Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

Advances in Applied Science Research, 2014, 5(2):387-393

ISSN: 0976-8610
CODEN (USA): AASRFC

Markov chain model to study the gene expression

[1]Amit Sharma and [2]Neeru Adlakha

[1]Department of Applied Mathematics & Humanities, S. V. National Institute of Technology, Surat, India
[2]Department of Mathematics, M. A. National Institute of Technology, Bhopal, India

---

ABSTRACT

This paper presents an approach for study of gene expressions based on Markov chain theory. A mathematical model has been proposed to study the transcription of DNA into mRNA and translation of mRNA into Proteins. It is assumed that the DNA, mRNA and Proteins are states in the model and initial state of the system is known. Based on the initial state, the successive states (without state feedback) are predicted using probabilities. The model is used to predict the final state Proteins of the system. The model is illustrated with the help of data set of asparagus maritimus RpoA gene, Accession: EU051382.1 and unidentified retrovirus gene, Accession: BD432460.

Key words: DNA, mRNA, Protein, Markov Chain Theory.

INTRODUCTION

The cells are the primordial units of all living organisms. In a biological cell, genes express continuously. The most fundamental property of cell of all living organisms is their ability to reproduce. Even more amazing is that each cell stores its own set of instructions for carrying out each of these activities [1].

The central dogma of molecular biology deals with the transfer of sequential information in organisms. According to it information cannot be transferred back from protein to either protein or nucleic acid. In other words, once information gets into protein, it can't flow back to nucleic acid. The dogma is a framework for understanding the transfer of sequence information between sequential information-carrying biopolymers, in the most common or general case, in living organisms [2]. The final information embedded in DNA and RNA is ultimately expressed as proteins. Proteins are one of the most important biological macromolecules made of 20 alphabets called as amino acids. They make most of the structures in cells. Normally, proteins act as the eyes and ears of cells. Hence, it is crucial for researchers to understand the basic phenomenon which underlies the functioning of these molecules called bio molecules which are the primers of life [3].

RNA interference has been exploited in disease therapy and control. The use of short interfering RNA mimics has been more successful [4]. The first application to reach clinical trials was in the treatment of muscular degeneration and respiratory syncytial virus [5]. RNAi has also been shown to be effective in the reversal of induced liver failure in mouse models [6, 7]. Many features of proteins for their classification and prediction have exploited [8]. With attributes like amino acid, di-peptide and tri-peptide composition, proteins can be classified into various levels of classes and subclasses. Similarly, using evolutionary information and statistical factor based scores an in depth analysis of proteins and their families have been achieved [9].

Regarding health issues in eukaryotic cells, genetic factors can be affected by energy insufficiency of oxygen which is the cause of weakness. A comparative study of gene expression has been carried out in healthy and weak persons. It is also manifested that *Single Cell Protein* (SCP) is not only provide a nutritional enhancer but also play an important role for other functioning in all living organism. It comes from microbial source and pointed out the production, processing and consumption for food supplements [10, 11]. A study on chloramphenicol acetyl

transferase, used as reporter gene in microbes, is responsible for chloramphenicol resistance in bacteria has been proposed. For DNA degradation of orange samples, a genomic DNA extraction method is developed and demonstrates the effect of microwave application [12, 13]. Molecular mechanics is primordial for the study of molecular modeling. A molecular modeling based model has been demonstrated using PM6 model in rosiglitazone metabolism [14].

MATERIALS AND METHODS

1. Preliminaries

Markov chain plays an important role to solve the complicated problems in many areas of science and technology such as: Polymers, Biology, Physics, Chemistry, Operations Research, Computer Networks etc. The application of Markov chain, in Chemical Engineering has been relatively diminutive [15]. It have extensively been dealt with in references [16, 17], mainly by mathematicians. Markov Chain Theory has great potential for applications in the field of Bioinformatics [18].

A Markov chain model can be described by the three-tuple: State space, Transition probability matrix and Initial state vector. The most important advantage is that physical models can be presented in a unified description via ***state vector*** and a ***one-step transition probability matrix***. The essence of the model is that if the initial state of a system is known, i.e. its present state and the probabilities to move forward to other states are also given then it is possible to predict the future state of the system ignoring its past history. In other words, it does not depend on the past history of the system for predicting the future; this is the key-element in Markov chains [19].

