTION
- Wine-making
- cheese making
- brewing
- dairy products
- food additives
AGRICULTURE
- microbiological means of protection of plants
- feed additives
- microbiological fertilizer
- therapeutic and diagnostic - preparations for veterinary medicine
- new methods of selection
Biotechnology
MEDICINE
- antibiotics
- enzymes of medical purpose
- enzymes for clinical diagnostics
- enzymes in the production of - semi-synthetic drugs- governmental funds
- vaccines
DEVELOPMENT OF ORE DEPOSITS
- microbial leaching of ore deposits
- microbial the accumulation of elements
ENERGY
- production of biogas and other fuels
- production of materials for
increasing oil recovery

Features of microorganisms

1. Universal distribution
2. High speed of growth and reproduction
3. A high degree of survivability (t=70-105C, radiation, NaCI=25-30%, drying, lack of oxygen
4. Microorganisms have a haploid genome, which allows to identify any mutations in the first generation
5. Incredible productivity. For example: During the day, 500 kg of soybean producing 5 kg of a protein. The yeast is able to produce 50 tons of a protein in the bioreactor .
6. Only several years are needed for deducing of microorganism strain.

Selection of microorganisms

Traditional methods:

Artificial mutagenesis.

The selection of strains on base of a productivity

The latest methods - genetic engineering

-The selection of a gene from one microorganism genome and its entry into the genome of another microorganism

- Synthesis gene artificially and introducing it into the genome of the organism

Gene annotation

- There are tens of thousands of nucleotide sequences of genomes of bacteria, biological function known only to ~55% of bacterial genes

- It means that some millions of genes have the known sequences but biological function of these sequences is unknown.

- **The task is to determine their biological function and, thus, to make available for use in biotechnology.**

Phylogenetic profile

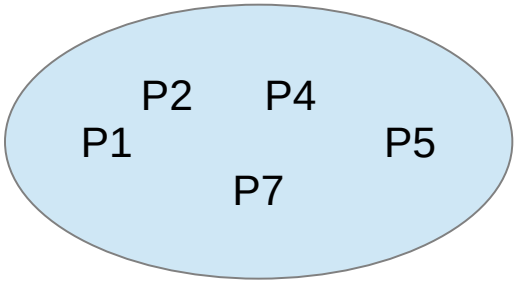
- The profile shows the presence or absence of a gene in selected (reference) genomes of bacteria

- Genes that have a similar profile to perform the same biological function or participate in the same metabolic pathway

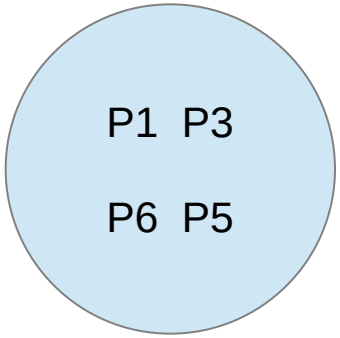
Annotation of gene by the method of functional groups.

- Known annotation methods are based on sequence similarity (dynamic programming, heuristic algorithms, HMM) and predict gene functions in the case of relatively high level of similarity (over 70%).
- **Orthologues** are genes having the similarity and performing the same biological function.
- **Paralogs** are genes that have considerable similarity, but various biological functions.
- Tasks is to find gene-**orthologues** on the background of a big number of genes-**paralogs**

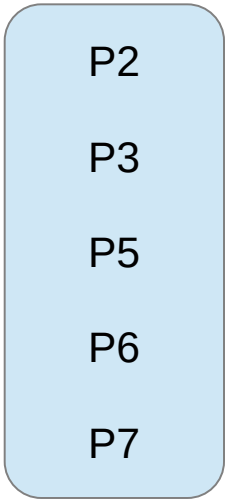
Аннотирование генов методом функциональных групп.
Филогенетический профиль гена (ФП).



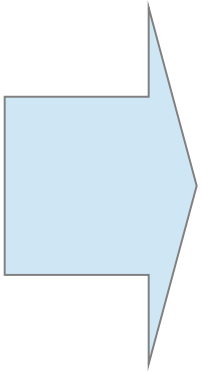
S.cerevisae (SC)



H.influenzae



B.subtilis

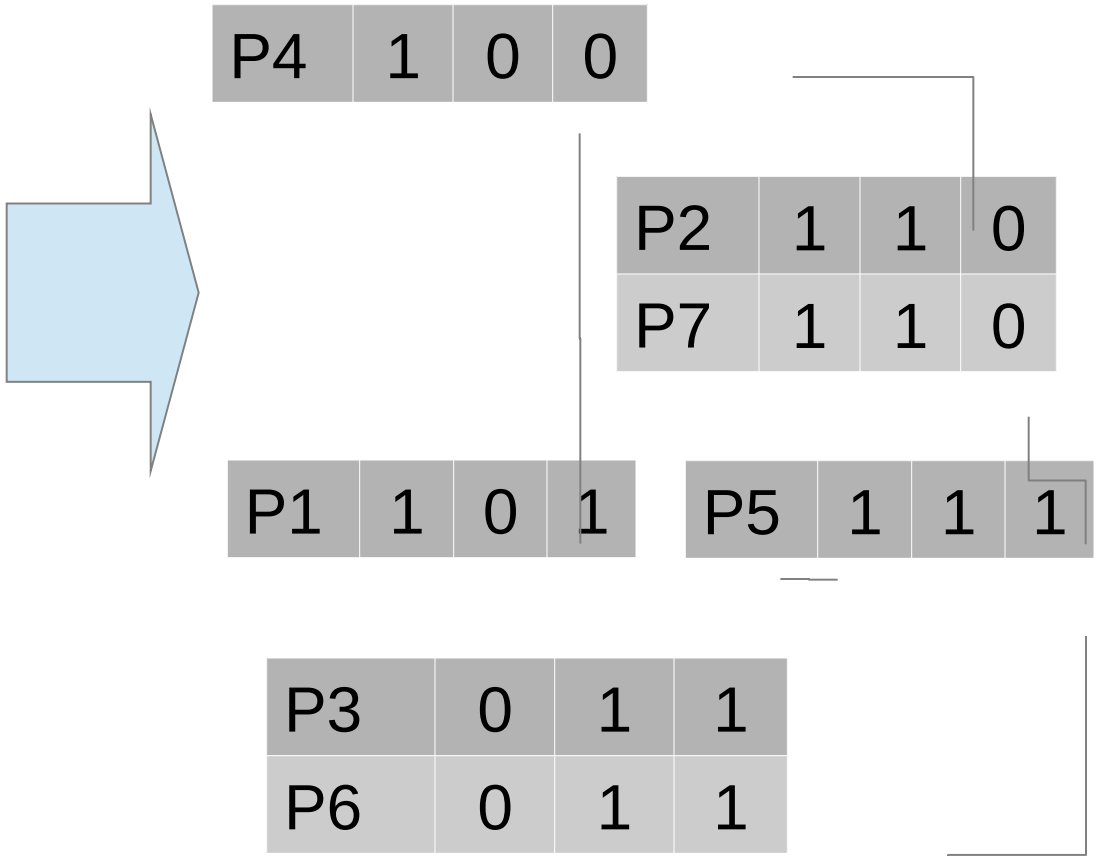


Филогенетические профили:			
	SC	**BS**	**HI**
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0

Pelligrini et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Biochemistry, 1999

Аннотирование генов методом функциональных групп.
Филогенетический профиль гена (ФП).

Филогенетические профили:			
	SC	**BS**	**HI**
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0



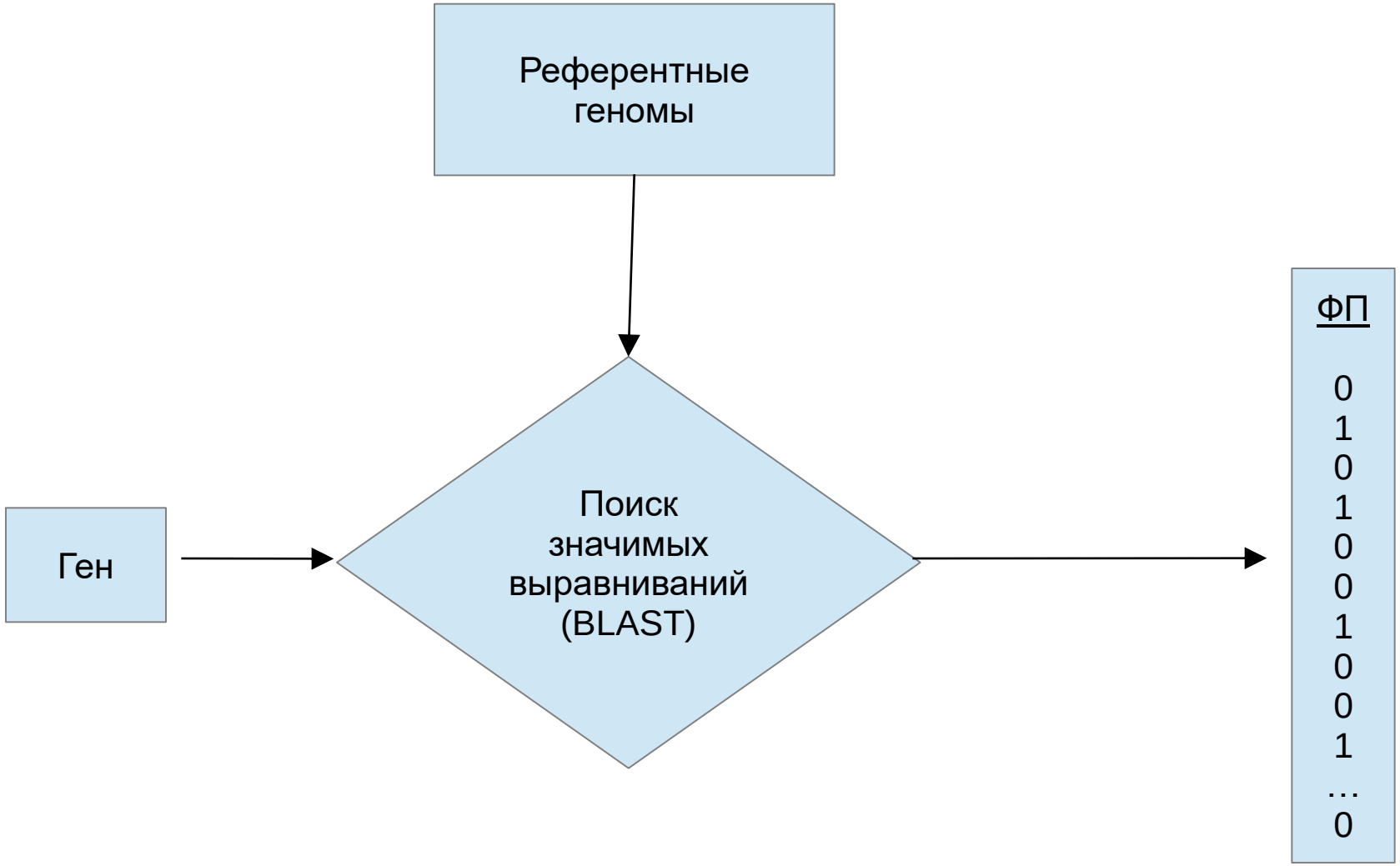
Основная идея:

Гены со сходными векторами могут иметь сходные функции.

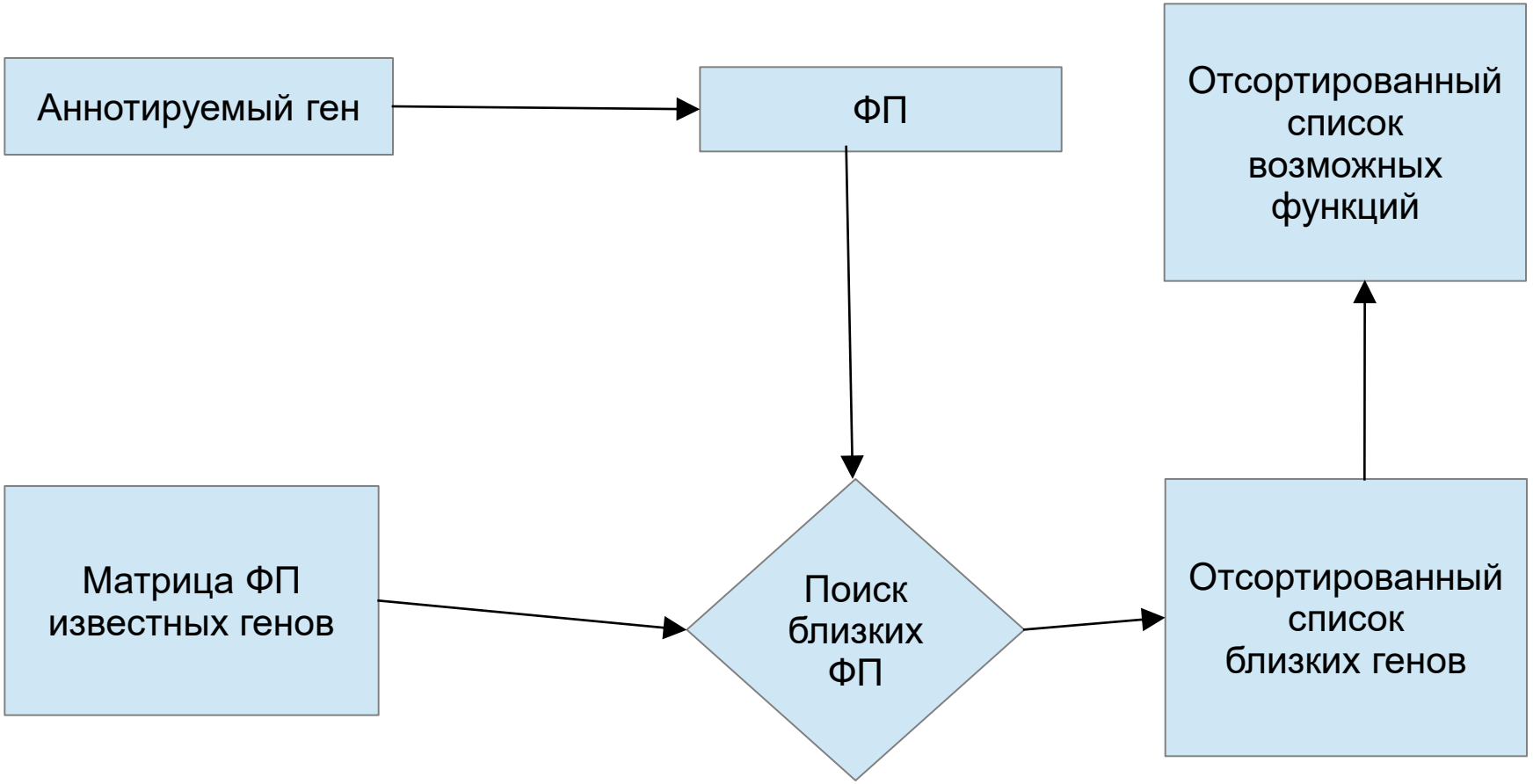
Аннотирование генов методом функциональных групп. Подготовка исходных данных

- Высоко гомологичные геномы, например, несколько штаммов одной бактерии, ухудшают точность предсказания — факты встречаемости в них гена не являются независимыми событиями.
- В качестве референтных геномов было выбрано 1200 из 2100 геномов.
- Для составления матрицы ФП генов с известными функциями были выбраны 3.7млн генов.

Аннотирование генов методом функциональных групп.
Алгоритм формирования ФП гена



Аннотирование генов методом функциональных групп.
Предсказание возможных функций гена



Аннотирование генов методом функциональных групп.
Предсказание функции гена. Пример.

