ISSN: 2230-9926

*Available online at http://www.journalijdr.com*

IJDR

*International Journal of Development Research*
*Vol. 07, Issue, 08, pp.14658-14662, August, 2017*

ORIGINAL RESEARCH ARTICLE OPEN ACCESS

# THE RECOGNITION SYSTEM OF SICKLE CELL ANEMIA BY USING HIDDEN MARKOV MODEL

**[1]Dr. Mohamed Soueycatt, [2]Dr.Ziad Kanaya, [3]Dr. Ahmad Younso and [4]Nour Azha**

[1]Professor of Faculty of Medical Engineering, AL-Andalus University for medical sciences, Tartus, Syria
[2]Associate Professor at Mathematics Department, Faculty of Science, Tishreen University, Lattakia, Syria
[3]Assistant Professor at Mathematical Statistics Department, Faculty of Science, Damascus University, Damascus, Syria
[4]Postgraduate Student, Department of Mathematics, Faculty Science, Tishreen University, Lattakia, Syria

**ARTICLE INFO**

***Article History:***
Received 14th May, 2017
Received in revised form
25th June, 2017
Accepted 22nd July, 2017
Published online 30th August, 2017

***Keywords:***

Bioinformatics,
Hidden Markov Model,
DNA sequences,
sickle cell Anemia,
Nucleotides,
Viterbi algorithm.

***Corresponding author***

**ABSTRACT**

The study of genetic mutations, that is responsible for diseases , is an important issue in genetics for its close relationship with the genetic evolution of living organisms , In this paper we present an algorithm that is bases on the Hidden Markov Models of recognition to the mutation that causes one of the most common genetic diseases, Sickle Cell disease, thus diagnoses the person state (infected, uninfected) , This method is applied to DNA sequence , Deoxyribonucleic acid, the practical application shows that the rate of recognition of an infected person equals (99%) and the rate of recognition of a healthy person equals (86.33%) ,All the code is written by using statistical program R.



**Citation: Dr. Mohamed Soueycatt, Dr. Ziad Kanaya, Dr. Ahmad Younso and Nour Azha. 2017.** "The recognition system of sickle cell anemia by using hidden markov mode.", *International Journal of Development Research*, 7, (08), 14658-14662.

## INTRODUCTION

The basic idea of the symbol recognition system is to simulate the computer in its work to recognition symbols of human visual system, that is, The computer artificial systems work similar to the human neurons dedicated to the perception of images and symbols as the various scenes stored somewhere in the brain memory they can be restored when needed, and this is what the computer will do when storing data in a way that simulates the work of the human brain. There are wide applications for this research, such as recognition people's identities, recognition car plates, recognizing bodies and victims of war, recognizing credit cards and finding out the counterfeit coin (Sharma, 2007).

Statistical systems have achieved remarkable results in recognition systems such as statistical analysis of multiple variables and stepwise regression systems based on reducing mean square error and Hidden Markov models the essence of our current research is comparing the image or symbol to be recognizable with a bank of stored information (images or symbols) then deciding if it's recognized or not , It's programmed recognizing the symbols in the Bioinformatics field (DNA sequences) where we will depend on the DNA sequence of a healthy person and then enter a DNA sequence for a person with sickle cell anemia and recognizing the sequence that caused the mutation.

**The Aims of research and its importance**

This research aims to design an algorithm that shows nucleotides that is responsible for generating DNA sequences that contain a genetic mutation that has caused one of the most common diseases of sickle cell anemia by using Markov's hidden system and its algorithms. The importance of this research is that it gives an incentive for many researches and applications that show the extent of confidentiality and human interaction with the computer in the field of dealing with genetic sequence by computer because of the high importance in the areas of genealogy, medical analysis, disease detection, and many other areas.

