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

2006－BIO－5 (12)
2006／6／15

ベイジアンネットワークによる突然変異モデルの作成と Belief Propagationによる推定

土井 慶紀† 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
〒169-8555 東京都新宿区大久保 3-4-1
†† 早稲田大学
E-mail: †yoshinori.doi@murata.eb.waseda.ac.jp, ††noboru.murata@eb.waseda.ac.jp

あらまし 太古のDNA修復タンパク質は現在のものよりも活性が高かったと考えられており，そのタンパク質が得られれば，癌などに対する遺伝子治療に役立つ可能性がある．そこで，現在分かっているタンパク質のアミノ酸配列情報から，太古のタンパク質のアミノ酸配列を推定することを考えている．複数のタンパク質から，その祖先にあたるタンパク質を推定するための突然変異モデルを，ベイジアンネットワークを用いて作成し，従来のBelief Propagationを改変することで，提案モデルに適用できるようにした．また，cross-validationを用いて最適なパラメータの選択も行った．計算機上で仮想的に，ある祖先タンパク質から特定の突然変異確率に従うタンパク質を複数作成し，作成したタンパク質に提案モデルを適用する事で祖先配列を推定した．その精度評価について報告する．

キーワード ベイジアンネットワーク，ベイズ推定，Belief Propagation，突然変異モデル

Buliding a New Mutation Model based on Bayesian Networks and estimation by a modified Belief Propagation algorithm

Yoshinori DOI† and Noboru MURATA††

† Waseda University
3-4-1 Ohkubo, Shinjuku, Tokyo 169-8555, JAPAN
†† Waseda University
E-mail: †yoshinori.doi@murata.eb.waseda.ac.jp, ††noboru.murata@eb.waseda.ac.jp

Abstract It is thought that ancient DNA repair proteins were more active than present ones, and obtaining the ancient DNA repair proteins would be helpful for hereditary diseases such as cancer. We propose a mutation model using Bayesian Networks and a modified the belief propagation algorithm, which is applied to estimate an ancient protein from multiple proteins that are present. Also, we applied the cross-validation to derive an optimum parameter. We virtually generated multiple proteins from an ancient protein following a specific mutation probability, and evaluated our model by estimating the ancient protein by the proposed model.

Key words Bayesian Networks, Bayes estimation, Belief Propagation, Mutation model

1. はじめに

全ての生物は傷付いたDNAを修復することの出来るタンパク質を持っている．DNA配列は紫外線などの要因から常に傷付けられているが，このタンパク質の働きにより常に修復されており，このタンパク質がないと生物はDNAの配列を保持することが出来ず死に至る．

ところで，太古の地球はオゾン層がなく紫外線が現在の地球よりも非常に強かった事などから，現在のDNA修復速度では生命活動を維持することが難しく，より修復速度が速い，すなわち活性が高かったであろうと考えられている．

より活性の高いDNA修復タンパク質を得ることができれば，現存する修復タンパク質の活性が低いために実現が出来ていない癌の遺伝子治療などに新たな道を示すことになる．そこで本報告では，現在わかっているDNA修復タンパク質のアミノ酸配列から太古のタンパク質のアミノ酸配列を推定することを目的とした突然変異モデルとその推定方法を提案する．

タンパク質のアミノ酸配列は進化の過程で置換，欠失，挿入を

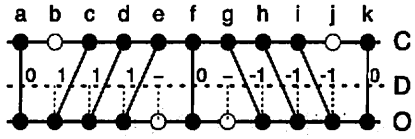

図 1 突然変異モデル (子配列が 1 つのとき)

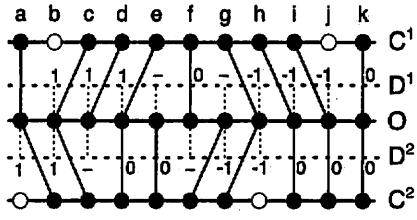

図 2 突然変異モデル (子配列が複数のとき)

起こしながら少しずつその配列を変化させていく. このとき, 欠失, 挿入は配列の長さを変化させ, ズレを生むことになる. 従って, このズレをとらえて推定する必要がある. 本報告では, 左右 2 枚の画像からの 3 次元立体構造を推定するステレオビジョン問題 [1], [2] で用いられるモデル構造をタンパク質に適用できるように改変した. ステレオビジョンでは左右 2 枚の画像のズレを考慮するモデルとなっているが, 本報告の提案モデルではアミノ酸配列のズレを考慮できるようにしてある.

