Profile Hidden Markov Model for Detection and Prediction of Hepatitis C Virus Mutation

Mohamed El Nahas[1] **Samar Kassim**[2] **Nabila Shikoun**[3]

[1]**Faculty of Engineering, Al Azhar University,**
Nasr city, Cairo, Egypt

[2]**Faculty of Medicine, Ain Shams University**
Abbassia, Cairo, Egypt

[3]**Faculty of Engineering, Al Azhar University,**
Nasr city, Cairo, Egypt

Abstract

Hepatitis C virus (HCV) is a widely spread disease all over the world. HCV has very high mutation rate that makes it resistant to antibodies. Modeling HCV to identify the virus mutation process is essential to its detection and predicting its evolution. This paper presents a model of HCV based on profile hidden Markov model (PHMM) architecture. An iterative model learning procedure is proposed and applied to both full-length sequence of virus and its very high variation (mutation) zone called NS5A. A pilot study on HCV dataset of type 4 is conducted which is of special concern in Egypt.

Keywords: *Hepatitis* ***C virus (HCV), Profile Hidden Markov Model (PHMM), Non-structure 5 A(NS5A)***

1- Introduction

Hepatitis C is a liver disease caused by the hepatitis C virus (HCV). It is widely spread disease all over the world. About 130–170 million people are chronically infected with HCV and more than 350 000 people die from hepatitis C-related liver diseases each year. HCV infection is found worldwide. Countries with high rates of chronic infection are Egypt (22%), Pakistan (4.8%) and China (3.2%). The main mode of transmission in these countries is attributed to unsafe injections using contaminated equipment. It is particularly menacing in Egypt. HCV is a major human health concern that causes fatal liver diseases. Currently no vaccine is available to prevent HCV infection [1].

The aim of this study is to identify a model of HCV genotype 4 genome using PHMM. This model shall be used for detection of HCV in blood samples. Moreover the HCV model will help in learning the mutation model of HCV. The mutation model presents new therapeutic targets as well as genomic information for designing vaccine candidates.

In this research we identify the profile hidden markov model (PHMM) from full length genomic sequences of 12 distinct HCV genotype 4. To reduce number of parameters of model (transition probability and length of the model from match states) and increase performance of the model, our approach selects a zone of HCV genome called non structure 5A (NS5A) which has more variation (mutation) than other zones in HCV genome and applied PHMM again.

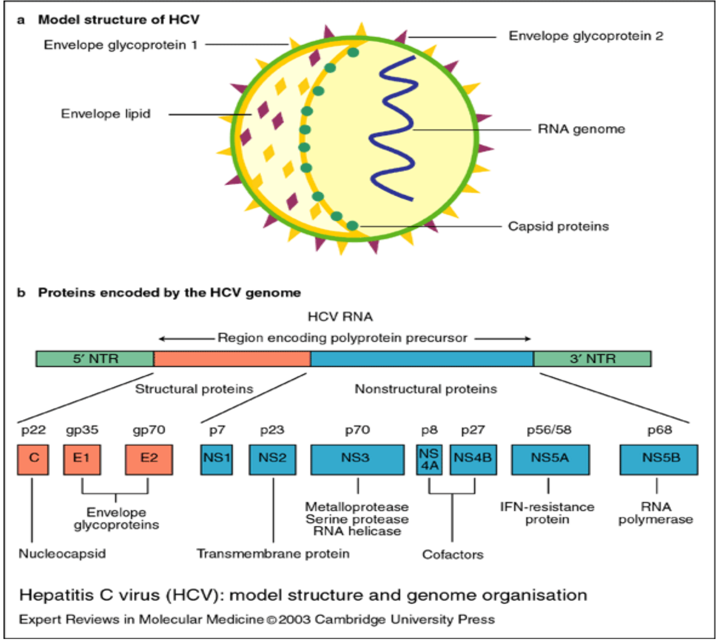
This paper is organized as follows: In section 2, is given related research for detecting hepatitis C virus. Section 3 describes HCV virus genome. Section 4 is divided into two subsections, first one presents a review of Profile Hidden Markov Model structure, and the second presents the suggested learning model. In section 5 presents data description and experimental results of PHMM. Section 6 concludes the paper with future research directions.