**The concept of hidden Markov models**

The hidden Markov models (HMMs) and it's algorithms are mainly inspired over ninety years from mathematical models which are as known as(Andrei Markov) the name of the scientist who has discovered them, which has appeared at the beginning of the twentieth century, called models of Markov, and this shows that the hidden Markov models (HMMs) are an extension of the Markov models usual (MMs), and the detection of this model is attributed to the researcher (Leonard E.Baum) and others , when they published a set of statistical articles in this regard in the second half of the sixties of the twentieth century . The first application of the hidden Markov model is in voice recognition (Speech Recognition) (Abdulla, 1999) and in the mid-seventies of the twentieth century, and in the second half of the eighties of the twentieth century, the use of hidden Markov model has begun in the Biological Sequence analysis especially (DNA) (Fonzo *et al*., 2007; Al-Khayyatt Younes, 2010) and since then the hidden Markov model has existence his presence in the field of bioinformatics.

Among the most important applications of hidden Markov models (Li xiaolin, 2000):

- Speech Recognition.
- Weather Modeling where we assumes the weather everyday.
- Word – Sense Disambiguation.
- DNA Sequence Modeling.
- Text Modeling and information extraction.
- Hand written Recognition.
- Network intrusion detection systems.

**Characterization of a Hidden Markov Model (Chen *et al*., 2005)**

Assuming we have markovian process $\{q_t\}$ with space states $S = \{S_1, S_2, \dots, S_N\}$ where (N) the number of hidden states in the model , and (M) The number of different observation symbols in the model we now give a formal characterization of a hidden markov sequence model or HMM in terms of its basic elements and parameters $\Theta = \{\pi, A, B\}$:

Transition probability $A = \{a_{ij}\}_{i,j \in S}$ of a homogeneous Markov chain with a total of N states:

$$a_{ij} = p(q_{t+1} = S_j / q_t = S_i)$$

Initial Markov chain state-occupation probabilities :$\pi_i = p(q_1 = S_i) \ \forall i \in S$ .

Observation probability distribution $P(O_t/S_i)$ , i = 1,2,…,N if $O_t$ is discrete , the distribution associated with each state gives the probabilities of symbolic observation $V = \{v_1, v_2, \dots, v_k\}$ :

$$b_i(k) = P(O_t = v_k / q_t = S_i) \ , i = 1 , 2 , \dots ,N$$

This model is used in the countable states and is widely used in weather recognition systems and Bioinformatics but if the observation probability distribution is continuous ,then the model is used in the uncountable states and is usually given by mixed natural density .

**Hidden Markov model types (**DanialJurafsky, 2000**)**

**Transitions HMM are divided into two parts**

**Ergodic** : It is possible to move from any state to any other state and clarify it as figure (1):

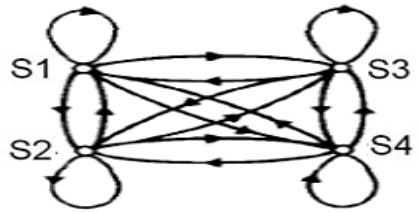


**Figure (1) Ergodic HMM**

**Left-Right:** You can move from the state to the state on the right only or go to the same state and clarify it as figure (2):

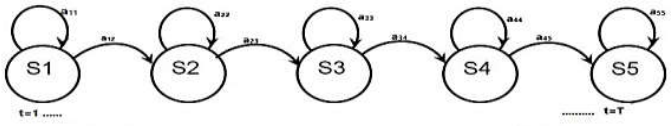


**Figure (2) Left-Right HMM**

**Hidden Markov Algorithms**

The previous version of the hidden Markov model has created three important main problems that must be resolved before using the model in applications. These problems are:

**The First problems (evaluation problem) (Abdulla, 1999; Grant, 2005):**

Given the observation sequence $O = O_1, O_2, \dots, O_T$ and model $\Theta = (A, B, \pi)$ , the problem calculating the probability of a series of observations for given model $P(O/\Theta)$ , and it's resolved by using the Forward – Backward algorithm.

**The Second problem (training problem)( Abdulla, 1999; Johanneson, 1999; Bilmes, 2002).**

Given the observation sequence $O = O_1, O_2, \dots, O_T$ and model $\Theta = (A, B, \pi)$ , The problem of selecting the best states series $Q = q_1, q_2, \dots, q_T$ and it's resolved by using the **Viterbi** algorithm (Batzoglou, 2010).

**Third problem (scale problem) (Bilmes, 2002; Rabiner, 1989)**

Modification of model elements $\Theta = (A, B, \pi)$ for getting the likelihood probability to $P(O/\Theta)$ and it's resolved by using the **Baum-Welch** algorithm.

