CONSTRUCTION AND APPLICATION BASED ON COMPRESSING DEPICTION IN PROFILE HIDDEN MARKOV MODEL

Zhijian Zhou [1,2] , Daoliang Li [2,3] , Li Li [2,3] , Zetian Fu [2,3,*]

[1] *College of Science, China Agriculture University, Beijing, China, 100083*

[2] *Key Laboratory of Modern Precision Agriculture System Integration, Research Ministry of Education, Beijing, China, 100083*

[3] *College of Information and Electrical Engineering, China Agricultural University, Beijing, China, 100083*

[*] *Corresponding author, Address: P.O.Box 209#, China Agricultural University, East Campus, Beijing 100083, P. R. China, Tel:+86-10-62736323, Email:fzt@cau.edu.cn*

Abstract: A method to express Profile Hidden Markov Model (Profile HMM) parameters with compressing matrix is presented, which is obtained by imposing the characteristics of both the state transfer and the character output in the Profile HMM.

Key words: Profile HMM, Multiple sequence alignment, Bioinformatics

1. INTRODUCTION

The Profile HMM is composed by a Markov chain including matching, insertion, delete states, and an observable stochastic process namely as observation chain. The state chain depicts the transfer relationship among different state. The observation chain represents the statistical association between the state and observation. Generally, the state in Markov process cannot be observed directly. It can only be understood through the observable process. The Profile HMM model was introduced to bioinformatics by Krogh (1994), and now was widely used in the Alignment of biological sequence.

A one grade Profile HHM with matching range as L is illustrated as Figure 1. This is a linear model. It progress only one direction from left to right. There are 3L+1 system states in the system namely as matching (M), insertion (I), and delete (D) respectively. For convenience in coding the program, two extra states, say as beginning (M_0) and ending (M_{L+1}), are involved in the model. They do not output any character to influence the model activity. The character set depends on the concerning object. For example, there are four characters in DNA sequence denoted by A, G, C, and T, and twenty characters for amino acid sequence. In the Figure 1, rectangle, diamond, and circle denote matching, insertion and delete state respectively. The arrow connected different states indicates the state transfer relation and direction. In an ascertain model, the transfer probability and character release between different state is wholly determinate.

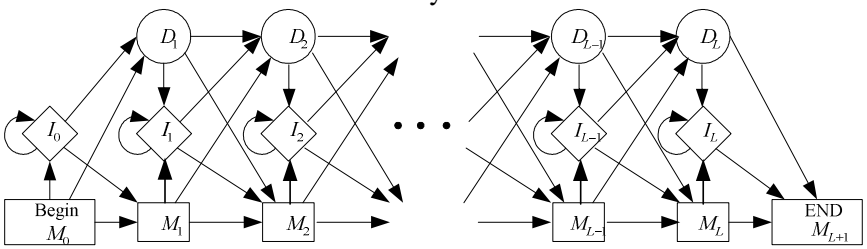

Figure 1 Model of one grade Profile HHM with matching range as L

2. COMPRESSION FORMS OF PROFILE HMM PARAMETERS

Based on the stochastic processes theory, a Profile HHM can be totally ascertained if $\lambda = \lambda(S, \Omega, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ is known. Where S denotes as state set, Ω is observation set, A is state transfer probability matrix, B is character output probability matrix, and π is initial state probability distribution function, respectively. Suppose a Profile HMM has matching range as L. Then the number of elements for S is $N = 3L + 1$. If two extra state, beginning and ending stats, are also included, the number of elements for S become as $N = 3(L+1)$. So the order of state transfer matrix A is $3(L+1) \times 3(L+1)$. It can be found from Figure 1, at any state such as M_l, three front states as inputs of the state are at the most. They locate at same layer, say as l-1. This means only three occasions, $M_{l-1}M_l, I_{l-1}M_l$ and $D_{l-1}M_l$, can turn up when the state transfer from M_{l-1} to M_l . Similarly, only three occasions, $M_l I_l, I_l I_l, D_l I_l$ and $M_{l-1}D_l, I_{l-1}D_l, D_{l-1}D_l$ for I_l and D_l state occur respectively. Remembering none of character output for delete state, the

