International Journal on Emerging Technologies **11**(1): 270-279(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Estimated Value of Hidden Markov Model Parameters for NS5 methyltransferase Protein of Dengue Virus

Nidhi Katiyar[1], Ravindra Nath[2] and Shashwat Katiyar[3]
[1]Dr. APJ Abdul Kalam Technical University (AKTU), Lucknow India.
[2]University Institute of Engineering, Technology, CSJM University Kanpur India.
[3]Institute of Bioscience and Biotechnology, CSJM University Kanpur India.

(Corresponding author: Nidhi Katiyar)
(Received 19 October 2019, Revised 19 December 2019, Accepted 24 December 2019)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In these days Dengue disease has become a serious health problem in all over the world. The dengue research also aims to provide better observation to limit the effect of dengue epidemics. Since several years, there is no good dengue vaccine available till now that can completely cure a patient but researches are going on to develop the good vaccine. In this study, we are targeting the NS5 Methyltransferase protein of dengue virus. It is the largest and most conserved protein of the dengue virus. Here, we present data of the influence of the methyltransferase domain (NS5-MMTase) of dengue virus (DENV). We have taken the complete data of (accession code) Methyltransferase protein, statistically consider each protein and apply a hidden Markov model approach (forward algorithm) to find which gene sequence is better for further analysis in bioinformatics field. In this study, we converted the protein sequence into the gene sequence and then calculated the probability of each character. In the second step, we draw the matrix with the help of HMM parameters and find the estimated value of observation sequence i.e. the value of P (O|λ). Although, most of the machine learning approaches were used in previous literature. But in this work we focus on the HMM technique because this is the mathematically strong and give the accurate result in most of cases. So, in the future we can use other machine learning, deep learning approaches on the bases of gene sequence and huge datasets of genome.

Keywords: Dengue virus, Methyltransferase protein, NS5 protein, Hidden Markov model.

Abbreviations: DENV, Dengue virus; NS, Non-structural protein; MTase, Methyltransferase; RdRp, RNA-dependent RNA-polymerase; HMM, Hidden Markov Model; PDB, Protein data bank.

I. INTRODUCTION

Dengue virus (DENV) is the cause of dengue fever. It is a mosquito-borne single positive-stranded RNA virus of the family *Flaviviridae*; genus *Flavivirus* [1, 2]. Five serotypes of the virus have been found, all of which can cause the full spectrum of disease [3, 4]. Nevertheless, scientists are finding their understanding of dengue virus may be simplistic, as rather than distinct antigenic groups there appears to be a continuum [5]. This same study identified 47 strains of dengue virus. Its genome is about 11000 bases of positive-sense single stranded RNA (ssRNA) that codes for three structural proteins (capsid protein C, membrane protein M, envelope protein E) and seven nonstructural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, NS5) [8]. It also includes short non-coding regions on both the 5' and 3' ends. Like the dengue virus fever cancer is one of the diseases where number of cases is increased year on year throughout in the India and all worlds. Cancer Statistics in United States of America says that in year of 2018, over 17 lakh new cases of cancer will be diagnosed. Among all the cancer disease types, motely in women, breast cancer is diagnosed most compared to other types of cancers like lung or cervix cancer etc. The situation is almost same in the India and in United States of America and other country [10].

In the last few decades, Dengue has extended globally. Researchers are identifying and studying the dengue viruses to understand the factors responsible for transmitting the viruses to humans. They are investigating the mechanism by which dengue virus replicates itself and also exploring the structure of the viral components, such as the Capsid membrane (C), and envelope proteins (E) [6]. On the other hand, Dengvaxia a tetravalent vaccine was developed by Sanofi Pasteur which was consists of genes encoding the premembrane (prM) and E proteins of dengue virus (DENV) serotypes 1-4 (DENV 1-4) [6-7].

NS5 methyltransferase domain is the largest and most conserved region (Fig. 1). It is also a bi-functional enzyme with a methyltransferase domain (MTase; residues 1–296) at its N-terminal end and a RNA-dependent RNA polymerase (RdRp; residues 320–900) at its C-terminal end [7-8]. A standard computational and other experimental drug design is used for lead compound optimization and in virtual screening studies to find novel biologically active molecules [11].

Dengue virus is an RNA virus that has very high mutation frequency with an average of 100 times higher than the genomic DNA mutations. The accumulation of mutations is a continuous process, together with the possibility of intermolecular recombination for their simultaneous infection with a different serotype that could cause the appearance of a new serotype of dengue virus [9].

NS5 is a large protein contains 67% amino acid similarity with DENV 1-4 and has multifunctional protein with several enzymatic activities [12].

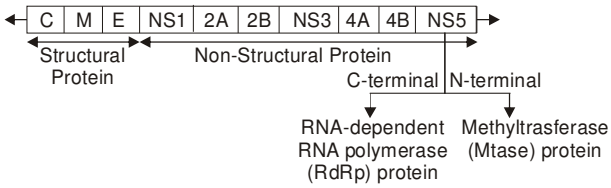


Fig. 1. The 5' and 3' untranslated regions and the arrangement of the genes encoding 3 structural and 7 nonstructural (NS) proteins are shown.

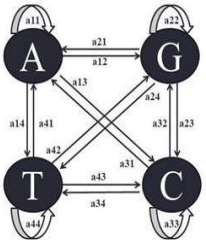
Here, we summarize recently published work on the NS5 protein, a very important enzyme drug target. NS5 is the largest (102 kDa) and the most conserved protein (with ~ 70% sequence identify) among the four serotypes expressed during infection by dengue virus [13]. The enzymes informed byNS5 show they have important functions in the replication of the virus, playing a key role and suggesting NS5 as a potential antiviral target. In this paper, we used NS5 MTase protein sequences and apply the HMM to find the estimated result. We are using this hidden markov model because it's frequently used in the Bioinformatics. In future work we will used other machine learning approaches like Genetic Algorithm (GAs), Metropolis, and Deep Learning Method etc. In general the main motivation for using GAs in the discovery of high-level prediction rules and find global maxima is that they perform a global search and handle better with attribute interaction than the greedy rule induction algorithms and dynamic programming often used in data mining etc. and other techniques [14]. Nowadays, the rapid progress in next generation sequencing technology and the simultaneous development of bioinformatics tools have allowed exploring novel genes coding for valuable enzymes or proteins [15].