**DNA sequences Concept**

DNA is the basis of all living organisms and has two types (DNA / RNA). It's quantity and complexity of packing vary according to the organism. The Deoxyribonucleic acid DNA is considered the genetic material of all cells of real and primitive - nucleus organisms, which are helix double strings. Each strings consists of many nucleotides (polynucleotides) figure (3),and the nucleotide consists of :

**Nitrogen base:** This base is divided into two types:

**Purines:** and the most common types of DNA in this dna molecule are adenine (A) and guanine (G), which are larger than the second class.

- Pyrimidine, which is smaller than the previous type and the most common types of it in DNA are Cytosine (C) and Thymine (T).
- Pentose sugar which it's called (Deoxyribose) .
- Phosphate Group ($PO_4$).

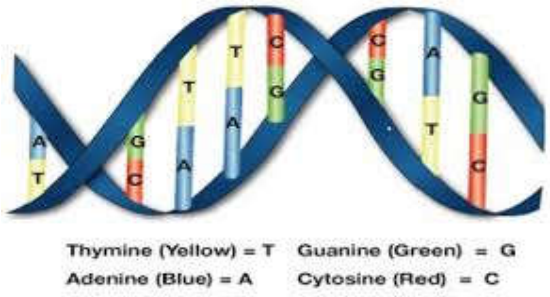


**Figure (3) DNA sequences model**

Each cell in the human body has the same DNA as every other cell , and it is organized into 46 chromosomes and there are about 3 billion nitrogen bases. Only 1-2% are genes and the other 98% are still studied to know and each gene , is two copies, one is a copy of the mother and the other is a copy of the father, and there are 99.6% similarity between any two persons, the difference is only 0.4%. This is about 12 million pairs of nitrogen bases and we share with the monkey 96% of our DNA. In this paper we will study DNA, which consists of A, T, C, and G **(**Bruce, 2002; Chris, 2004**)**

**The concept of mutation**

It is a change in the nucleotide bases sequence of genes to affect the phenotype. The mutation can be occurred at the molecular level to replace one nucleotide base instead of another. The mutation can occurred at the molecular level, replacing one nitrogen base rather than another, or adding a number of bases or deleting them , may be occurred at the mutation chromosome level so that the part of the chromosome can change and a sudden change in the number of chromosomes or method of their system. And this change causes the emergence of a new character. The mutation in it's genesis (chromosomal and genetic) may be spontaneously occurred as a result of dysfunction of the cell division. It has a defect in the number of chromosomes or their order system, and mutation can be occurred as a result of Exposure to radiation or using chemicals that affect chromosomes or genes (Burrus, 2004).

**Sickle cell anemia**

The Sickle cell disease is a hereditary disease that affects the red blood cells and changes their shape from circular globose to a lunar figure(4). These cells are attacked to each other, causing adhesion to the walls of the small blood vessels and can not pass to the cells and transport oxygen to them, which causes symptoms These cells are fragile, weak, and rapidly breakable and the symptoms are appeared in the fifth month of age(William *et al*., 2016).

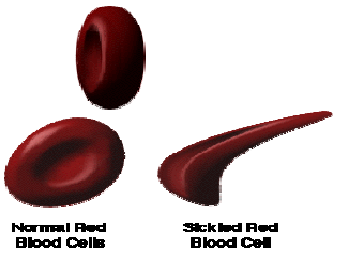


**Figure (4) Natural and sickle blood cells**

Nearly 300,000 children are born with a form of sickle cell disease every year, most of them in sub-Saharan Africa, and also in the other parts of the world such as the West Indies, and in 2013 caused to 176,000 deaths rate compared with 113,000 deaths rate in 1990 The disease occurs as a result of a change in the nucleotide bases in the DNA sequence, which , in turn changes the RNA sequence, which , in turn causes a change in protein as shown in Figure (5) (William *et al*., 2016; www.ncbi.com).