state transfer probability matrix as order as $3(L+1)\times 3(L+1)$ can be compressed to $9\times (L+1)$ without lost any information:

$$\overline{\boldsymbol{A}}_{9\times(L+1)}=\begin{pmatrix} a[M_0M_1] & a[M_1M_2] & \cdots & a[M_{l-1}M_l] & \cdots & a[M_{L-1}M_L] & a[M_LM_{L+1}] \\ a[I_0M_1] & a[I_1M_2] & \cdots & a[I_{l-1}M_l] & \cdots & a[I_{L-1}M_L] & a[I_LM_{L+1}] \\ 0 & a[D_1M_2] & \cdots & a[D_{l-1}M_l] & \cdots & a[D_{L-1}M_L] & a[D_LM_{L+1}] \\ a[M_1I_1] & a[M_2I_2] & \cdots & a[M_lI_l] & \cdots & a[M_LI_L] & 0 \\ a[I_1I_1] & a[I_2I_2] & \cdots & a[I_lI_l] & \cdots & a[I_LI_L] & 0 \\ a[D_1I_1] & a[D_2I_2] & \cdots & a[D_lI_l] & \cdots & a[D_lI_L] & 0 \\ a[M_0D_1] & a[M_1D_2] & \cdots & a[M_{l-1}D_l] & \cdots & a[M_{L-1}D_L] & 0 \\ a[I_0D_1] & a[I_1D_2] & \cdots & a[I_{l-1}D_l] & \cdots & a[I_{L-1}D_L] & 0 \\ 0 & a[D_1D_2] & \cdots & a[D_{l-1}D_l] & \cdots & a[D_{L-1}D_L] & 0 \end{pmatrix}_{9\times(L+1)}$$

Where $S=\{M_0,I_0,M_1,I_1,D_1,M_2,I_2,D_2\cdots,M_L,I_L,D_L,M_{L+1}\}$ are state set, and $a[X_iY_j]$ represents the one step transfer probability from X_i to Y_j. The elements sign as "0" means no corresponding state transfer, and elements in l column denote the probability of one step state transfer ending at l layer. Let $\overline{\mathbf{A}}_{9\times(L+1)}\equiv(\overline{a}_{ij})_{9\times(L+1)}$, $\overline{a}_{ij}=a[X_dY_j]$, $1\le i\le 9$, $1\le j\le L+1$, then following relations can be obtained ,

$$(1)\ i \bmod 3=\begin{cases}1, & X=M\\ 2, & X=I\\ 0, & X=D\end{cases} \tag{2-1}$$

$$(2)\ (d,Y)=\begin{cases}(j-1,M) & if\ 1\le i\le 3\\ (j,I) & if\ 4\le i\le 6\\ (j-1,D) & if\ 7\le i\le 9\end{cases} \tag{2-2}$$

$$(3)\ \overline{a}_{31}=\overline{a}_{91}=\overline{a}_{4(L+1)}=\cdots=\overline{a}_{9(L+1)}=0 \tag{2-3}$$

$$(4)\ a_{ij}+a_{(i+3)(j-1)}+a_{(i+6)j}=1, i=1,2,3;\quad j=2,3,4 \tag{2-4}$$

Similarly, let $\Omega=\{\omega_1,\omega_2,\cdots,\omega_K\}$ is observation character set, then character output probability matrix $\boldsymbol{B}_{K\times(3L+1)}$ can be compressed as

$$\overline{\boldsymbol{B}}_{K\times(2L+1)}=\begin{pmatrix} b_0^I[\omega_1] & b_1^M[\omega_1] & b_1^I[\omega_1] & b_2^M[\omega_1] & b_2^I[\omega_1] & \cdots & b_L^M[\omega_1] & b_L^I[\omega_1] \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ b_0^I[\omega_K] & b_1^M[\omega_K] & b_1^I[\omega_K] & b_2^M[\omega_K] & b_2^I[\omega_K] & \cdots & b_L^M[\omega_K] & b_L^I[\omega_K] \end{pmatrix}_{K\times(2L+1)} \tag{2-5}$$