Hidden Markov Model: Hidden Markov Model is the emerging theory and statically model [16]. The basic Markov Model is the process where each state corresponds to an observable occurrence and the state transition probabilities depend only on the current and precursor state. Hidden Markov Model is extended application for more complex processes, like speech recognition, bioinformatics and computational gene finding and forecasting is defined as technique of translating past experience into prediction of new things which come in future [17-18]. It tries to evaluate the magnitude and significance of forces and efforts that will affect future operating conditions in an enterprise. So, the estimation of type, quantity and quality of future work is termed as Forecasting [19]. A general Hidden Markov Model (HMM) consists of a finite set of states, an alphabet of output symbols, a set of state transition probabilities and a set of emission probabilities [20]. The emission probabilities specify the distribution of output symbols that may be emitted from each state. Therefore hidden markov model, are two stochastic processes; the process of moving between states and the second process of emitting an output sequence. The sequence of state transitions is a hidden process and is observed through the sequence of emitted symbols.

Here, we have defined the HMM parameters for solving this problem and applied this mathematical method on the bases of gene sequence. In first step we selected the MTase protein for computational analysis and we take the input of twenty protein sequence from PDB database. In this study, using the HMM approach to find the probability of DNA components of i.e. A, G, C, T nucleotides.

In bioinformatics, it has been used in sequence alignment, In-silico gene detection, structure prediction, data-mining literature, and so on [20]. Our problem has taken under the category of In-silico gene detection. In previous paper we have study the three main types of problems of Hidden Markov Model [21].

Hidden Markov Model is the probabilistic model based on probability theory for the observed sequence to find the hidden states and to make the discrete Markov chain like that ($O = O_1, O_2, O_3, \_ \_ \_ , O_T$). A Markov chain with 4 states (labeled A, G, C, T nucleotides to code the gene sequence) with selected transitions has been shown in (Fig. 2). We assume here, state A as 1, G as 2, C as 3 and T as 4. Here present the state transition function from the equation of HMM.



	A	G	C	T
A	a11	a12	a13	a14
G	a21	a22	a23	a24
C	a31	a32	a33	a34
T	a41	a42	a43	a44

A to A = a_{11}
A to G = a_{12}
A to C = a_{13}
A to T = a_{14}
.
.
.
T to T = a_{44}

Fig. 2. A Markov chain with 4 states of A, G, C, T characters and its transitions state.

An HMM is characterized by the following:

1. N, the number of states in the model.
2. M, the number of district observation symbols in each state.
3. The state transition probability distribution A = a_{ij}, where

$A = a_{ij} = P[q_t = S_j | q_{t-1} = S_i]$,

$1 \le i, j \le N a_{ij} \ge 0$ (1)

$\sum_{j=1}^{N} a_{ij} = 1$ (2)

4. The observation symbol probability distribution B = $\{b_i(k)\}$, where

$B = \{b_i(k)\} = P[V_k \text{ at } t | q_t = S_i]$,

$\le j \le N1 \le k \le M$ (3)

5. The initial state distribution $\pi = [\pi_i]$, where

$\pi_i = P[q_t = S_i]$,

$\le i \le N$ (4)

6. O is the observation, where T is the number of observation,

$O = O_1 O_2 \ldots\ldots\ldots O_T$ (5)

Given appropriate parameters of N, M, A, B and π, the HMM can be used as generator to given an observation sequence.

The specification of the three probability measures A, B, and π. For convenience, we use the compact notation.

$\boldsymbol{\lambda} = (A, B, \pi)$ (6)

Using forward method of HMM,

Initialization:

$\alpha_1(i) = \pi_i b_i (O_1)$,

$1 \le i \le N$.

Induction:

$\alpha_{t+1}(j) = \left[\sum_{j=1}^{N} \alpha_t(i) a_{ij}\right] b_j (O_{t+1})$,

$1 \le t \le T - 1 \quad 1 \le j \le N$.

Termination:

$P(O|\lambda) = \sum_{j=1}^{N} \alpha_T (i)$ (7)

So, $P(O|\boldsymbol{\lambda})$ is called the probability of observation sequence of model $\boldsymbol{\lambda}$.

II. METHODOLOGY

In this research, we are working on the NS5 MTase protein on the basis of resolution power of protein sequence. Selected protein sequences are fetched from protein data bank with all accession codes of NS5 MTase protein of dengue virus. When we are found the new candidate of drug then we take that protein, which has the highest resolution power (< 2.00 Å) in the crystallization structure. But in this methodology we will decide the target protein with the help of Hidden Markov Model approach.

In first step we used the online computational method to convert protein sequence into gene sequence. After that found the value of (A, B, π) parameters then we are using the forward algorithm of HMM for finding the value of probability P(O|λ) of observation sequence (i.e. gene sequence). In (Fig. 3) illustrates the flow chart, how we find the estimated value of P(O|λ) of a gene sequence of NS5 Methyltransferase protein of dengue virus.

Data availability: Coordinates and structure factors for DENV-2 NS5 Methyltransferase protein complexes are available in the Protein Data Bank (PDB). To retrieve the sequence of NS5 Methyltransferase (accession codes) of 1L9K, 1R6A, 2P1D, 2P3L, 2P3O, 2P3Q, 2P40, 2P41, 3EVG, 4VOQ, 5EIF, 4R05, 3MTE, 5E9Q, 3P8Z, 5EC8, 5EKX, 3P97, 5CUQ and 3MQ2 from RCSB protein data bank. In Table 1 showing the data of protein sequence which used as target pdb code, resolution power, residues count (length of sequence in amino acids) and last these references which used in previous literature [22-31].