以下,2 節で突然変異モデルの定義について述べ,3 節で Belief Propagation による推定の手法を説明する. その後,4 節で実験と考察を行い,5 節でまとめる.

2. 突然変異モデルの定義

2.1 確率モデル

1 つのアミノ酸配列が存在するとする. 進化の過程でこのアミノ酸配列は少しずつ変化をしていき, さらに複数の種類に分かれていく. このとき, 最初のアミノ酸配列を親配列, 進化してできた配列を子配列と呼ぶことにする.

同一の親配列から進化したと考えられる子配列が複数存在しているとき, この子配列から親配列を推定する事を考える.

アミノ酸配列は, 一般的に以下の 3 つの突然変異を起こす.

1. アミノ酸の置換
2. アミノ酸の欠失
3. アミノ酸の挿入

本報告では,2 と 3 によって生まれる位置のズレを Gap と呼ぶ. タンパク質はこの過程を繰り返す事で, 少しずつ進化をしていく. その際, 欠失, 挿入によって親配列と子配列の対応するアミノ酸の位置がズレていくので, それを考慮した確率モデルが必要となる.

本報告では, 図 1 のような 3 層構造のベイジアンネットワークモデルを考える.C は既知の子配列を表し, 以下のような配列長 N のベクトルである (−は Gap を表す).

$$\mathbf{C} = \{C_1, C_2, \cdots, C_s, \cdots, C_N\} \tag{1}$$

$$C_s \in A \quad (1 \leqq s \leqq N) \tag{2}$$

$$A = \{\text{20 種のアミノ酸}, -\} \tag{3}$$

子配列データの例(配列長 50)

MPAEMKSAASGSDPRSSGERDKALNLVLGQIERNFGKGSIMRLGDASRMR

O は配列長 N の親配列を表し, データ構造は子配列と同様である.

$$\mathbf{O} = \{O_1, O_2, \cdots, O_s, \cdots, O_N\} \tag{4}$$

$$O_s \in A \tag{5}$$

親配列は未知であるので, 親配列の s 番目の部位 O_s がアミノ酸 i であると仮定したときの確率を $P(O_s = i) = o_{si}$ とし, これを全ての s,i について推定する. ただし,$\sum_{i \in A} o_{si} = 1$ でなくてはいけない. 推定する親配列の確率表を以下に示しておく.

親配列データの確率表の例 (列ベクトルの和が 1)

	配列長 N			
ア	o_{1A}	o_{2A}	$\cdots$	o_{NA}
ミ	o_{1C}	o_{2C}	$\cdots$	o_{NC}
ノ	$\vdots$	$\vdots$	$\ddots$	$\vdots$
酸	o_{1-}	o_{2-}	$\cdots$	o_{N-}

D は親配列が子配列のどの部位に対応するか (深度と呼ぶ) を表す. 例えば, 図 1 の c の部位に注目すると, 親配列の b の部位と対応していることが分かる. このとき, 深度は 1 であるとする. どこまでの位置のズレを考慮するかは事前に決定しておく必要がある. 深度を左右 2 つ目までのズレを考慮するとしたときの例は以下のようになる. ただし, 深度も未知データである.

$$\mathbf{D} = \{D_1, D_2, \cdots, D_s, \cdots, D_N\} \tag{6}$$

$$D_s \in \{-2, -1, 0, 1, 2, -\} \quad (1 \leqq s \leqq N) \tag{7}$$

ここで, 図 1 では子配列が 1 つであるが, 本来 C は複数存在する. そのとき各個体について親配列は深度を持っていると仮定する. 子配列が 2 つの場合の例を図 2 に示す. このとき, 各子配列の配列長が異ならないように, アラインメントなどを行って長さを予め揃えておく必要がある (Gap が含まれていても良い). アラインメントについては 4.1 節で説明する.

