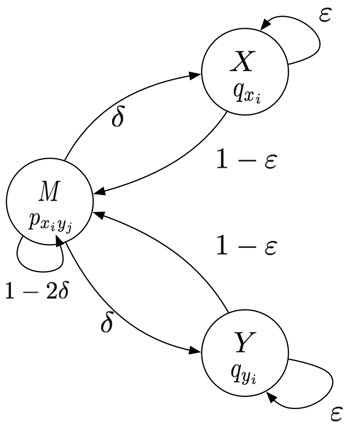

図 1　phmm の状態遷移図

1　Pairwise HMM

Pairwise HMM とは 2 つの sequence data（DNA の塩基配列やタンパク質の residue 配列）の類似度をモデル化した HMM である．これは discrete HMM の拡張であると考えることができるが，一般的な discrite HMM とは出力が 2 つのシンボルの組になるという点で異なる．隠れ状態の状態遷移図は以下のようになる．M は Match 状態，X, Y はそれぞれの配列の挿入状態を表す．

1.1　Notations

定式化する上で Notation を整理する．観測変数はシンボル列 $\boldsymbol{x}, \boldsymbol{y}$ である．

$$\boldsymbol{x} = (x^1, ..., x^{T_x}) \tag{1.1}$$
$$\boldsymbol{y} = (y^1, ..., y^{T_y}) \tag{1.2}$$
$$\tag{1.3}$$

観測変数 x^i, y^j に対応する隠れ変数 $\boldsymbol{z}^{ij}$ は以下のように定義する．ここで，x^i もしくは y^j の代わりに gap を許す（観測されたシンボルが対応しない場合がある）．

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{z}^{11} & \cdots & \boldsymbol{z}^{1T_y} \\ \vdots & \ddots & \vdots \\ \boldsymbol{z}^{T_x 1} & \cdots & \boldsymbol{z}^{T_x T_y} \end{pmatrix} \tag{1.4}$$

隠れ変数は $\boldsymbol{z}^{ij}$ は状態数 K に対して 1-of-K 符号化方式で表される．

$$\boldsymbol{z}^{ij} = (z_1^{ij}, \ldots, z_K^{ij}) \tag{1.5}$$
$$z_k^{ij} = \begin{cases} 1 & x^i, y^j \text{ は } k \text{ 番目の状態から出力される} \\ 0 & \text{otherwise} \end{cases} \tag{1.6}$$

また，表記の簡単のため，$\boldsymbol{z}^{i0} = \boldsymbol{z}^{0j} = 0$, 遷移の方向の offset として $\boldsymbol{\delta} = \{(\delta_x, \delta_y)\}$ と定義する．ここでは，$\boldsymbol{\delta} = \{(1,1),(0,1),(1,0)\}$ となる．これを用いて，隠れマルコフモデルの同時分布は以下ように表される．

$$p(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta}) = p(\boldsymbol{z}^{11}|\boldsymbol{\alpha}) \prod_{i,j=1}^{T_x,T_y} \prod_{(\delta_x,\delta_y)\in\boldsymbol{\delta}} p(\boldsymbol{z}^{ij}|\boldsymbol{z}^{i-\delta_x,j-\delta_y},\boldsymbol{\beta}) \prod_{i,j=1}^{T_x,T_y} p(x^i y^j|\boldsymbol{z}^{ij},\boldsymbol{\phi}) \tag{1.7}$$

上記で，$\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\phi})$ は HMM のパラメタである．

$$\begin{array}{ll} \text{初期確率:} & p(\boldsymbol{z}^{11}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \alpha_k^{z_k^{11}} \\ \text{出力確率:} & p(x^i,y^j|\boldsymbol{z}^{ij},\boldsymbol{\phi}) = \prod_{k=1}^{K} p(x^i,y^j|\boldsymbol{\phi}_k)^{z_k^{ij}} \\ \text{遷移確率:} & p(\boldsymbol{z}^{ij}|\boldsymbol{z}^{i-\delta_x,j-\delta_y},\boldsymbol{\beta}) = \prod_{l,k=1}^{L,K} \beta_{lk}^{z_l^{i-\delta_x,j-\delta_y} z_k^{ij}} \end{array} \tag{1.8}$$

ここで，

$$\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K) \in \mathbb{R}^K \tag{1.9}$$

$$\boldsymbol{\beta} = (\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta}_K) \in \mathbb{R}^K \times \mathbb{R}^K \text{ where } \ \boldsymbol{\beta}_k = (\beta_{k1},\ldots,\beta_{kK}) \in \mathbb{R}^K \tag{1.10}$$

$$\boldsymbol{\phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_K) \tag{1.11}$$

である．

1.2 Biterbi Algorithm

1.3 EM algorithm

EM algorithm によりパラメタの最適化を行う．EM アルゴリズムは以下のようなくり返しのアルゴリズムである．

$$\textbf{E-step} \tag{1.12}$$

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) = \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{old})}[\ln p(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})] \tag{1.13}$$

$$\tag{1.14}$$

$$\textbf{M-step} \tag{1.15}$$

$$\boldsymbol{\theta}^{new} \quad = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta},\boldsymbol{\theta}^{old}) \tag{1.16}$$