KEGG Entry: SEN1936

Функция: phage capsid protein

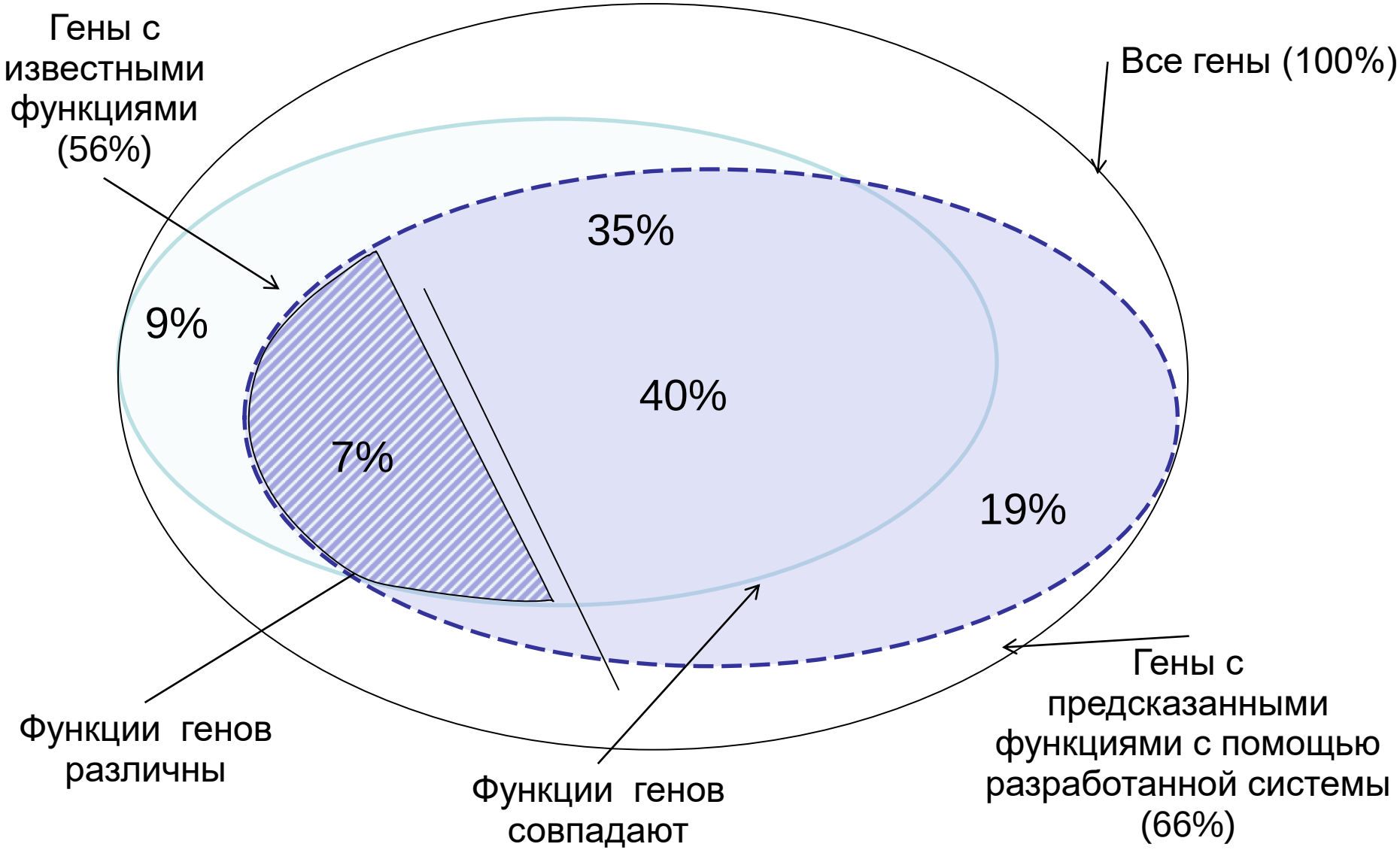
Вероятность	**Функция**
5.03427e-17	phage capsid protein
5.03427e-17	capsid
4.02153e-16	phage portal protein
4.02153e-16	Portal protein
4.02153e-16	HK97 family phage prohead protease
4.02153e-16	HK97 family phage portal protein
2.01035e-14	major capsid protein
1.40422e-13	phage protease
1.40422e-13	phage phi-105 ORF25-like protein
1.40422e-13	phage capsid protease

Аннотирование генов методом функциональных групп.
Исходные данные для оценки качества

Методы аннотирования	**Число геномов**
UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, InterPro, IMG-ER	38
BLAST, homology	28
GenDB, BLAST, COG, COGnitor	7
InterPro(Scan)	5
Все	104

Общее число генов: 375152.

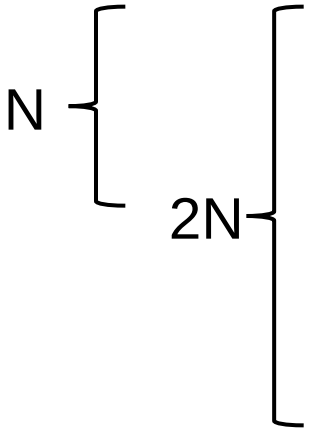
Аннотирование генов методом функциональных групп.
Оценка качества предсказания функций генов.



Аннотирование генов методом функциональных групп.
Оценка качества предсказания функций генов.

На какой позиции нашли известную функцию	% генов	Число генов
1	63,23%	108806
2	13,89%	23894
3	4,94%	8498
4	2,58%	4446
5	1,59%	2743
6	1,13%	1949
7	0,83%	1433
8	0,65%	1127
9	0,56%	962

Аннотирование генов методом функциональных групп.
Оценка качества предсказания группы на основе метаболических путей



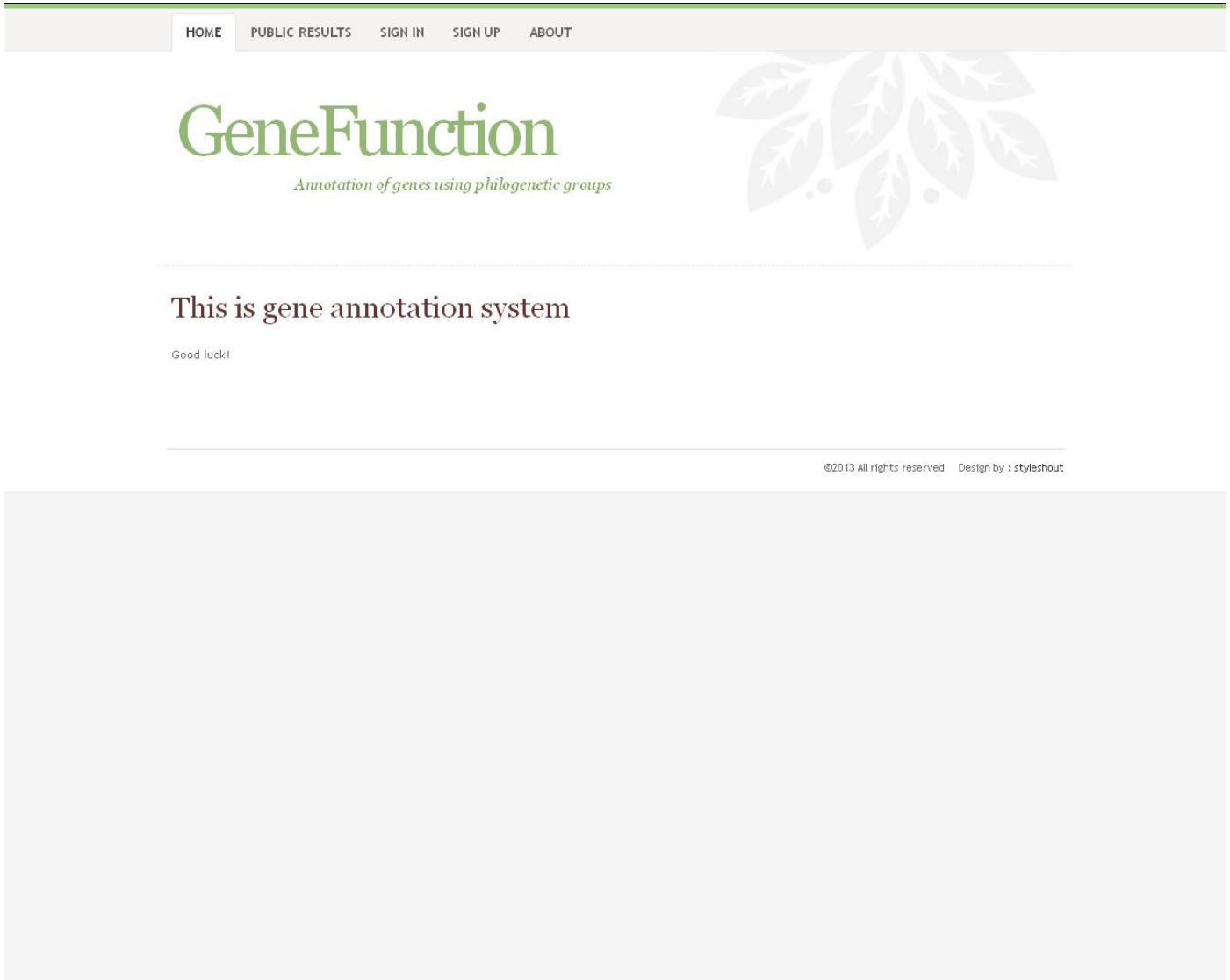
Вероятность	Функция
5.03427e-17	phage capsid protein
5.03427e-17	capsid
4.02153e-16	phage portal protein
4.02153e-16	Portal protein
4.02153e-16	HK97 family phage prohead protease
4.02153e-16	HK97 family phage portal protein
…	…

Длина метаболи… пути, N							
Первые N функций	0.496	0.276	0.197	0.218	0.498	0.358	0.268
Первые 2N функций	0.588	0.530	0.572	0.538	0.543	0.578	0.636

Аннотирование генов методом функциональных групп.

- Разработанная система предсказывает функции для 65% из всех рассматриваемых генов, при этом для 19% генов функция ранее была не определена.
- Для 7% генов предсказываемая функция отличается от существующей
- Нужно функционально аннотировать заново все уже известные бактериальные геномы и сделать доступными для биотехнологии десятки миллионов новых генов.

Web-сайт для аннотации бактериальных генов http://genefunction.ru

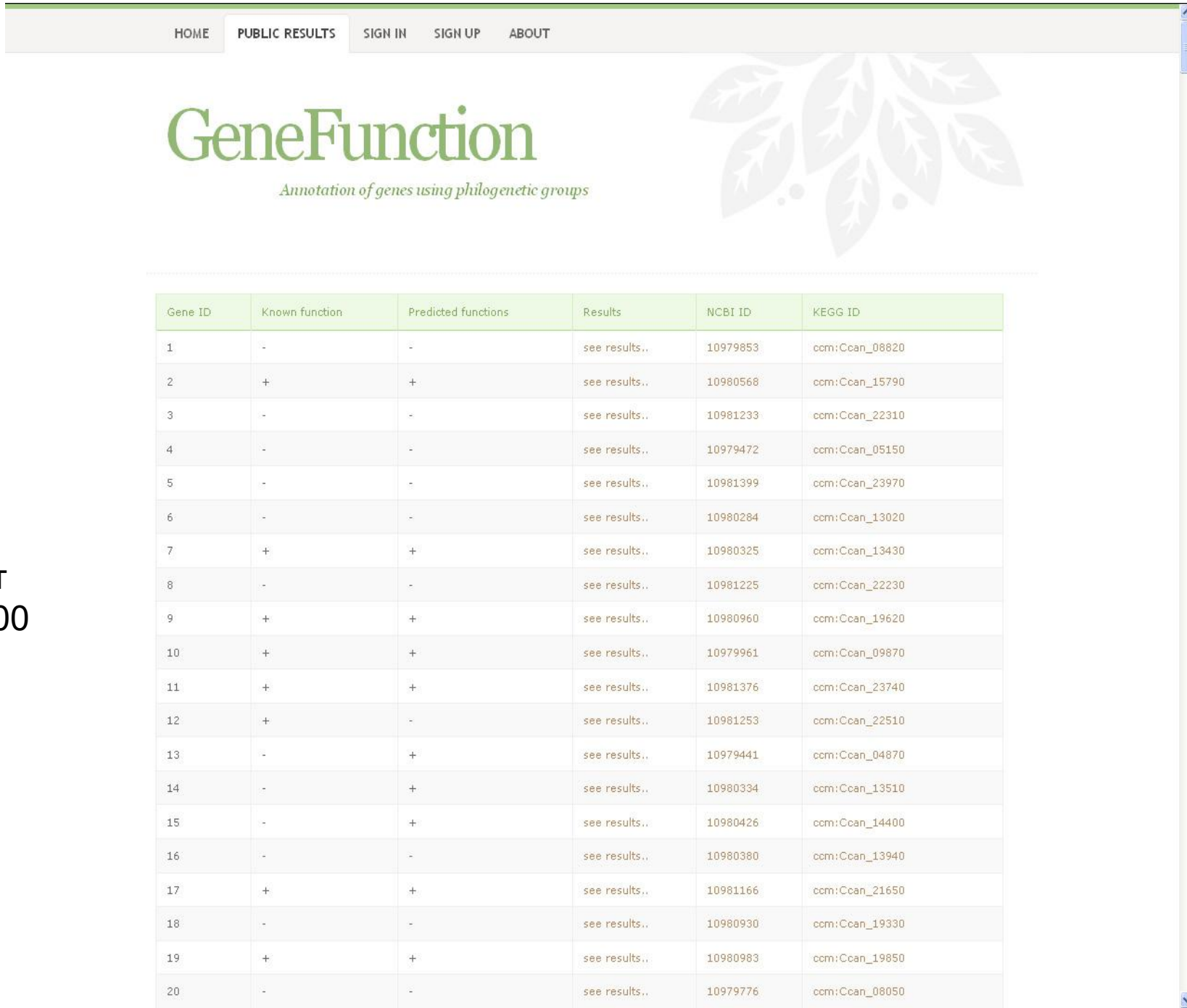


Web-сайт для аннотации бактериальных генов http://genefunction.ru

104 аннотированных генома бактерий

19% «новых» генов

Genome	RefSeq ID	KEGG ID
Acetobacterium woodii	NC_016894	T01732
Acholeplasma laidlawii	NC_010163	T00659
Aeromonas hydrophila	NC_008570	T00424
Alteromonas sp. SN2	NC_015554	T01513
Aminobacterium colombiense	NC_014011	T01202
Amycolatopsis mediterranei S699	NC_017186	T01822
Anaplasma phagocytophilum	NC_007797	T00327
Anoxybacillus flavithermus	NC_011567	T00790
Arcanobacterium haemolyticum	NC_014218	T01242
Arcobacter nitrofigilis	NC_014166	T01235
Atopobium parvulum	NC_013203	T00983
Azorhizobium caulinodans	NC_009937	T00609
Baumannia cicadellinicola	NC_007984	T00349
Beutenbergia cavernae	NC_012669	T00893
Candidatus Solibacter usitatus	NC_008536	T00412
Capnocytophaga canimorsus	NC_015846	T01582
Catenulispora acidiphila	NC_013131	T00967
Cellulophaga algicola	NC_014934	T01404
Chitinophaga pinensis	NC_013132	T00959
Chlamydophila abortus	NC_004552	T00242



Gene ID	Known function	Predicted functions	Results	NCBI ID	KEGG ID
1	-	-	see results..	10979853	ccm:Ccan_08820
2	+	+	see results..	10980568	ccm:Ccan_15790
3	-	-	see results..	10981233	ccm:Ccan_22310
4	-	-	see results..	10979472	ccm:Ccan_05150
5	-	-	see results..	10981399	ccm:Ccan_23970
6	-	-	see results..	10980284	ccm:Ccan_13020
7	+	+	see results..	10980325	ccm:Ccan_13430
8	-	-	see results..	10981225	ccm:Ccan_22230
9	+	+	see results..	10980960	ccm:Ccan_19620
10	+	+	see results..	10979961	ccm:Ccan_09870
11	+	+	see results..	10981376	ccm:Ccan_23740
12	+	-	see results..	10981253	ccm:Ccan_22510
13	-	+	see results..	10979441	ccm:Ccan_04870
14	-	+	see results..	10980334	ccm:Ccan_13510
15	-	+	see results..	10980426	ccm:Ccan_14400
16	-	-	see results..	10980380	ccm:Ccan_13940
17	+	+	see results..	10981166	ccm:Ccan_21650
18	-	-	see results..	10980930	ccm:Ccan_19330
19	+	+	see results..	10980983	ccm:Ccan_19850
20	-	-	see results..	10979776	ccm:Ccan_08050