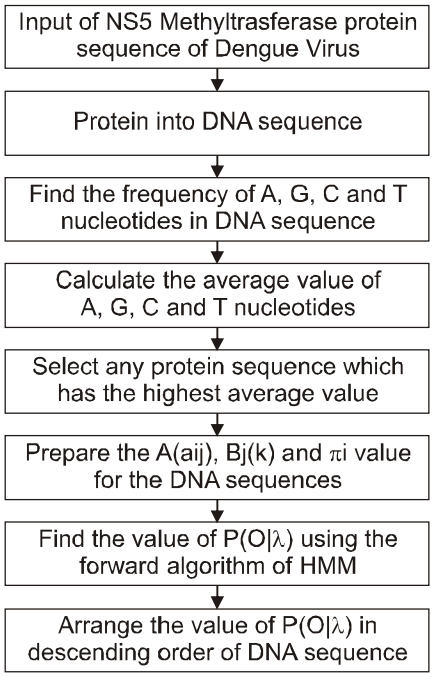


Fig. 3. Flowchart of Hidden Markov Model approach.

Table 1: List of the Dengue NS5 Methyltransferase protein sequence into DNA sequence for using online computational method.

S.No.	Pdb code	Resolution Power	Residues Count	References
1.	2P1D	2.90 Å	305	Egloff *et al.,* (2002) [22]
2.	2P3O	2.76 Å	305	Egloff *et al.,* (2007) [23]
3.	2P3Q	2.75 Å	305	Egloff *et al.,* (2007) [23]
4.	2P40	2.70 Å	305	Egloff *et al.,* (2007) [23]
5.	1R6A	2.60 Å	295	Benarroch *et al.,* (2004) [24]
6.	1L9K	2.40 Å	305	Egloff *et al.,* (2002) [22]
7.	4V0Q	2.30 Å	892	Zhao *et al.*, (2015) [25]
8.	3EVG	2.20 Å	275	Geiss *et al.*, (2009) [26]
9.	2P3L	2.20 Å	305	Egloff *et al.,* (2007) [23]
10.	5EIF	2.00 Å	552	Benmansour *et al.*, (2017) [27]
11.	4R05	2.00 Å	534	Brecher *et al.*, (2015) [28]
12.	3MTE	1.80 Å	444	Macmaster *et al.*, (2010) [29]
13.	2P41	1.80 Å	305	Egloff *et al.,* (2007) [23]
14.	5E9Q	1.79 Å	512	Benmansour *et al.*, (2017) [27]
15.	3P8Z	1.79 Å	534	Lim *et al.*, (2011) [30]
16.	5EC8	1.76 Å	552	Benmansour *et al.*, (2017) [27]
17.	5EKX	1.76 Å	552	Benmansour *et al.*, (2017) [27]
18.	3P97	1.74 Å	534	Lim *et al.*, (2011) [30]
19.	5CUQ	1.74 Å	534	Brecher *et al.*, (2015) [28]
20.	3MQ2	1.69 Å	436	Macmaster *et al.*, (2010) [29]

III. RESULTS AND DISCUSSION

In this work, we target the DENV NS5 MTase protein sequences. These selected MTase proteins are target of DENV-2 virus. Protein sequence has been converted into the DNA sequence. Reverse-translation of the protein sequence into DNA sequence using the online computational method (Translate DNA and RNA sequences to protein sequences) has been done. Now we have counted the A, G, C and T nucleotides and also show the probability of each A, G, C and T nucleotide in the given DNA sequence. (Table 2)

From (Table 2), find the result that DNA sequence corresponding the protein sequence highest probability of nucleotide A or average probability of each nucleotide. Now we apply the forward algorithm of Hidden Markov Model technique to find the HMM parameters (λ= A= a_{ij}, B= $b_j(k)$ and π= π_i) of each DNA sequence. Below the Table showing the matrix data are filled by the A, G, C and T counts in each gene sequence.

Table 2: A value of the A, G, C and T nucleotides in DNA sequence.

S.No.	PDB code	DNA Sequences	A	G	C	T	Total no. of nucleotides	Probability of each nucleotides (A, G, C, T)			
1.	2P1D	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
2.	1R6A	GGAA.....GACA	175	171	66	102	514	2.93	3.00	7.78	5.03
3.	3EVG	ATGG.....CACA	162	159	63	97	481	2.93	3.02	7.63	4.95
4.	3P8Z	GGCC.....GACG	308	314	128	186	936	3.08	2.98	7.31	5.03
5	1L9K	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
6	2P3L	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
7	2P3O	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
8	2P3Q	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
9	2P40	ATGG.....GACA	162	159	63	97	481	2.93	3.02	7.63	4.95
10	2P41	ATGG.....GACA	182	175	70	103	530	2.91	3.02	7.57	5.14
11	5EKX	GGAC.....GTAT	305	351	108	216	980	3.21	2.79	9.07	4.53
12	5EIF	GGAC.....GTAT	303	347	147	183	980	3.23	2.82	6.66	5.35
13	5EC8	GGAC.....GTAT	303	347	147	183	980	3.23	2.82	6.66	5.35
14	5E9Q	GAAC.....GACG	343	317	109	167	936	2.72	2.95	8.58	5.60
15	5CUQ	GGCC.....GACG	343	317	109	167	936	2.72	2.95	8.58	5.60
16	4V0Q	ATGG.....GAGA	557	501	221	346	1625	2.91	3.24	7.35	4.69
17	3P97	GGCC.....GACG	308	314	128	186	936	3.08	2.98	7.31	5.03
18	3MTE	ATGG.....AACA	286	198	82	189	755	2.63	3.81	9.20	3.99
19	3MQ2	ATGG.....ATCC	162	286	146	146	740	4.56	2.58	5.06	5.06
20	4R05	GGCC.....GACG	308	314	128	186	936	3.08	2.98	7.31	5.03

Sequence 1 with PDB id: 2P1D

(a) $A_{2P1D}=a_{ij}$

(b) $B_{2P1D}= b_{jk}$

(c) $\Pi i_{(2P1D)}$= [0.6, 0.2, 0.1, 0.1]

(a)

	A	G	C	T
A	0.390	0.219	0.176	0.214
G	0.299	0.428	0.074	0.198
C	0.436	0.140	0.253	0.169
T	0.259	0.538	0.057	0.144

(b)

	A	G	C	T
A	0.360	0.284	0.315	0.040
G	0.096	0.130	0.320	0.453
C	0.190	0.093	0.226	0.490
T	0.301	0.296	0.210	0.192

Sequence 2 with PDB id: 2P3O

(a) A_{2P3O}=aij

(b)B_{2P3O}=bjk

(c) $\Pi i_{(2P3O)}$ = [0.5, 0.2, 0.1, 0.2]

(a)