Where $b_j^X[\omega_k]$ is the probability of output ω_k at the state X_j. Let $\overline{\boldsymbol{B}}_{K\times(2L+1)} \equiv (b_{ij})_{K\times(2L+1)}$, $b_{ij} = b_d^X(\omega_i)$, $1 \le i \le K$, $1 \le j \le 2L+1$, following relations can be obtained ,

$$j \bmod 2 = \begin{cases} 1, & (X,d) = (I, \dfrac{j-1}{2}) \\ 0, & (X,d) = (M, \dfrac{j}{2}) \end{cases} \tag{2-6}$$

Therefore a profile HMM can be simplified by the compression state transfer probability matrix with order of 9× (L+1) and the compression character output probability matrix with order of K× (2L+1).

3. FORWARD ALGORITHM

Let us consider the observation sequence $O = (o_1, o_2, \cdots, o_T)$. The matching range is L in Profile HMM (λ). Based on compress state transfer probability matrix $\overline{\boldsymbol{A}}_{9\times(L+1)}$ and character output probability matrix $\overline{\boldsymbol{B}}_{K\times(2L+1)}$, forward algorithm can be obtained.

Definition 1 Let $\alpha_l^X(t) = P(o_1, o_2, \cdots, o_t, \text{end of } X_l \mid \lambda), X = M, I, D$ be the probability when part sequence $O_t = (o_1, o_2, \cdots o_t)$ output in X_l state at l ($1 \le l \le L$) layer. Then $(\alpha_l^M(t), \alpha_l^I(t), \alpha_l^D(t))^T$ is the probability vector for l layer, denoted as $\boldsymbol{\alpha}_l(t)$.

Definition 2 Let $\varphi(X_r) = (a[M_{r-1}X_r], a[I_{r-1}X_r], a[D_{r-1}X_r])^{\mathrm{T}}$, $X = M, D$ be a column vector composed by one step transfer probability from state X_{r-1} to X_r , and $\varphi(I_q) = (a[M_qI_q], a[I_qI_q], a[D_qI_q])^{\mathrm{T}}$ also be a column vector composed by one step transfer probability from state I_{q-1} to I_q . Thus $\boldsymbol{\varphi}(M_p) = (\overline{a}_{1p}, \overline{a}_{2p}, \overline{a}_{3p})^{\mathrm{T}}$, $\boldsymbol{\varphi}(I_q) = (\overline{a}_{4q}, \overline{a}_{5q}, \overline{a}_{6q})^{\mathrm{T}}$ and $\boldsymbol{\varphi}(D_r) = (\overline{a}_{7r}, \overline{a}_{8r}, \overline{a}_{9r})^{\mathrm{T}}$, if compress state transfer probability matrix is $\overline{\boldsymbol{A}}_{9\times(L+1)}$ and character output probability matrix is $\overline{\boldsymbol{B}}_{K\times(2L+1)}$.

Now the forward algorithm for Profile HMM can be express as,

1) Initiation

$$\boldsymbol{\alpha}_0(0) = (1,0,0)^{\mathrm{T}}$$
$$\boldsymbol{\alpha}_0(t) = (0,0,0)^{\mathrm{T}}, t = 1,2,\cdots,T+1 \tag{3-1}$$

$\boldsymbol{\alpha}_l(0) = (0,0,0)^{\mathrm{T}}, l = 1,2,\cdots,L+1$

2) Recursion calculation

$$\boldsymbol{\alpha}_l(t) = \begin{pmatrix} \boldsymbol{\alpha}_{l-1}^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(M_l)b_l^M(o_t) \\ \boldsymbol{\alpha}_l^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(I_l)b_l^I(o_t) \\ \boldsymbol{\alpha}_{l-1}^{\mathrm{T}}(t)\boldsymbol{\varphi}(D_l) \end{pmatrix} \quad t = 1,2,\cdots,T \ , \ l = 1,2,\cdots,L \quad (3\text{-}2)$$

where $b_d^M(o_t) = b_{i(2d)}, b_l^I(o_t) = b_{i(2d+1)}$, when $o_t = w_i$.

