Биоинформатика

Коротков Евгений Вадимович
Институт Биоинженерии, ФИЦ Биотехнологии РАН

bioinf@yandex.ru

Множественные выравнивания

Профили

There are essentially four major approaches to multiple sequence alignment:

1. Optimal global sequence alignment . It is NP-complete problem.

2. Progressive global alignment

3. Block-based global alignment

4. Motif-based local alignment

5. Markov models

Methods 2-5 don’t work for sequences which have more 1.5 mutation per amino acid.

Обобщение парного выравнивания

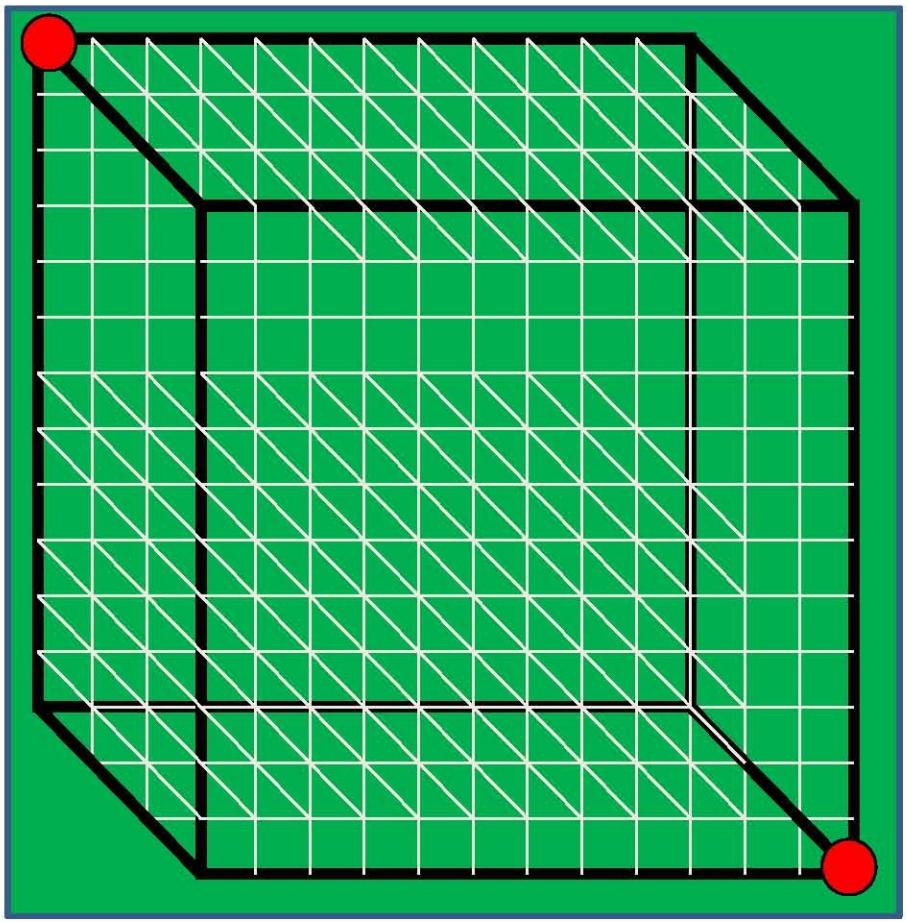
- Выравнивание 2-х последовательностей – двумерная матрица
- 3-х последовательностей – 3-х мерная.

A T _ G C G _
A _ C G T _ A
A T C A C _ A

- Задача: больше консервативных столбцов, лучше выравнивание

Глобальное выравнивание 3-х последовательностей

начало



конец

Алгоритм

- $s_{i,j,k} = \max \begin{cases} s_{i-1,j-1,k-1} + \delta(v_i, w_j, u_k) & \text{Нет гэпов} \\ s_{i-1,j-1,k} + \delta(v_i, w_j, \_) & \\ s_{i-1,j,k-1} + \delta(v_i, \_, u_k) & \text{Один гэп} \\ s_{i,j-1,k-1} + \delta(\_, w_j, u_k) & \\ s_{i-1,j,k} + \delta(v_i, \_, \_) & \\ s_{i,j-1,k} + \delta(\_, w_j, \_) & \text{Два гэпа} \\ s_{i,j,k-1} + \delta(\_, \_, u_k) & \end{cases}$

- $\delta(x, y, z)$ – запись в трехмерной матрице весов

Время работы алгоритма

- Для 3-х последовательностей длины n, время работы – $7n^3$; $O(n^3)$
- Для k последовательностей - $(2k-1)(n^k)$; $O(2kn^k)$

Множественное выравнивание порождает парные выравнивания

```
x: AC-GCGG-C
y: AC-GC-GAG
z: GCCGC-GAG
```

Порождает:

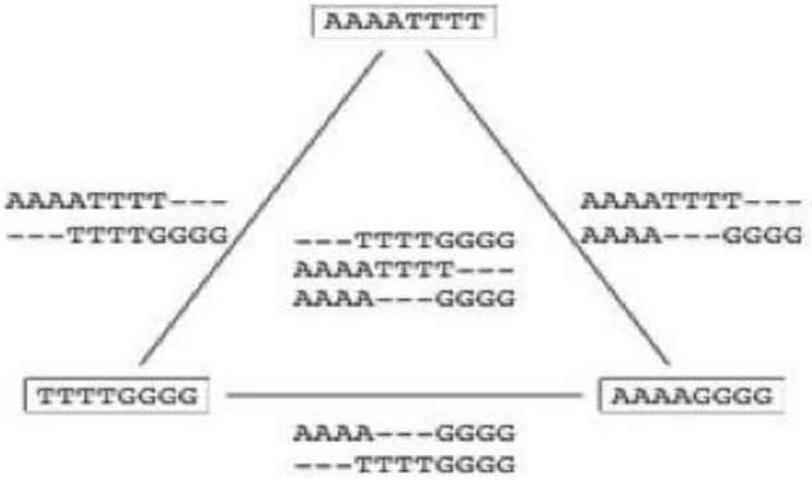
```
x: ACGCGG-C;   x: AC-GCGG-C;   y: AC-GCGAG
y: ACGC-GAC;   z: GCCGC-GAG;   z: GCCGCGAG
```

Обратная проблема

Имея 3 субъективных парных варнивания:

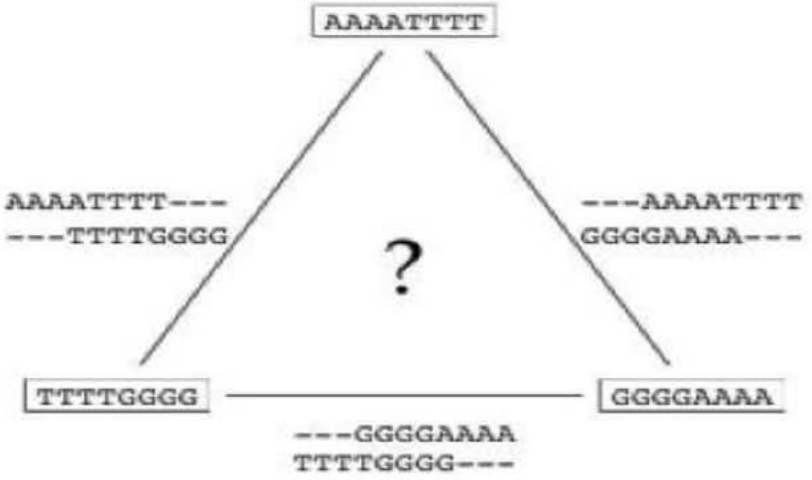
```
x: ACGCGG-C;   x: AC-GCGG-C;   y: AC-GCGAG
y: ACGC-GAC;   z: GCCGC-GAG;   z: GCCGCGAG
```

Хороший вариант



(a) Compatible pairwise alignments

Плохой вариант



(b) Incompatible pairwise alignments

Выравнивание выравниваний

```
x GGGCACTGCAT
y GGTTACGTC--     Alignment 1
z GGGAACTGCAG

w GGACGTACC--     Alignment 2
v GGACCT-----
```

Описание выравнивания

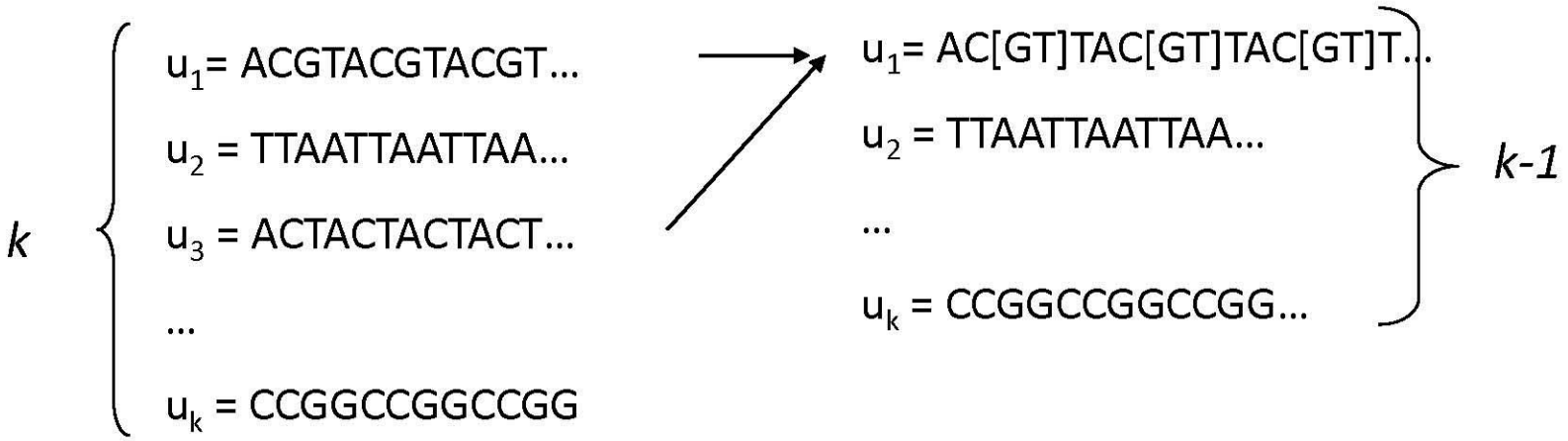
GTCTGA
GTCAGC
} GTC[TA]G[AC] – профиль

[X][5X]

```
x GGGCACTGCAT
y GGTTACGTC--
z GGGAACTGCAG
w GGACGTACC--
v GGACCT-----
  GGACACAGCAT – консенсус
```

Матрица частот – используется редко

Множественное выравнивание – жадный алгоритм



Время работы алгоритма на k последовательностях длины n – $O(n^2k^2)$

Прогрессивное выравнивание ClustalW

- Прогрессивное выравнивание – жадный алгоритм с более «умным» способом выбора пар.
- Три шага
  1.) Построить парные выравнивания
  2.) Построить дерево-подсказку
  3.) Прогрессивное выравнивание по дереву-подсказке

Шаг 1: Парные Выравнивания

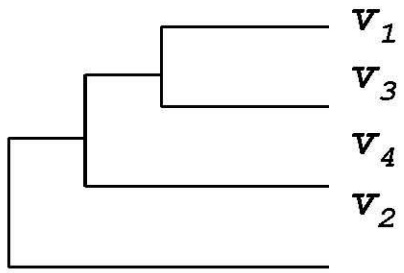
- Выравнивания пар порождают матрицу identity

	v_1	v_2	v_3	v_4
v_1				
v_2	.17	-		
v_3	.87	.28	-	
v_4	.59	.33	.62	-

(.17 значит идентичны на 17 %)

Шаг 2: Дерево-подсказка

	v_1	v_2	v_3	v_4
v_1				
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–



Далее вычислить:

$v_{1,3}$ = выравнивание (v_1, v_3)

$v_{1,3,4}$ = выравнивание $((v_{1,3}), v_4)$

$v_{1,2,3,4}$ = выравнивание $((v_{1,3,4}), v_2)$

Шаг 3: Прогрессивное выравнивание

- Выравниванием 2 наиболее близких последовательности.
- Следуя дереву - подсказке, довыравниваем следующую последовательность к имеющемуся выравниванию

```
FOS_RAT          PEEMSVTS-LDLTGGLPEATTPESEEAFTLPLLNDPEPK-PSLEPVKNISNMELKAEPFD
FOS_MOUSE        PEEMSVAS-LDLTGGLPEASTPESEEAFTLPLLNDPEPK-PSLEPVKSISNVELKAEPFD
FOS_CHICK        SEELAAATALDLG----APSPAAAEEAFALPLMTEAPPAVPPKEPSG--SGLELKAEPFD
FOSB_MOUSE       PGPGPLAEVRDLPG-----STSAKEDGFGWLLPPPPPPP-----------------LPFQ
FOSB_HUMAN       PGPGPLAEVRDLPG-----SAPAKEDGFSWLLPPPPPPP-----------------LPFQ
                 .   . :   ** .     :..  *:.*   *   . *                   **:
```

Точки и звезды отображают насколько консервативны столбцы.

Множественные Выравнивания: Взвешивание

- Количество полных совпадений
- Сумма по парам (SP-Score)
- Энтропия

Количество полных совпадений

AAA
AAA
AAT
ATC

- Хорошо только для очень близких последовательностей

Сумма по парам (SP-Score)

- Построим парное выравнивание по множественному
- Посчитаем веса всех этих парных выравниваний – $s(a_i, a_j)$
- Просуммируем:

$$s(a_1, \ldots, a_k) = \Sigma_{i,j}\, s\,(a_i, a_j)$$

Энтропия

- Определим вероятности букв в столбцах
  - $p_A = 1$, $p_T=p_G=p_C=0$ (1-ый столбец)
  - $p_A = 0.75$, $p_T = 0.25$, $p_G=p_C=0$ (2-ый столбец)
  - $p_A = 0.50$, $p_T = 0.25$, $p_C=0.25$ $p_G=0$ (3-ий столбец)
- Энтропия столбца будет равна

$$-\sum_{X=A,T,G,C} p_X \log p_X$$

AAA
AAA
AAT
ATC

Энтропия: Пример

Лучший вариант $entropy\begin{pmatrix} A \\ A \\ A \\ A \end{pmatrix} = 0$

Худший вариант $entropy\begin{pmatrix} A \\ T \\ G \\ C \end{pmatrix} = -\sum \frac{1}{4}\log\frac{1}{4} = -4(\frac{1}{4} * -2) = 2$

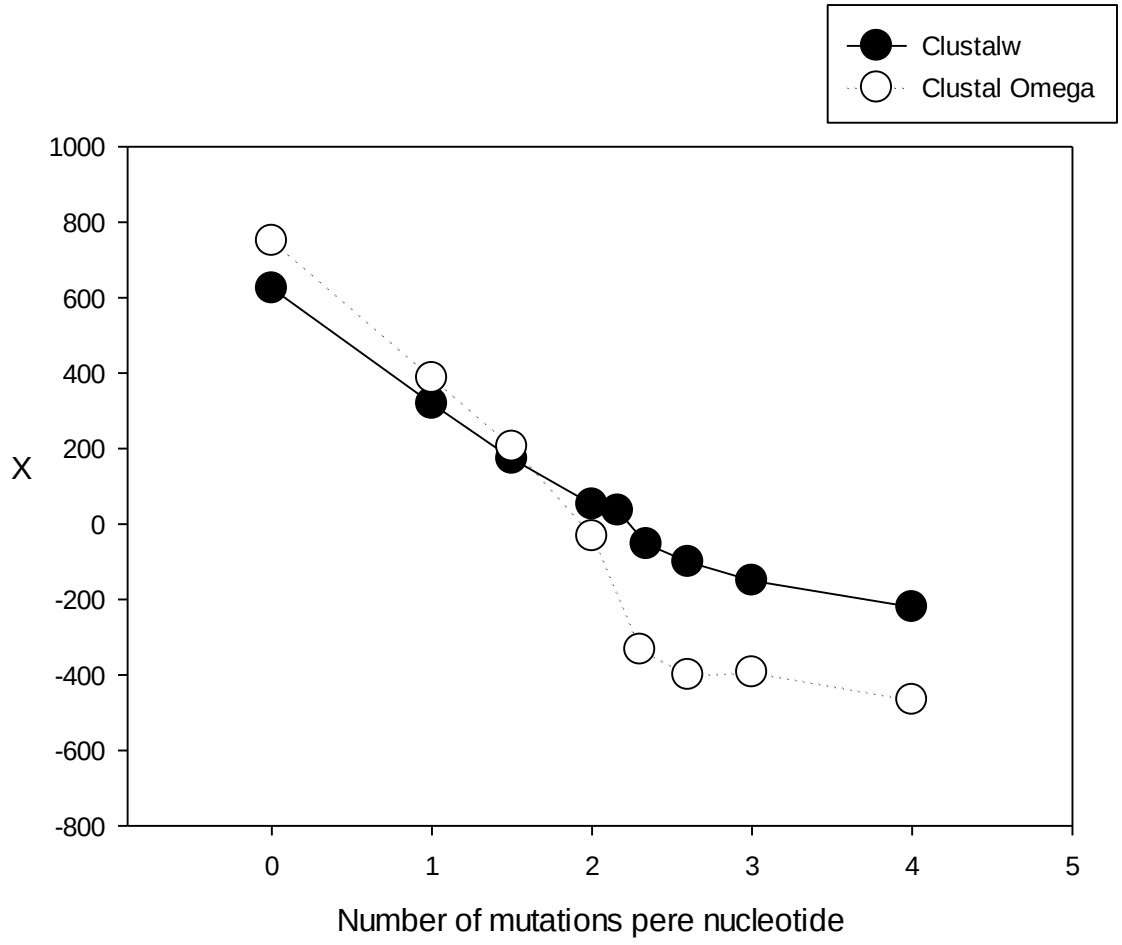
Энтропия: Пример

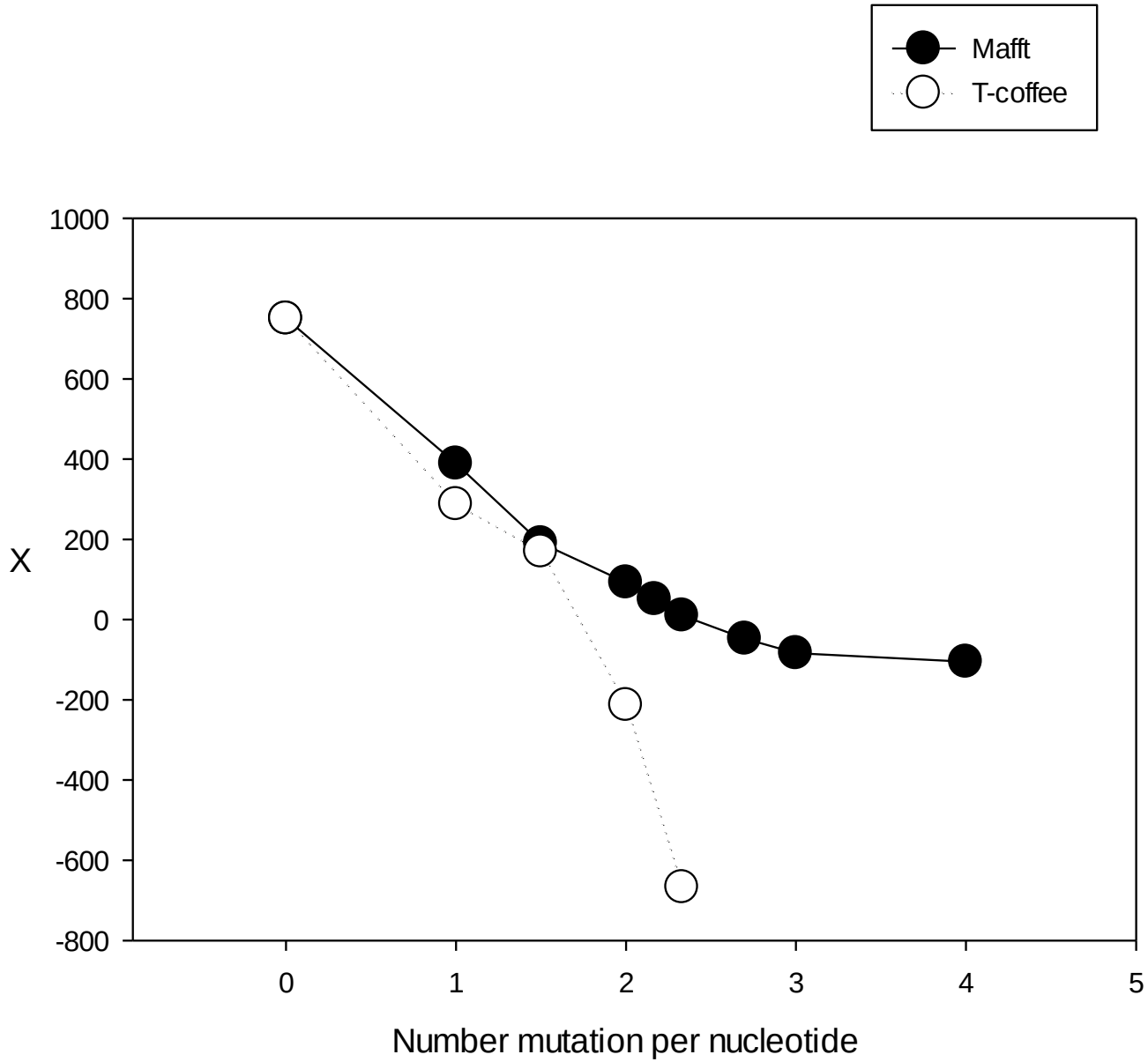
Энтропия столбца:

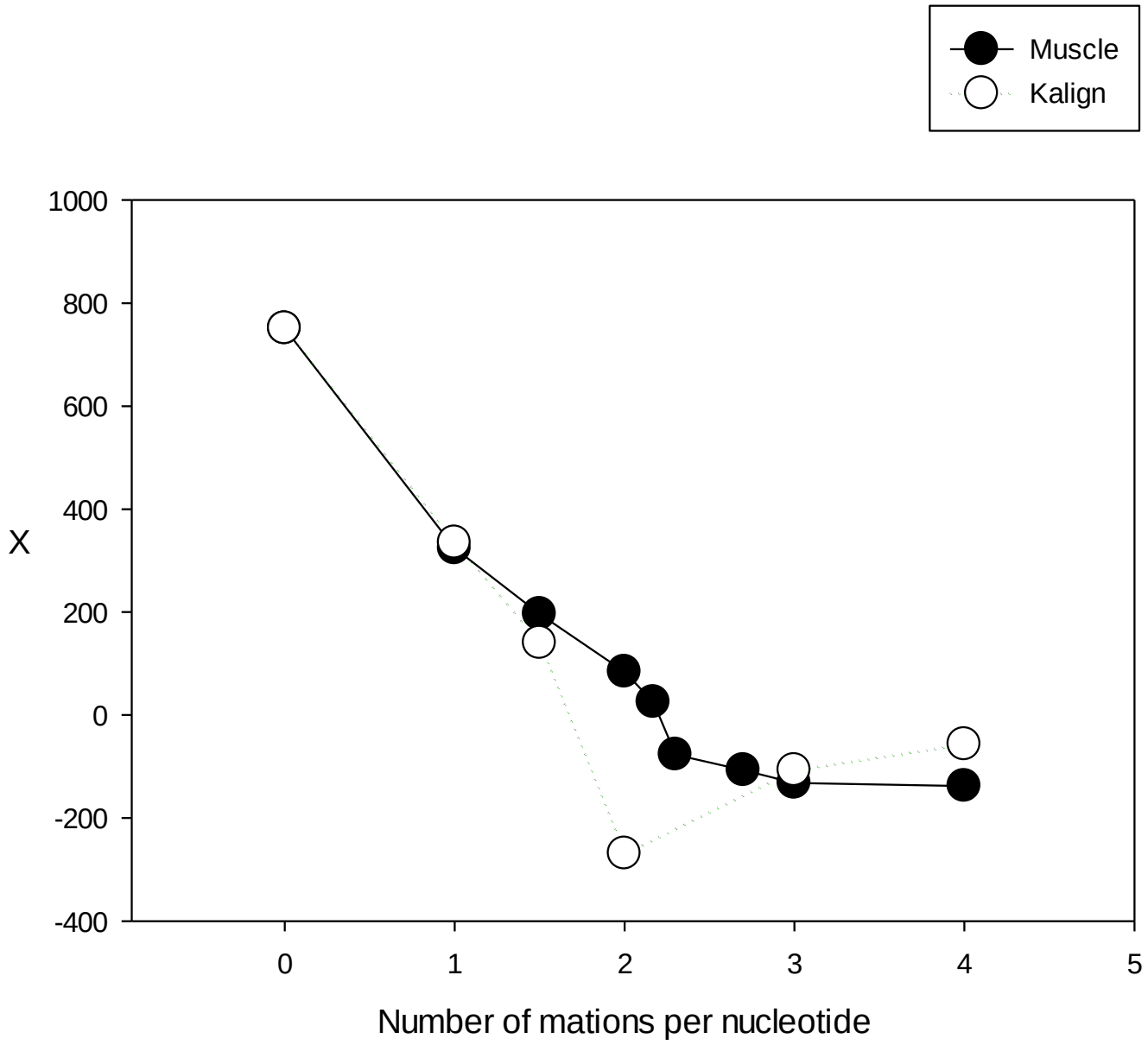
$$-(p_A \log p_A + p_C \log p_C + p_G \log p_G + p_T \log p_T)$$

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

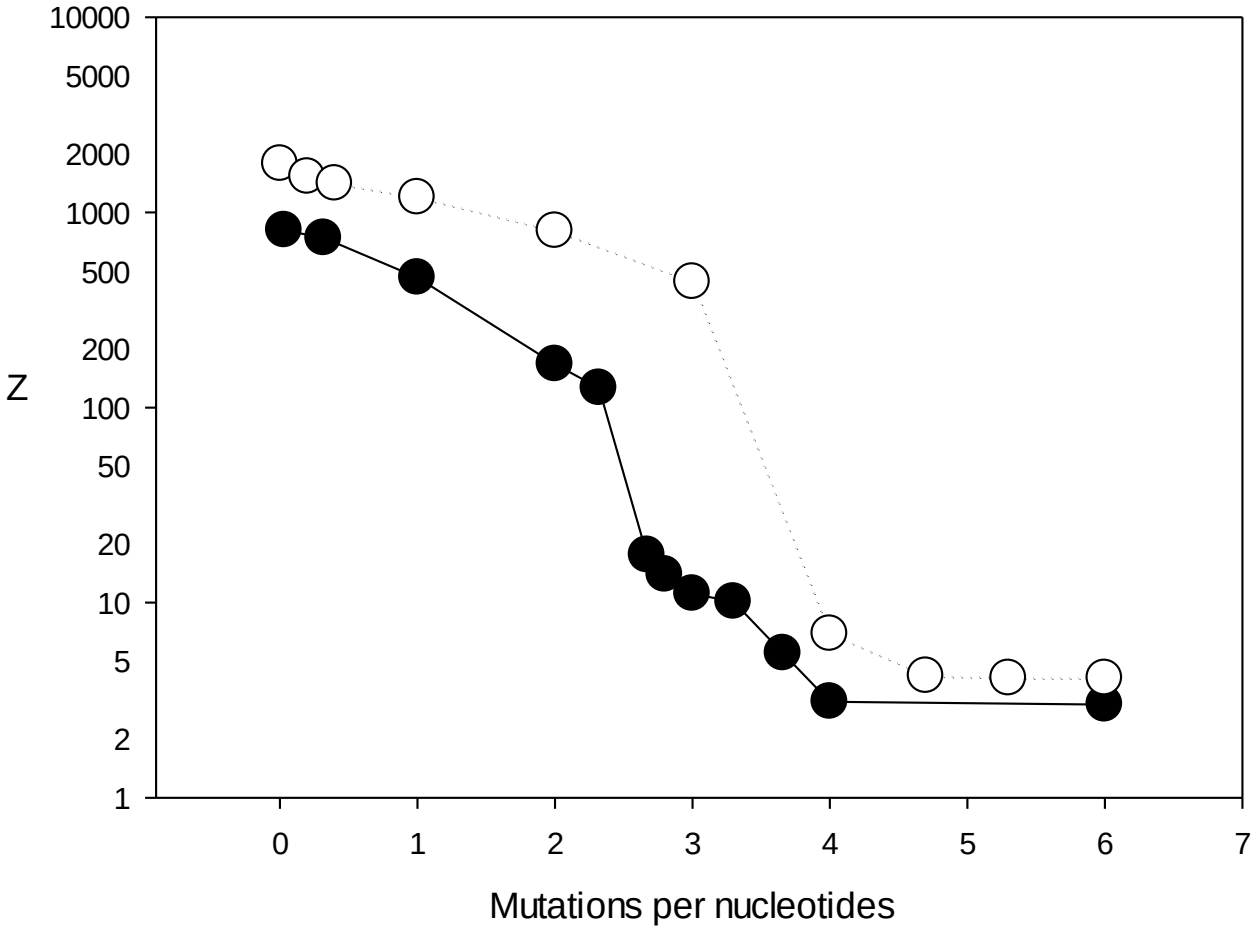
- Столбец 1 = -[1*log(1) + 0*log0 + 0*log0 +0*log0]
  = 0
- Столбец 2 = -[($^1/_4$)*log($^1/_4$) + ($^3/_4$)*log($^3/_4$) + 0*log0 + 0*log0]
  = -[($^1/_4$)*(-2) + ($^3/_4$)*(-.415)] = +0.811
- Столбец 3 = -[($^1/_4$)*log($^1/_4$)+($^1/_4$)*log($^1/_4$)+($^1/_4$)*log($^1/_4$)
+($^1/_4$)*log($^1/_4$)] = 4* -[($^1/_4$)*(-2)] = +2.0
- Энтропия выравнивания = 0 + 0.811 + 2.0 = +2.811


Clustalw
Clustal Omega
1000
800
600
400
200
0
-200
-400
-600
-800
X
0
1
2
3
4
5
Number of mutations pere nucleotide


Mafft
T-coffee
1000
800
600
400
200
0
-200
-400
-600
-800
X
0
1
2
3
4
5
Number mutation per nucleotide


Muscle
Kalign
1000
800
600
400
200
0
-200
-400
X
0
1
2
3
4
5
Number of mations per nucleotide

100_50, 500_250

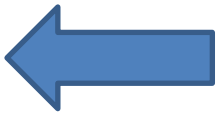


Algorithm

- How we can find the multiple alignment if the sequences have big number of mutations (more than 1.5 per amino acid or nucleotide)?
- Statistical important pairwise alignments, common “seeds” and “words” are absent.

Position Weight Matrix (PWM)

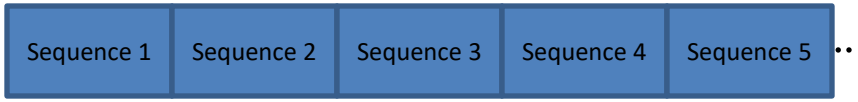
1	2	3	4	5
A	1.2	-2.3	3.1	-1.0
F	-1.2	2.2	2.5	-0.8
K	1.8	-0.9	0.1	3.3
M	2.3	3.9	-0.8	1.8
…	…	…	…	…



Multi alignment

```
MAHMV
RASSG
LSYPE
RFYAA
ASYVL
```

```
Sequence A   12345123451234512345123 45…
Sequence B
```



Task: to find PWM which describes multi alignment of analyzed sequences by west way.

Algorithm for search latent periodicity with indels

- **Main idea**:
- Generate the set of the random position weights matrixes (PWM)
- Calculate the alignment of the sequence with PWM's
- Improve the matrixes by genetic algorithm

Algorithm

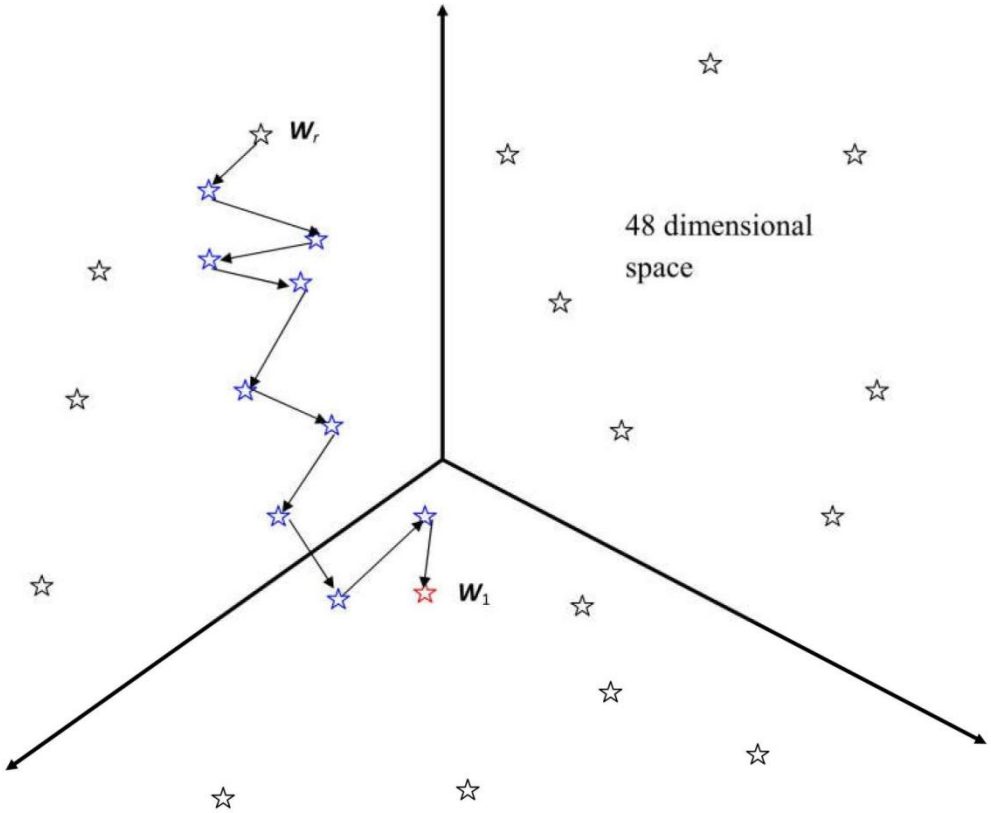
1 Creation of the set of random PWM's. Each PWM corresponds to some set of multi alignments.

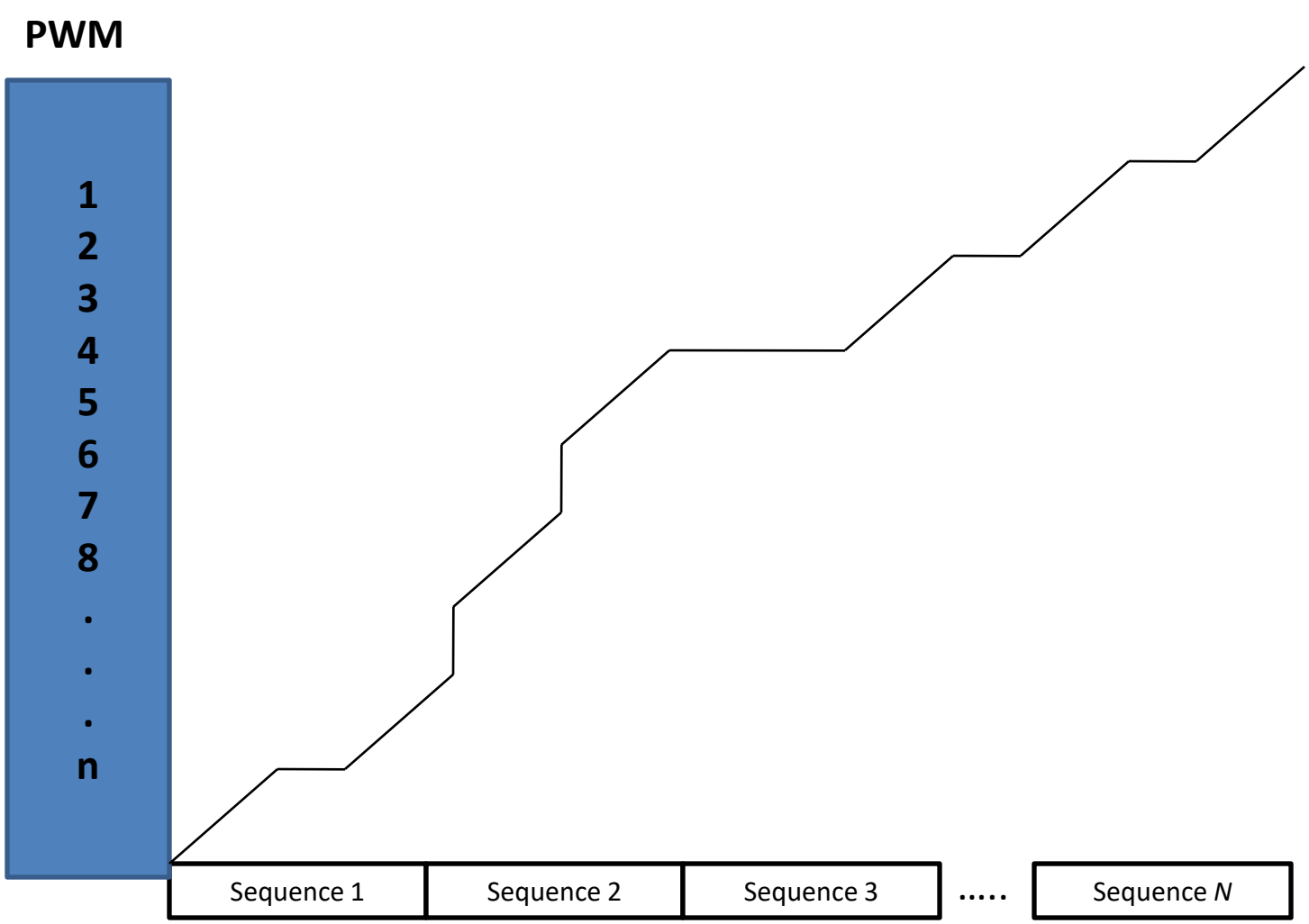
2 Union of all sequences in a single sequence one after another

3 Global alignment of united sequence with each random PWM.

4 Optimization of PWM by the genetic algorithm.

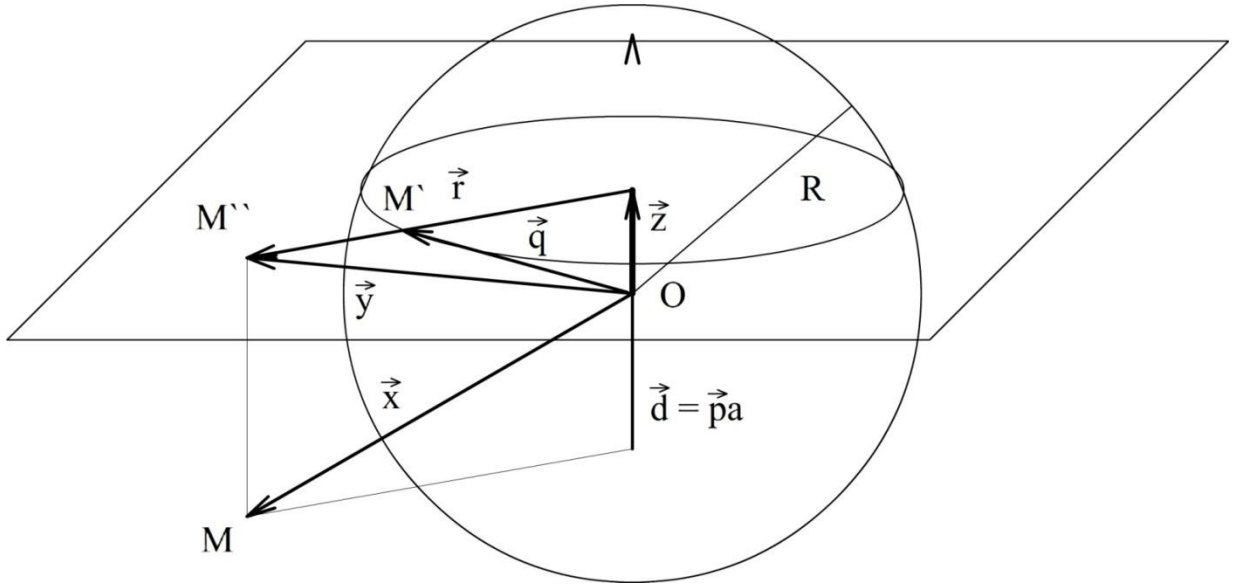
5 Selection of PWM with maximum of similarity function *F*


w_r
48 dimensional
space
w_1


PWM
1
2
3
4
5
6
7
8
.
.
.
n
Sequence 1
Sequence 2
Sequence 3
.....
Sequence N

Creation of the set of random PWM's

Space 20*n*



$$A = \sum_{i=1}^{20} \sum_{j=1}^{n} m(i, j)^2$$

$$B = \sum_{i=1}^{20} \sum_{j=1}^{n} m(i, j)\, p(i)\, f(j)$$

where $f(j)=1/n$; $p(i)=n(i)/N$, $n(i)$- number of aa i in sequence A, N – length of sequence A, n – length of period.