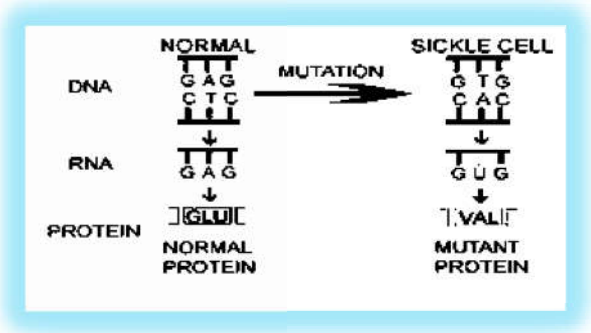


**Figure (5) Change nucleotide bases**

**Research methods and materials**

The R software that is used in connection with the National Center for Biotechnology Information (NCBI) it can be called through periodically updated the (Seqinr) package and the (HMM) package have been used packages and this research has done according to the following stages:

**Table (1) Database**

| infected | healthy |
|---|---|
| CTG ACT CCT GTG GAG AAG TCT | CTG ACT CCT GAG GAG AAG TCT |
| CTG ACT CCT CAC GAG AAG TCT | CTG ACT CCT CTC GAG AAG TCT |
| ..... | ..... |

**Database**: A database for a healthy person and another infected with sickle cell anemia has put as follows:

The histogram in the two states as shown in Figure (6)

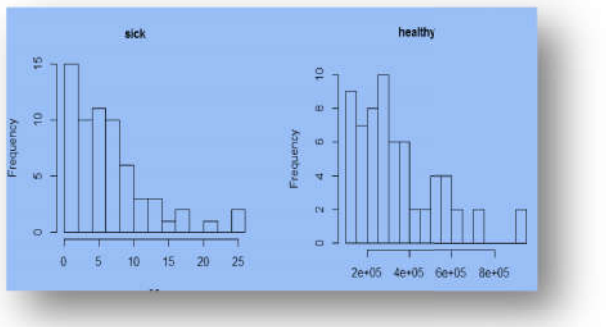


**Figure (6) histogram of Database**

The data of a healthy person is obtained from NCBI Bank by (CR541913) and the data of an infected person is also obtained from NCBI Bank by (CP012524) , then they are generated in a relating to R program , the transitive probabilities matrix is being between the observing states (sick , non- sick ) i.e. transition from GAG to GTG or from CTC to CAC as follow :

$$A = \begin{matrix} \\ SICK \\ non \quad SICK \end{matrix} \begin{pmatrix} SICK & non & SICK \\ 0.88 & & 0.12 \\ 0.10 & & 0.90 \end{pmatrix}$$

And the emission matrix

$$B = \begin{matrix} \\ SICK \\ non \quad SICK \end{matrix} \begin{pmatrix} A & C & G & T \\ 0.31 & 0.20 & 0.19 & 0.30 \\ 0.20 & 0.26 & 0.31 & 0.23 \end{pmatrix}$$

And the initial state probabilities

$\pi = \{0.25, 0.25, 0.25, 0.25\}$

**The recognition algorithm**

The following diagram (1) shows the algorithm for recognizing sickle cell anemia :

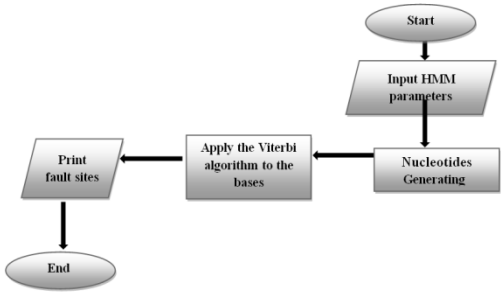


**Diagram (1) recognition algorithm**

**Stages of study and results**

**The practical application is done in two phases:**

The first aims to test the application algorithm on an experimental sample consists of 30 nucleotides that are generated by using the ready-mod function in R(Robert Gentleman& other, 2005)**:**
**Sample (nucleotides, 30, rep=TRUE)**

**It aims to know the quality of the applied algorithm.**

**The second**, we have generated 100, 200 and 300 nucleotides by using a function in R one time depending on a transitive probability matrix of a healthy person and another depending on a matrix of transitional probabilities matrix of an infected person and then using the Viterbi algorithm. The results were as in Table (2):