Геном каждой бактерии от 1000 до 5000 генов

- Старая и новая функции совпадают

Web-сайт для аннотации бактериальных генов http://genefunction.ru

Функция определена впервые

Change Point of TP

one type of TP another type of TP

Pair Change Point of TP

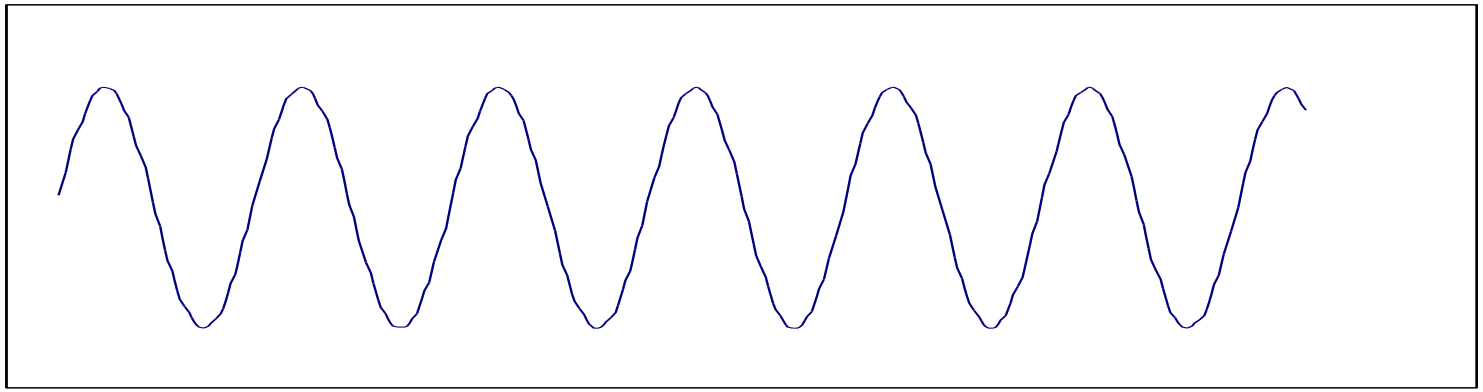
- **Task: to develop the mathematical method for revealing of the triplet periodicity (TP) change points and pair change points genes**

- **The method should use the gene sequence, any external parameters should be absent**

- **There are ~2400 types of triplet periodicity in genes**

Frenkel FE, Korotkov EV. Classification analysis of triplet periodicity in protein-coding regions of genes. Gene. 2008. 15;421(1-2):52-60. 2008

Gene triplet periodicity



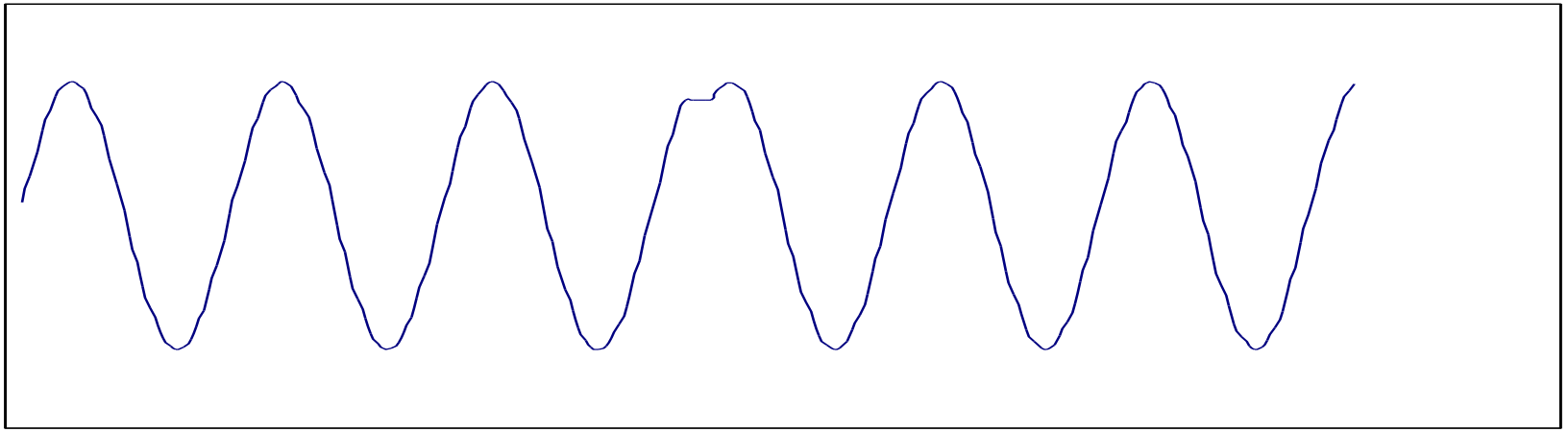
+1

123123123123123123123123123123123...

atggcttcgatccattcggctagagacatcgaatca

Triplet periodicity exists in gene if positions 1, 2 and 3 have the different base frequencies

Splicing of the two different types of the triplet periodicity

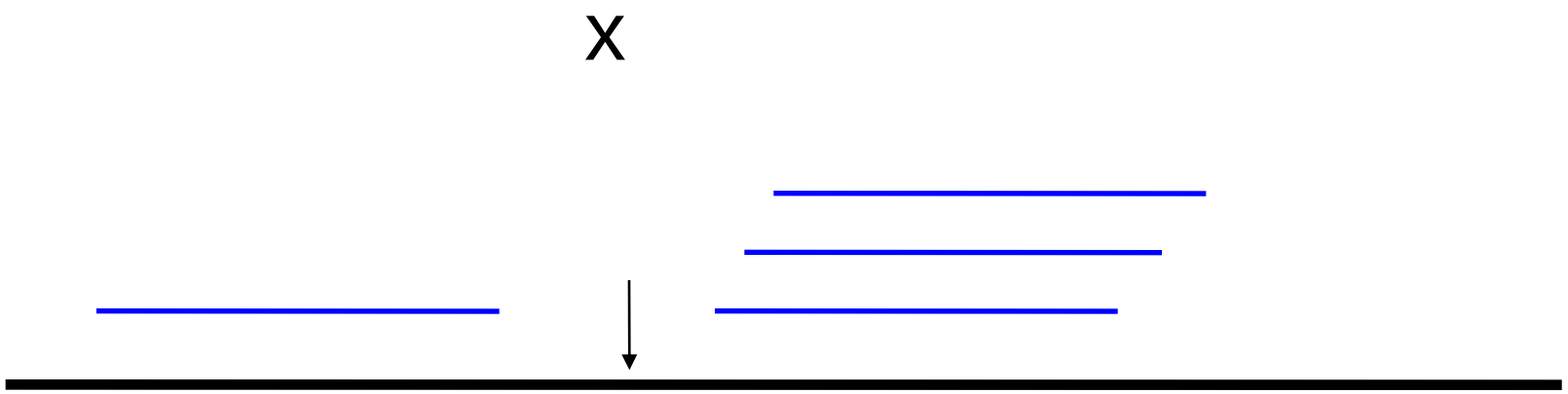


+1

123123123123123123123123123123123...

atggcttcgatccattcggctagagacatcgaatcat...

Algorithm of the search of the triplet periodicity splicing



Algorithm of the search of the triplet periodicity splicing

231231231231231231231231 N_3

312312312312312312312312 N_2

123123123123123123123123123123123123 N_1 and M_1

atctgatcgatggctagctagatttgatcgctggctcatcg

	1	2	3
a	3	0	1
t	2	2	1
c	0	2	2
g	1	2	2

M_1

	1	2	3
a	1	2	1
t	3	2	3
c	3	2	0
g	0	1	3

N_1

	1	2	3
a	1	1	2
t	3	3	2
c	0	3	2
g	3	0	1

N_2

	1	2	3
a	2	1	1
t	2	3	3
c	2	0	3
g	1	3	0

N_3

Conditions for triplet periodicity splicing

$$\begin{Bmatrix} V\{M_1(1,\ x),\ N_1(x+1,\ L)\} & \le V_0 \\ V\{M_1(1,\ x),\ N_2(x+1,\ L)\} & > V_0 \\ V\{M_1(1,\ x),\ N_3(x+1,\ L)\} & > V_0 \end{Bmatrix}$$

The splicing of the triplet periodicity is absent at the *x* position

$$\begin{Bmatrix} V\{M_1(1,\ x),\ N_2(x+1,\ L)\} & > V_0 \\ V\{M_1(1,\ x),\ N_1(x+1,\ L)\} & > V_0 \\ V\{M_1(1,\ x),\ N_3(x+1,\ L)\} & > V_0 \end{Bmatrix}$$

The splicing of the triplet periodicity is present at the *x* position

***V* – measure of dissimilarity of two compared matrixes**

Example of splicing of the triplet periodicity

- 123123123123123123123123123123123123 - RF *T*1
- 312312312312312312 - RF *T*2
- 231231231231231231 - RF *T*3
- atgatgatgatgatgatgcgtcgtcgtcgtcgtcgt

x

	1	2	3
a	6	0	0
t	0	6	0
c	0	0	0
g	0	0	6

M_1

	1	2	3
a	0	0	0
t	0	0	6
c	6	0	0
g	0	6	0

N_1

	1	2	3
a	0	0	0
t	0	6	0
c	0	0	6
g	6	0	0

N_2

	1	2	3
a	0	0	0
t	6	0	0
c	0	6	0
g	0	0	6

N_3

V-MEASURE

	1	2	3	
a	m_{11}	m_{21}	m_{31}	$x(1)$
t	m_{12}	m_{22}	m_{32}	$x(2)$
c	m_{13}	m_{23}	m_{33}	$x(3)$
g	m_{14}	m_{24}	m_{34}	$x(4)$
	$y(1)$	$y(2)$	$y(3)$	

$$p(i, j) = \frac{x(i)y(j)}{L^2}$$

$$L = \sum_{i=1}^{4} x(i) = \sum_{j=1}^{3} y(j)$$

V-MEASURE

$$Z(i,j) = \frac{m(i,j) - Lp(i,j)}{\sqrt{Lp(i,j)(1-p(i,j))}}$$

Z(i,j) has ~*N*(0,1) distribution

***Lp(i,j)*- expectation value,** ***Lp(i,j)(1-p(i,j))* - dispersion**

Z(i,j) calculated for M_1 matrix, and N_k, k=1,2,3

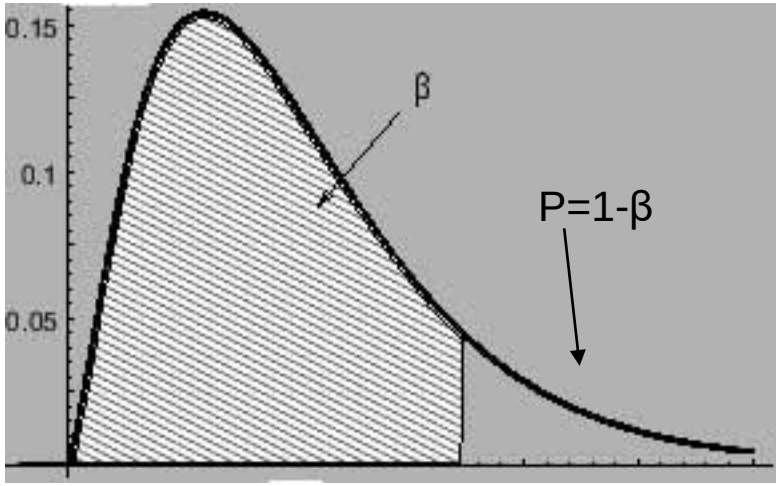
$$D(M_1, N_k) = \sum_{i=1}^{4}\sum_{j=1}^{3}\left(\frac{Z_{M1}(i,j) - Z_{Nk}(i,j)}{\sqrt{2}}\right)^2 \qquad k=1,2,3$$

$D(M_1, N_k)$ has ~χ2 distribution with 6 degrees of freedom

$Z(k)= \sqrt{2D(M_1, N_k)} - \sqrt{11.0}$

V-MEASURE

X^2 distribution



P11= prob($D(M_1,N_1)\geq D_0$)

P12= prob($D(M_1,N_2)\geq D_0$)

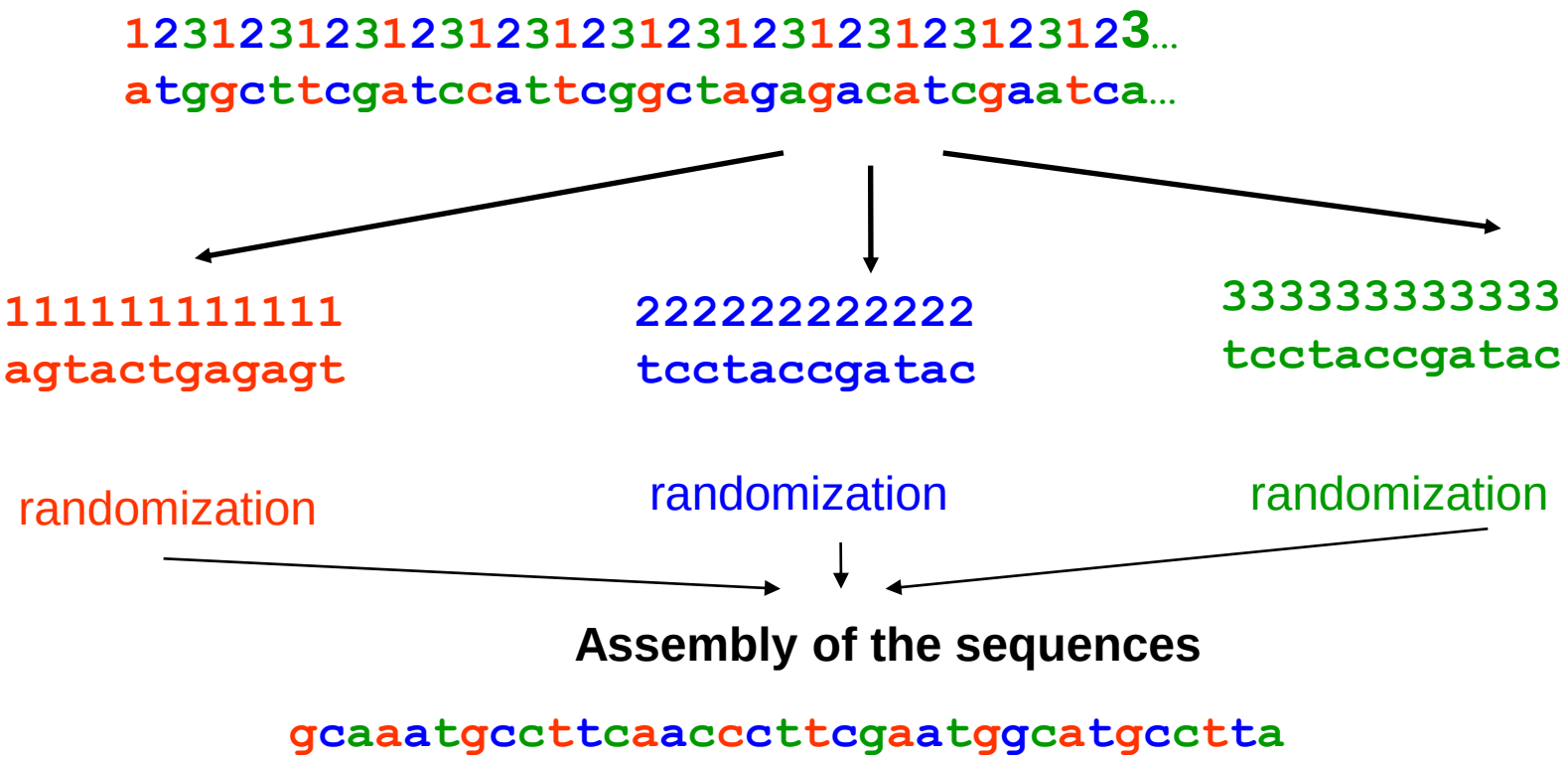
P13= prob($D(M_1,N_2)\geq D_0$)

