IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/rjes

Estimating the Order of Markov Chains Using Shannon Criterion and Akaike Criterion

Alaa Abdallah Mahmoed and Zinah Mudher Albazzaz

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

ARTICLE INFO

Article history:

Received	5 November 2025
Revised	5 November 2025
Accepted	12 January 2026
Available online	16 January 2026

Keywords:

Markov Chain
Order Estimation
Prediction
Markov Processes

ABSTRACT

Determining the order of a Markov chain is un important problem, as the correct decision to be made will be based on the specific order of that model. This research involves a Markov chain analysis of the gene sequence for the breast cancer-causing gene (BRCA1) and estimating its order using Shannon's Information Criterion and Akaike's Information Criteria. Selecting the appropriate order in modeling contributes to enhancing the understanding of the behavior of genes associated with breast cancer and enables early detection and the development of effective treatment strategies. The two aforementioned criteria were used, and the results showed that the order of this gene is the Second order using Shannon's criterion, and the Second -order model is the most accurate using Akaike's criteria

1. Introduction

In recent times, there has been growing interest in the study of systems that change randomly over time, which cannot be fully controlled or their future behavior predicted with certainty. These are known as stochastic processes. This interest stems from the fact that many phenomena exhibit non-deterministic behavior, reliant on chance. One of the most important stochastic processes is the Markov process, which holds a significant place in practical statistical applications. It has been used in various scientific fields such as medicine, psychology, physics, computer science, and others. The focus is often on cases where the index set, according to which the stochastic process changes, is of a discrete nature; at this point, the process is called a Markov chain.

Estimation problems concerning the order of a Markov chain are among the important issues, as the correct decision to be made will be based on the specific order determined for that model. This research involves the analysis of a Markov chain and the estimation of its order using statistical methods, including the Shannon Information Criterion and the Akaike Information Criterion.

Within the framework of our research, the order can be defined as the number of previous values upon which the probability of the next state depends. It represents the length of the historical period or the size of the

Corresponding author: drosamahannon@uomosul.edu.iq

https://doi.org/10.62933/kev7nx24

memory upon which the probability of the next state relies. A first-order Markov chain is one that possesses a memory of size one. It is a chain whose next state depends only on the one immediately preceding state and is independent of the more distant past. It can be expressed as follows :-

$P\{X_{\_n} = i \mid X_{n-1} = j, X_{n-2} = j_1, \ldots\} = P\{X_n = i \mid X_{n-1} = j = P_{(i,j)}$

The common (conventional) Markov chains are first-order chains. As for a second-order Markov chain, it is a chain whose outcome probabilities depend only on the two immediately preceding states. It is expressed as follows :-

$P\{X_n = i \mid X_{n-1} = j, X_{n-2} = j1, X_{n-3} = j_2, \ldots\} = P\{X_n = i \mid X_{n-1}\} = j, X_{n-2} = j_1\}$

If the probability of the next state depends on r immediately preceding states, then the chain is called a Markov chain of order r . Conversely, a zero-order Markov chain represents a sequence of independent events, possessing no prior memory or history to depend upon.

Tong , 1975]

2. The Theoretical Aspect

2.1 Information theory is one of the modern fields derived from probability theory, and it has proven its importance through its wide applications in the field of communication. Claude Shannon is considered the pioneer in this field, as he laid the first foundations for measuring the quantity of information using statistical methods. These foundations were later termed Shannon's measure of information.

This was followed by numerous studies that expanded the horizons of this theory, pointing to its potential applications and implementation in various fields such as mathematics, psychology, economics, biology, and others.

The Shannon information criterion is one of the methods for estimating the order of a Markov chain by using conditional uncertainty, also known as the entropy function (Entropy). The Shannon criterion is denoted by the symbol H and is a measure of the conditional uncertainty contained in the random variable X; therefore, it is a value and not a probability.