子配列が H 個体存在しているとき,h 個目の子配列を $\mathbf{C}^h$, そのときの深度を $\mathbf{D}^h$ とする. 親配列と深度 $\{\mathbf{O}, \mathbf{D}^1, \mathbf{D}^2, \cdots, \mathbf{D}^H\}$ の事後確率は以下の式 (8) のようになる.

$$P(\mathbf{O}, \{\mathbf{D}\}|\{\mathbf{C}\}) = \frac{P(\{\mathbf{C}\}|\mathbf{O}, \{\mathbf{D}\})P(\mathbf{O}, \{\mathbf{D}\})}{P(\{\mathbf{C}\})} \propto P(\{\mathbf{C}\}|\mathbf{O}, \{\mathbf{D}\})P(\mathbf{O}, \{\mathbf{D}\}) \tag{8}$$

ただし,$\{\mathbf{C}\}$ は $\mathbf{C}^1, \mathbf{C}^2, \cdots, \mathbf{C}^H$ を,$\{\mathbf{D}\}$ は $\mathbf{D}^1, \mathbf{D}^2, \cdots, \mathbf{D}^H$ を表す.

2.2 尤　　度

各子配列の間に依存関係がないと仮定すると,O, D を変数と見たときの子配列の尤度は以下の式 (9) のように各子配列の尤度の積となる.

$$P(\{\mathbf{C}\}|\mathbf{O},\{\mathbf{D}\}) = P(\mathbf{C}^1|\mathbf{O},\mathbf{D}^1)\times\cdots\times P(\mathbf{C}^H|\mathbf{O},\mathbf{D}^H) = \prod_{h=1}^{H} P(\mathbf{C}^h|\mathbf{O},\mathbf{D}^h) \quad (9)$$

本報告では, この尤度を以下の式 (10) ように仮定した.D_s^h は,h 個目の子配列の s 番目の深度を表し,C_s^h は h 個目の子配列の s 番目のアミノ酸を表す.

$$P(\mathbf{C}^h|\mathbf{O},\mathbf{D}^h) = \prod_s P(\mathbf{C}^h|O_s,D_s^h) \propto \prod_s F(O_s, C^h_{s+D_s^h}) \quad (10)$$

F は設計者が事前に決める確率である. タンパク質の問題では, 一般に突然変異確率が用いられる. 本報告では,F は以下の式 (11) のような確率とした.

$$F(O_s,D_s^h,\mathbf{C}^h) = \begin{cases} a & \text{if } O_s = C_{s+D_s^h} \\ b & \text{if } O_s \neq C_{s+D_s^h} \\ g & \text{if } D_s^h = - \end{cases} \quad (11)$$

a は親配列のアミノ酸が子配列に保存される確率,b は置換される確率で本報告では一様性を仮定した. また g は Gap penalty と呼び Gap が入る確率である. ただし,$a > b$ で $a + 19b + g = 1$ となる.

2.3 事前分布

次に事前分布 $P(\mathbf{O},\{\mathbf{D}\})$ を考える. 前節で各子配列の深度同士には依存関係がないと仮定した. さらに, 親配列と深度にも依存関係がないと仮定すると, 事前分布は以下の式 (12) のように分解出来る.

$$P(\mathbf{O},\{\mathbf{D}\}) = P(\mathbf{O})\prod_{h=1}^{H} P(\mathbf{D}^h) \quad (12)$$

ここで, 各 O_s は部位ごとに独立, 深度は隣りの影響を受けると仮定し,$P(\mathbf{O},\mathbf{D}^h)$ を考えると, 以下の式 (13) のようになる.

$$P(\mathbf{O},\mathbf{D}^h) = \prod_{s=1}^{N} P(O_s)\prod_{s=1}^{N-1} P(D_s^h, D_{s+1}^h) \quad (13)$$

$P(O_s)$ は 2.1 節の事前分布,$P(D_s^h, D_{s+1}^h)$ は設計者が事前に決定しておく相互確率である. 本報告では図 3 のように,D_s^h と D_{s+1}^h が近ければ値が大きく, 遠くなるほど値が小さくなる確率値としている (ただし,-のときは Gap penalty が入る).

このとき,D_s^h と D_{s+1}^h が近いときと遠いときの確率値の差を相互作用と呼び, これが強いほど差が大きく, 弱くなるほど差が小さくなっていくとして, 相互確率 $P(D_s^h, D_{s+1}^h)$ を制御するようにした. 図 4 にそのイメージ図を示す.

以上の結果から, 式 (8) は以下の式 (14) のようになる.