	A	G	C	T
A	0.306	0.247	0.2	0.246
G	0.355	0.315	0.094	0.236
C	0.460	0.150	0.207	0.182
T	0.271	0.565	0.061	0.103

(b)

	A	G	C	T
A	0.366	0.225	0.194	0.215
G	0.342	0.352	0.081	0.225
C	0.520	0.130	0.195	0.155
T	0.274	0.571	0.055	0.1

Sequence 3 with PDB id: 2P3Q

(a) A_{2P3Q}=aij

(b) B_{2P3Q}=bjk

(c) $\Pi i_{(2P3Q)}$ = [0.5, 0.3, 0.1, 0.1]

(a)

	A	G	C	T
A	0.310	0.245	0.2	0.245
G	0.356	0.317	0.090	0.237
C	0.461	0.150	0.210	0.179
T	0.270	0.566	0.063	0.101

(b)

	A	G	C	T
A	0.370	0.222	0.198	0.210
G	0.340	0.354	0.079	0.227
C	0.519	0.129	0.194	0.157
T	0.270	0.577	0.053	0.1

Sequence 4 with PDB id: 2P40

(a) A_{2P40}=aij

(b)B_{2P40}=bjk

(c) $\Pi i_{(2P40)}$ = [0.5, 0.3, 0.1, 0.1]

(a)

	A	G	C	T
A	0.311	0.250	0.2	0.239
G	0.355	0.320	0.090	0.235
C	0.461	0.152	0.207	0.180
T	0.272	0.565	0.060	0.103

(b)

	A	G	C	T
A	0.371	0.220	0.198	0.211
G	0.343	0.352	0.080	0.225
C	0.520	0.131	0.194	0.155
T	0.275	0.570	0.055	0.1

Sequence 5 with PDB id: 1R6A

(a) $A_{1R6A}=a_{ij}$

(b) $B_{1R6A} = b_{jk}$

(c) $\Pi i_{(1R6A)}$ = [0.3, 0.6, 0.25, 0.75]

(a)

	A	G	C	T
A	0.405	0.222	0.156	0.216
G	0.307	0.412	0.077	0.203
C	0.373	0.164	0.283	0.179
T	0.262	0.533	0.058	0.146

(b)

	A	G	C	T
A	0.073	0.262	0.262	0.365
G	0.500	0.024	0.131	0.345
C	0.500	0.222	0.28	0.249
T	0.306	0.600	0.68	0.26

Sequence 6 with PDB id: 1L9K

(a) A_{1L9K}=aij

(b) B_{1L9K}=bjk

(c) $\Pi i_{(1L9K)}$ = [0.5, 0.2, 0.1, 0.2]

(a)

	A	G	C	T
A	0.308	0.250	0.2	0.242
G	0.355	0.320	0.090	0.235
C	0.462	0.149	0.210	0.179
T	0.272	0.565	0.058	0.105

(b)

	A	G	C	T
A	0.370	0.220	0.196	0.214
G	0.340	0.350	0.080	0.230
C	0.519	0.127	0.194	0.160
T	0.270	0.575	0.055	0.1

Sequence 7 with PDB id: 4V0Q

(a) A_{4V0Q}=aij

(b) B_{4V0Q}=bjk

(c) $\Pi i_{(4V0Q)}$ = [0.5, 0.3, 0.1, 0.1]

(a)

	A	G	C	T
A	0.276	0.262	0.194	0.266
G	0.356	0.294	0.128	0.219
C	0.437	0.223	0.214	0.124
T	0.264	0.519	0.084	0.131

(b)

	A	G	C	T
A	0.289	0.261	0.189	0.259
G	0.375	0.290	0.118	0.214
C	0.449	0.222	0.210	0.117
T	0.488	0.531	0.081	0.128

Sequence 8 with PDB id: 3EVG

(a) A_{3EVG}=aij

(b) B_{3EVG}=bjk

(c) $\Pi i_{(3EVG)}$ = [0.5, 0.2, 0.1, 0.2]

(a)

	A	G	C	T
A	0.417	0.202	0.166	0.214
G	0.282	0.429	0.077	0.211
C	0.376	0.187	0.265	0.171
T	0.265	0.520	0.061	0.153

(b)

	A	G	C	T
A	0.346	0.285	0.202	0.165
G	0.290	0.134	0.225	0.345
C	0.255	0.292	0.234	0.218
T	0.395	0.072	0.21	0.313

Sequence 9 with PDB id: 2P3L

(a) A_{2P3L}=aij

(b) B_{2P3L}=bjk

(c) $\Pi i_{(2P3L)}$ = [0.6, 0.2, 0.1, 0.1]

(a)

	A	G	C	T
A	0.310	0.250	0.2	0.240
G	0.357	0.316	0.090	0.237
C	0.463	0.147	0.210	0.180
T	0.270	0.566	0.061	0.103

(b)

	A	G	C	T
A	0.371	0.219	0.195	0.215
G	0.342	0.350	0.078	0.230
C	0.520	0.129	0.195	0.156
T	0.272	0.573	0.055	0.1

Sequence 10 with PDB id: 5EIF

(a) A_{5EIF}=aij

(b) B_{5EIF}=bjk

(c) $\Pi i_{(5EIF)}$ = [0.3, 0.5, 0.1, 0.1]

(a)

	A	G	C	T
A	0.295	0.228	0.206	0.271
G	0.348	0.31	0.069	0.262
C	0.379	0.243	0.276	0.102
T	0.296	0.521	0.055	0.128

(b)

	A	G	C	T
A	0.305	0.24	0.195	0.26
G	0.370	0.335	0.058	0.237
C	0.391	0.238	0.272	0.099
T	0.288	0.531	0.054	0.127

Sequence 11 with PDB id: 4R05

(a) A_{4R05}=aij

(b) B_{4R05}=bjk

(c) $\Pi i_{(4R05)}$ = [0.1, 0.6, 0.2, 0.1]

(a)

	A	G	C	T
A	0.288	0.232	0.211	0.269
G	0.348	0.334	0.077	0.241
C	0.372	0.240	0.275	0.113
T	0.291	0.529	0.057	0.123

(b)

