# 過去のアミノ酸配列を推定する手法の開発：
# 進化過程反転による系統樹全ノードのアミノ酸配列予測

理学専攻・情報科学コース
g1640652　高野伶美

## 1　はじめに

ゲノムとは，細胞を構成するタンパク質などの鎖状高分子の設計図が書き込まれている生体高分子である．多くの生物において，この化学的実態は DNA である．DNA も 4 種類の塩基から成る鎖状高分子であり，この塩基の並び順そのものが暗号化されたタンパク質の設計図になっている．

DNA は様々な要因により，変異を起こしたり損傷されたりすることがある．DNA の変異はアミノ酸配列の変異に繋がり得る．こうした変異が子孫に受け継がれ，長い時間をかけて集団にもたらされた遺伝的構成の変化を，進化と呼ぶ．

近年，多様な生物のゲノムが数多く決定されてきたことにより，全生物種において，遺伝の仕組みやタンパク質の合成等の生物学の一般原理が当てはまることが明らかになった．このことから，すべての生物は共通祖先から進化してきたと考えられている．

共通祖先の配列や系統関係を知ることは，生物学の様々な分野において有益である．しかし，基本的に進化の過程の配列は残っておらず，直接解析することは不可能である[1].

そのため，観測された遺伝的形質からそれらを予測する手法が開発されてきた．遺伝的形質の一つとしてタンパク質が挙げられる．タンパク質とは，20 種類のアミノ酸を組み合わせて構成される鎖状高分子である．タンパク質を構成するアミノ酸配列を対象とするデータ群とした場合，アミノ酸配列が類似しているものほど近縁関係にあると仮定して系統樹を描くことができる[2]．これは，進化にかかった時間とタンパク質を構成するアミノ酸配列の相違度がおおよそ比例するという性質に基づいている．

アミノ酸に起こる変異には置換・挿入・欠失がある．図 1 では上段のアミノ酸配列から下段のアミノ酸配列への変異を示している．アミノ酸はそれぞれ一文字のアルファベットで表されるため，アミノ酸配列は文字列として表記することができる．配列内でのアミノ酸の位置をサイトという．オレンジ色の矢印で示したサイトはアミノ酸 M がアミノ酸 L になる置換を示している．また，青い矢印で示したサイトでは上段の配列に存在したアミノ酸 I が下段では欠失しており，緑の矢印で示したサイトでは逆に，上段にはなかったアミノ酸 Q が下段では挿入されている．

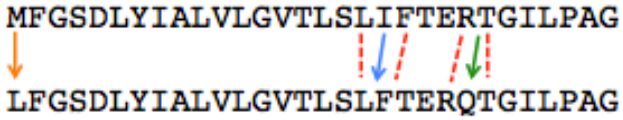


図 1:アミノ酸配列に起こる変異．矢印で示したサイトに変異が起こっており，左から置換・欠失・挿入．

アミノ酸配列群から系統樹を推定し，各進化過程でのアミノ酸配列を予測する手法として ANCESCON がある[3].

本研究では，動的計画法を用いて系統樹を描画し，そこから各ノードにおけるアミノ酸配列，つまり各進化段階におけるアミノ酸配列を予測する手法を提案する．また，本手法の適用により，ANCESCON よりも高い精度で過去のアミノ酸配列を推定することを目指す．

## 2　手法

### 2.1　系統樹の作成

#### 2.1.1　動的計画法を用いたアラインメント

アミノ酸配列間の類似度はその一致度として求めることができるが，挿入や欠失により，対応するサイトがずれていることがあるため，単に同じサイト同士を比較することはできない．このため，まずはアミノ酸配列間のサイトを対応付ける必要がある．そのような手法としてアラインメントがある．

アラインメントは動的計画法を用いて実装することができる[4]．配列 1 のサイト $i$ までと配列 2 のサイト $j$ までを並べたときのアラインメントスコア $F(i,j)$ は，$F(1,j)=\delta_{x_1 y_j}$　$F(i,1)=\delta_{x_i y_1}$ として，ギャップ定数 $d$ を用いて

$$F(i,j)=max\begin{cases}F(i-1,j-1)+\delta_{x_i y_j},\\F(i-1,j)-d,\\F(i,j-1)-d.\end{cases}$$

という式で表される．配列 1 の長さが $N$，配列 2 の長さが $M$ としたとき，$max(F(N,j),F(i,M))$を配列全体のアラインメントスコアとする．このアラインメントスコアとして採用された座標が対応する最後のサイトとしてバックトラックすることでアラインメントすることができる．$d$ を含む二式は，対応するアミノ酸がないことを示しており，アラインメントの際にはギャップ文字'-'で表す．配列 1，配列 2 をアラインメントした結果，図 2 に示したような，長さの同じ配列 a 1，a 2 が生成される．


図 2: 配列 1 と配列 2 をアラインメントすると，同じ長さのアラインメント配列 a1，a2 が生成される．赤字で示したものはアラインメントの結果挿入されたギャップ文字.