In the discrete processes, the mathematical formulation of the complex problem can be expressed. Let there are finite or countably infinite number of states of a system. Also, let the states are $S_1, S_2, S_3, S_4, S_5 \ldots . S_i \ldots . S_R, S_P$ where S reveals for state [20, 21]. Let $X(t)$ be a discrete random variable which reveals the states of the system with respect to time. In a discrete process, the quantity t intimates the number of steps from time zero and $X(t)$ designates the fact that the system has occupied some state at step t. $X(t)$ can be assigned any of the values corresponding to the states $S_1, S_2, S_3, S_4, S_5 \ldots . S_i$, i.e., $X(t) = S_i$ where state i was occupied by the system on step t [22, 23]. The notations used in the present formulation are as given below:

2. Nomenclature

SS	State space
S_D	State of DNA
S_R	State of mRNA
S_P	State of Proteins
X	Function of time
P_{DP}	Two-step transition probability function
m, n	Steps or time, where $m \neq n$
$S_D(n)$	Probability which is occurred by the system of state DNA at time n or step n
$S_R(n)$	Probability which is occurred by the system of state mRNA at time n or step n
$S_P(n)$	Probability which is occurred by the system of state Protein at time n or step n

The following assumptions have been made in the present model.

i. In the present model, three states have been considered, namely DNA, mRNA and Proteins.

ii. Also, it is assumed that DNA completely changes into mRNA and in turns into Proteins in a normal cell but in the case of retroviruses, it does not follow.

iii. It is assumed that no more DNA is being transcript into mRNA and whatever RNA is available in the cell is being translated into Proteins.

3. Mathematical Model

In this model, Markov Chain Theory is used to show the change of one state to another state. In this theory, the future state of a system can be predicted on the basis of present state ignoring its past history. A cell has DNA, mRNA and Proteins which is considered as system in the model. In the present approach, DNA, mRNA and the Proteins are considered as states. Now the problem is how to construct the basic elements of Markov chain depicting

the change of the states DNA to mRNA and in turns to Proteins. Suppose S_D be the state of DNA, S_R be the state of mRNA and S_P be the state of Proteins. For this, let SS be the state space, we have

$$SS = [S_D, S_R, S_P] \quad \ldots(1)$$

For the state space the probability of states S_D, S_R and S_P is given by

$$0 \le \text{Prob}\{S_i\} \le 1 \qquad \text{where } i = D, R \text{ and } P \quad \ldots(2)$$

The expression for transition to Proteins state provided that the DNA and mRNA states exists can be expressed as

$$0 \le \text{Prob}\{S_P \,|\, S_D, S_R\} \le 1 \quad \ldots(3)$$

In the above equation, the state S_D must be occupied before occupying each of others. As per our assumption, the entire DNA converts completely into mRNA and all the mRNA converts completely into protein, therefore we have

$$\text{Prob}_f\{S_P \,|\, S_D, S_R\} = 1 \quad \ldots(4)$$

Since, the state S_D change into S_R and S_R changes into S_P, which has two steps to complete the whole process, then the two-step transition probability matrix can be constructed. Suppose P_{DP} is the two-step transition probability function then we have,

$$P_{DP} = \text{prob}\{S_P \mid S_D, S_R\} \qquad \text{For all } D, R \,\&\, S \quad \ldots(5)$$

Further, the two-step transition probability function P_{DP} is independent of time or time-homogeneous then the probability of a transition from one given sate to another state depends solely on the states. Also, there are three states namely DNA, RNA and Proteins, where the DNA state is given then the two-step transition probabilities in matrix form can be given as

$$P = P_{DP} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} \quad \ldots(6)$$

Where P_{DP} denotes the probability of transition from state DNA to state Proteins. Finally, the initial state vector, a function which gives the probability that the system is initially in state i, i.e., at time zero or at step zero, the system is initially in state i. Thus, the initial state vector is

$$S_i(0) = \text{Prob}\{S_i\} \qquad \text{where } i = D, R \,\&\, P \quad \ldots(7)$$

