Model Selection for Pairwise Hidden Markov Models

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

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

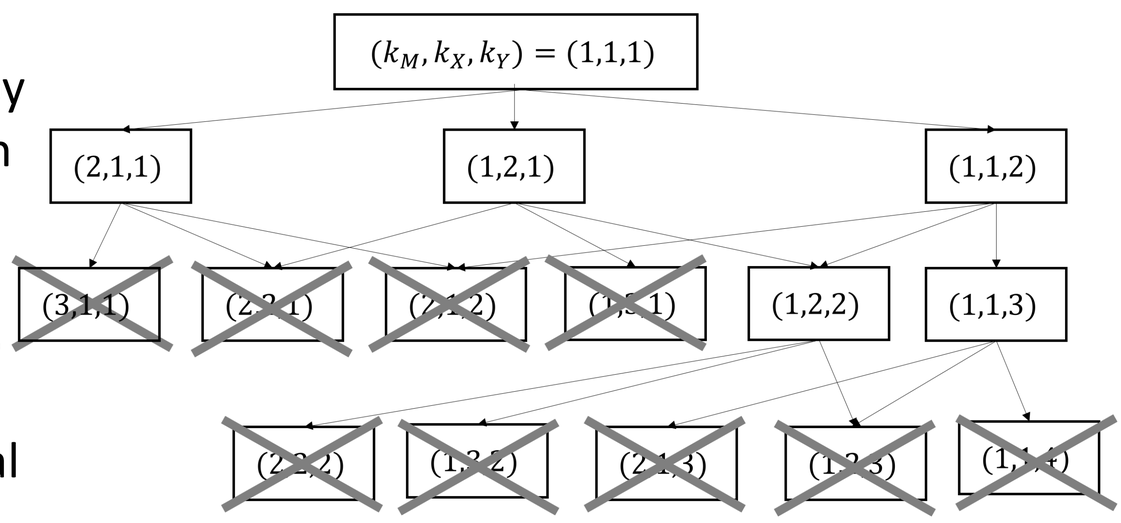
Background

Pairwise sequence alignment is a fundamental technique in comparative study of nucleotides and amino acids. Although extensive research has been done in this area, the model selection problem remains untouched. There are at least two distinct motivations for the model selection: model flexibility and model interpretation. Due to the fact that DNA regions, for example, have functional diversity, underlying model for the alignment problem should be more flexible. We believe that constructing appropriately complex model enables more accurate inference of alignment posterior probability. Additionally, we assume the model selection enable us to extract biological knowledge by interpreting the resulting model, e.g. by inspecting posterior distribution of hidden states in some specific regions.

Methods

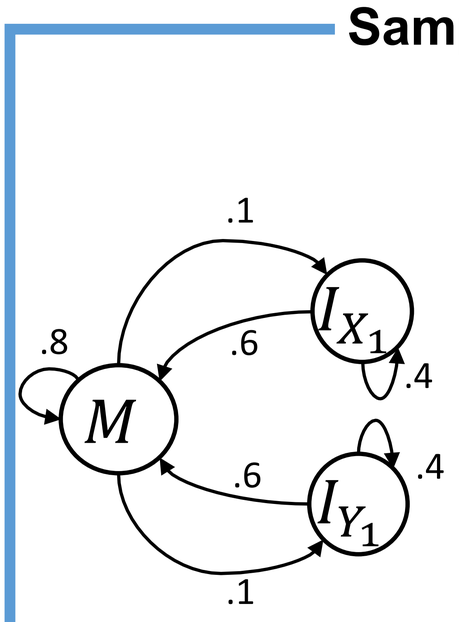
We introduce Factorized Asymptotic Bayesian Pairwise Hidden Markov Models (FAB-PHMM), inspired by FAB-HMM [Fujimaki+2012]. This model automatically chooses the optimal number of hidden states by maximizing lower-bound of Factorized Information Criterion (FIC), an asymptotically consistent approximation of marginal log-likelihood. This model does not have tunable hyper-parameters, thus resulting model is chosen highly subjectively. Our main contributions in terms of methods are 1) extending FAB optimization to PHMM, and 2) constructing a novel theoretically appropriate algorithm, incremental model search, based on shrinkage effect.

