Non-Homogeneous Hidden Markov Model (NNHMM) for better phylogeny estimation and multiple sequence alignment of sequence

Qingyuan Liu

1. Introduction

Hidden Markov Model (HMM), since this idea has been developed, has been widely used for homology detection, protein family assignment, multiple sequence alignment (MSA) and phylogenic placement etc. due to its strong statistical foundation and flexibility. The HMM idea is first proposed by Ruslan L. Stratonovich[1] in 1960, and the mathematics behind HMM is developed by L.E.Baum[2,3,4,5,6] over the course of 6 years from 1966 to 1972. HMM is a statistical model that modeling a system assumed to be a Markov process with unobserved states (hidden states)[7]. A loose definition for Markov process is that a process that is memoryless, as what is going to happen next does not depend on previous states; it only depends on present states. In order to let the model be functioning, the parameters of the model are usually trained over a set of sequences. A lot of models were built based on HMM theory such as homogeneous (Classical) HMM, which is the easiest and the most constrained HMM; profile HMM, widely used in MSA of protein sequence; and NHHMM, which is one of the less constrained HMM. We are going to discuss all three models in detail in the later chapters.

2. Profile HMM

Krogh et al. introduced profile HMM idea in 1994[8]. The profile HMM was developed as a generalized statistical model for a family of proteins such that for any sequence of amino acids, the model would provide a probability of generating that specific sequence using the model in a way that the probability the model defined would be higher if amino acids sequences are within the protein family. For a consensus column, a 'match' state describes the distribution state of the residues. An 'insert' state allows inserting a new residue into the sequence while 'delete' state allows skipping one of the 'match' states. The model is illustrated in Figure 1[8]. There are two parameters in the model. One is the transition probability between states, the other is the emission probability of emitting certain residue. The probability of generating a sequence of amino acids x_1, x_2, … x_L by following a path of states q_0, q_1, . . . q_N, q_{N+1} through the model is defined as equation 1[8]:

$$Prob(x_1 \ldots x_l, q_0 \ldots q_{N+1}|model) = P(M_{N+1}|q_N) \times \prod_{i=1}^{N} p(q_i|q_{i-1}) \cdot p(x_{l(i)}|q_i) \quad (1)$$

where l(i) denotes the index of in the sequence x_1 … x_L. $P(M_{N+1}|q_N)$ denotes the last transition probability to the end state M_{N+1} from q_N. $p(q_i|q_{i-1})$ denotes the transition probability from q_{i-1} state to q_i state. $p(x_{l(i)}|q_i)$ denotes the emission probability of emitting $x_{l(i)}$ residue at state q_i. For delete state, the emission probability parameter would always be 1.

As there can be multiple pathways of generating sequences from the model, with each pathway has different probability, the most probable pathway and the highest probability of generating this sequence using this specific parameter can be both found. An algorithm called Viterbi algorithm would provide the pathway with the highest probability.

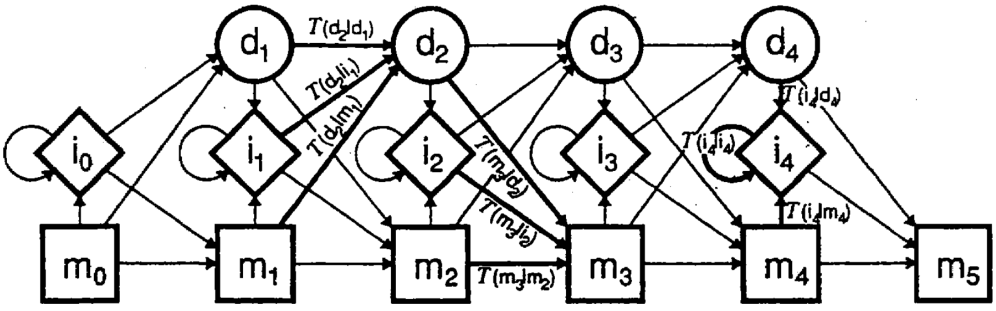

Figure 1. Profile HMM. M for match state, i for insert state and d for delete state.