The average amount of information for an experiment with K possible states and successive probabilities is :

[Lindiey ,1956]

$H = E(-\log_2 p) = -\sum p_j \log_2 p_j$

Its properties include :

1- Possesses the property of additivity

Is a positive function 2-

Is a monotonically decreasing function 3-

Is continuous and defined 4-

[Finesso ,1991]

Algorithm for Shannon's Information Criterion for Estimating Markov Chain Order :

[Alamin ,2009]

First: Testing for Zero Order :-

1- Formulate the hypotheses

$H_0: k = 0$

$H_1: k \geq 1$

2- Calculate $\hat{H}_1$ using the following formula :-

$\hat{H}_1 = \log_2 N_2 - (1/N_2) [\sum_j n_j \log_2 nj]$

3- Calculate $\hat{H}_2$ using the following formula:-

$\hat{H}_2 = (1/N_2) [\sum_i n_i \log_2 n_i - \sum_{i,j} n_{ij} \log_2 n_{ij}]$

4- Find the value of $\hat{T}_1$ using the following formula:-

$\hat{T}_1 = \hat{H}_1 - \hat{H}_2$

5- Find the value of Z_1 using the following formula

$Z_1 = 2 (\log_e 2) N_2 \hat{T}_1$

6- Compare the value of Z_1 with the tabulated value of χ^2 with degrees of freedom (c-1) and a specific significance level, where c is the number of states in the chain.

7- If $Z_1 < \chi^2$ tab(α, c-1), then the null hypothesis is accepted, meaning the chain is of zero order (independent).

8- If χ^2tab(α, c-1) < Z_1, then the null hypothesis is rejected, meaning the chain is of first order or higher.

Second: Testing for First Order (The hypotheses become as follows)

1- Formulate the hypotheses

H_0: k = 1

H_1: k ≥ 2

2- Calculate $\hat{H}$(pairs) using the following formula

$$\hat{H}(pairs) = \log_2 N_3 - (1/N_3) [\sum_{i,j} n_{ij} \log_2 n_{ij}]$$

3- Calculate $\hat{H}$(single) using the following formula

$$\hat{H}(single) = \log_2 N_3 - (1/N_3) [\sum_j n_{j.} \log_2 n_{j.}]$$

4- Calculate $\hat{H}_2$ using the following formula

$$\hat{H}_2 = \hat{H}(single) - \hat{H}(pairs)$$

5- Calculate $\hat{H}$(Triplets) using the following formula

$$\hat{H}(Triplets) = \log_2 N_3 - (1/N_3) [\sum_{i,j,k} n_{ijk} \log_2 n_{ijk}]$$

6- Calculate $\hat{H}_3$ using the following formula

$$\hat{H}_3 = (1/N_3) [\sum_{j,k} n_{jk} \log_2 n_{jk} - \sum_{i,j,k} n_{ijk} \log_2 n_{ijk}]$$

7- Find the value of $\hat{T}_2$ using the following formula

$$\hat{T}_2 = \hat{H}_2 - \hat{H}_3$$

8- Find the value of Z_2 using the following formula

$$Z_2 = 2 (\log\_e\ 2) N_3 \hat{T}_2$$

9- Compare the value of Z_2 with the tabulated χ^2 value with degrees of freedom $(c-1)^2$ and a specific significance level. If $Z_2 < \chi^2$_tab(α, $(c-1)^2$) then the null hypothesis is accepted, meaning the chain is of first order

10- If χ^2_tab(α, $(c-1)^2$) < Z_2 then the null hypothesis is rejected, meaning the chain is of second order or higher.