2- Related Research

Bioinformatics aims to improve current knowledge and understanding of biological and molecular entities. Pattern recognition and representation of motifs is a fundamental problem in bioinformatics and bioinformatics for diseases [2]. Several researches have been conducted to unravel information and useful patterns in a database for detecting hepatitis C virus. Leung et al. [3] present a data mining framework which includes molecular evolution analysis, clustering, feature selection, classifier learning and classification applied on Hepatitis B virus (HBV) DNA sequence. Jilani et al. [4] introduce an automatic diagnosis system based on neural network for Hepatitis C virus. This automatic diagnosis system deals with the mixture of feature extraction and classification. ElHefnawi et al. [5] implement a novel approach for extracting features including informative markers from mutations in the non-structural 5A protein (NS5A), specifically its Interferon sensitivity determining region (ISDR) and V3 regions, and use a novel bioinformatics approach for pattern recognition on the NS5A protein and its motifs to find biomarkers for response prediction using class association rules and comparing the Predications of the different features. Yasin et al. [6] used logistic regression model to investigate factors that contribute significantly in the classification of HCV cases. Jacob et al. [7] evaluate the performance of twenty classification algorithm on the cancerous HCV dataset that comprises individual medical cases from UCI Machine Learning Repository . Njouom et al. [8] conducted phylogenetic analyses of NS5B gene sequences from HCV-infected inhabitants of a remote area of south-west CAR which indicated that (82.8 %) were infected with (HCV-4), (8.6 %) with (HCV-2) and (8.6 %) (HCV-1). Where HCV-4 strains were highly heterogeneous, an evolutionary analysis using the coalescent approach was used to estimate the epidemic history of these HCV-4 strains.

3- HCV Virus Genome

The HCV genome is an enveloped structure approximately 50 nm in diameter. HCV is a positive single stranded enveloped RNA virus belonging to the *Flaviviridae* family with an average length of 9600 bases and carries a single, long open reading frame (ORF) flanked by 5' and 3' non-translated regions. The ORF encodes a polyprotein of ~ 3000 amino acids that is processed into three structural proteins (Envelopes 1 and 2 and p7) and six non-structural proteins named NS2-NS5B [9] as shown in Figure 1 [10].


a Model structure of HCV

b Proteins encoded by the HCV genome

Hepatitis C virus (HCV): model structure and genome organisation

Expert Reviews in Molecular Medicine ©2003 Cambridge University Press

Figure 1 Hepatitis C virus (HCV): model structure and genome organization1 [10].

HCV is classified into eleven major genotypes (designated 1-11), many subtypes (designated a, b, c, etc.), and about 100 different strains (numbered 1, 2, 3, etc.) based on the genomic sequence heterogeneity [11].

Genotypes 1-3 have a worldwide distribution. Types 1a and 1b are the most common, accounting for about 60% of global infections. They predominate in Northern Europe and North America, and in Southern and Eastern Europe and Japan, respectively. Type 2 is less frequently represented than type 1. Type 3 is endemic in south-east Asia and is variably distributed in different countries. Genotype 4 (HCV4) is principally found in the Middle East and Africa, particularly Egypt, which represent more than 90% of infections due to genotype 4 worldwide [12].

HCV has high rates of replication and mutation that promote chronicity and the development of resistance to antiviral therapy. Within an individual, viral mutations produce closely related strains called quasi-species [10].

4- Profile Hidden Markov Model

4.1 Review

Hidden Markov models have become one of the most statistically powerful methods used to model sequence alignment. A special type of left-to-right HMMs called profile HMM (PHMM) is commonly used to model multiple alignments. The architecture of PHMM was introduced by Krogh (1994) [13]. PHMM is well suited to the popular "profile" methods for searching databases using multiple sequence alignments instead of single query sequences [14]. The profile is weight matrix that for each position in a group of aligned sequences, assigns a score for each of the twenty possible amino acid residues. Each row in the profile regarded as a "match state" and the values in the row as the emission probabilities for each of the twenty possible amino acid residues. The position specific gap weights represent transition probabilities for moving to an insert or delete state from a match state [15].