	A	G	C	T
A	0.300	0.245	0.198	0.257
G	0.342	0.357	0.069	0.232
C	0.382	0.234	0.295	0.089
T	0.285	0.535	0.054	0.126

Sequence 12 with PDB id: 3MTE

(a) A_{3MTE}=aij

(b) B_{3MTE}=bjk

(c) $\Pi i_{(3MTE)}$ = [0.5, 0.3, 0.25, 0.75]

(a)

	A	G	C	T
A	0.330	0.237	0.116	0.315
G	0.417	0.208	0.164	0.208
C	0.365	0.268	0.170	0.195
T	0.372	0.344	0.045	0.237

(b)

	A	G	C	T
A	0.339	0.248	0.111	0.300
G	0.414	0.212	0.171	0.202
C	0.348	0.302	0.162	0.186
T	0.378	0.329	0.043	0.248

Sequence 13 with PDB id: 2P41

(a) A_{2P41}=aij

(b) B_{2P41}=bjk

(c) $\Pi i_{(2P41)}$ = [0.5, 0.3, 0.1, 0.1]

(a)

	A	G	C	T
A	0.307	0.250	0.2	0.243
G	0.355	0.320	0.090	0.235
C	0.462	0.150	0.208	0.180
T	0.271	0.566	0.061	0.102

(b)

	A	G	C	T
A	0.368	0.222	0.196	0.212
G	0.340	0.352	0.079	0.227
C	0.519	0.129	0.194	0.157
T	0.272	0.572	0.054	0.1

Sequence 14 with PDB id: 5E9Q

(a) A_{5E9Q}=aij

(b) B_{5E9Q}=bjk

(c) $\Pi i_{(5E9Q)}$ = [0.2, 0.6, 0.1, 0.1]

(a)

	A	G	C	T
A	0.295	0.226	0.207	0.272
G	0.340	0.342	0.063	0.255
C	0.379	0.241	0.276	0.104
T	0.296	0.523	0.056	0.125

(b)

	A	G	C	T
A	0.311	0.24	0.199	0.25
G	0.373	0.330	0.057	0.237
C	0.386	0.226	0.267	0.096
T	0.285	0.533	0.053	0.127

Sequence 15 with PDB id:3P8Z

(a) A_{3P8Z}=aij

(b) B_{3P8Z}=bjk

(c)$\Pi i_{(3P8Z)}$ = [0.1, 0.5, 0.3, 0.1]

(a)

	A	G	C	T
A	0.291	0.230	0.211	0.268
G	0.345	0.334	0.080	0.241
C	0.373	0.241	0.275	0.111
T	0.292	0.528	0.055	0.125

(b)

	A	G	C	T
A	0.303	0.245	0.196	0.256
G	0.340	0.359	0.069	0.232
C	0.391	0.237	0.289	0.083
T	0.283	0.537	0.054	0.126

Sequence 16 with PDB id: 5EC8

(a) A_{5EC8}=aij

(b)B_{5EC8}=bjk

(c) $\Pi i_{(5EC8)}$ = [0.2, 0.5, 0.2, 0.1]

(a)

	A	G	C	T
A	0.296	0.225	0.206	0.273
G	0.343	0.340	0.067	0.250
C	0.380	0.242	0.275	0.103
T	0.294	0.520	0.058	0.128

(b)

	A	G	C	T
A	0.305	0.24	0.195	0.26
G	0.375	0.330	0.057	0.238
C	0.396	0.232	0.273	0.099
T	0.288	0.530	0.055	0.127

Sequence 17 with PDB id: 5EKX

(a) A_{5EKX}=aij

(b) B_{5EKX}=bjk

(c) $\Pi i_{(5EKX)}$ = [0.5, 0.2, 0.2, 0.1]

(a)

	A	G	C	T
A	0.290	0.236	0.209	0.263
G	0.356	0.315	0.082	0.246
C	0.378	0.257	0.272	0.090
T	0.306	0.519	0.054	0.120

(b)

	A	G	C	T
A	0.303	0.245	0.199	0.253
G	0.350	0.335	0.075	0.240
C	0.388	0.240	0.287	0.085
T	0.298	0.528	0.051	0.123

Sequence 18 with PDB id: 3P97

(a) A_{3P97}=aij

(b) B_{3P97}=bjk

(c) $\Pi i_{(3P97)}$ = [0.1, 0.6, 0.2, 0.1]

(a)

	A	G	C	T
A	0.291	0.230	0.211	0.268
G	0.347	0.333	0.078	0.241
C	0.370	0.241	0.274	0.112
T	0.292	0.528	0.057	0.123

(b)

	A	G	C	T
A	0.300	0.245	0.198	0.257
G	0.340	0.357	0.079	0.233
C	0.380	0.230	0.292	0.098
T	0.285	0.535	0.054	0.126

Sequence 19 with PDB id: 5CUQ

(a) A_{5CUQ}=aij

(b) B_{5CUQ}=bjk

(c) $\Pi i_{(5CUQ)}$ = [0.2, 0.6, 0.1, 0.1]

(a)

	A	G	C	T
A	0.289	0.231	0.210	0.270
G	0.345	0.332	0.080	0.243
C	0.372	0.242	0.274	0.112
T	0.293	0.528	0.055	0.124

(b)

	A	G	C	T
A	0.302	0.243	0.197	0.256
G	0.339	0.358	0.069	0.232
C	0.378	0.227	0.287	0.083
T	0.286	0.538	0.050	0.126

Sequence 20 with PDB id: 3MQ2

(a) A_{3MQ2}=aij

(b) B_{3MQ2}=bjk

(c) $\Pi i_{(3MQ2)}$ = [0.3, 0.2, 0.1, 0.4]

(a)

	A	G	C	T
A	0.194	0.337	0.207	0.259
G	0.284	0.261	0.223	0.230
C	0.237	0.388	0.244	0.129
T	0.112	0.605	0.112	0.169

(b)

	A	G	C	T
A	0.192	0.349	0.216	0.240
G	0.275	0.262	0.227	0.234
C	0.224	0.408	0.244	0.122
T	0.106	0.626	0.106	0.16

Using HMM parameters (**λ**= A, B, π) and applying forward algorithm computational method. Computational results are showing in Fig. 4.