This algorithm can be generalized to test for order (m) as follows :

Third: Testing for Order (m) :

Formulate the hypotheses 1-

H_0: k = m - 1

H_1: k ≥ m

2- Calculate $\hat{H}$(m folds) using the following formula:

$$\hat{H}(m\ folds) = \log_2 N_{m+1} - (1/N_{m+1}) [\sum_{i,j,\ldots,m} n_{ij\ldots m} \log_2 n_{ij\ldots m}]$$

3-Calculate $\hat{H}$(Triplets) using the following formula:

$$\hat{H}(Triplets) = \log_2 N_{m+1} - (1/N_{m+1}) [\sum_{j,k,\ldots,m} n_{jk\ldots m} \log_2 n_{jk\ldots m}]$$

4- Calculate $\hat{H}_m$ using the following formula

$$\hat{H}_m = \hat{H}_{(m\ folds)} - \hat{H}_{(Triplets)}$$

5-Calculate $\hat{H}_{(m+1\ folds)}$ using the following formula

$$\hat{H}_{(m+1\ folds)} = \log_2 N_{m+1} - (1/N_{m+1}) [\sum_{i,j,\ldots,m+1} n_{ij\ldots m+1} \log_2 n_{ij\ldots m+1}]$$

Where $\hat{H}_{(m+1\ folds)}$ is the conditional uncertainty given knowledge of the previous m outcomes.

6-Calculate $\hat{H}_{m+1}$ using the following formula:

$$\hat{H}_{m+1} = (1/N_{m+1}) [\sum\_{j,k,\ldots,m+1} n_{jk\ldots m+1} \log_2 n_{jk\ldots m+1} - \sum_{i,j,\ldots,m+1} n_{ij\ldots m+1} \log_2 n_{ij\ldots m+1}]$$

7-Find the value of $\hat{T}_m$ using the following formula

$$\hat{T}_m = \hat{H}_m - \hat{H}_{m+1}$$

8-Find the value of Z_m using the following formula

$$Z_m = 2 (\log_e 2) N_{m+1} \hat{T}_m$$

9-Compare the value of Z_m with the tabulated χ^2 value with degrees of freedom $c^{m-1}(c-1)^2$. If $Z_m < \chi^2_{tab}(\alpha, c^{m-1}(c-1)^2)$, then the null hypothesis is accepted, meaning the chain is of order m-1.

10-If $\chi^2_{tab}(\alpha, c^{m-1}(c-1)^2) < Z\_m$ then the null hypothesis is rejected, meaning the chain is of order m or higher .

Akaike's Information Criterion (AIC) and Its Application Step

Akaike's Information Criterion (AIC) is considered one of the most important statistical tools used for comparing models and selecting the most appropriate one. This criterion strikes a balance between the model's goodness of fit and the number of parameters used within it. In other words, it seeks to achieve an equilibrium between estimation accuracy and model simplicity. The core premise is that the model which explains the data well using the fewest number of parameters is deemed the most efficient.

The general formula for Akaike's Information Criterion is :

$AIC = 2k - 2\ln(L)$

Where

P : The number of estimated parameters in the model .

L : The maximum value of the likelihood function

Steps for Applying Akaike's Information Criterion

1- Specify the Models This involves formulating a set of potential models to explain the data .

2- Estimate the Parameters: Use the Maximum Likelihood method to estimate the parameters for each model .

3- Calculate the AIC value for each model using the following formula

$AIC = 2k - 2\ln(L)$

4- Compare the Models The model with the lowest AIC value is considered the most suitable.

5- Interpret the Results Preference is given to the model that combines structural simplicity with the ability to accurately explain the data, while avoiding the excessive use of variables.

It is crucial to note that Akaike's Information Criterion does not provide an "absolute quality" measure for a model; rather, it is used exclusively for comparing models applied to the same dataset. [Akaike , 1974]

3.The Application

3.1 Shannon Information

The order of the nucleotide sequence for the breast cancer gene (BRCA1), consisting of 1416 bases, was estimated. This serves as a practical application of the theoretical aspect mentioned earlier, using Shannon's information criterion and Akaike's criterion. These sequences were obtained from the National Center for Biotechnology Information (NCBI).

