TEST Engineering & Management

March-April 2020
ISSN: 0193-4120 Page No. 1235 - 1242

Multiple Sequence Alignment with Hidden Markov Model for Diabetic Genome

[1]Deneshkumar V, [2]*Manoprabha M, [3]Senthamarai Kannan K

[1, 2, 3]Department of Statistics, Manonmaniam Sundaranar University, Abishekapatti,
Tirunelveli-627012, Tamilnadu, India
[1]vdenesh77@gmail.com, [2*]manoprabhamurugan@gmail.com,[3]senkannan2002@gmail.com

Article Info
Volume 83
Page Number: 1235 - 1242
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 14 March 2020*

Abstract

Diabetes is one of the chronic diseases which occur when the pancreas is not able to secret insulin. Insulin is an important factor that transforms glucose in to energy. Analysing the multiple DNA sequence of diabetes is helpful in deriving more information about the disease. Profile Hidden Markov Model has a wide application in molecular biology. Thus we emphasized the use of PHMM for this Multiple Sequence Alignment (MSA). The main objective of this paper is to find the sequence pattern which the disease follows, estimating the parameters using Baum-Welch algorithm and finding the best optimal path using Viterbi algorithm. All valuable information from the sequences is obtained using PHMM.

***Keywords:** MSA, PHMM, Baum-Welch, Viterbi, Diabetes and DNA Sequence.*

1. Introduction

Genetic information's are stored as the sequences of nucleotides in DNA which are represented by symbols. DNA is made up of molecule called nucleotides. The information in DNA is stored as a code made up of four nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Sequence alignment is used to match the homologous nucleotides of two sequences. One of the major tasks in MSA is to compare three or more DNA sequences and find similarities, or differences, and infer structural, functional and evolutionary relationship.

Markov chain is a linear collection of symbols, chosen from a finite set, each symbol occurs at every position with a specified probability. The Markov chain may exist in one of a set of hidden states at any given time, with the probability of making a transition to another one of these hidden states. It then requires a hidden Markov model. A variability model of nucleotides in particular position of family is called profile. Profile HMM is architecture that is suitable for modelling sequence profile. This model consist of match, insert, delete as three hidden states and it is used for position-specific symbol frequencies, symbol insertion and symbol deletions respectively (Yoon 2009).

India is the capital of diabetes of the world, it consists of over 60 million adults with diabetes and 30 million remains undiagnosed. Thus it increased the risk of complications in patients and premature death. It is a challenge for researchers to detect it earlier. Multiple Sequence Analysis is helpful in the prevention of the disease. The knowledge from DNA sequence analysis is more important in the field of biological research. It has vast applications in the field of medical diagnosis, forensic biology, biotechnology and biological systematic. Rabiner (1989) initially developed a HMM in speech recognition for classification, clustering and segmentation. Krogh (1998) has developed HMM models for sequence alignments in biological sequence analysis and also presented some statistical approaches to find the similarity between two sequences. A biological sequence is modelled by stochastic process which is moved from first state to the next state and each state emits the element of sequence based on emission probability distribution Przytycka and Zheng (2006). The homologies of a sequence alignment or protein structures are detected by profiles and it has many parameters. The HMM is composed of a number of states with corresponding positions in columns of a multiple alignment. Each state emits the symbols according to the emission probabilities and the states are interrelated by state transition probabilities. The major problem in this model is to setting the position specific residue scores, score gaps insertions and combining structural and multiple sequence information. Starting from initial state, a sequence of states generated and moving from state to state according to the state transition probabilities until an end state is reached Eddy (1996). Kalpana and sasikumar (2016) studied about the application of profile and pair Hidden Markov Model explained with some numerical illustrations. Ramanathan (2006) has discussed the generative sequences characterized by a set of observable sequences through HMM. The HMM can be used to model stochastic processes where the non-observable state of the system is governed by a Markov process. The observable sequences of system have an underlying probabilistic dependence.