Output file

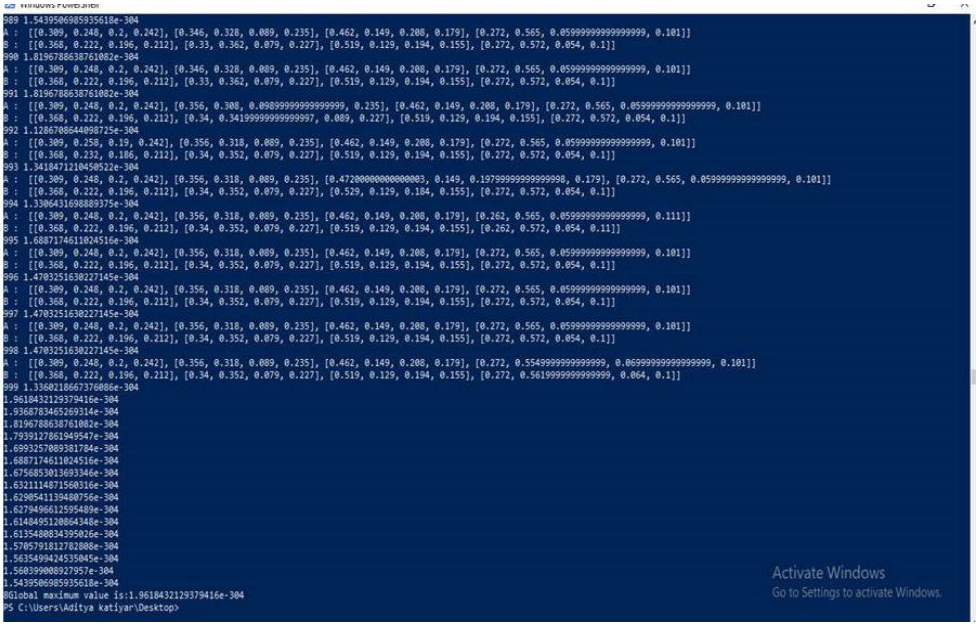


Fig. 4. Output of all possible confirmations of P(O| **λ**).

In above, Fig. 4 showing the computational result via forward algorithm. Using HMM parameters we simplify the estimated value of observation sequence. In back of this figure we write the forward algorithm code to get the accurate result of observation sequence of P(O| **λ**).

Here, we are showing this Table 3 how to get that result. PDB code is denoting the protein sequence code (accession code) resolution power is showed the crystallization structure of protein and last column is value of P(O| **λ**) of each protein (gene) sequence.

Table 3: Comparative table of resolution power and computational result of P(O| λ) using HMM forward. algorithm.

S.No.	PDB code	Resolution Power	Estimated Value of P(O| **λ**)
1.	3MTE	1.80 Å	$9.12571297574780e^{-211}$
2.	3MQ2	1.69 Å	$2.070484515632\ e^{-213}$
3.	5CUQ	1.74 Å	$3.168163846644452e^{-258}$
4.	5E9Q	1.79 Å	$9.97682150936174e^{-259}$
5.	4R05	2.00Å	$8.05668799576945e^{-269}$
6.	3P97	1.74Å	$8.084576150979e^{-279}$
7.	3P8Z	1.79 Å	$68.92936623203153e^{-279}$
8.	5EIF	2.00 Å	$4.663302228006e^{-278}$
9.	3EVG	2.20 Å	$3.435601382845763e^{-281}$
10.	5EC8	1.76Å	$3.84478566459856e^{-281}$
11.	5EKX	1.76Å	$7.372982666530e^{-281}$
12.	1R6A	2.60Å	$6.854042143922399e^{-300}$
13.	2P3O	2.76Å	$1.9618432129379416e^{-304}$
14.	2P3Q	2.75Å	$1.882298836123327e^{-304}$
15.	2P40	2.70 Å	$1.9618432129379416e^{-304}$
16.	1L9K	2.40Å	$1.9618432129379416e^{-304}$
17.	2P3L	2.20Å	$1.9618432129379416e^{-304}$
18.	2P41	1.80Å	$1.9618432129379416e^{-304}$
19.	4V0Q	2.30Å	$6.474723842632e^{-311}$
20.	2P1D	2.90Å	$5.29e^{-321}$

As shown in Table 3 compare the value of P(O| **λ**) using the computational method of forward algorithm. In other word we can say that resolution power of protein sequence in terms of non-decreasing value of P(O| **λ**) and found that highest probability also contains the average resolution power. Here using the P(O| **λ**) is probability of observation sequence and Fig. 5 presenting the graph of global maximum value of P(O| **λ**) probability of observation sequence in non-descending order.

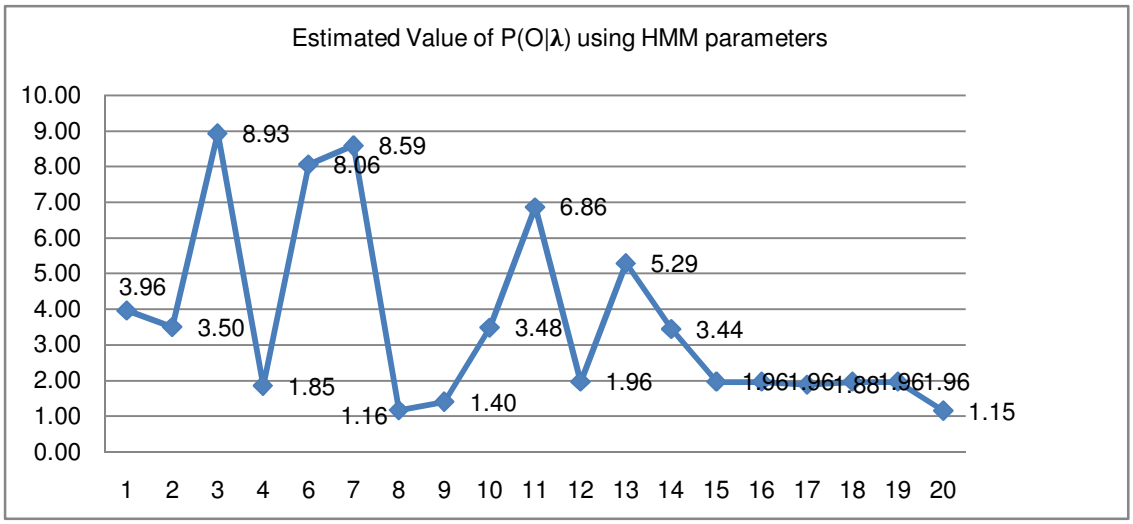


Fig. 5. Showing the graph of P(O| λ) value with the iteration (itr) of observation sequence.