This can be arranged in row vector form of the initial state vector as

$$S(0) = [S_D(0), S_R(0), S_P(0)] \quad \ldots(8)$$

Similarly, if there are n steps in the initial state vector then the row vector form is

$$S(n) = [S_D(n), S_R(n), S_P(n)] \quad \ldots(9)$$

Now multiplying P_{DP} and $S(n)$ from equation (6) & (9) respectively, gives us the new row vector $S(n+1)$, that is, the probability of occupying state Proteins S_P at $(n+1)$ is

$$S_P(n+1) = S(n)P_{DP} \quad \ldots(10)$$

Where P_{DP} is the probability of transition from state DNA to state Proteins and presents a recurrence relation and it can be expressed in matrix notation as

$$S(n+1) = S(n)P \quad \ldots(11)$$

Using iteration method, equation (10) can be written as

$$S(n+1) = S(0)P^{n+1} \quad \text{where } n = 0,\ 1,\ 2,\ 3..... \quad \ldots(12)$$

Now, from equation (6) & (8), putting the value of P and $S(0)$ in equation (12) respectively, we get

$$S(n+1) = [S_D(n), S_R(n), S_P(n)] \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}^{n+1} \quad \ldots(13)$$

The above equation (13) shows that if the initial state of gene, i.e., DNA is known and the probability of moving forward from the state DNA to state mRNA and the state mRNA to state Proteins is given then the state Proteins of the system can be predicted.

RESULTS AND DISCUSSION

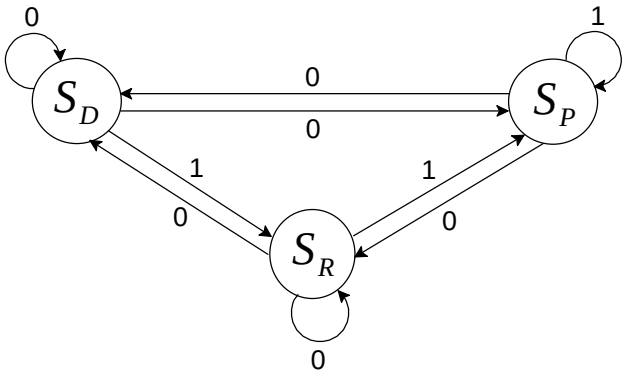
Since there are three states DNA, mRNA and Proteins, where DNA is the initial state and Proteins is the future state, thus the state space $[S_D, S_R, S_P]$ is finite. Secondly, there are two steps as DNA transcript into mRNA and mRNA translate into Proteins, then the two step transition probability is

$$P_{DP} = \begin{array}{c} \\ S_D \\ S_R \\ S_P \end{array} \begin{array}{c} \begin{array}{ccc} S_D & S_R & S_P \end{array} \\ \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}^2 \end{array}$$

Also, the initial state vector is

$$S(0) = [1, 0, 0]$$

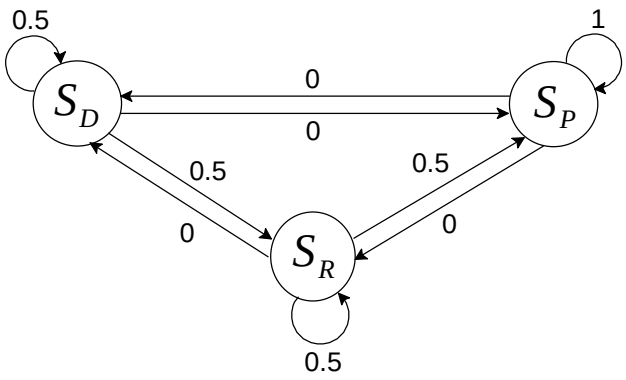
The graphical representation of transition probabilities are given below:



$$P_{DP} = \begin{array}{c} \\ S_D \\ S_R \\ S_P \end{array} \begin{array}{c} \begin{array}{ccc} S_D & S_R & S_P \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}^2 \end{array}$$

$$S_P = [1\ 0\ 0] \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}^2 = [0\ 0\ 1]$$