To estimate parameters of the profile HMM, either the parameters is handcrafted or use Expectation-Maximization (EM) algorithm to generate the parameters of the model: transition probability and emission probability. The later method requires some existed and related sequences (seed sequences) to generate all the parameters. The model needs to be 'trained' on a very large set of seed sequences to get a relatively good parameter for the model. As its requirement for seed sequences, finding a good set of seed sequences for training the model can sometimes be very time-consuming. A new model called "Ultra-large multiple sequence alignment using Phylogeny-aware Profiles" (UPP) developed by *Nguyen, Mirarab, and Warnow* using a random subset of sequences to build a "backbone alignment" and construct a profile HMM on the backbone alignment. This model improved the original profile HMM as it does not require picking seed sequences separately beforehand. More EM algorithm would be discussed in section 3.

So far, by using profile HMM, a protein can be determined for its family it belongs to, as a different family would have different parameters for the profile HMM and it would generate different most probable pathway.

3. Homogeneous HMM and difference comparing to profile HMM

The homogeneous HMM or the classical homogeneous HMM (sometimes be considered as the default/basic model of HMM) are considered to be the simplest form of HMM. The homogeneous HMM process consists of two processes, a hidden process which evolves under first-order Markov process (the outcome of next state may only depend on current state), and an observed process which is only dependent on the current hidden state. The hidden process can be viewed as the 'Match'/'Delete'/'Insert' state in the profile HMM with transition probability in between. The observed process is the emission probability of the regarded states, such as 'Match' or 'Delete' or 'Insert'. A graphical model is illustrated on figure 2[9].

Two assumptions lie under the HMM:

$$P(z_t|z_{1:t-1}) = P(z_t|z_{t-1}) \qquad (2)^9$$

$$P(y_t|z_{1:t}) = P(y_t|z_t) \qquad (3)^9$$

C_t is the hidden state at t, $C_{1:t-1}$ is the hidden state sequence, C_1, C_2 ... C_{t-1}. X_t is the observed state at t, $X_{1:t-1}$ is the observed state sequence, X_1, X_2 ... X_{t-1}.

All the parameters of the models are trained using an EM algorithm named Baum-Welch algorithm. The Baum-Welch algorithm is, in essence, an EM algorithm but specifically for solving HMM. The algorithm iteratively calculates the parameter of the model until the desired level of convergence.

One thing that is different between basic HMM and profile HMM is that: after the model has been set up, profile HMM can have a unlimited number of states for a pathway due to the inner-loop of the insert state while homogeneous HMM only has a limited number of states. Therefore, if basic HMM is used for protein sequence alignment, it has limitation over sequences with long consecutive inserting states.

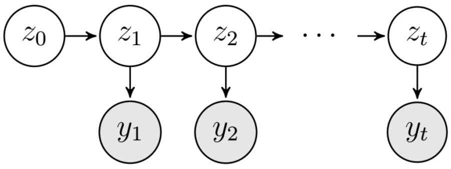

Figure2: An HMM underlying the sequence of data values (z_0, ..., z_t). z_i is the hidden state for the observation y_i, $P(y_i|z_i)$ is the probability of emission of y_i in state z_i, and $P(z_i|z_{i-1})$ is the probability of transition from state z_{i-1} to state z_i.

For both of the methods, the additional limitation would be the homogeneity of the model. All insert state has the same emission parameter. And for classic HMM, the transition probabilities between states are fixed too. However, it is definitely not the case for real biology. Some part of the genetic sequence or protein sequence has a different profile comparing to the rest of the sequence. Taking a very common example, there is a promoter region in genetic sequence near the start of the gene called CpG island where this region has a higher than normal C-G emission frequency. Using an 'insert' state's emission parameter that is trained from an overall sequence would definitely not give out an increased in C or G emission probability. Some protein sequences have a lot of acidic or basic amino acid residue near the activation site for providing redox reaction advantage. The whole sequence profile cannot efficiently represent the profile of some specific region.