#### 2.1.2　アラインメントスコアによる系統樹の推定

対象となるデータ群を総当たりでアラインメントし，アラインメントスコアを比較することで，最も似ている配列

のペアを，最も近縁であるとして結び，親ノードを作ることを繰り返して系統樹を推定する．

次の近縁を求める際は，結ばれた二つのノードの配列の代わりに，親ノードの配列を用いて総当たりを行う．

### 2.2 親ノードのアミノ酸配列の仮決定

アラインメントされた配列 a1，a2 を用いて，親ノードにおけるアミノ酸配列を推定する．a1 と a2 のサイト n を比較し，同じであれば親ノードのサイト n も同じ文字であるとし，異なれば決めることができないということで図 3 のように候補として両方の文字を保持する．

a1，a2 の少なくとも一方がサイト n で候補を持っている場合，a1 と a2 のサイト n における候補の中に重複があれば，それを親ノードのサイト n として採用する．重複がなければ全てを候補として保持する．

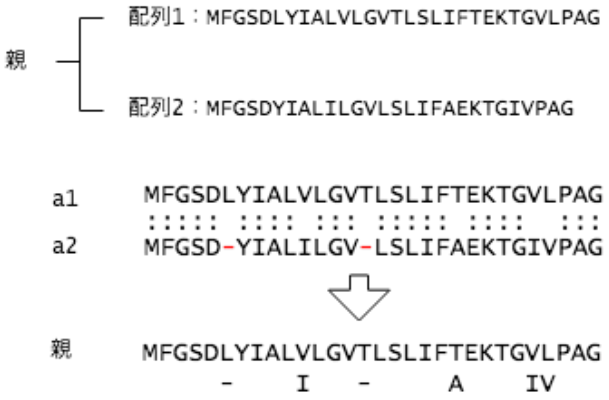


図 3: 配列 1 と配列 2 から予測したノード「親」におけるアミノ酸配列．アラインメントの結果アミノ酸が一致しているサイトを重点で示した．重点がないサイトは親において候補を持っている．

### 2.3 各ノードにおけるアミノ酸配列の尤度の決定

各ノードにおいて，候補を持つサイトは子孫ノードで候補のアミノ酸を持っている．つまり，分岐していく過程で候補のアミノ酸の全てに進化していくことになる．このことから，アミノ酸置換行列である PAM1 を用いて，進化の起こりやすさを算出した[5].

## 3 精度検証実験と結果

進化の起点とするアミノ酸配列と任意に形を決めた系統樹を用意し，PAM1 を用いて変化させることで子ノードのアミノ酸配列を決めることを繰り返し，擬似的な進化系統のデータを作成した．

このデータの変異後のアミノ酸配列を用い，本手法を適用することで，各ノードのアミノ酸配列を正しく予測できたか確認し，ANCESCON の適用結果と比較した．

この予測結果について，各ノードを，(1)葉を結んだもの，(2)中央付近で葉に近いもの，(3)根に近いもの，(4)根にあたるものの 4 種類に分類し，精度を検証した．

結果，4 種類のノードすべてで本手法が高い精度で予測することができていた．

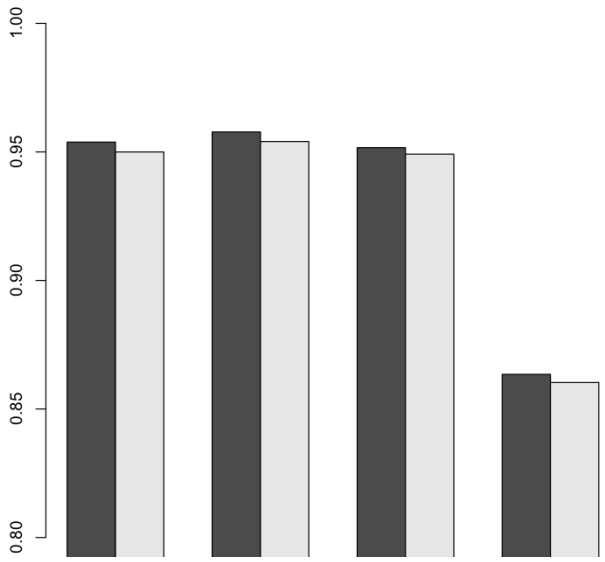


図 4 :本手法と ANCESCON の精度比較．黒いバーが本手法，グレーのバーが ANCESSCON

## 4 今後の課題

結果で挙げたように(4)の根にあたるノードでの精度が著しく落ちていたため，原因を検証し，精度向上を計りたい．

また，本研究の手法では文字列としてのアミノ酸配列の類似度のみを考えているが，本来タンパク質は立体構造を取ることで機能を持つものである．そのため，アミノ酸配列は直前，直後のアミノ酸との関係だけではなく，立体構造を取った時に近くに位置するアミノ酸の影響も受けるため，文字列としての小さな変化が機能的に大きな変化につながることがある．しかし，現在立体構造の判明していないタンパク質も未だ多く存在する．このため，離れた位置にあるアミノ酸と変異の相関関係をアミノ酸配列から構造の違いも予測することができれば，より本来の進化系統に近い，機能的に似ているものを近縁として予測することができるのではないかと考える．

## 参考文献

[1] Joseph W Thornton, Resurrecting ancient genes: experimental analysis of extinct molecules, Nature Reviews Genetics(2004), 5:366:375

[2] D.Sadava(Eds.), 大学生物学の教科書 第 4 巻 進化生物学 石崎泰樹・斎藤成也(訳), ブルーバックス(2014), pp.112-116

[3] Wei Cai, Jimin Pei, and Nick V. Grishin, Reconstruction of ancestral protein sequences and its applications, BMC Evol Biol.(2004), 17;4:33

[4] Needleman, Saul B. & Wunsch, Christian D. (1970)"A general method applicable to the search for similarities in the amino acid sequence of two proteins"JMB 48 (3): 443–53.

[5] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt (1997) Atlas of Protein Sequence and Structure 第5巻, *National Biomedical Research Foundation,* pp.348