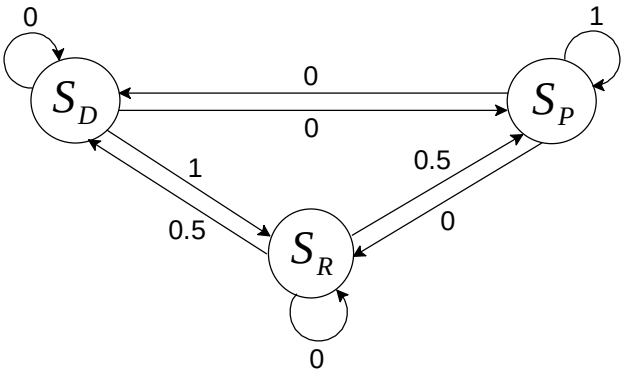
Fig. 1(a): State transition diagram and transition probability matrix in a normal cell



$$P_{DP} = \begin{array}{c} \\ S_D \\ S_R \\ S_P \end{array} \begin{array}{c} \begin{array}{ccc} S_D & S_R & S_P \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}^2 \end{array}$$

$$S_P = [1\ 0\ 0] \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}^2 = [0.25\ 0.50\ 0.25]$$

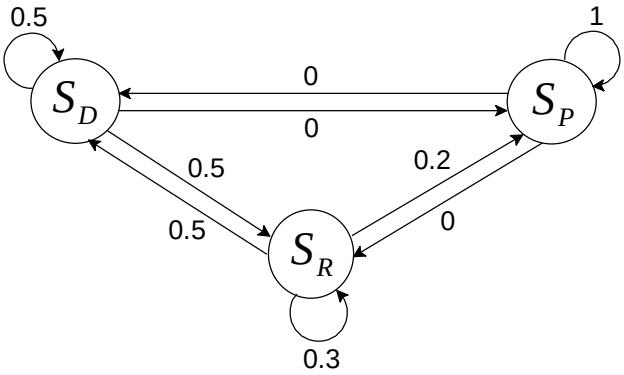
Fig. 1(b): State transition diagram and transition probability matrix in a normal cell



$$P_{DP} = \begin{array}{c} \\ S_D \\ S_R \\ S_P \end{array} \begin{array}{c} \begin{array}{ccc} S_D & S_R & S_P \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}^2 \end{array}$$

$$S_P = [1\ 0\ 0] \begin{bmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 1 \end{bmatrix}^2 = [0.5\ 0\ 0.5]$$

Fig. 2(a): State transition diagram and transition probability matrix in retroviruses



$$P_{DP} = \begin{array}{c} \\ S_D \\ S_R \\ S_P \end{array} \begin{array}{c} \begin{array}{ccc} S_D & S_R & S_P \end{array} \\ \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.3 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}^2 \end{array}$$

$$S_P = [1\ 0\ 0] \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.3 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}^2 = [0.5\ 0.4\ 0.1]$$

Fig. 2(b): State transition diagram and transition probability matrix in retroviruses

In Fig. 1(a), the DNA state has completely changed into mRNA state and in turns into Proteins state. Also, there are some examples which show the possibilities of DNA, mRNA and Proteins as states in the cell. Regarding this, Fig. 1(b) shows that only 25 percent DNA state transcripts into mRNA state and 50 percent of whole transcripted mRNA state change into Proteins state. In this, we get 25 percent of Proteins as final state. We take the data set of gene RpoA in E. Coli from NCBI [24], which follows our model; Accession: EU051382, GI: 158021634, GenBank: EU051382.1, Organism: Asparagus maritimus. There are 359 base pair (bp) in the gene. The DNA sequence is taken in FASTA format and using the software "EMBOSS Transeq", we get the 6 frame Proteins sequences. We found that the entire DNA sequence transcripts into mRNA sequence and mRNA sequence translates into Proteins as given below:

```
>EU051382.1_1 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-
petD intergenic spacer, partial sequence; chloroplast
SYEN*IFSHRRCKTDIGHSTEAFCNQFT*E*TFILI*IY*NLFIFFLYFINYILKV*RKN
FIFYLRHAIRIFAE*IIIKFMYLGRIHFRRDYSLDTYIVVFHD*IHLKKT*S*RFINR*X

>EU051382.1_2 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-
petD intergenic spacer, partial sequence; chloroplast
LMKIEYFRTEDVKQILDTLQKHFAINLPKNKLSF*FKSIRIFSYSFYIL*IIF*RSKEKT
```

