Model Selection on Pairwise Hidden Markov Models using Factorized Information Criterion D1-32

Taikai Takeda[1], Michiaki Hamada[1,2]

[1]Waseda University, [2]AIST-Waseda CBBD OIL

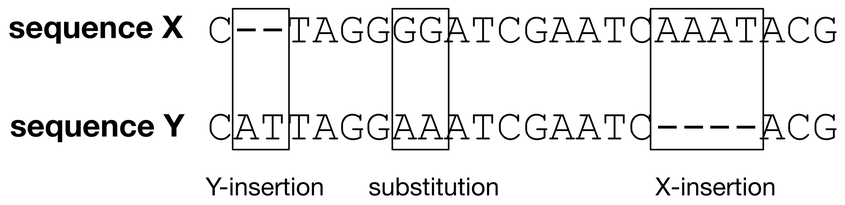
Abstract

Although extensive research has been done on biological sequence alignment problem, model selection problem remains untouched. We assume the model selection enables us to extract biological knowledge by interpreting resulting models, e.g. by comparing resulting models of pairwise DNA sequence alignment between some different pairs of species. As a model selection method, we introduce Factorized Asymptotic Bayesian Pairwise Hidden Markov Models (FAB-PHMM), based on asymptotic consistent information criterion with model evidence. We conducted an experiment on a synthetic dataset to illustrate model selection capability of proposed method. On a real DNA sequence data experiment, we observed that much more complex models are selected than previously utilized models, and the result is consistent with evolutionary distance.

Background

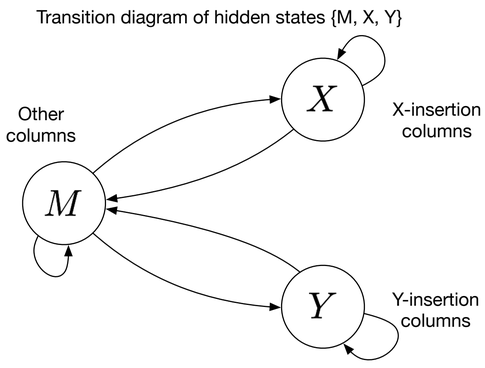
Pairwise sequence alignment

Pairwise sequence alignment aims to assess similarity of two sequences by introducing "gap". Gaps are represented as "-" and the gap columns have the nucleotides only in single side. In score-based alignment [Smith+1981], a similarity score is calculated as the sum of column scores; match columns make positive contribution and insertion/substitution columns affect negatively. The optimal alignment is given by Dynamic Programming, which seek the optimal gap insertions that maximize the similarity scores. Similarity of the sequence pair can be assessed by the score of optimal alignment.



Pairwise Hidden Markov Models (PHMM)

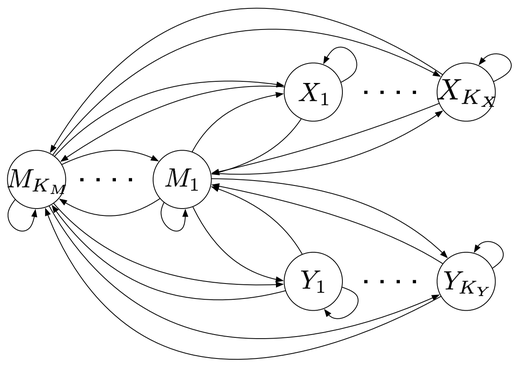
Pairwise Hidden Markov Models are probabilistic models [Durbin+, 1999] of pairwise sequence alignment. Using this model, alignment can be obtained via MAP decoding. In addition, parameters can be learned using EM algorithm.



Problem Setting

Question: Which is the "best" model of PHMM?

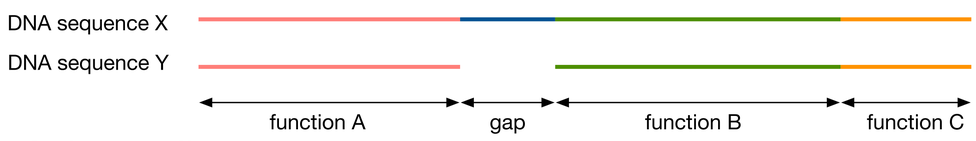
The model of PHMM is parameterized by the number of hidden states for each type $M = (K_M, K_X, K_Y)$. Our goal is to select the best model M that maximize a marginal likelihood given a observed data.



Motivations

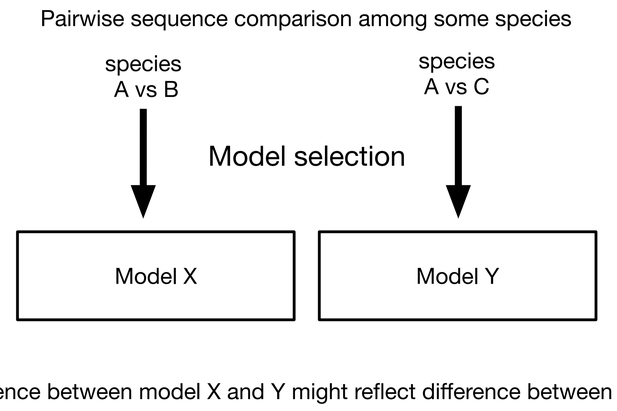
1) Biological function annotation

Each DNA region has different biological functions. Assuming that each of those regions are generated from different probabilistic distributions corresponding to a specific hidden states, we can annotate the regions with function type by inspecting posterior of hidden states.



