社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (13)
2006／6／15

# EM法を用いた太古のDNA修復タンパク質RecAの推定

永井 友理† 胡桃坂 仁志‡ 村田 昇‡

†‡早稲田大学 大学院 理工学研究科 電気・情報生命専攻
‡早稲田大学

〒169-8555 東京都新宿区大久保3-4-1

E-mail: †ynagai@akane.waseda.jp, ‡hitoshi.kurumizaka@eb.waseda.ac.jp ‡noboru.murata@eb.waseda.ac.jp

**あらまし** DNAが紫外線等で損傷した場合，それを修復するのに中心的役割を担うRecAタンパク質が存在する．太古の地球は現在よりも強い電磁波を浴びていたので，過去のRecAタンパク質の方が現在よりも修復活性が強かったのではないかと考えられる．そこで，現在の微生物のRecAタンパク質のアミノ酸配列を比較することで過去のタンパク質の推定を行なわれている．MCMC法を用いて，子配列となる複数のアミノ酸配列のグループ化を行い，各グループの多数決により，親配列を推定する手法が提案されている．シミュレーションを行なった結果，十分な精度が得られなかった．これは子配列のグループ化が不十分なためと考えられる．そのため本問題に適した新たなモデルを構築し，EM法を用いて解くことを提案する．

**キーワード** MCMC，Gibbs sampling，EM法，RecAタンパク質

# Inference of ancestral RecA protein by EM algorithm

Yuri NAGAI† Hitoshi KURUMIZAKA‡ and Noboru MURATA‡

†‡Waseda University 1-2-3 Ohkubo, Shinjuku-ku, Tokyo, 169-8555 Japan
‡Waseda University

E-mail: †ynagai@akane.waseda.jp, ‡hitoshi.kurumizaka@eb.waseda.ac.jp ‡noboru.murata@eb.waseda.ac.jp

**Abstract** Protein RecA plays a center role to restore DNA when it is damaged because of ultraviolet rays .

It is thought that the repair activity of the ancient RecA protein was stronger than that of present one, because ancient earth was exposed by a strong electromagnetic radiation . A past protein was estimated by comparing the amino acid sequences of the RecA protein exists in a present microorganism.

Currently, parents arrays are estimated from the majority decision of those groups, classified by some known amino acid sequences, using MCMC method. However, the result of the simulation shows the poor accuracy.

Therefore, instead of using MCMC, we use the EM method to estimate how to classify proteins.

**Keyword** MCMC, Gibbs sampling, EM algorithm, RecA protein

## 1. 研究目的

ある生物をその生物足らしめるのに必須な遺伝情報をゲノムといい，その実体はDNAの塩基配列である．もしDNAが紫外線などで損傷してしまった場合，癌などの病気の原因となることが知られている．しかし生物にはDNAの損傷を修復する機構が存在し，なかでも原核生物の二重鎖切断損傷の修復にはRecAタンパク質が中心的役割を果たしている．

太古の地球は現在よりも強い電磁波（放射線，紫外線）にさらされ，これによって二重鎖切断が誘導されるので，過去のRecAタンパク質のほうが現在のRecAタンパク質よりも二重鎖切断修復活性が強かったのではないかと考えられる．もし，太古の活性の強いRecAタンパク質を推定することができれば，癌の遺伝子治療や遺伝子組み換え食品を安全に作るためなどに活用できるだろうと考えられる．そこで，現在のバクテリアの持つRecAタンパク質を比較して過去のRecAタンパク質を推定するというのがこの研究の目的である．

タンパク質は 20 種類のアミノ酸の並び方によってその形や働きが決まる．そこで，現在存在する異なるバクテリアが持つRecAタンパク質のアミノ酸配列（子配列）を比較することで過去のタンパク質（親配

列）の推定を行った．本報告では，EM 法を使用したモデルを提案し，Markov chain Monte Carlo(MCMC)法を使用した従来のモデルとの比較を行った．2 章で手法の説明を行い，3 章に実験結果，4 章にまとめを記した．