In above graph, we shown the estimated value of observation sequence P(O| λ) using the HMM parameters. In x-axis we take the 1-20 gene sequence and y-axis showing maximum value of P(O|λ)**.** The estimated global maximum value using the HMM parameters is lie between (8.00-9.00) value is $8.93e^{-211}$ and lowest value lie between (1.00-2.00) value $1.15e^{-321}$ where E is the exponential value.

NS5 protein has played an important role in the dengue virus replication [12]. Information about the NS5 protein is important for drug design and development and also for effects of mutations on the NS5. Comparison and classification of protein structure are essential to understand protein function. Due to the problem in computational capabilities and the increasing amount of data structures, it is not possible to compare all the existing structures using standard methods. Researchers have found that NS5 MTase is one of the most important proteins of dengue virus replication. Based on the result of function prediction, NS5 proved to be one of the most important parts of dengue virus replication. Some researchers have used NS5 MTase for dengue drugs [32-34]. In mutations of protein structure can cause changes in the secondary and tertiary structure of proteins. Changes in the secondary and tertiary structures affect the protein function and folding. The change of secondary and tertiary structure can affect the molecular docking simulation during drug design process [35]. The present study and recent works are based on this MTase protein and finding the parameters of HMM. Then using the forward algorithm to find the value of P(O| λ) of each DNA sequence of NS5 MTase protein of dengue virus. In previous work we have taken the same protein of dengue virus to find the novel drug molecules as anti-dengue compounds using *in-silico* protein-lig and molecular docking method [36]. In this study we have done the analysis for DNA sequence of NS5 Methyltransferase protein with the help of computational methods.

IV. CONCLUSION

In this work, we used Hidden Markov Model for analysis of DNA sequence of NS5 Methyltransferase protein of dengue virus. The Hidden Markov Model approach plays a major role in the bioinformatics field as well as in the prediction theory of many mathematical and engineering methods. Using forward algorithm HMM approach, the Table 3 showing the highest value of P(O| λ). Conclusion of this method we say that estimated value of observation sequence is depends upon the average value of A, G, C and T characters of gene sequence. On the basis of DNA base pairs we draw the a_{ij}, $b_j(k)$ π_i matrix and conclude the highest global maxima value of P(O| λ). Analyze the Fig. 5 showing the possible confirmations of each gene sequence. Novelty of this work, we perform in-silico work using HMM a computational method to find the best gene sequence of (accession code) of NS5 MTase of dengue virus. HMM technique much not used in this type of research work and this method usage to solve the problem of large data sets of proteins, genes and also used the whole genome of any organism.

V. FUTURE SCOPE

In a further study, we will apply other approaches like Genetic Algorithm (GA), Artificial Neural Network (ANN) and etc. The result of HMM and other machine learning approaches succeed then researchers will be used for further research and study of NS5 Methyltransferase protein of dengue virus for drug discovery.

ACKNOWLEDGEMENTS

This work is a part of Ph.D. research work.

Conflicts of Interest. The authors declare no conflicts of interest.

REFERENCES

[1]. Rodenhuis-Zybert, I. A., Wilschut, J., & Smit, J. M. (2010). Dengue virus life cycle: viral and host factors modulating infectivity. *Cellular and molecular life sciences*, *67*(16), 2773-2786.

[2]. Normile, D (2013). Tropical medicine. Surprising new dengue virus throws a spanner in disease control efforts. *Science*, *342* (6157), 342-415.

[3]. Dwivedi, V. D., Tripathi, I. P., Tripathi, R. C., Bharadwaj, S., & Mishra, S. K. (2017). Genomics, proteomics and evolution of dengue virus. *Briefings in functional genomics*, *16*(4), 217-227.

[4]. Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L. & Myers, M. F. (2013). The global distribution and burden of dengue. *Nature*, *496*(7446), 504-507.

