High Order and PWM Based Hidden Markov Model (HOP-HMM)

August 2, 2017

Abstract

Introduction

[enhancer background]
[PWMs and motif to classify tissue specific enhancers]
[k-mer to classify tissue specific enhancers]
[HMM to classify tissue specific enhancers]
[other machine learning work to classify tissue specific enhancers]
[Why HOP-HMM might be better]

Methods

Setup

Let us consider a high order emission base-states and PWM emission sub-states HMM from a dataset of N observations sequences $X = (X_1, ..., X_N)$ where each observation sequence is L observations long $X_i = \left(x_1^i, ..., x_L^i\right)$. let the space of observation $\chi = \{1, 2, ..., n\}$. We assume an underlying hidden variable sequences $Y = (Y_1, ..., Y_N)$ where each underlying sequence is also L variables long $Y_i = y_1^i, ..., y_L^i$. Let the space of underlying states be $\Upsilon = \{1, 2, ..., m\} \times \{0, 1, ..., k\}$

Emission and Transition

Underlying states emit the observed sequence are of two types: base-states and their sub-states. We mark the j'th base-state as $(j, 0)$ for $j \in \{1, ..., m\}$ and its l'th sub-state as (j, l) for $l \in \{1, ..., k\}$. Denote the base-state emission order by o, meaning a base-state emits a letter sampled from an emission matrix E that depends on previous $o - 1$ letters.

Sub-state emits multiple letters sampled from a PWM that is fixed and isn't learned in the training. DenoteW_l the PWM of the l'th sub-states, which is shared between the l'th sub-states of all base-states.

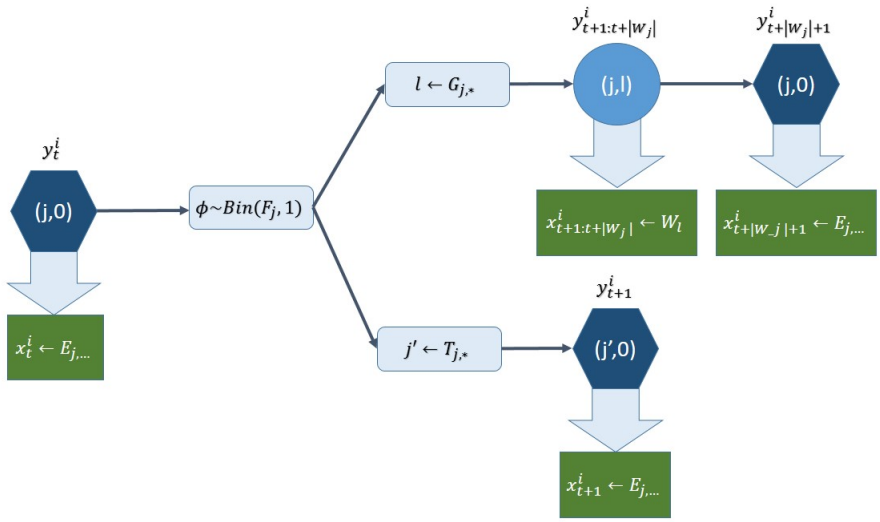

Figure 4: emission and transition process between base-state. The upper flow represent transition from base-state to the same base-state, through a sub-state that emits a motif. The lower flow represent transition between two base-state using the T transition matrix, similar to the conventional HMM.

After emitting a single letter, the j'th base-state has a probability F_j to make a transition into one of its sub-state and emit a motif and probability $1 - F_j$ to make a transition into one of the base-states and emit a single letter. The distribution of transitions between base-states is set by T matrix, and between base-state to its sub-states by G matrix. After emitting a motif in a sub-state, the next state will be the sub-state's base-state where it will emit a single letter.