1- Testing the Order Using Shannon's Information Criterion

To test the independence of this sequence, the following hypothesis was tested:

H_0: K=0

H_1: K≥1

The transition matrix for pairs and their marginal totals was formed, as shown in Table (1)

Table (1): Frequency Transition Matrix for Pairs of Nitrogenous Bases and Their Marginal Totals

j \ I	A	C	G	T	ni .
A	73	14	124	64	275
C	25	7	3	47	82
G	129	27	99	182	437
T	47	34	212	328	621
n.j	274	82	438	621	1415

The sum of all possible values for these pairs represents the state space, which is

State Space = {AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}

From step (2) first of the algorithm, the value is:

$\hat{H}_1 = 1.74188393$

From step (3), the value is:

$\hat{H}_2 = 1.64137219$

From step (4), we obtain the value of $\hat{T}$

$\hat{T} = 0.10051174$

From step (5), the value of Z_1 is:

$Z_1 = 197.164489$

Where the value of Z_1 is compared with the tabulated value χ^2_tab with degrees of freedom $(C-1)^2$ and a significance level of $\alpha = 0.05$, which equals:

$\chi^2_{tab}(0.05, 9) = 16.919$

Given that C, the number of states in the sequence, is four (A, C, G, T), and since the calculated Z_1 value is greater than χ^2_{tab} (0.05, 9), the null hypothesis is rejected. This means the sequence is not independent; it is of first order or higher. The hypothesis then becomes :

H_0: K = 1

H_1: K ≥ 2

The frequency transition matrix for triplets of nitrogenous bases (nijk) was formed, as shown in Table(2)

Table (2): Frequency Transition Matrix for Triplets of Nitrogenous Bases (nijk)

i	A				C				G				T			
j \ K	T	G	C	A	T	G	C	A	T	G	C	A	A	G	T	C
A	17	4	8	24	7	0	1	5	62	30	11	21	38	16	6	4
C	3	17	3	2	2	0	3	2	1	1	1	0	27	14	2	4
G	28	26	2	37	16	3	1	7	30	11	4	54	98	60	7	17
T	15	21	1	10	21	0	2	11	89	57	11	54	146	122	19	22

Their marginal totals are:

n.j.				n.jk				
nij.								
47	255	53	129	14	124	64	73	25
82								
129	27	99	18	104	3	99	212	
437								

47 34 211 32 63 47 182 327 619

The state space represents all possible values for the triplets

State Space = {AAA, AAC, AAG, AAT, ACA, ACC, ACG, ACT, AGA, AGC, AGG, AGT, ... TTT}

From step (2) secondly of the algorithm, the value is:

$\hat{H}$(Pairs) = 3.38418552

From step (3), the value is:

$\hat{H}$(Single) = 1.74191855

From step (4), the value is:

$\hat{H}_2$ = 1.64137219

From step (5), the value was found:

$\hat{H}$(Triplets) = 4.97808317

From step (6), the value $\hat{H}_3$ was calculated:

$\hat{H}_3$ = 1.59288079

From step (7), the value $\hat{T}_2$ is calculated:

$\hat{T}_2$ = 0.0484914

From step (8), the value Z_2 is calculated:

Z_2 = 95.1919125

Then the value Z_2 is compared with the tabulated value χ^2_{tab} with degrees of freedom $C^{\wedge}(K-1)(C-1)^2$ and a significance level of α = 0.05, which equals

χ^2_{tab} (0.05, 36) = 50.998

Since the calculated Z_2 value is greater than χ^2_{tab}, the null hypothesis is rejected and the alternative hypothesis is accepted, meaning the sequence is of second order or higher. The hypothesis becomes:

H_0: K = 2

H_1: K ≥ 3

To perform this test, the frequency transition matrix for quadruplets (nijkl) was determined, as in Table(3)

Table (3): Frequency Transition Matrix for Quadruplets of Nitrogenous Bases (nijkl)