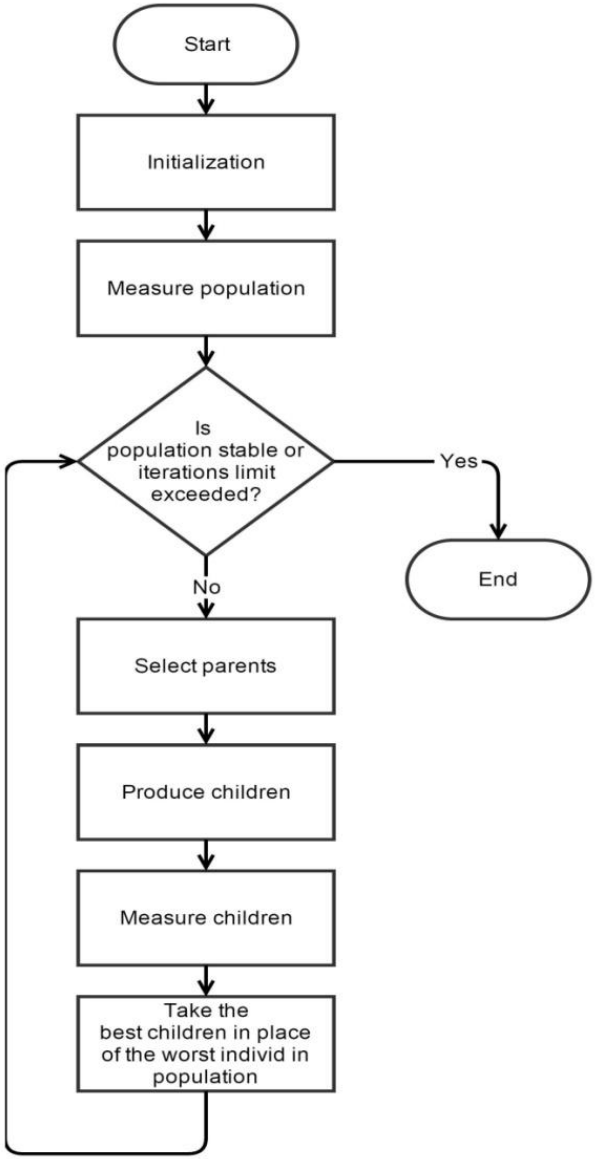
$$F(i, j) = \max \begin{Bmatrix} F(i-1, j-1) + m(a(i), b(j)) \\ F_1(i-1, j-1) - d \\ F_2(i-1, j-1) - d \end{Bmatrix}$$

$$F_1(i, j) = \max \begin{Bmatrix} F(i-1, j) - d \\ F_1(i-1, j) - e \end{Bmatrix}$$

$$F_2(i, j) = \max \begin{Bmatrix} F(i, j-1) - d \\ F_2(i, j-1) - e \end{Bmatrix}$$

$$Z = \frac{F_{\max} - M\left[\vec{F}\right]}{\sqrt{D\left[\vec{F}_{\max}\right]}}$$

Improving the matrixes by genetic algorithms



Volume of PWM population is 500 matrixes for each *n*.
n – period length

Goal is to increase the $F_{\max}$.

Types of tandem repeats

1. Perfect periods

GCAT GCAT GCAT GCAT GCAT…

GCAT
GCAT
GCAT
GCAT
GCAT

There are no changes per symbol

2. Periods with indels and changes of the symbols

GGAT GCAG *CAT GCGT G*AT…

GGAT
GCAG
***CAT**
GCGT
G*AT

There are 1.0 or less changes per symbol

Types of tandem repeats

3. Latent periods without indels

GCAT CGAC TACT AGGT GGTA…

GCAT
CGAC
TACT
AGGT
GGTA…

There are 1.0 or more changes per symbol

4. Latent periods with indels

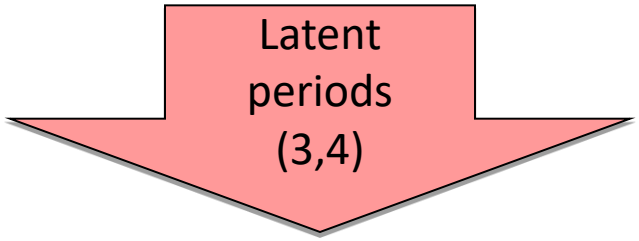
G*AT CGAC TA*T AGGT GG*A…

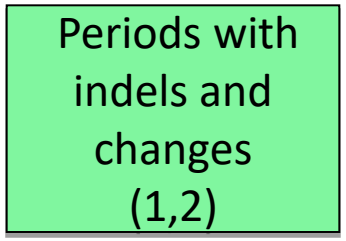
G*AT
CGAC
TA*T
AGGT
GG*A…

There are 1.0 or more changes per symbol

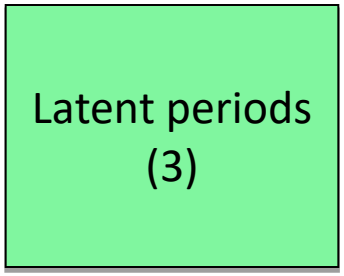
Methods applied for periodicity search

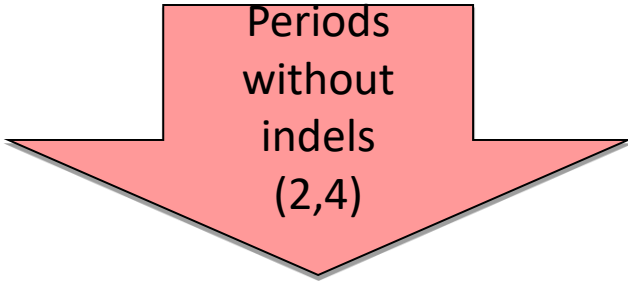
1. The dynamic programming



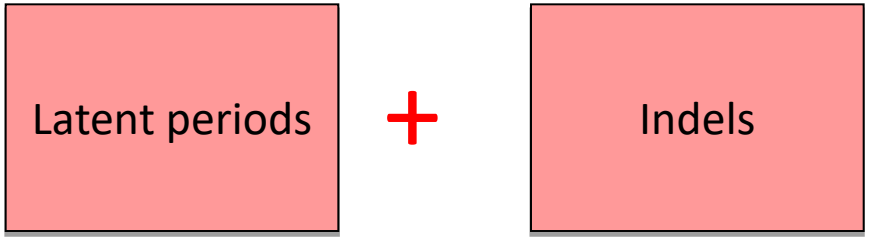


1. Spectral Approach (Fourier transform, Information Decomposition,…)

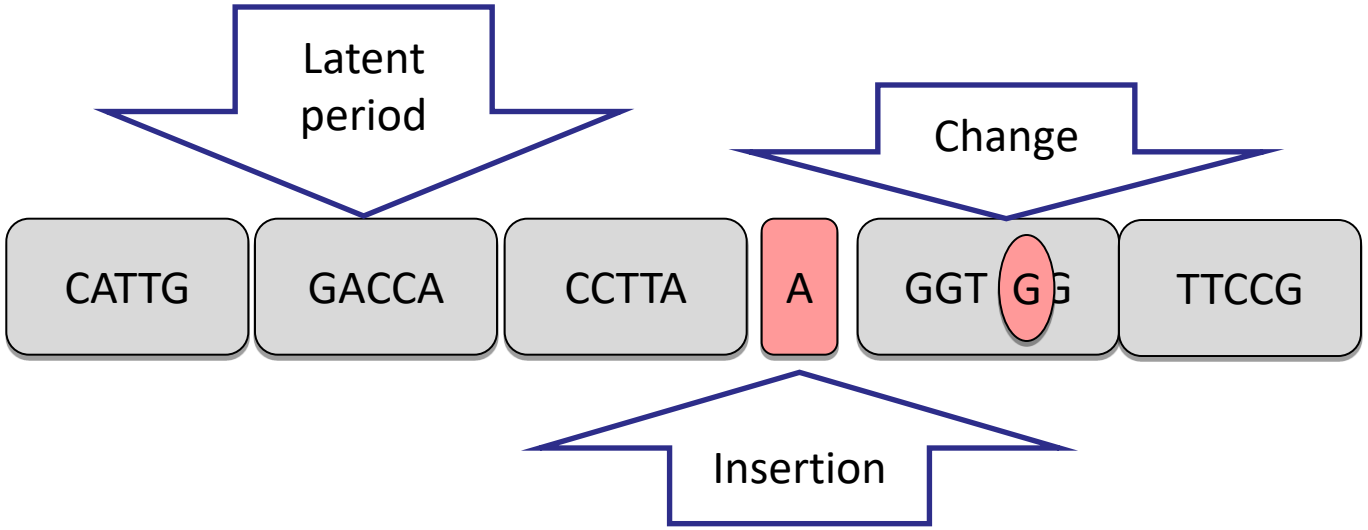




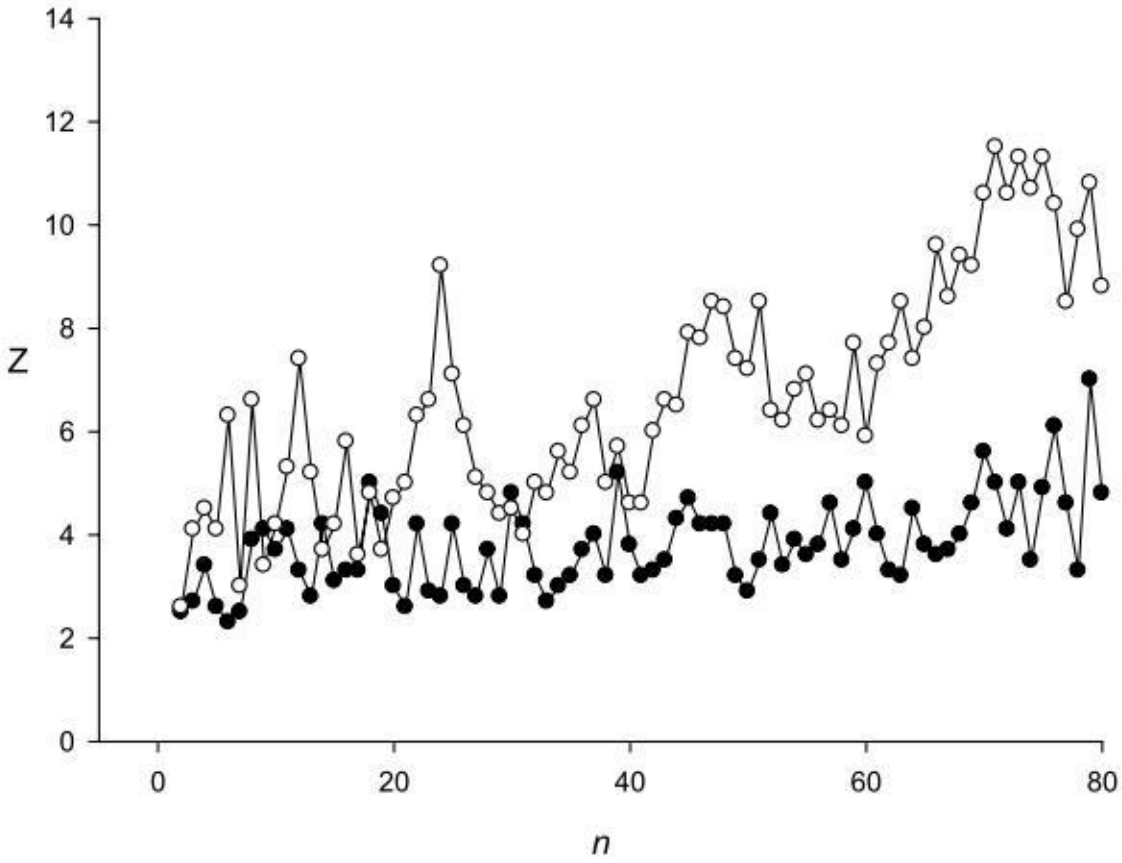
1. ?



Type periodicity which we are searching



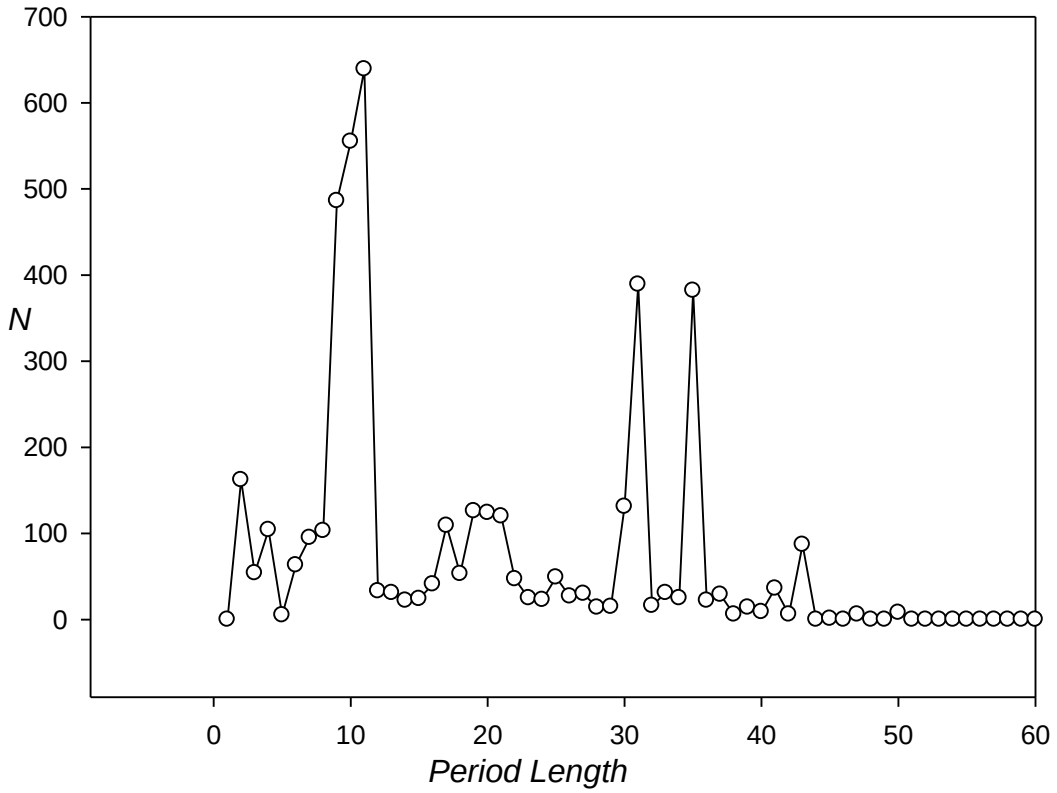
Test sequence $(actgattcgagctagtaacgggct)_{30}$ with 1.5 random mutation per nucleotide.

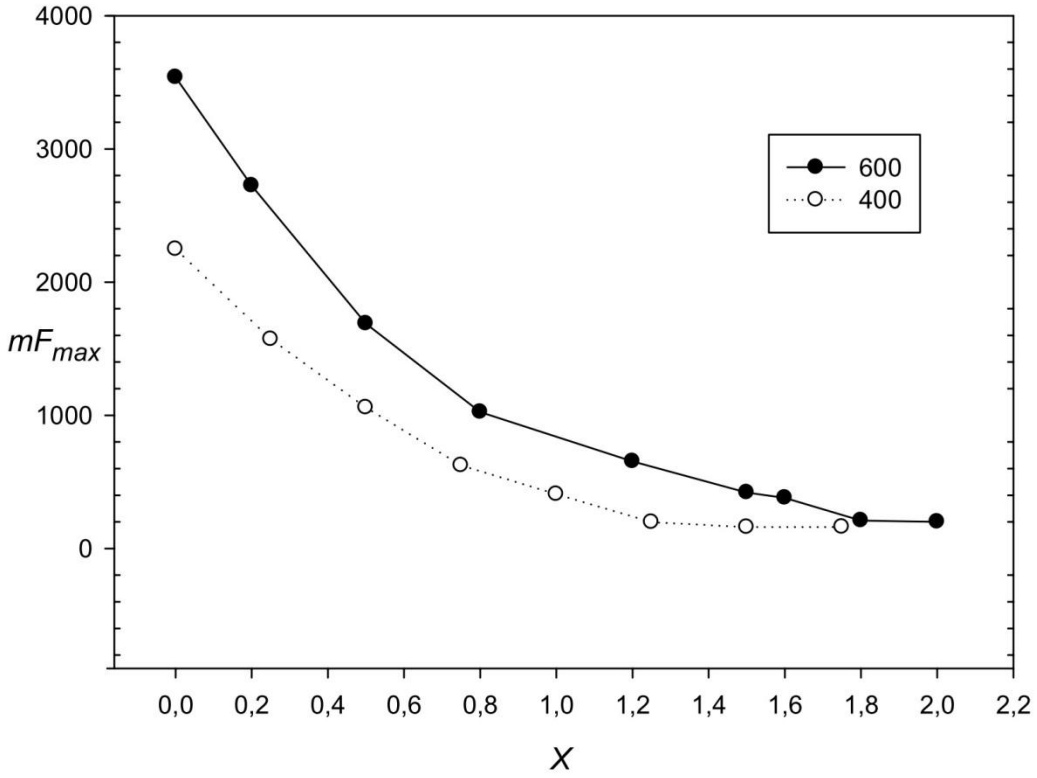


Regions with periodicity in C.elegans genome

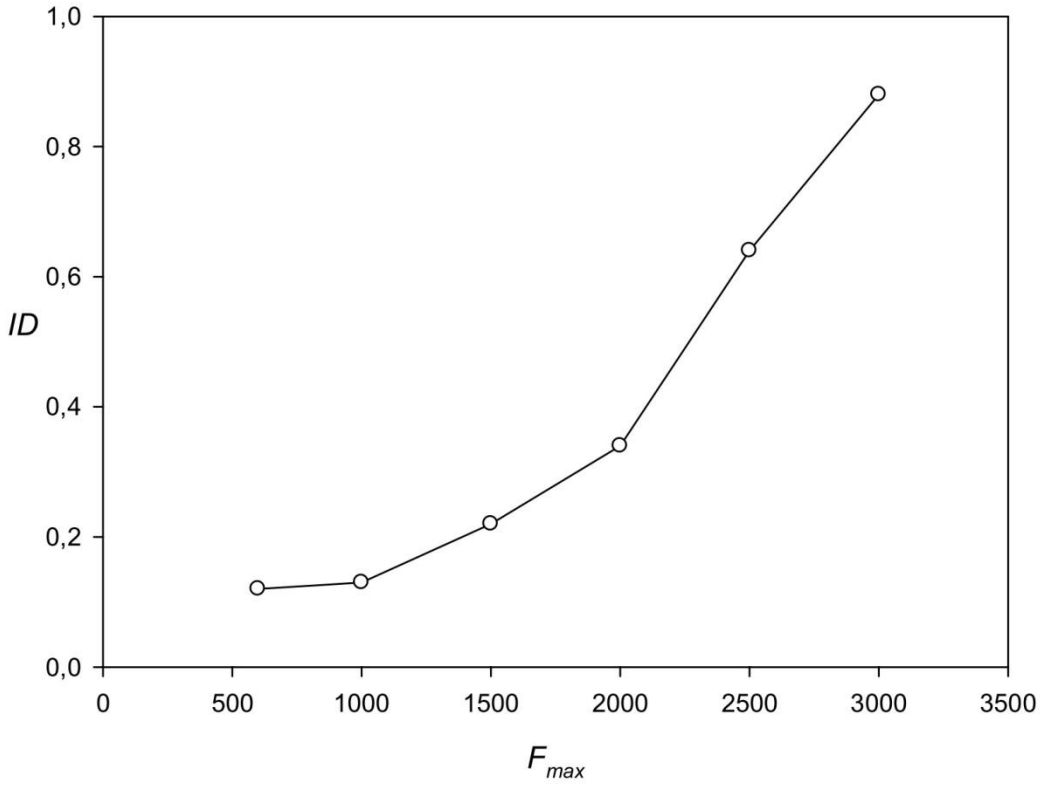
25360 regions having a periodicity with length of 2 to 50 bases

About 12% genome have periodicity of different length



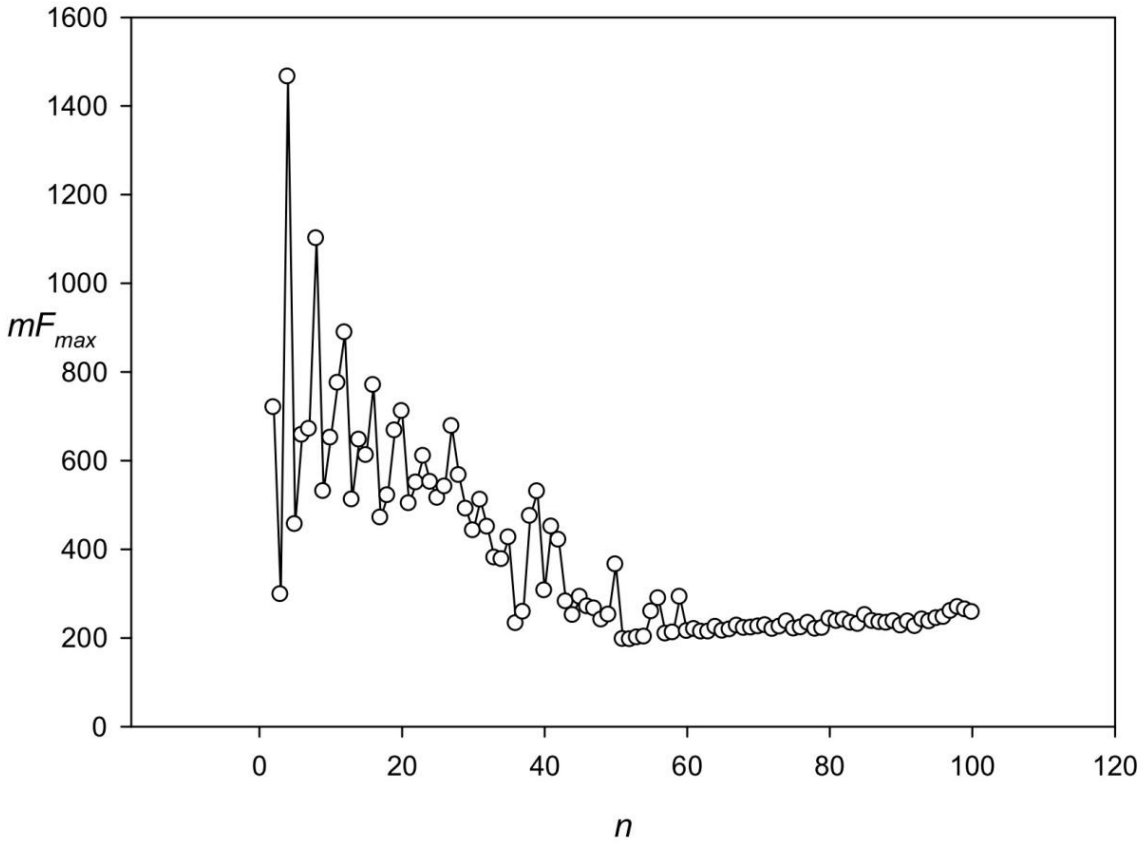


Influence of base changes on $mF_{max}(20)$ for sequences 400 and 600 base pairs. *X* is the number of base changes per 1 nucleotide. The period length equals to 20 b.p.



