CHAPTER VI

SEQUENCE ALIGNMENT USING HIDDEN MARKOV MODELS

6.1 INTRODUCTION

Modeling in bioinformatics means any representation that simulates biological processes. Bioinformatics techniques promise to provide information that explores the function of life, explains and improves the treatment of inherited and acquired diseases, aids in fighting famine and world hunger and helps to a better environment. To modeling the structure of biological processes, it is better to understand how these structure and process perform their own functions. Furthermore, these models can be used for

- Designing compounds to modify or enhance the function of biological molecules for medical and industrial purposes.
- Simulating a biological process for decision making purposes.
- Better understanding of the process enables to make any changes in the condition of the experiment within the biological process and observe the outcomes.

In many biological processes, bioinformatics is a very important component to translate new information about genes and proteins into improved health care. Bioinformatics tools are very handy for the researchers to collect huge amounts
of information about biological structures, mechanisms and systems, human health, and to find underlying patterns and relationships which are helpful to know how biological processes works. To overcome the challenging task, many statistical methods and applications have been extensively used with the help of stochastic modeling, represents something that may be either an abstract or a concrete entity. The study of Markov Chains (MCs) was initiated in early 1900s by Markov, who laid the foundation for the theory of stochastic processes.

A Markov process is a particular case of stochastic process, where the state at every time belongs to a finite set, the evolution occurs in a discrete time and the probability distribution of a state at a given time is explicitly dependent only on the last states and not on all the others. Thus, Markov chains can be used to model any biological sequences in bioinformatics setup. For illustration, consider a stationary first order Markov chain ‘M’ (a Markov chain is a first order Markov process for which the probability of the future state is directly dependent only on the present state and the past states are irrelevant once the present state is given) and Let $A_{u\rightarrow v} = \text{Pr}(X_t = v | X_{t-1} = u)$, where $A$ is called the transition probability matrix for M. The dimensions of the matrix $A$ are $|S| \times |S|$, where $S$ denotes the state space. For DNA sequences, $A$ denotes the order 4 x 4. Then the probability that the sequence $v = (v_0, v_1 \ldots v_t)$ was generated by M is,

$$
\text{Pr}(v_0, v_1, \ldots v_t) = \text{Pr}(v_0) A_{v_0, v_1} A_{v_1, v_2} \ldots A_{v_{t-1}, v_t} = \text{Pr}(v_0) \prod_{j=1}^{t} A_{v_{j-1}, v_j} \quad \ldots (6.1)
$$
where \( \Pr(v_0) \) is estimated by the frequency of \( v_0 \) in the genome; and \( A_{u,v} \) is estimated by \( \frac{N_{u,v}}{N_u} \), where \( N_{u,v} \) is the number of occurrences in the genome and

\[
N_u = \sum_j N_{u,j}.
\]

Earlier, HMMs have been widely investigated as a representation of stochastic functions of MCs [21] [26] [34] [38]. Its initial development was predominated by theoretical reasonings that attempt to solve problems pertaining to the issues of uniqueness and identifiability. HMMs did not gain much popularity until early 1970s, when Baum et al., successfully applied the technique to speech recognition by developing an efficient training algorithm for HMMs [11]. But, recently various signal processing models and algorithms have been used in biological sequence analysis, among which the HMMs are well-known for their effectiveness in modeling the correlations between adjacent symbols, domains, or events and they have been extensively used by Rabiner [73] in speech recognition. Since the tremendous growth of HMMs applied in various fields, HMMs and their variants have been used in gene prediction [64], pairwise and multiple sequence alignment [25], [69], base-calling [57], modeling DNA sequencing errors [59], protein secondary structure prediction [93], RNA structural alignment [95], fast non-coding RNA annotation [90] and many others.
6.2 HIDDEN MARKOV MODEL TERMINOLOGY

A Hidden Markov Model is a generalization of Markov chain, in which each internal state is not directly observable (the term hidden) but emits an observable random output external state, also called emission, according to a given stationary probability law. The HMM consists of two stochastic processes: an invisible process of hidden states and a visible process of observed symbols. The hidden states form a Markov chain, and the probability distribution of the observed symbol depends on the underlying state. Hence, an HMM is otherwise called as doubly-embedded stochastic process [15].