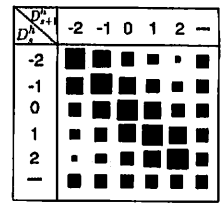

図 3　D_s^h と D_{s+1}^h の相互確率行列 (イメージ図)
(■が大きい方が値が大きい確率行列)

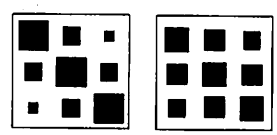

図 4　相互作用の強さによる相互確率の違い (イメージ図)

$$P(\mathbf{O},\{\mathbf{D}\}|\{\mathbf{C}\}) \propto \prod_{h=1}^{H}\left\{\prod_{s=1}^{N} P(\mathbf{C}^h|O_s,D_s^h)P(O_s)\prod_{s=1}^{N-1} p(D_s^h,D_{s+1}^h)\right\} \quad (14)$$

3. Belief Propagation による推定

3.1 Belief Propagation

推定に用いる Belief Propagation(BP) [2] [3] [5] についてまとめておく. 一般の BP は以下のように定式化される.

$\mathbf{x} = (x_1, x_2, \cdots, x_N)$ が未観測データ,$\mathbf{y} = (y_1, y_2, \cdots, y_N)$ が観測データであるとする. このとき, 未観測データの結合分布が以下の式 (15) のように書けるとする.

$$P(\mathbf{x}|\mathbf{y}) \propto \prod_s \phi_s(x_s,y_s)\prod_{(st)}\psi_{st}(x_s,x_t) \quad (15)$$

(st) は隣り合う部位の組み合わせを表す.

$\phi_s(x_s,y_s)$ は s 番目の部位の未観測データ x_s における local evidence を表し, 一般に観測データの局所的な尤度 $P(y_s|x_s)$ が使われる.$\psi_{st}(x_s,x_t)$ は x_s と x_t の compatibility matrix と呼ばれ, 隣り合う部位の相互作用を表す.

このとき, 以下の式 (16) で, 部位 t が持つ部位 s の情報を s へ伝搬するメッセージ m の更新式を定義する.

$$m_{ts}(x_s) \leftarrow k\max_{x_t}\phi_s(x_t,y_t)\psi_{st}(x_s,x_t)\prod_{u\in n(t)\setminus s} m_{ut}(x_t) \quad (16)$$

k は正規化項であり,$n(t)$ は t と隣り合う全ての部位を表し,$n(t)\setminus s$ は s 以外の t と隣り合う全ての部位を表す.

全てのノード間のメッセージを計算し, 以下の式 (17) を解くことで, 推定したい未観測データの MAP 解が得られる.

$$b_s(x_s) = k\phi_s(x_s,y_s)\prod_{t\in n(s)} m_{ts}(x_t) \quad (17)$$

$$x_s^{MAP} = \operatorname*{argmax}_{x_s} b_s(x_s) \quad (18)$$

これは max-product BP と呼ばれ，メッセージの更新則に max を利用する [2]．これに対し，sum-product BP と呼ばれる更新則 (max の所に sum を利用する) も存在し，未観測データの周辺尤度が得られる [3], [5]．

3.2 突然変異モデルの更新則

式 (9),(12) から，$P(\mathbf{O}, \{\mathbf{D}\}|\{\mathbf{C}\})$ は以下のように分解出来る．

$$P(\mathbf{O}, \{\mathbf{D}\}|\{\mathbf{C}\}) = \prod_{h=1}^{H} P(\mathbf{O}, \mathbf{D}^h|\mathbf{C}^h) \tag{19}$$

このとき式 (14) より，各 h についての $P(\mathbf{O}, \mathbf{D}^h|\mathbf{C}^h)$ は以下の式 (20) のようになる．

$$P(\mathbf{O}, \mathbf{D}^h|\mathbf{C}^h) \propto \prod_{s=1}^{N} P(\mathbf{C}^h|O_s, D_s^h)P(O_s) \prod_{(st)} P(D_s^h, D_t^h) \tag{20}$$