3) Ending

Thus

$$\alpha_{L+1}^M(T+1) = \alpha_L^M(T)a_{1(L+1)} + \alpha_L^I(T)a_{2(L+1)} + \alpha_L^D(T)a_{3(L+1)} \quad (3\text{-}3)$$

and the probability is

$$P(O \mid \lambda) = \alpha_{L+1}^M(T+1) \quad (3\text{-}4)$$

4. APPLICATION OF PROFILE HMM

An example of compress state transfer probability matrix and character output probability matrix for known multiple sequence comparison is shown below. Suppose DAN sequences have be aliment as Table 1. Matching states locate in first, second, and sixth column. Thus m is 5, T is 6 and L is 3. The corresponding Profile HMM framework is depicted in Figure 2.

Table 7 Alignment of multiple DNA sequence

	1	2	3	4	5	6
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	A	A	-
gnat	-	-	A	A	A	C
goat	A	G	-	-	-	C

In order to avoid zero in probability calculation, pseudo counting is adapted. For example, character A appears 4 times in first column, the probability to output A for this state is $b_1^M(A) = \frac{4+1}{4+4} = 0.625$. Character G does not show up and the probability to output G is $b_1^M(G) = \frac{0+1}{4+4} = 0.125$.

Table 2 list character output probability at the state in which state transfer real occurs. As the same rule used, the compress state transfer probability matrix can be obtained as matrix (4-1).

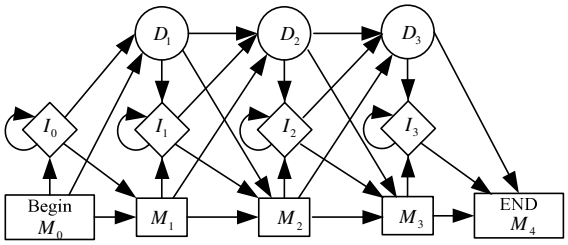

Figure 7 Profile HMM frameworks for the case of three states

Table 2. Frequency of character output

	$b_1^M(\omega_k)$	$b_2^M(\omega_k)$	$b_2^I(\omega_k)$	$b_3^M(\omega_k)$
A	0.625	0.143	0.333	0.125
G	0.125	0.571	0.333	0.125
C	0.125	0.143	0.111	0.625
T	0.125	0.143	0.222	0.125

$$\boldsymbol{A} = \begin{pmatrix} 0.625 & 0.571 & 0.500 & 0.833 \\ 0.333 & 0.333 & 0.500 & 0.500 \\ 0.000 & 0.250 & 0.200 & 0.667 \\ 0.143 & 0.333 & 0.167 & 0.000 \\ 0.333 & 0.167 & 0.500 & 0.000 \\ 0.250 & 0.600 & 0.333 & 0.000 \\ 0.250 & 0.286 & 0.167 & 0.000 \\ 0.333 & 0.333 & 0.333 & 0.000 \\ 0.000 & 0.500 & 0.200 & 0.000 \end{pmatrix} \quad (4\text{-}1)$$

Suppose $O = AGC$ is a new observation sequence. Thus the probability that the sequence can be observed by using the model as shown above is $P(O \mid \lambda) \approx 0.0373$

In practical application, the observation sequence can be classed according to its probability for different model λ_i that established by corresponding training data.

ACKNOWLEDGEMENTS

This work supported by the grant (10631070) from the National Natural Science Foundation of China.

REFERENCES

Gong Guang-lu, Qian Min-Ping，2004，Application stochastic process course and stochastic model in intelligent calculation, Qinghua Publisher, Beijing.

Katoh A，Mian I S，Haussler D.，1994，A hidden Markov model that finds genes in E.coli DNA. Nucleic Acids Research，22:4768-4778.

Krogh A，Brown M，Mian I S，Sjolander K，and Haussler D.，1994，Hidden Markov models in computational biology: applications to protein modeling，Journal of Molecular Biology，235:1501-1531.

Wang Yu-Fei, Shi Ding-Hua，2006，Bioinformatiic intelligent arithmetic and its application, Chemistry industry Publisher, Beijing.