[5]. Balmaseda, A., Standish, K., Mercado, J. C., Matute, J. C., Tellez, Y., Saborío, S., ... & Holmes, E. C. (2010). Trends in patterns of dengue transmission over 4 years in a pediatric cohort study in Nicaragua. *The Journal of infectious diseases*, *201*(1), 5-14.
[6]. Lescar, J., Luo, D., Xu, T., Sampath, A., Lim, S. P., Canard, B., & Vasudevan, S. G. (2008). Towards the design of antiviral inhibitors against flaviviruses: the case for the multifunctional NS3 protein from Dengue virus as a target. *Antiviral research*, *80*(2), 94-101.
[7]. Vannice, K. S., Durbin, A., & Hombach, J. (2016). Status of vaccine research and development of vaccines for dengue. *Vaccine*, *34*(26), 2934-2938.
[8]. Herrero, L. J., Zakhary, A., Gahan, M. E., Nelson, M. A., Herring, B. L., Hapel, A. J., & Taylor, A. (2013). Dengue virus therapeutic intervention strategies based on viral, vector and host factors involved in disease pathogenesis. *Pharmacology & therapeutics*, *137*(2), 266-282.
[9]. Monath, T. P. (1994) Dengue: The risk to developed and developing countries. *Proc. National Academy of Sciences, Washington USA, 91*, 2395-2400
[10]. Prasad K. S., & Miryala, R. (2019). Histopathological Image Classification Using Deep Learning Techniques. *International Journal on Emerging Technologies,10*(2), 467-473.
[11]. Khan, N. Gautam, G., & Gupta A. K. (2019). In Silico Protein-Ligand Docking Studies on Thiazolidinediones analogs as Potential Aldose reductase Inhibitors. *Biological Forum- An International Journal, 11*(1), 77-83.
[12]. Noble, C. G., & Shi, P. Y. (2012). Structural biology of dengue virus enzymes: towards rational design of therapeutics. *Antiviral research*, *96*(2), 115-126.
[13]. Lim, S. P., Noble, C. G., & Shi, P. Y. (2015). The dengue virus NS5 protein as a target for drug discovery. *Antiviral research*, *119*, 57-67.
[14]. Jain, N., & Sharma, V. (2012). Distance Weight Optimization of Association Rule Mining with Improved Genetic algorithm. *International Journal of Electrical, Electronics and Computer Engineering, 1*(2), 25-27.
[15]. Nguyen, K. H. V., Nguyen, T. T., Truong, N. M., & Do, T. H. (2019). Application of Bioinformatic tools for Prediction of Active pH and Temperature Stability of Endoglucanases Based on Coding Sequences from Metagenomic DNA data. *Biological Forum- An International Journal, 11*(2), 14-20.
[16]. Rabiner, L. R., Wilpon, J. G., & Juang, B. H. (1986). A model-based connected-digit recognition system using either hidden Markov models or templates. *Computer Speech & Language*, *1*(2), 167-197.
[17]. Pedersen, J. S., & Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, *19*(2), 219-227.
[18]. Eddy, S. R. (1996). Hidden Markov Models. *Curr. Opin. Struct. Biol., 6*, 361-365.
[19]. Kumar, R., Lamba, N., Sharma A., & Chugh, P. (2013). Comparative analysis for measures of forecast accuracy 'A Case Study'. *International Journal on Emerging Technologies, 4*(1), 26-31.
[20]. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.
[21]. Katiyar, N., Nath, R. (2016). Methods for Protein Structure Prediction and Its Application in Drug Design Using Hidden Markov Model. *International Journal of Emerging Trends in Science and Technology, 3*(03), 3676-3684.
[22]. Egloff, M. P., Benarroch, D., Selisko, B., Romette, J. L., & Canard, B. (2002). An RNA cap (nucleoside-2′-O-)-methyltransferase in the flavivirus RNA polymerase NS5: crystal structure and functional characterization. *The EMBO journal*, *21*(11), 2757-2768.
[23]. Egloff, M. P., Decroly, E., Malet, H., Selisko, B., Benarroch, D., Ferron, F., & Canard, B. (2007). Structural and functional analysis of methylation and 5′-RNA sequence requirements of short capped RNAs by the methyltransferase domain of dengue virus NS5. *Journal of Molecular Biology*, *372*(3), 723-736.
[24]. Benarroch, D., Egloff, M. P., Mulard, L., Guerreiro, C., Romette, J. L., & Canard, B. (2004). A structural basis for the inhibition of the NS5 dengue virus mRNA 2′-O-methyltransferase domain by ribavirin 5′-triphosphate. *Journal of Biological Chemistry*, *279*(34), 35638-35643.
[25]. Zhao, Y., Soh, T. S., Zheng, J., Chan, K. W. K., Phoo, W. W., Lee, C. C., ... & Shi, P. Y. (2015). A crystal structure of the dengue virus NS5 protein reveals a novel inter-domain interface essential for protein flexibility and virus replication. *PLoS pathogens*, *11*(3), 1-27.
[26]. Geiss, B.J., Thompson, A. A., Andrews, A. J., Sons, R. L., Gari, H. H., Keenan, S. M., & Peersen, O. B. (2009) Crystal Structure of Dengue-2 Virus Methyltransferase Complexed With S-adenosyl-l-homocysteine. *Journal of Molecular Biology, 385*(5), 1643-1654.
[27]. Benmansour, F., Trist, I., Coutard, B., Decroly, E., Querat, G., Brancale, A., & Barral, K. (2017). Discovery of novel dengue virus NS5 methyltransferase non-nucleoside inhibitors by fragment-based drug design. *European journal of medicinal chemistry*, *125*, 865-880.
[28]. Brecher, M. B., Li, Z., Zhang, J., Chen, H., Lin, Q., Liu, B., & Li, H. (2015). Refolding of a fully functional flavivirus methyltransferase revealed that S-adenosyl methionine but not S-adenosyl homocysteine is copurified with flavivirus methyltransferase. *Protein Science*, *24*(1), 117-128.
[29]. Macmaster, R., Zelinskaya, N., Savic, M., Rankin, C. R., & Conn, G. L. (2010). Structural insights into the function of aminoglycoside-resistance A1408 16S rRNA methyltransferases from antibiotic-producing and human pathogenic bacteria. *Nucleic acids research*, *38*(21), 7791-7799.
[30]. Lim, S. P., Sonntag, L. S., Noble, C., Nilar, S. H., Ng, R. H., Zou, G., & Bodenreider, C. (2011). Small molecule inhibitors that selectively block dengue virus methyltransferase. *Journal of Biological Chemistry*, *286*(8), 6233-6240.
[31]. Brecher, M., Chen, H., Li, Z., Banavali, N. K., Jones, S. A., Zhang, J., & Li, H. (2015). Identification and characterization of novel broad-spectrum inhibitors of the flavivirus methyltransferase. *ACS infectious diseases*, *1*(8), 340-349.
[32]. Podvinec, M., Lim, S. P., Schmidt, T., Scarsi, M., Wen, D., Sonntag, L. S., & Schwede, T. (2010). Novel inhibitors of dengue virus methyltransferase: discovery by in vitro-driven virtual screening on a desktop computer grid. *Journal of Medicinal Chemistry*, *53*(4), 1483-1495.
[33]. Puspitasari, R. D., & Tambunan, U. S. F. (2017). In silico mutation analysis of non-structural protein-5 (NS5) dengue virus. In *IOP Conference Series: Materials Science and Engineering, 188*(1), 1-6.
[34]. Simmons, C. P., Farrar, J. J., van Vinh Chau, N., & Wills, B. (2012). Dengue. *New England Journal of Medicine*, *366*(15), 1423-1432.
[35]. Halstead, S. B. (2007). Dengue. *The lancet*, *370*, 1644-1652.
[36]. Katiyar, N., & Nath, R. (2018). In-Silico Protein-Ligand Docking Studies against the NS5 Methyltransferase Protein of Dengue Virus. *International Journal of Computer Trends and Technology (IJCTT)*, *62*(1), 21-29.

How to cite this article: Katiyar, Nidhi, Nath, Ravindra and Katiyar, Shashwat (2020). Estimated Value of Hidden Markov Model Parameters for NS5 methyltransferase Protein of Dengue Virus. *International Journal on Emerging Technologies*, *11*(1): 270–279.