HMM are applied in various fields like sequence comparison, structure prediction, detecting gene recombination and gene finding. A HMM consists of different hidden states and each state emits a residue when it is visited. Every state contains a transition and emission probability. Like transition probability emission probability also depends on the past Benjamin and Bateman (2007). Auer and Doerge (2010) proposed the new technology to analyse the RNA sequences and it was used in characterizing and quantifying entire genomes. The data generated using this technique is more informative, but some essential concepts like randomization, replication and blocking should be paid attention for any properly analyzed design. Kalpana and Sasikumar (2015) have discussed the HMM model for cancer sequence alignment. They used multiple sequence alignment for sequence architecture and used all basic algorithms in HMM to find the optimal sequence path and sequence probability fit. Sonnhammer et al., (1998) have predicted the protein sequences using HMM with seven states which was estimated by the maximum likelihood and a discriminative method. This method was also applied for large protein family and achieved the high accuracy. Nath and Jain (2011) discussed about the parameter estimation of HMM. The most important problem in HMM is to optimize the model parameter to describe how the observation sequence comes about. The traditional method that is applied to estimate the parameter in the HMM model is Baum-Welch algorithm. Mulia, et al. (2012) discussed about the profile HMM based MSA for DNA sequences. They tested the applicability of PHMM for MSA with a task and showed that it works well.

Blasiak and Rangwala (2011) have discussed the HMM in classifying the sequence of amino acids in to structural classes. They have used Baum-Welch algorithm, Gibbs algorithm and a variational algorithm to infer the model parameters. Bartolomeo, et al., (2011) have applied HMM in the progression of liver cirrhosis to HCC. They have estimated the transition probability of different stages in liver cirrhosis patients. The database used in this study was affected by misclassification. HMM is used in finding the transition probability between two states inspite of misclassification. Petersen, et al. (2017) modelled HMM in sepsis progression. HMM is considered as one of the important tool in disease progression. Early research modelled HMM in the sepsis progression with homogenous group of patients. This study is modelled significantly on heterogeneous groups. Birney (2001) investigated the biomolecular sequence using HMM which deals with gene-prediction. This paper briefly described about the techniques used in the sequence analysis. Sakakibara (2003) proposed the pair HMM on tree structures for structural alignment of RNA and identified the non-coding RNA regions on the genome. The effectiveness and the complexity issue of PHMMTSs on structural alignment are demonstrated. Stanke and Waack (2003) have introduced new program called AUGUSTUS for predicting protein genes in eukaryotic DNA sequences. The program was based on HMM, which predicts better than the existing program that does not perform well on longer sequences. Fonzo, et al. (2007) have explained HMM in bioinformatics and also the problems faced while undergoing HMM is reviewed in detail. Eddy (1995) has discussed the algorithms of HMM for multiple sequence alignment and it was compared with simulated method and other existing procedures. This experiment was done on ten different protein families based on the comparison this simulated method was good in multiple sequence alignment.

Benyacoub, et al. (2014) have developed a new model for classification under the assumption of both the states and observations are discrete. HMM has a vast application but has some restriction in learning supervised problems. Bonneville and Jin (2013) have identified the epigenetic regulation patterns for estrogen receptor α target genes computational approach. They have illustrated the application of HMM in genome-wide high-throughput genomic data to study epigenetic influences on E2/ER α regulation in breast cancer. Based on the report of Sean R. Eddy and Boer (2016) described about HMM for sequence alignment than other sequence alignment method. The basic algorithms like the forward, backward and the veterbi were used in HMM was explained. Nimmy et al., (2018) have investigated about DNA discontinuity, which may lead to cause of harmful diseases. Tuberculosis is one of the critical diseases which are caused due to some breakage in DNA sequence. Hidden Markov chain, linear transformation and Box-Cox transformation were used to predict the break in a long DNA. Among these predictive model HMM provided faster result with more accuracy and reliability. Tamposis et al., (2018) applied the semi supervised HMMs for sequence analysis. This algorithm is used for all labeled, unlabeled and partially labeled data. HMM works under the concept of EM algorithm, where the missing labels are considered as missing data. The result showed a significant prediction than other classifiers. Bottolo and Richardson (2019) discussed about gene hunting using hidden Markov model knockoffs. They reviewed and motivated to make knockoff for genetic applications. The hidden Markov model is used to generate knockoffs which are helpful in capturing the DNA pattern variations.