Since HMMs offer a more sophisticated and systematic approach for the study of protein models, it can be visualized as a finite state machine. Like an ordinary profile, the finite state machines typically move through a series of states and produce some kind of output either when the machine has reached a particular state or when it is moving from one state to another state. The HMM generates a protein sequence by emitting amino acids as it advances through a series of states. Each state has a counterpart of amino acid emission probabilities similar to those described in a profile model, and there are also transition probabilities moving from one state to another state.

The Human Genome Project (HGP) provides a vast amount of data available about proteins, DNA, and RNA. The modernized laboratory methods of studying the structures and functions of those molecules are no longer to keep up with the rate of new information. As an effect, molecular biologists and scientists have turned to statistical methods capable of analyzing huge
amounts of data, with the help of computer programs and software tools to implement these methods. Since the HMM is a dynamic statistical profile built from the analysis of a training dataset, its applications in biological sequence analysis involves the following primary roles [94].

- **Sequence Identification**
- **Sequence Classification (PHMM) and**
- **Generation of Multiple Alignments (PHMM)**

### 6.2.1. SEQUENCE IDENTIFICATION

The goal of the Hidden Markov Model is to set apart sequences that match the consensus sequences from those that do not match. The form, length, and location of the target sequence will define the structure of the HMM used for its identification. Since, Hidden Markov modeling is a generic approach to sequences identification, the target sequence must be initiated before the development of the model. The target sequence possesses two kinds of forms such as (i) regional identification and (ii) specific short sequence identification. The former target category is concerned with the location of relatively large regions of genetic code; an example is the gene promoter region. The latter target category forms a small subsequence; an example is a basal promoter element within a gene sequence or a restriction enzyme site. These two forms of targets utter the structure and complexity of the HMM. The identification of regional targets in genomic sequences involves the construction of a trained model via sequence alignment of known target
regions. Identification procedures are based on distinguishing trained targets against non target sequences. In general, any Hidden Markov Model possesses three categories of states, namely, main, insert, and delete that depend on the target sequence composition. The diagrammatic representation of these states is shown in Figure 6.1.

Figure 6.1: Representation of states of Probabilities for a HMM

In above figure, a possible hidden markov model for the protein ACCY is represented as a sequence of probabilities. The numbers in the boxes show the probability that an amino acid occurs in a particular state, and the numbers next to the directed arcs show probabilities which connect the states. The probability of ACCY is shown as a highlighted path through the model. The description of three states are as follows: the main state represents the non-gapped regions of the sequences; the insert states represents regions not indicating with the consensus sequence; and whereas the delete state represents no emissions of respective states.
While applying Hidden Markov Models to nucleic acid sequence identification, the biological data under sequence analysis is very limited to four alphabets such as A, G, C and T. Given only four bases per state, there is a 25% likelihood of a random occurrence matching the consensus base. As a consequence, a consensus base per state has an equivalent background likelihood of 25% at each state. This value represents a very high background to distinguish a consensus score. When applied to protein analysis, the alphabet is expanded by default to 20, accounting for each amino acid. This increase in the size of the alphabet significantly lowers the random likelihood of a unit matching the target at each state of the HMM.

Hence it is concluded that the larger the alphabet, the lower the likelihood of random occurrences matching the consensus sequence under analysis. This is especially applicable for sequences of the considerable length. Also, the greater the target length and the larger the alphabet, the less likely that the target sequence will occur in the biological sequence analysis data.

6.2.2. SEQUENCE CLASSIFICATION

Sequence classification is generally concerned with the identification of protein domains. More specifically, these domains are classified as conserved structures, or discrete functional units. When the sequence classification method is applied to model family identification, Profile Hidden Markov Models (PHMMs) are to be considered. The profile applicability of Hidden Markov Models is suited to many situations and is primarily involved with the
classification of protein sequences into an encompassing family. The term *profile* applies to the structure of the model, which acts as an identification mechanism enhancing the primary features of the family under the biological sequence analysis. In general, the profile contains the specific information relating to sequence composition, insert and delete regions, regions of conservation, and the categorization of residues associated with their respective positions. More details of the applications of the Profile Hidden Markov Models (PHMMs) can be found in the literature (Baldi, et al., 1994) for the identification of the globin, immunoglobulins and kinase family members.