ここで，式 (15) と式 (20) を比較すると，ϕ_s, ψ_{st} はそれぞれ以下の式 (21),(22) のようになる．

$$\phi_s = P(\mathbf{C}^h|O_s, D_s^h)P(O_s) \tag{21}$$

$$\psi_{st} = P(D_s^h, D_t^h) \tag{22}$$

従って，max-product BP のメッセージ m の更新式は，以下の式 (23) のようになる．

$$m_{ts}(\mathbf{O}, D_s^h) \leftarrow k \max_{D_t^h} \left\{ P(\mathbf{C}^h|O_t, D_t^h)P(O_t) P(D_s^h, D_t^h) \prod_{u \in n(t)\backslash s} m_{ut}(\mathbf{O}, D_t^h) \right\} \tag{23}$$

ここで，式 (23) を，$\mathbf{O}$ について max を取ることで周辺化する．

$$\max_{\mathbf{O}} m_{ts}(\mathbf{O}, D_s^h) \leftarrow k \max_{O_t, D_t^h} \left\{ P(\mathbf{C}^h|O_t, D_t^h)P(O_t) P(D_s^h, D_t^h) \prod_{u \in n(t)\backslash s} \max_{\mathbf{O}} m_{ut}(\mathbf{O}, D_t^h) \right\} \tag{24}$$

ここで，$\max_{\mathbf{O}} m_{ut}(\mathbf{O}, D_t^h)$ は O_t には依存していないので，事前に max を取り $m_{ut}(D_t^h)$ としておくことで，式 (24) は以下の式 (25) のように変形できる．

$$\max_{O_t} m_{ts}(O_t, D_s^h) \leftarrow k \max_{O_t, D_t^h} \left\{ P(\mathbf{C}^h|O_t, D_t^h)P(O_t) P(D_s^h, D_t^h) \prod_{u \in n(t)\backslash s} m_{ut}(D_t^h) \right\} \tag{25}$$

$\max_{O_t} m_{ts}(O_t, D_s^h) = m_{ts}(D_s^h)$ とすると，最終的な更新式は以下の式 (26) のようになる．

$$m_{ts}(D_s^h) \leftarrow k \max_{O_t, D_t^h} \left\{ P(\mathbf{C}^h|O_t, D_t^h)P(O_t) P(D_s^h, D_t^h) \prod_{u \in n(t)\backslash s} m_{ut}(D_t^h) \right\} \tag{26}$$

D_s^h の MAP 解は以下の式 (27),(28) を解けば求められる．

$$b_s(D_s^h) = k \max_{O_s} P(\mathbf{C}^h|O_s, D_s^h)P(O_s) \prod_{t \in n(s)} m_{ts}(D_s^h) \tag{27}$$

$$D_s^{h\,MAP} = \operatorname*{argmax}_{D_s^h} b_s(D_s^h) \tag{28}$$

子配列 $\mathbf{C}^h$ に対する親配列の事後確率は，以下の式 (29) を解く事で得られる．

$$P(O_s|\mathbf{C}^h) = k \max_{D_s^h} P(\mathbf{C}^h|O_s, D_s^h)P(O_s) \prod_{t \in n(s)} m_{ts}(D_s^h) \tag{29}$$

さらに，子配列は複数存在するので，最終的に親配列の事後確率は，以下の式 (30) のようになる．

$$P(O_s|\{\mathbf{C}\}) = k \prod_{h=1}^{H} \left\{ \max_{D_s^h} P(\mathbf{C}^h|O_s, D_s^h)P(O_s) \prod_{t \in n(s)} m_{ts}(D_s^h) \right\} \tag{30}$$

この結果をメッセージの更新則の $P(O_s)$ にフィードバックすることによって，より良い親配列を逐次的に推定する．

3.3 突然変異モデルの BP アルゴリズム

入力　アラインメントした上で配列長を N に揃えた H 個の子配列データ

初期化　全ての h, s に対して

$m_{ts}(D_s^h) \leftarrow 1 \quad (^\forall t)$

$P(O_s) \leftarrow$ 初期値を設定

反復計算　以下を親配列の事後分布 $P(O_s|\{\mathbf{C}\})$ が収束するまで繰り返す

- message の更新

全ての h, s に対して式 (25) を計算し，$m_{ts}(D_s^h)$ を更新する

- 事後分布 $P(O_s|\{\mathbf{C}\})$ の推定と $P(O_s)$ の更新

全ての s に対して式 (30) を計算し，$P(O_s)$ を更新する

$P(O_s) \leftarrow P(O_s|\{\mathbf{C}\})$

出力　得られた事後分布 $P(O_s|\{C\})$ より，親配列の MAP 解を推定する

全ての s に対して

$O_s^{MAP} = \operatorname*{argmax}_{O_s} P(O_s|\{C\})$

図 5　変異モデルに対する BP 法のアルゴリズム

$P(\mathbf{C}^h|O_s, D_s^h), P(D_s^h, D_t^h)$ は既知の確率として設定し，$P(O_s)$ は初期値を定める．このとき，変異モデルのアルゴリズムは図 5 のようになる．