Kl / IJ	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	5	5	6	8	1	0	0	7	8	3	5	8	0	3	6	8
AC	1	1	3	0	0	1	0	0	0	0	0	0	0	1	5	2
AG	4	1	10	6	3	1	1	6	17	2	3	8	6	2	18	36
AT	0	0	4	0	1	1	0	4	6	3	3	4	4	5	4	25
CA	1	0	1	0	1	1	0	1	5	2	1	9	0	0	0	3
CC	1	0	1	0	1	1	0	1	0	0	0	0	0	0	1	1
CG	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
CT	0	0	3	1	0	0	0	2	5	5	3	1	3	3	8	13
GA	13	2	15	7	2	0	0	0	4	3	19	36	4	1	5	18
GC	0	1	6	0	0	0	0	1	0	1	1	1	1	0	4	11
GG	17	1	25	11	2	0	1	1	8	0	0	3	2	0	10	18
GT	3	0	5	9	3	0	0	4	17	1	9	33	2	3	31	62
TA	5	1	2	2	1	0	0	0	4	3	5	9	0	2	4	9
TC	0	1	7	3	1	1	0	0	0	0	0	0	3	1	4	13

TG	16	0	27	11	2	0	0	9	28	2	8	19	9	4	32	44
TT	7	1	9	5	7	1	0	11	26	2	42	51	13	8	79	65

It is noted from Table (3) that it contains frequencies less than (5), and their percentage exceeded the allowable percentage of 20%. Therefore, the merging method was used to solve this problem. Table (3) shows the frequency transition matrix for quadruplets of nitrogenous bases nijkl after merging .

Frequency Transition Matrix for Quadruplets of Nitrogenous Bases nijkl after Mergi

Kl / ij	A+C	G+T	A+C	G+T	A+C	G+T	A+C	G+T
AA+Ac	12	17	2	7	11	13	4	21
AG+AT	5	20	6	11	28	16	17	83
CA+CC	2	2	4	2	7	10	0	5
CG+cT	0	4	0	3	11	4	7	21
GA+Gc	16	28	2	0	8	57	6	38
GG+GT	21	50	5	6	26	45	7	121
TA+TC	7	14	3	0	7	14	6	30
TG+TT	24	52	10	20	58	120	33	220

The marginal totals for this matrix are

n.jk

78	49	3	31
186	144	42	87
3	22	10	13
50	43	7	20
155	109	46	65
281	199	82	134
81	57	24	39
510	403	106	250

n.jkl

37	2
66	6
7	4
6	5
34	18
102	14
13	7
159	26

nijk.

17
37
8
18
39
29
21
104

The state space represents the set of all possible values for the quadruplets:

State Space = {AAAA, AAAC, AAAG, AAAT, AACA, AACC, AACG, AACT, ... TTTT}

From step (2) thirdly of the algorithm, the value is:

$\hat{H}$(Triplets) = 4.9780831

From step (3), the value is calculated:

$\hat{H}$(Pairs) = 1.87668070

The difference between the two previous values is H_3:

H_3 = 3.1014024

From step (5), the value is calculated:

$\hat{H}$(Fourfolds) = 5.1502172

From step (6), the value $\hat{H}_4$ is calculated:

$\hat{H}_4$ = 1.4849822495287

From step (7), the value $\hat{T}_3$ is found:

$\hat{T}_3$ = 1.6164202

From step (8), the value $\hat{Z}_3$ is calculated:

$\hat{Z}_3$ = 6.94654802

Then the calculated value $\hat{Z}_3$ is compared with the tabulated value χ^2_tab at degrees of freedom C^(K-1)(C-1)² and a significance level of $\alpha = 0.05$, which equals:

χ^2_{tab} (0.05, 144) = 171.785

Since the calculated value $\hat{Z}_3$ is less than χ^2_{tab}, the null hypothesis is accepted, stating that the sequence of this gene is of second order. Table (4) shows a summary of the results of estimating the Markov chain order for the breast cancer gene using Shannon's information criterion.