The PHMMs resembles the principles of HMMs with the term *Profile* being attributed to its trained structure representing a profile of the target under biological sequence analysis. The representation of a generalized structure of PHMM is shown in Figure 6.2. The principal steps applied in Profile Hidden Markov Model generation are summarized as below.

- To gain a lucid understanding of the target domain for identification of the following parameters: Length, Composition, Location and Deviation trends.
- To use the most representative dataset available in any web portal.
- To model the variability of the insert regions and consensus regions as main states.
- To use pseudocounts for further analysis.
- To apply Dirichlet techniques for more complex sequences.
The above figure demonstrates the major features of PHMM, derived from multiple sequence alignments. When PHMM is applied to protein sequence analysis, it is noted that there are three kinds of states represented by three different shapes.

The description of the shapes is as follows:

- The squares are called match states, and the amino acids emitted from them form the conserved primary structure of a protein. These amino acids are similar as those in the common ancestor or, if not, are the results of substitutions.

- The diamond shapes are called insert states, which emits amino acids result from insertions.

- The circles are called delete states, used for model deletions.
It includes the transitions from state to state progress from left to right through the model, with the exception of the self-loops on the diamond insertion states. The self-loops allow deletions of any length to fit the model, regardless of the length of the other sequences in the family.

As a whole, it is concluded that the Profile Hidden Markov Model evolved out of the principal steps summarized above will serve as the basis for classification of sequences based on the trained model. Using this framework, the task of candidate sequences to protein families may provide contextual insights into their structures and functions.

### 6.2.3 Generation of Multiple Alignments

The most important use of Hidden Markov Models is to automatically create a multiple alignment from a group of unaligned sequences. The generation of multiple alignments is a task of great significance, with correlations to homologous studies. Thus, multiple alignments is the process of taking a group of sequences and identifying amino acids which are homologous, structurally or functionally similar.

When multiple alignments are applied using Profile HMMS, many algorithms are to considered; however, the Viterbi algorithm is primarily chosen to align candidate sequences to the pre-constructed model or to find the most likely path through the HMM for each sequence. Each match state in the HMM corresponds to a column in the multiple alignment. A delete state is represented by a dash and whereas, amino acids from insert states are either not shown or are displayed in lower case letters.
An alternative method to Viterbi, known as posterior decoding, whereby considering all the positions in the model and calculate the probability that each amino acid occurs at each position in the model. This produces the table of probabilities for all possible pairs of amino acids and positions in the model. From the constructed table, it is possible to find the highest probability path through the model for any protein, based on all possible places that any amino acid can occur. The key steps in performing this process using HMMS are as follows:

- To develop a trained model by studying preformed multiple alignments consisting of the training dataset.
- To use dynamic programming techniques to score candidate sequences using the Viterbi algorithm.
- To include the candidate as part of the multiple alignments generated depending of the score from the Viterbi output.

The representation of a path sequence EGGR, is best described by Viterbi algorithm shown as an illustration in Figure 6.3.

In general, any sequence can be represented by a path through the model. The probability of any sequence, given the model, is computed by multiplying the emission and transition probabilities along the path. For the above sequence EGGR, there may exits more than one path through the model. In a real model, many different state paths through a model can generate the same sequence. Therefore, the correct probability of a sequence is the sum of probabilities overall of the possible state paths.
Figure 6.3: Representation of HMM with Multiple paths

To determine the most likely path through the specified model, the Viterbi algorithm is employed in the form of a component matrix shown in Figure 6.4. The columns of the matrix are indexed by the states in the model, and the rows are indexed by the sequence. The deletion states does not play a significant role, since by definition, they have a zero probability of emitting an amino acid. The elements of the matrix are initialized to zero and the initial steps specified below are applied.