4. 実験と考察

4.1 アラインメント

アラインメントとはDNAやアミノ酸配列の似ている部分を揃えることである[4][6]. 例を図6に示す. 図6で, 一番左側では上下で一致している部分が1箇所であるアミノ酸配列に対して,3箇所にGap(-)を入れることにより一致している部分を3箇所に増やす事ができる. さらに,AからTへアミノ酸が置換したとすると, 一致している箇所を3箇所に保ったまま, 挿入するGapを1つに減らすこともできる.

以上のように, アラインメントとは適切にGapを挿入することで, 2つ(もしくは複数)の配列の相同性を高める手法のことである.

```
MADG      MA--DG      MA-DG
MTEDG  →  M-TEDG  →   MTEDG
*         *   **      *  **
```

図6 簡単なアラインメントの例(*は一致した箇所を表す)

4.2 準　　備

この節では, 計算機上で人工的に作成したデータを用いて,3.2節で示したアルゴリズムを適用し, どの程度正しい親配列が得られるかを検証する.

まず適当に親配列を選び, そこから計算機上で子配列を10個体生成する. このとき, アミノ酸が変異する確率は0.15で, 突然変異確率はPAM250[8]を用いた. さらに,Gapの確率も0.15とした. 親配列と生成した子配列の一部を以下に示す.

親配列 (長さ $N = 369$ の配列の一部を示す)

```
···MARTTDDSKKAAPAAGTADEAQKQKALKHVLTQIKRNFGEGAIMRLGENTRIRV···
```

子配列 (上から[1]~[10]の10個体の子配列の一部を示している)

```
···MAAYLGTDDSKKDANDASGTYAIYEAAEKQKALNMQVLTQIKNHPFGGIMRLGT···
···MARTTWDSKKAAPAAPTWDEAQKQPQALSPVLTEIKRNFHGAIMRLQGENTRIQ···
···MARTTDDSKKAAPAAGTADIAKKKQRALMVLTQIKRNFGEGCAQIMRFGETHTR···
···MARTTDDDSKRHAAAAGTPHYEAKKLKMVLTQIRNFGEGAIGRGFTRIRVWTVP···
···MYRTTDDSKKAAASGTGEEAQWQSALMVLTQIKRAQFGEGCIMRGPNTKMIGRV···
···MMRNTTDDYSKKAAAATKDEQYQATKMWLQRMKGEGAMRLGNRIRVTYWSGAIR···
···MARTDSKAANPAAGTADEMKQKMLYKGVLTQIKRFGEGAIMRLGENTRRRHEYV···
···MRARTTHSKKAAVLAEETTADEAQKQKALLKMVLTQVWIRNCGEGALMRFGENT···
···MARTFDDSKAAPAAGWQDEAQQKAKMVKQQIKKDNCGEGAQIMCLGENHRERVE···
···MARTTDDWKKAAPAAGTADEAPQLVWKAPRLKCVLCQIKQNGFGEQAIMRQLGE···
```

この子配列からアラインメントを取った結果の一部を以下に示しておく. この結果は,CLUSTALWを提供するwebサイト[7]によって行っている.

アラインメント結果 (上の10個体の子配列のアラインメント結果)

```
···MAAYLGTDDSKKDANDASGTYAIYEAAEKQKALNMQVLTQIKNHPFGG-IMRLG···
···-MARTTWDSKKAAPAAPTWD--EAQK---QPQALSPVLTEIKRN----FHGAIM···
···-MARTTDDSKKAAPAAGTAD--IAK----MKQRALMVLTQIKRN-FGEGCAQIM···
···-MARTTDDDSKRHAAAAG-------TPHYEAKKLKMVLTQIR--NFGEGAIGRG···
···-MYRTTDD-SKKAAASGT-------GEEAQWQSALMVLTQIKRAQFGEGCIMRG···
···MMRNTTDDYSKKAAAATKDEQ---YQAT------KMWLQ----RMKGEGA-MRL···
···-MART-DSKAAWPAAGTAD-------ENKQKMLYKGVLTQIKR--FGEGA-IMR···
···-MRARTTHSKKAAVLAEETTA---DEAQKQKALLKMVLTQVWIRNCGEGALMRF···
···-MARTFDDSKAAPAAGWQD-------EAQQK--AKMVKQQIKKDNCGEGAQIMC···
···-MARTTDDWKKAAPAAGTAD--EAPQLVWKAPRLKCVLCQIKQNGFGE-QAIMR···
```

BPの適用の際に, このアラインメントの結果を利用する.