## 2. 推定モデルの説明

### 2.1 MCMC 法による手法

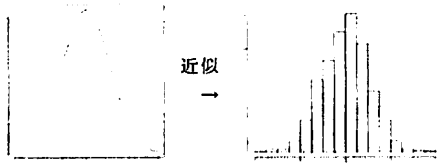
#### 2.1.1 MCMC 法[3]

ある推定量 $\theta$ が事後分布 $\pi(\theta)$をもつとき，$\theta$ の期待値は，

$$E[\theta]=\int\theta\cdot\pi(\theta)d\theta \qquad (1)$$

である．しかし，積分計算を行うのは困難である[3]．もし，確率分布 $\pi(\theta)$をもとに推定量 $\theta$ の独立な標本を得ることができれば，式(2)のように期待値計算を近似できる(図 1)．

$$E[\theta]=\int\theta\cdot\pi(\theta)d\theta\approx\frac{1}{N}\sum_{t=1}^{N}\theta^{(t)} \qquad (2)$$




真の値が推定量 $\theta$ である確率　推定量 $\theta$ が選ばれた回数

図 1　期待値計算の近似

そこで，以下に述べるように，$\pi(\theta)$からの確率標本 $\theta^{(t)}$をサンプリングする手法としてよく用いられる手法が MCMC 法である．

Markov 連鎖には，適当な初期値からはじめて十分な回数を繰り返していくと，確率標本の分布が正則条件の下で不変分布に収束していくという性質がある．この不変分布が $\pi(\theta)$になるように Markov 連鎖を構成することにより，Markov 連鎖の確率標本 $\theta^{(t)}$を $\pi(\theta)$からの確率標本とすることができる．

本報告では MCMC 法の特別な場合である Gibbs sampling という手法を用いた．構成したい分布は $\pi(\theta)$であり，データ $x$ がある時，$\theta$ の事後分布の確率密度関数は $\pi(\theta|x)$である．$\theta$ は，$\theta=(\theta_1,...,\theta_p)$ といくつかのブロックに分割できるとする．また，$\theta$ の $i$ 番目のみを除いた要素を $\theta_{-i}$ と定義する．

$$\theta_{-i}=(\theta_1,..,\theta_{i-1},\theta_{i+1},..,\theta_p) \qquad (3)$$

$\theta_{-i}$ と $x$ が与えられたときの条件付き確率分布の確率密度関数を $\pi(\theta_i|\theta_{-i},x)$とし，この条件付き分布からのサンプリングが容易であると仮定する．このとき Gibbs sampler とは以下のようなアルゴリズムである．

1) 初期値 $\theta^{(0)}=(\theta_1^{(0)},\theta_2^{(0)},..,\theta_p^{(0)})$を決め，$t$=1 とおく．
2) $t$=1,..,$p$ について

$$\theta_i^{(t)}\sim\pi(\theta_i|\theta_{-i}^{(t)},x) \qquad (4)$$

$$\theta_{-i}^{(t)}=(\theta_1^{(t)},..,\theta_{i-1}^{(t)},\theta_{i+1}^{(t-1)},..,\theta_p^{(t-1)}) \qquad (5)$$

を発生させる．
3) $t$ を $t$+1 として 2）にもどる．

2，3 を繰り返し，十分大きな数 $T$ について $t\geqq T$ のとき，$\theta^{(t)}=(\theta_1^{(t)},\theta_2^{(t)},..,\theta_p^{(t)})$を $\pi(\theta)$の確率標本とする．

#### 2.1.2 モデルの構築

Gibbs sampling を用い Pritchard が提案したモデル[1]に基づいて推定を行った(図 2)．これは，子配列のグループ分けを推定し，各グループの子配列の多数決により親配列を推定するというものである．

