

# Protein Structure Analysis Using Continuous Density Hidden Markov Models

Morihiro Hayashida<sup>1</sup>

morihiro@kuicr.kyoto-u.ac.jp

Nobuhisa Ueda<sup>1</sup>

ueda@kuicr.kyoto-u.ac.jp

Katsuhisa Horimoto<sup>2</sup>

horimoto@post.saga-med.ac.jp

Tatsuya Akutsu<sup>1</sup>

takutsu@kuicr.kyoto-u.ac.jp

<sup>1</sup> Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

<sup>2</sup> Laboratory of Mathematics, Saga Medical School, 5-1-1 Nabeshima, Saga, Saga 849-8501, Japan

**Keywords:** continuous density HMMs, Gaussian model, alpha helix

## 1 Introduction

Hidden Markov models [2] (HMMs) have been successfully applied to Bioinformatics such as gene finding, remote homology detection and secondary structure prediction. On the other hand, continuous density HMMs have been widely used in the field of speech recognition. Though continuous density HMMs were not applied to Bioinformatics so far, they may also be useful in Bioinformatics. Currently, we are developing methods for applying continuous density HMMs to analysis of three-dimensional structures of proteins. In this poster abstract, we report a preliminary result on application of continuous density HMMs to analysis of protein structures.

## 2 Method

Hidden Markov models with discrete observations have probabilities for output characters. On the other hand, output probabilities of continuous density HMMs are defined by Gaussian distributions. The probability to output a vector  $\mathbf{O}_t$  at time  $t$  in state  $i$  is

$$b_i(\mathbf{O}_t) = (2\pi)^{-\frac{d}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{O}_t - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{O}_t - \boldsymbol{\mu}_i)\right),$$

where  $\boldsymbol{\mu}_i$  is a mean vector of state  $i$ ,  $\Sigma_i$  is a covariance matrix, and  $d$  is the dimension of vectors.

### 2.1 Parameter estimation

We can estimate continuous density HMM parameters using the EM(Expectation-Maximization) algorithm [1], such as discrete HMMs. Among the parameters, initial probabilities and transition probabilities are updated by the same formulae. On the other hand, the parameters of the output probabilities are updated as follows

$$\tilde{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^T \gamma_t(i) \mathbf{O}_t}{\sum_{t=1}^T \gamma_t(i)}, \quad \tilde{\Sigma}_i = \frac{\sum_{t=1}^T \gamma_t(i) (\mathbf{O}_t - \boldsymbol{\mu}_i) (\mathbf{O}_t - \boldsymbol{\mu}_i)'}{\sum_{t=1}^T \gamma_t(i)},$$

where  $\gamma_t(i)$  is the probability of being state  $i$  at time  $t$ , given an observation sequence  $\mathbf{O}_1, \dots, \mathbf{O}_T$ . In addition, we used updating formulae for many observation sequences, and in order to avoid underflows we applied a scaling method and used logarithm of probabilities.