Comparison of developed algorithm with the program T-REKs (Jorda & Kajava 2009). *ID* shows the part of periodicities regions which can find the T-REKs. We can assume that the results are the same if the T-REKs detects at least 50% of the number of periods and the period length differs not by more than one base.

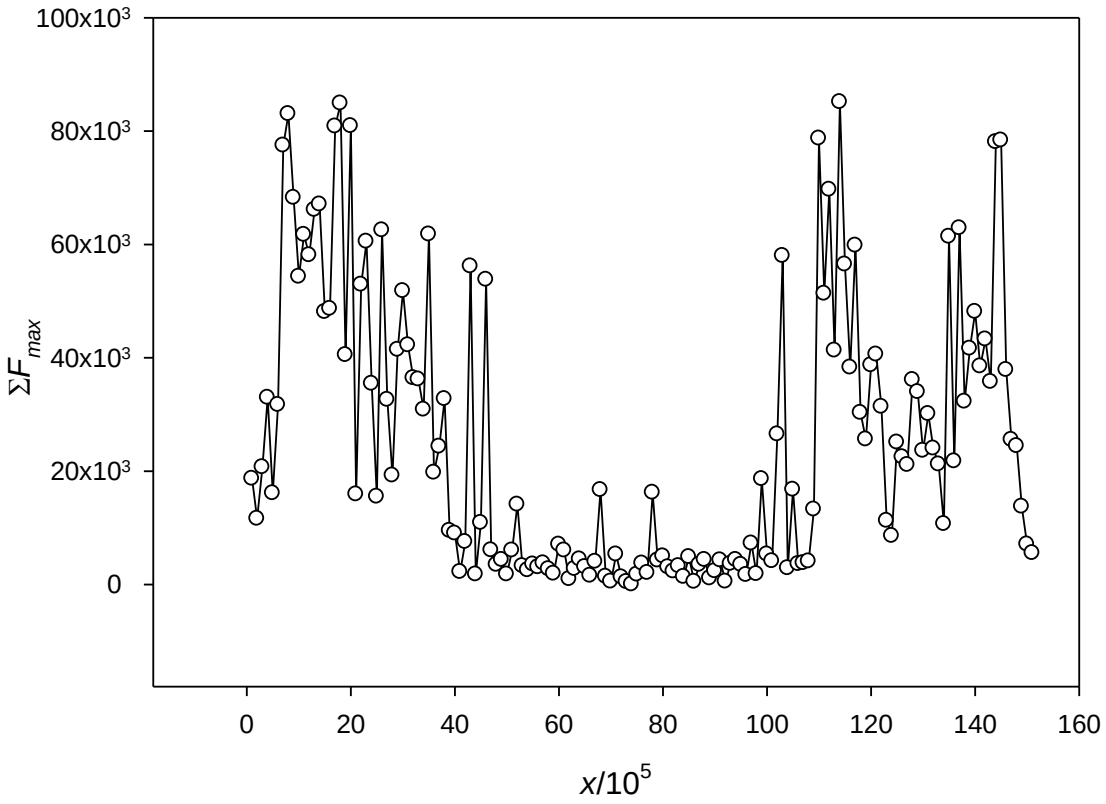
$mF_{max}(n)$ spectrum for fragment of the sequence NC_003279.8 from chromosome 1 of the C.elegans genome. The coordinates of fragment are: 887101-887688.



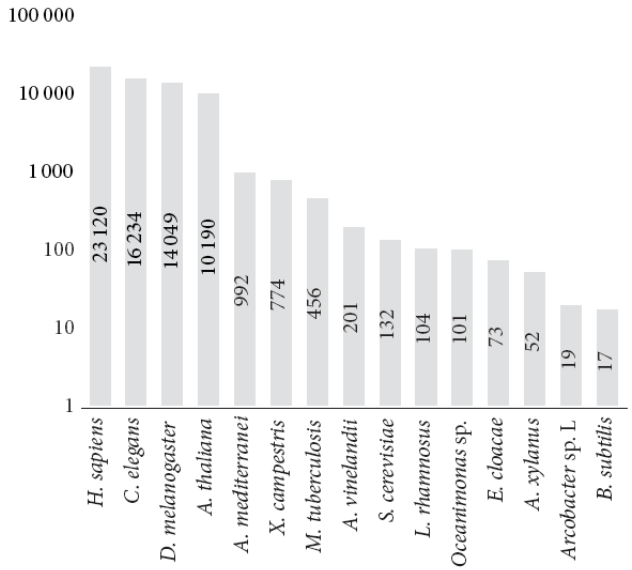
Alignment of sequences S_1 and S for the fragment of sequences NC_003279.8 from chromosome 1 of the C.elegans genome. The coordinates of fragment are: 887101-887688.

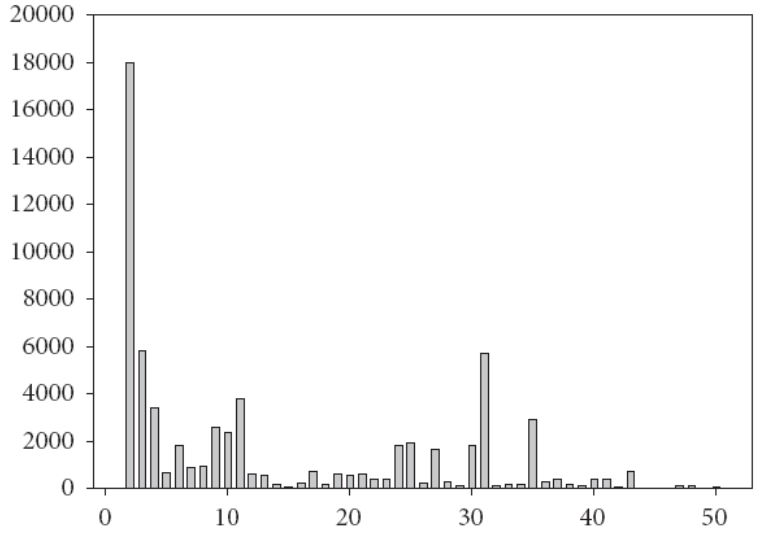
234123412341234123412341234123412341**234123412341234123412341234123412341234123412341234123412341234123412341234 ACCTGCATGCTTTCCTACCTATTCGCATACTTAGGCGTACCTACC*ACAAGCTTAGCTACCTACAAAACTGCCTTC*TGCCAACCTGCATGACTGCG
123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341 TACCTTTTCGCCTGCTTAGGTGCCTACC*ACAAGCCTCCCT*TTTGTCGATATTCCTACGCGCCTATG*GCATACCTACATGCCTACTTGTTTGAGT
234123412*341234123412341234123412**23412341**2341234123412341**12341234123*412341*2341234123*412341234 CGATACATATCATGCCTACGCGCCTACCTTGGCCTACCTTGGCCTACAATCCTGAAATTGCACTTTTCCTGTGCCTAGCTTGTACACTATACTTA**
123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341 TGCCTGCCT**TTACGCGCCTTC*TCCGCGCTTATCTGCCTACAAGCATACCTTC*TGCCTCCTTGCATGCCTACATACCTATTAGCCTGCTTAC**
234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412341234123412 GCGAACCTACC*ACAAGCCTAGCTGCCTACAAACCTCCCTTC*TGCCTATATGCCTACGCACCTAACT*CCTGCCTGCCTATACGTCTACCT*CATG
341234123412341234123412341234123412341234123412341234123412341234123412341234123412341*234123412341234123 C*TGAAAACAGACCTGCCT*TGTACCTACCTATGCGCCTACCTA*AGATTTGCCT*CCTACGTGCCTGGCTGCCTATTTTGCCTATAGGCATTACTG
341234123412341234123412341 TCAGTGCGTCTACCCAGA*GCCTACCT

$\sum F_{max}$ for the chromosome 1 of the genome *C.elegans* . x is the position in the chromosome. $\sum F_{max}$ is the sum of mF_{max} for all sequences with periodicity which are found in the region with coordinates form $(x\text{-}1)/10^5$ to $x/10^5$.



Database of periodic DNA regions in major genomes





http://victoria.biengi.ac.ru/cgi-bin/indelper/index.cgi

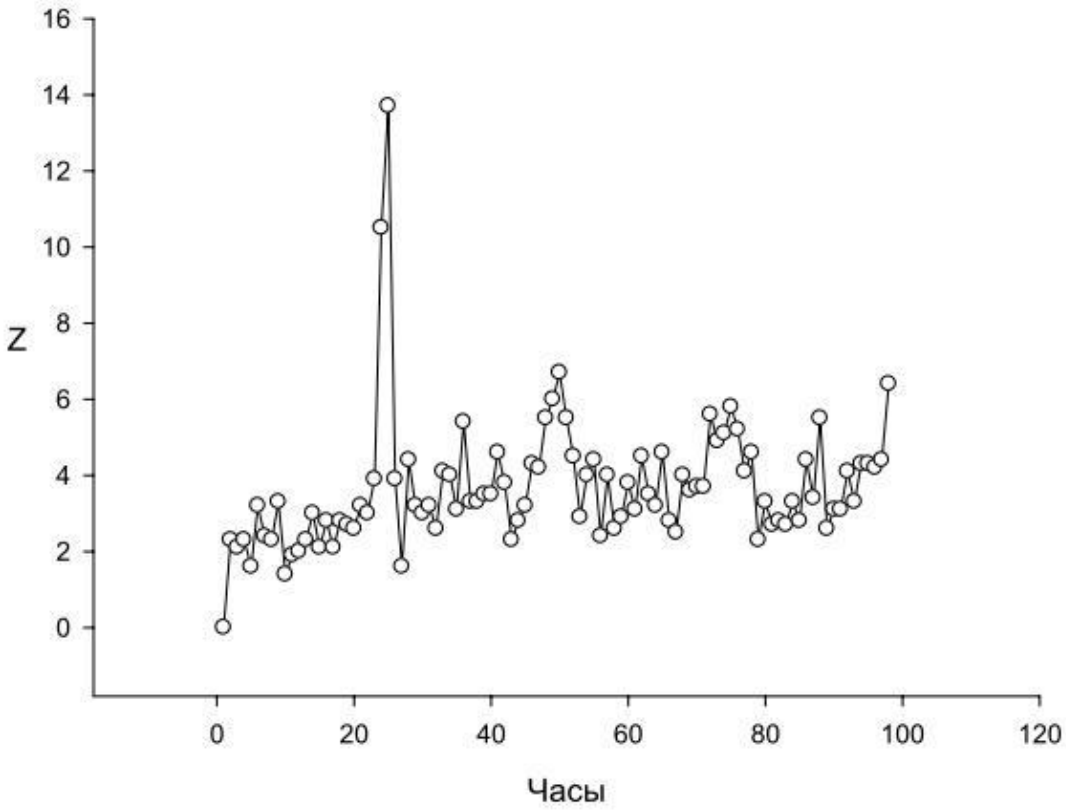
Coding of the sequence $s(i)=x_2(i)-x_1(i)$ for Euro/US$ rates

$x_1(i)$ -beginning of the candle (occurs at the beginning of each hour)

$x_2(i)$ -end of the candle (occurs at the end of each hour)

K	N	I	M	T	R	S	L	Y	F
-0.03700 -0.01140	-0.01140 -0.00780	-0.00780 -0.00550	-0.00550 -0.00390	-0.00390- 0.00260	-0.00260- 0.00160	-0.00160- 0.00100	0.00100- 0.00040	0.00040- 0.00010	-0.00010 0.00020
C	W	P	H	Q	V	A	D	E	G
0.00020 0.00060	0.00060 0.00100	0.00100 0.00150	0.00150 0.00230	0.00230 0.00320	0.00320 0.00460	0.00460 0.00610	0.00610 0.00820	0.00820 0.01140	0.01140 5.00000

Z(*n*) for Eur/US$ for candle equal to 1 hour

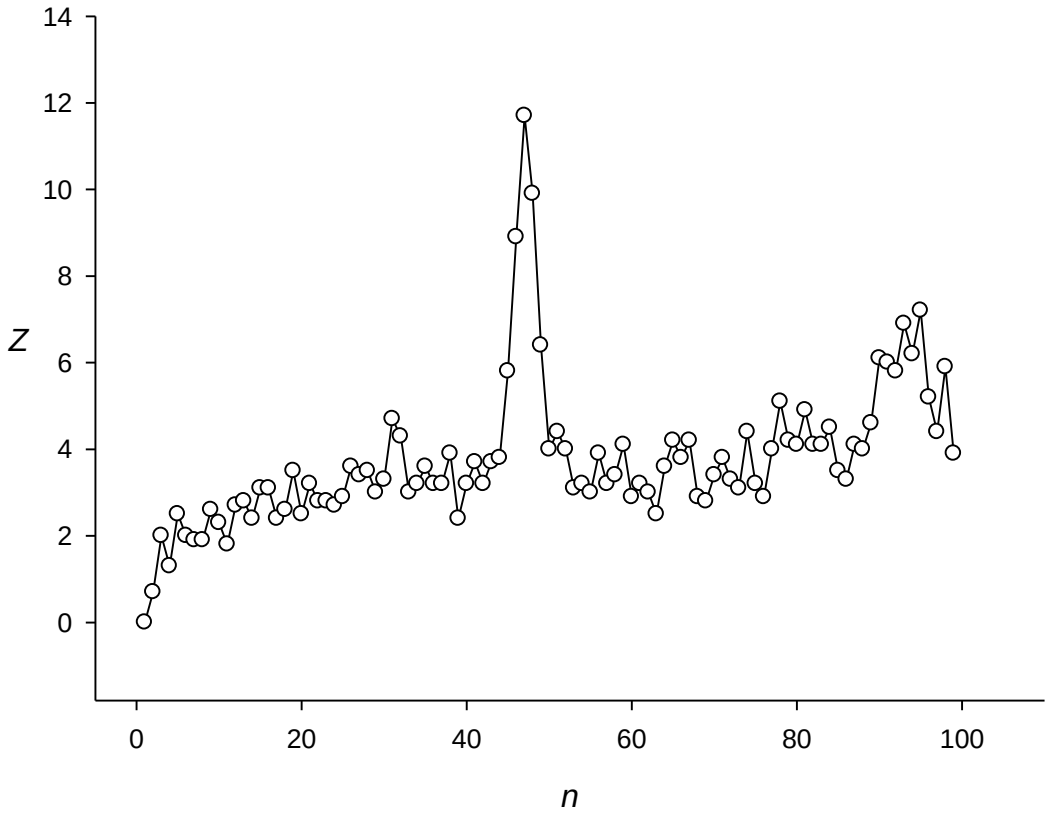


Period equals to 24 hours was found for last 20 days. We have same period for another intervals. Alignment contains 46 indels of different size.

Fragment of the PWM for n=24 hours from 1-12 position of period

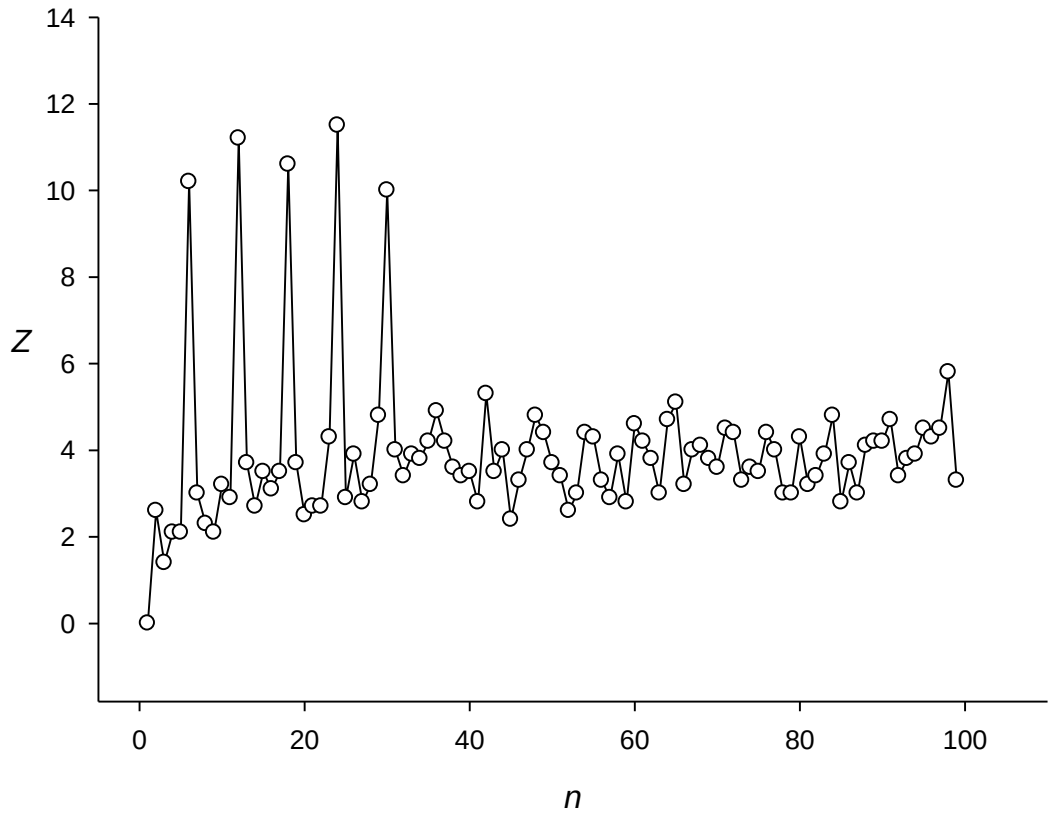
	1	2	3	4	5	6	7	8	9	10	11	12
K	2.2	2.5	1.9	0.1	1.9	2.4	2.4	2.5	-3.5	1.3	-0.5	-1.7
N	2.6	-1.5	1.4	-2.0	1.7	0.9	0.9	1.7	2.4	1.8	2.1	-3.8
I	1.6	0.7	-1.1	2.0	2.2	-1.1	0.7	1.3	2.1	-0.5	-1.1	0.1
M	-0.6	-0.6	-0.6	-2.4	0.0	0.6	1.2	1.2	-2.4	0.6	2.1	-0.6
T	-0.6	-1.8	1.3	0.7	-1.8	1.6	-3.6	-3.0	-2.4	0.7	2.0	-1.8
R	-2.9	0.9	-2.2	-1.6	0.9	1.5	0.2	-0.4	1.7	1.8	-1.6	-1.0
S	-1.8	1.8	-0.6	-2.4	-1.2	-3.0	-0.6	-2.4	-1.8	-1.8	0.6	1.2
L	-2.3	-2.3	-1.7	0.7	-1.1	-1.7	-0.5	-1.1	-3.5	-0.5	2.1	0.1
Y	0.4	-3.4	1.9	-1.5	-0.9	-3.4	-2.7	-3.4	-2.1	-2.1	0.4	-0.9
F	-2.2	1.0	-2.8	-1.5	-2.8	-2.2	-0.9	-2.2	-2.2	-3.4	-2.8	0.3
C	0.9	1.5	-0.3	-4.1	-1.6	-3.4	-2.8	-2.2	-1.0	-0.3	1.5	-3.4
W	-3.4	-1.6	-0.3	-1.6	-1.6	-1.6	-2.8	-2.8	-2.8	-1.6	0.3	-1.0
P	-2.9	-1.6	-2.3	0.2	-1.6	-0.4	-4.1	-1.0	1.4	0.2	-2.3	-2.3
H	0.3	-0.3	-3.4	1.9	-1.0	-1.6	-3.4	1.5	-1.6	-0.3	-2.8	-2.2
Q	-1.2	-1.2	1.8	-0.6	-3.6	0.6	-4.2	-1.8	2.0	-3.6	-3.0	2.1
V	-2.3	-0.5	1.6	-1.7	-2.3	-2.3	-1.7	-1.7	0.1	-1.7	-4.1	2.2
A	1.3	0.7	-1.8	1.6	1.9	-1.2	1.8	1.8	-2.4	-1.2	-2.4	2.3
D	0.4	-0.1	0.4	1.0	-1.9	2.1	-0.1	-1.9	2.0	-1.3	1.0	1.0
E	-1.6	0.2	1.5	-0.4	0.9	-1.0	2.5	2.2	-2.9	2.5	-2.9	-0.4
G	0.7	-1.7	0.7	2.5	2.2	2.2	2.2	-0.5	2.2	1.6	1.4	-1.1