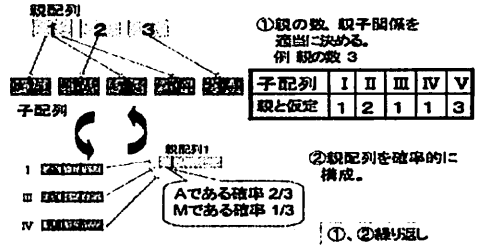



図 2　Pritchard のモデル

データ配列，パラメータを以下のように定義する．まず，既知の子配列を $\mathcal{X}$ とする．

$$\mathcal{X}=\begin{bmatrix}\mathcal{X}_1\\ \vdots\\ \mathcal{X}_h\\ \vdots\\ \mathcal{X}_H\end{bmatrix}=\begin{bmatrix}x_1^1 & \cdots & x_n^1 & \cdots & x_N^1\\ \vdots & \ddots & & & \vdots\\ x_1^h & & x_n^h & & x_N^h\\ \vdots & & & \ddots & \vdots\\ x_1^H & \cdots & x_n^H & \cdots & x_N^H\end{bmatrix} \qquad (6)$$

$\mathcal{X}_h$:子配列 $h$

$x_n^h\in A$　$A\in$｛アミノ酸 20 種，“X”｝

$H$：子配列の数　$1 \leqq h \leqq H$

$N$：配列長　　$1 \leqq n \leqq N$

データの長さは異なる場合があるので，アラインメント[4]をとるなどして事前に揃える必要がある(詳細は3.1)．また，"X"はアミノ酸が欠失した状態である．

次に，$H$個の子配列がどの親配列から発生したかを表す変数$\theta_1$を考える．

$$\theta_1 = \{a_1, a_2, .., a_h, ..., a_H\}, \quad a_h \in \{1,2,...,K\} \tag{7}$$

$K$:親配列の数　$1 \leqq k \leqq K$

最後に，親配列がどのような配列かを表す変数$\theta_2$を考える(図3)．

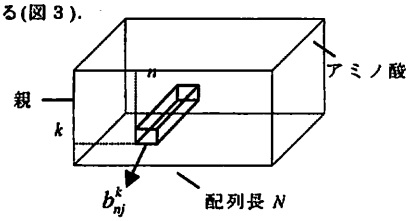



図3　パラメータ$\theta_2$

$$\theta^k_{2n} = b^k_n = (b^k_{nA}, .., b^k_{nj}, ..., b^k_{nY}), \quad j \in \mathrm{A} \tag{8}$$

$b^k_{nj}$は親配列$k$上の$n$番目の位置に，アミノ酸$j$が存在する確率であり，式(9)の条件を満たす．

$$\sum_{j \in \mathrm{A}} b^k_{nj} = 1 \tag{9}$$

以下に，Gibbs Samplingのアルゴリズムを記述する．

1) $\theta_1$の初期値$\theta_1^{(0)}$を適当に決める．

2) step1　全ての$k$，$n$に対して

$$b_n^{k(t)} sample \ D(\lambda_A + \beta^k_{nA}, ..., \lambda_j + \beta^k_{nj}, ..., \lambda_Y + \beta^k_{nY} \mid \theta_1^{(t-1)}, X_h) \tag{10}$$

$D$はディリクレ分布を表す．

$\lambda_j$：事前に知識として分かっているアミノ酸$j$の出現頻度．本報告では全て1にしている．

$\beta^k_{nj}$：親が$k$とラベルされた子配列のn番目の位置のアミノ酸が$j$である回数．

3) step2　全ての$h$に対して

$$a_h^{(m)} \sim \Pr(a_h = k \mid X_h, b_n^{k(m)}) = \frac{\Pr(X_h \mid b_n^{k(m)}, a_h = k)}{\sum_{k'=1}^{K} \Pr(X_h \mid b_n^{(m)}, a_h = k')} \tag{11}$$