- *F1=-logP11*
- *F2=-logP12*
- *F3=-logP13*

*F*1>*F*0; *F*2>*F*0; *F*3>*F*0

Monte-Carlo calculations for cutoff level of the *F1* and *F2*

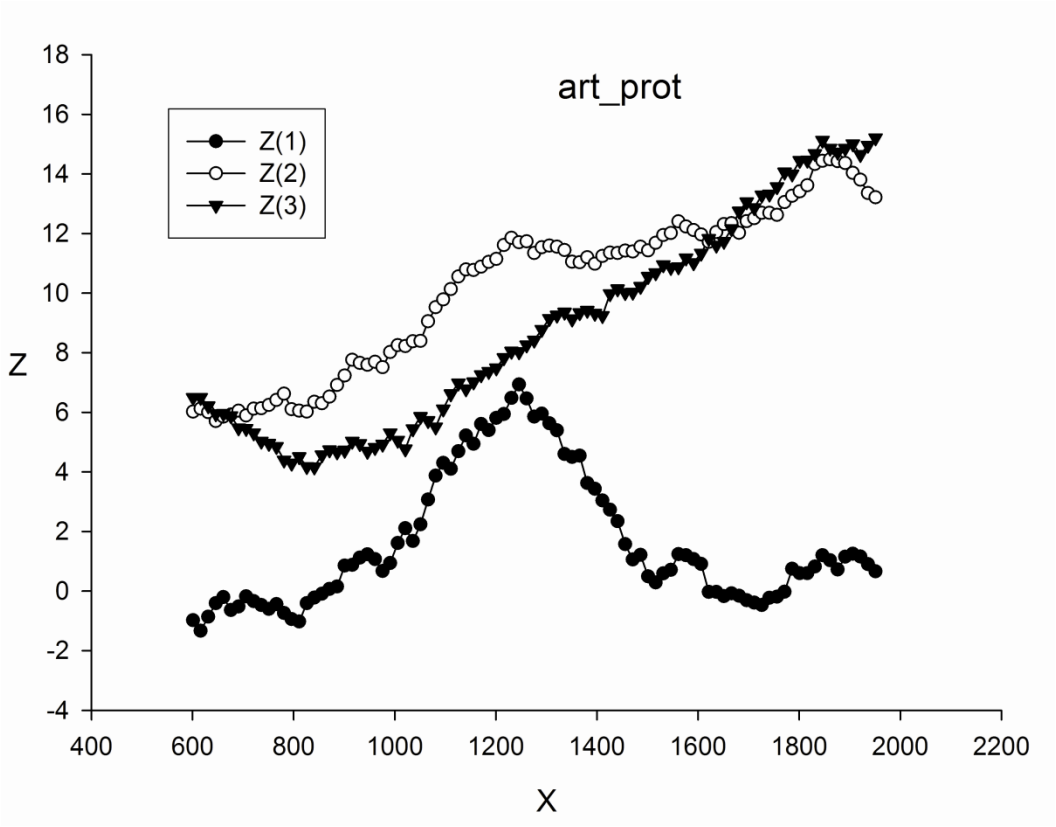
Random bank of gene sequences was produced with saving of the triplet periodicity of the genes.



Monte-Carlo calculations for cutoff level of the *F0*

***F*0=5.0** has 5% of the false positives

Z(1), *Z*(2) and *Z*(3) for artificial gene



The first part (1-1224 bp) is the first half of the gene PD1767 coding the DNA **topoisomerase** from genome of the *X.fastidiosa.* The second part of the artificial gene (1225-2553 bp) is the first half of the gene XAC4270 coding the **glycerol-3-phosphate acyltransferase** from genome of the X.axonopodis.

Uniformity of triplet periodicity in gene sequence

$X_1 = 1+3n$, $n = 0,1,2,3,\ldots$

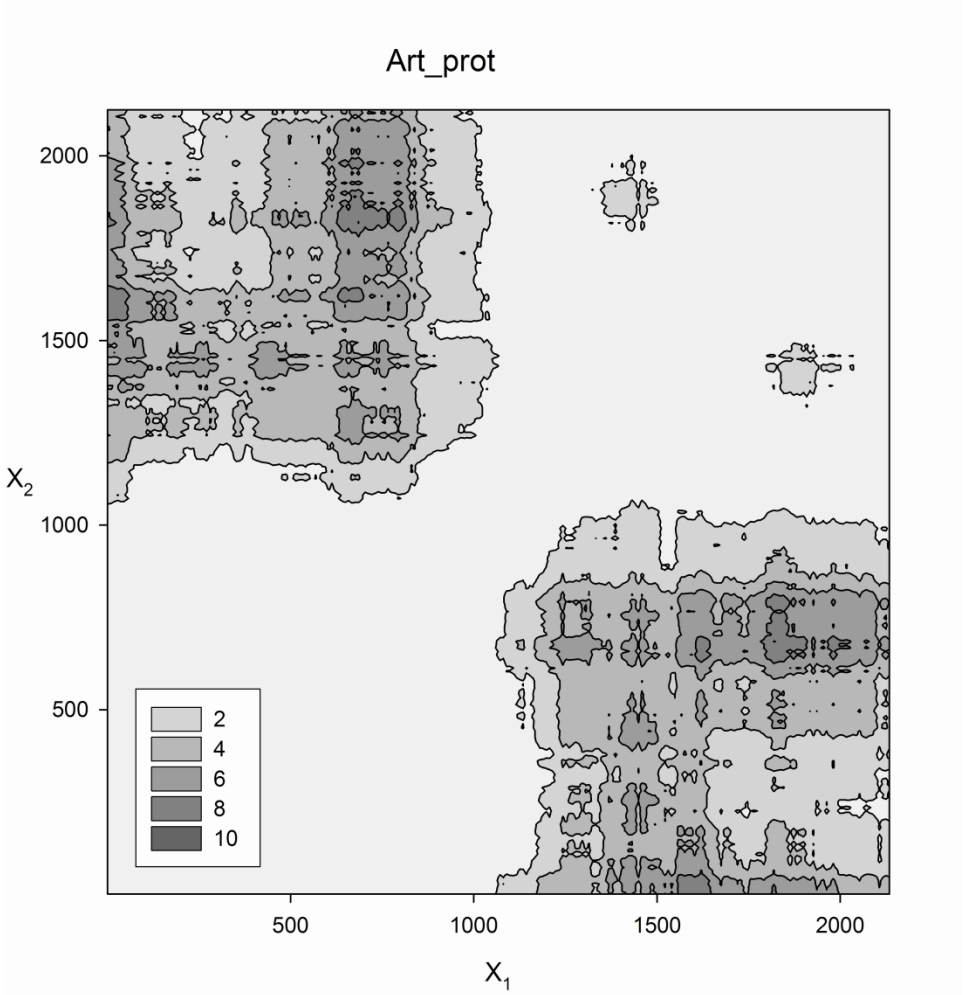
$X_2 = 1+3n$, $n = 0,1,2,3,\ldots$

	1	2	3
a			
t			
c			
g			

We calculated *M1* and *N1* matrixes and calculated $D(M_1,N_1)$

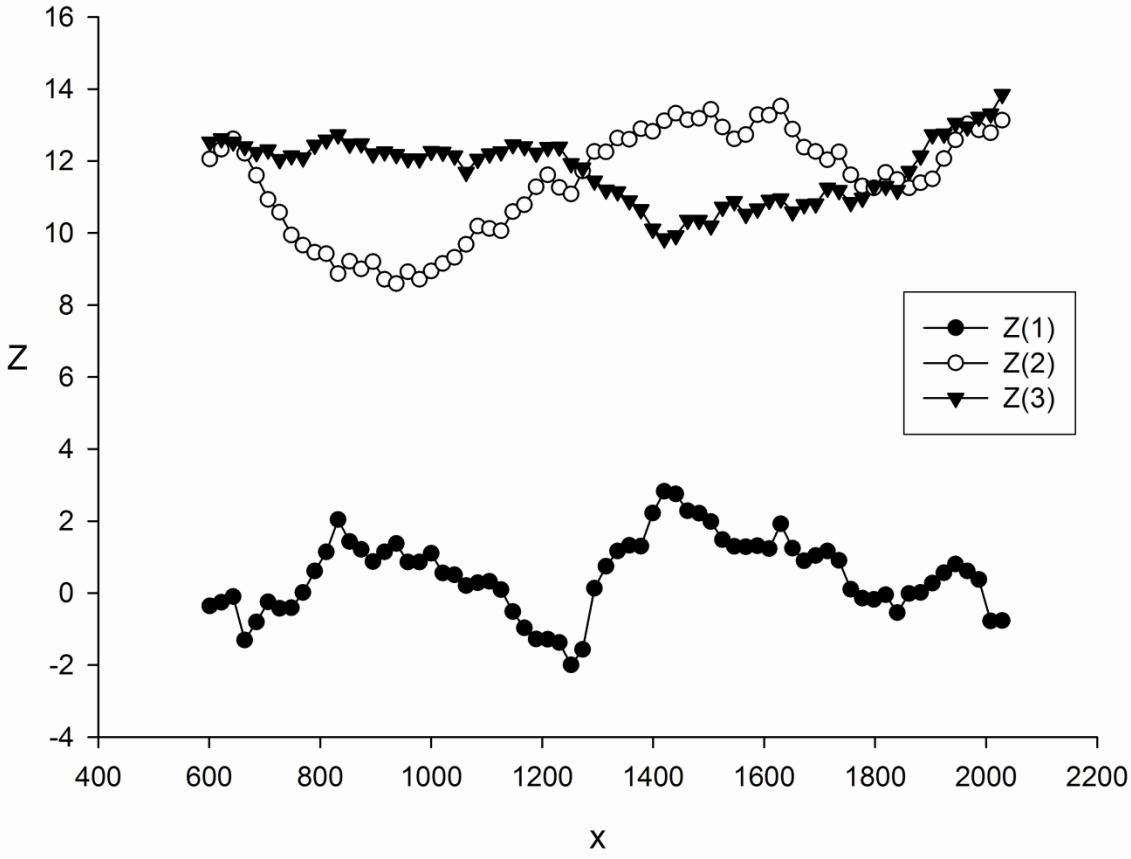
Than *D(M1,N1)* was transformed to the *N(0,1)* distribution *Z*(1)

Contour plot for artificial sequence



Contour plot shows the difference of two matrixes of the triplet periodicity. *X*1 and *X*2 are the first bases of codon in gene.

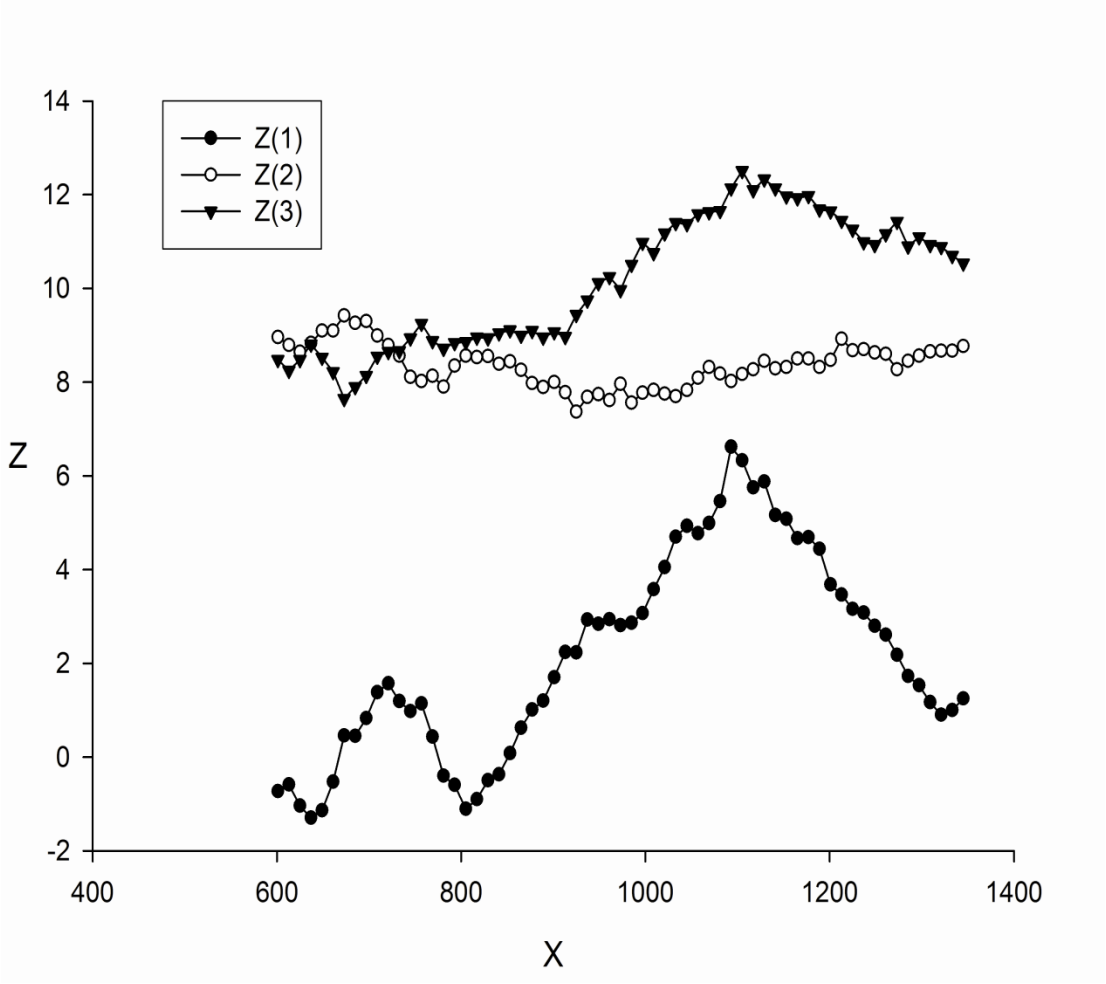
Z(1), Z(2) and Z(3) for Acid345_0008 gene



It is a typical case of gene with uniform TP.