2) Evolutionary distance

We assume resulting model reflect evolutionary distance; selected model of evolutionary further species should have complex model. The selected model might summarize evolutionary behavior of them.



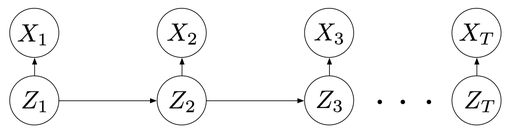
Difference between model X and Y might reflect difference between B and C

Methods

Formulation of HMM and PHMM

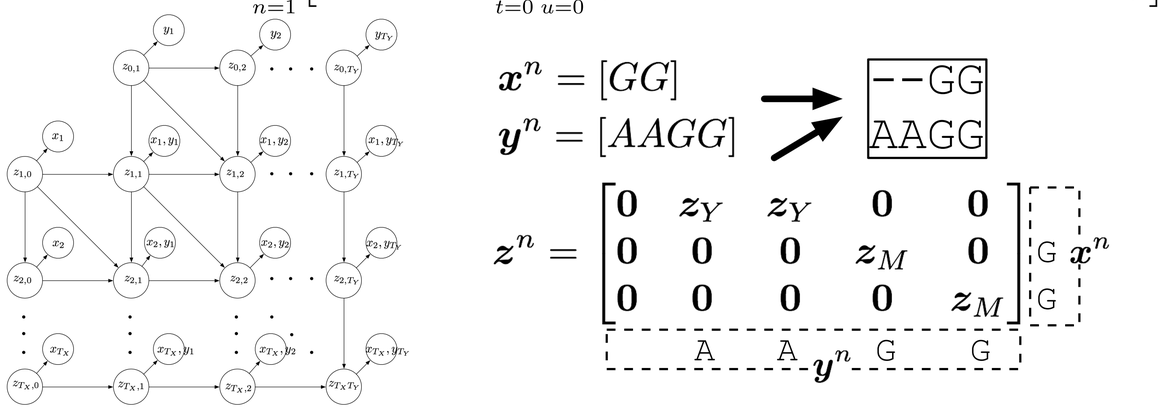
HMM

$$\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[\log p(\boldsymbol{z}_1^n|\boldsymbol{\alpha}) + \sum_{t=1}^{T} \Big(\log p(\boldsymbol{z}_t^n|\boldsymbol{z}_{t-1}^n, \boldsymbol{\beta}) + \log p(x_t^n|\boldsymbol{z}_t^n, \boldsymbol{\phi}) \Big) \right]$$



PHMM

$$\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[\log p(\boldsymbol{z}_{\text{in}}^n|\boldsymbol{\alpha}) + \sum_{t=0}^{T_X} \sum_{u=0}^{T_Y} \Big(\log p(\boldsymbol{z}_{tu}^n|pa(\boldsymbol{z}_{tu}^n), \boldsymbol{\beta}) + \log p(x_t^n, y_u^n|\boldsymbol{z}_{tu}^n, \boldsymbol{\phi}) \Big) \right]$$



Factorized Asymptotic Bayesian Pairwise Hidden Markov Models (FAB-PHMM)

Following the FAB-HMM [Fujimaki+2012], we apply Factorized Information Criterion (FIC) for PHMM. Our goal is to maximize FIC, which is an asymptotically consistent approximation of marginal likelihood.

Maximization of marginal likelihood

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} p(\boldsymbol{x}, \boldsymbol{y}|\mathcal{M}) \qquad p(\boldsymbol{x}, \boldsymbol{y}|\mathcal{M}) = \int \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$$

Variational Lower Bound

We introduce variational distribution $q(\boldsymbol{z})$to take lower-bound of the marginal likelihood.

$$p(\boldsymbol{x}, \boldsymbol{y}|\mathcal{M}) \geq \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \log \left(\frac{p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\mathcal{M})}{q(\boldsymbol{z})} \right).$$

Factorized Information Criterion (FIC)

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\mathcal{M}) = \int \prod_{n=1}^{N} p(\boldsymbol{z}_{\text{in}}^n|\boldsymbol{\alpha}) \prod_{t=0}^{T_X} \prod_{u=0}^{T_Y} \prod_{k=1}^{K} p_k(\boldsymbol{z}_{tu}^n|\boldsymbol{\beta}_k)^{pa(\boldsymbol{z}_{tu}^n)_k} p(x_t^n, y_u^n|\boldsymbol{z}_{tu}^n, \boldsymbol{\phi}_k)^{z_{tuk}^n} p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$$