2. Methodology

Let y_t represents a collection of random variables depending on t. We say that y_t is parameterized by t and a parameterized collection of random variables is known as a stochastic process. A stochastic process has the Markov property if the future is conditionally independent of the past given the present. A stochastic process that has the Markov property is called a Markov process. A Markov chain is a Markov process for which the random variables take only countably many values. The sample space of the individual random variables of a stochastic process is referred to as the state space, so a Markov chain has a countable state space. To specify a finite state space, one needs to specify a distribution for where the chain starts, and a set of conditional probabilities that specify how the chain moves from one state to another (Cavan Reilly 2009).

2.1 Hidden Markov Model

The fundamental idea behind a hidden Markov model is that there is a Markov process we cannot observe that determines the probability distribution for what we do observe. Thus a hidden Markov model is specified by the transition density of the Markov chain and the probability laws that govern what we observe given the state of the Markov chain. Given such a model, we want to estimate any parameters that occur in the model. We would also like to determine what is the most likely sequence for the hidden process. Finally we may want the probability distribution for the hidden states at every location.

Let y_t represent the observed value of the process at location t for $t = 1,\ldots.,T$, θ_t the value of the hidden process at location t and let ϕ represent parameters necessary to determine the probability distribution for y_t given θ_t and θ_t given θ_{t-1}. In our applications, y_t will either be an amino acid or nucleotide and the hidden process will determine the probability distribution of observing different letters. Our model is then described by the sets of probability distributions $p(y_t\,|\,\theta_t,\phi)$ and $p(\,\theta_t\,|\,\theta_{t-1},\phi\,)$. A crucial component of this model is that the y_t are independent given the set of θ_t and θ only depends directly on its neighbours θ_{t-1} and θ_{t+1}.

The various distribution in which we are interested are $p(\phi\,|y_1,\ldots.,y_T)$, $p(\,\theta_t|y_1,\ldots.,y_T)$ for all t and $p(\theta_1,\ldots\ldots,\theta_T|y_1,\ldots\ldots,y_t)$. We will adopt a Bayesian perspective, so that we treat θ_t as a random variable. (Cavan Reilly 2009).

A profile HMM is a certain type of HMM with a structure that is suitable for representing the profiles of a MSA. It can be obtained from a multiple alignment of protein or DNA sequences and effectively represents the common pattern and other statistical properties.

2.2 Multiple Sequence Alignment

MSA may be formally defined as a two-dimensional table in which each row represents a nucleic acid sequence, and the columns are the individual residue positions. Alignment of several sequences has lead to many important results regarding common sequence patterns or motifs in nucleic acids. One of the common goals of building MSA is to characterize gene families and identify the shared region of homology. MSA also helps to classify sequences in to families. All the sequences in such a family may have been derived from some common ancestral sequence, indicating an evolutionary relationship. MSA helps to predict the secondary and tertiary structures for new sequences, and identify templates for threading and homology modelling, which are methods for 3-D structure prediction.

Consensus sequences are a useful way of representing patterns, but they are even more deterministic than regular expressions. They are a succinct way of representing the information present in a MSA, but they abstract only the most prominent of such information and discard all the rest of it. Sequence logos, though requiring specialized software and hardware, are a way of writing consensus sequences using probabilistic information. To build a logo, we start with an aligned set of sequences. The residues that occur most frequently at each position are identified, and they form the consensus sequence, which is displayed most prominently in the logo. (Gautham 2006).