Table (4): Shows the results of estimating the Markov chain order for the breast cancer gene using Shannon's information criterion.

hypothesis	Zi	d. f	$\chi^2(\text{tab}(\alpha, \text{df})$	decision
$H_0: k = 0$ $H_1: k \geq 1$	$Z1 = 197.164489$	9	$\chi^2_{\text{tab}(0.05,9)} = 16.919$	H_0 *Refuse* $k \geq 1$
$H_0: k = 1$ $H_1: k \geq 2$	$Z2 = 95.1919125$	36	$\chi^2_{\text{tab}(0.05,36)} = 50.998$	H_0 *Refuse* $k \geq 2$
$H_0: k = 2$ $H_1: k \geq 3$	$Z3 = 6.94654802$	144	$\chi^2_{\text{tab}\ (0.05,144)} = 171.785$	H_0 Accept $k = 2$

3.2 Rank Test Using the Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a statistical tool used to select the most appropriate model from among several competing models. It is based on the principle of balancing fit quality and model simplicity, thereby avoiding both over-complexity and excessive simplification (Akaike, 1974). In the context of our research, this criterion was used to test the accuracy of the model and select the suitable model for the gene sequence causing breast cancer.

First-Order Test

The transition frequency matrix for nitrogenous base pairs (BRCA1):

	A	C	G	T	Ni.
A	73	14	124	64	275
C	25	7	3	47	82
G	129	27	99	182	437
T	47	34	212	328	621
n.j	274	82	438	621	1415

After calculating the frequency matrix, we proceed to calculate the probability transition matrix:

	T	G	C	A
A	0.232	0.450	0.050	0.265
C	0.573	0.036	0.085	0.304
G	0.416	0.226	0.061	0.295
T	0.528	0.341	0.054	0.075

After calculating the probabilities and using the AIC formula:

$AIC = 2p - 2\ln(L)$

Where p is the number of parameters, calculated as $p = N^K(N-1)$

N: is the number of states

K: is the order (rank) being tested

L: is the maximum likelihood value, found using the formula.

$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} [P(X_t = j \mid X_{t-1} = i)]^{nij}$

After calculating the value of ln(L) using the above formula, it was found to be:

$\ln(L) = \sum nij \; Ln \; Pij$

ln(L) = -2862.20

$p = N^K (N-1) = 4^1 * (4-1) = 4 * 3 = 12$

AIC = 2P -2ln (L)

AIC = 5736.41

This is the AIC value for the first-order test.

Second-Order Test

The triple frequency matrix for nitrogenous base pairs (BRCA1)

	A	C	G	T	ni
AA	24	8	24	17	73
AC	5	1	0	8	14
AG	21	11	30	62	124
AT	4	6	16	38	64
CA	2	3	17	3	25
CC	2	3	0	2	7
CG	0	1	1	1	3
CT	4	2	14	27	47
GA	37	2	62	28	129
GC	7	1	3	16	27
GG	54	4	11	30	99
GT	17	7	60	28	112
TA	10	1	21	15	47
TC	11	2	0	21	34
TG	54	11	57	89	211
TT	22	19	122	164	327

After calculating the expected frequency matrix, we find the probability matrix

Probability Matrix for Nitrogenous Base Pairs

	A	C	G	T
AA	0.328	0.109	0.328	0.232
AC	0.357	0.071	0	0.571
AG	0.169	0.088	0.241	0.5
AT	0.062	0.093	0.25	0.593
CA	0.08	0.12	0.68	0.12
CC	0.285	0.428	0	0.285
CG	0	0.333	0.333	0.333

CT	0.085	0.042	0.297	0.574
GA	0.286	0.015	0.480	0.217
GC	0.259	0.037	0.111	0.592
GG	0.242	0.040	0.111	0.303
GT	0.151	0.062	0.535	0.25
TA	0.212	0.021	0.446	0.319
TC	0.323	0.058	0	0.617
TG	0.255	0.052	0.270	0.421
TT	0.067	0.058	0.373	0.501

We calculate the value of ln(L) as illustrated above and according to the previous steps. It was found that the value of ln(L) is

$\ln(L) = \sum nij\ Ln\ Pij$

ln(L) = -1589.02

We find the value of p for the second order according to the equation :

$p = N^K (N-1) = 4^2 * (4-1) = 16 * 3 = 48$

2p = 2 * 48 = 96

AIC = 2p - 2 ln(L)

AIC =3274.05

This is the AIC value for the second-order test .