Z(*n*) for Eur/US$ for candle equal to 30 minutes



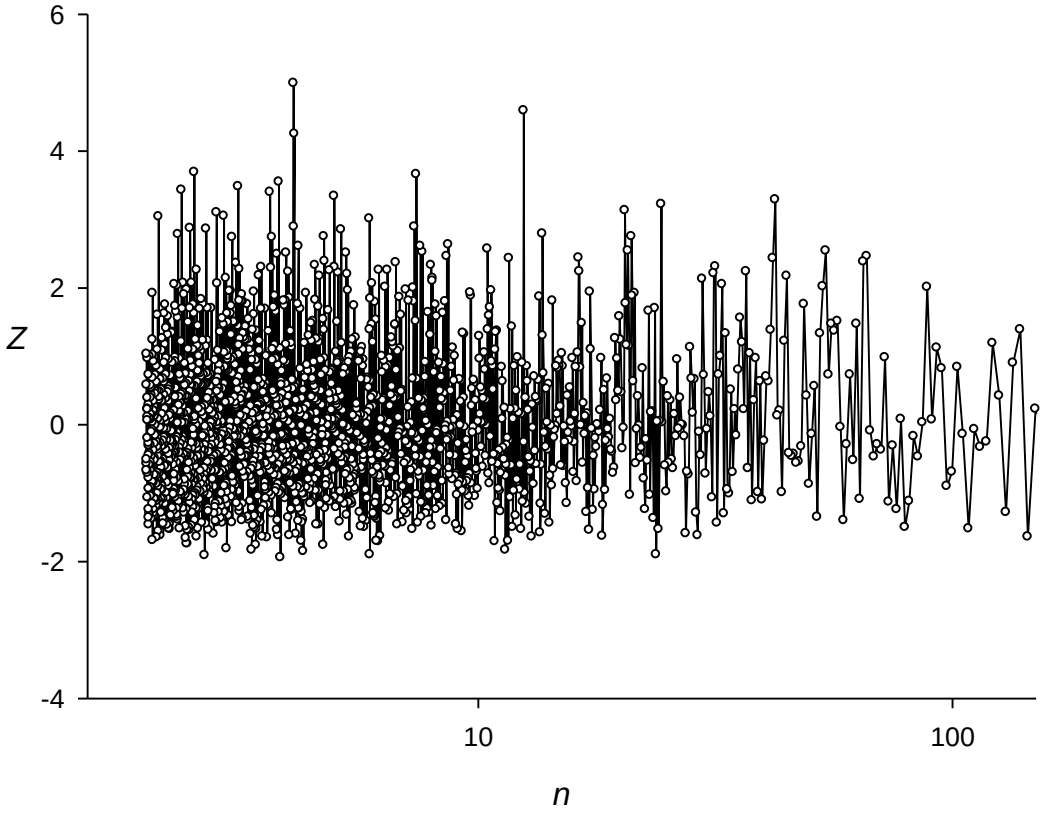
n – length of period

Z(*n*) for Eur/US$ for candle equal to 4 hours



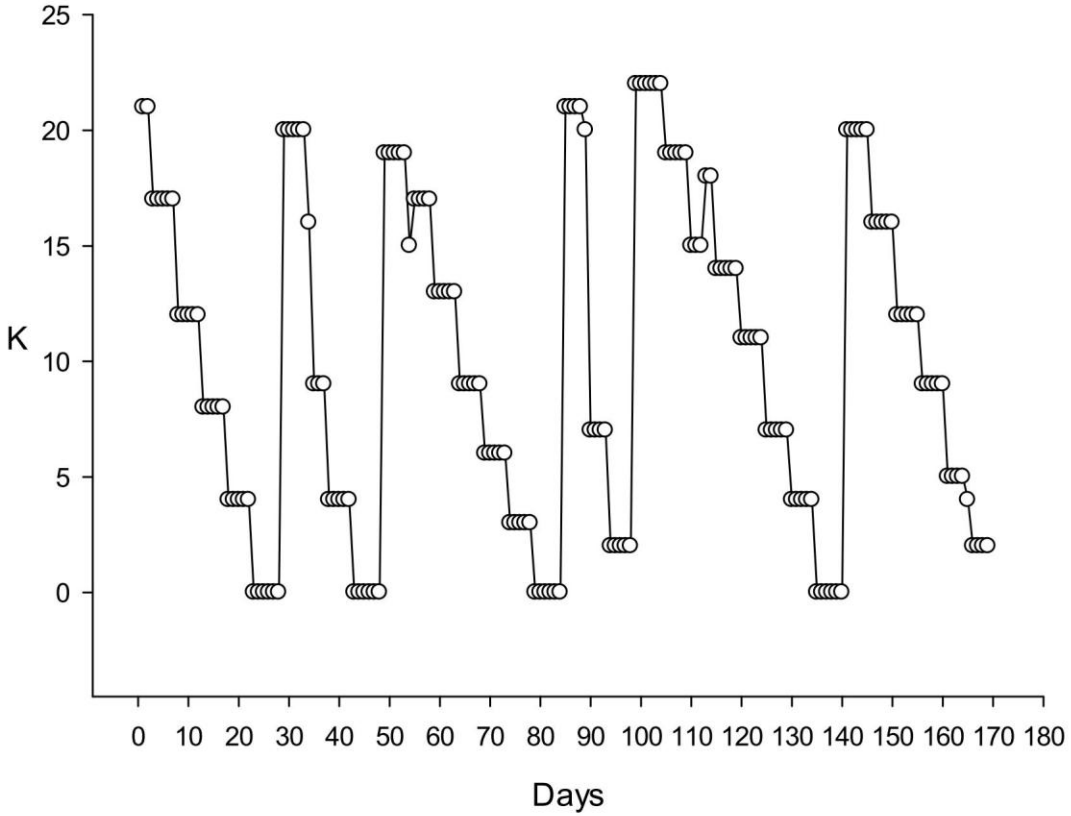
n – length of period

Fourier transform for Eur/US$ for candle equal to 1 hour

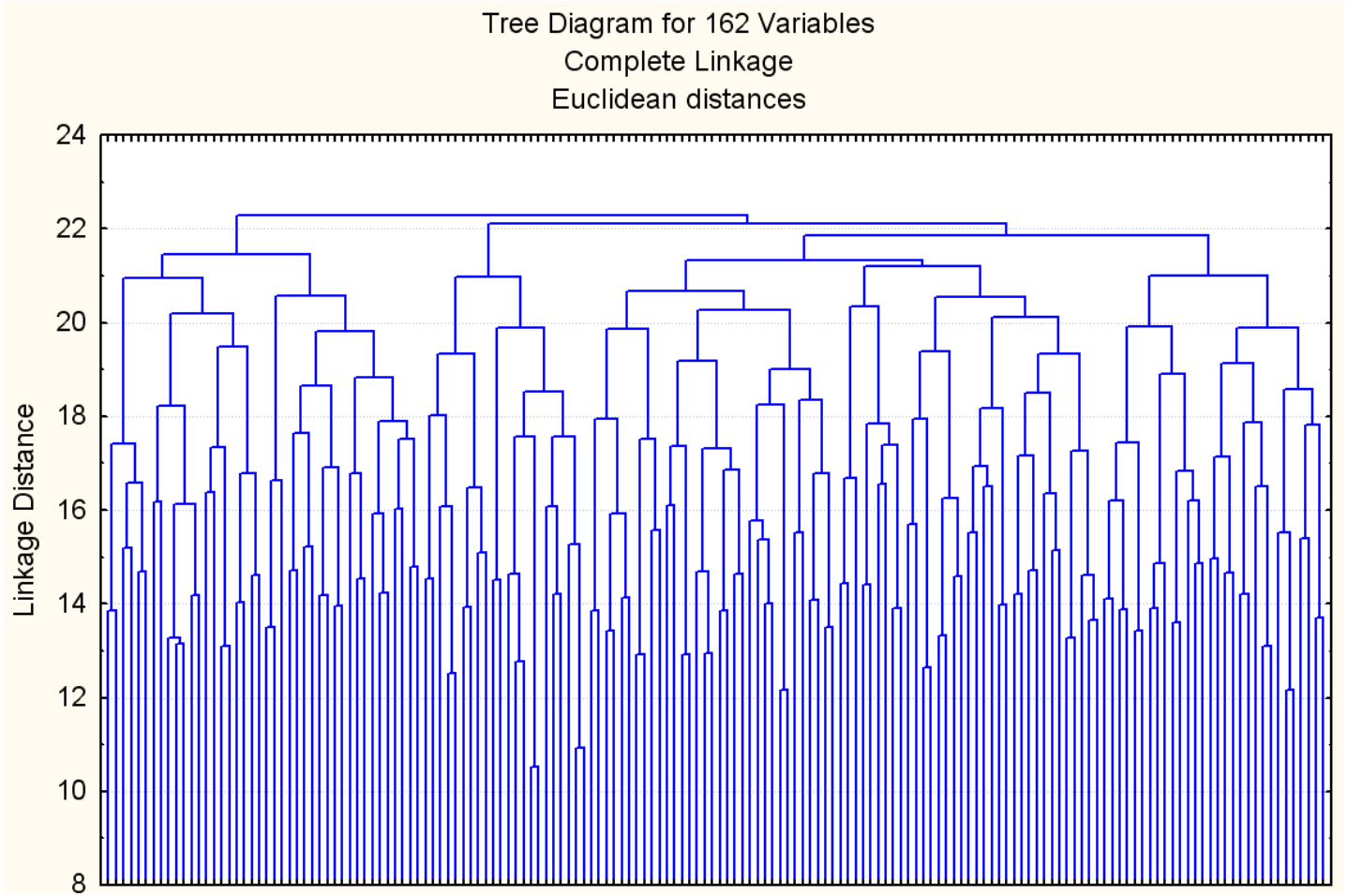


Phase shift of period equal to 24 hours

K- position of PWM which corresponds to the first candle of each day (0.00-1.00 hour)



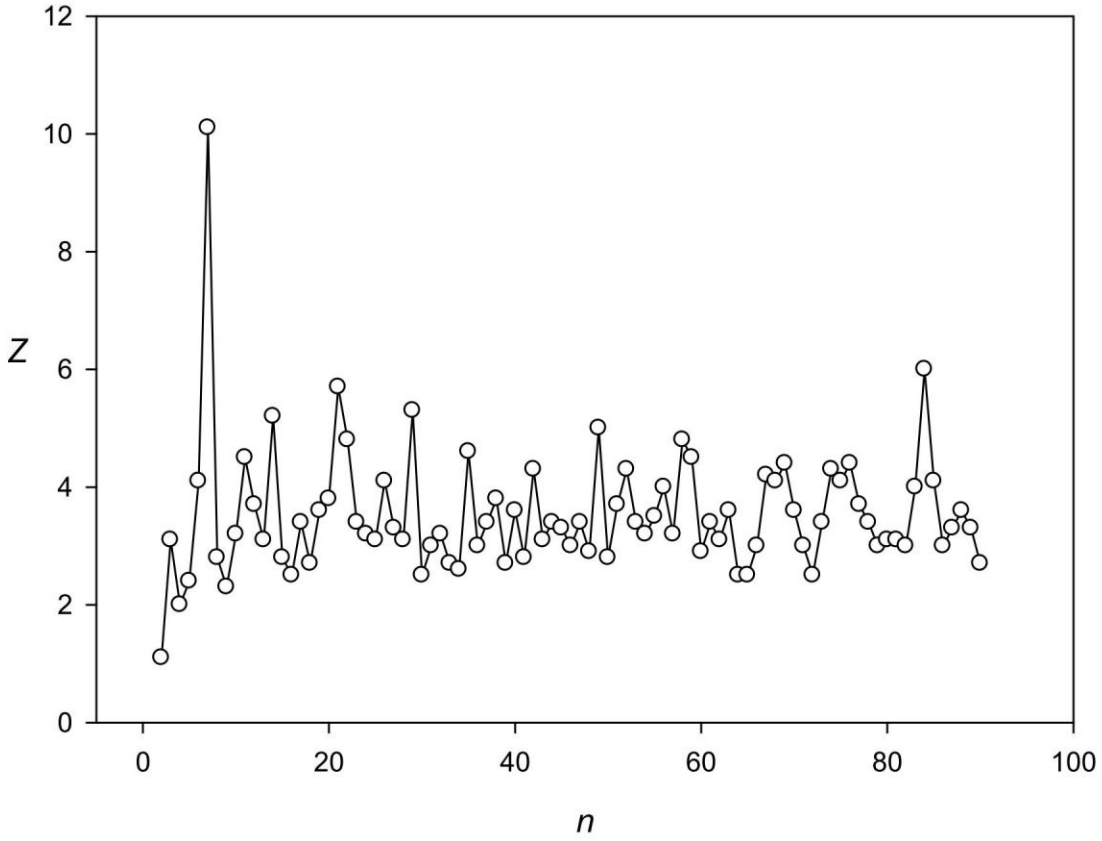
Classification periods equal to 24 hours



Multiple alignment periods equal to 24 hours

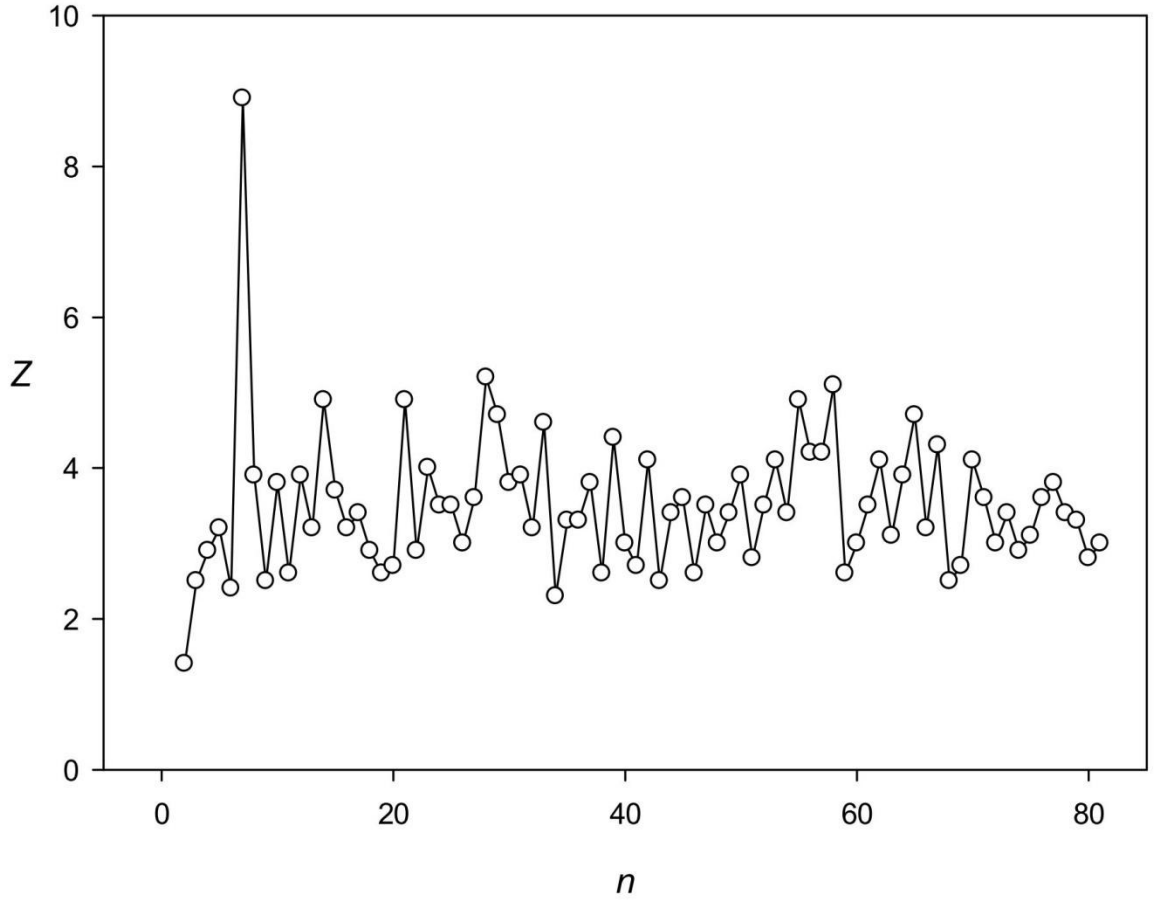
```
J=  1   WMLPA........VK.IA.MMKNAENWD.AWL.MRG
J=  2   PFQTQ........KR.GN.RNGLKLMDG.VWT.FQI
J=  3   PRRQW........KQ.MD.NWKWKMKND.RRM.MNG
J=  4   VTLVE........AE.LM.WALVVRMAW.FKA.NFD
J=  5   RMTAQ........MN.LP.KGDIEGNFE.PFL.RGL
J=  6   FWNIQ........NM.AR.DGAKQKAMW.LLT.FIP
J=  7   VRWFA........GG.ET.NGGRGWMGV.RFT.MLI
J=  8   WWWAN........AK.WA.GKMIPIVIP.WPQ.DMW
J=  9   QTFWM........DE.RK.INFPEALIM.EFF.WML
J= 10   QAPLD........KF.NK.VWGANRPID.VPK.PQV
J= 11   LMLEW........FG.MV.KKQGKGTRL.MDF.RWN
J= 12   VTFTE........KT.KG.NKFNNKVRF.RDP.FTQ
J= 13   IPTTN........EE.KI.GKNKTGVLK.WWQ.FFL
J= 14   FVMPE........KG.EP.MKPNGVMWG.RTW.TPN
J= 15   MMWEE........IK.AD.MDTIFRGFT.LVT.VWA
J= 16   MIWFI........VE.TI.DVNVGKGQL.QMV.LCL
J= 17   QKCIA........AQ.LP.NKGEMQRII.VRP.PRQ
J= 18   LRLMR........WK.VM.VRGIGMDFE.MQA.TVQ
J= 19   FRRRV........GD.KL.LDRLGTLLF.ALF.PGI
J= 20   PWWTT........PE.TW.IGKLGKLNR.PSS.LFF
J= 21   IFLAE........KQ.DM.KFQMGVIEF.LFF.I.E
J= 22   DEPTW........EL.VG.NIDFFTQPV.LDL.FSQ
J= 23   PTLVL........LG.NI.QGSKIFPAQ.QVW.NQP
J= 24   GNTLT........DN.SF.EMTPWSQND.FFQ.TLT
J= 25   WEWWT........SD.AQ.WFPGGNDTW.WPF.PLA
J= 26   PTIAF........MQ.TI.TWMQELENI.SNF.FTE
J= 27   ADAQS........TK.KV.NQDEWTKED.GWS.PFT
J= 28   WPTFS........VM.TD.NSAAGSKGD.QWP.ALQ
J= 29   NAFPF........VM.GE.TKNTNGGIP.TSW.LWT
J= 30   WPFFF........EA.PN.ENDAPKGQN.WTD.QFF
J= 31   LLLLS........AL.QS.FSNNKKWVW.EQL.STD
J= 32   DMSSQ........LN.KN.QDMGGTGNR.PSW.FLQ
```

The spectrum of $Z(n)$ obtained for the stock of Bank of America.