- $P(EI0) = (\text{Start trans x EI0emiss})$ as the first element
- $P(GI1) = ((I0 \rightarrow D0)\text{trans x (D0}\rightarrow I1)\text{trans x (GI1emiss))$ & $P(GM1) = ((I0\rightarrow M1)\text{trans x (GM1emiss))}$
- $PMAX(I1,M1)$ back pointer set to I0
Repeat (2)-(3) substituting amino acids until matrix completes.

Table 6.1: Matrix Representation Using Viterbi algorithm with Multiple paths

<table>
<thead>
<tr>
<th>sequence</th>
<th>I₀</th>
<th>I₁</th>
<th>M₁</th>
<th>I₂</th>
<th>M₂</th>
<th>I₃</th>
<th>M₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>0.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.003</td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.16</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.021</td>
<td>0.32</td>
</tr>
</tbody>
</table>

As a result, this section concludes that the HMMs represent a framework for identification and classification of biological data. Statistically, the multiplicative product of the sequence probability states defines the score for the sequence where the composition of each state acts independently of the others. This assumption of independence is not necessarily true. For example, the amino acids of similar properties are clustered together, and hydrophobic regions highlight this problem. Identification of biologically significant chain interactions within the polypeptide chain cannot be carried out with models in this form. Therefore, internal interactions representing hydrogen bonding or disulphide bridging cannot be predicted and the modeling of a target family represents a generic profile allowing for the classification of candidate sequences.

Multiple alignment generation is considered to be highly time consuming but can be executed efficiently using the PHMM.
In most cases of applications of hidden markov models in bioinformatics, a fictitious inversion occurs between causes and effects when dealing with emissions. The basic definition and underlying assumptions are as follows.

**DEFINITION**

Let us formally define a HMM [15], denote by the observed symbol sequence $X = x_1, x_2 \ldots x_L$ and the underlying state sequence $Y = y_1, y_2 \ldots y_L$, where $y_n$ is the underlying state of the $n^{th}$ observation $x_n$. Each symbol $x_n$ takes a finite number of possible values from the set of observations $O = O_1, O_2 \ldots O_N$, and each state $y_n$ takes one of the values from the set of states $S = \{1, 2 \ldots M\}$, where $N$ and $M$ denotes the number of distinct observations and the number of distinct states in the model, respectively.

Since the hidden markov model is a statistical model used to describe the evolution of observable events, the sequence is assumed to be a time homogeneous first order markov chain. It implies that the probability of entering state $j$ in the next time point depends only on the current state $i$, and the probability does not change over time. Therefore,

$$P\{y_{n+1} = j | y_n = i, y_{n-1} = i_{n-1}, \ldots, y_1 = i_1\} = P\{y_{n+1} = j | y_n = i\} = t(i, j) \quad \ldots (6.2)$$

for all states $i, j \in S$ and $\forall n \geq 1$. The fixed probability for making a transition from state $i$ to state $j$, is called the transition probability, denoted by $t(i, j)$. For the initial state $y_1$, the initial state probability is denoted as $p(i) = P\{y_1 = i\}$,
\( \forall i \in S. \) Then, the probability that the \( n^{th} \) observation will be \( x_n = x \), depends only on the underlying sequence state \( y_n \).

Hence,

\[
P\{x_n = x | y_n = i, y_{n-1}, x_{n-1}, \ldots \} = P\{x_n = x | y_n = i\} = e(x|i)
\]  

for all possible observations \( x \in O \), all state \( i \in S \), and all \( n \geq 1 \). This is called the emission probability of \( x \) at state \( i \), denoted by \( e(x|i) \). Then, the three probability measures \( t(i, j), \pi(i) \), and \( e(x|i) \) completely specifies an HMM. Based on these parameters, the probability that the HMM will generate the observation sequence \( X = x_1, x_2 \ldots x_L \) with the underlying state sequence \( Y = y_1, y_2 \ldots y_L \) is computed. Then, the joint probability of \( P\{x, y | \Theta\} \) is computed as below.

\[
P\{x, y | \Theta\} = P\{x|y, \Theta\} \cdot P\{y | \Theta\}
\]

where,

\[
P\{x|y, \Theta\} = e(x_1 | y_1) e(x_2 | y_2) \ldots e(x_L | y_L)
\]

\[
P\{y | \Theta\} = \pi(y_1) t(y_1, y_2) \ldots t(y_{L-1}, y_L)
\]