incremental model search

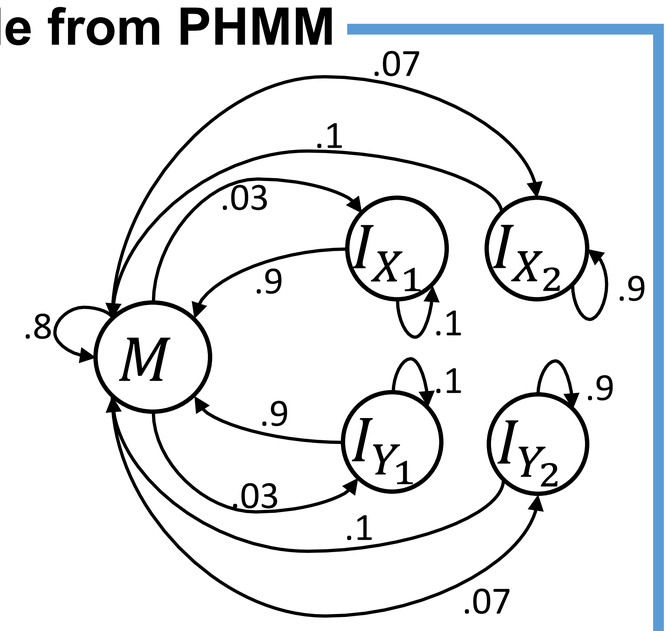


Results

Sample from PHMM



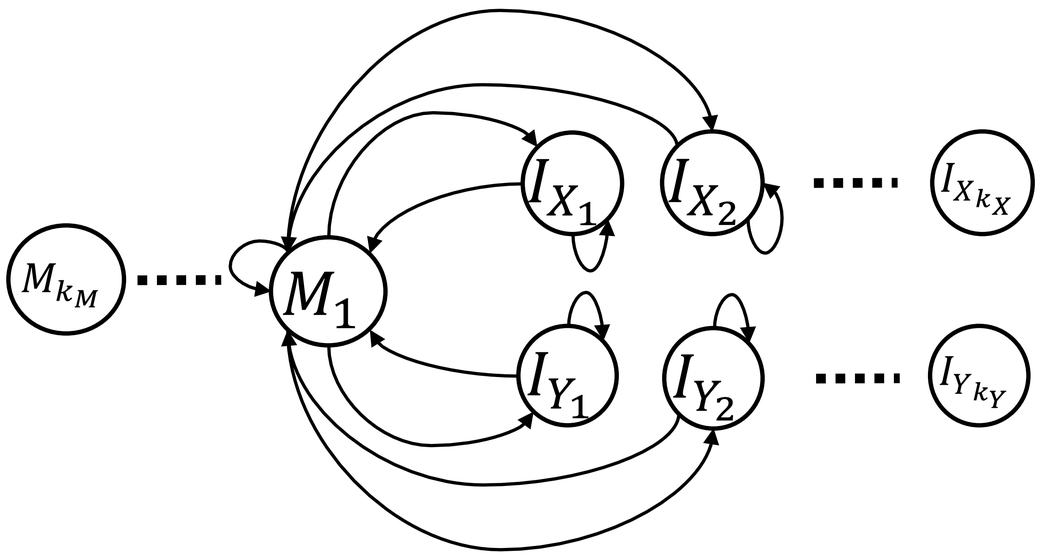
small-size model



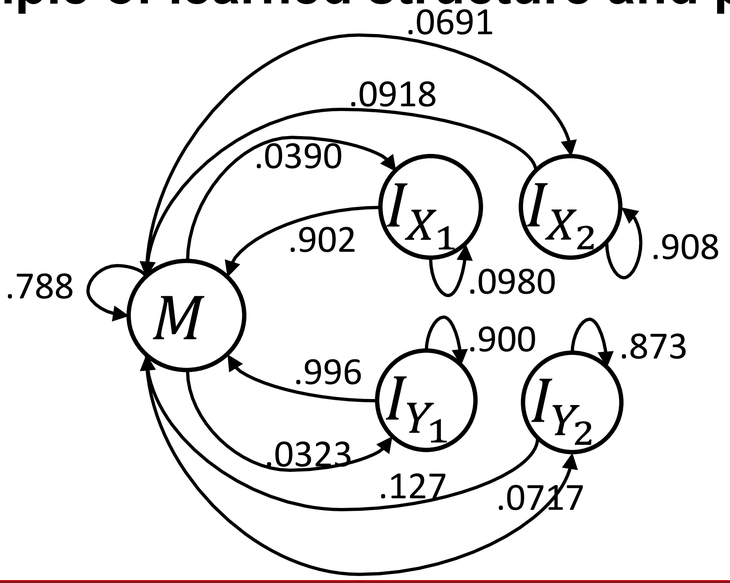
medium-size model

generate sample

Train on FAB-PHMM



Example of learned structure and parameters



	small model	medium model
incremental algorithm	0.96	**0.90**
decremental algorithm [Fujimaki+2012]	0.96	0.05

Model selection accuracy of two algorithms on two models:
Accuracy is measured by the proportion of accurately predicted number of hidden states ($\mathrm{k_M}, k_X, k_Y$). Incremental algorithm accurately predict original structure while decremental algorithm failed when original model size is larger.

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