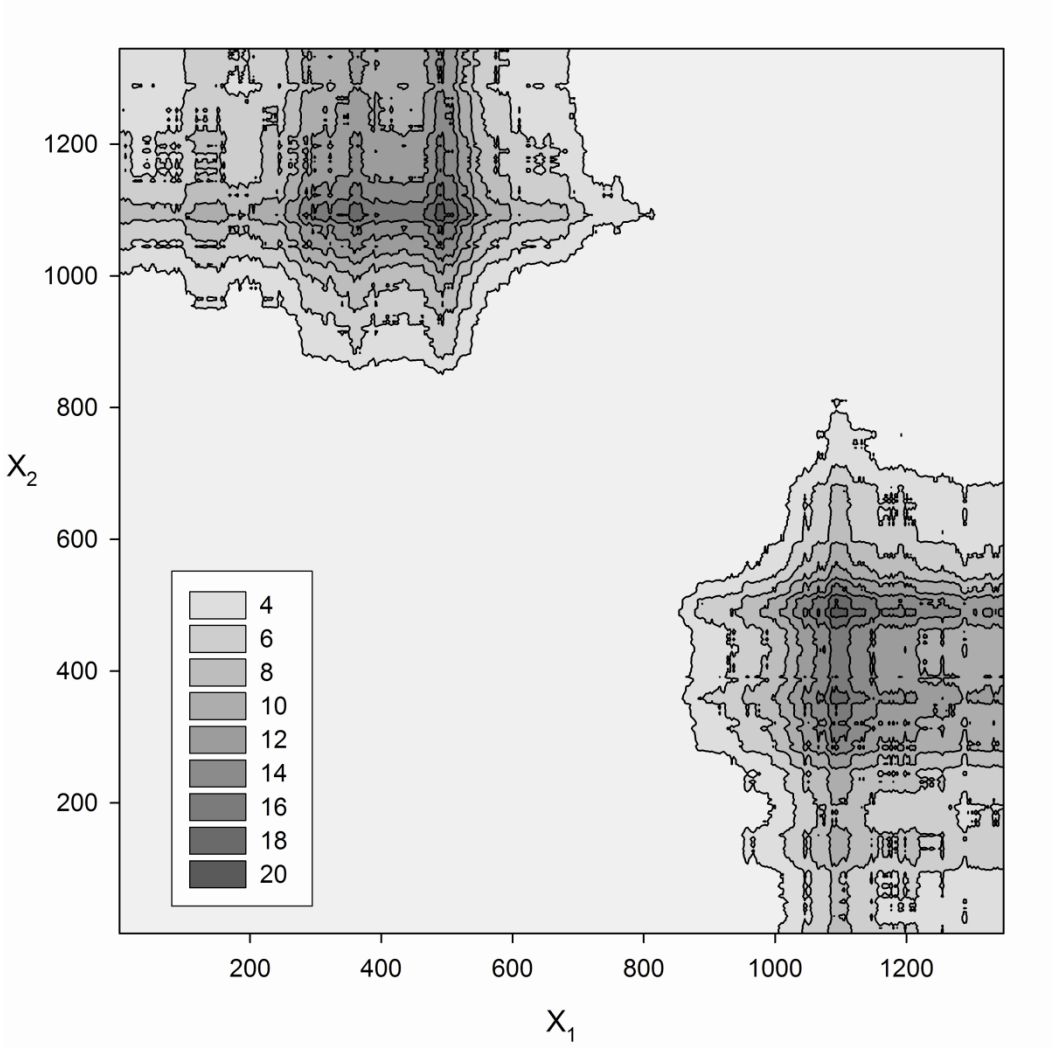
Gene Acid345_0008 is coding **DNA gyrase subunit B from** *A.bacterium* genome.

Splicing of two different TP's for ECP_0691 gene



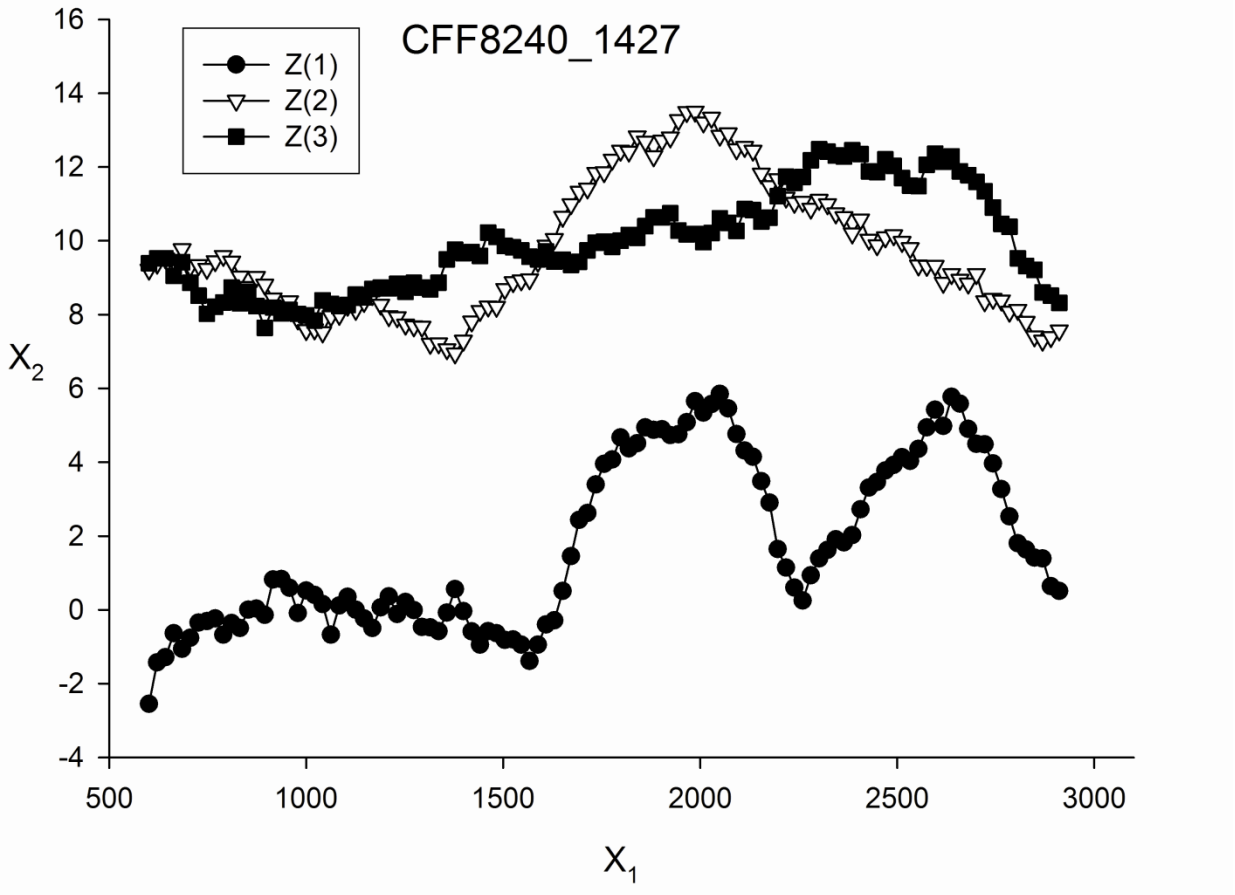
This gene is coding the **N-acetylglucosamine-** specific IIA component from E.coli_336 genome

Contour plot for ECP_0691 gene



Contour plot shows the difference of two matrixes of the triplet periodicity. *X*1 and *X*2 are the first bases of codon in gene.

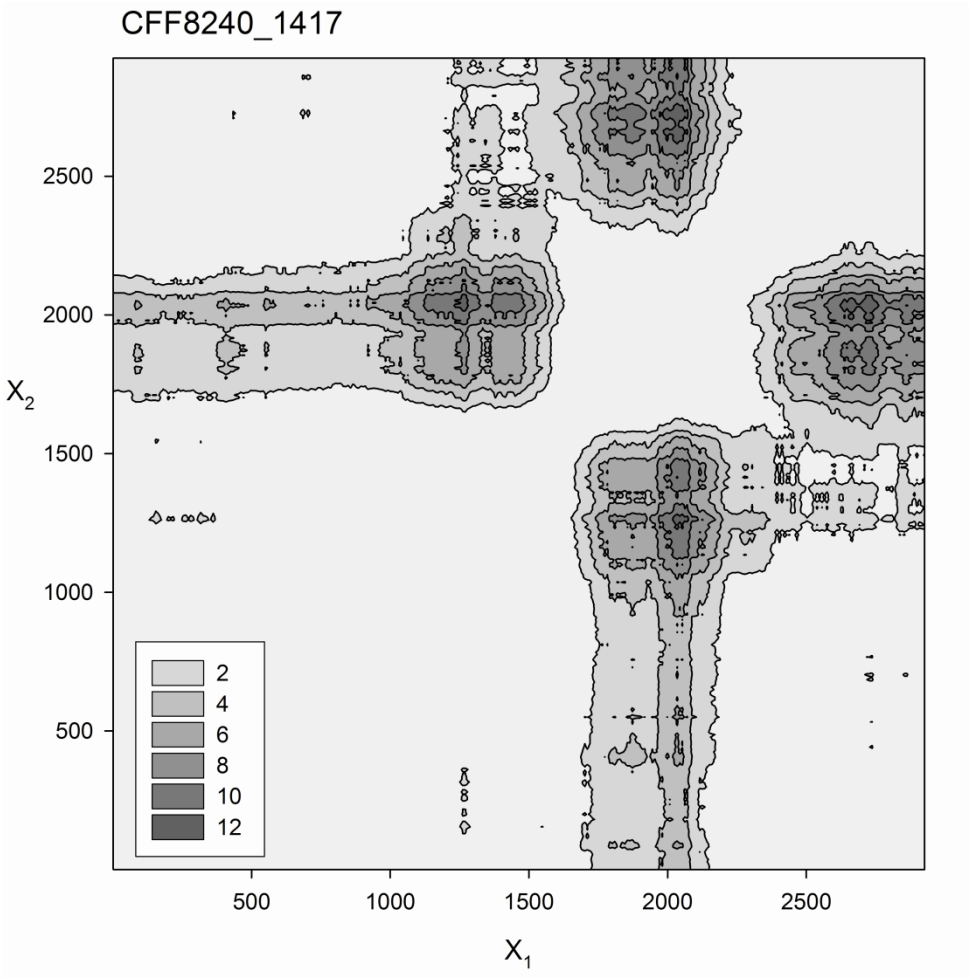
Z(1), Z(2) and Z(3) for CFF8240_1417 with two TPS.



It is possible to see that the first TPS is for *x*≈2100 bp and second TPS is for *x*≈2700 bp.

This gene is coding the **serine protease** from *C.fetus* genome.

Contour plot for the CFF8240_1417 gene



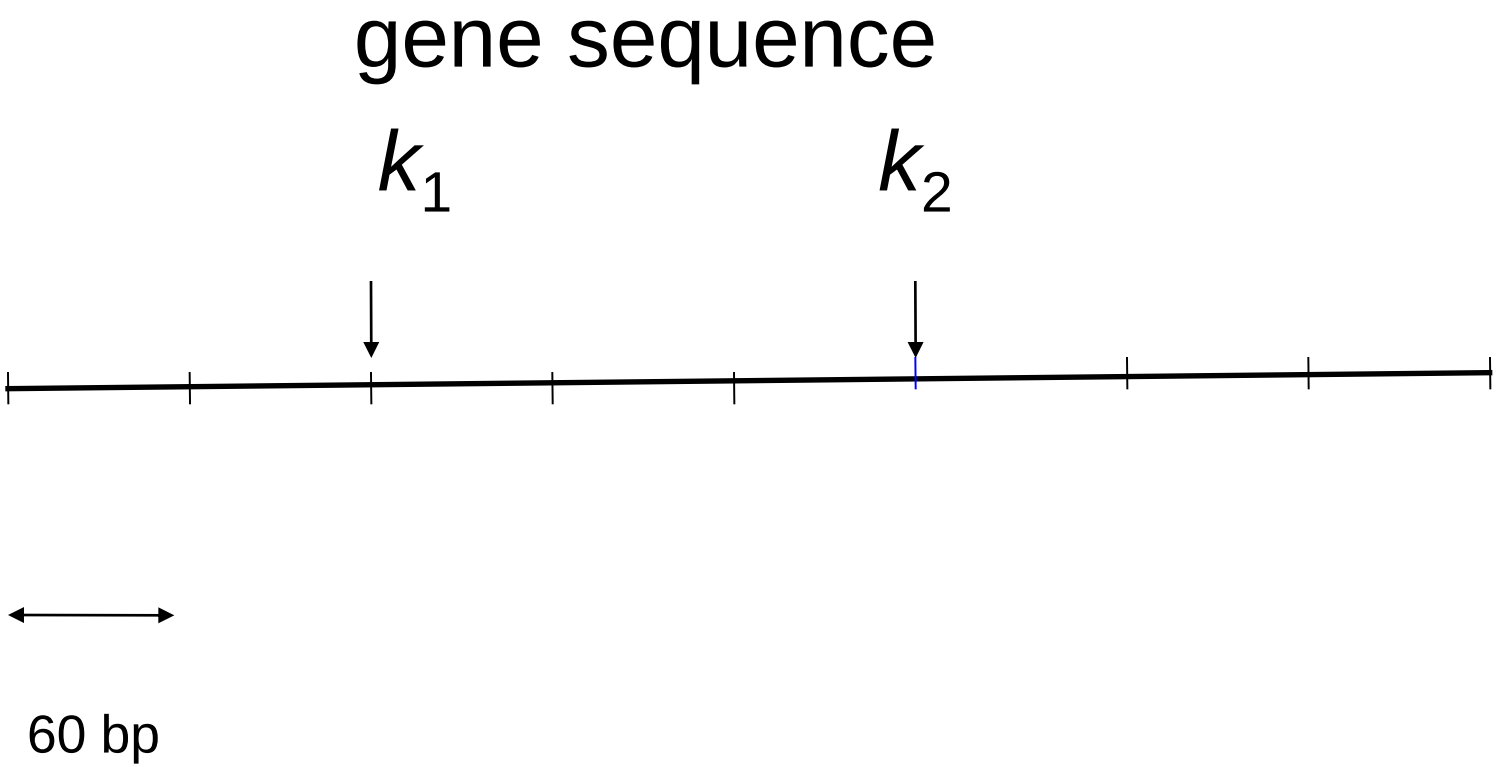
Number of genes with TPS found in KEGG with the same biological function

Gene function definition	Number of genes
translation initiation factor IF-2	411
acriflavin resistance protein	384
PE-PGRS family protein	304
ABC transporter related	264
TonB-dependent receptor	255
major facilitator transporter	245
Serine/threonine protein kinase	217
methyl-accepting chemotaxis sensory transducer	209
integral membrane protein	205
TPR repeat-containing protein	197
binding-protein-dependent transport systems inner membrane component	196
exodeoxyribonuclease VII large subunit (EC:3.1.11.6)	189
glycosyl transferase family protein	173
methyl-accepting chemotaxis protein	149
PE-PGRS family protein	148
RNA polymerase sigma factor RpoD	147

Results of search of the TP splicing

- **$4{,}01 \times 10^6$ genes are collected in the Kegg-48 data bank**
- **311221 genes contain triplet periodicity splicing (5% false positives)**
- **Triplet periodicity change points can be the reflection of the splicing of parts of genes in the new gene**