4.3 シミュレーション結果

まず比較対象として,3.1節で示したアラインメントの結果から各部位ごとに多数決を取ることで親配列を推定した. さらに, 推定した配列が, どの程度真の親配列と一致しているかを数値的に示すため, 誤差率を以下の計算で行った.

$$\text{誤差率} = \{1 - (\text{正答数})/(\text{推定した配列の長さ})\} \qquad (31)$$

推定結果が親配列を完全に一致すると誤差率は0となる. その結果, 多数決では誤差率は0.123となった.

次に,$P_p(O_s, C_{s+D_s^h})$ を以下の式(32)のような一様分布, 深度は左右2つ目までのズレを考慮するとして, $P(D_s^h, D_t^h)$ を変化させながらシミュレーションを行った.$P(O_s)$ の初期値は, アラインメントの結果 s 番目の部位に最も多かったアミノ酸の確率を0.5, それ以外を0.025とした. その結果を図7に示す. 図7は4.2節で示した10個の子配列を用いて,$P(D_s^h, D_t^h)$ の相互作用の強さとGap penaltyの値を少しずつ変化させながら350回実験を行い, その誤差率のバラツキを見た図である. 横軸は誤差率, 縦軸は頻度を表している.

$$P_p(O_s, C_{s+D_s^h}) = \begin{cases} 0.01 & \text{if } O_s \neq C_{s+D_s^h} \\ 0.80 & \text{if } O_s = C_{s+D_s^h} \\ 0.01 & \text{if } D_s^h = - \end{cases} \qquad (32)$$

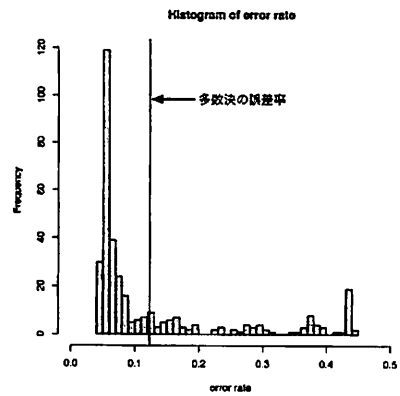

図7 $P(D_s^h, D_t^h)$ の変化による誤差率のバラツキ

図7から,$P(D_s^h, D_t^h)$ の値が分からなくても, 約70%の確率で多数決に比べ誤差率が減少する事がわかる. 今回の実験で最も誤差率が低くなったときの値は0.0423であり,95%以上が元の親配列と一致した. そのときの $P(D_s^h, D_t^h)$ の確率値を表1に示す.

4.4 パラメータ選択

今回の実験では真のパラメータが分からないときでも, 約7割の確率で多数決よりも良い親配列を推定できる事がわかった. 逆に言えばこれは約3割の確率で多数決よりも悪い親配列になる事を意味しており, 最悪の場合誤差率が45%になってしまう.

表 1 誤差率が最小のときの $P(D_s^h, D_t^h)$ の確率値 (– は Gap を表す)

D_s^h \ D_t^h	2	1	0	-1	-2	–
2	8.76e-1	6.00e-2	3.60e-3	2.16e-4	1.30e-5	6.00e-2
1	2.85e-2	8.44e-1	6.20e-2	3.72e-3	2.23e-4	6.20e-2
0	1.71e-3	2.86e-2	8.42e-1	6.21e-2	3.73e-3	6.21e-2
-1	1.03e-4	1.72e-3	2.86e-2	8.45e-1	6.21e-2	6.21e-2
-2	6.17e-6	1.03e-4	1.72e-3	2.86e-2	9.07e-1	6.21e-2
–	8.06e-2	1.21e-1	2.42e-1	1.21e-1	1.94e-1	2.42e-1

そこで, パラメータを適切に選ぶ必要がある. 今回は leave-one-out の Cross-Validation を用いてパラメータの変化と誤差率の変化を比較してみた.Cross-Validation では以下の式 (33) のような計算を行って, 誤差率を算出した.