**BASIC ASSUMPTIONS**

More details of the applications of HMMs in bioinformatics have been reported in following literature – Thompson (1995); Churchill (1989); Choo, K.H., Jong, J.C and Zhang, L. (2004); Eddy, S.R. (1995). The main assumptions occurring in the use of Hidden Markov Models are described as follows:
Let $Q = \{q_t\}_{t=0}^T$ be the hidden state sequence in the specified interval $0 \leq t \leq T$, where $q_t \in \mathcal{S}$. There are three major assumptions made in the analysis of HMM problems which are,

1. **The Markov assumption:**

   This assumption states that the next state depends only on the current state, which means that the transition probabilities are defined by,
   \[
   P(q_{t+1} = j | q_t = i, q_{t-1} = l, q_{t-2} = m, ... q_0 = n) = P(q_{t+1} = j | q_t = i) = P_{ij} \tag{6.7}
   \]

   In general, the next state might depend on the past $k$ states, thereby giving rise to a $k^{th}$ order hidden markov model. However, such models are more complicated to analyze than the preceding first order HMMs.

2. **The Stationarity assumption:**

   This assumption states that the state transition probabilities are independent of the actual time the transitions takes place. Thus, for any two times $t_1$ and $t_2$,
   \[
   P(q_{t_1+1} = j | q_{t_1} = i) = P(q_{t_2+1} = j | q_{t_2} = i) = P_{ij} \tag{6.8}
   \]

3. **The Observation independence assumption:**

   This assumption states that the current observation or output is statistically independent of the previous observations. This, if the observation sequences are $O = v_1, v_2, \ldots, v_T$, then
   \[
   P(O | q_1, q_2, \ldots, q_T, \lambda) = \prod_{t=1}^{T} P(v_t | q_t, \lambda) \tag{6.9}
   \]

   When the above three assumptions holds, then the joint probability distribution of $P(Q, O)$ is obtained by,
\[
P(Q,O) = \prod_{i=1}^{T} P(q_i \mid q_{i-1}) P(v_i \mid q_i)
\]  

6.3. A SIMPLE HMM FOR MODELING GENES

Based on the earlier literature works, it is clear that HMMs can be effectively used for representing biological sequences. For illustration, let us consider an HMM that models protein-coding genes in eukaryotes. It is known that many protein-coding regions display codon bias. The non uniform usage of codons results in different symbol statistics for different codon positions [13], and it is a source of the period-3 property in the coding regions [85]. Therefore, it is essential to incorporate these codon statistics when modeling protein-coding genes and building a gene finder.

Figure 6.4 shows a toy HMM for modeling eukaryotic genes. The given hidden markov model seeks to capture the statistical difference in exons and introns. In this process, the HMM has four states, where E_1, E_2, and E_3 are used to model the base statistics in exons. Each E_k uses a different set of emission probabilities to reflect the symbol statistics at the kth position of a codon. The state I is used to model the base statistics in introns. In this example, if the structures and important characteristics of the biological sequences are known, then building the corresponding HMM is quietly simple and ease.

To illustrate the use of HMM for modeling protein coding genes, let us consider a new DNA sequence \( X = x_1 \ldots x_{19} = \text{ATGCAGCTGCATAGCACTT} \). Initially, the two common questions arises to fulfill the objective, that is to find out whether this DNA sequence is a coding gene or not, and how can predict
the locations of the exons and introns in the given sequence? Now, the solution to the first question is by computing the observation probability of $x$ based on the given HMM that models coding genes. If this probability is high, it implies that this DNA sequence is likely to be a coding gene. Otherwise, it is concluded that $x$ is unlikely to be a coding gene, since it does not contain the statistical properties that are typically observed in protein-coding genes.