Third-Order Test

The transition matrix for nitrogenous base quadruplets (BRCA1)

	A	C	G	T	.n.j
AAA	5	5	6	8	24
AAC	1	0	0	7	8
AAG	8	3	5	8	24
AAT	0	3	6	8	17
ACA	1	1	3	0	5
ACC	0	1	0	0	1
ACG	0	0	0	0	0
ACT	0	1	5	2	8
AGA	4	1	10	6	21
AGC	3	1	1	6	11
AGG	17	2	3	8	30
AGT	6	2	18	36	62
ATA	0	0	4	0	4
ATC	1	1	0	4	6
ATG	6	3	3	4	16
ATT	4	5	4	25	38

After calculating the expected frequency matrix, we calculate the probability matrix:

Probability Matrix for Nitrogenous Base Quadruplets

	A	C	G	T
AAA	0.208	0.208	333. 0	0.25
AAC	0.125	0	0.875	0
AAG	0.333	0.125	333. 0	0.208
AAT	0	0.176	470. 0	0.352
ACA	0.2	0.2	0	0.6
ACC	0	1	0	0
ACG	0	0	0	0
ACT	0	0.125	0.25	0.625
AGA	0.190	0.047	0.285	0.476
AGC	0.272	0.090	545. 0	0.090
AGG	0.566	0.066	266. 0	0.1
AGT	0.096	0.032	580. 0	0.290
ATA	0	0	0	1
ATC	1.065	0.166	666 .0	0
ATG	0.375	0.187	0.25	0.187
ATT	0.105	0.131	0.657	0.105

After calculating both the frequency and probability matrices, we calculate the value of ln(L) according to the previous steps. It was found that the value of ln(L) for this test is :

$\ln(L) = \sum nij\ Ln\ Pij$

ln(L) = -241.883

AIC = 2p - 2 ln(L)

We find the value of p for the third order according to the equation :

$p = N^K (N-1) = 4^3 * (4-1) = 64 * 3 = 192$

2p = 2 * 192 = 384

$p = N^K (N-1) = 4^3 * (4-1) = 64 * 3 = 192$

AIC = 675.767

This is the AIC value for this test

Comparison and Decision-making table

Alc	Ln(L)	(P) Parameters	(K) Rank
5736.41	-2862.20	12	First
3274.05	-1589.02	48	Second
675.767	-241.883	192	Third

Conclusions

1- The results obtained using Shannon's information criterion indicate that the Markov chain for the breast cancer gene is of the second order, meaning it depends on its previous state. This was the primary objective of the research.

2- since the Alc value of the second model is the lowest, the best and most likely model to represent this time series is the second-order Markov model. Although the second-order model improved the probability logarithm from (675.767) to (3274.05), the large penalty imposed on it due to its huge number of parameters made it generally worse

3- Regarding preference, Shannon's criterion is considered superior for determining the order, as Akaike's criterion does not provide an absolute measure of model quality but is used only for comparison between models applied to the same data.

Recommendations

1- Adoption of the research results by the National Center for Biotechnology Information

2- (NCBI) for use in predicting the occurrence of the disease and expediting decision-making.

References

[1] Al-Azdi, Iman Suleiman Mohammed. (2002). "Estimating the Order of Markov Chains with Application to DNA Sequences." Unpublished Master's Thesis, College of Computer Science and Mathematics, University of Mosul, Iraq.

[2] Al-Khayyat, Basil Younis Dhanoon. (2011). "Markov Modeling with Practical Applications." Ibn Al-Atheer Publishing House, University of Mosul, 2011.

[3] Al-Amin, Israa Abduljawad Saleh. (2009). "Estimating the Order of a Markov Chain for Weather Conditions in the City of Mosul Using Shannon's Information Criterion and Backpropagation Network." Unpublished Master's Thesis, College of Computer Science and Mathematics, University of Mosul, Iraq.

[4] Basawa, Ishwar V., Prakasarao, B. L. S., (1980): "Statistical Inference for Stochastic Processes", Academic Press, London, Newyork.