For a better model, we need to do the HMM non-homogeneously.

4. Non-Homogeneous HMM (NHHMM)

The classic HMM has fixed transitional parameter and emission parameter for all states. However, for a lot of cases, not only in biology, such as geology (precipitation)[11], all the parameter is not fixed and it will affect by some other factors. Therefore, the parameters need to be changed based on some covariate. In other words, a set of observed factor (covariate) would have effects on the model's parameter. To extend this idea, two possible NHHMM model are possible. The first model is that the covariate only affects the transition probability; the other model extends from the first model where the covariate can affect both transition probability and emission probability. The special case of covariate only affects emission probability lies beneath the second model of the non-homogeneous models. Both non-homogeneous models are illustrated in figure 3[9]:

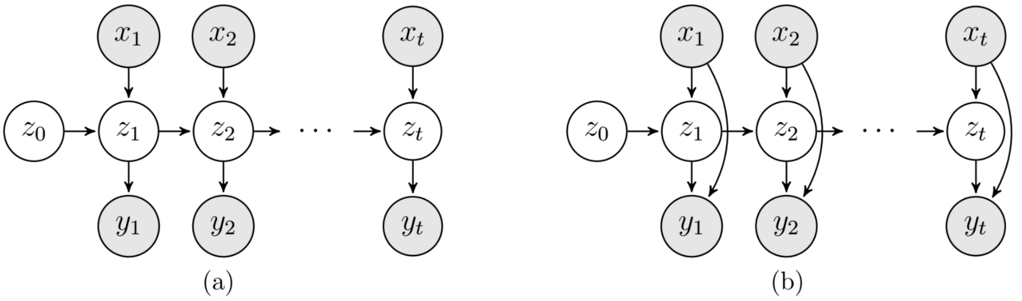

Figure 3: a) first non-homogeneous HMM, with X covariate only affect transition probability. b) second non-homogeneous HMM with X covariate affect both transition probability and emission probability.

For the NHHMM only affect transition probability, two assumptions is updated from equation (2) and equation (3) to equation (4) and equation (5):

$$P(z_t|z_{1:t-1}, x_{1:t}) = P(z_t|z_{t-1}, x_t) \qquad (4)^9$$

$$P(y_t|z_{1:t}) = P(y_t|z_t) \qquad (5)^9$$

For the NHHMM affect both transition probability and emission probability, the two assumptions is updated as equation (6) and equation (7):

$$P(z_t|z_{1:t-1}, x_{1:t}) = P(z_t|z_{t-1}, x_t) \qquad (6)^9$$

$$P(y_t|z_{1:t}, x_{1:t}) = P(y_t|z_t, x_t) \qquad (7)^9$$

where x_t is the covariate at time t and $x_{1:t}$ is the covariate state sequence $x_1, x_2, \ldots x_t$.

So far, there has not been many successfully implementation of non-homogeneous HMM onto biological data. *Fatemeh Zamanzad Ghavidel et al.*[10] proposed to use NH-HMM to model the DNA sequence with single nucleotide polymorphism (SNP) and the covariate of the system is how far the residue is adjacent to the last occurrence of SNPs. Apart from that, this NHHMM idea is widely used for modeling the precipitation[11].

5. Potential improvement of MSA using NHHMM

One thing to notice is that the HMM that used for MSA is profile HMM. The NHHMM that mentioned early is only non-homogeneous on HMM. A modification is required to convert NHHMM to a non-homogeneous profile HMM.

The reason that non-homogeneous model may get to play is largely due to the extra information in the data provided. There is more in the sequence than the sequence itself. Different region of the genes would have a different profile; protein sub-sequences with a different structure may have a different profile etc. A non-homogeneous model can be visualized as a combination of a few homogeneous models that highly accurately model a specific part of the system.