![Figure 6.4: A Toy HMM for Modeling Eukaryotic Genes](image)

Next, the solution to the second question is about first predicting the state sequence $y$ in the HMM that best describes $x$. Once the best $y$ is inferred, it is straightforward to predict the locations of the exons and introns. For example, assume that the optimal state sequence $y$ is as shown in Figure 6.5. This implies that the first nine bases $x_1, \ldots x_9$ belong to the first exon, the following four bases $x_{10}, \ldots x_{13}$ belong to an intron, and the last six bases $x_{14}, \ldots x_{19}$
belong to another exon. Hence, the above example clearly states the use of HMMs, a formal probabilistic framework for analyzing biological sequences.

### 6.4. CONSTRUCTION OF HIDDEN MARKOV MODEL

Another tricky problem is how to create an HMM in the first place, given a particular set of related training sequences. It is necessary to estimate the amino acid emission distributions in each state and all state-state transition probabilities from a set of related training sequences.

If the state paths for all the training sequences are known, the emission and transition probabilities in the model can be calculated by computing their expected value; observing the number of times each transmission or emission occurs in the training set and dividing by the sum of all the transition probabilities or all the emission probabilities.

If the state paths are unknown, finding the best model given the training set is an optimization problem which has no closed form solution. It must be solved by the iterative methods.

The algorithms used to do this are closely related to the scoring algorithms and the goal is to find the model parameters which maximize the probability of all sequences in the training data set. In other words, the desired model is a model against which all the sequences in the training set will have the best possible scores. The parameters are re-estimated after every iteration by computing a score for each training sequence against the previous set of model parameters.
The Baum-Welsch algorithm is a variation of the forward algorithm, where it begins with an initial model and then calculates a score for each sequence in the training set over all possible paths through this model. During the next iteration, a new set of expected emission and transition probabilities is calculated, as described above for the case when state paths are known. The updated parameters replace those in the initial model, and the training sequences are scored against the new model.

As in many iterative methods, convergence indicates only that a local maximum has been found. Several heuristic methods have been developed to deal with the problem and one approach is to begin with several initial models and proceed to build several models in parallel. When the models converge at several different local optima, the probability of each model given the training set is computed, and the model with the highest probability prevails.

6.5. VARIANTS OF HIDDEN MARKOV MODELS

The basic concepts of hidden markov models and its applications in biological sequence analysis were discussed in preceding sections of this chapter. After addressing the fundamentals of HMMS, this section introduces three types of HMM variants that modify and extend the basic model to meet the needs of various applications in biological sequence analysis.

The three types of HMM variants [15] are,

- Profile – Hidden Markov Models
- Pair – Hidden Markov Models and
- Context-Sensitive Hidden Markov Models
Among them, this section addressed the first two commonly used hidden markov model variants, which have been widely useful in various sequence analysis problems and its applications play a significant role in bioinformatics setup.

6.5.1. PROFILE HMMs AND ITS APPLICATIONS

Let us consider a multiple sequence alignment of proteins or DNA sequences that belong to the same functional family. To build an HMM that can effectively represent the common patterns, motifs, and other statistical properties in the given alignment, then one model that is especially useful for representing the profile of a multiple sequence alignment is the Profile-HMM. These are HMMs with a specific architecture that is suitable for modeling sequence profiles. Unlike ordinary HMMs, Profile-HMMs have a strictly linear left-to-right structure that does not contain any cycles. It generally possess three types of hidden states, namely, match states, insert states, and delete states to describe position-specific symbol frequencies, symbol insertions, and symbol deletions respectively.

For illustration, let us consider a profile hidden markov model based on the multiple alignments shown in Figure 6.5 (a). From the below figure, it is seen that the given alignment has five columns, where the base frequencies in the respective columns are different from each other. The \( k \)th match state \( M_k \) in the profile-HMM is used to describe the symbol frequencies in the \( k \)th column of the alignment. It is called a 'match' state, since it is used to represent the case when a symbol in a new observation sequence matches the symbol in the
consensus sequence of the original alignment. As a result, the number of
match states in the resulting profile-HMM is identical to the length of the
consensus sequence. The emission probability \( e(x|M_k) \) at the \( k^{th} \) match state
\( M_k \) reflects the observed symbol frequencies in the \( k^{th} \) consensus column. By
interconnecting the match states \( M_1, M_2 \ldots M_5 \), we obtain an ungapped HMM as
shown in Figure 6.5 (b). This ungapped HMM can represent DNA sequences
that match the consensus sequence of the alignment without any gap, and it
serves as the backbone of the final profile-HMM that is to be constructed.