Estep では，事後分布 $p(\boldsymbol{z}^{ij}|\boldsymbol{x})$ と $p(\boldsymbol{z}^{ij},\boldsymbol{z}^{i-\delta_x,j-\delta_y}|\boldsymbol{x})$ を求める必要がある．この事後分布は forward-backward アルゴリズムで効率的に求めることが出来る．forward 変数 f と backward 変数 b を以下のように定義する．この 2 つの変数から，事後分布を求めることが出来る．ここでは，表記の簡単のためパラメタは明記しない．

$$f^{ij} = p(x^1,\cdots,x^i,y^1,\cdots,y^j,\boldsymbol{z}^{ij}) \tag{1.17}$$

$$b^{ij} = p(x^{i+1},\cdots,x^{T_x},y^{j+1},\cdots,y^{T_y}|\boldsymbol{z}^{ij}) \tag{1.18}$$

$$\boldsymbol{\gamma}^{\boldsymbol{ij}} = p(\boldsymbol{z}^{ij}|\boldsymbol{x},\boldsymbol{y}) = f^{ij}b^{ij}/p(\boldsymbol{x},\boldsymbol{y}) \tag{1.19}$$

$$\gamma_k^{ij} = p(z_k^{ij} = 1|\boldsymbol{x},\boldsymbol{y}) \tag{1.20}$$

$$\boldsymbol{\xi}^{\boldsymbol{ij\delta}} = p(\boldsymbol{z}^{ij},\boldsymbol{z}^{i-\delta_x,j-\delta_y}|\boldsymbol{x},\boldsymbol{y}) = f^{i-\delta_x,j-\delta_y} p(\boldsymbol{z}^{ij}|\boldsymbol{z}^{i-\delta_x,j-\delta_y}) p(x^{ij}|\boldsymbol{z}^{ij}) b^{ij}/p(\boldsymbol{x},\boldsymbol{y}) \tag{1.21}$$

$$\xi_{lk}^{ij\delta} = p(z_k^{ij} = 1, z_l^{i-\delta_x,j-\delta_y} = 1|\boldsymbol{x},\boldsymbol{y}) \tag{1.22}$$

ここで，表記の簡単のため $\boldsymbol{\gamma}, \boldsymbol{\xi}$ を定義した．すると，これらは DP(Dynamic Programming) で計算できる．

$$f^{11} = p(\boldsymbol{z}^{11})p(x^1, y^1|\boldsymbol{z}^{11}) \tag{1.23}$$

$$f^{ij} = p(x^i, x^j|\boldsymbol{z}^{ij}) \sum_{(\delta_x,\delta_y)\in\boldsymbol{\delta}} \sum_{\boldsymbol{z}^{i-\delta_x,j-\delta_y}} p(\boldsymbol{z}^{ij}|\boldsymbol{z}^{i-\delta_x,j-\delta_y}) f^{i-\delta_x,j-\delta_y} \tag{1.24}$$

$$b^{T_xT_y} = \mathbf{1} \tag{1.25}$$

$$b^{ij} = \sum_{(\delta_x,\delta_y)\in\boldsymbol{\delta}} \sum_{\boldsymbol{z}^{i+\delta_x,j+\delta_y}} p(x^{i+\delta_x}, x^{j+\delta_y}|\boldsymbol{z}^{i+\delta_x,j+\delta_y}) p(\boldsymbol{z}^{i+\delta_x,j+\delta_y}|\boldsymbol{z}^{ij}) b^{i+\delta_x,j+\delta_y} \tag{1.26}$$

Mstep では，この事後分布 $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^{old})$ を用いて Q 関数を最大化するパラメタ $\boldsymbol{\theta}$ を求める．

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}^{old})}[\ln p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})] \tag{1.27}$$

$$\boldsymbol{\theta}^{new} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) \tag{1.28}$$

それぞれのパラメタは以下の通り計算できる．

$$\alpha_k = \gamma_k^{11} / \sum_k \gamma_k^{11} \tag{1.29}$$

$$\beta_{lk} = \sum_{ij\boldsymbol{\delta}} \xi_{lk}^{ij\delta} / \sum_{ij\boldsymbol{\delta}k} b\xi_{lk}^{ij\delta} \tag{1.30}$$

$$\boldsymbol{\theta}_k^* = \operatorname*{argmax}_{\boldsymbol{\theta}_k} \sum_{ijk} \gamma_k^{ij} p(x_i, y_j|\boldsymbol{\theta}_k, z_k^{ij} = 1) \tag{1.31}$$

$\boldsymbol{\theta}$ は，emission の分布の形 $p(x, y|\boldsymbol{\theta}, \boldsymbol{z})$ を仮定することで得ることが出来る．ここでは，シンボルを $s^1 ... s^D$ の D 種類として，s^m, s^n の組み合わせを出力する確率を μ^{mn} に持つ多項分布を仮定する．

$$p(x, y|\boldsymbol{\phi}) = \prod_{mn} \mu_{mn}^{I(x,s^m)I(y,s^n)} \tag{1.32}$$

このとき，パラメータ $\boldsymbol{\phi} = \boldsymbol{\mu}$ はラグランジュの未定乗数法を用いて以下のように求められる．

$$\mu^{mn} = \sum_{ij} \gamma^{ij} I(x^i, s^m) I(y^j, s^n) / \sum_{ijmn} \gamma^{ij} I(x^i, s^m) I(y^j, s^n) \tag{1.33}$$

以上を用いて，パラメタの最適化を行うことができる．