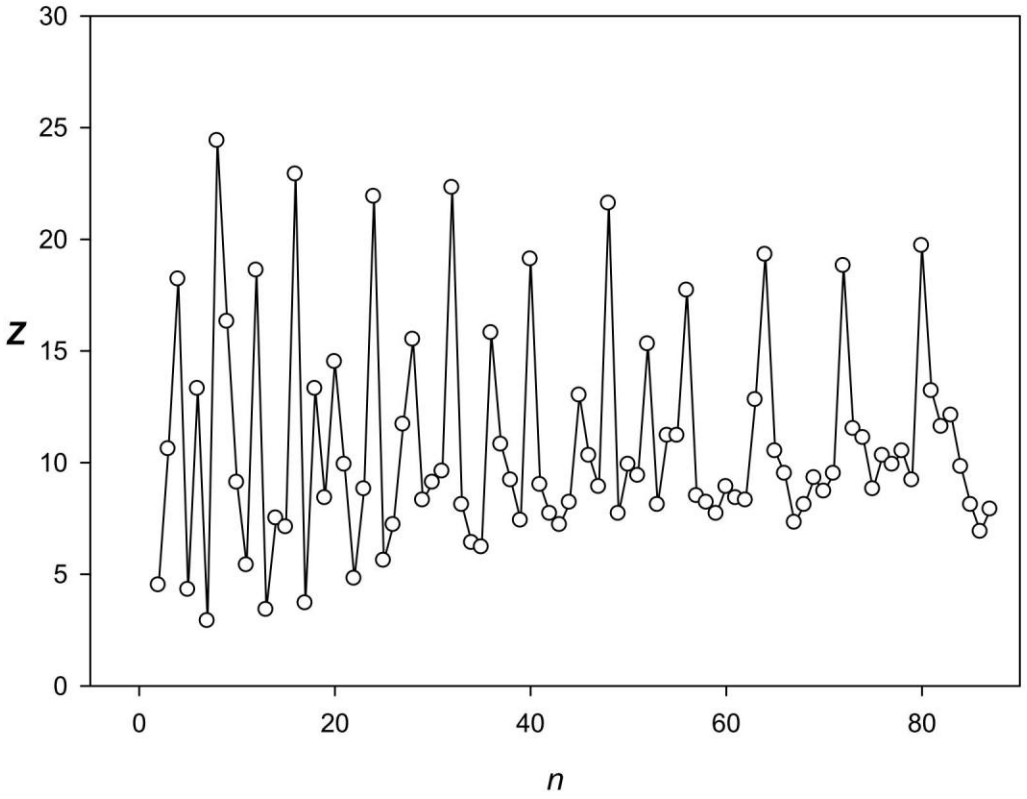
Trading Time - 7 hours every day

The spectrum of $Z(n)$ obtained for the stock of Microsoft corp.



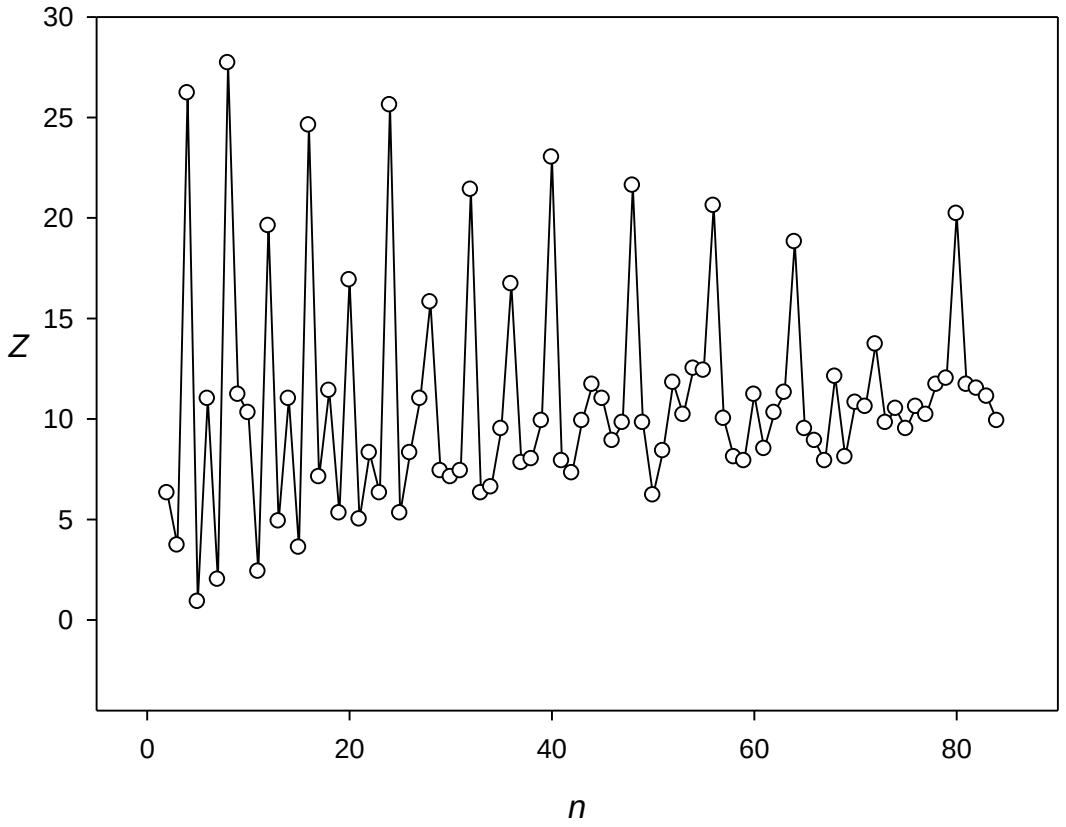
Trading Time - 7 hours every day

The spectrum of $Z(n)$ obtained for the S&P500.



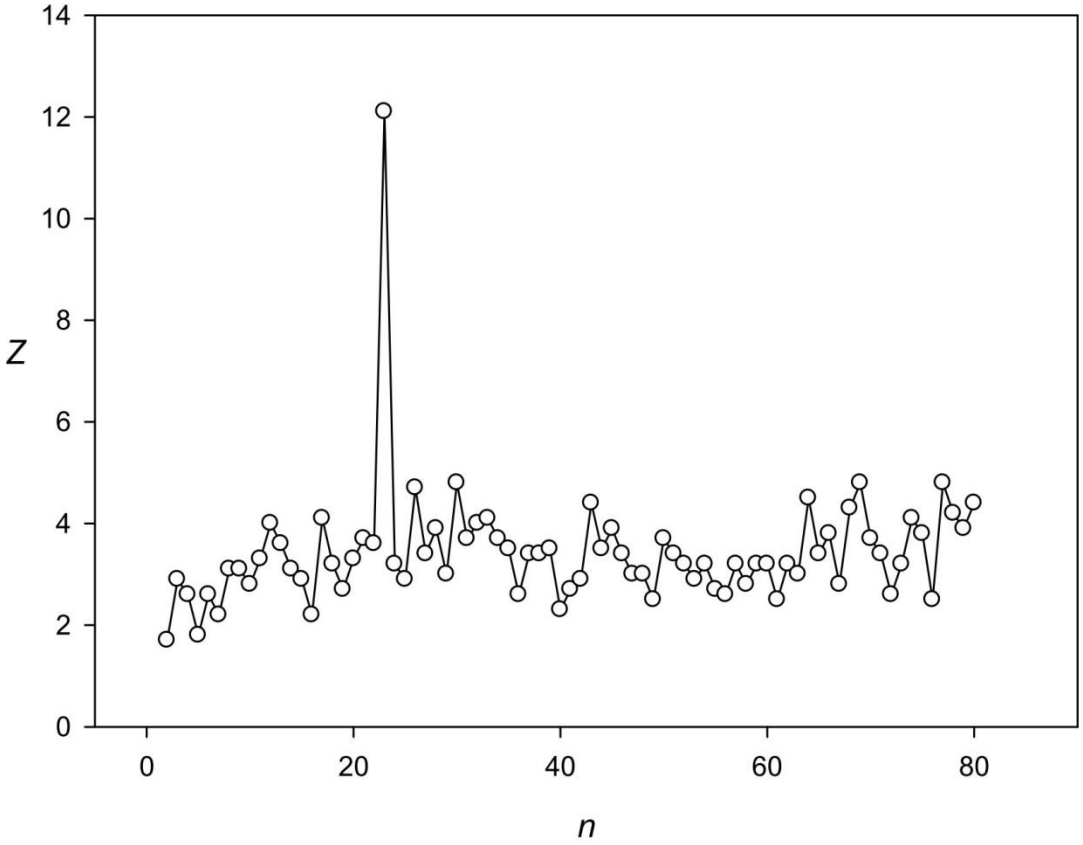
Trading Time - 8 hours every day

The spectrum of $Z(n)$ obtained for the NASDAG.



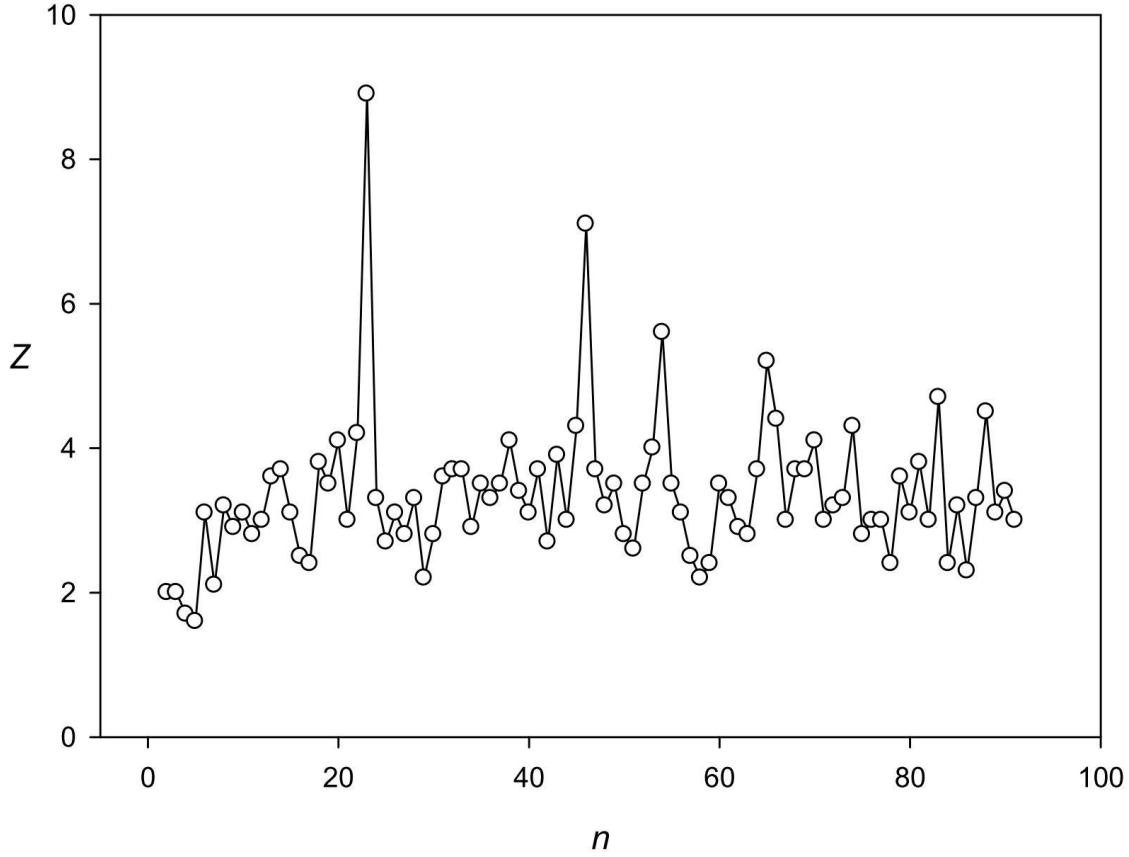
Trading Time - 8 hours every day

The spectrum of $Z(n)$ obtained for the Gold price.

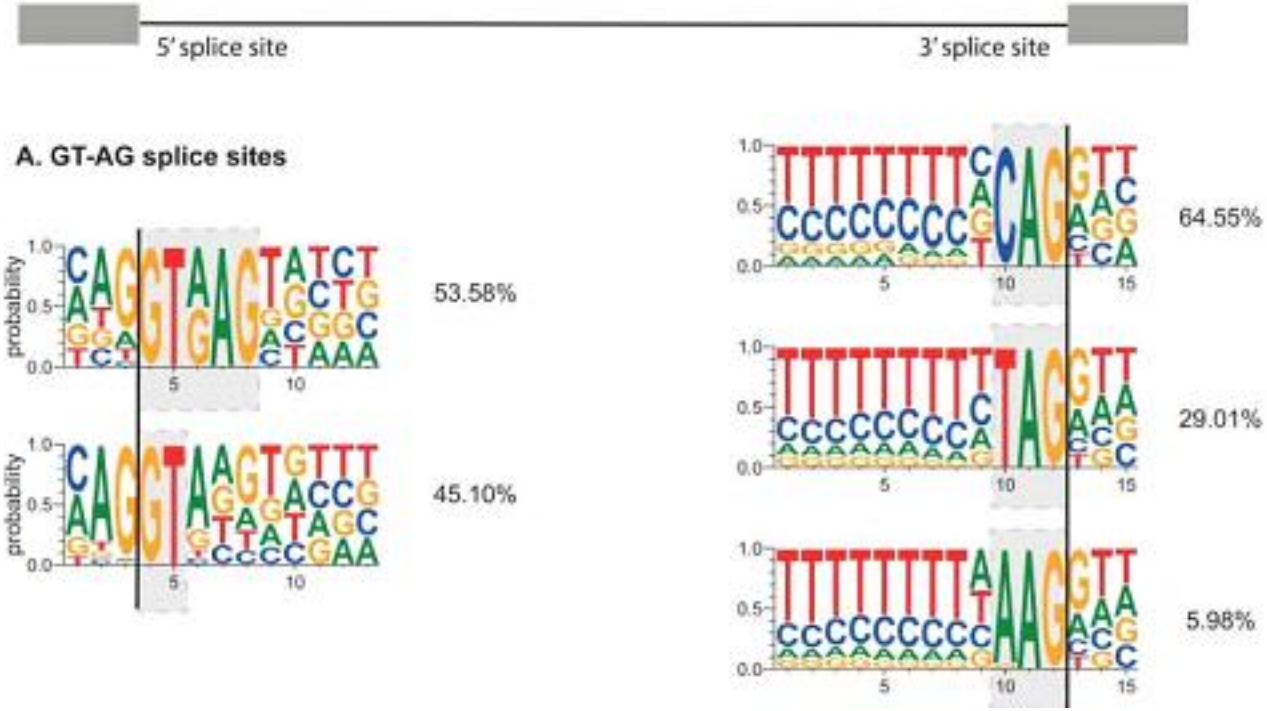


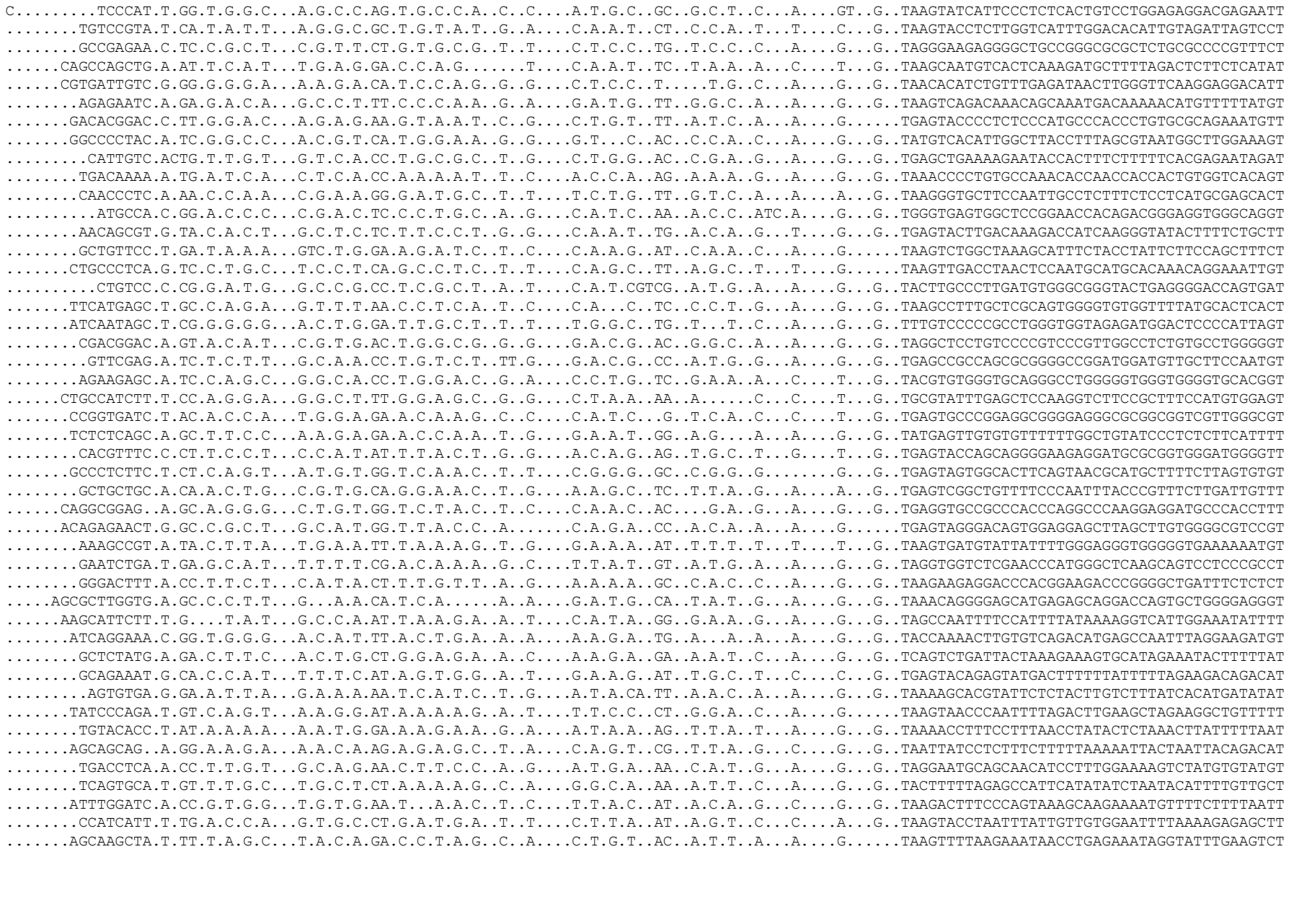
Trading Time - 23 hours every day

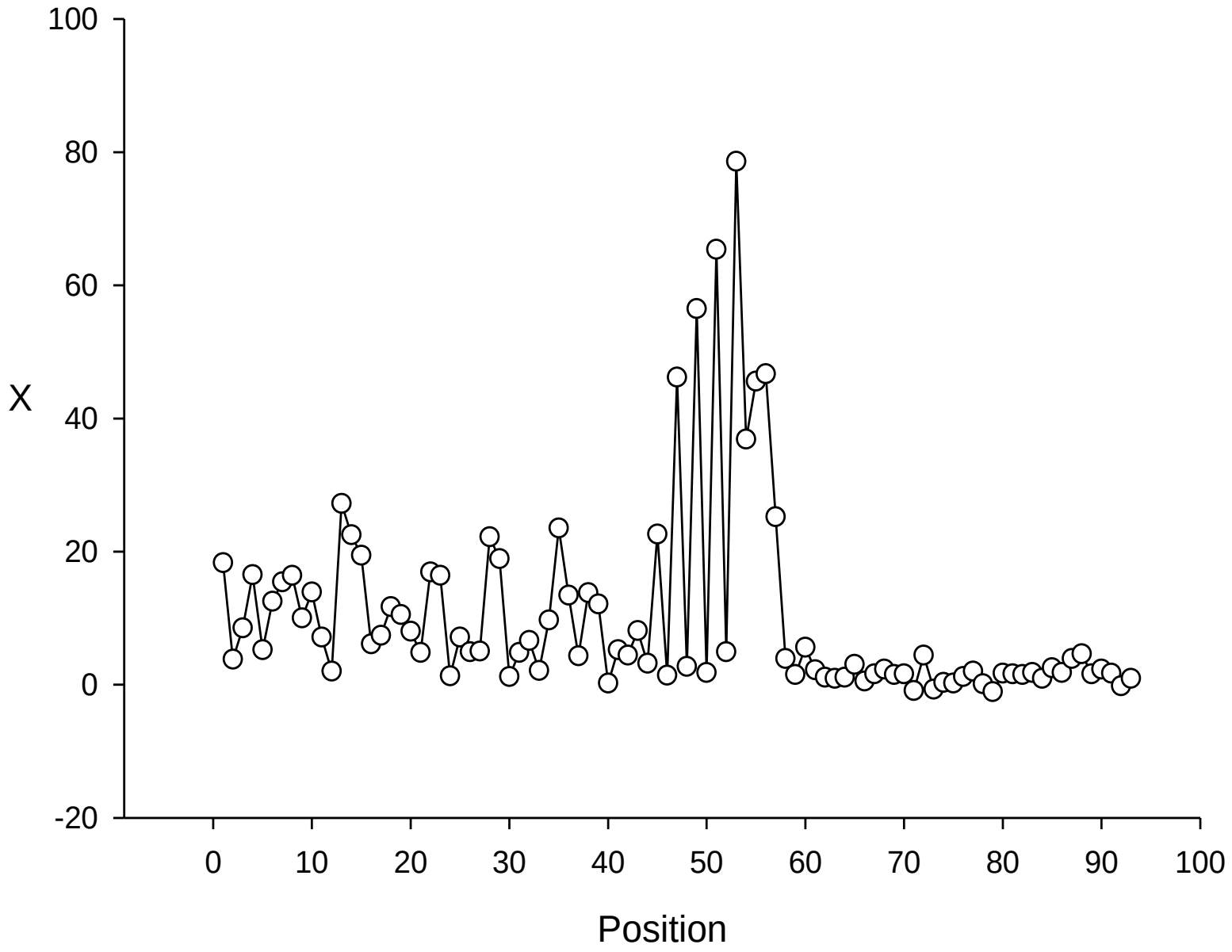
The spectrum of $Z(n)$ obtained for the Silver price.