CV 誤差率

$$= \sum_{i=1}^{10} \left\{ 1 - \frac{\text{子配列 } [i] \text{ と推定した配列の正答数}}{\text{子配列 } [i] \text{ 以外で推定した配列の長さ}} \right\} \quad (33)$$

その結果を図 8 に示す. 図 8 の index は $P(D_s^h, D_t^h)$ のみを少しずつ変化させたときの実験番号で,index が同じ箇所では同じ $P(D_s^h, D_t^h)$ を用いている. 縦軸は左が Cross-Validation による誤差, 右が真の誤差率を表している.

図 8 の結果より,$P(D_s^h, D_t^h)$ が同じ箇所で,Cross-Validation による誤差と真の誤差の値は異なるものの, 全体の概形はほぼ一致する. この事から,Cross-Validation によってパラメータ選択を行う事ができる.

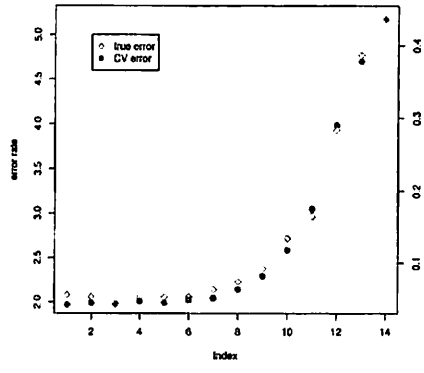

図 8 Cross-Validation と真の誤差の比較

4.5 考 察

今回の実験から, パラメータを適切に選ぶことができれば, 提案した突然変異モデルを用いることで, 多数決に比べてかなり良い親配列を推定する事ができることがわかった. さらに,Cross-Validation を事前に行っておく事で, より良いパラメータ選択も行える.

今回突然変異確率 $P_p(O_s, C_{s+D_t^h})$ に一様分布を用いた. これは, 実際のタンパク質データでは, 真の突然変異確率を推定する事が困難であるからである. 一様分布を用いてもかなり良い推定を行えた事から, ある程度ロバストな推定を行うことができると考えられる.

5. むすび

本報告では, 祖先が同一であるタンパク質のアミノ酸配列から, その祖先のアミノ酸配列を推定するための突然変異モデルを提案した.

さらに,Belief Propagation を事前分布を考慮できるように改変し, 提案モデルに適用できるようにした. その結果, 多数決に比べてかなり良い推定結果を得る事ができた.

ステレオビジョンではセグメンテーションやエッジなどで領域が分割できるとき, メッセージがその領域を超えて伝播されるのを制御することがあり, その制御に用いる情報を cue と呼ぶ [2]. タンパク質もドメインと呼ばれる構造での領域分割ができ, これを cue として導入してやる事により, 提案モデルをさらに改良する事が可能ではないかと考えている.

文 献

[1] S.Birchfield and C.Tomasi. "A Pixel Dissimilarity Measure That Is Insensitive to Image Sampling", IEEE transactions on pattern analysis and machine intelligence, vol.20, no.4, 1998

[2] J. Sun, N. Zheng and H. Shum. "Stereo Matching Using Belief Propagation", IEEE transactions on pattern analysis and machine intelligence, vol.25, no.6, 2003.

[3] J. S. Yedidia, W. T. Freeman and Y. Weiss. "Understanding Belief Propagation and its Generalizations", Technical Reports (Mitsubishi electric research labolatories), TR-2001-22, 2002.

[4] R. Durbin, S. R. Rddy, A. Krogh, G. Mitchison 著, 明渓遼也, 浅井潔, 矢田哲士訳. バイオインフォマティクス-確率もデルによる遺伝子配列解析-, 医学出版, 2001.

[5] 田中和之, 樺島祥介, 田中利幸, "確率・統計モデルが切り開く推論と学習の新しいパラダイム", 電子情報通信学会誌, vol.88, no.9, 2005

[6] 宮田勝 監修, 斉藤輪太郎 著. バイオインフォマティクスの基礎, サイエンス社, 2005.

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

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