2.3. Parameter Estimation

For parameter estimation if there is prior information about the parameters, we could incorporate such information in the usual fashion. A popular approach to parameter estimation is to use the EM algorithm to estimate the parameters in the model, and then use these estimates as if they were known. The most popular implementation of the EM algorithm for hidden Markov models is called the Baum Welch algorithm. (Cavan Reilly 2009).

2.3.1. The Baum-Welch Algorithm

The Baum-Welch algorithm starts from an initial set of model parameters θ_0. In each iteration, it changes the parameters as follows:

1. Calculate the estimated number of times in each transition and emission is used to generate the training set T in an HMM whose parameters are θ_k.
2. Use the frequencies obtained in step 1 to reestimate the parameters of the model, resulting in a new set of parameters θ_{k+1}.

The first step of the algorithm can be viewed as creating a new annotated training set $T^{(k)}$, where for each unannotated sequence $X \in T$, we add every possible pair (X, H) of the sequence X and any state path, weighted by the conditional probability $Pr(H|X, \theta_k)$ of the path H in the model with parameters θ_k, given the sequence X. The second step then estimates new parameters θ_{k+1}, as in the supervised scenario, based on the new training set $T^{(k)}$. The Baum-Welch algorithm achieves the same result in $O(nm^2)$ time per iteration using the forward and backward algorithm to avoid explicitly creating this exponentially large training set. Baum has shown that the likelihood of the training set improves in each iteration of this algorithm. However, this does not guarantee that the Baum-Welch algorithm reaches optimal model parameters: it may instead reach a local maximum or a saddle point in the parameter space. A modification of the Baum-Welch algorithm, called Viterbi training, is also often used in practice. In the first step of the algorithm, instead of considering all possible paths through the model, we only consider the most probable path. However, this algorithm is not guaranteed to increase the likelihood of the observed data in each step. The Baum-Welch algorithm can also be used in the semi supervised scenario. (Mandoiu and Zelikovsky (2008)).

2.3.2. The Viterbi Algorithm

Once the HMM topology is set and its parameters trained, we can use it to find genes in a newly unlabeled DNA sequence X. In other word, we seek an appropriate state path H^* that best explains how the model could have produced X; this process is called HMM decoding. The simplest measure of "best" is to find the path that has the maximum probability in the HMM, given the sequence X. Recall that the model gives the joint probabilities $Pr(H, X)$ for all sequence, and as such, it also gives the posterior probability $Pr(H, X) = Pr(H, X)/Pr(X,$ for every possible state path H through the model, conditioned on the sequence X. We will seek the path with maximum posterior probability. Given that the denominator $Pr(X)$ is constant in the conditional probability formula for a given sequence X, maximizing the posterior probability is equivalent to finding the state path H* that maximizes the joint probability $Pr(H^*, X)$.

The most probable state path can be found in time linear in the sequence length by the Viterbi algorithm. This simple dynamic programming algorithm computes the optimal paths for all prefixes of X; when we move from the i −length prefix to the $(i + 1)$ −length prefix, we need only add one edge to one of the precomputed optimal paths for the i −length prefix.

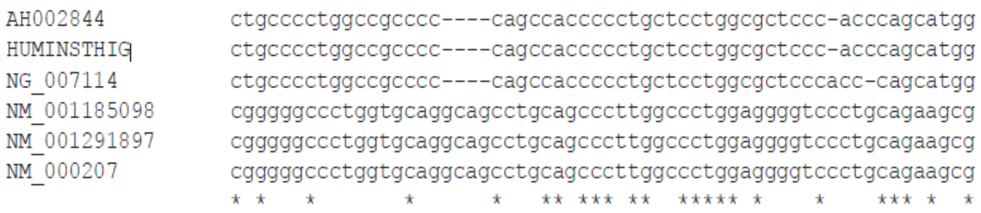
For every position i in the sequence and every state k, the algorithm finds the most probable state path $h_1, \ldots, h_i$ to generate the first i symbols of X, provided that $h_i = k$. The value $V[i, k]$ stores the joint probability $Pr(h_1, \ldots, h_i, x_1, \ldots, x_i)$ of this optimal state path. Again, if $h_1, \ldots, h_i$ is the most probable state path generating $x_1, \ldots, x_i$ that ends in state h_i, then $h_1, \ldots, h_{i-1}$ must be the most probable state path generating $x_1, \ldots, x_{i-1}$ and ending in state h_{i-1}. To compute $V[i, k]$, we consider all possible states as candidates for the second-to-last state, h_{i-1} and select the one that leads to the most probable state path, as expressed in the following recurrence:

$$V[i,k] = \begin{cases} s_k . e_{k,x_1}, & if\ i = 1 \\ max_l V[i-1,l]. a_{l,k} . e_{k,x_i,} & otherwise \end{cases} (1)$$

The probability $Pr(H^*, X)$ is then the maximum over all states k of $V[n, k]$, and the most probable state path H^* can be traced back through the dynamic programming table by standard techniques. The running time of the algorithm is $O(nm^2)$, here *n*, *m* denoted by length of the sequence and the number of state in the HMM. (Mandoiu and Zelikovsky (2008)).

3. Results and Discussion

For performing the multiple sequence alignment we used the six different DNA sequences of diabetic patients. These DNA sequence are collected from GenBank which are openly accessible. To generate the alignment between these sequences we used Clustal Omega which is a multiple sequence alignment program. The alignment generated for this dataset is given in figure 1.



```
AH002844        ctgcccctggccgcccc----cagccacccctgctcctggcgctccc-acccagcatgg
HUMINSTHIG      ctgcccctggccgcccc----cagccacccctgctcctggcgctccc-acccagcatgg
NG_007114       ctgcccctggccgcccc----cagccacccctgctcctggcgctcccacc-cagcatgg
NM_001185098    cggggccctggtgcaggcagcctgcagcccttggccctggaggggtccctgcagaagcg
NM_001291897    cggggccctggtgcaggcagcctgcagcccttggccctggaggggtccctgcagaagcg
NM_000207       cggggccctggtgcaggcagcctgcagcccttggccctggaggggtccctgcagaagcg
                * *   *       *      * **  *** **  ***** *    *    *** *  *
```

Figure 1: Multiple sequence alignment of six diabetic DNA sequences

Schneider and Stephens invented the sequence logos which graphically represents the consensus sequences using probabilistic information. It is helpful in studying the order of predominance of each residue at each position, the probability of each residue at that position, the amount of information present at each position. The residues that occur most frequently at each position are identified, and they form the consensus sequence. Sequence logos given in figure 2 provides rich information in a single figure, these are generated using WebLogo.

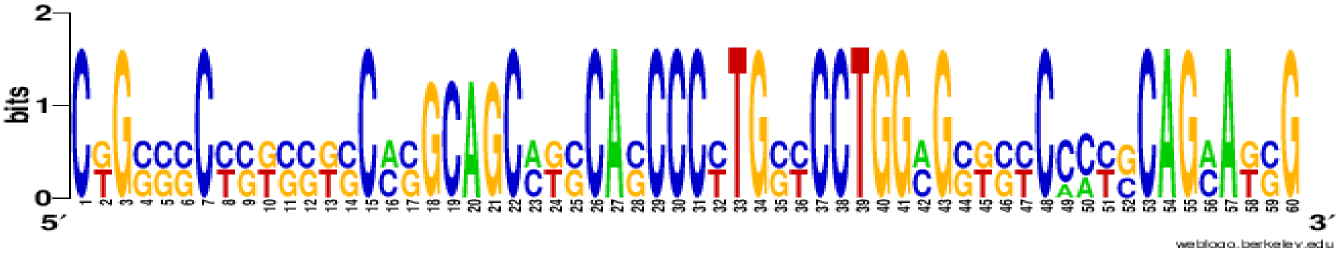


Figure 2: The sequence logos for an arbitrary set of aligned DNA sequences

Table 1: Percent Identity Matrix for the DNA Sequence

Sl.No	Sequence Name	Probability
1	AH002844	100.00 89.28 88.11 81.34 70.38 69.78
2	HUMINSTHIG	89.28 100.00 99.02 81.34 70.38 69.78
3	NG_007114	88.11 99.02 100.00 81.65 70.58 70.00
4	NM_001185098	81.34 81.34 81.65 100.00 93.14 95.48
5	NM_001291897	70.38 70.38 70.58 93.14 100.00 93.76
6	NM_000207	69.78 69.78 70.00 95.48 93.76 100.00