$$\Pr(X_h \mid b_n^{k(m)}, a_h = k) = \prod_{n=1}^{N} b^k_{nx_n^h} \tag{12}$$

2，3を十分な回数繰り返した後の標本の期待値を計算する．

### 2.1　EM法による手法

#### 2.1.2　EM法

MCMC法使用モデルでは，親が複数の場合うまく推定できなかったので，EM法[2]によって親配列を推定することを提案した．

EM法とはデータに未観測データがある場合に，モデルのパラメータを尤度が大きくなるように逐次更新していくことで，最尤推定値を求める方法である．そこで，観測できたデータを子配列，観測できなかったデータを親子関係・親配列と考える．そして，アミノ酸の突然変異確率$P$を変化させていくことによりデータに最もよく「あてはまる」$P$を推測していく．$P$は親配列のアミノ酸$j$が子のアミノ酸$l$に変異する確率である．式(13)のように定義する．

$$P_{jl} = \begin{bmatrix} P_{AA} & \cdots & P_{jA} & \cdots & P_{XA} \\ \vdots & \ddots & & & \vdots \\ P_{Al} & & P_{jl} & & P_{Xl} \\ \vdots & & & \ddots & \vdots \\ P_{AX} & \cdots & P_{jX} & \cdots & P_{XX} \end{bmatrix} \tag{13}$$

$$\sum_j P_{jl} = 1, \quad l \in \mathrm{A} \tag{14}$$

親子関係，親配列は未知である．子$h$の親が親$k$である確率を式(15)のように定義する．

$$\Pr(a_h = k) = a^{ih}_k \tag{15}$$

$$\sum_k a^{ih}_k = 1 \tag{16}$$

親配列を$y$とおき子配列と同様に定義すると，

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_k \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} y_1^1 & \cdots & y_n^1 & \cdots & y_N^1 \\ \vdots & \ddots & & & \vdots \\ y_1^k & & y_n^k & & y_N^k \\ \vdots & & & \ddots & \vdots \\ y_1^K & \cdots & y_n^K & \cdots & y_N^K \end{bmatrix} \tag{17}$$

となる．親$k$の$n$番目の位置にアミノ酸$j$が存在する確率を式(18)のように定義する．

$$\Pr(y^k_n = j) = b^k_{nj} \tag{18}$$

$$\sum_{j \in \mathrm{A}} b^k_{nj} = 1 \tag{19}$$

以上のことを用いて，子供$h$の親が$k$，親$k$の$n$番

目のアミノ酸が $j$ である時，子供 $h$ の $n$ 番目のアミノ酸が $l$ である確率 $p'$ は次のようになるとする．

$$p'(x_n^h = l, a_h = k, y_n^k = j; P) = a_k^{ih} \cdot b_{nj}^k \cdot P_{jl} \qquad (20)$$

すると，子供 $h$ の $n$ 番目のアミノ酸が $l$ である時，子供 $h$ の親が $k$，親 $k$ の $n$ 番目のアミノ酸が $j$ である確率 $p$ は次のようになる．

$$p(a_h = k, y_n^k = j \mid x_n^h = l; P) = \frac{a_k^{ih} \cdot b_{nj}^k \cdot P_{jl}}{\sum_{j \in A} \sum_{k=1}^{K} \left\{ a_k^{ih} \cdot b_{nj}^k \cdot P_{jl} \right\}} \qquad (21)$$

以下にアルゴリズムを記述する．

初期化：初期値 $a_k^{ih(t)}, b_{nj}^{k(t)}, P^{(t)}$ を設定し，$t \leftarrow 0$ とする．

反復計算：以下を収束するまで繰り返す．

Eステップ（期待値計算）

$t$ 回目の反復の $a_k^{ih}, b_{nj}^k, P$ を $a_k^{ih(t)}, b_{nj}^{k(t)}, P^{(t)}$ とする． $Q$ 関数を計算すると，式(22)のようになる．

$$Q(a_k^{ih}, b_{nj}^k, P \mid a_k^{ih(t)}, b_{nj}^{k(t)}, P^{(t)}) = \frac{1}{HN} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j=1}^{21} \sum_{k=1}^{K} \{ p \cdot \log p' \} \qquad (22)$$

$$= \frac{1}{HN} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j=1}^{21} \sum_{k=1}^{K} \frac{a_k^{ih} \cdot b_{nj}^k \cdot P_{jl}}{\sum_{j=1}^{21} \sum_{k=1}^{K} \{ a_k^{ih} \cdot b_{nj}^k \cdot P_{jl} \}} \cdot \log(a_k^{ih} \cdot b_{nj}^k \cdot P_{jl})$$