**Table 2. Application results**

| Recognition% | **number** | state | Number of bases |
|---|---|---|---|
| %100 | 100 | sick | **100** |
| %100 | 100 | Non-sick | |
| %100 | 100 | sick | **200** |
| %100 | 100 | Non-sick | |
| %99 | 297 | sick | **300** |
| %86.33 | 259 | Non-sick | |

The recognition locations are determined based on the Viterbi algorithm and the number of bases that equal to 300 as in Table (3):

**Table 3. location of recognizing and error**

| error | Location | state |
|---|---|---|
| 177-176 | 175-1 | sick |
| 180 | 179-178 | |
| | 300-181 | |
| 41-1 | 300-42 | Non-sick |

The results of the study have a good efficiency in the recognition of DNA sequence of a healthy person and an infected one , and the exact determination of the locations that caused the defect. thus, it can diagnose a sick state depending on the proposed algorithm. This study can also be also applied to non-genetic diseases if DNA sequence has recognized thus , This study opens fields for pathological diagnosis depending on DNA sequences then determining the locations that need treatment by knowing the corresponding RNA sequence and then the corresponding protein.

## REFERENCES

Sharma Amit Kumar and Kishor Rama 2007. "*Pattern recognition : Different available approaches* " , proceeding of National conference on challenges & opportunities in information technology (COIT2007) , Mandi Gobindrh .

Abdulla,W.H, and Kasabov, N.K. 1999. "*The concept of hidden markov models in speech recognition* " ,Dept of Knowledge engineering Lab. Dept. information science , College of engineering , University of Otage , New Zealand .

Fonzo, V., Aluffi-Pentini, F. and Parisi, V, 2007. "*Hidden Markov Models in Bioinformatics* ",Vol.2,No.1,Euro. Bio . Park, University di Roma , Roma , Italy.

Al-Khayyatt Younes, 2010. "*Markovian Modeling and its application*" , Vol.2, Almousel , Iraq .

Li xiaolin , Parizeau Mark and Plamondon Rejean, 2000. " *Training Hidden Markov Models with multiple*

Transactions on PAMI , Vol .22 , No.4,pp.371 -377.

Chen ,Q .El-Sawah , A., Joslin ,C. and Georganas , Nicolas , D. 2005. " *A dynamic gesture interface for virtual environments based on hidden markov models* " proceeding of IEEE International Workshop on Haptic Audio Visual Environments and their applications Ottawa , Ontario , Canada .

DanialJurafsky, J.H.M 2000. "*Speech and language Processing, An introduction to Natural language Processing, Computational Linguistics, and Speech Recognition.Prentice Hall*" Upper Saddle River, New Jersey 07458.

Grant,G. and Ewens ,W. 2005. " *Statistical methods in Bioinformatics* " , Second Edition , University of Pennsylvania , Philadelphia, USA .

Johanneson, P. 1999. "*Rain flow analysis of Switching Markov Loads* " , Phd thesis , Lund institute of Technology , Lund

Bilmes, J. 2002. "What HMM's Can Do " UWEE Technical Report Number UWEETR -2002-0003 .

Batzoglou, S.2010. "*Hidden Markov Models and Viterbi algorithm* " Scribed by John C.Mu .

Rabiner ,L.R.1989. "*A Tutorial on Hidden Markov Models and Selected Applications in speech recognition* " , Proceeding of IEEE, Vol.77 , No.2, pp.257-286.

Bruce, Alberts, & other 2002." *Molecular Biology of the Cell. 4th Ed*"., Garland Science. Newyork. USA.

Chris R. Calladine& other 2004. *"Understanding Dna , The Molecule & How It Works, Third Edition*" , Elsevier Ltd.

Burrus V, Waldor M 2004. "*Shaping bacterial genomes with integrative and conjugative elements*". *Res.Microbiol*376 - 86 . doi :10.1016 /j.resmic. 2004.01.012. PMID 15207870.

William C. Shiel Jr., MD, FACP, FACR 2016. " *Sickle Cell Disease (Sickle Cell Anemia)*" , http://www.medicinenet.com

Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013, www.ncbi.com

Robert Gentleman& other 2005. "*Bioinformatics and Computational Biology Solutions Using R and Bioconductor* " , Springer Science-Business Media, Inc. .

**Websites**

1.www.ncbi.com
2. www.R-blogger.com

*******