Table 1 shows that the sequence similarity of all the six DNA sequence can be studied using the percent identity matrix. And also shows that the diabetic DNA sequence in human has the higher probability to follows the same pattern. This percent identity matrix was created by Clustal 2.1. Percent identity is used to estimate the similarity between two different sequences. The DNA sequence from the same family is expected to have higher percent identity.

Table 2: Transition Probability matrix for estimating the parameter

Hidden states	M	I	D
M	0.5010294	0.2853530	0.2136176
I	0.2428036	0.3384280	0.4187684
D	0.1694115	0.3237429	0.5068456

In the set of DNA sequence, each residues belongs to any one of the three specified hidden states; Match state (M), Insert state (I) and Delete state (D). The next important step in HMM construction after sequence alignment is parameter estimation, which are obtained from the transition and emission probabilities produced by the implementation of Baum-Welch algorithm. The values obtained in table 2 are the maximum likelihood function which shows the probability of transition from one state to another state or remains the same over a period of time. The change in the position of each residue may lead to sequence evolution. And the effect of a particular residue in particular position strongly depends on its neighbor residue which is the concept of a Markov model.

Table 3: Emission Probability Matrix using Baum-Welch algorithm

Hidden states	Observable states			
	A	C	G	T
M	0.04155045	0.5919802	0.1663403	0.2001290
I	0.20600955	0.2657379	0.3136564	0.2145962
D	0.33287647	0.1552212	0.3555312	0.1563711

The emission probability represents the relationship between the hidden state and the observable state. And the values in Table 3 refer to the probability of an observation representing the hidden state of the model for that specific state transition. This shows that each hidden state M, I and D has different emission probabilities for each nucleotide A, C, G and T in DNA sequences.

Table 4: Viterbi Path of Hidden States

Sl.No	Sequence of hidden states
2	"M" "I" "D" "D" "I" "I" "D" "D" "D" "I" "I" "I" "D" "D" "M" "M" "D" "D" "D" "D" "D" "M" "M"
24	"M" "M" "M" "D" "D" "M" "M" "M" "M" "I" "I" "M" "M" "M" "M" "I" "D" "D" "D" "D" "D" "D" "D"
47	"D" "I" "D" "I" "M" "M" "M" "D" "D" "D" "D" "D" "D" "I" "D" "D" "D" "D" "D" "D" "D" "D" "I"
70	"M" "D" "I" "I" "D" "D" "D" "D" "D" "M" "M" "D" "D" "D" "M" "M" "M" "M" "D" "D" "D" "D" "M"
93	"M" "M" "M" "M" "M" "M" "I" "D" "D" "D" "D" "M" "M" "M" "M" "M" "I" "I" "D" "D" "D" "D" "D"
116	"M" "M" "D" "D" "D" "D" "I" "D" "I" "I" "D" "D" "D" "M" "M" "D" "D" "D" "D" "D" "I" "M" "M"
139	"M" "M" "M" "D" "D" "D" "I" "I" "I" "D" "D" "D" "D" "D" "I" "M" "M" "M" "M" "M" "D" "D" "D"
162	"D" "D" "D" "M" "M" "M" "M" "D" "I" "I" "I" "M" "M" "M" "M" "M" "M" "M" "M" "D" "D" "M" "I"
185	"D" "M" "M" "D" "D" "D" "M" "M" "M" "I" "I" "I" "I" "M" "M" "M" "M" "M" "M" "M" "M" "I" "M"
208	"M" "I" "I" "I" "D" "D" "D" "D" "D" "M" "M" "M" "M" "D" "D" "D" "D" "D" "D" "M" "M" "M" "M"
231	"I" "D" "D" "D" "M" "M" "M" "M" "M" "M"