Trading Time - 23 hours every day


5' splice site
3' splice site
A. GT-AG splice sites
probability
1.0
0.5
0.0
5
10
15
53.58%
45.10%
64.55%
29.01%
5.98%




100
80
60
40
20
0
-20
X
0
10
20
30
40
50
60
70
80
90
100
Position

CONSENSUS=TTACCGCATCACCAGCGACTGCAACATCAACTTGC
ACCTTACGGTAACGGGGTAAGTTGCAAAAATCAGAAGGAGAAAA
TTAGTAATTATGCA

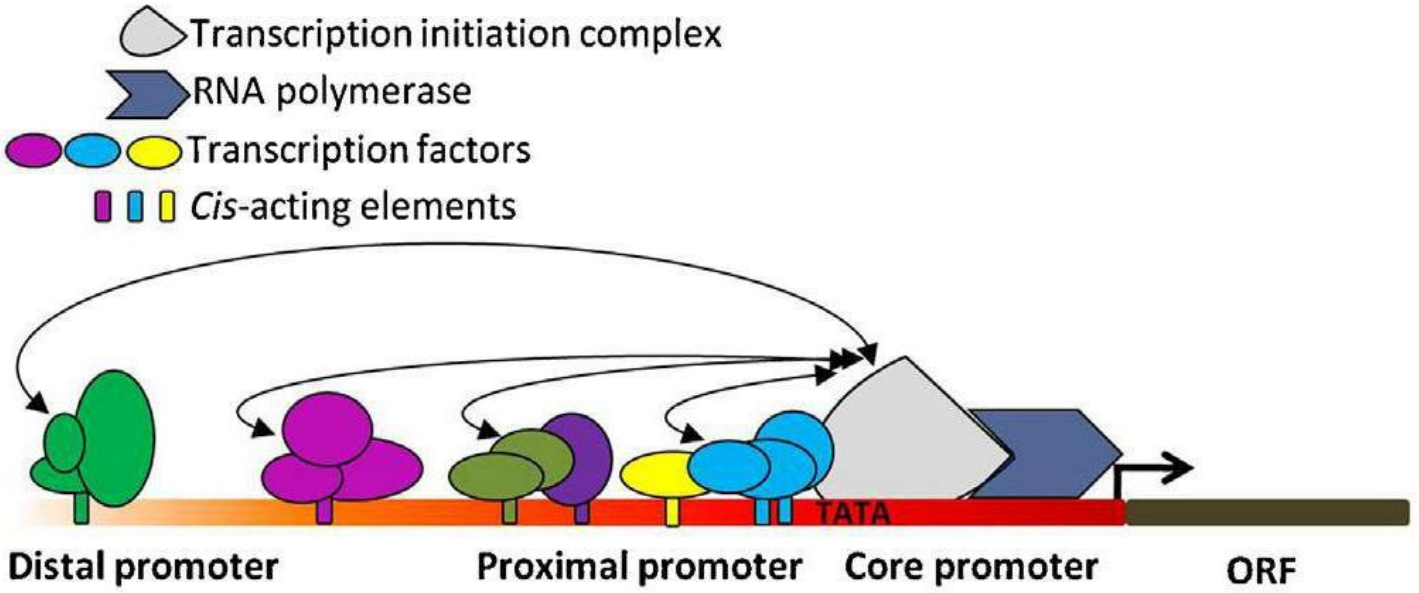
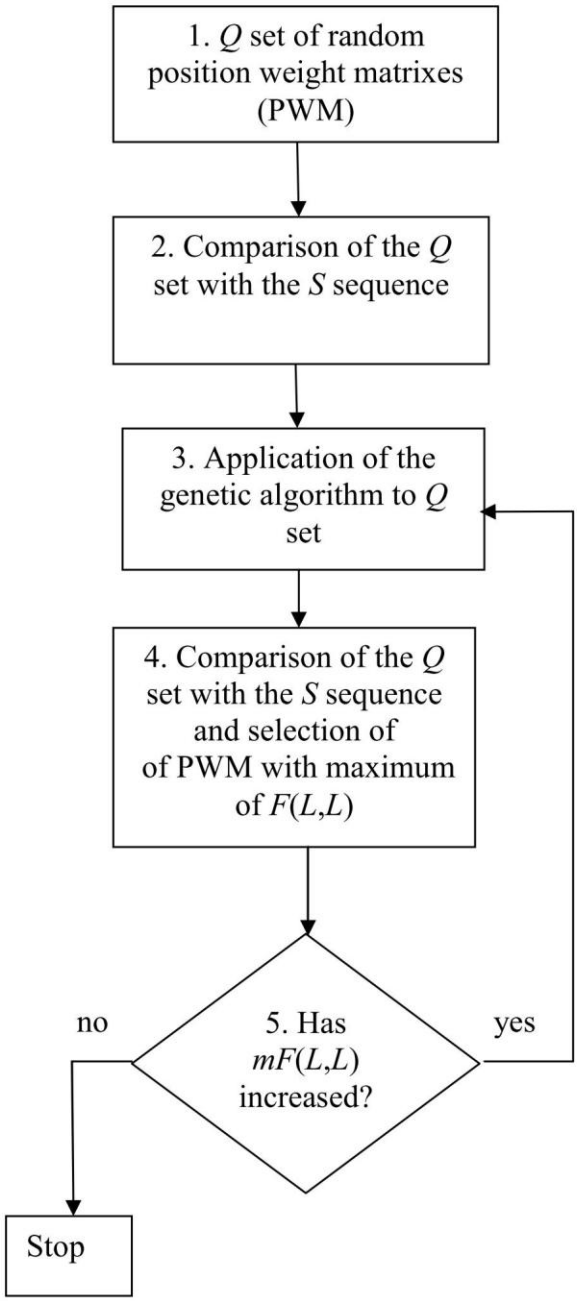
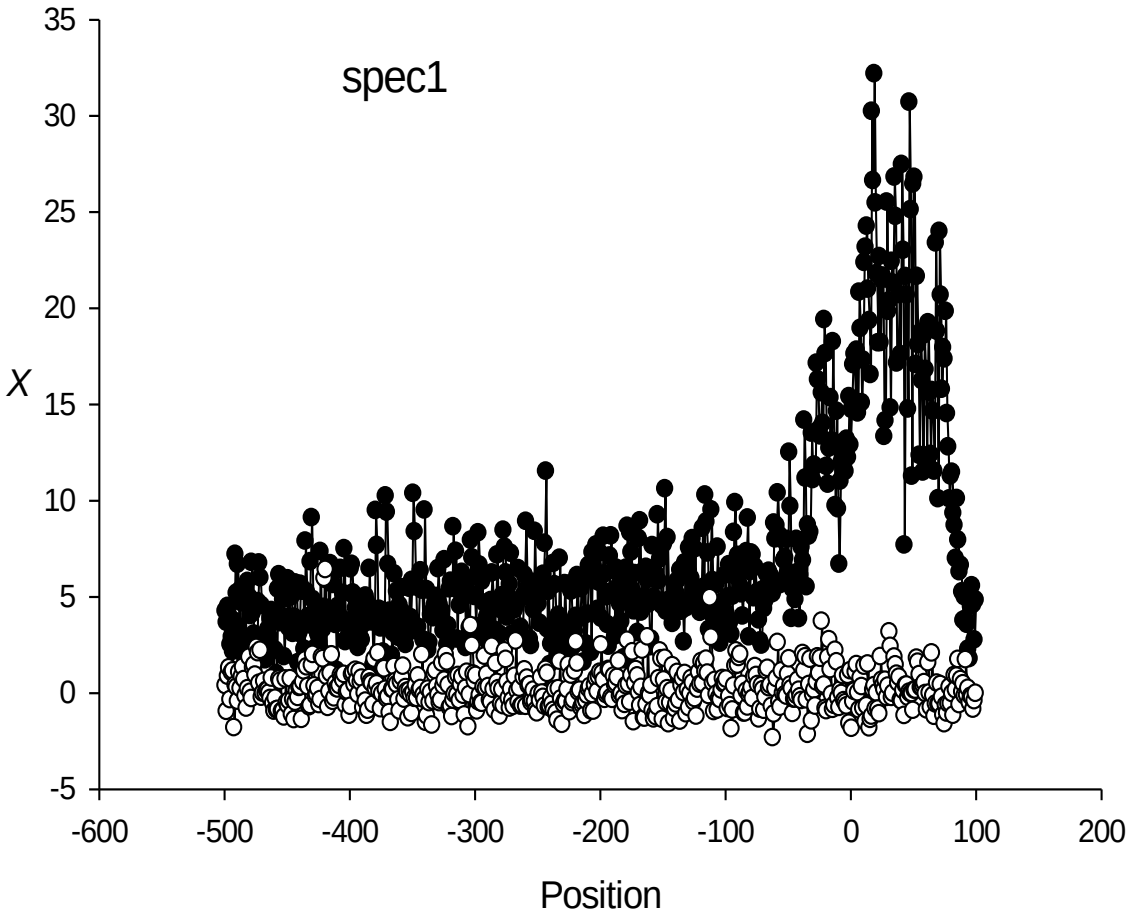
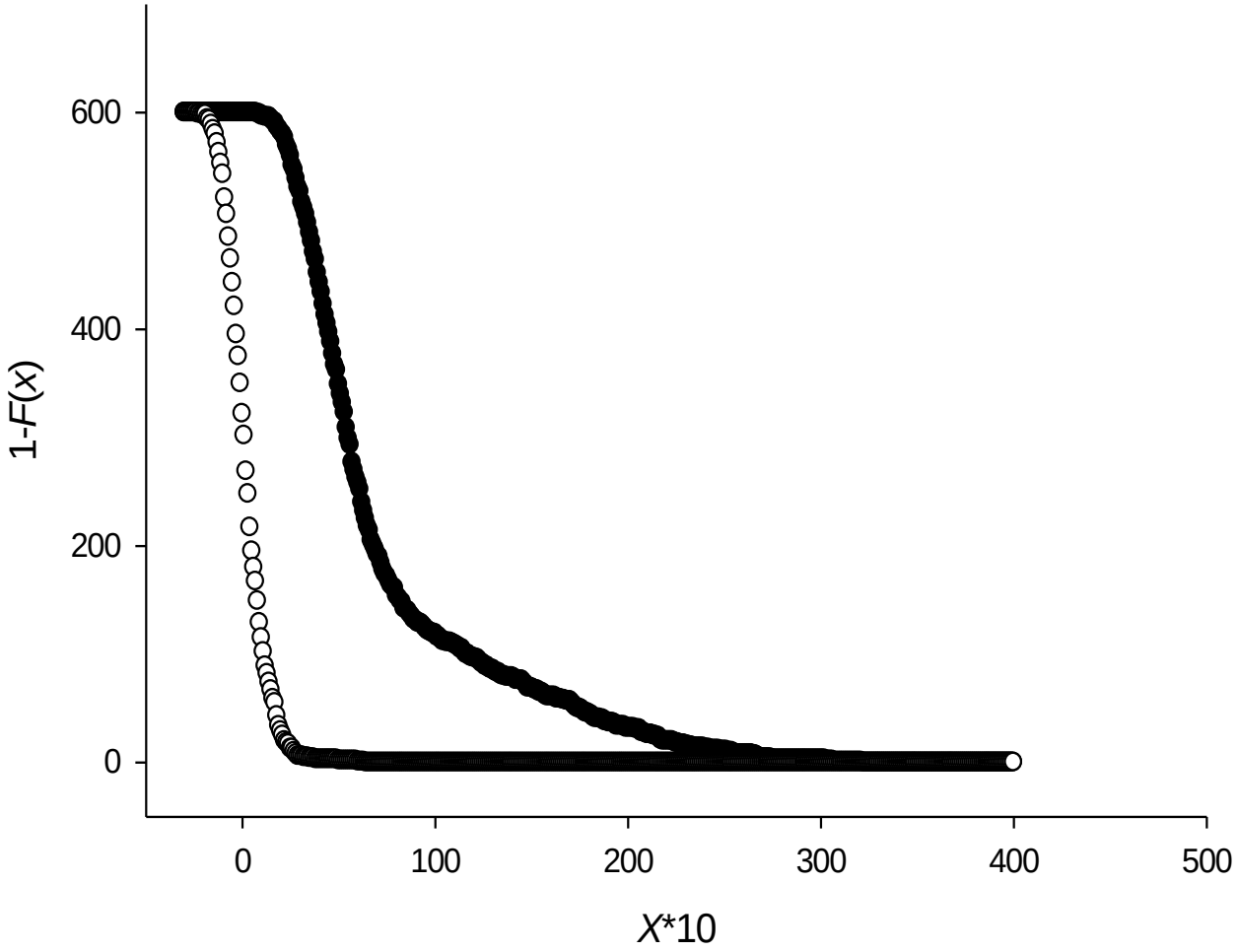


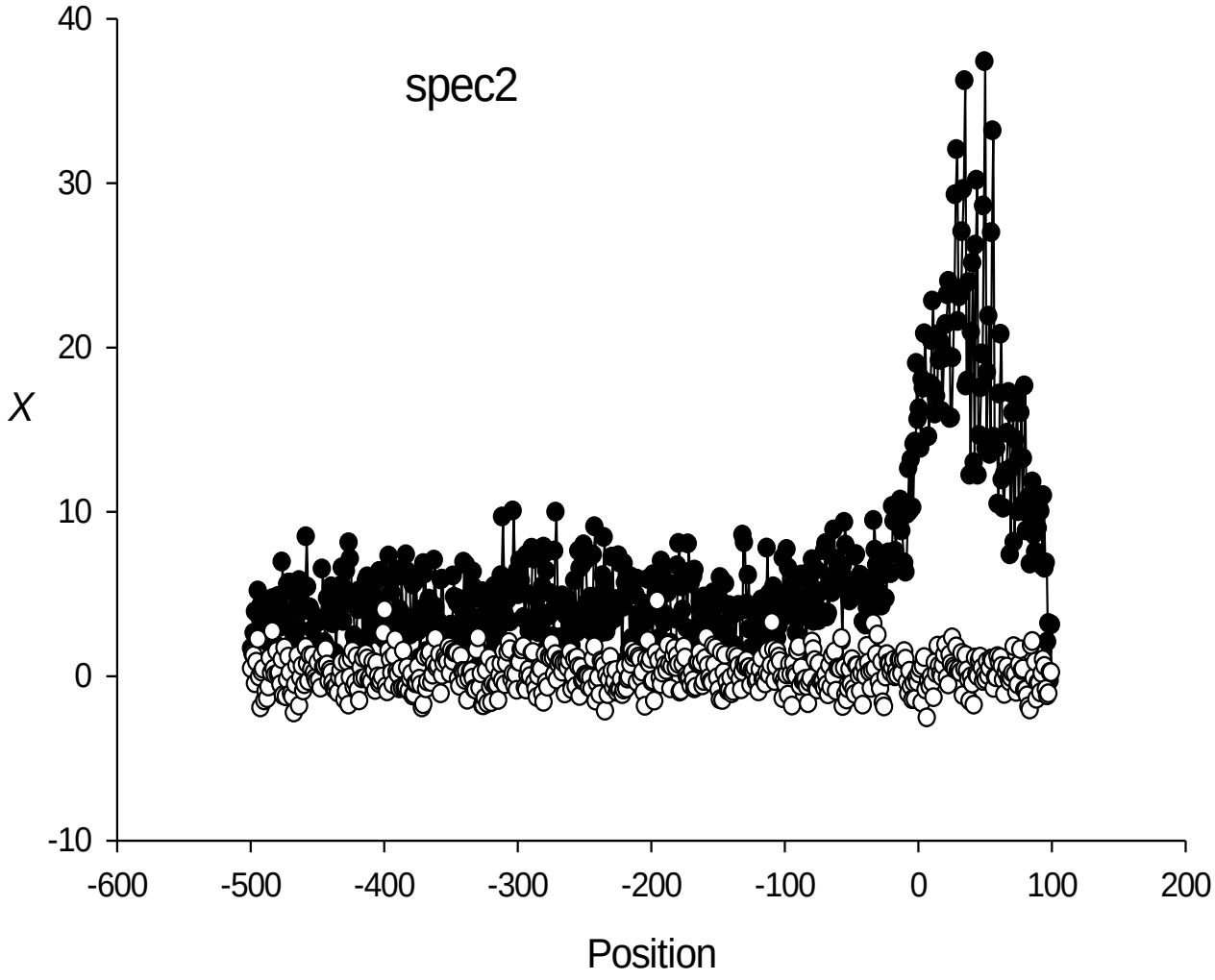
Fig. 1. Simplified model of transcriptional regulation of protein-encoding genes.

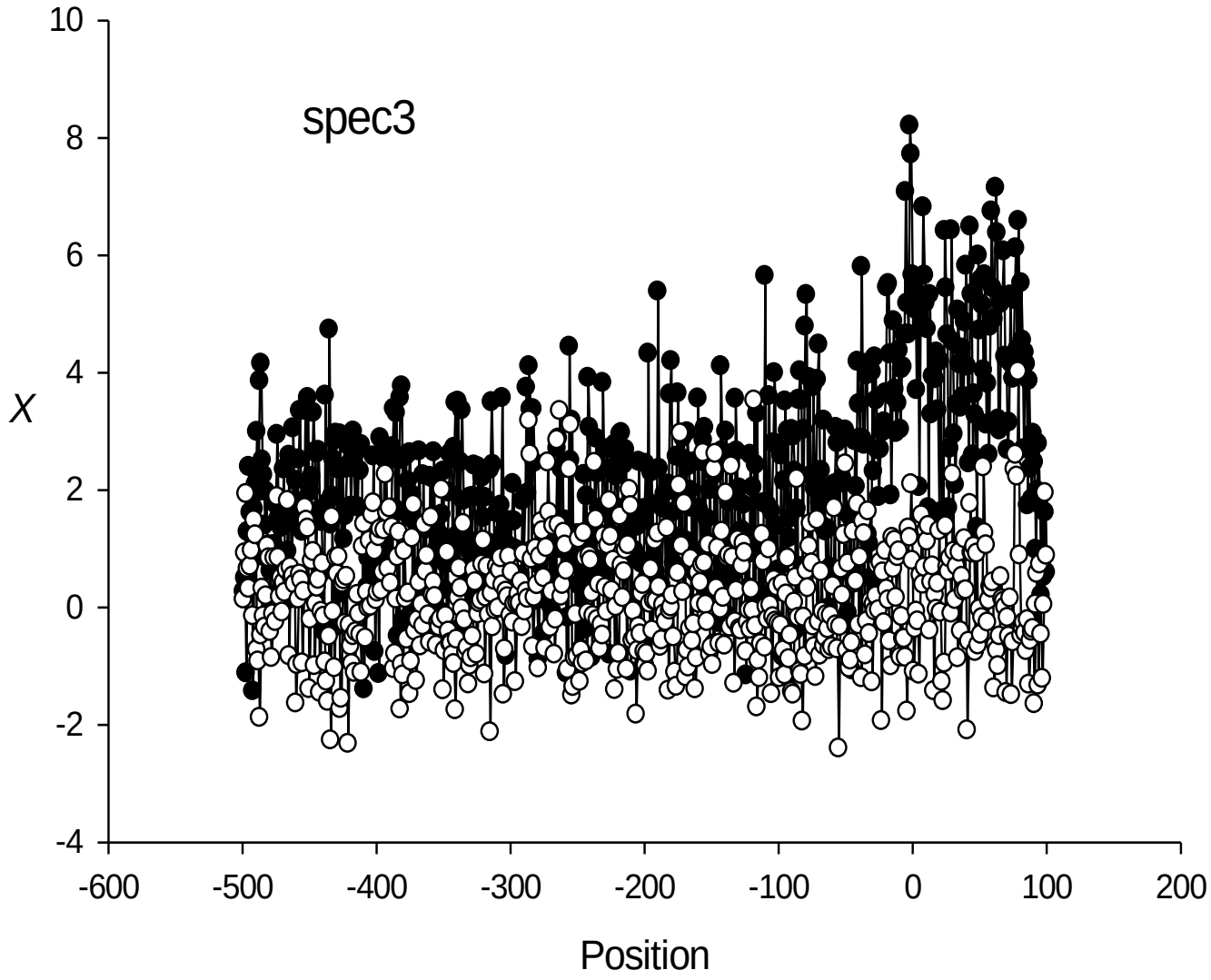

1. Q set of random position weight matrixes (PWM)
2. Comparison of the Q set with the S sequence
3. Application of the genetic algorithm to Q set
4. Comparison of the Q set with the S sequence and selection of of PWM with maximum of F(L,L)
5. Has mF(L,L) increased?
no
yes
Stop