HMM is formally defined as a 5 –tuple $(S, \Omega, P, \Phi, \pi)$ where

$S = \{s_1, s_2, \ldots, s_N\}$ is a finite set of N states (hidden states)

$\Omega = \{o_1, o_2, \ldots, o_N\}$ is a finite set of M possible symbols (observed states which is one of 20 amino acids)

$P = \{p_{ij}\}$is the set of state-transition probabilities, p_{ij} is the probability that the system goes from state s_i to state s_j

$\Phi = \{\varphi_i(o_k)\}$ are the observation probabilities,

$\varphi_i(o_k)$ is the probability that the symbol o_k is emitted when the system is in state s_i

$\pi = \{\pi_i\}$ are the initial state probabilities.

π_i is the probability that the system starts in state s_i

Because the states and output sequence are understood, the parameters of an HMM denote by $\lambda = (P, \Phi, \pi)$

There are three tasks solved by HMM: aligning, scoring sequences with the model and learning to estimate best parameters for the model [16].

Once a HMM drawn, the standard dynamic programming algorithms used for aligning and scoring sequences with the model. These algorithms called Forward (for scoring), Viterbi (for alignment) and Baum-Welch is used for learning [14].

Profile Hidden Markov Model (PHMM) is probably the most popular application of HMM in molecular biology for detecting remote homology between proteins. PHMM turn a multiple sequence alignment into a position-specific scoring system suitable for searching databases for remotely homologous sequences [14].

The PHMM architecture shown in Figure 2 [17], consists of match states M, deletion states D, and insertion states I, flanking states (S, N, B, E, C, T) are used for proper modeling of the ends of the sequence, either for global, local or fragment alignment of the profile. S, B, E, and T are silent (don't emission symbols), while N and C are used to insert symbols at the flanks.

A PHMM for a motif of length L contains L match states. Match state M_i emits i-th motif residue, while insertion state I_i emits background residues after this i-th residue. Each state M_i defines an emission probability distribution. M_i emits residue j with probability M_{ij}. All states I_i emit residues according to a common background distribution π. I_i emits residue j with probability $M_{ij} = \pi_j$. Deletion state D_i allows i-th motif residue to be skipped; it is non-emitting.

At each position i of a motif, there are seven allowed transitions $M_i \to M_{i+1}$, $M_i \to I_i$, $M_i \to D_{i+1}$, $I_i \to I_i$, $I_i \to M_{i+1}$, $D_i \to D_{i+1}$, and $D_i \to M_{i+1}$. A PHMM can generate a state path by first following a transition $B \to M_i$ then extending the path by transitions as described above until it reaches E following a final transition $M_i' \to E$. Match and insertion states on the path emit residues according to their emission probabilities. Flanking insert states (N and C) used for local profile alignment [14].

To force a global alignment setting the looping transition probabilities in the flanking insert states to zero.

$N \to B = C \to T = 0$ $N \to N = C \to C = 1$ [17].

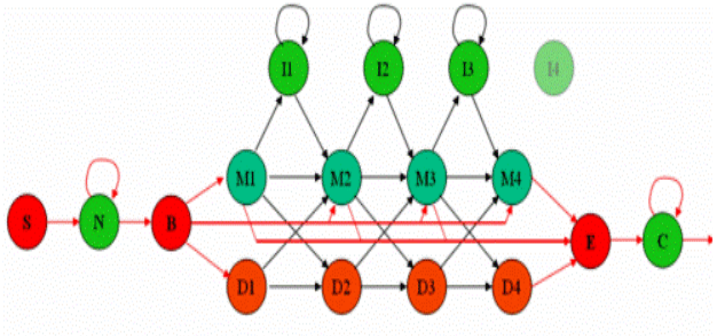


Figure 2 Architecture of PHMM [17]

4.2 Model Learning

The objective of model learning process is to estimate the parameters of PHMM from a training set. This set contains n DNA sequences each of which is labeled by its HCV genomic type. The Baum Welch algorithm is generally accepted to estimate PHMM parameters. However, this algorithm assumes that the model length is known, which not the case in this work. Hence, we have to adapt the learning procedure to search for the optimal model length. In this work we use MATLAB bioinformatics tool box functions. The learning procedure is detailed in the following steps:

1- Input training examples which consist of n sequences of full-length of 12 distinct subtypes of 4 genotype.
2- Apply Multiple Sequence Alignment (MSA) to training examples [18]. MSA performs by using a heuristic search known as progressive technique (also known as the hierarchical or tree method). The MATLAB function used is MSA=*multialign(Sequences).*
3- Preprocess data to filter unknown symbols. Sometimes, a character 'x' is found in training sequences which do not map to any of 20 amino acid, so it is replaced by one of amino acid by using MSA.
4- Initialize structure for PHMM of MSA. The initial model structure and length are defined using information derived from the alignment together with its prior knowledge of the general nature of proteins. The MATLAB function used is Model=*hmmprofstruct(Length)* .
5- Estimate the PHMM parameters from training sequences. All the parameters in the PHMM (i.e. the transition probabilities and the amino acid distributions) are estimated from a set of aligned sequences to maximize the likelihood of the observed sequences in the family. The likelihood of observed sequences is defined as:

   P(sequences | model) = P(sequence 1| model) * … * P(sequence n |model)

   The MATLAB function used is *hmmprofestimate (Model, MSA).*
6- Score the model. Scoring is used to assign a score with respect to the model to any query sequence, the better the score, the higher the chance that the query sequence is a member (homologue) of the protein family represented by the model. Scores are computed using log-odd ratios for emission probabilities and log probabilities for state transitions. The MATLAB function used is *Score = hmmprofalign(Model, Seq).*
7- Repeat steps 5 to 6 and compare the score recorded: until there is no change in score, and record length of the model and maximum score obtained from all training sequence.
8- Validate the model: Randomly select samples out of training sequences from genotype 1 to 6, and generate a set of fake sequences
9- Score the model performance based on test samples.

These steps are graphically represented as shown in Figure 3.

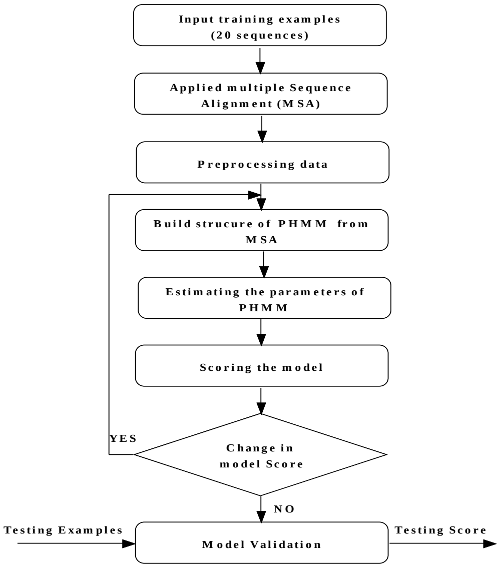


Figure 3 Model Learning of HCV

The steps for learning the model of NS5A regions are as follows:

1- Extract NS5A from training sequences, and make MSA of these sequences.
2- Initialize structure of PHMM for NS5A.
3- Estimate the parameters of a PHMM from NS5A training sequences.
4- Score the PHMM model using NS5A query sequences.
5- Repeat steps 2 to 4 and compare the score recorded until there is no change in score values.
6- To validate the model take samples out of training sequences from NS5A genotype 1 to 6, and NS5A fake sequence.
7- Score the model performance based o test samples.

These steps are graphically represented as shown in Figure 4.

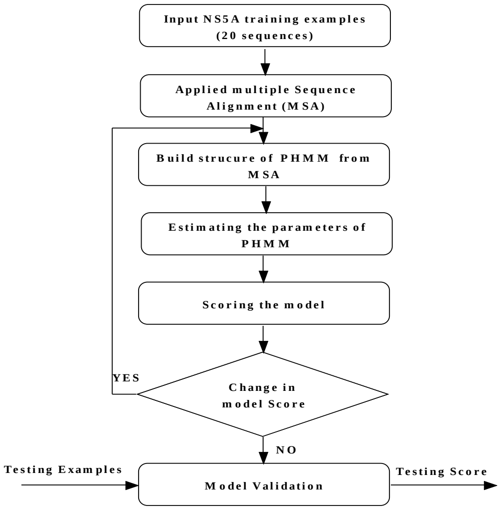

Input NS5A training examples
(20 sequences)
Applied multiple Sequence Alignment (MSA)
Build strucure of PHMM from MSA
Estimating the parameters of PHMM
Scoring the model
YES
Change in model Score
NO
Testing Examples
Model Validation
Testing Score