Mステップ（最大化）

Eステップで求めた $Q(a_k^{ih}, b_{nj}^k, P \mid a_k^{ih(t)}, b_{nj}^{k(t)}, P^{(t)})$ を最大にする $a_k^{ih}, b_{nj}^k, P$ を $a_k^{ih(t+1)}, b_{nj}^{k(t+1)}, P^{(t+1)}$ とする．

$$(a_k^{ih(t+1)}, b_{nj}^{k(t+1)}, P^{(t+1)}) = \arg\max Q\{(a_k^{ih}, b_{nj}^k, P) \mid (a_k^{ih(t)}, b_{nj}^{k(t)}, P^{(t)})\} \qquad (23)$$

## 3. シミュレーション実験

### 3.1 アラインメント

子配列データとしてはタンパク質のアミノ酸配列を用いるが，その長さは一定ではない．そこで，アラインメントを行い，適当な長さに揃える必要がある[4]．

アラインメントとは，"X"（アミノ酸が欠失した状態であることを表す）を挿入することにより，2つまたは複数の配列を，一致するアミノ酸が最も多くなるように整列させる手法である．アラインメントの候補は複数存在するため，スコア行列(突然変異確率を考慮した行列)に基づいて，減点を行う．例えば，図4のような長さの異なる2配列が存在するとする． case 1 の場合なら，Lが2カ所消失し，N，LがYに，PがNに変異している．図5のスコア行列によるとNがYに変わる減点は 1，LがYに変わる減点は 2，PがNに変わる減点は 3，Lが消失する減点は 2 なので，case 1 の減点は合計 10 となる．同様に，case 2 についても減点を算出してやり，減点の最も少ないものを最終的なアラインメントとして採用する．

| | | |
|---|---|---|
| | sequence1 | LANLANP |
| | sequence2 | LAYYALNNL |
| | ↓ | |
| case 1 | sequence1 | LANLAXNPX |
| | sequence2 | LAYYALNNL |
| case 2 | sequence1 | LANXXLANP |
| | sequence2 | LAYKALNNL |

図 4 アラインメント

| | A | L | N | P | Y | X |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| L | 3 | 0 | | | | |
| N | 1 | 6 | 0 | | | |
| P | 2 | 8 | 3 | 0 | | |
| Y | 5 | 2 | 1 | 6 | 0 | |
| X | 2 | 2 | 2 | 2 | 2 | 2 |

図 5 スコア行列

スコア行列は生物学上意味のあるデータでなければならない．そこで，先祖が共通のタンパク質を集め，置換の頻度を調べて求めた置換行列[4]を使用する．しかし，集めるタンパク質によって置換頻度は一定ではないので，どれが RecA の進化の過程を最も良く表しているかを事前知識のみで判断するのは困難である．そこで，本報告では，最も一般的な PAM250 行列[5]，BLOSUM50 行列[4]の2つのスコア行列を使用した．

### 3.2 検証方法

既知の RecA の配列を 3.2.1 に示したように確率的に変異させて人工的な子配列データを作成した．MCMC 法，と EM 法の2つの手法で親配列を推定し，両者の比較を行った．

#### 3.2.1 子配列の作成方法

子配列を作成する際，親のアミノ酸がそのまま子に受け継がれる確率を 0.8，突然変異が起きる確率を 0.15，アミノ酸の挿入が起きる確率を 0.025，欠

失が起きる確率を 0.025 とした．変異が起きる場合は，PAM 行列または BLOSUM 行列の値を確率値にしたものを使った．1 個の親から 5 個ずつ合計 15 個の配列を作った．