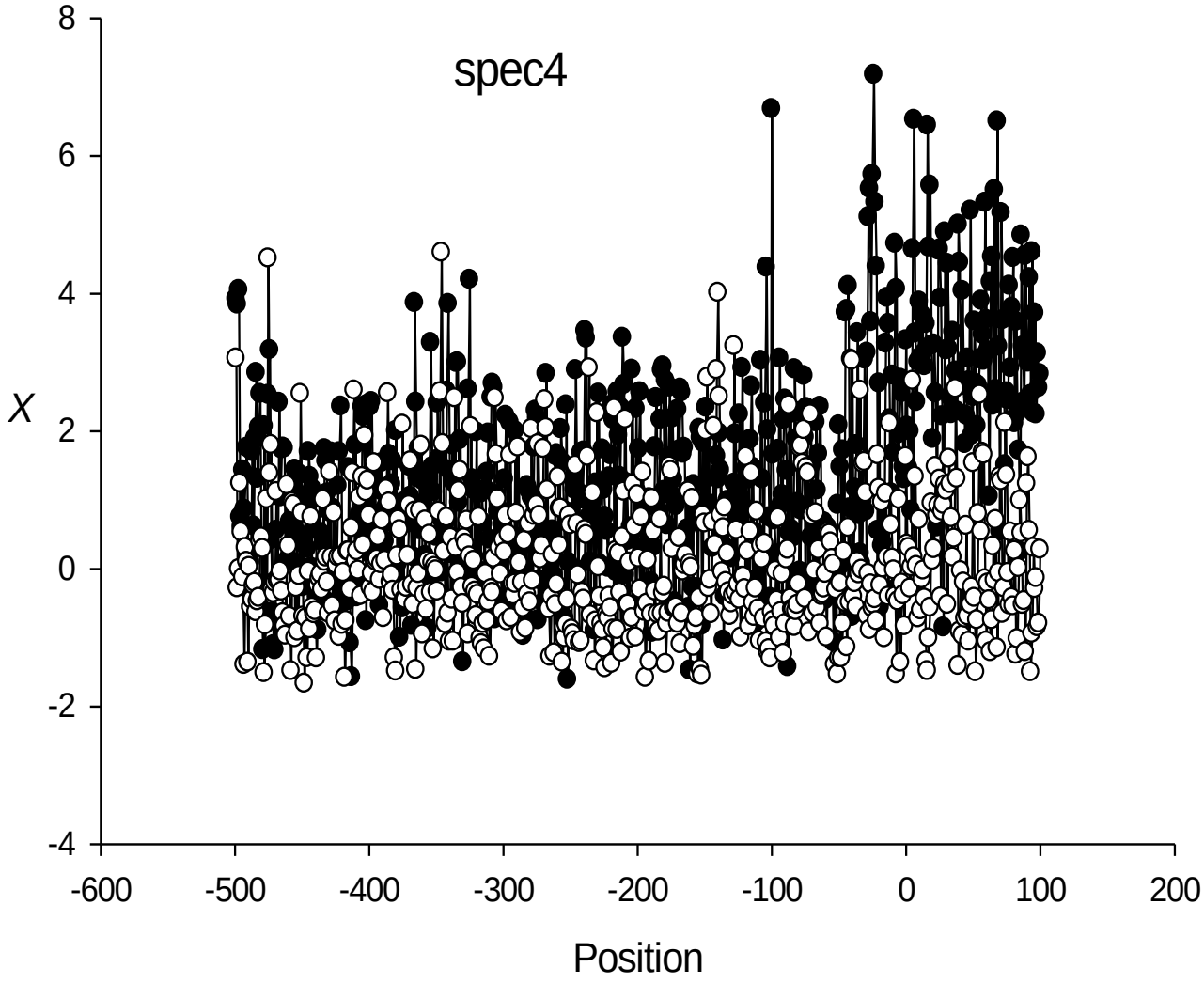
N	Volume of promoter group	Volume of random group
1	7247	32
2	1773	28
3	633	21
4	201	24
5	279	18
6	182	24
7	152	28
8	152	33
9	141	30
10	138	23
11	138	35
12	103	18
13	85	20
14	124	22
15	95	25
16	77	26
17	95	21
18	118	23
19	44	29
20	90	31
21	81	30
22	23	25
23	54	32
24	81	24
25	24	18

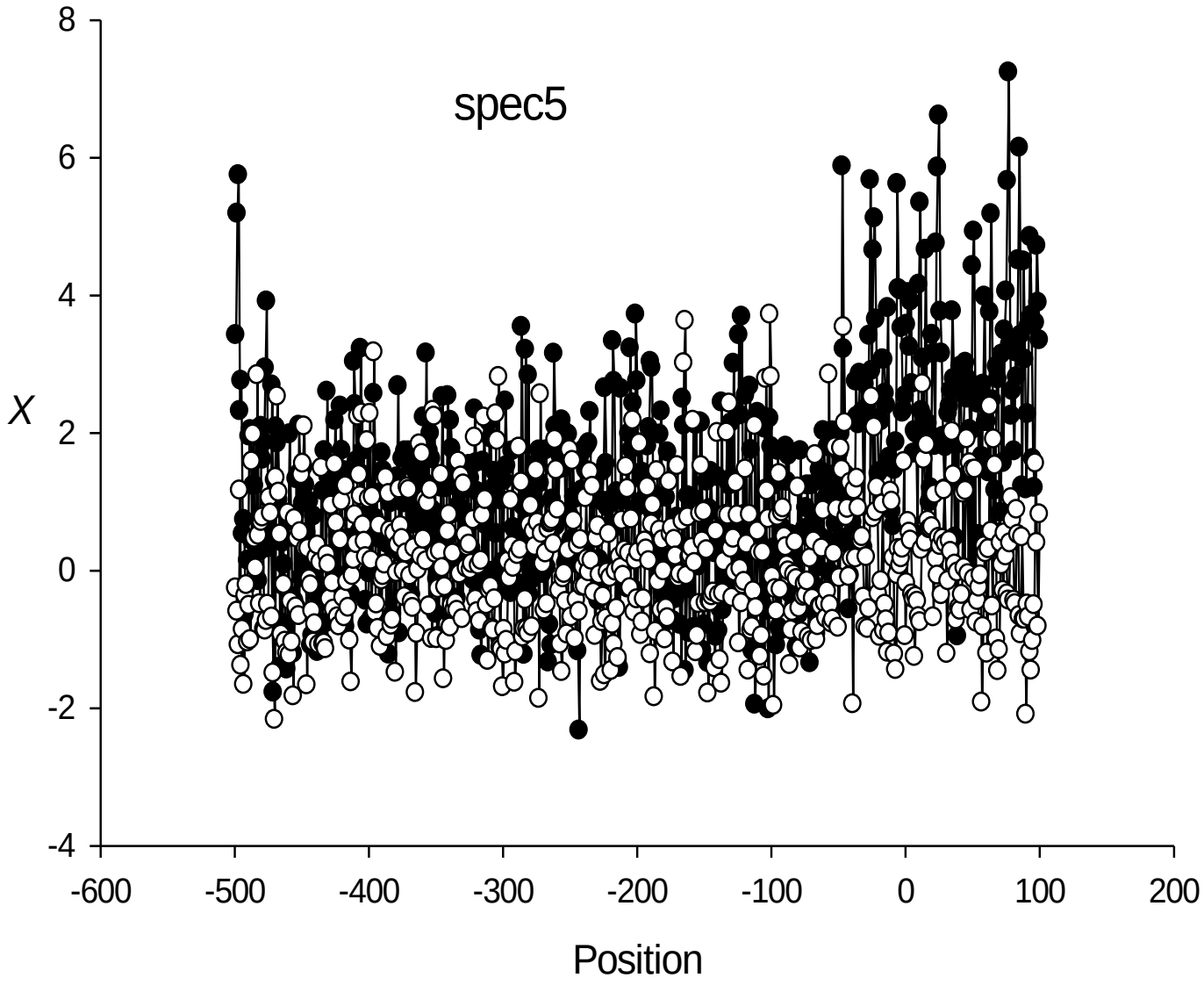

spec1
X
35
30
25
20
15
10
5
0
-5
-600
-500
-400
-300
-200
-100
0
100
200
Position

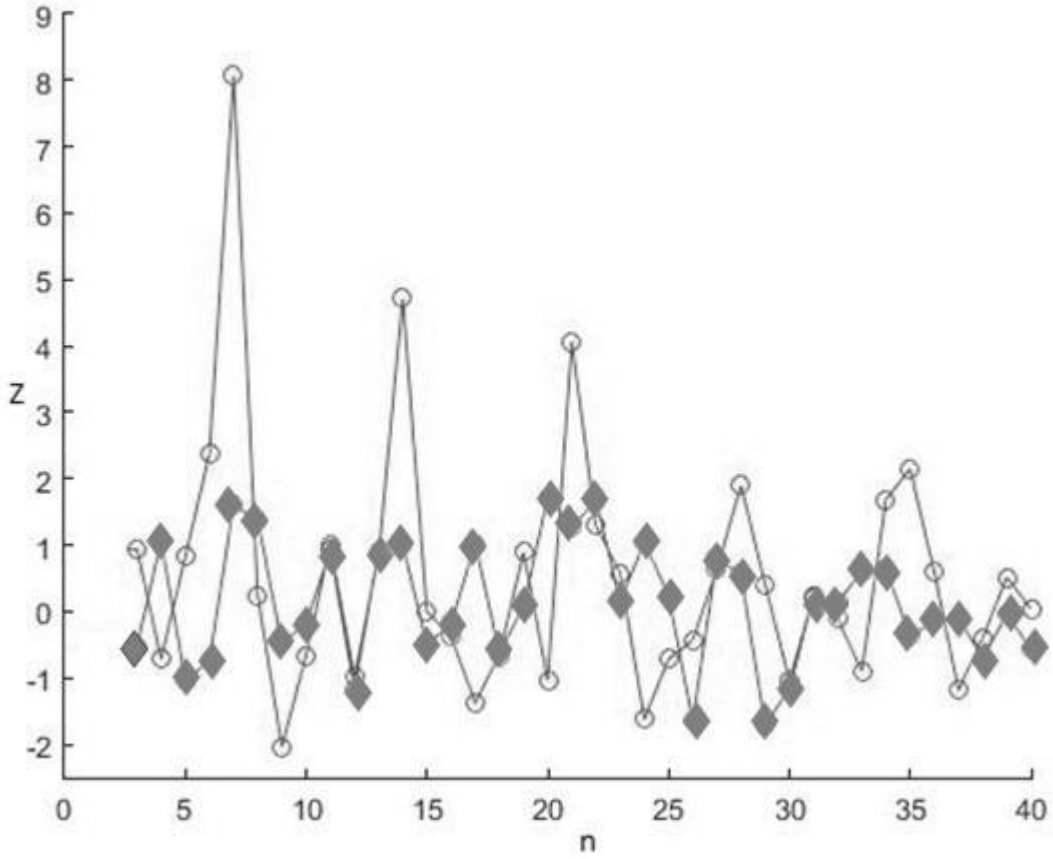

600
400
200
0
1-F(x)
0
100
200
300
400
500
X*10


spec2
40
30
20
10
0
-10
X
-600
-500
-400
-300
-200
-100
0
100
200
Position

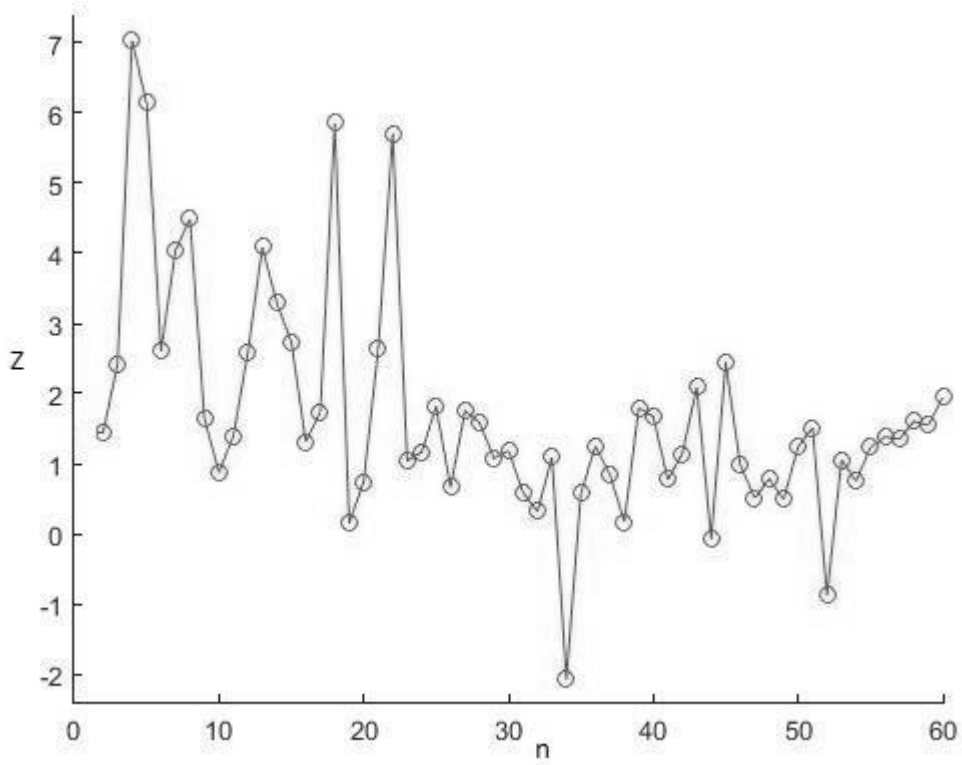

spec3
X
Position
10
8
6
4
2
0
-2
-4
-600
-500
-400
-300
-200
-100
0
100
200


spec4
X
8
6
4
2
0
-2
-4
-600
-500
-400
-300
-200
-100
0
100
200
Position

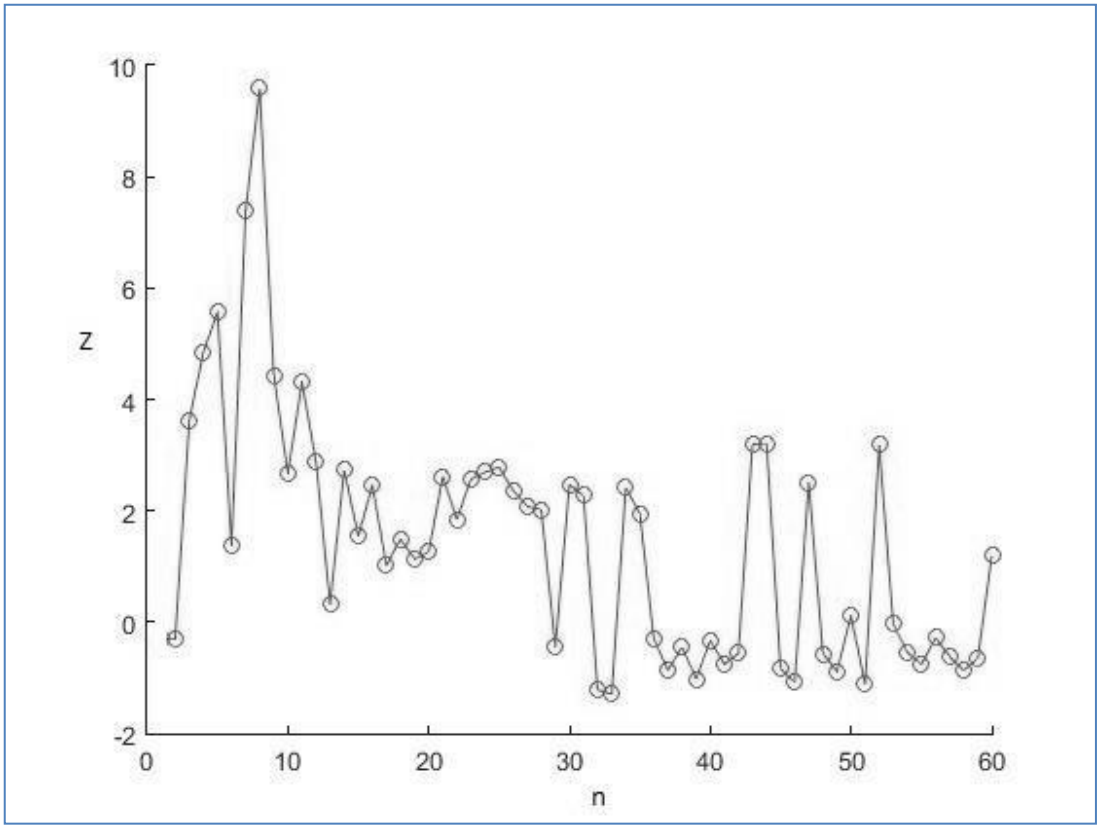

spec5
X
8
6
4
2
0
-2
-4
-600
-500
-400
-300
-200
-100
0
100
200
Position



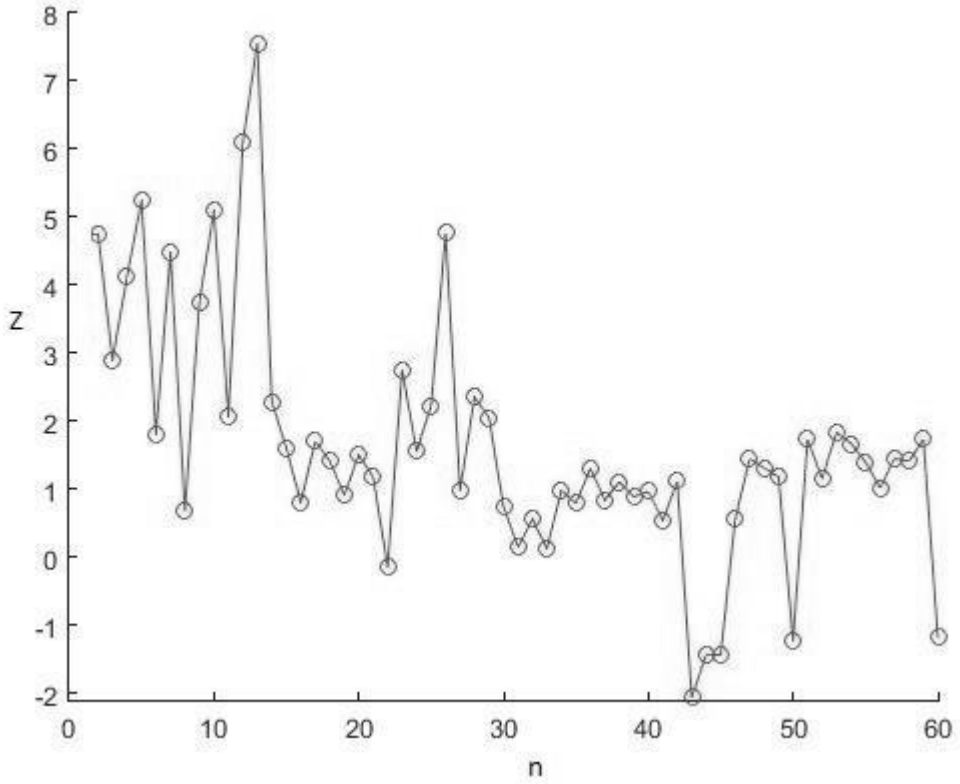
Graph *Z(n) for an artificial sequence containing* 60% random base substitutions with the addition of 8 inserts and 12 deletions (circles) and *Z(n) for a random* sequence (rhombus).



Graph Z (n) for the poem A.S. Pushkin "I remember a wonderful moment...".



Graph *Z(n) for the William Blake's poem "Spring".*



Graph Z (n) for the poem by Robert Frost "Fire and ice".

Lengths of fuzzy periods found in 95 works of poetry of different authors.

Author	Number of analyzed works of poetry	Lengths of fuzzyperiods found in the works of poetry
Pushkin A.	15	2,4,7,10
YeseninS.	10	2,6,8,11
Blok A.	5	3,5,10,11
Tutchev F.	5	2,7,8,15
Fet A.	5	2,4,6,9
Mayakovsky V.	7	2,3,5,8
Shakespeare W.	5	10,15,18,37
Byron D.	5	8,16,28
Frost P.	20	5,8,13,15,20
Blake W.	18	4,6,8,10,11,23