[5] Finesso, L. (1991): "Consistent Estimation of the Order for Markov and Hidden Markov Chain", Unpublished Ph. D. Thesis, Dept of Electrical Engineering, Harvard University.

[6] Lindley, D.V. (1956): "Onameasure of the Information Provided by an Experiment", Ann. Math. Statist .27, 986-1005.

[7] Tong, H., (1975): "Determination of the order of a Markov Chain by using Akaike's Information Criterion",J. Ap.Prob.12,488-497.

Appendix A

Markov chain of the breast cancer gene

"ATGTCTCTTTGGTTTCAGTAGTTCAGAGGCGGAGGTTGGGAGGAAGC
TGCTTGGTAGCAGCCTGCAGTTGAAGAGAAAGTGTCATTTTAAAAGCGCGTCAGT
GTGTGGAAGGAAATGCTGGGTTACAGATAGGTTTAATGCTTTAGGAAGGATTTTGT
GAGGCACCCAAGTTTGGAGTTTCTTGTGTTGGGAAATCCAGTGTGGTGGCAGACAG
TATTCATTTCCCCTTTAAAGTATCTAGTTTGTATCTTTGAGTAGGTTTTTCTAGAG
TTCTTGCTTGAGTGTGTATTTCAGTCTATGGTGGAGTGATGTGATGAGAAACTTCT
TTGTGAGGATTCAGAGCAGTGAGGAACAGCTTAGAGAGGATCAGAGTGAGGAAGAG
GAAAGAAATGAAAAATTCTTGTTTTATGATTTTATTTTTGTCACAAACTGAAATCT
TAGTTTGAGTGTGTCTTTTTTGAGTTGCTTCTGGTGTGAATTTGAAAATGCAGTTT
GGAAGTGATTTGGGAATTGAGCTTTGGGACACTCTTCTTGTTTGAACTTTCAGGGT
GCTTTGTGTTTGGAAGTGAGGATGGAATTTTGTGTATTTTATGAGTTTTTGTAATT
GTGAGTCTTGTAACTGGAATTTGGAGTTGATTCTGAGTGTGAGTTCATTTTTAAAC
TGCTAGGAAAGATTTTGTGGTTTTTGTGAGTTTTTGTTGAGGAGTATGGGAAGATA
GTTTGGCTTAGCTGCTTGTGATTTTGTATTCTTTTTTGTTGTTTAAATTTTTGTGA
GTAAACTGAAATGATAGAAAAGCAGTTGAAGGAGGGAGCTTTTGGGATTAAGTGTT
TTTAAGAGTTTTCTGAAACTGCTGTGttiTTTGAAGGGATTTGGATTAGATTTATTT
TGAAGTTTCTTTTGGTGTATTGGTGGATTATTTAGCTTTAGTGTTTGTTGTGGTTT
GTTAGTTGGTATTTGGATTAGGAGTTGAAGAGTTGTGAGGTTTTGTGATAGTTTTG
GAGTGTTTTGTTGAATCTGAAGGTTGAGGAATTTGTTGAAGATTTGGAGTTTTGTGT
AGTTTTGTGTTGTTGAGTTTGATTTTGGGTTTTGGAGGTGTTGTTGAATGAGGAAGT
TTTGGATTGAGTTTTGATTTTGTTGGAGTTGTTTTTGATTTGTTGTTGGAGATGTTT
TTGGTTTGGAGTGTGTTTGTTTTGGAGTTGGTTTGTTTGGAGTGTTGTTGTTGGAGT
TTTGTTGAGTTTGTTTGGTGTTTGTGGTTTTGGAGGTTTTGTGTAGTTTGGTGAGGA
TGTTTTTGAGTGTGGAGTGTGTGTTTGGTTTGTTTGTTGGAGTTTTGTTTGGTTTTG
TTTGAGTTTGTTTTGTTGGAGTTTTGGTTTGGAGTTTTTGGTGTGTTGGAGGTTTTGT
TTGGAGTTTTGTGTTTTG"

NO.Of Observition =1416
No. Of Cases =K=4
L =2