LFFIFGTRSVFSRNRS**NSCI*GGFTLEGTIP*IPTSWYFTIESI*KRPKVRDLSIGXX

>EU051382.1_3 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-petD intergenic spacer, partial sequence; chloroplast
L*KLNIFAQKM*NRYWTLYRSILQSIYLRINFHFNLNLLESFHILFIFYKLYSKGLKKKL
YFLSSARDPYFRGIDHNKIHVSREDSL*KGLFPRYLHRGISRLNPFKKDLKLEIYQ*VX

>EU051382.1_4 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-petD intergenic spacer, partial sequence; chloroplast
LPIDKSLTLGLF*MDSIVKYHDVGI*GIVPSKVNPP*IHEFYYDLFRENTDRVPKIKNKV
FSLDL*NIIYKI*KEYEKILIDLN*NESLFLGKLIAKCFCRVSNICFTSSVRKYSIFIR

>EU051382.1_5 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-petD intergenic spacer, partial sequence; chloroplast
XTY**ISNFRSFLNGFNREIPRCRYLGNSPF*SESSLDT*ILL*SIPRKYGSRAEDKK*S
FFFRPLEYNL*NIKRI*KDSNRFKLK*KFILR*IDCKMLL*SVQYLFYIFCAKIFNFHKX

>EU051382.1_6 Asparagus maritimus RpoA (rpoA) gene, partial cds; and rpoA-petD intergenic spacer, partial sequence; chloroplast
XYLLINL*L*VFFKWIQS*NTTM*VSRE*SLLK*ILPRYMNFIMIYSAKIRIACRR*KIK
FFL*TFRI*FIKYKKNMKRF**I*IKMKVYS*VN*LQNASVECPISVLHLLCENIQFS*X

In Fig. 2(a), only 50 percent DNA state transcripts into mRNA state but whole transcripted mRNA state translates into Proteins state. In this, we get 50 percent of Proteins as final state. Also, the probability of S_D and S_R remaining in state S_D and S_R are 0 respectively, the probability of S_D reaching in state S_R is 1. The probability of S_R reaching in state S_D and S_P is 0.5, which shows that the reverse transcription as mRNA state reverse transcripts into DNA state. The probability of S_P remaining in state S_D and S_R is 0 and the probability of S_P remaining in state S_P is 1. In Fig. 2(b), only 50 percent DNA state transcripts into mRNA state and only 40 percent of whole transcripted mRNA state translate into Proteins state. In this, we get 10 percent of Proteins as final state. For this, we get the gene of unidentified retrovirus "an infective endogenous retrovirus and its association with demyelinating diseases and other diseases", Accession: BD432460, GI: 92376536, GenBank: BD432460.1, Organism: unidentified retrovirus. There are 129 base pair (bp) in the gene. Again, the DNA sequences are taken in FASTA format from NCBI [25] and using the software "EMBOSS Transeq", we get the 6 frame Proteins sequences as given below:

>BD432460.1_1 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
RLSG*RLTLPDRLGSPLDHHGCRASGNSHNGRSPAI*RQPSWT

>BD432460.1_2 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
GLAAED*RCPIASEAP*TITDAELRVTLTMEDPQPYEDNLAGR

>BD432460.1_3 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
A*RLKTDAARSPRKPPRPSRMPSFG*LSQWKIPSHMKTT*LDX

>BD432460.1_4 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
RPARLSSYGWGSSIVRVTRSSASVMV*GASEAIGQRQSSAAKP

>BD432460.1_5 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
SS*VVFIWLGIFHCESYPKLGIRDGLGGFRGDRAASVFSR*AX

>BD432460.1_6 AN INFECTIVE ENDOGENOUS RETROVIRUS AND ITS ASSOCIATION WITH DEMYELINATING DISEASES AND OTHER DISEASES
VQLGCLHMAGDLPL*ELPEARHP*WSRGLPRRSGSVSLQPLSR