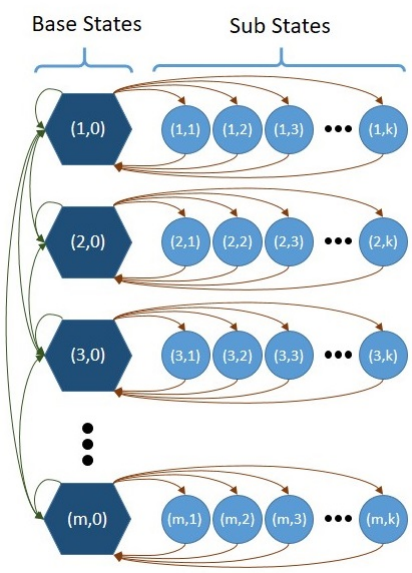

Figure 1: The hidden variable states graph of the HOP-HMM. The left hexagons represent base-states, and the circles in the right part of each row's represent its sub-states.

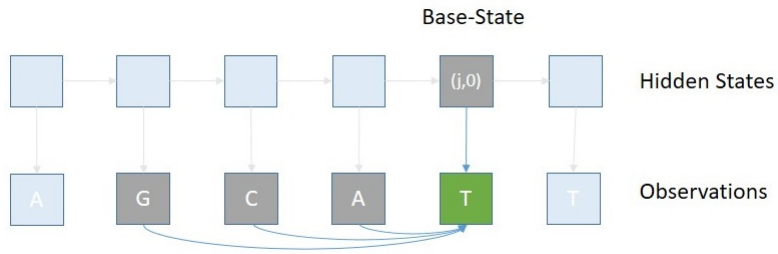

Figure 2: High order emission of base-states. Each emission is dependent on the hidden base-state and o-1 previous observations.

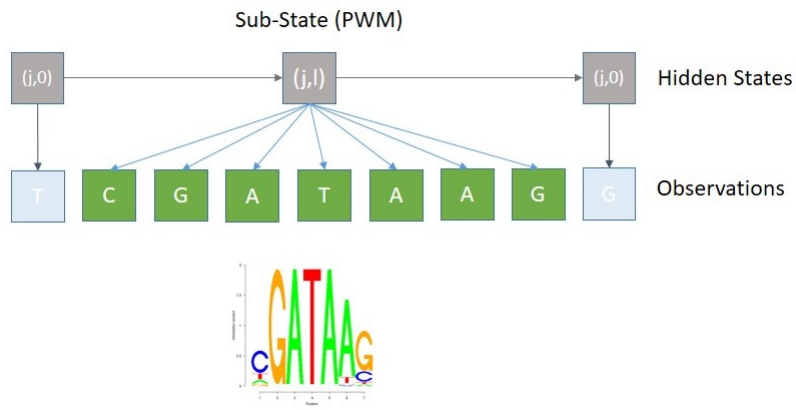

Figure 3: PWM emission of sub-states.

Parameters

An HOP-HMM $\theta = \{\pi, E, T, G, F\}$ is parameterized by:

- π : $m \times 1$ initial base-state distribution vector

$$\pi_j = P(y_1^i = j)$$

- E : $m \times \underbrace{4 \times 4 \times ... \times 4}_{o\ times}$ the base-state high order emission probability matrix

$$E_{j,b_1,b_2,...,b_o} = P\left(x_t^i = b_o | y_t^i = (j,0), x_{t-o+1}^i = b_1, ..., x_{t-1}^i = b_{o-1}\right)$$

- T : $m \times m$ the transition probability matrix

$$T_{j_1,j_2} = P\left(y_t^i = (j_2,0) | y_{t-1}^i = (j_1,0)\right)$$

- G : $m \times k$ the sub-state entry probability matrix

$$G_{j,l} = P\left(y_t^i = (j,l) | y_{t-1}^i = (j,0), l > 0\right)$$

- F : $m \times 1$ the probability to enter one of the sub-states from a base state

$$F_j = P\left(l > 0 | y_{t-1}^i = (j,0), y_t^i = (j,l)\right)$$

note: since these describe distributions: $\sum_{b \in \{1,...,n\}} E_{i,b_1,b_2,...,b_{o-1},b} = 1$, $\sum_{l \in \{1,...,k\}} G_{j,l} = 1$ and $\sum_{j_2 \in \{1,...,m\}} T_{j_1,j_2} = 1$

note 2: we marked here the index of the sequence by $i \in \{1, ..., N\}$

EM Algorithm

E-Step

Forward Algorithm

denote $L_M(\overline{x})$ as the likelihood of motif $\overline{x}$, i.e. the probability that $\overline{x}$ was generate by PWM M

$$L_M(\overline{x}) = P(\overline{x}|M) = \prod_{i\in\{1,\dots,|\overline{x}|\}} M_{\overline{x}_i,i,}$$