Pair Change Points



Pair Change Points

$$W_1 = \sum_{1 \le i < k_1} \sum_{1 \le j < k_2} Sim_{ij}(1,1) + r \sum_{1 \le i < k_1} \sum_{k_1 \le j \le k_2} Dif_{ij} + \sum_{1 \le i < k_1} \sum_{k_2 < j \le K} Sim_{ij}(1,1)$$

$$+ \sum_{k_1 \le i \le k_2} \sum_{k_1 \le j \le k_2} Sim_{ij}(1,1) + r \sum_{k_1 \le i \le k_2} \sum_{k_2 < j \le K} Dif_{ij} + \sum_{k_2 < j \le K} \sum_{k_2 < j \le K} Sim_{ij}(1,1)$$

$$W_2 = \sum_{1 \le i \le K} \sum_{1 \le j \le K} Sim_{ij}(1,1)$$

$$W = W_1 - W_2$$

Measure of similarity of TP matrixes

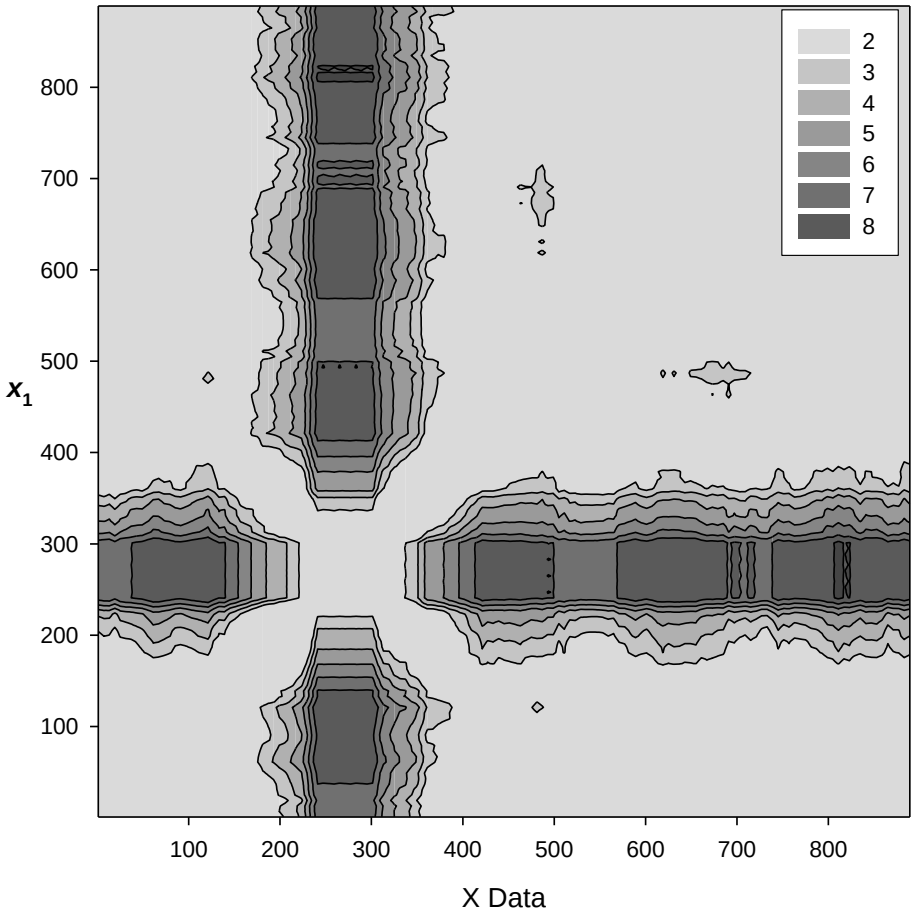
$$z_{1k}(i, j) = v_1(i, j) w_k(i, j)$$

$$f(z) = \pi^{-1} K_0(|z|)$$

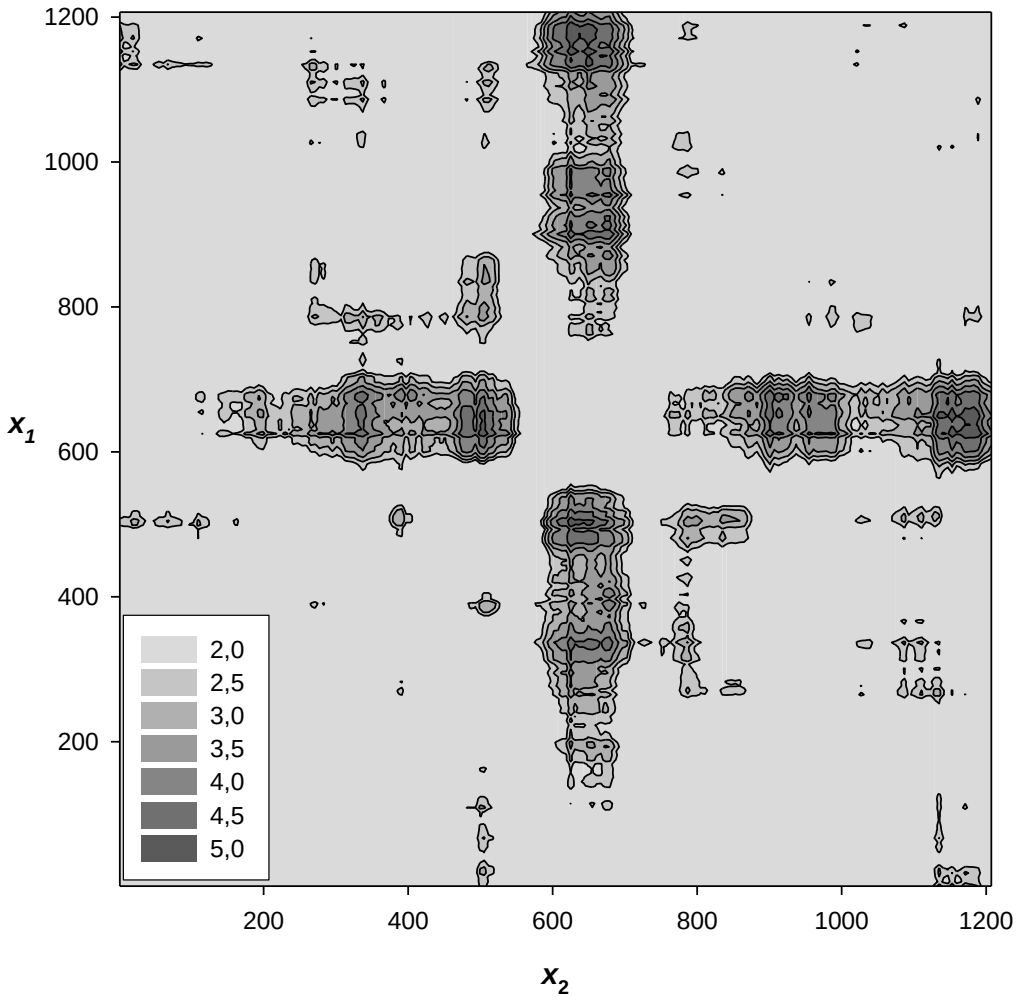
$$P(x > x_{1,k}(i, j)) = P(z > z_{1,k}(i, j))$$

$$D(1, k) = \sum_{i=1}^{4} \sum_{j=1}^{3} x_{1,k}(i, j)$$

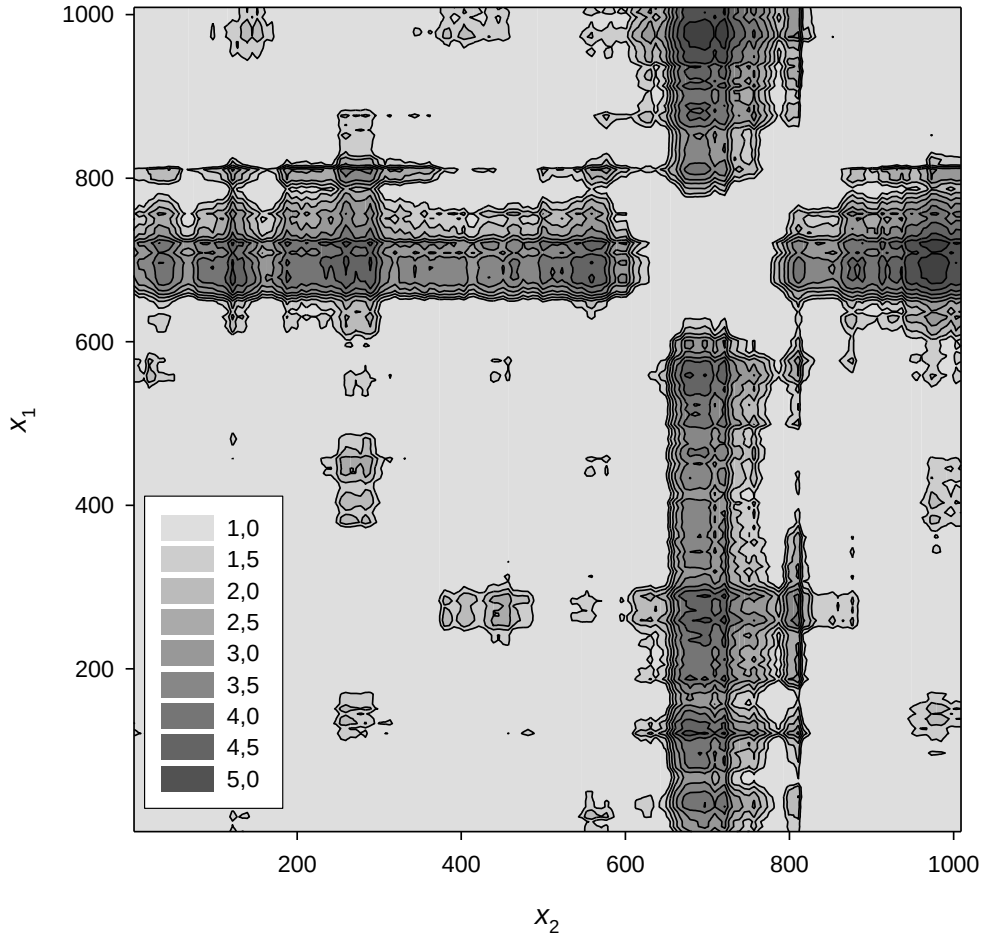
Gene coding of the chitosanase (KEGG ID is BSU26890) with artificial insertion of 180 bp length after 240th bp.



Gene coding the glycerol-3-phosphate permease from *B.subtilis* genome (BSU02140)



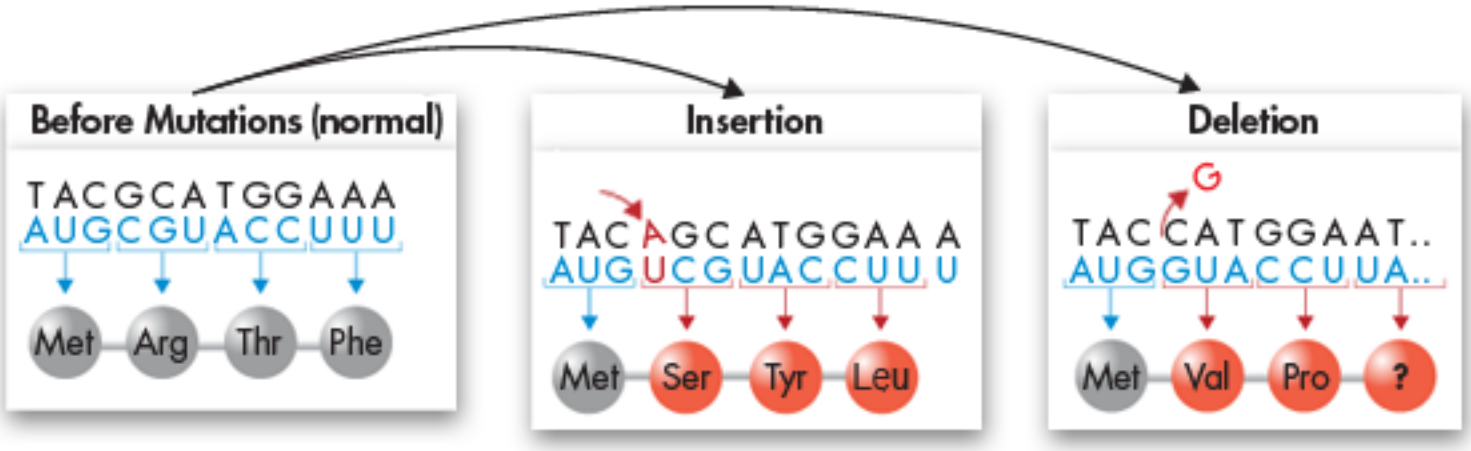
Gene coding the flagellar motor switch protein from *B.subtilis* genome (BSU16320).



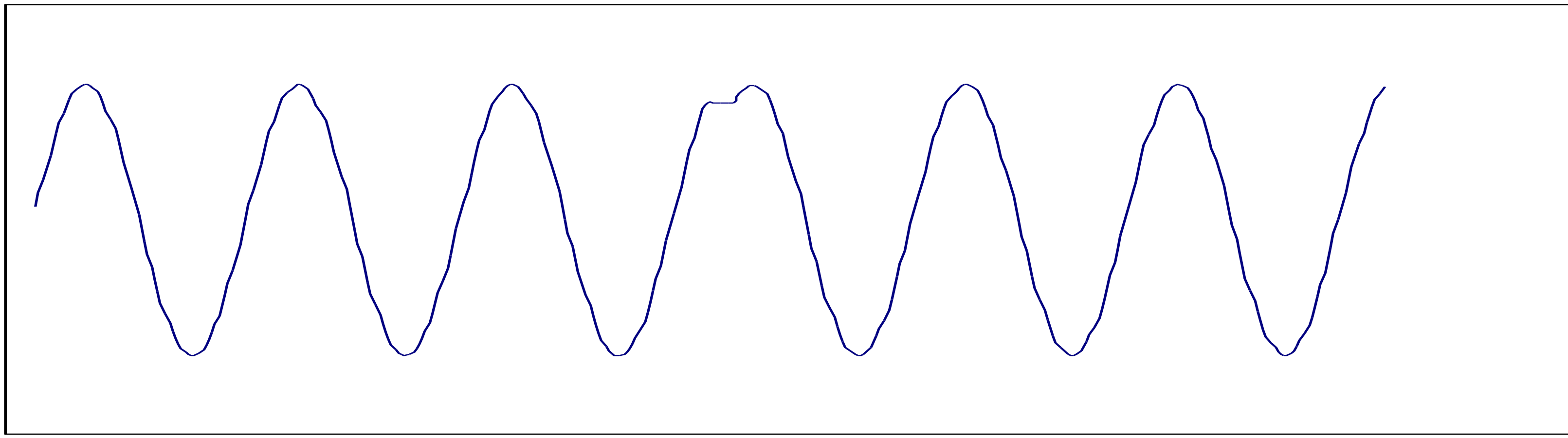
- **17 bacterial genomes were analysed**
- **4% genes contains the pair change points**

Frameshift mutation

Frameshift mutations can change every amino acid that follows the point of the mutation and can alter a protein so much that it is unable to perform its normal functions.



Shift of the triplet periodicity



+1

1231231231231231231231231231231231231...

atggcttcgatccattcgAgctagagacatcgaatca...

base insertion

Triplet periodicity matrix M(3,16)

123123123123123123123123123...

atgatgatgatgatgatgatgatgatg...

	N	**1**	**2**	**3**
aa	1	0	0	0
ta	2	0	0	0
ca	3	0	0	0
ga	4	50	0	0
at	5	0	50	0
tt	6	0	0	0
ct	7	0	0	0
gt	8	0	0	0
ac	9	0	0	0
tc	10	0	0	0
cc	11	0	0	0
gc	12	0	0	0
ag	13	0	0	0
tg	14	0	0	50
cg	15	0	0	0
gg	16	0	0	0

1=>*12*

2=>*23*

3=>*31*

Matrix change after reading frame shift

$S=\{atg\}_{50}$.