Once the ungapped HMM is constructed, then add insert states \( I_k \) and
delete states \( D_k \) to the model for insertions and deletions in new observation
sequences. Let us first consider the case when the observed DNA sequence is
longer than the consensus sequence of the original alignment. In this case, if
the sequences are aligned, there will be one or more bases in the observed DNA
sequence that are not present in the consensus sequence. These additional
symbols are modeled by the insert states. The insert state \( I_k \) is used to handle
the symbols that are inserted between the \( k^{th} \) and the \((k + 1)^{th}\) positions in the
consensus sequence.

Now, let us consider the case when the new observed sequence is shorter
than the consensus sequence. In this case, there will be one or more bases in
the consensus sequence that are not present in the observed DNA sequence.
The \( k^{th} \) delete state \( D_k \) is used to handle the deletion of the \( k^{th} \) symbol in the
original consensus sequence. As delete states represent symbols that are
missing, \( D_k \) is a non-emitting state, or a silent state, which is simply used as a
place-holder that interconnects the neighboring states. After adding the insert states and the delete states to the ungapped HMM in Figure 6.5(b), then the final profile-HMM is obtained and shown in Figure 6.5(c).

Due to the expediency and effectiveness in representing sequence profiles, Profile-HMMs have been widely used for modeling and analyzing biological sequences. Recently, there exits publicly available software packages, such as HMMER [25] and SAM [53], that can be used to build and train suitable
models. To have a library of ready-made Profile HMMs for known sequence families, currently two libraries such as PROSITE database [42] [78] and the Pfam database [30] [80] have compiled a large number of Profile-HMMs for various protein families.

The primary goal is to search a sequence database to determine additional homologues that belong to the same family. For example, by querying a new protein sequence against Pfam or PROSITE, it can be easily find out whether the sequence contains any of the known protein domains.

In addition, the Profile HMMs are used to compare two multiple sequence alignments or sequence profiles, instead of comparing a single sequence against a multiple alignment or a profile. For example, COACH [90] allows to compare sequence alignments, by building a Profile-HMM from one alignment and aligning the other alignment to the constructed Profile –HMM.

The other program called HHSearch and PROFILE COMPARER were applied for generalizing the traditional pairwise sequence alignment algorithm using two Profile-HMMs, and provides a tool for scoring and aligning sequences produced by popular software tools such as HMMER and SAM.

6.5.2. PAIR HMMs AND ITS APPLICATIONS

In biological sequence analysis, it is inevitable to compare two sequences to find out whether these sequences are functionally related. Sequence similarity is often a good indicator of their functional relevance, and methods for quantitatively measuring the similarity of two proteins or DNA sequences
have been of interest to many researchers. A classical approach for comparing two biological sequences is to align them based on their similarity, compute their alignment score, and evaluate the statistical significance of the predicted alignment.

The Pair-Hidden Markov Model is a variant of the basic HMM that is especially useful for finding sequence alignments and evaluating the significance of the aligned symbols. Unlike the original hidden markov model, which generates only a single sequence, a Pair-HMM generates an aligned pair of sequences. The primary intention of pair hidden markov models is fulfilled in two aspects [15], that is

- To find the best alignment between the sequences, it is crucial to define a reasonable scoring scheme for ranking different alignments.
- Based on the respective scoring scheme, it is highly remarkable to choose the alignment that maximizes the alignment score.

Since a pair hidden markov model generates an aligned pair of sequences, a probabilistic framework of this model consists of three parameters, depicted in Figure 6.6. The constructed pair hidden markov model was then used to find the optimal sequence alignment, computation of overall alignment probability, and estimation of the reliability of the individual alignment regions. Using pair hidden markov models, to describe specific indel length distributions has been shown to be very useful for finding accurate alignments of non-coding DNA sequences.
From the above figure, it is clear that the model possess three parameters described as follows,

**Three states: M, I, J**

- State $M$ matches one letter from each sequence
- State $I$ inserts a gap to the second sequence
- State $J$ inserts a gap to the first sequence

**Emission probabilities**: $P_{x_i,y_j}, Q_{x_i}, Q_{y_j}$ and

**Transition probabilities**: $\delta$ and $\epsilon$.

Also, the model contains a start state and a final state. From the start state, there is initial transition probability $d$ from the start state to state $I$ and $J$, and $1-2d$ to $M$. In addition, in order to terminate, there is transition probability $t$ from state $M, I$ and $J$ to the final state and from final state back to itself. In the
case that we have to take into account of $t$, the transition probabilities between states might have to adjust accordingly.