Although the non-homogeneous model better represents the system in theory, there are still a lot of problems that non-homogeneous model cannot fix. Taking an example of long ranged deletion. Assume a sequence that lost 10 nucleotides due to deletion happened during evolution, which is very common; and each deletion state has a transition probability of 0.5; the probability of the path with 10 consecutive deleting stage is going to be 0.5^{10} which ended up to be a very small probability. And the path with the largest probability may become the wrong one. In order to solve this problem, the transition probability of deletion state at i (D_i) may depend on not only the state in i-1 (D_{i-1},M_{i-1}, I_{i-1}) but state in i-2(D_{i-2},M_{i-2}, I_{i-2}) or even more previous state. As the classical HMM assumed the Markov process is first ordered, the state only depends on its direct previous state, incorporating non-homogeneous model with higher ordered HMM may provide a better way for representing the sequence and better MSA alignment.

Another limitation of using NHHMM is that this model when compared to classic HMM, has the most degree of freedom in it. The higher degree of freedom, the more data the model would require for training its parameters. Insufficient amount of data may result NHHMM in a poorly trained parameters thus does not model the system correctly.

Future Research:

A non-homogeneous profile HMM needs to be constructed based on NHHMM and profile HMM. How to combine both of the models is strictly required. A higher ordered profile HMM can be researched on and combination of the higher ordered profile HMM and NHHMM can be the final ultimate goal. If there were a way to construct a non-homogeneous profile HMM with all its parameters, a comparison study between non-homogeneous profile HMM and homogeneous profile HMM would be worth looking into it. An overall evaluation of non-homogeneous profile HMM would required to see if it could capable of provide a better statistical model for a family of sequences.

Reference:

1. Stratonovich, R. L. "Conditional Markov Processes." *Theory of Probability & Its Applications* 5, no. 2 (1960): 156-78. doi:10.1137/1105015.
2. Baum, Leonard E., and Ted Petrie. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." *The Annals of Mathematical Statistics* 37, no. 6 (1966): 1554-563. doi:10.1214/aoms/1177699147.
3. Baum, Leonard E., and J. A. Eagon. "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology." *Bulletin of the American Mathematical Society* 73, no. 3 (1967): 360-64. doi:10.1090/s0002-9904-1967-11751-8.
4. Baum, Leonard E., and George Sell. "Growth transformations for functions on manifolds." *Pacific Journal of Mathematics* 27, no. 2 (1968): 211-27. doi:10.2140/pjm.1968.27.211.
5. Baum, Leonard E., Ted Petrie, George Soules, and Norman Weiss. "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains." *The Annals of Mathematical Statistics* 41, no. 1 (1970): 164-71. doi:10.1214/aoms/1177697196.
6. Baum, L.E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process". *Inequalities*. 3: 1–8.
7. "Hidden Markov model." Wikipedia. May 01, 2017. Accessed May 03, 2017. https://en.wikipedia.org/wiki/Hidden_Markov_model#cite_note-Stratonovich1960-6.
8. Krogh, Anders, Michael Brown, I.saira Mian, Kimmen Sjölander, and David Haussler. "Hidden Markov Models in Computational Biology." Journal of Molecular Biology 235, no. 5 (1994): 1501-531. doi:10.1006/jmbi.1994.1104.
9. Sarkar, Abhra, Anindya Bhadra, and Bani K. Mallick. "Nonparametric Bayesian Approaches to Non-homogeneous Hidden Markov Models." (n.d.): n. pag. 8 May 2012. Web. 4 Apr. 2017.
10. Ghavidel, Fatemeh Zamanzad, Jargen Claesen, and Tomasz Burzykowski. "A Nonhomogeneous Hidden Markov Model for Gene Mapping Based on Next-Generation Sequencing Data." *Journal of Computational Biology* 22.2 (2015): 178-88. Web.
11. Hughes, J. P., P. Guttorp, and S. P. Charles. "A non-homogeneous hidden Markov model for precipitation occurrence." Journal of the Royal Statistical Society: Series C (Applied Statistics) 48, no. 1 (1999): 15-30. doi:10.1111/1467-9876.00136.