	N	1	2	3
aa	1	0	0	0
ta	2	0	0	0
ca	3	0	0	0
ga	4	50	0	0
at	5	0	50	0
tt	6	0	0	0
ct	7	0	0	0
gt	8	0	0	0
ac	9	0	0	0
tc	10	0	0	0
cc	11	0	0	0
gc	12	0	0	0
ag	13	0	0	0
tg	14	0	0	50
cg	15	0	0	0
gg	16	0	0	0

$S=\{atg\}_{25}\{tga\}_{25}$

	N	1	2	3
aa	1	0	0	0
ta	2	0	0	0
ca	3	0	0	0
ga	4	25	0	25
at	5	25	25	0
tt	6	0	0	0
ct	7	0	0	0
gt	8	0	0	0
ac	9	0	0	0
tc	10	0	0	0
cc	11	0	0	0
gc	12	0	0	0
ag	13	0	0	0
tg	14	0	25	25
cg	15	0	0	0
gg	16	0	0	0

Optimization of matrix $W(3,16)$

It is unknown

ancestor

atg gat cga tcg att tcg cgc tac ttc

sequence S_0

$W_0(3,16)$

It is known

descendant

atg gat cga tcg Gat ttc gcg cta ctt

sequence S

$W(3,16)$

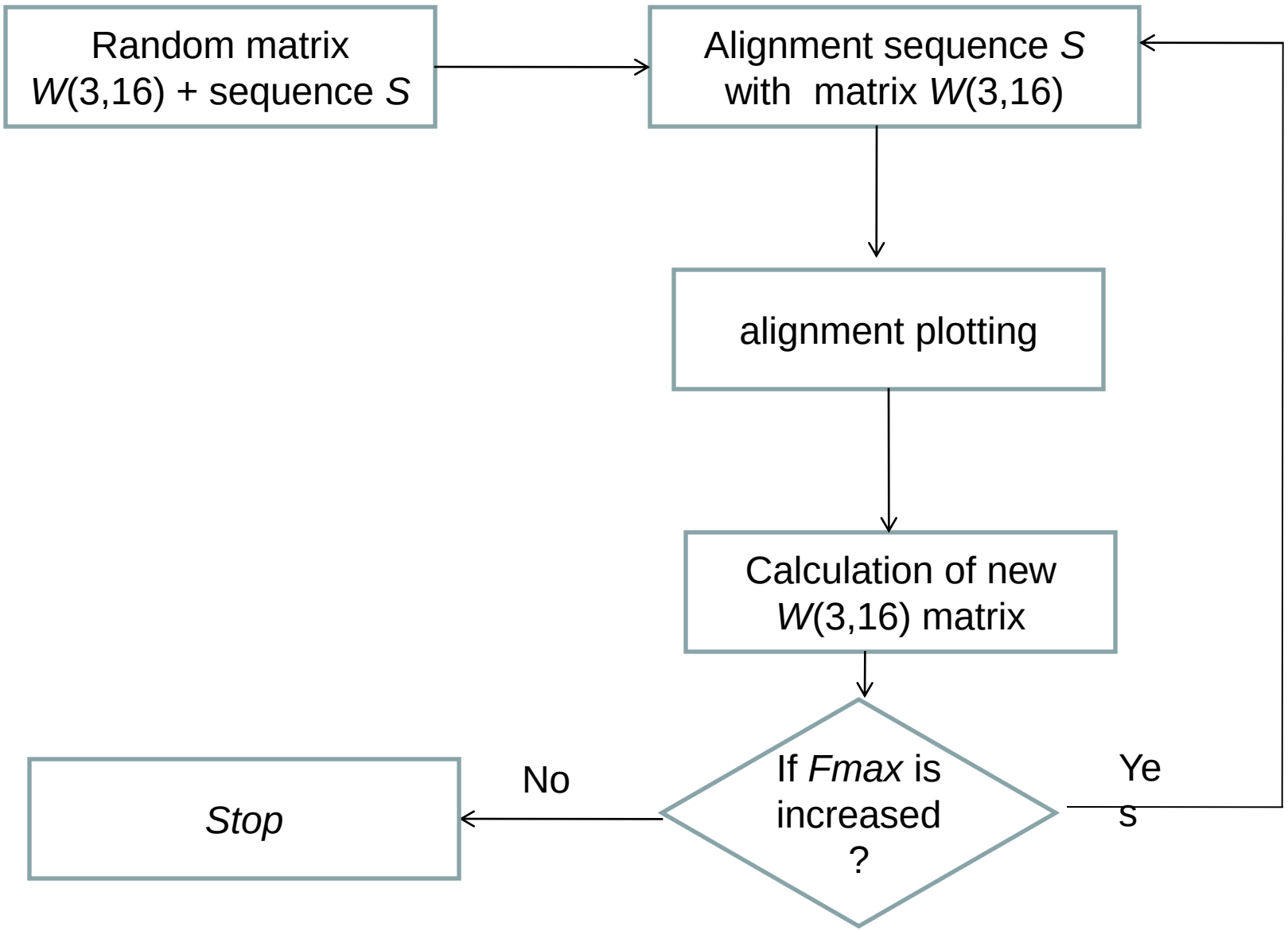
Search for the best triple matrix

1.To find the reading frame shifts in the *S* sequence, you need to know the matrix $W_0(3,16)$. But the ancestral sequence S_0 in most cases is not available and the matrix $W_0(3,16)$ is not known.

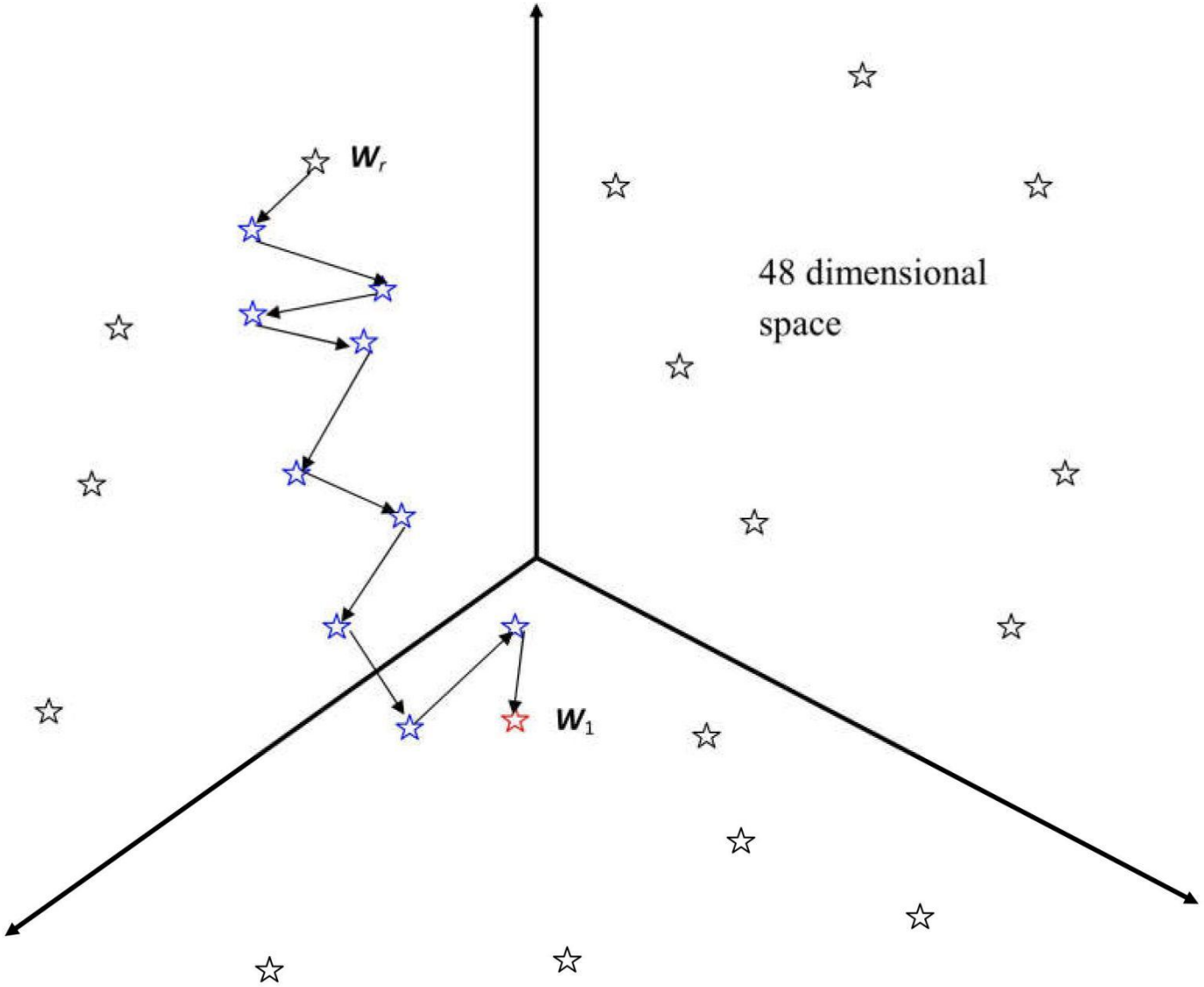
2. The task is to apply the optimization procedure and find the best approximation to $W_0(3,16)$.

3. Optimization procedure - a genetic algorithm and dynamic programming.

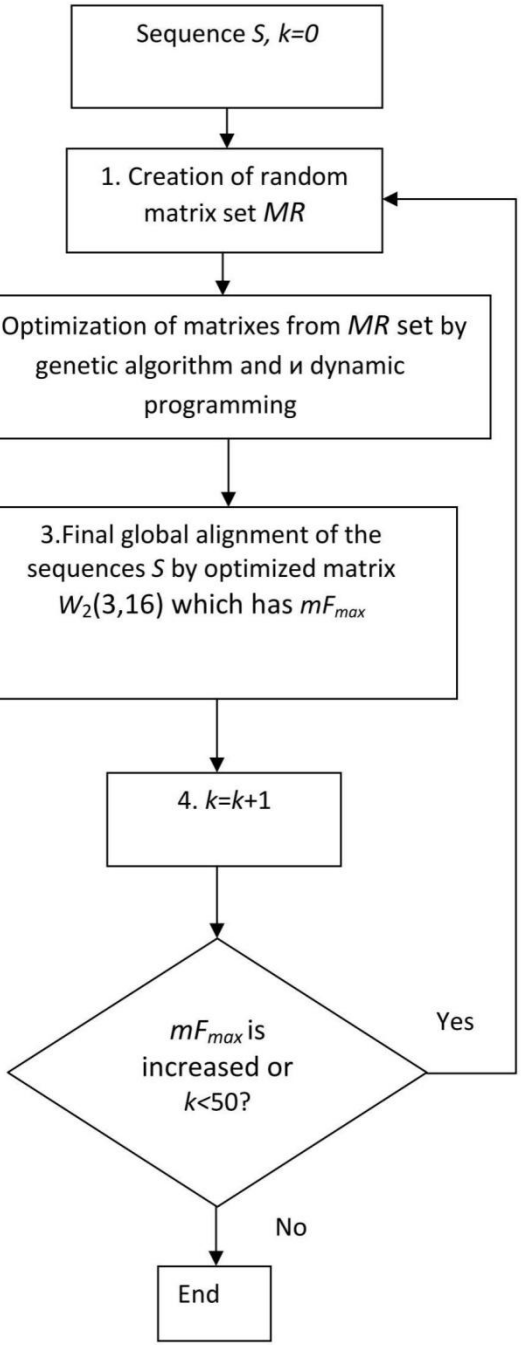
Optimization of the matrix *M*(3,16)



Optimization of the matrix *M*(3,16)



Optimization of matrix *W*(3,16)



For details see:
Korotkov EV et.al. DNA Research, 2019, https://doi.org/10.1093/dnares/dsy046

http://victoria.biengi.ac.ru/cgi-bin/frameshift

Database of potential frameshifts ?

Query parameters

Organism Homo sapiens

Gene symbol or Transcript ID

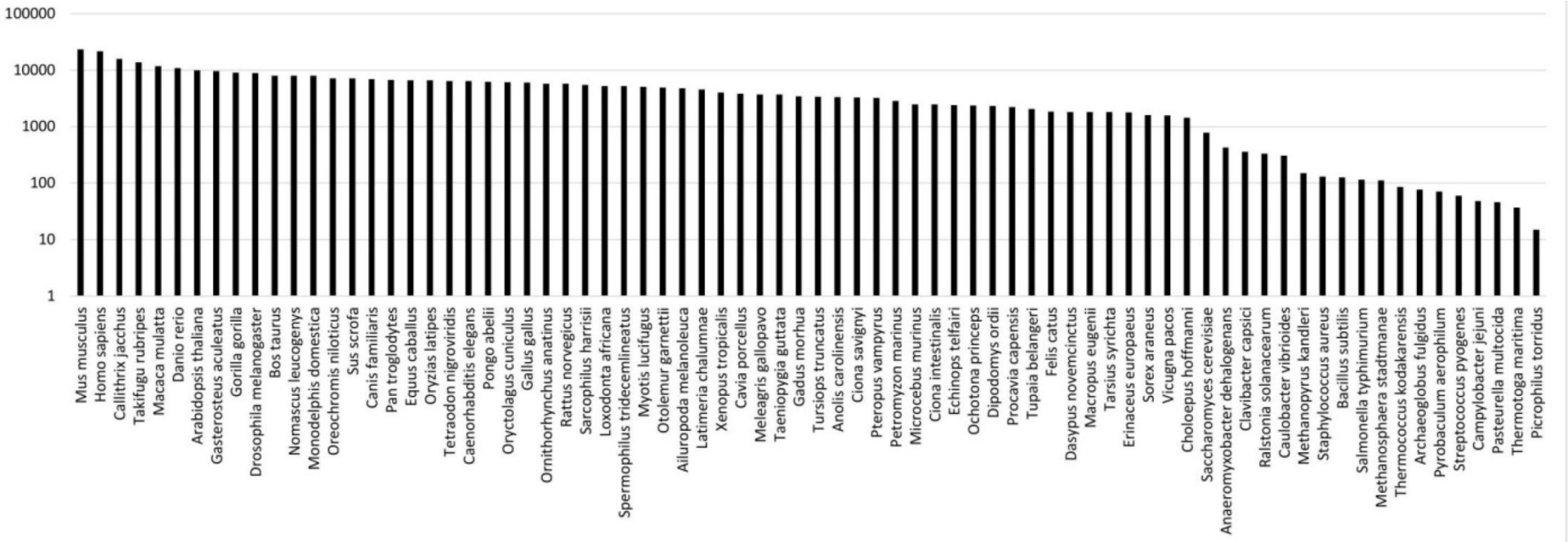
Get results

Results

Organism	Gene	Transcript
Homo sapiens	CASP12	ENST00000447913 »
Homo sapiens	CASP12	ENST00000447913 »
Homo sapiens	CASP12	ENST00000447913 »
Homo sapiens	CYP2D7	ENST00000574062 »
Homo sapiens	IGHV4-59	ENST00000390629 »
Homo sapiens	IGHV4-59	ENST00000390629 »

The number of cds with potential shifts of the reading frame from different genomes

http://victoria.biengi.ac.ru/cgi-bin/frameshift/



http://victoria.biengi.ac.ru/cgi-bin/frameshift

Get results

Results

Organism	Gene	Transcript
Ailuropoda melanoleuca	TNIP3	ENSAMET00000003373 »

Transcript

ENSAMET00000003373

Gene

TNIP3

Correlation matrix

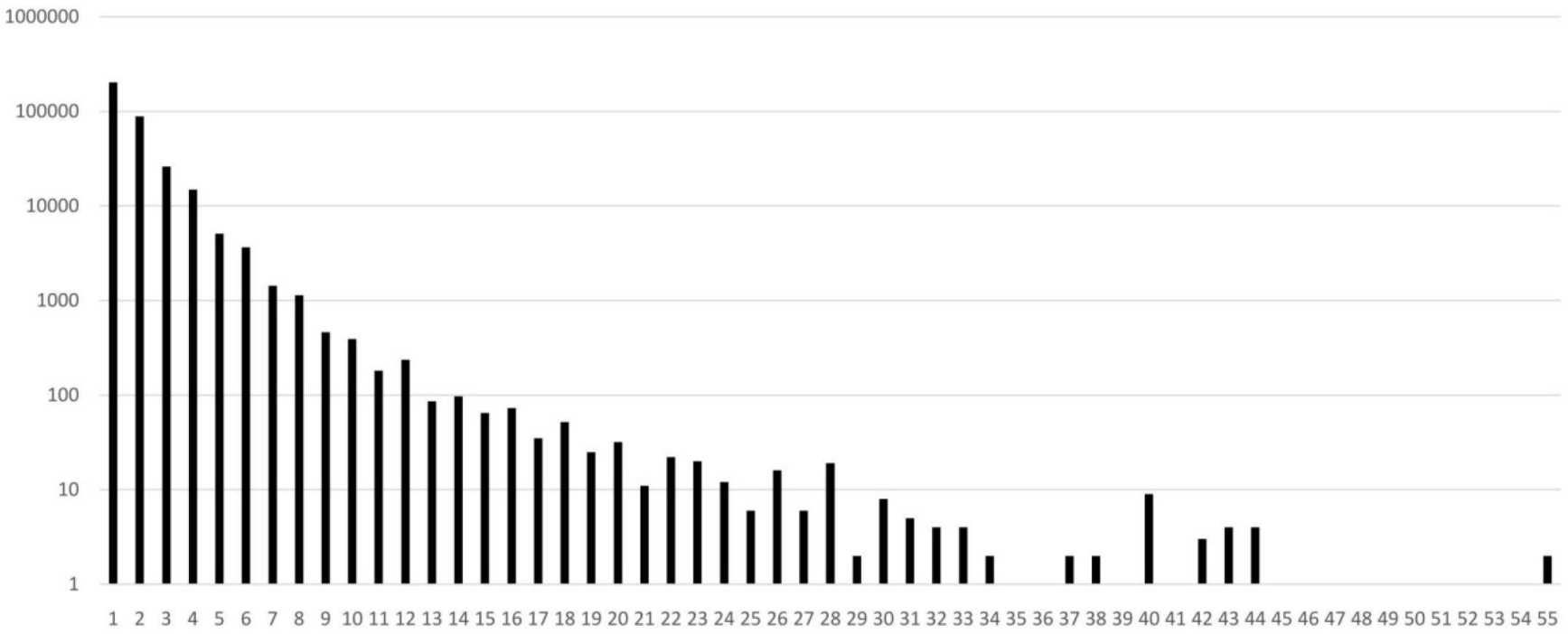
	Position 1			
	A	T	C	G
A	0.8	-0.5	-2.3	8.2
T	-4.4	1.7	3.4	6.8
C	-3.7	-0.9	-1.7	-4.5
G	-6.4	-1.9	-1.2	-1.9