### 3.2.2 推定方法

3.2.1 で示した方法で異なるデータを 10 個ずつ生成し，そのデータを用いて以下の 1 から 8 の方法で 10 回ずつ推定を行った．アラインメントをとる際 clustalW[6]という web site を使用した．

BLOSUM によって生成されたデータを

| 方法 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| アラインメント | PAM | BLOSUM | PAM | BLOSUM |
| 推定方法 | EM | EM | MCMC | MCMC |

PAM によって生成されたデータを

| 方法 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| アラインメント | PAM | BLOSUM | PAM | BLOSUM |
| 推定方法 | EM | EM | MCMC | MCMC |

その際，真の親配列と推定した配列の一致率，配列の長さ，そして，EM 法で導出される突然変異確率が実際の突然変異確率に対してどの程度異なるかに注目した．

## 3.3 実験結果

### 3.3.1 実験精度

それぞれの方法の下で推定した配列と真の配列の一致率[%]を図 6 にプロットした．横軸の番号は推定方法，縦軸は[%]を表している．なお，平均と分散は表 1 に示した．ここで，EM 法と MCMC 法の推定結果に有意差が存在するかを見るために，分散値，平均値の検定を行った．分散値の検定は有意水準 5% の F 検定，平均値の検定は Welch の検定を用いた．その結果を表 2 に示す．MCMC 法の推定結果よりも EM 法の方が，分散が小さく高い精度を示していることがわかった．

また，有意水準 5%の F 検定，t 検定を用いて，EM 法で実際と異なる確率過程を仮定してしまった時と正しい確率過程を仮定した場合の推定精度に有意差があるか検定を行った．結果，誤った場合でも同じ程度の精度を出せることが示された(表 3)．

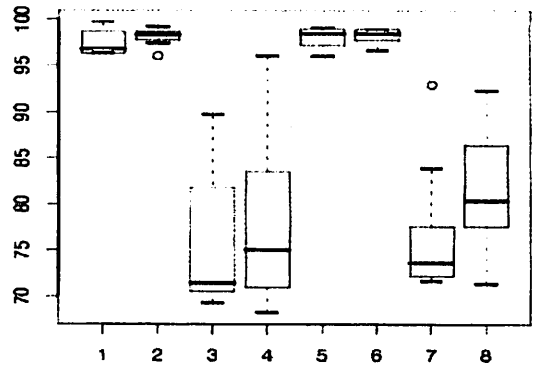



図 6 推定精度[%]の比較

表 1 推定精度[%]の平均，分散

| 方法 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 平均 | 97.4 | 98.2 | 75.8 | 77.7 | 98.2 | 98.0 | 81.6 | 76.7 |
| 分散 | 1.61 | 0.87 | 52 | 81 | 0.54 | 1.06 | 50 | 46 |

表 2 EM 法と MCMC 法の推定精度

| 推定方法 | 平均 | 分散 |
|---|---|---|
| EM (1, 2, 5, 6) | 98.0 | 1.05 |
| MCMC (3, 4, 7, 8) | 78.0 | 58.03 |
| 検定結果 | 有意差あり | 有意差あり |

表 3 EM 法で前提条件が異なる時の推定精度

| 前提条件 | 平均 | 分散 |
|---|---|---|
| 正しい過程を仮定(2, 5) | 98.2 | 0.070 |
| 誤った過程を仮定(1, 6) | 97.7 | 0.216 |
| 検定結果 | 有意差なし | 有意差なし |

### 3.3.2 配列長

真の親配列の長さ 352 に対する推定された配列の長さのばらつきを表した(図 7)．縦軸は配列の長さ(アミノ酸の数)，横軸は方法を表す．これは，推定精度と同様，PAM でアラインメントをとり EM 法で推定を行った場合に，真値と同じ長さの配列が推測される確率が高いことを示している．