Figure 4 Model learning of NS5A

5- Pilot Study

The main objective of this pilot study is to identify the model of Hepatitis C Virus spread in Egypt as a first step to identify its mutation model. Genotype 4 (HCV4) is particularly principally found in Egypt, which represent more than 90% of infections worldwide [12]. For this purpose, a data set representing HCV4 is collected and used for to identify its model. Then, the learning procedure described in section 4.2 is applied on this real world data set to identify the model and evaluate its validity.

5.1 Data Description

The dataset contains the full-length genomic sequences of 20 distinct hepatitis C virus (HCV) genotype 4 isolates/subtypes 4a(7 sequences), 4f(2 sequences) , 4f(2 sequences), and one sequence for each 4l, 4t, 4n, 4o, 4k, 4b, 4m, 4p and 4g.

The resulting genomes varied between 2969 and 3011 amino acid (aa) length and each contains a single ORF.

The data is obtained from the site "http://www.ncbi.nlm.nih.gov/protein" and it presented in Table 1 which contains virus name, genebank and sequence length.

The data set was found to contain the character 'x' which undefined as amino acid, to overcome this problem and replace character 'x' with suitable amino acid followed this steps

1- Determine sequence number and positions numbers which contain character 'x' in original sequences.
2- Apply global alignment to all sequences MSA, and determine sequence number, position number of 'x' and the most character repeated
3- Replace character 'x' with suitable amino acid found in step 2

Table 1: The Training Data Set (http://www.ncbi.nlm.nih.gov/protein)

Sequence length	Genebank	Virus name
3008	ADF97233.1	hcv4a_1
3008	CAA72338.1	hcv4a_2
3008	O39929.3	hcv4a_3
3009	ABD75824.1	hcv4a_4
3008	ABD75826.1	hcv4a_5
3008	ABD75829.1	hcv4a_6
3008	ABD75830.1	hcv4a_7
3010	ABU68272.1	hcv4f_1
2969	ABU68271.1	hcv4f_2
3007	ABD75828.1	hcv4d_1
3006	ACS29436.1	hcv4d_2
3006	ACT66295.1	hcv4l
3007	ACT66294.1	hcv4t
3008	ACS29440.1	hcv4n
3005	ACS29439.1	hcv4o

3011	ACS29437.1	hcv4k
3011	ACS29434.1	hcv4b
3006	ACS29432.1	hcv4m
3007	ACS29430.1	hcv4p
3008	ACS29431.1	hcv4g

The validity of the PHMM assessed with a new test data obtained from the site indicated in Table 2.

Table 2: The Testing Data Set (http://www.ncbi.nlm.nih.gov/protein)

Virus name	Genebank	Sequence length
hcv1	NP_671491.1	3011
hcv2	YP_001469630.1	3033
hcv3	YP_001469631.1	3021
hcv4	YP_001469632.1	3008
hcv5	YP_001469633.1	3014
hcv6	YP_001469634.1	3019

5.2 Experimental Results

1- After applying the derived PHMM model to all 20 distinct full length of HCV4 subtypes, the results show that maximum score is (9611.8) at a number of match states (length of the model) 3000 as shown in Figure 5.

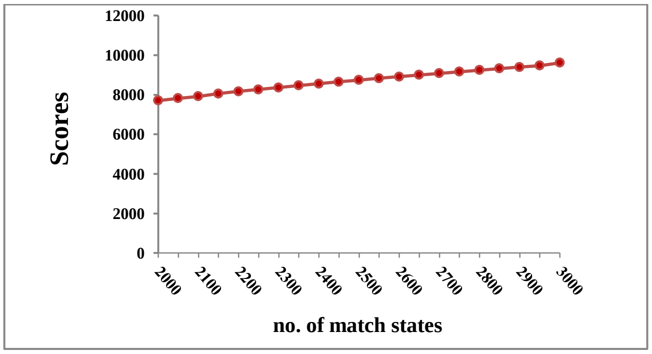


Figure 5 Relation between number of match states in PHMM and scores for HCV sequence

2- To validate PHMM model using test data which consist of fake sequence and 6 sequences of HCV with distinct types from 1 to 6. The maximum score record to HCV4 where PHMM is designed to this genotype as shown in Figure 6.