$\alpha - N \times m \times L$

we generate α by iterating over $t = 1, 2, ..., L$ for $t = 1$:

$\alpha_{i,j,1} = P(y_1^i = j, x_1^i) =$

$$\begin{aligned}
\alpha_{i,j,t} =& P(y_t^i = j, x_{1:t}^i) = \\
=& \underbrace{\sum_{j'\in\{1,\dots,m\}} \alpha_{i,j',t-1} \cdot (1 - F_{j'}) \cdot T_{j',j} \cdot E_{j,x^i_{t-o+1},\dots,x^i_t}}_{\text{base-state step}} \\
&+ \underbrace{\sum_{l\in\{1,\dots,k\}} \alpha_{i,j,t-|W_l|-1} \cdot F_j \cdot G_{j,l} \cdot L_{W_{l,}}\left(x^i_{t-|W_l|},\dots,x^i_{t-1}\right) \cdot E_{j,x^i_{t-o+1},\dots,x^i_t}}_{\text{sub-state step}}
\end{aligned}$$

Backward Algorithm

$\alpha - N \times m \times L$

$$\begin{aligned}
\beta_{i,j,t} =& P(y_t^i = j, x_{t+1:L}^i) = \\
=& \underbrace{\sum_{u\in\{1,\dots,m\}} (1 - F_u) \cdot T_{j,u} \cdot E_{u,x^i_{t-o+2},\dots,x^i_{t+1}} \cdot \beta_{i,u,t+1}}_{\text{base-state step}} \\
&+ \underbrace{\sum_{l\in\{1,\dots,k\}} F_j \cdot G_{j,l} \cdot L_{W_{l,}}\left(x^i_{t+1},\dots,x^i_{t+|W_l|}\right) \cdot E_{j,x^i_{t-o+|W_v|+2},\dots,x^i_{t+|W_v|+1}} \cdot \beta_{i,j,t+|W_l|+1}}_{\text{sub-state step}}
\end{aligned}$$

M-Step

First we calculate auxiliary variables:

$$\begin{aligned}
\psi_{i,j,l,t} =& P\left(y_t^i = (j,0), y^i_{t+1:t+|W_l|} = (j,l), X_i\right) \\
=& \alpha_{i,j,t} \cdot F_j \cdot G_{j,l} \cdot L_{W_{l,}}\left(x^i_{t+1},\dots,x^i_{t+|W_l|}\right) \cdot E_{j,x^i_{t+|W_l|-o+2},\dots,x^i_{t+|W_l|+1}} \cdot \beta_{i,j,t+|W_l|+1}
\end{aligned}$$

$$\begin{aligned}
\gamma_{i,j,t} =& P\left(y_t^i = (j,0)|X_i\right) \\
=& \frac{\sum_{i\in\{1,\dots,N\}} P\left(y_t^i = (j,0), X_i\right)}{P\left(X_i\right)} \\
=& \frac{\alpha_{i,j,t} \cdot \beta_{i,j,t}}{\sum_{j'\in\{1,\dots,m\}} \left(\alpha_{i,j',t} \cdot \beta_{i,j',t} + \sum_{l\in\{1,\dots,k\}} \sum_{s\in\{1,\dots,|W_l|\}} \psi_{i,j',l,t-s}\right)}
\end{aligned}$$

TODO: does different t gives different $P(X_i) = \sum_{j' \in \{1,\ldots,m\}} \alpha_{i,j',t} \cdot \beta_{i,j',t}$? Should it?