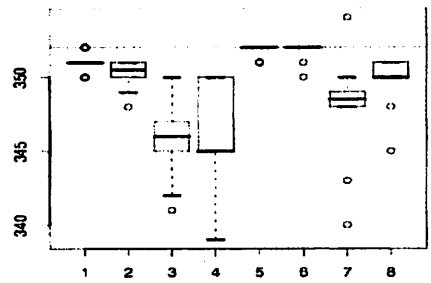



図 7 配列長の比較

3.3.3 突然変異確率の分散

推定した突然変異確率と実際の突然変異確率の差の二乗のヒストグラムを図 8 に示す．上が正しい確率過程を仮定した場合，下は誤った確率過程を仮定した場合である．平均値と分散値は表 4 に示した(有意水準 5%の F 検定，t 検定)．両者に差はなく，誤った確率過程を仮定した場合でも同じ程度の精度で突然変異確率の推定を行えることがわかった．

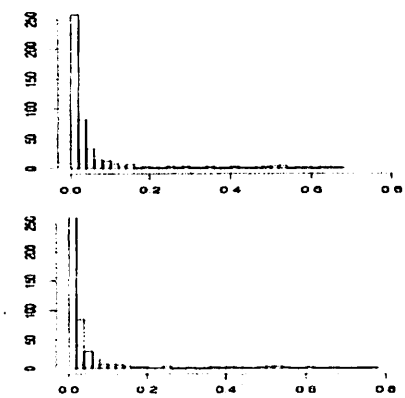


図 8 突然変異確率の誤差

表 4 突然変異確率の誤差

| | 平均 | 分散 |
|---|---|---|
| 正しい過程を仮定 | 0.0506 | 0.0124 |
| 誤った過程を仮定 | 0.0515 | 0.0131 |
| 検定結果 | 有意差なし | 有意差なし |

3.3.4 考察

今回は複数の親から子配列を作成したが，複数ではなく 1 個の親から作成した 10 個の子配列を用いて 1 個の親を推定した場合，MCMC 法使用モデルの推定結果は EM 法とほぼ同等の精度を維持していた．しかし，MCMC 法使用モデルで複数の親配列を推測すると精度が大幅に悪くなった．これは親子関係を断定的に 1 つの値に持たせてしまったためと考えられる．今回提案した EM 法使用モデルは，突然変異確率など新しいパラメータを導入したが，親子関係を確率的に記述したためにうまくいったと考えられる．

## 4. 終わりに

今回，EM 法では最初に仮定した確率過程が誤っていても，正しい確率過程を仮定した時と変わらない推定精度を持つことがわかった．しかし，実際のタンパク質の配列データを使っていく上で，配列に適した置換行列を知ることは重要である．例えば，BLOSUM は長くて似ていない配列のアラインメントには適していない．したがって，実データでは生物学的見地も必要となっていくと思われる．

また，子配列をいくつのグループに分けるのが最適かを事前に決定することは困難であり，今後どのようにモデル選択を行っていくかも考える必要がある．

## 文 献

[1] Jonathan.K.Pritchard, Matthew Stephens and Ppeter Donnelly Inference of Population Structure Using Multilocus Genotype Data Genetics Society of America Genetics 155:945-959, 2000

[2] 金芳，田栗正章，手塚集，椿島祥介，上田修功，計算統計 I 確率計算の新しい手法，岩波書店，2003

[3] 伊庭幸人，種村正美，大森裕浩，和合肇，佐藤整尚，高橋明彦，計算統計 II マルコフ連鎖モンテカルロ法とその周辺，岩波書店，2005

[4] Richard Durbin, Sean R. Eddy, Anders Krogh, Graeme Mitchison, 阿久津達也，浅井潔，矢田哲士，バイオインフォマティックス 確率モデルによる遺伝子配列解析，医学出版，2001

[5] http://www.cmbi.kun.nl/gvteach/aainfo/pam250.shtml

[6] http://align.genome.jp/