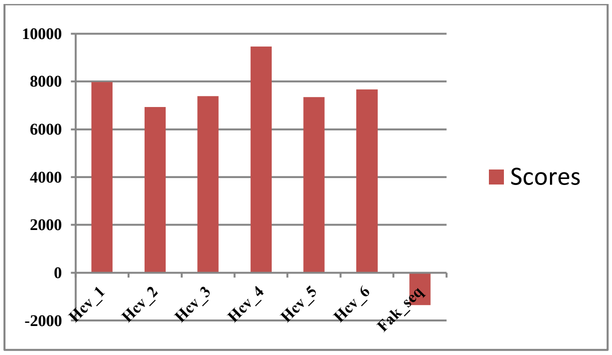


Figure 6 Relation between PHMM model and testing data for HCV sequence

3- Table 3 records the values of the score from testing data. It’s shown that the maximum value is 9453.6 marked at HCV genotype 4. And all other types of HCV marked values are less than this maximum value. This means that the PHMM model succeeded to identify HCV4 from other genotypes.

Table 3: Show percentage of classified HCV genotype

HCV_type	Score
HCV_1	**7970.5**
HCV_2	**6930.3**
HCV_3	**7388.6**
Hcv_4	**9453.6**
HCV_5	**7345.8**
HCV_6	**7664.8**
Fak_seq	**-1347.3**

4- When applying the derived PHMM model for the zone NS5A of HCV4 subtypes, the results show that maximum score is (1389.3) at a number of match states 440 as shown in Figure 7.

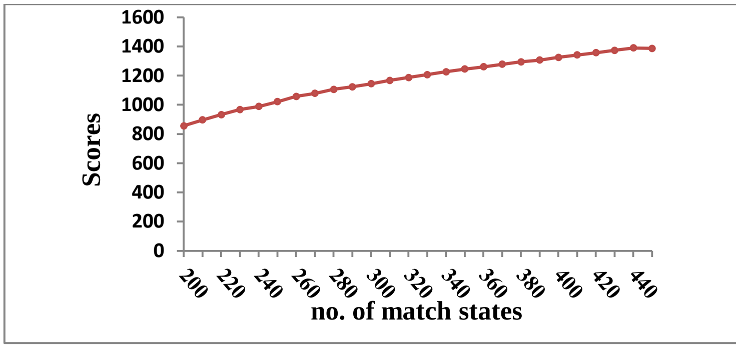


Figure 7 Relation between number of match states and scores in NS5A

5- To validate PHMM model using test data which consist of fake sequence and 6 sequences of NS5A with distinct types from 1 to 6. The maximum score record to NS5A genotype 4 as shown in Figure 8.

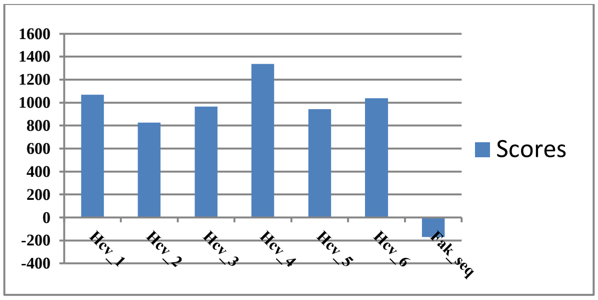


Figure 8 Relation between PHMM model and testing data for NS5A sequence

6- Table 4 records the values of the score from testing data. The maximum value is 1336.7 marked for NS5A genotype 4. And all other types of NS5A mark values much less than this maximum value. Which mean that the PHMM model has succeeded to identify NS5A4 more than other genotypes.