$$
\begin{aligned}
\xi_{i,j_1,j_2,t} =& P\left(y_t^i = (j_1,0), y_{t+1}^i = (j_2,0) | X_i\right) \\
=& \frac{P\left(y_t^i = (j_1,0), y_{t+1}^i = (j_2,0), X\right)}{P(X_i)} \\
=& \frac{\alpha_{i,j_1,t} \cdot (1-F_{j_1}) \cdot T_{j_1,j_2} \cdot E_{j_2,x^i_{t-o+2},\ldots,x^i_{t+1}} \cdot \beta_{i,j_2,t+1}}{\sum_{j_1',j_2' \in \{1,\ldots,N\}} \alpha_{i,j_1',t} \cdot \left(1-F_{j_1'}\right) \cdot T_{j_1',j_2'} \cdot E_{j_2',x^i_{t-o+2},\ldots,x^i_{t+1}} \cdot \beta_{i,j_2',t+1} + \sum_{j',\in\{1,\ldots,N\}\, l \in \{1,\ldots,k\}} \left(\sum_{s \in \{0,\ldots,|W_l|\}} \psi_{i,j',l,t-s} \right)}
\end{aligned}
$$

$$
\begin{aligned}
\eta_{i,j,l,t} =& P\left(y_t^i = (j,0), y_{t+1:t+|W_l|}^i = (j,l) | X_i\right) \\
=& \frac{P\left(y_t^i = (j,0), y_{t+1:t+|W_l|}^i = (j,l), X_i\right)}{P(X_i)} \\
=& \frac{\psi_{i,j,l,t}}{\sum_{j' \in \{1,\ldots,m\}} \left(\alpha_{i,j',t} \cdot \beta_{i,j',t} + \sum_{l \in \{1,\ldots,k\}} \sum_{s \in \{1,\ldots,|W_l|\}} \psi_{i,j',l,t-s} \right)}
\end{aligned}
$$

We use the temporary auxiliary variables to calculate the θ_{max} that maximizes likelihood of the observations.

$E_{j,b_1,b_2,\ldots,b_o} = \frac{\sum_{i \in [N]\, t \in [L]} \gamma_{i,j,t} \cdot \mathbf{1}_{b_1,\ldots,b_o}(x^i_{t-o+1},\ldots,x^i_t)}{\sum_{i \in [N]\, t \in [L]} \gamma_{i,j,t}}$

$T_{j_1,j_2} = \frac{\sum_{i \in [N]\, t \in [L]} \xi_{i,j_1,j_2,t}}{\sum_{i \in [N]\, t \in [L]} \gamma_{i,j_1,t}}$

$F_j = \frac{\sum_{i \in [N]\, t \in [L]\, l \in [k]} \eta_{i,j,l,t}}{\sum_{i \in [N]\, t \in [L]} \gamma_{i,j,t}}$

$G_{j,l} = \frac{\sum_{i \in [N]\, t \in [L]} \eta_{i,j,l,t}}{\sum_{i \in [N]\, t \in [L],\, l' \in [k]} \eta_{i,j,l',t}}$

$\pi_j = \frac{\gamma_{j,1}}{\sum_{j' \in [m]} \gamma_{j',1}}$

[Roadmap enhancers preprocessing]
[training on roadmap data]
[classification of regulation modules]

Results

[test accuracy on roadmap enhancers]
[prediction on roadmap regulation modules]
[Whole genome classification?]
[Was HOP-HMM better?]

Possible Applications

labeled enhancer seqs from multiple motifs-> EM to learn E M F per floor + setting $T = \mathbb{I}_{m \times m}$ -> posterior of whole genome with sliding window -> classify whole genome
learn E M F -> check correlation with TF expression
run EM on whole genome -> posterior of whole genome -> check correlation of posterior to ChIP-Seq of histone modifications
E M F T-> posterior of whole genome -> see if known critical SNPs are critical in classification

Discussion