CONCLUSION

In this paper, a Markov chain is employed to propose a model for gene expression when the state space, initial state vector and the two-step transition probability matrix are given. All the three properties are satisfied to predict the state Proteins from state mRNA and in turn from the state DNA. Thus, a different approach is used to find the Proteins state. It is also shown that if the initial state of a system is known and the probability to move forward from one state to another state is also given then the future state of the system can be predicted. The model is also suitable for retroviruses as the probability to move from mRNA to DNA is given, which is shown in figure 2(a) and 2(b). Such models can be developed and employed to predict gene expressions in sequences of other organisms and generate information and knowledge which may be useful for development of protocols for diagnosis, prevention and treatment of diseases.

Acknowledgements
The first author is grateful to Council of Scientific & Industrial Research (CSIR), New Delhi, India, award no - 09/1007(0002)/2009 for giving financial assistance as JRF/SRF.

REFERENCES

[1] Berg JM, Tymoczko JL, Stryer L, 5th Ed., *Biochemistry*, WH Freeman and Company, **2002**, pp. 118–119 and 781–808.
[2] Crick F, *Nature*, **1970**, 227, 5258.
[3] Crick F, *Symp Soc Exp Bio*, **1958**, 12.
[4] Paddison P, Caudy A, Hannon G, *Proc Nat Acad Sci*, **2002**, 99, 3.
[5] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM, *Nature*, **2005**, 433, 7027.
[6] Zender L, Hutker S, Liedtke C, Tillmann HL, Zender S, Mundt B, Waltemathe M, Gosling T, Flemming P, Malek NP, Trautwein C, Manns MP, Kuhnel F, Kubicka S, *Proc Nat Acad Sci*, **2003**, 100, 13.
[7] Sah D, *Life Sci*, **2006**, 79, 19.
[8] Bhasin M, Raghava GPS, *Nuc A Res*, **2005**, 33, 2.
[9] Mundraa P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD, *Patt Recog Let*, **2007**, 28, 13.
[10] Zamanloo K, Abolghasemi A, Zaeifizadeh M, *Euro J Exp Bio*, **2013,** 3, 1.
[11] Adedayo MR, Ajiboye EA, Akintunde JK, Odaibo A, *Ad App Sci Res*, **2011**, 2, 5.
[12] Tharian JA, Padmapriya R, Thirunalasundari T, *Ad Ap Sci Res*, **2013**, 4, 4.
[13] Abdolmaleki F, Assadi MM, Ezzatpanah H, Honarvar M, *Euro J Exp Bio*, **2013**, 3, 6.
[14] Kumar A, Kumar S, Jain S, Kumar P, *Der Pharm Sini*, **2010**, 1, 2.
[15] Parzen E, *Modern Probability Theory and Its Applications*, John Wiley & Sons, Inc., **1960**.
[16] Bharucha-Reid AT, *Elements of the Theory of Markov Processes and their Applications*, McGraw-Hill Book Company, Inc., **1960**.
[17] Howard RA, *Dynamic Programming and Markov Processes*, The M.I.T. Press, **1960**.
[18] Kemeny JG, Snel1 JL, *Finite Markov Chains*, D.Van Nostrand Company, Inc., **1960**.
[19] Parzen E, *Stochastic Processes*, Holden-Day, Inc., **1962**.
[20] Lowry GG, *Markov chains and Monte Carlo Calculations in Polymer Science*, Marcel Dekker, Inc., New York, **1970**.
[21] Norman MF, *Markov Processes and Learning Models*, Academic Press, **1972**.
[22] Cinlar E, *Introduction to Stochastic Processes*, Prentice-Hall, Inc, **1975**.
[23] Kovalenko IN, Kuznetsov NY, Shurenkov VM, *Models of Random Processes*, CRC Press, Inc., Boca Raton, New York, Florida, **1996**.
[24] http://www.ncbi.nlm.nih.gov/nuccore/EU051382.1
[25] http://www.ncbi.nlm.nih.gov/nuccore/BD432460.1