	Position 2			
	A	T	C	G
A	-4.8	-2.4	0.5	-1.1
T	-5.3	-1.0	4.3	0.0
C	-3.7	-0.0	2.5	-2.3
G	0.8	-2.3	5.6	3.5

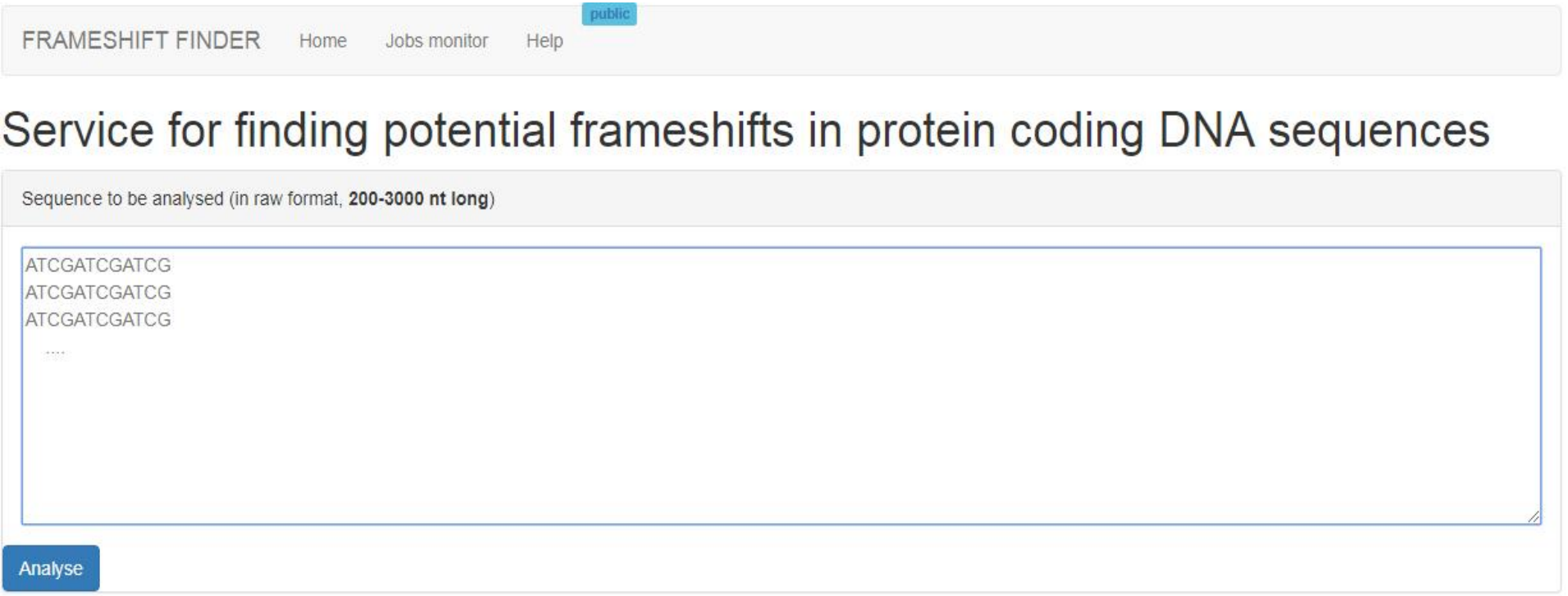
	Position 3			
	A	T	C	G
A	-0.8	-0.5	-4.8	-2.5
T	-6.6	4.0	-0.9	-2.3
C	3.1	7.3	-0.5	1.0
G	6.2	-1.5	-2.7	-5.7

```
1141 AAGCGAAGGAGTTGATGTCCCTGACTGCAGAAGAACTGTATCAGCTCCAGTTGTTAGACC
1141 2312312312312312312312312312312312312312312312312312312312

1201 TAAAGATACATACGTAAAGGAAGAAAGCTCGAAACATTTTGATCCTTGCAACTCGGTGGA
1201 2312312312312312312312312312312312312312312312312312312312

1261 ATTCTTGGATTTGGCTCATAGTTCCGAAAGCCAGGAGACCATATCAAGCATGGGAGAACA
1261 2312312312312312312312312312312312312312312312312312312312

1321 ATCAGATAACCTTTTTGAACAGAGAAAAGATACAGAAAACATGGAGGATTGC*CAGAATC
1321 2312312312312312312312312312312312312312312312312312312312

1380 TTTTGAAGCCATGTAGTCTGTGTGAGAAAAGACCACGAAACGGGAACATTATTCACGGGA
1381 2312312312312312312312312312312312312312312312312312312312

1440 GGACAGGCCATCTTGTCACTTGTTTTCATTGTGCCAGAAGACTAAAGAAGGCTGGGGCTT
1441 2312312312312312312312312312312312312312312312312312312312

1500 CTTGTCCTATTTGCAAGAAAGAGATTCAGTTGGTTATTAAGGTTTTTGTAGCATAGCTGA
1501 2312312312312312312312312312312312312312312312312312312312

1560 ATCAGTGAATTACAGACAGATAGTAACAGAGCTGGTCATGTGTCCAAATATCAGTATTGA
1561 2312312312312312312312312312312312312312312312312312312312

1620 GCATCTCTACGCAGGGGTGACCCATTTCATACTTTGATTTATTCATGTGAACTTTTATGT
1621 2312312312312312312312312312312312312312312312312312312312
```

Distribution of *cds* according to the number of potential frameshifts



http://victoria.biengi.ac.ru/fsfinder/

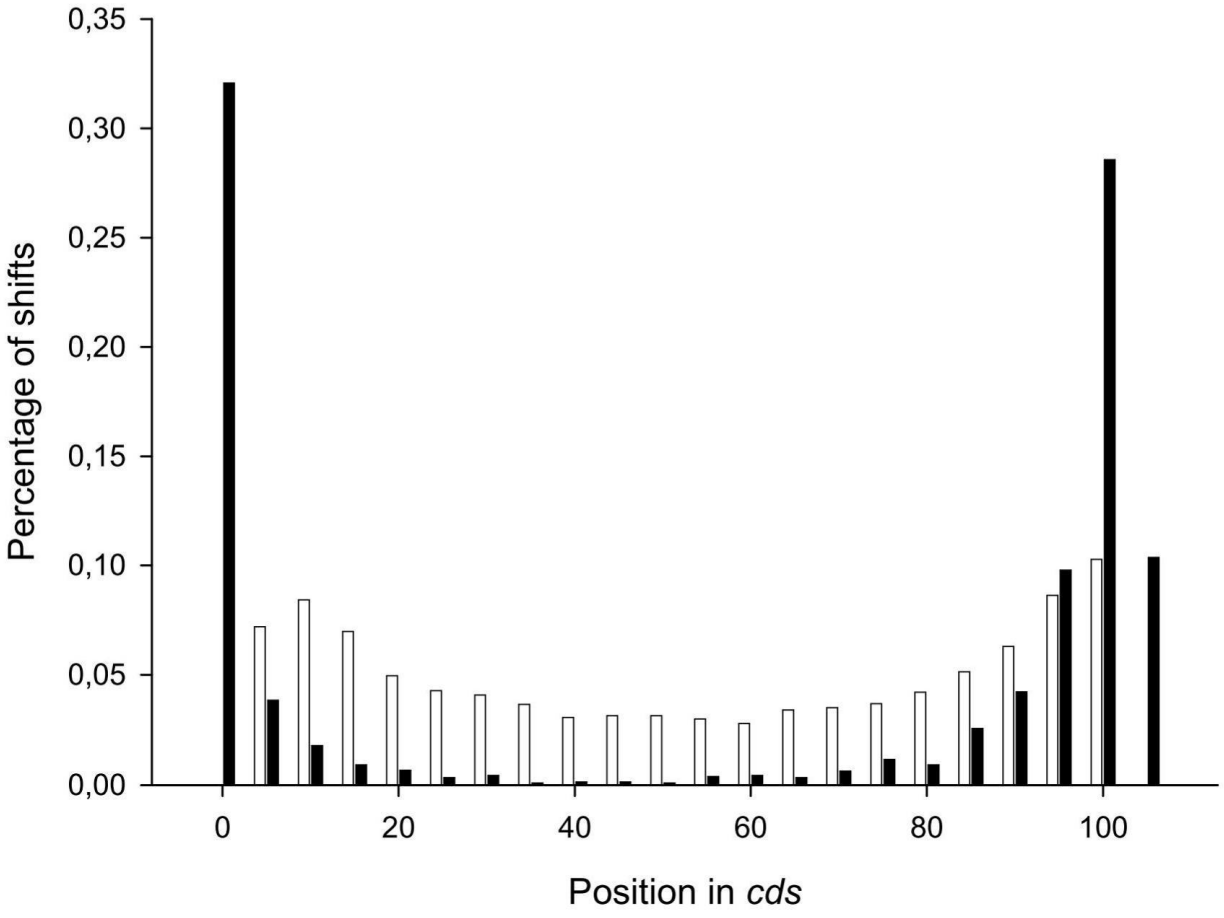


Comparison of results for 6 genomes with work Antonov et.al. [1].

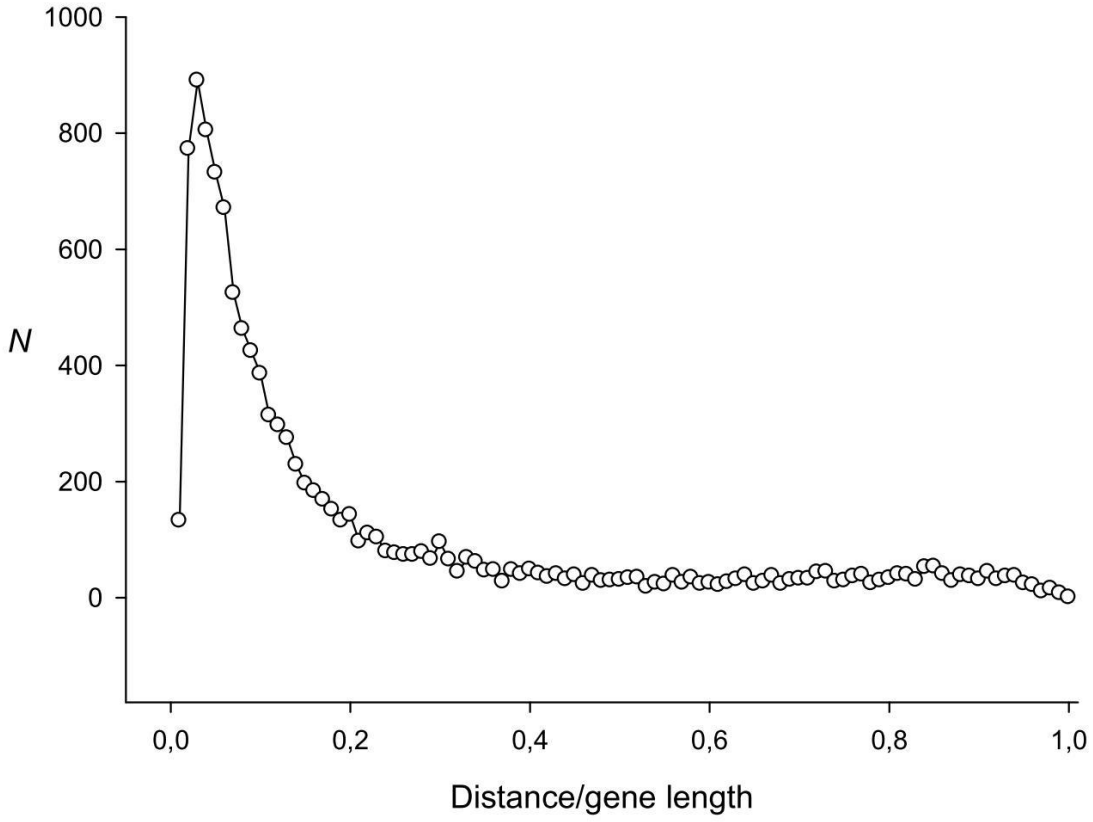
Species name	Number of potential framesfift mutations	Number of cds with potential framesfift mutations	Number of potential framesfift mutations from [1]
A.thaliana	14954	9930	2067
C.elegans	10411	5941	611
D.melanogaster	31873	8833	2616
H.sapiens	20795	13285	7395
R.Norvegicus	9811	5768	703
X.tropicalis	6518	4228	529

[1] Antonov, I., Baranov, P., and Borodovsky, M. 2013, GeneTack database: genes with frameshifts in prokaryotic genomes and eukaryotic mRNA sequences. *Nucleic Acids Res.*, **41**, D152-6.

Distribution of shifts position in the sequence of a cds



Distribution of the distance between paired compensating shifts of the triplet periodicity phase in the *A.thaliana* genome.



Results

1. Number potential frameshifts is approximately 21% of all analyzed *cds* of the genomes.

2. The type I and type II error rates were estimated as 11 and 30%, respectively

Publication:
Korotkov EV et.al. *Search for potential reading frameshifts in cds from Arabidopsis thaliana and other genomes*. DNA Research, 2019, https://doi.org/10.1093/dnares/dsy046

1. Frenkel FE, Korotkov EV. Using triplet periodicity of nucleotide sequences for finding potential reading frame shifts in genes. DNA Res. 2009;16:105-114.
2. EV Korotkov, MA Korotkova "Bioinformatics and search of shifts of reading frame in genes" Information technologies and computation systems (Russian), №1, pp.1-23, 2010.
3. EV Korotkov, MA Korotkova «Study of the triplet periodicity phase shifts in genes, Journal of Integrative Bioinformatics, v.7,131-141, 2010
4. Rudenko VM and Korotkov EV. Monte-Carlo applications fro search of potential shifts of reading frame in genes. Mathematical Biology and bioinformatics, v. 6, pp. 79-91, 2011
5. M.A.Korotkova, N.A. Kudryashov, E.V.Korotkov. An approach for searching insertions in bacterial genes leading to the phase shift of triplet periodicity. *Genomics, Proteomics &Bioinformatics*, v.9, pp.158-170, 2011.
6. Yu.M.Suvorova V.M. Rudenko, E.V.Korotkov. Detection change points of triplet periodicity of gene, *Gene.* v.491, pp.58-64, 2012.
7. Pugatcheva VM Korotkov AE, Korotkov EV E.B. Search of pair points of shifts of triplet periodicity in genes from 17 bacterial genomes. Mathematical Biology and bioinformatics V. 7, №2, 2012
8. Suvorova YM, Korotkova MA, Korotkov EV. Study of the Paired Change Points in Bacterial Genes IEEE/ACM Transactions on Computational Biology and Bioinformatics; v.11(5), pp.955-964. DOI:10.1109/TCBB.2014.2321154
9. Pugacheva V, Frenkel F, Korotkov E. Investigation of phase shifts for different period lengths in the genomes of C. elegans, D. melanogaster and S. cerevisiae. Comput Biol Chem. v.51, p.12-21. 2014. doi: 10.1016/j.compbiolchem.2014.03.004.
10. Golishev MA, Korotkov EV Developing of the Computer Method for Annotation of Bacterial Genes. Advances in Bioinformatics, 2015.