↓ Laplace approximate on each terms and ignore asymptotically small terms w.r.t. N

$$FIC(\boldsymbol{x}, \boldsymbol{y}|\mathcal{M}) = \max_{q(\boldsymbol{z})} \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \Bigg(\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\bar{\boldsymbol{\theta}}) - \frac{D_\alpha}{2} \log N - \sum_{k=1}^{K} \frac{D_{\beta_k}}{2} \log(\sum_{n,t,u} z_{tuk}^n) - \sum_{k=1}^{K} \frac{D_{\phi_k}}{2} \log(\sum_{n,t,u} z_{tuk}^n) - \log q(\boldsymbol{z}) \Bigg)$$

FIC Lower Bound

We optimize FICLB through EM-like iteration. This lower-bound is obtained by using 1) linear approximation of logarithm function $\log a \leq \log b - (a - b)/b$ and 2) optimality of ML estimator $\log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\bar{\boldsymbol{\theta}}) \geq \log p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})$

$$FIC \geq FICLB(\boldsymbol{x}, \boldsymbol{y}, q, \tilde{q}, \boldsymbol{\theta}, \mathcal{M}) = \sum_{n} \sum_{\boldsymbol{z}^n} q(\boldsymbol{z}^n) \Bigg(\log p(\boldsymbol{x}^n, \boldsymbol{y}^n, \boldsymbol{z}^n|\boldsymbol{\theta}) + \sum_{tuk} z_{tuk}^n \log \delta_{tuk} - \log q(\boldsymbol{z}) \Bigg)$$

$$- \frac{D_\alpha}{2} \log N - \sum_{k=1}^{K} \frac{D_{\beta_k}}{2} \log \Big(\sum_{n,t,u} \tilde{q}(z_{tuk}^n) - 1 \Big) - \sum_{k=1}^{K} \frac{D_{\phi_k}}{2} \log \Big(\sum_{n,t,u \setminus (*, T_X^n, T_Y^n)} \tilde{q}(z_{tuk}^n) - 1 \Big)$$

$$\delta_{tuk} = \begin{cases} \exp\left(- \dfrac{D_{\phi_k}}{2 \sum_{ntu} z_{tuk}^n} \right) & \text{if } t = T_X \text{ and } u = T_Y \\ \exp\left(- \dfrac{D_{\phi_k}}{2 \sum_{ntu} \tilde{q}(z_{tuk}^n)} - \dfrac{D_{\beta_k}}{2 \sum_{ntu \setminus (*, T_X^n, T_Y^n)} \tilde{q} z_{tuk}^n} \right) & \text{otherwise} \end{cases}$$

Heuristic model Pruning

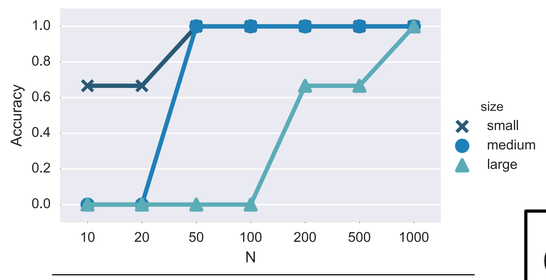
Algorithm 1 The model pruning algorithm

Input: data $(\boldsymbol{x}, \boldsymbol{y})$, initial model $\mathcal{M} = (K_M, K_X, K_Y)$, initial variational distribution q, initial parameter set $\boldsymbol{\theta}$ and pruning threshold ϵ

while not converged **do**
  $\tilde{q} \leftarrow q$
  $q \leftarrow \arg\max_q FICLB(\boldsymbol{x}, \boldsymbol{y}, q, \tilde{q}, \boldsymbol{\theta}, \mathcal{M})$
  for all k that suffice $\sum_{n,t,u} q(z_{tuk}^n) \leq \epsilon$ **do**
    delete the k-th hidden state of the model $\mathcal{M}$
  end for
  $\boldsymbol{\theta} \leftarrow \arg\max_{\boldsymbol{\theta}} FICLB(\boldsymbol{x}, \boldsymbol{y}, q, \tilde{q}, \boldsymbol{\theta}, \mathcal{M})$
end while

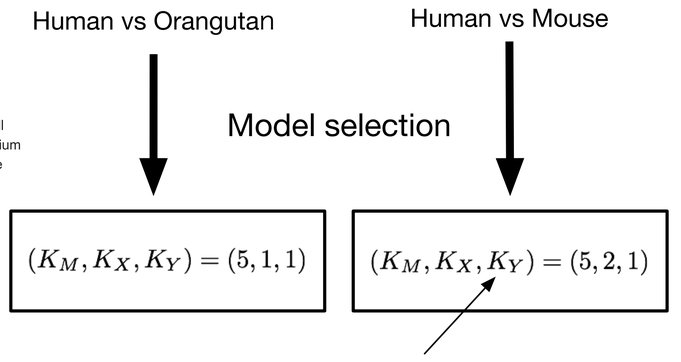
Experiments

Model selection on synthetic data



Size	$M^* = (K_M^*, K_X^*, K_Y^*)$
small	(1,1,1)
medium	(1,2,2)
large	(3,2,2)

Real DNA sequence



55N-06-10, 3-4-1, Okubo Shinjuku-ku, Tokyo 169-8555, Japan. E-mail: 297.1951@gmail.com