Table 4: Show percentage of classified NS5A genotype

NS5A_type	Score
NS5A_1	1070.2
NS5A_2	824.8
NS5A_3	964.4
NS5A_4	1336.7
NS5A_5	943.8
NS5A_6	1038.5
Fak_seq	-169.639

6. Conclusion and Future Work

In this paper, the model of the HCV has been identified. The learning process of PHMM model for the full length RNA result in a lengthy model which is prone to prediction errors and time consuming. This model could be impractical for large data set of HCV sequences. In this research we present two approaches to reduce the model length and increase its prediction accuracy. The first approach iteratively applies the Baum-Welch learning to search for the best model length. The second approach apply combine the first approach and selects the region NS5A, which has large variation in amino acid for each position in MSA. The validity of both approaches has been demonstrated in a pilot study based on real world data set. The future work shall study the impact of feature selection methods on mutation model identification and HCV detection in Egypt.

References

[1] WHO http://www.who.int/en/2012

[2] N. M. Luscombe, D. Greenbaum, M. Gerstein, “What is Bioinformatics? A Proposed Definition and Overview of the Field,” Method Inform Med, vol. 40, pp. 346–358, 2001.

[3] K. Leung, K. Lee, J. Wang, E. Y. Ng, H. L. Chan, S. K. Tsui, T. S. Mok, P. C. Tse, J. J. Sung, “Data Mining on DNA Sequences of Hepatitis B Virus,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp. 428–440, 2009.

[4] T. A. Jilani, H. Yasin, M. M. Yasin, “PCA-ANN for Classification of Hepatitis C Patients,” International Journal of Computer Applications (0975 – 8887), vol. 14, no. 7, 2011.

[5] M. M. ElHefnawi, S. Zada and I. A. El-Azab, ”Prediction of prognostic biomarkers for Interferon-based therapy to Hepatitis C Virus patients: a meta-analysis of the NS5A protein in subtypes 1a, 1b, and 3a,” Virology Journal, vol.7, pp.130-138, 2010.

[6] H. Yasin, T. A. Jilani, M. Danish , “Hepaiitis-C Classification using Data Mining Techniques,” International Journal of Computer Applications (0975 – 8887), vol. 24, no. 3, 2011.

[7] S. G. Jacob, p. Nancy, R. G. Ramani, “Efficient classifier for classification of HCV clinical data through data mining algorithms and techniques,” ICCA, pp. 105-110, 2012.

[8] R. Njouom , E. Frost, S. Deslandes, F. M. Yaya, A. C. Labbe´, R. Pouillot, P. M. lesso, S. Mbadingai, D. Rousset, J. Pe´ pin, “Predominance of hepatitis C virus genotype 4 infection and rapid transmission between 1935 and 1965 in the Central African Republic,” Journal of General Virology, vol. 90, pp. 2452–2456, 2009.

[9] C. Combet, N. Garnier, C. Charavay, D. Grando, D. Crisan, J. Lopez, A. Dehne-Garcia, C. Geourjon, E. Bettler, C. Hulo, P. Le Mercier, R. Bartenschlager, H. Diepolder, D. Moradpour, J.M. Pawlotsky, C. M. Rice, C. Tre´po, F. Penin, G. Dele´age ., "euHCVdb: the European hepatitis C virus database Nucleic," Acids Research, vol. 35, 2007.

[10] http://www.expertreviews.org/

[11] A. J. Pérez-Berná, J. Guillén,M. R. Moreno, A. I. Gómez-Sánchez, G. Pabst, P. Laggner, J. Villalaín, "Glycoprotein with Membranes. Biophysical Characterization," Biophysical Journal vol. 94, pp. 4737–4750, 2008.

[12] S. M. Kamal and I. A. Nasser, "Hepatitis C Genotype 4: What We Know and What We Don’t Yet Know," HEPATOLOGY, vol.47, pp.1371-1383, 2008.

[13] A. Krogh, M. Brown, I.S. Mian, K. Sjölander, P. Haussler, “Hidden Markov Models in Computational Biology,” J. Mol. Bio. vol. 235, pp. 1501-1531, 1994.

[14] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.

[15] S. R. Eddy “Profile hidden Markov models,” Bioinformatics review, vol. 14, no. 9, pp. 755-763, 1998.

[16] O. G. Ibe, Markov processes for stochastic modeling. Elsevier Academic press, 2009.

[17] http://www.mathworks.com/2012

[18] C. Dewey, “Multiple Sequence Alignment : Task Definition,” 2011.