Table 5: Sequence probability fit in the PHMM

Diabetes	Hypertension	Cardiac Disease	Renal Disease	Obesity
0.933	0.39	0.498	0.441	0.492
0.852	0.56	0.42	0.498	0.327
0.893	0.467	0.421	0.478	0.431
0.99	0.47	0.46	0.395	0.43
0.86	0.433	0.41	0.466	0.417

Table 4 reveals that most likely sequence of hidden states computed from Viterbi algorithm. The derived PHMM is applied to 25 different types of DNA sequence from 5 family of human disease sequence. The probability score shows how the sequence fit with the PHMM. The maximum score denotes that these families of sequence perfectly fit with the derived PHMM. From table 5 only the diabetes family of sequence has the highest score and the rest disease sequence has the least score which means that only the diabetes diseases fit with the derived PHMM. It is also clear that the diabetic patient does not have any higher chance of being affected from the above mentioned diseases. Because their sequence patterns are different and does not have any higher probability to be same.

4. Conclusion

Multiple sequence alignment is one of the most important techniques used in discovering new patterns in the sequence. Aligning several sequences together gives more biological information which is useful in characterizing the gene families. Here we used the PHMMs in identifying the probabilistic pattern of a diabetic family of sequence; based on the sequence comparison by Baum-Welch algorithm the transition probability and the emission probability are estimated. From the results, the chance of each residue in each state and the relationship between the hidden state and observable states are elucidated. The best path of the hidden states is discovered using Viterbi algorithm. Using the sequence probability fit a new sequence is evaluated whether it has any membership with the aligned family of sequence. Thus PHMMs is one of the most appropriate tools in extracting all the statistical information from a multiple diabetic sequence.

References

[1] P. L. Auer andR. W. Doerge, “Statistical Design and Analysis of RNA Sequencing Data”, Genetic Society of America, vol. 185, (2010), pp. 405-416.

[2] N. Bartolomeo, P. Trerotoli and G. Serio, “Progression of Liver cirrhosis to HCC: An Application of Hidden Markov Model”, BMC Medical Research Methodology, vol. 11, no. 38, (2011), pp. 1-8.

[3] B. Benyacoub, S. El Bernoussi, A. Zoglat and E. M. Ismail, “Classification with Hidden Markov Model”, Applied Mathematical Sciences, vol. 8, no. 50, (2014), pp. 2483-2496.

[4] E. Birney, “Hidden Markov Models in Biological Sequence Analysis”, International Business Machines Corporation, vol. 45, no. 3/4, (2001), pp. 449-454.

[5] S. Blasiak and H. A. Rangwala, “Hidden Markov Model Variant for Sequence Classification”, Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, (2011), pp. 1192-1197.

[6] B. S. Bockler and A. Bateman, “An Introduction to Hidden Markov Models”, Current Protocols in Bioinformatics, (2007), pp. A.3A.1-A.3A.9.

[7] J. Boer, “Multiple Alignment using Hidden Markov Models”, Proteins 4, (2016), pp. 14.

[8] R. Bonnevilleand V. X. Jin, “A Hidden Markov Model to Identify Combinatorial Epigenetic Regulation Patterns for Estrogen Receptor α target genes”, Advance Access Publication, vol. 29, no. 1, (2013), pp. 22-28.

[9] L. Bottolo and S. Richardson, “Discussion of ‘Gene hunting with hidden Markov model knockoffs’”, Biometrika, vol. 106, no. 1, (2019), pp. 19-22.

[10] S. R. Eddy, “Multiple Alignment using Hidden Markov Models”, ISMB-95, AAAI Press, Menlo Park, CA, (1995), pp. 114-120.

[11] S. R. Eddy, “Hidden Markov Models. Current Opinion in Structural Biology”, vol. 6, (1996), pp. 361-365.