For illustration, let us consider a simple topology of a Pair - Hidden Markov Model [15], which traverses between the states $I_x$, $I_z$, and $A$, to simultaneously generate two aligned sequences $x$ and $z$, shown in Figure 6.7.

![Diagram of Pair - Hidden Markov Model](image)

<table>
<thead>
<tr>
<th>$I_x$: insertion in $x$ (seq 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_z$: insertion in $z$ (seq 2)</td>
</tr>
<tr>
<td>$A$: aligned symbols in $x$ and $z$</td>
</tr>
</tbody>
</table>

$x$ (seq 1) : T T C C G − −  
z$ (seq 2) : − − C C G T T  

$y$ (states): $I_x$ $I_x$ $A$ $A$ $A$ $I_z$ $I_z$

**Figure 6.7: Illustration of a Pair – Hidden Markov Model**

From the above figure, it is seen that a pair-hmm generates an aligned pair of sequences simultaneously, where the underlying state sequence $y$, uniquely determines the pairwise alignments between $x$ and $z$. Also, it is noted that the state $A$ generates an aligned pair of two symbols $x_i$ and $z_j$, where $x_i$ is inserted in $x$ and $z_j$ is inserted in $z$. And $x_1$ and $x_2$ are individually emitted at $I_x$; hence they are not aligned to any bases in $z$. The pairs $(x_3, z_1)$, $(x_4, z_2)$ and $(x_5, z_3)$ are jointly emitted at $A$, and therefore the bases in the respective pairs are aligned to each other. Thus, $z_4$ and $z_5$ are individually emitted at $I_z$, as unaligned bases.
Finally, based on the pair-hmm framework, the primary task of finding the best alignment between sequences $x$ and $z$ reduces to the problem of finding the following optimal state sequence,

\[
y^* = \arg \max_{y} P\{y|x, z, \Theta\}
\]  \hspace{1cm} \ldots (6.11)

Then, the optimal state sequence $y^*$ can be found using dynamic programming approach, by a simple modification of the Viterbi algorithm [25]. The computational complexity of the resulting algorithm leads $O(L_x, L_z)$, where $L_x$ and $L_z$ are the lengths of two sequences $x$ and $z$ respectively.

Furthermore, pair hidden markov models have many potential advantages and extensively used by various publicly available softwares and tools. Some of them are as follows:

More recently, a method called **MCALIGN2** [88] adopted pair-hmms with a slightly different structures, for global pairwise alignment of non-coding DNA segments. Another state-of-the-art multiple sequence alignment algorithm called **ProbCons** [24] uses a pair-hmm to compute the posterior alignment probabilities. Pair-HMMs have been widely used for gene prediction. For example, **Pairagon+N-SCAN_EST** provides a convenient pipeline for gene annotation by combining a pair-HMM using the prediction algorithm [6].

In addition, the **GPHMM** [69] provides a convenient probabilistic framework for comparative gene prediction by combining the pair-HMM and the generalized HMM. The comparative gene finders such as **SLAM** [01] and **TWAIN** [61] are implemented based on GPHMM used to compare two DNA sequences and jointly analyze their gene structures and functions.
CONCLUSION

In this chapter, a tutorial review of Hidden Markov Models with their applications in biological sequence analysis and problems in molecular biology is well presented with illustration. The chapter especially focused two variants of Hidden Markov Models such as Profile-HMMs and Pair-HMMs with illustration in sequence analysis. The chapter ends with the applications of variants of HMMs, which have been extensively used for finding pairwise alignment of proteins and DNA sequences. Also, it focused various publicly available software packages used for modeling and analyzing biological sequences.