[12] V. D. Fonzo, F. Aluffi-Pentini and V. Parisi, “Hidden Markov Models in Bioinformatics”, Current Bioinformatics, vol. 2, (2007), pp. 49-61.

[13] N. Gautham, “Bioinformatics Databases and Algorithms”, Narosa Publishing House Pvt. Ltd, (2006).

[14] A. Krogh, “An Introduction to Hidden Markov Models for Biological Sequences”, Computational Methods in Molecular Biology, (1998), pp. 45-63.

[15] I. I. Mandoiu and A. Zelikovsky, “Bioinformatics Algorithms Techniques and Applications”, John Wiley &Sons, Inc., Hoboken, New Jersey, (2008).

[16] S. Mulia, D. Mishra, andT. Jena, “Profile HMM based Multiple Sequence Alignment for DNA Sequences”, International Conference on Modelling Optimization and Computing (ICMOC-2012), Procedia Engineering, vol. 38,(2012), pp. 1783-1787.

[17] R. Nath and R. Jain, “Parameter Estimation of Hidden Markov Models using Go with the Winner Algorithms”, International Journal of Computer Applications, vol. 18, no. 5, (2011), pp. 11-15.

[18] S. F. Nimmy, M. G. Sarowar, N. Dey, A. S. Ashour and K. C. Santosh, “Investigation of DNA discontinuity for detecting tuberculosis”, Journal of Ambient Intelligence and Humanized Computing, (2018), pp. 1-15.

[19] B. K. Petersen, M. B. Mayhew, K. O. E. Ogbuefi, J. D. Greene, V. X. Liu and P. Ray, “Modeling Sepsis Progression using Hidden Markov Models”, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, (2017), pp. 1-5.

[20] T. Przytycka and J. Zheng, “Hidden Markov Models”, In: eLS. John Wiley & Sons, (2006), DOI: 10.1002/9780470015902.a0005267.pub2.

[21] L. R. Rabiner, “A tutorial on Hidden Markov Models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77, no. 2, (1989), pp. 257-286.

[22] N. Ramanathan, “Application of Hidden Markov Models”, University of Maryland, (2006), CMSC 828 J.

[23] C. Reilly, “Statistics in Human Genetics and Molecular Biology”, CRC Press, Taylor & Francis Group, Boca Raton, London, New York, (2009).

[24] Y. Sakakibara, “Pair Hidden Markov Models on Tree Structures”, Oxford University Press, vol. 19, no. 1, (2003), pp. i232-i240.

[25] R. Sasikumar and V. Kalpana, “Profile Hidden Markov Model for Sequence Alignment to Cancer Sequence”, Global Journal of Pure and Applied Mathematics, vol. 11, no. 5, (2015), pp. 3665-3675.

[26] R. Sasikumar and V. Kalpana, “Hidden Markov Model in Biological Sequence Analysis - A Systematic Review”, International Journal of Scientific and Innovative Mathematical Research, vol. 4, no. 3, (2016), pp. 1-7.

[27] E. L. L. Sonnhammer, G. Von Heijne and A. Krogh, “A Hidden Markov Model for Predicting Transmembrane Helices in Protein Sequences”, In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, American Association for Artificial Intelligence Press, Menlo Park, CA, (1998), pp. 175-182.

[28] M. Stanke and S. Waack, “Gene Prediction with a Hidden Markov Model and a New Intron Submodel”, Oxford University Press, vol. 19, no. 2, (2003), pp. ii215-ii225.

[29] I. A. Tamposis, K. D. Tsirigos, M. C. Theodoropoulou, P. I. Kontou, and P. G. Bagos, “Semi-supervised learning of Hidden Markov Models for biological sequence analysis”, Bioinformatics, vol. 35, no. 13, (2018), pp. 2208-2215.

[30] B. J. Yoon, “Hidden Markov Models and their Applications in Biological Sequence Analysis”, Current Genomics, vol. 10, (2009), pp. 402-415.